

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Towards Multi-lingual Multi-modal Dialogue Systems

Permalink

<https://escholarship.org/uc/item/9d0155nn>

Author

Zhou, Mingyang

Publication Date

2022

Peer reviewed|Thesis/dissertation

Towards Multi-lingual Multi-modal Dialogue Systems

By

MINGYANG ZHOU

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Zhou Yu, Chair

Kenji Sagae

Premkumar Devanbu

Committee in Charge

2022

Copyright © 2022 by

Mingyang Zhou

All rights reserved.

*To someone very important . . .
a nice dedication.*

CONTENTS

Abstract	v
Acknowledgments	vii
1 Introduction	1
1.1 Overview	1
1.2 Ground Vision and Action:	2
1.3 Cross-lingual Cross-modal Representation Learning	3
2 Ground Vision and Action	5
2.1 Balance Language Generation and Policy Learning for Visual dialogueue Systems	5
2.1.1 Introduction	5
2.1.2 Related Work	6
2.1.3 Model	8
2.1.4 Learning	10
2.1.5 Experiments	12
2.2 Image Synthesis from Dialogue	18
2.2.1 Introduction	18
2.2.2 Related Work	20
2.2.3 Dataset	21
2.2.4 Model	27
2.2.5 Experiment	29
3 Cross-lingual Cross-modal Representation Learning	34
3.1 Augment Machine Translation with Vision	34
3.1.1 Introduction	34
3.1.2 Related Work	35
3.1.3 Model	37
3.1.4 Experiment	40

3.2	Universal Cross-lingual Cross-modal representation learning	45
3.2.1	Introduction	45
3.2.2	Related Work	46
3.2.3	Model	48
3.2.4	Result	52
3.3	Unsupervised Cross-modal Representation Learning	60
3.3.1	Introduction	60
3.3.2	Related Work	63
3.3.3	Model	64
3.3.4	Experiment	70
4	Conclusion and Future Work	77
4.1	Summary	77
4.1.1	Ground Vision and Action	78
4.1.2	Cross-Lingual Cross-modal Representation Learning	79
4.2	Future Directions	80
4.2.1	Learning with Less Supervision	80
4.2.2	Unifying Modalities for Multi-modal Agent	80
4.2.3	Lifelong Interactive Learning	81

ABSTRACT

Towards Multi-lingual Multi-modal Dialogue Systems

Building dialogue systems that can communicate with human is a vital challenge for artificial intelligence. Existing dialog systems, such as Amazon Alexa and Google Assistant, often only interact with human users in a single language, limiting their application to conduct situated conversations requiring visual perception or interacting with users speaking different languages. In this dissertation, we present our exploration on building multi-lingual grounded dialog system that can understand and interact with the world with information from various channels (vision and language) to solve real-world tasks. My research effort on building multi-lingual and multi-modal dialogue system is mainly divided into two different direction:

First, we explore how to facilitate an agent to connect vision and language to actions in an interactive environment. We proposes an alternative learning procedure between supervised learning and reinforcement learning to train task-oriented visual dialogue systems, which achieves better balance between dialog response quality and policy effectiveness. Then, we introduce the task of image synthesis from dialog, which combines visual grounded language generation and conditional text-to-image generation into a unified problem.

Second, we aim to build multi-modal intelligent agents that can communicate with people who speak different languages via connecting vision to multilingual texts. our first attempt in this direction is augmenting machine translation with images by learning visually grounded text embedding, which tackles the challenges of translating ambiguous words with text-only context. Next, we propose a multi-modal multi-lingual pre-training framework intending to learn task-acoustic cross-modal cross-lingual representation. Besides, we have also proposed a retrieval based unsupervised vision and language pre-training method to address the challenges of collecting parallel image and text corpus for robust cross-modal representation learning.

Finally we summarize the findings of the existing work and discuss the future plan to push multi-modal multi-lingual interactive AI agent research further.

ACKNOWLEDGMENTS

Looking back at the past five years., it has been a challenging journey to pursue a Ph.D. degree, both mentally and physically. However, I cannot imagine accomplishing it by myself without the help and support of many people. The first person I want to thank is my advisor Zhou Yu, who is the best advisor I could ask for. She allowed me to join UC Davis and conduct natural language processing research when I had no related previous research experience. Besides teaching me a lot about how to do good research, she also shapes me to become a better person that is more confident, more responsible, and more self-disciplined. I feel proud to be the first Ph.D. student under her supervision. I also want to thank Professor Yong Jae Lee, who has provided me with much valuable advice on computer vision techniques in my research. He also sets a great model for me to do a good presentation for research. I appreciate all the help from him. I appreciate for professor Kenji Sagae, and Premkumar Devanbua’s service as my Ph.D. dissertation committee. They provide valuable feedback to help me to accomplish this dissertation.

During my Ph.D., I am fortunate to collaborate with many talented minds on research. I want to thank my intern mentors at Microsoft: Luwei Zhou, Linjie Li, Yu Cheng, Shuohang Wang, and Jingjing Liu. Luwei helps me define the multi-lingual multi-modal pre-training project and keeps pushing me to polish the research to conduct impactful work. Linjie is my coding advisor at Microsoft. She always patiently help with all my detailed questions on training the big neural network models. I also want to thank Yu, Shuohang, and Jingjing for always providing great feedback on my weekly research progress and a great contribution to the paper writing. For my second internship at Facebook AI, I am fortunate to work with Licheng Yu, Amanpreet Singh, Mengjiao Wang, and Ning Zhang. Licheng is a very nice research mentor who provides valuable feedback to my internship project and teaches me how to do good AI research. Aman is very knowledgeable about conducting efficient pre-training and has helped me with engineering details when implementing the unsupervised vision and language pre-training framework. Mengjiao has always contributed brilliant ideas on our weekly meeting. Ning is very supportive whenever I have research or internship-related questions. I also want

to thank Anna Rohrbach and Grace Luo, who have had valuable discussions with me on the news image captioning research.

I am thankful for the experience of working with my teammates on the Alexa Prize Challenge, which is one of the best memory I have during my Ph.D.: Chun-yen Chen, Dian Yu, Weiming Wen, Yimang Yang, Jiaping Zhang, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Lyer, Giritheja Sreenivasulu, Ashwini Bhandare, Runxiang Cheng, Kaihui Liang, Xueyuan Lu, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen. Winning the trophy of the 2018 Alexa Prize is sweet, but the more valuable experience is to accomplish this journey with all of you. I will never forget the hackathons we go through, the funny dialogue we read, and the pizzas we had together.

I also owe my sincere thanks to my labmates and college friends at UC Davis for making my graduate life much easier with many fun chats, hot-pot gatherings, and board game nights: Kun Qian, Yu Li, Weiyan Shi, Qingyang Wu, Qingxiaoyang Zhu, Junheng Chen, Hai Yu, Minmeng Tang, Xinwei Li, Xuan Ding, Can Cui, and Yanqiu Zhou. Finally, I also thank my external friends: Yangdi Zhou, Xiaokai Sun, Yi Chen, Quan Huang, and Yuke Liao, for always cheering me up when I encountered difficulties during my Ph.D.

Finally, I cannot accomplish my Ph.D. without the constant support from my family. My dearest wife Yewetao Wu and my 2-year-old boy Allen Zhou, bring the brightest happiness into my life, which is a major motivation for me to go through hard times. My parents are another motivation for me as they set up great examples of good professors that always explore challenging research questions and share knowledge to the younger generation. In addition, my parents-in-law have traveled overseas multiple times during my Ph.D. to take care of our family, which makes the work-life balance much better especially when we have the new born. Last but not least, I want to thank the company of my three lovely cats: Saturn, Jupiter, and Rocket. I cannot go through the long night of code debugging without having them sit aside.

Chapter 1

Introduction

1.1 Overview

Having an intelligent assistant that can communicate with humans to serve their needs is a fundamental challenge in Artificial Intelligence (AI) research. Recently, owing to the development of deep learning techniques and the large scale datasets, we have witnessed a great advancement of dialogue systems. Nowadays, conversational agents have been deployed in millions of smart devices such as Alexa, Google home assistant, and Smart phones (e.g. Siri) to serve as the personal assistants or the chat companions for human users. Although tremendous success has been achieved, there are still major limitations. The majority of current dialogue systems can only process and communicate with language context, which limits their application to conversational tasks that require situational understanding such as language-guided visual navigation or fashion shopping assistant. Additionally, while there are more than 6500 different languages used in our world, the dialogue systems are mainly studied on English. In order to broaden the access of such AI techniques to non-English speakers, it is essential to build conversational AI agents that can communicate in multiple languages. To address these limitations, we aim to build multi-lingual multi-modal dialogue systems that learns to process context from multi-modal signals (vision and language) and communicate in various languages via interacting with real users. In this dissertation, we introduce our effort to approach this goal in two

different research directions:

- **Ground Vision and Action:** we build multi-modal dialogue systems that can ground conversations in a visual environment and adopt optimal actions to improve task success. we also collect a new benchmark that helps the dialogue system to learn cross-modal grounding via simultaneously handling vision generation from textual context and text generation from visual context in a unified conversational task. [Chapter 2]
- **Cross-lingual Cross-modal Representation Learning:** To enable dialogue systems become multi-lingual speakers, we conduct researches to align vision and various languages in a learned semantic space. Specifically, we research multi-modal machine translation and cross-lingual cross-modal pre-training techniques to learn joint representations across languages and modalities. we have also introduced how to learn robust universal cross-modal representation without parallel image-text pairs. [Chapter 3]

Below we give an overview of our past research as the initial exploration to build multi-lingual multi-modal dialogue systems via interactive learning.

1.2 Ground Vision and Action:

As humans, we communicate with other people in natural language while perceiving and taking actions. However, existing dialogue systems, such as Alexa and Google Assistant, only focus on language understanding and generation, limiting their capability to help humans in the visual world. My research aims to facilitate the dialogue system with the ability to “see” and “act”, such that it can be applied to some more situated real-world tasks (e.g, blind people navigation assistant and online shopping assistant). Specifically, we situate the interactive agent in game settings such that the agent needs to hold a meaningful conversation with human players over visual concept in order to accomplish predefined goals.

Our first work addresses the challenge of training neural sequence-to-sequence framework to jointly learn between dialogue policy and language generation. We explore this problem

in a image guessing game. Unlike traditional methods that learn the dialogue policy and language generation on language decoder, we separate the action space for dialogue policy learning from that of language generation. Our training curriculum have lead to better balance between the language quality and task successful rate.

In the second work, we introduce a new task called GanDraw which is a collaborative game that combines language to vision generation and vision to language generation into a unified test bed. With this dataset, we hope to inspire the creation of the next-generation of interactive AI agent that can simultaneously handle various types of cross-modal grounding with a unified framework.

1.3 Cross-lingual Cross-modal Representation Learning

The world we navigate through is a multi-modal and multi-lingual kaleidoscope. Besides augmenting the dialogue system to “see” and “act”, it is also valuable to make the dialogue system a multi-lingual “speaker” to better serve different language users. However, the existing study on multi-modal interactive intelligent systems are biased towards English. To better serve people with different linguistic backgrounds from our global community, it is critical to learn cross-lingual cross-modal alignment. Given that vision is the shared signal perceived by everyone in the world, we are dedicated to study how to connect different languages by grounded them onto vision.

The initial effort in this direction is spent on introducing vision to augment neural machine translation. Our proposed visual-attention grounding mechanism helps to learn more accurate representations of salient words that are semantically associated to images, and thus lead to better translation quality. Next, we take a step further to learn general task-acoustic cross-lingual cross-modal representation via pre-training from multilingual image-text pairs. The pre-training framework demonstrate superior performance over task-specific methods on several benchmarks on multilingual vision and language tasks. Connecting vision and multi-lingual text requires large amount of parallel image-text pairs which is expensive to collect. We propose unsupervised unsupervised vision and language

pre-training to learn robust cross-modal representation from unpaired image and text corpus.

Chapter 2

Ground Vision and Action

2.1 Balance Language Generation and Policy Learning for Visual dialogue Systems

2.1.1 Introduction

Visually-grounded conversational artificial intelligence (AI) is an important field that explores the extent intelligent systems are able to hold meaningful conversations regarding visual content. Visually-grounded conversational AI can be applied to a wide range of real-world tasks, including assisting blind people to navigate their surroundings, online recommendation systems, and analysing mass amounts of visual media through natural language. Current approaches to these tasks involve an end-to-end framework that maps the multi-modal context to a deep vector in order to decode a natural dialogue response. This framework can be trained through supervised learning (SL) with the objective of maximizing the distribution of the response given a human-human dialogue history. Given a large amount of conversational data, the neural end-to-end system can effectively learn to generate coherent and natural language.

In this work we focus on building neural model to tackle task-oriented visual dialogue system, where we situate the agent in an image guessing task [1] to evaluates the model's ability to retrieve visual content via conversing in natural language. To obtain an optimal dialogue policy, reinforcement learning (RL) is introduced to enable the neural end-to-end

framework to model a more effective action distribution by exploring different dialogue strategies. A typical way to apply RL on a dialogue system is to assign a task-related reward to influence the utterance generation process by treating each output word as the action step. However, A significant limitation of this approach is that it is difficult to achieve an optimal dialogue policy that can both effectively complete the external goal and generate natural utterances.

we propose a novel learning curriculum to address the challenge of joint learning between the dialogue policy and language generation for task-oriented dialogue systems. In our framework, we separate the training of the image retrieval policy from dialogue generation by applying RL, with the goal of achieving an optimal policy for guessing the target image at every turn. In addition, we apply a language model objective function to optimize the utterance generator to mitigate language degeneration. We specifically study this framework in the image guessing task, GuessWhich, where a conversational agent attempts to guess a target image by asking a series of questions. When compared to state-of-the-art RL visual dialogue systems, our method achieves superior performance in both task-accomplishment and dialogue quality.

2.1.2 Related Work

Visual dialogue systems Visual dialogue systems are an emerging area of interdisciplinary research that attracts both the vision and language communities due to the potential applications. [2] proposed a visual dialogue task in which a conversational agent attempts to answer questions regarding an assigned image based on a dialogue history. To approach this task, they initially collected data by having two people chat about an image with one person acting as the questioner and the other as the answerer. GuessWhich [1] extends VisDial with the goal to build an agent that learns how to identify a target image through questions and answers. [3] additionally introduced a game in which a series of yes-or-no questions are asked by an agent in order to locate an object in an image. Many researchers approached these tasks via reinforcement learning (RL) with the goal of obtaining an optimal dialogue policy. [4], for example, designed three rewards with respect to the goals of task achievement, efficiency, and question informativeness, in order

to help the agent achieve an effective question generation policy for the GuessWhat game. [5] applies reinforcement learning in the GuessWhich task and demonstrates a moderate improvement in accuracy compared to the supervised learning approach. Both methods apply RL on a neural end-to-end pipeline to jointly influence the language generation and dialogue policy. Due the challenge of designing an appropriate reward for language generation, these methods generate responses that deviate from human natural language. [6], proposed an approach involving hierarchical reinforcement learning and state-adaptation techniques that enable the agent to learn an optimal and efficient multi-modal policy. The bottleneck of [6]’s method, however, is that the system response is retrieved from a predefined human-written or system-generated utterance. The number of predefined responses are limited, therefore, this method does not easily generalize to other tasks in real-world settings. We address these limitations by applying RL on a reduced, yet more relevant action space, while optimizing the dialogue generator in a supervised fashion. We alternatively optimize policy learning to language generation to combine the two tasks together.

RL on Task-oriented dialogue System Various RL-based models have been proposed to train task-oriented dialogue systems [7]. In order to build a traditional modular-based dialogue system, researchers first identify the semantic representation, such as the dialogue acts and slots in user utterances. Then they accumulate these semantic representations over time to track the dialogue state. Finally they apply RL to learn an optimized dialogue policy given the dialogue state [8, 9]. Such modular-based dialogue systems are effective in narrow task domains, such as searching a bus route schedule or reserving a restaurant through natural language, but they fail to generalize to complex settings where the size of the action space increases. Owing to the development of deep learning, RL on neural sequence-to-sequence models has been explored in more complex dialogue domains such as open-domain conversation [10] and negotiation [11]. However, due to the difficulty of assigning appropriate rewards when operating in a large action space, these frameworks cannot generate fluent dialogue utterances. [12] proposed a novel latent action RL framework to marry the advantage of a module based approach and

sequence-to-sequence approach. They learned the optimal dialogue policy in a complex task-oriented dialogue domain while achieving decent conversation quality. Here, we study a similar issue in a multi-modal task-oriented dialogue scenario. We propose an iterative approach using RL to optimize the dialogue policy and SL to optimize the generation of the system response.

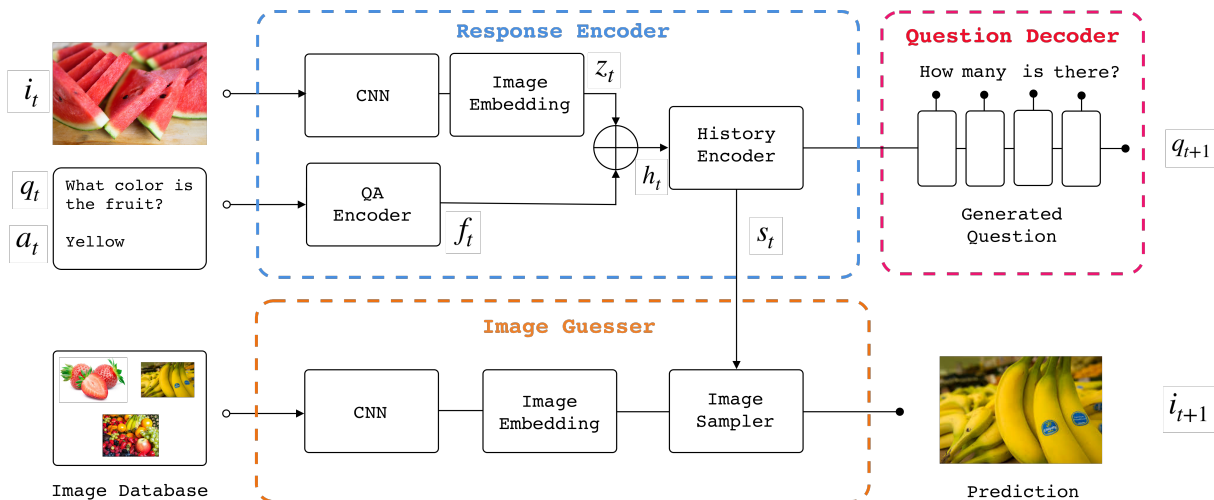


Figure 2.1: The proposed end-to-end framework of the conversation agent for GuessWhich task-oriented visual dialogue task

2.1.3 Model

Problem Setting

In the GuessWhich problem, we aim to build an agent (Q-Bot) that attempts to guess an image i_{tgt} that another agent (A-Bot) knows by asking it a series of questions. At the beginning of the conversation, the Q-Bot is primed with a short caption c of the target image that is only known by A-Bot. At every round t , the Q-Bot generates a question q_t to elicit as much information as possible about the target image and the A-Bot provides an appropriate answer i_t with regard to q_t and the target image. In the end, the agent guesses the target image among a set of images considering the entire conversation. In addition, our dialogue system also guesses a candidate image i_t out of an image database $\mathcal{I} = \{i_k\}_{k=0}^m$ at every turn. This action models the process of sequentially updating the visual belief state on the target image based on the latest dialogue history. Conditioned on the current guessed image and the prior dialogue contexts, the system will generate an

optimal question in order to get the maximum information from A-Bot that can strengthen the system’s belief on the target image. At the end of the conversation, our Q-Bot will guess the target image based on the multimodal contexts $s_n = (q_{1:n}, a_{1:n}, i_{1:n}, c)$ consisting of the dialogue history and the trajectory of guessed images.

Model Architecture

Our Q-Bot is constructed on top of a hierarchical encoder-decoder framework [13], which consists of three major components: The **Response Encoder**, the **Question Decoder**, and the **Image Guesser**.

Response Encoder The goal of the response encoder is to append the question q_t , the answer a_t , and the guessed image i_t received at current round to the dialogue history and obtain an updated vector representation of the multimodal context s_t . The image i_t is encoded with a pre-trained convolutional neural network VGG-16 [14] followed by a linear embedding layer and the image feature vector denoted as z_t . For the question and answer pair at the current round (q_t, a_t) , we map them to a hidden state vector f_t through the LSTM based *QA Encoder*. We then apply a linear projection on the concatenation of f_t and z_t in order to obtain the multi-modal context vector h_t for the current round. The context vector is then passed through another LSTM encoder: *History Encoder* generates an updated dialogue history representation $s_t = \text{HistoryEnc}(h_t, s_{t-1})$. We denote the trainable parameters for Response Encoder as θ_e .

Question Decoder The question decoder is a two-layer LSTM network initialized with the most updated dialogue history representation vector s_t from the response encoder. It will sequentially sample the words to come up with the next question q_t . The learned parameters for question decoder are denoted as θ_d .

Image Guesser The Image Guesser attempts to identify the candidate image that best aligns with the dialogue history. Given an image database $\mathcal{I} = \{i_k\}_{k=0}^m$ where we sample the candidate image, we first extract the image feature representations $\{z_k\}_{k=0}^m$ for all candidate images with the convolutional neural network and image embedding layer defined in response encoder. Then, we can sample a candidate image i_k for the current turn based on the euclidean distance $d(z_k, s_t)$ between the image feature of the candidate

image and the current dialogue history vector. The image with the smallest euclidean distance is selected as the guess i_t at the current round. The associated parameters for image guesser are defined as θ_g .

2.1.4 Learning

We follow a two-stage training fashion as introduced in many previous end-to-end RL dialogue systems [5, 4, 12], where we first pre-train the dialogue framework with a supervised objective then apply reinforcement learning to learn an optimal policy to retrieve the target image. The Supervised pre-training is a critical step that facilitates an effective policy exploration for RL training, as it is difficult to explore a complex action space with limited prior knowledge. During RL training, we introduce an alternative learning method between dialogue policy exploration and natural utterance generation that addresses the issue of language degeneration in previous RL based visual dialogue systems [5]. We introduce each training method as follows.

Supervised Pre-training

During the supervised pre-training process, we jointly optimize the objective to generate questions and also predict target image features from dialogue contexts. The task of question generation is optimized by maximizing the log conditional probability of the next question dependent on a ground truth dialogue for every round of the conversation. For the image feature prediction, we minimize the mean square error (MSE) between the target image feature z_{tgt} and the dialogue context vector s_t at each round. The joint loss function for supervised pre-training is:

$$\mathcal{L}_{SL}(\theta_r, \theta_d, \theta_g) = \alpha \sum_{t=0}^n \log p(q_t | s_t) + \beta \sum_{t=0}^n \text{MSE}(z_{tgt}, s_t) \quad (2.1)$$

Where α and β are weights assigned to the objective function of each task in the joint objective function. With SL pre-training process, the dialogue system is facilitated with the ability to estimate a visual object and emit a natural language sentence given a dialogue context.

Reinforcement Learning on Image Retrieval

In our framework, we treat the sequence of image guess through out the conversation as a partially observable markov decision process and train a policy network via RL to obtain an optimal strategy to retrieve the target image. We formally describe state, policy, action, rewards, and the training procedures in our pipeline.

State The dialogue states in our framework consist of a combination of multimodal contexts, including the image caption c , the dialogue history with A-Bot $[q_1, a_2, \dots, q_t, a_t]$, and the image guessing trajectories $[i_1, i_2, \dots, i_t]$.

Policy The dialogue policy $\pi_{\theta_r, \theta_g}(i_t|S_t)$ is a stochastic policy that samples the candidate image to guess from an image set based on the previous dialogue histories. The policy is learned from response encoder and image generator which is parameterized via θ_r and θ_g .

Action The full action space is the number of images in the database that we can sample to guess an image. As the pre-trained process already enables the system to approximate a target image feature z_{tgt} with the dialogue history representation vector s_t , we reduce the action space to the top-K nearest images, s_t , based upon the euclidean distance. The probability to sample an image i_j is gained by applying a softmax function over the top-K candidates on their distance to s_t : $\pi(j) = \frac{e^{-d_j}}{\sum_{k=1}^K e^{-d_k}}$. d_j represents the mean-square-distance between the j -th image and the dialogue history state vector s_t .

Rewards We use the ranking percentile of the target image with respect to the dialogue history vector s_t as the reward signal to credit the guess at each turn. The goal is to maximize the expectation value of the discounted return $\mathbb{E}[\sum_{t=1}^n \gamma^t r_t]$ over the n-round conversation. r_t is the ranking percentile of target image at round t and γ is the discounted factor between $(0, 1)$.

Training Procedure Inspired from the RL training process on the iterative image retrieval framework [15], we apply the policy improvement theory [16] to estimate an improved policy $\pi^*(s_t)$ from an existing policy $\pi(s_t)$ obtained from the pre-trained dialogue system. Given a dialogue state s_t and the action a_t derived from the existing policy, the value estimated by the current policy for taking the action i_t is $Q_\pi(s_t, i_t) = \mathbb{E}[\sum_{t'=t}^n \gamma^{t'} r_{t'}]$.

To improve this, we explore a different action $i_t^* \neq i_t$ such that a larger policy value $Q_\pi(s_t, i_t^*) > Q_\pi(s_t, i_t)$ estimated with the current policy is achieved. Then we can adjust the existing policy $\pi(s_t)$ to a new policy $\pi^*(s_t)$ that executes that optimal action i_t^* given the current dialogue state. The parameters of the policy can be effectively optimized via a cross entropy loss function conditioned on the derived optimal action i_t^* :

$$\mathcal{L}_{RL}(\theta_r, \theta_g) = \mathbb{E}\left[-\sum_{t=1}^n \log(\pi_{\theta_r, \theta_g}(i_t^* | s_t))\right] \quad (2.2)$$

Compared to the previous RL visual-grounded conversational agent, [5], there are several advantages of conducting policy learning on the action level of guessing the image. First, the action space of the top-k nearest neighbors are much smaller compared to the vocabulary size of the output words which reduces the difficulty to explore optimal strategies. Second, only the parameters of response encoder and image guesser will be optimized during the RL training stage. The question decoder stays intact so that it is less likely for the dialogue system to suffer from language deviation.

2.1.5 Experiments

Evaluation Setting

We evaluate our visual dialogue systems in two different settings: **AI-AI setting** and **Human-AI** setting.

AI-AI Setting We evaluate the performance of our task-oriented dialogue system by playing the image guessing game, GuessWhich, with an automatic answer bot. Our conversational agent’s goal is to locate the target image out of the 9,628 test images by interacting with the other player in five conversation exchanges. We evaluate agent on both goal achievement and utterance generation quality using two automatic evaluation metrics Percentile Mean Rank (PMR) and perplexity respectively. PMR estimates how good the agent can rank the target image against other candidates in the test database based on its current dialogue state. Perplexity estimates the closeness of the generated response to a reference utterance given a dialogue context from the VisDial dataset.

Human-AI Setting realistic conversational scenario, we also make our agent play the image guessing game with human users. The games are set up as 20-image guessing games

where the agent attempts to guess a target image outside of a pool of 20 candidate images by asking a human player 5 rounds of questions. The objective of the human player is to play the role of answer bot and answer agent’s question with respect to the target image. In this setting, the performance of the agent on task accomplishment is evaluated by the game win rates. The quality of the dialogues are manually rated on four criteria: fluency, comprehension, diversity and relevance. Fluency defines the naturalness and readability of the generated question in English. Comprehension represents the consistency of the generated question with respect to the previous dialogue context. Diversity evaluates the uniqueness of the questions generated within one game. Relevance presents how well the asked question is related to the target image and the given caption.

In order to evaluate the effectiveness of the model, we designed three human evaluation tasks. Six college students were recruited to conduct the evaluation. Each student evaluated 100 games using the ground truth captions and 30 games using human generated captions. An additional three evaluators each completed 30 rounds of the relevancy experiment.

Ground Truth Captions We generated 100 image guessing games that used the ground truth captions to ensure a consistent amount of information is supplied across all human evaluators. Each game consists of a randomly selected set of 20 images from the VisDial Dataset, with one image randomly chosen as the target. For each game, we test three different models, each twice, resulting in a total of 600 evaluated games from the 100 generated games. We keep the identity of the models anonymous to the evaluator.

During each game, the human evaluator is presented with a target image the agent is trying to guess. Five rounds of Q&A take place in which the agent asks a question to elicit information and the human evaluator responds with a relevant truthful answer. At the end of each game, the evaluator is asked to rate the conversation on four criteria: fluency, relevance, comprehension and diversity.

Human Generated Captions In order to distinguish SL-Q-IG and RL-Q-IG in a more natural setting, we generate an additional 30 games, similar to the previous human evaluation task, except when beginning the game, the evaluator is asked to provide the caption

for the target image instead of using the ground truth.

Relevance Experiment We noticed that the human evaluators found rating dialogues on the relevance criteria challenging and nuanced. In order to reduce the difficulty of rating dialogues using the relevance criteria, we designed a separate experiment in which, using the conversations obtained from the previous 600 evaluated ground truth games, a human evaluator is presented with three complete conversations side by side at each round. The evaluator then selects the most relevant conversation out of the three that corresponds to the image caption. Each of the three conversations have the same caption, however, correspond to a different model, thus allowing for an effective comparison between the relevancy of each model.

Baseline Models

We compare the performance of our model with state-of-the-art task-oriented visual dialogue systems. Meanwhile we also perform an ablation study to evaluate the contribution of different designs in our framework. We introduce each model as follows:

SL-Q: The dialogue agent from [5], which is trained with a joint supervised learning objective function for language generation and image prediction.

RL-Q: The dialogue agent from [5] which is fine-tuned on a trained SL-Q by applying RL to the action space of output word vocabulary.

SL-Q-IG: The dialogue agent from this framework is build on top of the SL-Q. Compared to SL-Q, SL-Q-IG has an additional image guesser module that makes a guess on target image at every round. SL-Q-IG also has an image encoder which fuses the guessed candidate image into the dialogue history tracker. We only train this model with the supervised learning objective introduced equation 2.1.

RL-Q-IG: We use RL method to fine-tune SL-Q-IG. The RL method used is applied on action space of guessing candidate image. We alternate the model to optimize towards dialogue policy learning and language generation.

RL-Q-IG-NA: We fine-tune SL-Q-IG by applying RL to the action space of guessing candidate image and only optimized with policy learning objective function alone.

RL-Q-IG-W: The dialogue agent from our framework, which is fine-tuned on a trained

SL-Q-IG by applying reinforcement learning on output word vocabulary. It follows the same training procedures as RL-Q to conduct policy learning.

All the SL dialogue agents are trained on the VisDial Dataset with the default setting from [5] for 40 epochs. The RL dialogue agents are then fine-tuned on their corresponding SL dialogue agents for another 20 epochs. We evaluate every model on AI-AI image guessing games with the same answer bot, trained on the Visdial Dataset with the objective of visual question answering. We only evaluate RL-Q, SL-Q-IG and RL-Q-IG in human evaluation.

Result

Model	PMR	Perplexity
SL-Q	90.07%	79.49
SL-Q-IG	96.09%	61.42
RL-Q	94.78%	544.97
RL-Q-IG	96.81%	54.66
RL-Q-IG-NA	96.88%	363.88
RL-Q-IG-W	96.65%	227.35

Table 2.1: RL-Q-IG-NA performs best in PMR and RL-Q-IG perform best in perplexity

Results on AI-AI Image Guess Game It is clear from Table 2.2 that our dialogue system significantly outperforms the baseline models from [5] in terms of PMR on every round of the dialogue. PMR estimates how good the agent can rank the target image against other candidates in the test database. The biggest improvement gap is observed between SL-Q-IG and SL-Q. In comparison to SL-Q, SL-Q-IG tracks the additional context from the previously guessed images which leads to a better estimation of the target image. RL-Q-IG has better performance compared to SL-Q-IG in terms of PMR. This suggests that fine-tuning dialogue systems with RL can further improve the success of guessing the correct image. The best image retrieval result is achieved by RL-Q-IG-NA, as the objective function of RL-Q-IG-NA is based solely on policy learning without consideration for the dialogue generation quality. Although our framework achieved an

improved image retrieval accuracy, we observed, however, that there is little improvement gained in PMR after additional rounds of conversation. We suspect this is partially due to the fact that images from MSCOCO are composed of a diverse selection objects and background scenes, thus making images easily distinguishable with a detailed caption. In cases where candidate images are visually similar or the given caption is not informative, additional rounds of dialogue are necessary to identify the target image.

Model	Win	Fluency	Relevance	Comprehension	Diversity
RL-Q	59.6	4.19	3.22	2.60	2.50
SL-Q-IG	62.7	4.18	3.96	3.18	3.22
RL-Q-IG	67.5	4.40	4.02	3.50	3.25

Table 2.2: Evaluation results on the human-AI image guessing game initialized with ground truth captions

Model	Win	Fluency	Relevance	Comprehension	Diversity
RL-Q	29.2	4.04	2.88	2.71	2.29
SL-Q-IG	40.6	4.16	3.19	2.75	2.69
RL-Q-IG	67.6	4.23	3.74	3.32	3.06

Table 2.3: Evaluation results on the human-AI image guessing game initialized with human generated captions

While achieving higher image retrieval accuracy, we also observe a marginal increase of perplexity from SL-Q to RL-Q in Table 2.1, thus demonstrating that there is a bottleneck when applying RL to improve the language generation. By decoupling the policy learning from the language generation and alternatively optimizing the dialogue policy and language model, our RL-Q-IG avoids language deviation while still achieving an optimal dialogue policy for the image retrieval task. To further evaluate the contribution from the RL and alternative training curriculum, we conduct two ablation studies. RL-Q-IG-NA is fine-tuned with a policy learning objective that excludes alternatively applying the language model loss. While RL-Q-IG-NA only achieves an incremental improvement over the full framework RL-Q-IG in terms of the PMR rate with less than 0.1%, it suffers from

a dramatic increase of perplexity from 61.42 to 363.88, thus suggesting that alternatively applying the supervised learning objective can prevent the language model from deviating from the human language distribution. We additionally apply policy learning on the question decoder of SL-Q-IG and follow the RL fine-tuning process in [5] to train the agent, RL-Q-IG-W. While applying word-level RL enables RL-Q to achieve a moderate improvement over SL-Q in terms of PMR, we did not observe, the same degree of advantage in RL-Q-IG-W over SL-Q-IG. Additionally, RL-Q-IG-W is affected by a marginal increase in perplexity in comparison to the SL pre-trained agent, which approves the drawbacks of applying RL on a large action space in language generation.


Image	RL-Q-IG	SL-Q-IG	RL-Q
	Q: are bananas in bunch?	Q: what color are planes?	Q: is this outside?
	A: yes it's in bunch	A: no planes there	A: yes it is
	Q: are they on table?	Q: are they in city?	Q: is there any people?
	A: yes they are	A: yes it is	A: yes ...
	Q: is this in kitchen?	Q: are they in baskets?	Q: is this in zoo?
	A: I'm not sure	A: yes they are	A: no it is not
	Q: any people?	Q: any people?	Q: are there any people?
	A: Yes there are several people	A: yes there are	A: yes there are
	Q: any other fruits?	Q: animals?	Q: is there any people?
	A: NO, only bananas	A: no animals	A: yes ...

Table 2.4: A dialogue example with the ground truth caption: **bunches of bananas hang on a wall and arranged for sale**. **blue** indicates ideal relevant questions and **orange** indicates less relevant questions.

Results on Human-AI Image Guess Game The performance of a dialogue agent evaluated with a user simulator does not necessarily reflect its performance on human users [3]. We conduct human evaluation on different dialogue agents. From the results summarized in Table 2.2 and Table 2.3, we observe a consistent optimal performance of our method from conversations with AI agent to conversations with real human. Our RL-Q-IG significantly outperforms the baseline RL agent in all criteria for both settings. RL-Q-IG’s advantage over SL-Q-IG is not significant in the game when agents are primed

with ground truth image caption. This observation correlates with the result in the Human-AI game, as both RL-Q-IG and SL-Q-IG achieve superior PMR over 96% when presented with the ground truth caption. However, if a human generated caption is given, the performance of the SL pre-trained agent suffers a big drop in all metrics except fluency while our RL agent maintains similar performance. Applying RL to fine-tune the dialogue system enables the agent to generate more consistent dialogues in unseen scenarios. We also notice a degradation of the baseline RL agent from its performance with the user simulator, which suggests deviation from natural language is due to the sub-optimal RL training on a large action space.

Besides a marginal improvement over the RL baseline model and SL pretrained agent in terms of decreased repetition and grammar mistakes, there is a distinct superiority in regards to the relevance to the image caption in the questions generated from our RL agent. For example, in Table 2.4, we demonstrate the three dialogues generated by RL-Q-IG, SL-Q-IG and RL-Q on one game. Given the image caption *bunches of bananas hang on a wall and arranged for sale.*, RL-Q and SL-Q-IG ask very general questions that are not related to the caption such as “*planes*”, “*zoo*” and “*animals*”. In comparison, our agent asks high-quality questions regarding the caption that covers “*bananas*” and “*fruits*”. These questions help our RL agent obtain useful information to guess the target image. We credit the positive result to the dialogue policy, which explores multiple paths to conduct the conversation. The optimal path will involve a set of questions that obtains the maximum information of the target image such that it can construct the best estimation of the target image.

2.2 Image Synthesis from Dialogue

2.2.1 Introduction

Building an interactive intelligent system that can communicate with humans to form a shared understanding of a rich visual scene can lead to a wide range of applications, such as understanding the surroundings of a robot in a human inaccessible region through natural language conversations or designing fashion products based on users’ language guidance.

Existing research on visual dialogue systems [17, 3] and iterative conditional text-to-image generation [18, 19] attempt to resolve only half of the above problem with the focus on the single-directional grounding from one modality to another in an interactive scenario. We take a step further to propose a new task that integrates both research fields to a unified testbed that simulates the full process of constructing a "shared view" through natural language conversations between two parties.

We propose GanDraw, a collaborative drawing game to study how to construct a shared understanding of a partially observable visual concept through dialogue. The game is played by two players, a Teller and a Drawer. The goal of the game is for the Drawer to recreate the target image perceived by the Teller, through an exchange of conversations. An example of this task is demonstrated in figure 2.2. To succeed in this game, the teller must accurately ground the visual scene of the target image into multiple rounds of description and the drawer needs to fully understand the teller’s instruction and apply corresponding changes to the drawing. As describing a real image with rich scene accurately in natural language is quite challenging, the drawer needs to carry out an effective conversation with the teller by learning how and when to ask appropriate questions in order to clarify unclear instructions and missing details.

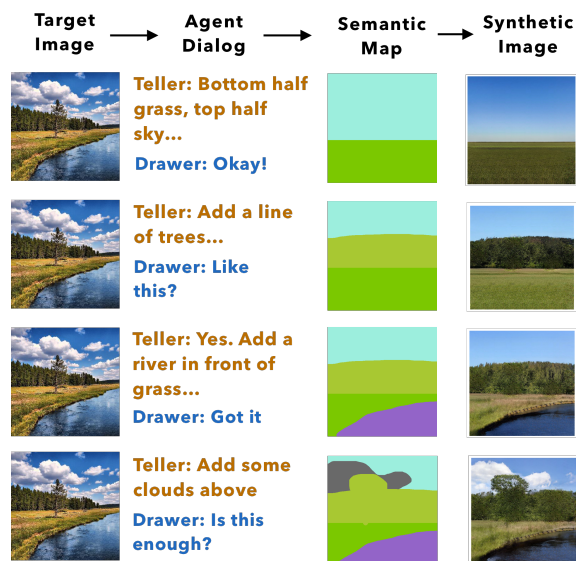


Figure 2.2: Overview of the new multi-modal visual dialogue task that we propose. In this task, two participants cooperate to recreate a realistic image by communicating in natural language.

2.2.2 Related Work

Visually-Grounded Language Generation The task of the teller is closely related to visual-grounded language generation. Applications for generating language with a single visual-context input include image captioning [20, 21], video description generation [22, 23], and visual storytelling [24]. Visual question answering (VQA) [25] takes a step toward building a collaborative agent with one round of human-machine interaction, where the agent must interpret the answer from multiple modalities. Das et al. extended VQA by proposing the visual dialogue task [17] which has multiple rounds of human-machine interaction. Our GanDraw dataset is distinguished from the previous visual dialogue dataset due to the additional challenge of generating a photo-realistic image from a dialogue utterance on the Drawer’s side. This additional challenge further pushes research to build machine learning models that can handle various type of interplay between vision and language. The game setting of GanDraw is highly inspired from Codraw [26], a task that involves a Drawer placing predefined images and objects within a scene whilst holding a conversation with the Teller. However, our task has a significant difference from Codraw on how the language is grounded to vision. Codraw’s image editing is implemented by placing a set of limited predefined objects with a few properties on a virtual space, which means that the language to vision model prediction will be in a coordinate space for their task. However, GanDraw involves natural free-hand drawing in a photo-realistic image domain which involves a more challenging mapping from language to vision. We have observed the complexity of our task has evoked richer dialogue between participants.

Conditional Image Generation Conditional text-to-image generation with GANs can be extended to build our drawer role. Reed [27] proposed the first conditional GAN that generates an image given a caption. StackGAN [28] improved the synthetic image quality by breaking down the generation process into two stages. [29] extended StackGAN with an attention mechanism that can learn to attune to different text inputs to synthesize various spatial locations on the generated image. Instead of using a single caption, Chatpainter [30] uses the entire dialogue history to generate images thereby leveraging richer information. Language based image editing (LBIE) is a more related field with

our work where the generation of the target image will be conditioned on both a textual description and a source image. Existing LBIE work [31, 32, 33] focuses on editing images with a single instruction without addressing the challenge of iterative image editing, which involves a sequence of textual instructions. Recently, there are several attempts to resolve the sequential text-to-image generation. Story GAN [19] addresses the problem of story visualization by generating discrete animation frames from a story script mapped to a corresponding frame. GeNeVA GAN [18] is another similar work which composes an image step by step based upon a sequence of instructions. Both systems, however, focus entirely on iterative image generation from text without the ability to interact with other agents, which ultimately fails to maximize the information that they could obtain for image synthesis. Both pipelines also work on animation or highly simplified synthetic images leading to significantly limited textual instructions and operations on the image. We address these issues in our GanDraw task where we study realistic image generation from interactive conversations. This requires the model to handle complex mapping between text and real image while eliciting more information from the other party through dialogue.

2.2.3 Dataset

GanDraw Game

GanDraw is a cooperative game between a teller and a drawer. When the game begins, the teller is randomly assigned with a target image. The drawer is given access to a drawing tool and a blank background. The teller initializes the conversation with a caption to describe part of the target image to the drawer. Then the drawer can either perform the editing to the background image according to teller’s instructions or ask a clarification question to the teller if the given instruction is not clear. After the teller gets the feedback from the drawer, s/he will answer any clarification question raised and continue introducing other parts of the target image that hasn’t been covered before. Additionally, the teller also has access to drawer’s image at each turn, and can choose to give correction suggestions if the generated image is not well aligned with the target image. When the teller feels the drawer’s image is close enough to the target image, s/he

would terminate the game.

One big challenge in this task is to design the drawing tool that can enable the drawer to easily perform editing on a realistic image. Owing to the recent advances in conditional GAN, we propose to use GauGAN [34], which can synthesize high-quality realistic images from semantic label maps. Each colored label in the semantic map is aligned with an object in the synthetic image and changes applied to the color label in semantic map are also reflected in the generated synthetic image. Therefore, as shown in figure. 2.3, we build a drawing tool interface that allows users to paint on a semantic label space where we then convert the semantic label image to a realistic image with a pre-trained GauGAN model to simulate the process of directly editing the real image. This interface helps to simplify the complex mapping from language to vision to a problem that focuses on interpreting the geometric layout of mentioned items from language. Although the drawer does not have fine-grained control on the other visual attributes such as colors, brightness, or texture, we provide them with a list of style selection (e.g sunset view, winter view, etc) that they can apply to the image to change the general outlook of the image with respect to teller’s description.

Data Collection

Target Image Selection As the image editing tool is built from GauGAN [35], we follow their paper to use the model that is pre-trained on natural landscape images which demonstrates the superior quality on the generated synthetic image. Images generated from GAUGAN using datasets with a more complex scenario, such as MSCOCO [36], resulted in severe artifacts. Hence, we limit the target image selection to the natural landscape domain. We downloaded 50,000 landscape images from the Flickr API to provide a large pool for selecting high quality target images. Following GauGAN’s paper we predefine 21 nature item labels that we want to keep in the drawing interface. We apply a pretrained semantic segmentation network from DeepLabV2 [37] on the downloaded landscape images to generate semantic labeling map for them and discard images that contain the labels out side of the predefined ones. We individually curated images that contained between 2-6 semantic labels in-order to provide a diverse selection of landscape

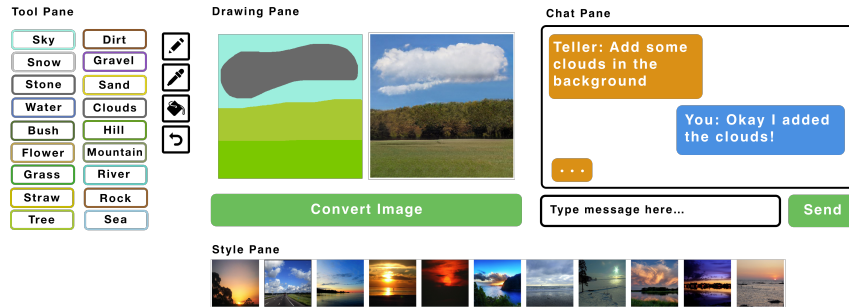


Figure 2.3: Data Collection Interface for the Drawer.

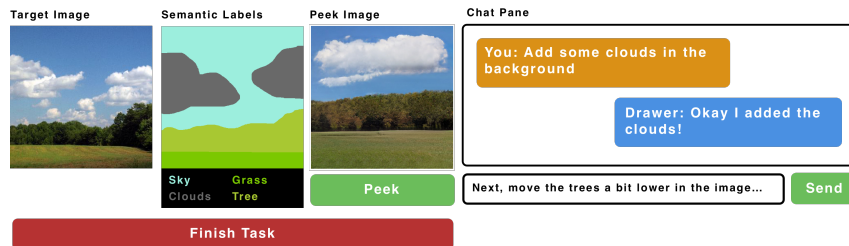


Figure 2.4: Data Collection Interface for Teller.

images that contains a rich scene while not making the image too challenging for the drawer to recreate with just the language description from the teller. After filtering, we select 240 diverse target images of natural landscapes to be used in our task. For each target image, we also manually edit the segmentation map that is generated from the pretrained DeepLabV2 model as the ground truth segmentation map associated with the target image. The ground truth segmentation map is later used to compute the automatic evaluation score that we give to measure the performance of the drawer’s drawing. Although the final pool of target image is relatively small, it still covers a wide range of different combinations from the limited predefined nature items. Additionally, after collecting multiple conversations for each target image, we populate the dataset to contain over 7K image-to-text pairs which is suited to build machine learning model to explore cross-modal grounding and generation in a constrained domain.

Interface We designed two primary interfaces (see Fig. 2.3 and Fig. 2.4) using the *React Task Demo* from the ParlAI framework [38]. Both the Drawer and Teller’s interface feature a chat-pane on the right side where they can communicate in a turn-wise fashion. From the Teller’s left side, they can see a) the target image they are instructed to sequentially

describe, b) the ground truth semantic labels for the target image, as often semantic labels can be ambiguous in the natural image (is it a tree or a bush?), and c) an option to peek at what the Drawer has currently drawn. In comparison, from the Drawer’s left side, they can see a) a canvas in which they can paint, b) a palette in which they can select different semantic labels and different textures/styles, c) a toolbox used for switching drawing tools, such as changing brush size, and d) an option to preview their synthesized image from their semantic labels. At every turn, each side is only allowed to send a single message with a capped length of 140 characters. The length cap is set to prevent the Teller from giving too detailed description at once, so the information can be added interactively through conversations.

Post Processing At the end of each game, the performance of the participants is evaluated with the SegScene Similarity score that we introduce in Sec.2.2.3 We instruct the teller that the final drawing must satisfy two conditions in order to be considered as a qualified drawing that can recover the target image: (1) The final drawing must contain all of the scene labels in the target image. (2) The relative positioning between each pair of scene labels must align with the relative position of the content in the target image. To assure high quality data, we also conduct a post processing step by filtering out the drawings that have a lower than 1.0 SegScene Similarity Score. Additionally, we also have two annotators go through all of the collected data and filter out the conversations that do not meet the two criteria previously set for qualified recovery. As a result, we collect 5 games for each target image which sums to 1200 dialogues.

Dataset Analysis

Se describe details of the analysis we have on dataset collected from GanDraw. The GanDraw dataset consists of 1,240 games, with over 7K utterances and corresponding images. The average number of turns per game is 6.0. The average utterance length of the drawer is 5.7 words, while the average utterance length of the teller is 15.4 words. To understand how dialogues can help with the drawing, we identify the dialogue policies that drawers and tellers apply to form an effective communication. We identify four different dialogue strategies tellers and drawers adopt to effectively exchange information in our

dataset. We describe them in details below:

Describe Image: teller’s dialogue strategy to communicate his general understanding of the target image to the drawer. The teller also adopts this strategy to provide answers when the drawer raises question on the target image.

Correct Drawing: teller applies this strategy when he captures a critical difference between the drawer’s generated image and the target image. In this case, a teller will provide direct instruction on how to change the items in the drawing in order to make it look closer to the target image.(e.g *No. Remove land from bottom of page.*)

Elicit Information: a drawer practices this strategy when he gets unclear instructions from a teller or when he wants to confirm an interpretation that he has based on the description from the teller. This strategy is usually carried by asking a clarification question.

Request Correction: a drawer applies this strategy when he is not confident whether he has followed teller’s instruction appropriately. The drawer will then raise a yes or no question to confirm his drawing is ok. (e.g *is the mountain big enough?*)

We attempt to annotate every utterance in our dataset with one or more of the above dialogue strategies. If none of the dialogue strategy is matched with the utterance, we will just assign it to *other* type. We found that dialogues where tellers correct drawers more often achieved better final images compared to dialogues where tellers only describe the target images without correcting drawers. This suggests that a teller using both strategies leads to more successful task outcome. Yet even with tellers leading the conversation, it is also critical for drawers to effectively use strategies that make the teller provide more accurate information about the target image. For example, we found that whenever the drawer requests for a correction, the teller has a high chance to apply *correct drawing* descriptions afterwards. Leveraging elicited information by asking for clarification (e.g *how far up the image is the mountain?*) also helps the teller to give clearer instructions.

SegScene Similarity

We propose metrics to measure the similarity between the Drawer’s image with the target image. As the real image from Drawer is synthesized from GauGAN, the degree of realism is not controllable by the Drawer. Therefore, we primarily compare the semantic label image that the Drawer reconstructs from Teller’s instructions to the ground truth semantic label on the target image. We follow classic semantic segmentation literature [39] by computing the MeanIoU between the two semantic segmentation maps. Additionally, we also compute a pairwise similarity (PSim) score to evaluate the relative positioning between each label in the target image. We compute the final SegScene Similarity score as a weighted sum of the meanIoU and PSim in order to gain a score that is between 0 and 5.

MeanIoU Suppose the set of labels contained in Drawer semantic labeling image is L_d and the set of labels contained in target semantic labeling image is L_t . The first thing we want to measure is how many shared semantic labels there are and how much overlap there is between the shared label regions of the two images. Therefore, we compute the mean IoU between L_d and L_t :

$$\text{meanIoU}(L_d, L_t) = \frac{\sum_{l^s \in L_d \cap L_t} \text{IoU}(l_d^s, l_t^s)}{|\{L_d \cup L_t\}|}, \quad (2.3)$$

where l_d^s and l_t^s stands for the shared label l^s in Drawer’s image and Teller’s image, respectively.

PSim As it is hard to describe all aspects within an image, achieving a high meanIoU with only textual instruction is extremely challenging. We thus also compute a pairwise similarity (PSim) score between every pair of shared labels $l^i, l^j \in \{L_d \cap L_t\}$. It serves as a straight-forward evaluation metric to reward the correct relative positioning between the major elements from the drawers’ image:

$$\text{PSim} = \frac{\sum_{l^i, l^j \in L_d \cap L_t} h(l_d^i, l_d^j, l_t^i, l_t^j)}{|L_d \cup L_t|(|L_d \cup L_t| - 1)}, \quad (2.4)$$

where l_d^i, l_d^j represents the shared label l^i, l^j in Drawer’s image. l_t^i, l_t^j represents the shared label in target image. $h(l_d^i, l_d^j, l_t^i, l_t^j)$ stands for the relative positional similarity score

between the shared labels l^i and l^j in Drawer’s image and that in the target image. The equation is defined below:

$$h(l_d^i, l_d^j, l_t^i, l_t^j) = \mathbf{1}_{(x_d^i - x_d^j)(x_t^i - x_t^j) > 0} + \mathbf{1}_{(y_d^i - y_d^j)(y_t^i - y_t^j) > 0} \quad (2.5)$$

where the x coordinate and y coordinate for l^i and l^j are gained as the center of mass of the regions for l^i and l^j .

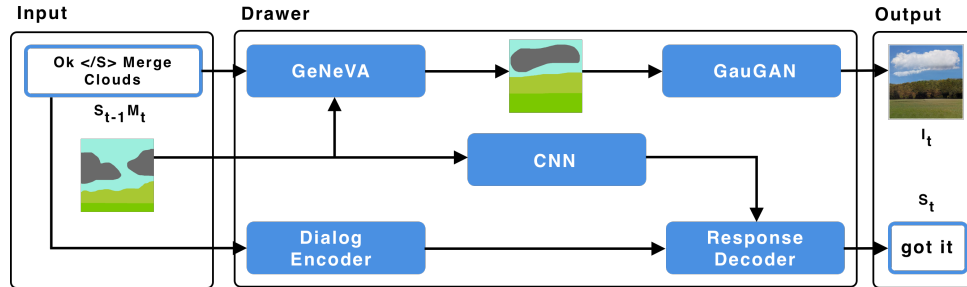


Figure 2.5: Neural Drawer’s architecture. It takes previous utterance context and the previously generated image as input. It then generates the output image I_t through a two stage process and next utterance s_t through a separate encoder decoder pipeline.

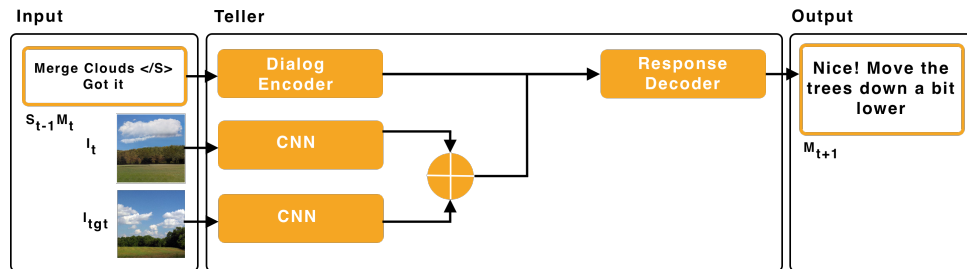


Figure 2.6: Neural Teller’s architecture. It takes the previous turns along with drawer’s generated image and target image as input. The model then run through a multimodal encoder decoder pipeline to generate the next utterance.

2.2.4 Model

Given a target image I_{tgt} , a teller initializes the conversation with a general instruction M_1 on how to recreate part of the target image. At each turn t , a drawer follows the instruction M_t from the teller to edit the previous drawing I_{t-1} into a new picture I_t . The drawer needs to send a feedback message S_t back to the teller in order to carry on the conversation. The feedback message can either be a general response like "done", or

clarification question for more information from the teller. The teller processes the edited image I_t and the feedback message S_t from the drawer, and compares them with the target image I_{tgt} to determine the next information M_{t+1} back to the drawer. This message can be a combination of answering drawer’s question, giving correction suggestion based on the observed difference between I_t and I_{tgt} . The conversation continues until the teller is confident that the image created by the drawer is similar enough to the target image.

In the above problem definition, the teller and the drawer perform multiple complex vision-language tasks, including visual question answering, visual question generation, dialogue history tracking and image synthesis. To the best of our knowledge, no single existing model can be directly applied to solve our task automatically. We thus build an automatic teller and drawer extending other language-vision models.

Neural Teller

We convert a relative image caption [40] neural model to construct the baseline of the teller, which is demonstrated in figure 2.6. The original model can capture the differences between two input images and generate a natural language to describe them which simulates the process of teller giving correction suggestions to the drawer by comparing their drawing with the target image. Specifically, at each turn the drawing generated from the drawer I_t and the target image I_{tgt} are encoded through a pretrained convolutional neural network (CNN) to generate feature representations. The generated features of the two images are fused through a fusion module and then applied to a long short-term memory network (LSTM) which generates a sentence to describe the difference between the two images. To enable the model to also condition the language generation on the dialogue context, we add a Bi-directional LSTM dialogue encoder to encode the previous utterances.

Neural Drawer

The objective of the drawer is highly correlated with the problem of iterative text-to-image generation, where the model needs to condition the image generation on a sequence of text instructions and the previous generated image. We build the baseline for our drawer from the state-of-the-art iterative text-to-image GAN model GeNeVA [18] (displayed in

figure 2.5) with several critical changes. First, we break the problem of text to natural image generation into a two-stage process. In the first stage, we generate a semantic labeling map from the text through GeNeVA. Then, the generated segmentation map will be processed with the pre-trained GauGAN model [35] to generate the real image. The two-stage generation process converts a complex task into two easier sub-problems that can be better handled by different GAN models. Our second critical change is that we construct a separate encoder-decoder pipeline to enable the drawer to talk. The utterance generated by the drawer is conditioned on both dialogue history and the previous generated segmentation map through a LSTM based dialogue encoder and a shallow CNN image encoder. This separate encoder-decoder pipeline is optimized with the maximum-likelihood objective.

2.2.5 Experiment

We evaluate our neural models with a fully automatic protocol and by having the model play with human players in the GanDraw game. The performance of the model in each setting is evaluated with the SegScene Similarity Score.

Automatic Evaluation Setting

We first introduce the three types of our automatic evaluation setting:

Image Generation from Recorded dialogues: The neural drawer model can be evaluated against the human-human conversation from the dataset, where the utterances from both sides at each turn are considered as the instruction to guide the neural drawer to generate the image corresponding to that turn. In this setting, the neural drawer model is considered as a "silent" drawer, where it would not generate any utterances but simply focus on generating images from the dialogue script.

Image Generation Without Context: In order to understand the importance of the visual and dialogue context for the success of the task, we modify the neural drawer baseline to only generate the image at each step with the text instruction at that turn. We evaluate this modified neural drawer with the ground truth dialogues following the same setting as the first evaluation to make a fair comparison.

Machine-Machine Game: To evaluate the neural teller and neural drawer in an interactive setting, we let the neural teller and neural drawer to play the GanDraw game between themselves by iterating through all target images once.

Human Evaluation Setting

We also evaluate the performance of our teller and drawer baselines by playing the GanDraw task with human players. We take all 25 images from the test split of our dataset to form the pool of the target images. The teller model and the drawer model are then paired with human drawers and human tellers separately to play 25 games. We have different game settings for the teller model and the drawer model when they play with human.

Neural Teller-Human Drawer: when the teller model plays with the human drawer, we allow the model to access all the intermediate generated images from the human drawer. With the full observable context, we want to verify whether the teller can adopt different strategies to carry on the conversation, such as provide correction suggestion or describe new concept on the target image that is not covered in the game.

Human Teller-Neural Drawer: When the drawer model interacts with a human teller, we only let the human teller peek the intermediate images generated by the drawer model twice. By doing this, we want to reduce the amount of visual information the human teller can receive in order to encourage the drawer model to adopt effective dialogue strategies such as elicit information and request correcting introduced in Sec.2.2.3 to gain more information from tellers to re-create the target image.

We launched the human evaluation task on AMT and collected 25 games for each setting. We filter out the games where Turkers did not attempt to collaborate with our neural models to assure all the evaluation results do not have any major human-side mistakes.

Teller	Drawer	MeanIoU	Psim	SegScene Sim
dialogue	Machine	0.18	0.25	1.11
dialogue	Machine w/o v+d	0.14	0.21	0.92
Machine	Machine	0.13	0.22	0.92
Machine	Human	0.19	0.25	1.12
Human	Machine	0.18	0.26	1.13
Human	Human	0.40	0.61	2.63

Table 2.5: Our baseline model’s results from both human evaluation and automatic evaluation protocol. **Machine** under the Teller or Drawer represents the neural model we develop for that role. **Machine w/o v+d** is the Drawer model that generates image without any visual or dialogue context. The line with **dialogue** under the teller is the experiment for iterative image generation from the collected dialogues in the dataset.

Results and Analysis

Table 2.5 summarizes the evaluation results of the performance from our neural baselines under the evaluation of human-machine setting and full automatic evaluation setting. There is a considerable gap between the performance of the neural models and humans, which indicates the need for additional work to improve the model performance on our task. Performance of the neural drawer when they are played with human partners or when they have access to ground truth dialogue contexts are relatively the same. However, when the neural teller and the neural drawer played with each other, errors from both sides caused the worst SegScene Similarity score compared to other settings. The reduced performance from the neural drawer when it is restrained to only generate an image with the instruction from each turn indicates the importance of utilizing dialogue and visual context to succeed in our task. To better understand the model’s limitations, we additionally reviewed the games played between the neural models and their human partners.

Neural Tellers Upon reviewing the games, we found that the neural tellers failed to track information introduced previously. This results in repetition both on the utterance level and the object level. For example, in figure. 2.7, the neural teller mentions to *add a small patch of grass* twice. This could be caused by the LSTM-based encoder lacking the

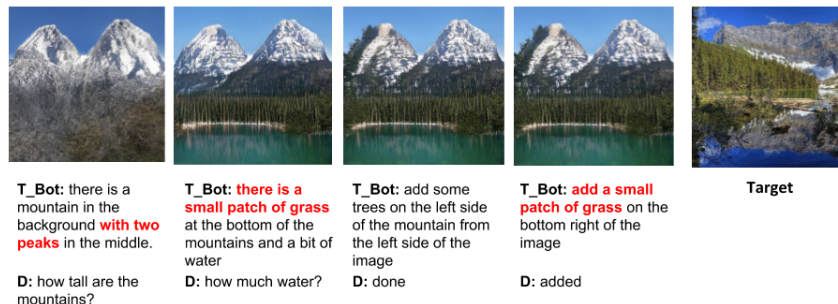


Figure 2.7: A qualitative example of the game our neural teller played with the human drawer. We highlight the text where the neural teller makes a mistake with **red color**.

ability to track extended context. Using a transformer-based pre-trained language model may reduce this issue. The neural teller also lacks accurate understanding of the geometric layout of the object in the target image. It often captures the right item from the target image but provides an incorrect description of its location, size, or shape. In figure.2.7, it fails to recognize the shape of the mountain and describe it as "two peak". This indicates GanDraw requires the model not only describe the visual attributes associated item, but also understand the geometric properties inside the image.

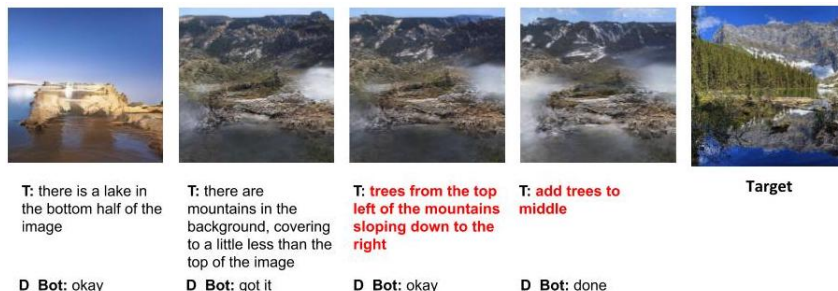


Figure 2.8: A qualitative example of the game our neural drawer played with the human teller. We highlight the text where the neural drawer does not follow teller’s instruction with **red color**.

Neural Drawer The neural drawer seems to be good at capturing individual concepts when the description is short. However, it struggles to interpret several concepts mentioned in longer sentences. When the human teller tries to describe multiple objects in one sentence, it usually generates one of them. Besides, the neural drawer lacks the ability to ground the fine-grained change instructions to the image. Often times it only follows the instruction that gives a general description on new objects but ignores the correction

suggestion. An example is displayed in figure 2.8. On the third turn, when the teller tells the neural drawer to add trees to the middle, the model ignores the teller's request. This suggests that the conditional GAN model cannot ground text descriptions that involve various levels of details to appropriate changes in the generated image. Finally, the language generated from the drawer are generic phrases, mostly "ok" and "done". This might be caused by the implicit bias within the GanDraw task where the drawer predominantly takes a passive role by following teller instruction and only replying with generic information. It could also result from the language generator in the neural drawer being trained with the maximum likelihood objective, which has been found to generate generic response in many applications.[41]

Chapter 3

Cross-lingual Cross-modal Representation Learning

3.1 Augment Machine Translation with Vision

3.1.1 Introduction

Multimodal machine translation is the problem of translating sentences paired with images into a different target language [42]. In this setting, translation is expected to be more accurate compared to purely text-based translation, as the visual context could help resolve ambiguous multi-sense words. Examples of real-world applications of multimodal (vision plus text) translation include translating multimedia news, web product information, and movie subtitles.

Several previous endeavours [43, 44, 45] have demonstrated improved translation quality when utilizing images. However, how to effectively integrate the visual information still remains a challenging problem. For instance, in the WMT 2017 multimodal machine translation challenge [46], methods that incorporated visual information did not outperform pure text-based approaches with a big margin.

We propose a new model called Visual Attention Grounding Neural Machine Translation (VAG-NMT) to leverage visual information more effectively. We train VAG-NMT with a multitask learning mechanism that simultaneously optimizes two objectives: (1) learning a translation model, and (2) constructing a vision-language joint semantic em-

bedding. In this model, we develop a visual attention mechanism to learn an attention vector that values the words that have closer semantic relatedness with the visual context. The attention vector is then projected to the shared embedding space to initialize the translation decoder such that the source sentence words that are more related to the visual semantics have more influence during the decoding stage. When evaluated on the benchmark Multi30K and the Ambiguous COCO datasets, our VAG-NMT model demonstrates competitive performance compared to existing state-of-the-art multimodal machine translation systems.

3.1.2 Related Work

In the machine translation literature, there are two major streams for integrating visual information: approaches that (1) employ separate attention for different (text and vision) modalities, and (2) fuse visual information into the NMT model as part of the input. The first line of work learns independent context vectors from a sequence of text encoder hidden states and a set of location-preserving visual features extracted from a pre-trained convnet, and both sets of attentions affect the decoder’s translation [44, 47]. The second line of work instead extracts a global semantic feature and initializes either the NMT encoder or decoder to fuse the visual context [48, 49]. While both approaches demonstrate significant improvement over their Text-Only NMT baselines, they still perform worse than the best monomodal machine translation system from the WMT 2017 shared task [50].

The model that performs best in the multimodal machine translation task employed image context in a different way. [43] combine region features extracted from a region-proposal network [51] with the word sequence feature as the input to the encoder, which leads to significant improvement over their NMT baseline. The best multimodal machine translation system in WMT 2017 [52] performs element-wise multiplication of the target language embedding with an affine transformation of the convnet image feature vector as the mixed input to the decoder. While this method outperforms all other methods in WMT 2017 shared task workshop, the advantage over the monomodal translation system is still minor.

The proposed visual context grounding process in our model is closely related to the

literature on multimodal shared space learning. [53] propose a neural language model to learn a visual-semantic embedding space by optimizing a ranking objective, where the distributed representation helps generate image captions. [54] densely align different objects in the image with their corresponding text captions in the shared space, which further improves the quality of the generated caption. In later work, multimodal shared space learning was extended to multimodal multilingual shared space learning. [55] learn a multi-modal multilingual shared space through optimization of a modified pairwise contrastive function, where the extra multilingual signal in the shared space leads to improvements in image-sentence ranking and semantic textual similarity task. [56] extend the work from [55] by using the image as the pivot point to learn the multilingual multimodal shared space, which does not require large parallel corpora during training. Finally, [45] is the first to integrate the idea of multimodal shared space learning to help multimodal machine translation. Their multi-task architecture called “imagination” shares an encoder between a primary task of the classical encoder-decoder NMT and an auxiliary task of visual feature reconstruction.

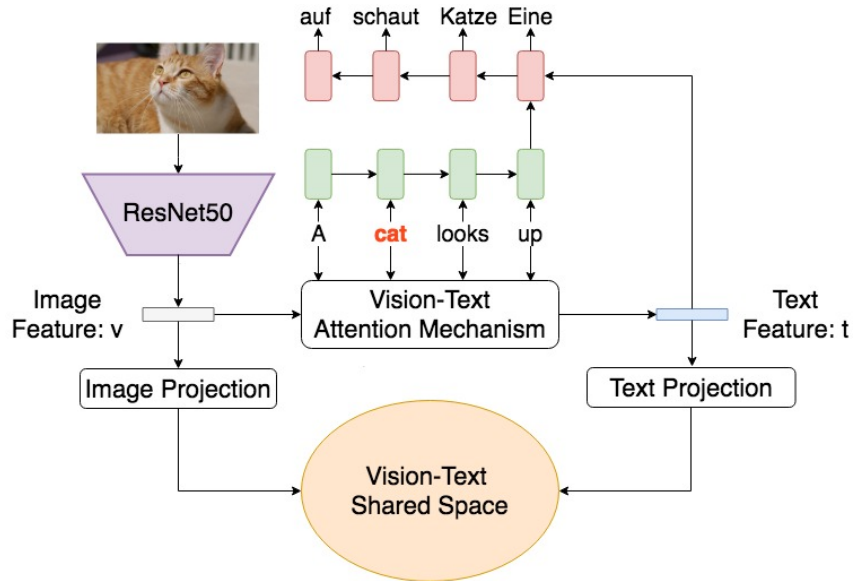


Figure 3.1: An overview of the VAG-NMT structure

3.1.3 Model

Given a set of parallel sentences in language X and Y , and a set of corresponding images V paired with each sentence pair, the model aims to translate sentences $\{x_i\}_{i=1}^N \in X$ in language X to sentences $\{y_i\}_{i=1}^N \in Y$ in language Y with the assistance of images $\{v_i\}_{i=1}^N \in V$.

We treat the problem of multi-modal machine translation as a joint optimization of two tasks: (1) learning a robust translation model and (2) constructing a visual-language shared embedding that grounds the visual semantics with text. Figure 3.1 shows an overview of our VAG-NMT model. We adopt a state-of-the-art attention-based sequence-to-sequence structure [57] for translation. For the joint embedding, we obtain the text representation using a weighted sum of hidden states from the encoder of the sequence-to-sequence model and we obtain the image representation from a pre-trained convnet. We learn the weights using a visual attention mechanism, which represents the semantic relatedness between the image and each word in the encoded text. We learn the shared space with a ranking loss and the translation model with a cross entropy loss.

The joint objective function is defined as:

$$J(\theta_T, \phi_V) = \alpha J_T(\theta_T) + (1 - \alpha) J_V(\phi_V) \quad (3.1)$$

where J_T is the objective function for the sequence-to-sequence model, J_V is the objective function for joint embedding learning, θ_T are the parameters in the translation model, and ϕ_V are the parameters for the shared vision-language embedding learning, and α determines the contribution of the machine translation loss versus the visual grounding loss. Both J_T and J_V share the parameters of the encoder from the neural machine translation model.

Encoder

We first encode an n -length source sentence $\{x\}$, as a sequence of tokens $x = \{x_1, x_2, \dots, x_n\}$, with a bidirectional GRU [58, 59]. Each token is represented by a one-hot vector, which is then mapped into an embedding e_i through a pre-trained embedding matrix. The bidirectional GRU processes the embedding tokens in two directions: left-to-right (forward)

and right-to-left (backward). At every time step, the encoder’s GRU cell generates two corresponding hidden state vectors: $\vec{h}_i = \overrightarrow{GRU}(h_{i-1}, e_i)$ and $\overleftarrow{h}_i = \overleftarrow{GRU}(h_{i-1}, e_i)$. The two hidden state vectors are then concatenated together to serve as the encoder hidden state vector of the source token at step i : $h_i = [\overleftarrow{h}_i, \vec{h}_i]$.

Shared Embedding Objective

After encoding the source sentence, we project both the image and text into the shared space to find a good distributed representation that can capture the semantic meaning across the two modalities. Previous work has shown that learning a multimodal representation is effective for grounding knowledge between two modalities [53, 60]. Therefore, we expect the shared encoder between the two objectives to facilitate the integration of the two modalities and positively influence translation during decoding.

To project the image and the source sentence to a shared space, we obtain the visual embedding (v) from the pool5 layer of ResNet50 [61] pre-trained on ImageNet classification [62], and the source sentence embedding using the weighted sum of encoder hidden state vectors ($\{h_i\}$) to represent the entire source sentence (t). We project each $\{h_i\}$ to the shared space through an embedding layer. As different words in the source sentence will have different importance, we employ a visual-language attention mechanism—inspired by the attention mechanism applied in sequence-to-sequence models [57]—to emphasize words that have the stronger semantic connection with the image. For example, the highlighted word “cat” in the source sentence in Fig. 3.1 has the more semantic connection with the image.

Specifically, we produce a set of weights $\beta = \{\beta_1, \beta_2, \dots, \beta_n\}$ with our visual-attention mechanism, where the attention weight β_i for the i ’th word is computed as:

$$\beta_i = \frac{\exp(z_i)}{\sum_{l=1}^N \exp(z_l)}, \quad (3.2)$$

and $z_i = \tanh(W_v v) \cdot \tanh(W_h h_i)$ is computed by taking the dot product between the transformed encoder hidden state vector h_i and the transformed image feature vector v , and W_v and W_h are the association transformation parameters.

Then, we can get a weighted sum of the encoder hidden state vectors $t = \sum_{i=1}^n \beta_i h_i$ to represent the semantic meaning of the entire source sentence. The next step is to project

the source sentence feature vector t and the image feature vector v into the same shared space. The projected vector for text is: $t_{emb} = \tanh(W_{t_{emb}}t + b_{t_{emb}})$ and the projected vector for image is: $v_{emb} = \tanh(W_{v_{emb}}v + b_{v_{emb}})$.

We follow previous work on visual semantic embedding [53] to minimize a pairwise ranking loss to learn the shared space:

$$J_V(\phi_V) = \sum_p \sum_k \max\{0, \gamma - s(v_p, t_p) + s(v_p, t_{k \neq p})\} + \sum_k \sum_p \max\{0, \gamma - s(t_k, v_k) + s(t_k, v_{p \neq k})\} \quad (3.3)$$

where γ is a margin, and s is the cosine distance between two vectors in the shared space. $t_{k \neq p}$ and $v_{p \neq k}$ are the contrastive examples with respect to the selected image and the selected source text, respectively. When the loss decreases, the distance between a paired image and sentence will drop while the distance between an unpaired image and sentence will increase.

In addition to grounding the visual context into the shared encoder through the multimodal shared space learning, we also initialize the decoder with the learned attention vector t such that the words that have more relatedness with the visual semantics will have more impact during the decoding (translation) stage. However, we may not want to solely rely on only a few most important words. Thus, to produce the initial hidden state of the decoder, we take a weighted average of the attention vector t and the mean of encoder hidden states:

$$s_0 = \tanh(W_{init}(\lambda t + (1 - \lambda) \frac{1}{N} \sum_i^N h_i)), \quad (3.4)$$

where λ determines the contribution from each vector. Through our experiments, we find the best value for λ is 0.5.

Translation Objective

During the decoding stage, at each time step j , the decoder generates a decoder hidden state s_j from a conditional GRU cell [63] whose input is the previously generated translation token y_{j-1} , the previous decoder hidden state s_{j-1} , and the context vector c_j at the current time step:

$$s_j = \text{cGRU}(s_{j-1}, y_{j-1}, c_j) \quad (3.5)$$

The context vector c_j is a weighted sum of the encoder hidden state vectors, and captures the relevant source words that the decoder should focus on when generating the current translated token y_j . The weight associated with each encoder hidden state is determined by a feed-forward network. From the hidden state s_j we can predict the conditional distribution of the next token y_j with a fully-connected layer W_o given the previous token’s language embedding e_{j-1} , the current hidden state d_j and the context vector for current step c_j :

$$p(y_j|y_{j-1}, x) = \text{softmax}(W_o o_t) \quad (3.6)$$

where $o_t = \tanh(W_e e_{j-1} + W_d d_j + W_c c_j)$. The three inputs are transformed with W_e , W_d , and W_c , respectively and then summed before being fed into the output layer.

We train the translation objective by optimizing a cross entropy loss function:

$$J_T(\theta_T) = - \sum_j \log p(y_j|y_{j-1}, x) \quad (3.7)$$

By optimizing the objective of the translation and the multimodal shared space learning tasks jointly along with the visual-language attention mechanism, we can simultaneously learn a general mapping between the linguistic signals in two languages and grounding of relevant visual content in the text to improve the translation.

3.1.4 Experiment

Experiment Setting and Dataset

We evaluate our proposed model on three datasets: Multi30K [42], Ambiguous COCO [46], and our newly-collected IKEA dataset. The Multi30K dataset is the largest existing human-labeled dataset for multimodal machine translation. It consists of 31,014 images, where each image is annotated with an English caption and manual translations of image captions in German and French. There are 29,000 instances for training, 1,014 instances for validation, and 1,000 for testing. Additionally, we also evaluate our model on the Ambiguous COCO Dataset collected in the WMT2017 multimodal machine translation challenge [46]. It contains 461 images from the MSCOCO dataset [64], whose captions contain verbs with ambiguous meanings.

We pre-process all English, French, and German sentences by normalizing the punctuation, lower casing, and tokenizing with the Moses toolkit. A Byte-Pair-Encoding (BPE) [65] operation with 10K merge operations is learned from the pre-processed data and then applied to segment words. We restore the original words by concatenating the subwords segmented by BPE in post-processing. During training, we apply early stopping if there is no improvement in BLEU score on validation data for 10 validation steps. We apply beam search decoding to generate translation with beam size equal to 12. We evaluate the performance of all models using BLEU [66] and METEOR [67]. The setting used in IKEA dataset is the same as Multi30K, except that we lower the default batch size from 32 to 12; since IKEA dataset has long sentences and large variance in sentence length, we use smaller batches to make the training more stable. We run all models five times with different random seeds and report the mean and standard deviation.

Results

Method	English → German		English → French	
	BLEU	METEOR	BLEU	METEOR
Imagination [45]	30.2	51.2	N/A	N/A
LIUMCVC [52]	31.1 ± 0.7	52.2 ± 0.4	52.7 ± 0.9	69.5 ± 0.7
Text-Only NMT	31.6 ± 0.5	52.2 ± 0.3	53.5 ± 0.7	70.0 ± 0.7
VAG-NMT	31.6 ± 0.3	52.2 ± 0.3	53.8 ± 0.3	70.3 ± 0.5

Table 3.1: Translation results on the Multi30K dataset

Method	English → German		English → French	
	BLEU	METEOR	BLEU	METEOR
Imagination [45]	28.0	48.1	N/A	N/A
LIUMCVC [52]	27.1 ± 0.9	47.2 ± 0.6	43.5 ± 1.2	63.2 ± 0.9
Text-Only NMT	27.9 ± 0.6	47.8 ± 0.6	44.6 ± 0.6	64.2 ± 0.5
VAG-NMT	28.3 ± 0.6	48.0 ± 0.5	45.0 ± 0.4	64.7 ± 0.4

Table 3.2: Translation results on the Ambiguous COCO dataset

We compare the performance of our model against the state-of-the-art multimodal machine translation approaches and the text-only baseline. The idea of our model is inspired by the "Imagination" model [45], which unlike our models, simply averages the encoder hidden states for visual grounding learning. As "Imagination" does not report its performance on Multi30K 2017 and Ambiguous COCO in its original paper, we directly use their result reported in the WMT 2017 shared task as a comparison. LIUMCVC is the best multimodal machine translation model in WMT 2017 multimodal machine translation challenge and exploits visual information with several different methods. We always compare our VAG-NMT with the method that has been reported to have the best performance on each dataset.

Our VAG-NMT surpasses the results of the "Imagination" model and the LIUMCVC's model by a noticeable margin in terms of BLEU score on both the Multi30K dataset (Table 3.1) and the Ambiguous COCO dataset (Table 3.2). The METEOR score of our VAG-NMT is slightly worse than that of "Imagination" for English -> German on Ambiguous COCO Dataset. This is likely because the "Imagination" result was produced by ensembling the result of multiple runs, which typically leads to 1-2 higher BLEU and METEOR points compared to a single run. Thus, we expect our VAG-NMT to outperform the "Imagination" baseline if we also use an ensemble.

We observe that our multimodal VAG-NMT model has equal or slightly better result compared to the text-only neural machine translation model on the Multi30K dataset. On the Ambiguous COCO dataset, our VAG-NMT demonstrates clearer improvement over this text-only baseline. We suspect this is because Multi30K does not have many cases where images can help improve translation quality, as most of the image captions are short and simple. In contrast, Ambiguous COCO was purposely curated such that the verbs in the captions can have ambiguous meaning. Thus, visual context will play a more important role in Ambiguous COCO; namely, to help clarify the sense of the source text and guide the translation to select the correct word in the target language.



Source Caption: a tennis player is moving to the side and is gripping his **racquet** with both hands .

Text-Only NMT: ein tennisspieler bewegt sich um die seite und greift mit beiden händen an **den boden** .

VAG-NMT: ein tennisspieler bewegt sich zur seite und greift mit beiden händen **den schläger** .



Source Caption: three skiers skiing on a hill with two going down the hill and one moving **up** the hill .

Text-Only NMT: drei skifahrer fahren auf skiern einen hügel hinunter und eine person fährt den hügel **hinunter** .

VAG-NMT: drei skifahrer auf einem hügel fahren einen hügel hinunter und ein bewegt sich den hügel **hinauf** .



Source Caption: a blue , yellow and green surfboard **sticking out** of a sandy beach .

Text-Only NMT: ein blau , gelb und grünes surfbrett **streckt aus** einem sandstrand .

VAG-NMT: ein blau , gelb und grüner surfbrett **springt aus** einem sandstrand .

Figure 3.2: Translations generated by VAG-NMT and Text-Only NMT. VAG-NMT performs better in the first two examples, while Text-Only NMT performs better in the third example. We highlight the words that distinguish the two systems' results in **red** and **blue**.

Analysis and Discussion

In the first row of Figure 3.3, the attention mechanism assigns high weights to the words "skiing", "snowboarding", and "snow". In the second row, it assigns high attention to "rafting" or "raft" for every caption of the three images. These examples demonstrate evidence that our attention mechanism learns to assign high weights to words that have corresponding visual semantics in the image.



a person is skiing or snowboarding down a mountainside .



a mountain climber trekking through the snow with a pick and a blue hat .



the snowboarder is jumping in the snow .



two women are water rafting .



three people in a blue raft on a river of brown water .



people in rafts watch as two men fall out of their own rafts .

Figure 3.3: The first and second rows show the three closest images to the caption *a person is skiing or snowboarding down a mountainside* and *two woman are water rafting*, respectively. The original caption is listed under each image. We highlight the three words with highest attention in red.

We also find that our visual grounding attention captures the dependency between the words that have strong visual semantic relatedness. For example, in Figure 3.3, words, such as “raft”, “river”, and “water”, with high attention appear in the image together. This shows that the visual dependence information is encoded into the weighted sum of attention vectors which is applied to initialize the translation decoder. When we apply the sequence-to-sequence model to translate a long sentence, the encoded visual dependence information strengthens the connection between the words with visual semantic relatedness. Such connections mitigate the problem of standard sequence-to-sequence models tending to forget distant history. This hypothesis is supported by the fact that our VAG-NMT outperforms all the other methods on the IKEA dataset which has long sentences.

Lastly, in Figure 3.2 we provide some qualitative comparisons between the translations from VAG-NMT and Text-Only NMT. In the first example, our VAG-NMT properly translates the word “racquet” to “den schläger”, while the Text-Only NMT mistranslated it to “den boden” which means “ground” in English. We suspect the attention mechanism and visual shared space capture the visual dependence between the word “tennis” and

“racquet”. In the second example, our VAG-NMT model correctly translates the preposition “up” to “hinauf” but Text-Only NMT mistranslates it to “hinunter” which means “down” in English. We consistently observe that VAG-NMT translates prepositions better than Text-Only NMT. We think it is because the pre-trained convnet features captured the relative object position that leads to a better preposition choice. Finally, in the third example, we show a failure case where Text-Only NMT generates a better translation. Our VAG-NMT mistranslates the verb phrase “sticking out” to “springt aus” which means “jump out” in German, while Text-Only NMT translates to “streckt aus”, which is correct. We find that VAG-NMT often makes mistakes when translating verbs. We think it is because the image vectors are pre-trained on an object classification task, which does not have any human action information.

3.2 Universal Cross-lingual Cross-modal representation learning

3.2.1 Introduction

Vision-and-language pre-training has achieved impressive success in learning multimodal representations between vision and language. To generalize this success to non-English languages, we introduce UC² (Universal Cross-lingual Cross-modal pre-training), the first machine translation-augmented framework for cross-lingual cross-modal representation learning. To tackle the scarcity problem of multilingual captions for image datasets, we first augment existing English-only datasets with other languages via machine translation (MT). Then we extend the standard Masked Language Modeling and Image-Text Matching training objectives to multilingual setting, where alignment between different languages is captured through shared visual context (*i.e.*, using image as pivot). To facilitate the learning of a joint embedding space of images and all languages of interest, we further propose two novel pre-training tasks, namely Masked Region-to-Token Modeling (MRTM) and Visual Translation Language Modeling (VTLM), leveraging MT-enhanced translated data. MRTM encourages fine-grained alignment between words and image re-

gions, by sharing the embedding space of word tokens and region labels (*i.e.*, object class predictions from an object detector). VTLM is designed to jointly learn cross-lingual cross-modal mapping from parallel textual corpora and paired images. Extensive experiments demonstrate that our proposed UC² framework achieves new state of the art over multiple mainstream benchmarks such as Multi30k [68, 69, 70] and COCO [36, 71, 72] across multilingual image-text retrieval and visual question answering (VQA) tasks.

3.2.2 Related Work

Vision-Language Pre-training. There is a growing interest in building generic pre-trained BERT-like [73] models for Vision and Language (V+L) tasks. Early work such as ViLBERT [74] and LXMERT [75] propose a two-stream architecture that encodes visual and textual input through two separate Transformers, and then fuse the two modalities by a cross-modal Transformer. Later work such as VL-BERT [76], Unicoder-VL[77] and UNITER [78] introduce a single-stream architecture that uses one Transformer to encode concatenated input from both modalities simultaneously. Later, Unified VLP [79] applies to both understanding and generation tasks. Further improvements are proposed on using different input features [80, 81] and multi-task learning [82].

Multimodal Multilingual Learning. Existing studies arching over multilingual and multimodal aspects mainly focus on two tasks: cross-modal retrieval and multimodal machine translation (MT). [83] introduces a multimodal multilingual approach by aligning images and captions in different languages to English captions. Unlike previous work using languages as a pivoting point, [84] learns a shared embedding space that forces representations of different languages towards the pivot image representation. Later work focuses on scaling to more languages via character-based word-embedding [85] or shared language-acoustic embedding [86]. SMALR [87] proposes a scalable multilingual model to learn visually aligned word embeddings, for better balance between multilingual capacity and task performance.

Multimodal MT exploits visual information to improve language translations. Earlier work introduces vision to an LSTM-based neural MT model via attention to visual context [88, 89], or fusion [90], or multi-task learning [91, 92]. Lately, Transformer-based [93]

models are proposed [94, 95]. There is also an growing interest in unsupervised multimodal MT [96, 97], where translation between monolingual corpus is augmented via pivoting on image.

While successful in individual tasks, these models are usually trained on small amount of data, which limits its extension to other tasks or languages. To learn task-agnostic universal representations across vision and multilingual text, M³P[98] introduces the first pre-training framework that alternatively optimizes the model on multi-modal monolingual corpus and mono-modal multilingual corpus. While M³P achieves better performance compared to task-specific methods, the alignment between vision and Non-English languages is hard to capture, as the model is learned via using English as the anchor point. To strengthen the alignment between vision and all languages, we propose to pre-train a unified architecture where sentences in different languages are grounded on shared visual context.

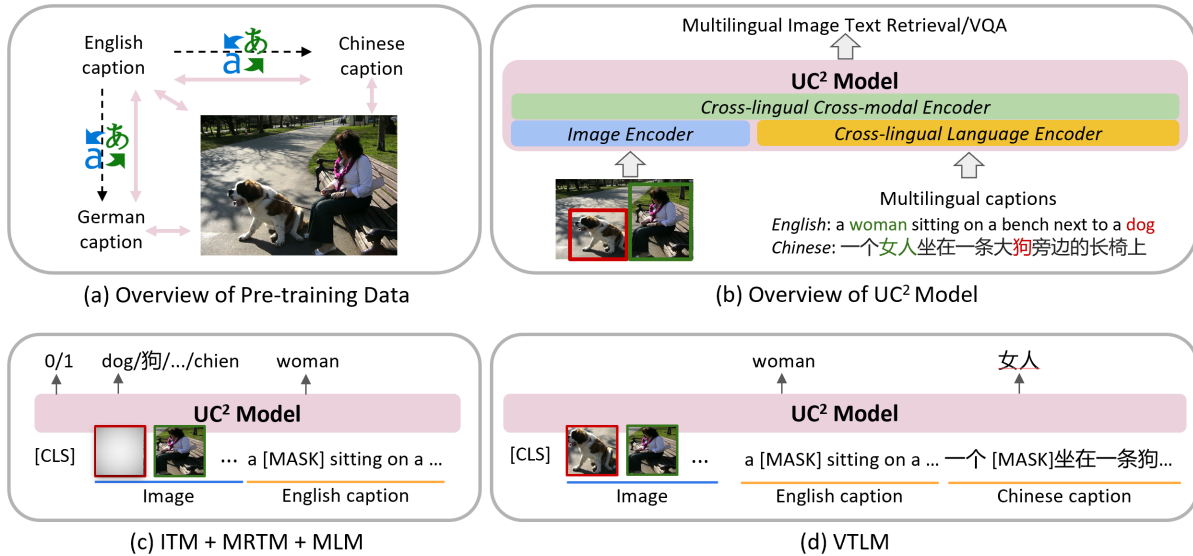


Figure 3.4: An overview of UC² model. Figure (a) shows the construction of multilingual multimodal pre-training corpus via machine translation. (b) depicts the overall UC² framework, which is pre-trained with a massive corpus of multilingual caption-image pairs. Figure (c) and (d) illustrate details of four pre-training tasks.

3.2.3 Model

In this section, we start with introducing our machine translation augmented dataset that enables large-scale cross-lingual pre-training. We then go over the proposed UC² model and our designed pre-training objectives for universal representation learning across vision and languages.

Machine Translation Augmented Dataset

Our multilingual image-text paired data is collected via augmenting the captions from the Conceptual Captions dataset [99] with a set of machine translated¹ captions in other languages $\mathbf{L} = \{l_1, l_2, \dots, l_n\}$. Specifically, we translate the original English captions into five different languages (German, French, Czech, Japanese, and Chinese), which covers languages required for all the downstream tasks studied in this work. Note that with recent advances on machine translation for low-resource languages, we can further expand the dataset to more languages, which we leave for future work. With this data augmentation, we obtained 3.3 million images, each paired with captions in six languages, as the process shown in Figure 3.4 (a). This one-to-many mapping greatly facilitates the learning of alignment between visual content and semantics from each language through image as a shared anchor. By introducing translated data into model pre-training, our method yields significant improvement over the baseline with MT tools applied only on downstream tasks. Next, we elaborate how to leverage these data for cross-lingual cross-modal pre-training.

Model Overview

UC² extends monolingual language encoder of V+L frameworks, such as UNITER [78], to cross-lingual encoder [100], as shown in Figure 3.4 (b). The visual feature is extracted from an image encoder and the language feature is obtained from a general cross-lingual language encoder. The multimodal features are then combined into a sequence and fed to a multi-layer Transformer to produce contextualized cross-modal and cross-lingual representations.

¹We use Microsoft Azure Translation API Service and will release the translated captions.

Image Encoder. Given an input image, we first obtain a sequence of image region features $\mathbf{v} = \{v_1, v_2, \dots, v_m\}$ with Faster R-CNN [101]. For each region, we also extract location features via a 7-dimensional vector: $\mathbf{p} = [x_1, y_1, x_2, y_2, w, h, w * h]$, which denotes the normalized top left coordinates, bottom right coordinates, width, height, and the area of the detected region box. The region feature and location feature are fed through separate fully-connected (FC) layers to be projected into the same dimension as the text embedding space, followed by a layer-normalization (LN) layer. The final representation of the region feature is then obtained via summing up the projected region feature and location feature.

Cross-lingual Language Encoder. We follow XLM-R [100] to tokenize an input sentence T^{l_i} in language l_i to BPE tokens $\mathbf{t}^{l_i} = \{t_1^{l_i}, t_2^{l_i}, \dots, t_n^{l_i}\}$ using Sentence Piece model [102]. We then project each token to its embedding based on the XLM-R vocabulary and word embeddings. The final representation of each token is obtained via summing up its word embedding, segment embedding, and position embedding as in XLM-R, followed by another Layer Normalization.

Pre-training Tasks

For model training, we employ four pre-training objectives to train on large multilingual image-text paired data: Masked Language Modeling (MLM), Image-Text Matching (ITM), Masked Region-to-Token Modeling (MRTM), and Visual Translation Language Modeling (VTLM), as shown in Figure 3.4 (c) and (d). We continuously optimize our model with the four objectives on multilingual image-text pairs to capture the cross-modal alignment between vision and different languages. As the translated captions are associated to the same image, cross-lingual alignment is also enforced using visual context as the anchor.

Masked Language Modeling (MLM). Given a set of image regions $\mathbf{v} = \{v_1, v_2, \dots, v_m\}$ and its associated caption words $\mathbf{w}^{l_i} = \{w_1^{l_i}, \dots, w_T^{l_i}\}$ in language $l_i \in \mathbf{L}$, and mask indices as $m \in \mathbb{N}^M$, we randomly mask a word $w_m^{l_i}$ with the probability of 15% and replace the masked word with a special token [mask]. The objective is to predict the masked word $w_m^{l_i}$ based on the surrounding words $w_{\setminus m}$ and all image regions \mathbf{v} , by minimizing the

negative log-likelihood:

$$\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{(w^i, v) \sim D} \log P_{\theta}(w_m^i | w_{\setminus m}^i, v), \quad (3.8)$$

where θ is the learnable parameters. Each pair $(\mathbf{w}^i, \mathbf{v})$ is sampled from the whole training set D . The caption for each language is sampled with even probability $p = 1/|\mathbf{L}|$.

Image-Text Matching (ITM). ITM has been widely used in vision-and-language pre-training [78, 77, 74, 76] to learn instance-level alignment between image and sentence. The output of the special token [cls] is fed through a FC layer and a sigmoid function to predict a score $s_{\theta}(w^i, v)$ between 0 and 1, which predicts whether the input image \mathbf{v} and the text input \mathbf{w}^i are semantically matched. During training, we sample positive and negative pairs from the dataset D with equal probability at each step. The negative image-text pair is created by replacing the image or text in a matched pair with a randomly-selected distractor from the same mini-batch. The objective is optimized with binary cross-entropy loss:

$$\begin{aligned} \mathcal{L}_{ITM}(\theta) = & -\mathbb{E}_{(w^i, v) \sim D} [y \log s_{\theta}(w^i, v) \\ & + (1 - y) \log(1 - s_{\theta}(w^i, v))] \end{aligned}$$

where $y \in 0, 1$ indicates whether the input image-text pair is a positive or negative sample. The deployment of MLM and ITM serves as our base model. Next, we introduce two novel objectives to further enhance cross-lingual cross-modal representation learning.

Masked Region-to-Token Modeling (MRLM) This new objective aims to classify each masked region to its “pseudo” object label, which is the (sub-word) token in our word vocabulary that associates with the original object label. Compared to the MRM objective from previous work [74, 77, 78], MRTM leverages additional semantic association between object labels and captions to capture semantic alignment between vision and language. More formally, given an image region $\mathbf{v}_i \in \mathbf{v}$, we set its probability for being masked out as 15% (as in [73]). For each masked region, the region feature vector is either replaced by a zero-initialized vector v_m (90% probability) or remains the same (10%). Then we predict the associated “pseudo” object label $c_{v_m}^i$ on the masked region based on the observation of surrounding image regions $v_{\setminus m}$ and the paired caption w^i in language l_i , by minimizing

the negative log-likelihood:

$$\mathcal{L}_{MRTM}(\theta) = -\mathbb{E}_{(w^{l_i}, v) \sim D} \log P_{\theta}(c_{v_m}^{l_i} | w^{l_i}, v_{\setminus m}) \quad (3.9)$$

To facilitate learning of a joint embedding space between vision and language, we warm up the *image encoder* to make sure the output visual embedding shares the same embedding space as word embeddings. Specifically, each image region is projected to an image region feature $v_i \in \mathbb{R}^p$ through the *image encoder*, with the same dimension as the word embedding vector. We then extract the word embedding vectors from XLM-R that correspond to the k object categories $\mathbf{c} = \{c_1, c_2, \dots, c_k\}$ defined by the object detector. We compute the cosine similarity between the projected image feature v_i with the k word embedding vectors followed by a softmax function, resulting in a normalized distribution $h_{\theta_I}(v_i) \in \mathbb{R}^k$ that indicates the prediction on what semantics are mapped in the region. We then maximize the similarity between this predicted distribution and the ‘‘GT’’ object probability distribution from the object detector output $g(v_i) \in \mathbb{R}^K$, by minimizing their KL divergence:

$$\mathcal{L}_{EA}(\theta_I) = D_{KL}(g(v_i) || h_{\theta_I}(v_i)), \quad (3.10)$$

where θ_I is the learnable parameters of the *image Encoder*.

Visual Translation Language Modeling (VTLM) All the objectives mentioned so far operate on image and *monolingual* input, without considering cross-lingual objectives. The correspondence between languages is vital for cross-lingual generalization (*i.e.*, the zero-shot setting in our experiments), clearly observed from existing work on language understanding [100]. Our proposed methods so far unexceptionally learn cross-lingual correspondence *indirectly* through the image focal point, which might not be sufficient. We hence propose visual translation language modeling (VTLM), which directly and jointly learns the alignment between visual context and text in different languages.

In VTLM, given an image \mathbf{v} and a pair of captions (w^{l_i}, w^{l_j}) in two different languages, the goal is to predict masked caption tokens from both languages. One of the two languages is always English, as English captions in our pre-training data are directly from [99], while captions in other languages are translated by MT, therefore less reliable. Under this bilingual framework, model input size only grows linearly with more languages.

Besides, as our model is initialized with the weights of a powerful pre-trained multilingual model, it has already learned a good alignment between different linguistic words to some extent. Applying random masking strategy in VTLM is sub-optimal, as the model can make a correct prediction by simply translating words from one language to another, without taking into account the visual information from image. To encourage the model to fully consider visual context, we introduce a strategy called *co-masking*, where we simultaneously mask out tokens with similar semantic meanings from paired captions to prevent easy translations.

There are a few steps in *co-masking*. First, we apply Fast Align [103] to learn the word alignment between two different languages (l_i, l_j) from the noisy parallel corpus that was created using machine translation. Then, during the pre-training stage, we follow the same strategy as in MLM to randomly mask a token $w_m^{l_i}$ from the caption of one language. For the paired caption in the other language l_j , we mask the aligned word tokens $w_k^{l_j}$ that are predicted from Fast Align. [103] The final objective is again to predict masked tokens from both languages by minimizing the negative log-likelihood:

$$\mathcal{L}_{VTLM}(\theta) = -\mathbb{E}_{(w^{l_i}, w^{l_j}, v) \sim D} \log P_{\theta}(w_m^{l_i}, w_k^{l_j} | w_{\setminus m}^{l_i}, w_{\setminus k}^{l_j}, v) \quad (3.11)$$

3.2.4 Result

We first compare UC² to various SOTA with or without pre-training on the two downstream tasks. Then, We conduct ablation experiments to study the effectiveness of MRTM and VTLM, as well as the impact of the number of languages used for pre-training. Finally, we visualize the alignments between visual context and cross-lingual text context learned by our pre-trained UC² model.

Evaluation on Multilingual Retrieval

We compare UC² with state-of-the-art methods on image retrieval and text retrieval in two different settings:

- **Cross-lingual Zero-Shot Transfer:** Assume we only have training data in English for downstream task, we evaluate the ability of pre-trained UC² in transferring learned knowledge from English to other languages.

Method	Flickr30K				MSCOCO			Meta-Ave
	EN	DE	FR	CS	EN	ZH	JA	
<i>SOTA without pre-training</i>								
EmbN[104]	72.0	60.3	54.8	46.3	76.8	73.2	73.5	65.3
PAR.EmbN [105]	69.0	62.6	60.6	54.1	78.3	76.0	74.8	67.9
S-LIWE [85]	76.3	72.1	63.4	59.4	80.9	73.6	70.0	70.8
MULE [86]	70.3	64.1	62.3	57.7	79.0	75.9	75.6	69.3
SMALR [87]	74.5	69.8	65.9	64.8	81.5	77.5	76.7	73.0
<i>Cross-Lingual Zero-Shot Transfer</i>								
M ³ P[106]	86	48.8	39.4	38.8	87.4	55.8	54.4	58.7
UC ²	87.2	74.9	74	67.9	88.1	82	71.7	78.0
<i>Translate-Test</i>								
UNITER _{CC} [78]	87.7	81.2	81.9	80.2	88.4	87.3	82.2	84.1
<i>All-Language</i>								
M ³ P[106]	86.7	82.0	73.5	70.2	88.0	81.8	86.8	81.3
UC ²	88.2	84.5	83.9	81.2	88.1	89.8	87.5	86.2

Table 3.3: Evaluation results on image-text retrieval over Flickr30K and MSCOCO datasets across different languages. We highlight the MSCOCO results for MULE and SMALR in blue as they are using different dev/test splits of MSCOCO compared to other models.

- **All-Language:** We finetune the pre-trained model on merged training data of all languages.

Besides reporting AR on each language, we also compute the Meta-Ave (average of AR across all languages over two datasets) to reflect the overall performance in this task. Given that we have access to pre-trained machine translation models, we also introduce a strong translate-test baseline UNITER_{CC} based on [78], which is pre-trained on Conceptual Conception English data and finetuned on English training data in the downstream tasks. By translating the test data from other languages to English, UNITER_{CC} can be directly applied for text/image retrieval. Results are summarized in Table 3.3.

Our model on the all-language setting achieves a significant improvement over all task-

specific methods without pre-training, showing the effectiveness of cross-lingual cross-modal pre-training in learning universal representation across vision and different languages. Our model also demonstrates a superior transferability. When finetuned on English dataset only, we observe an absolute gain of 19.3% on Meta-Ave across different languages over M3P via better transition of the learned knowledge from English to other languages. Compared to the best non-pretrained models trained on data in each language, our cross-lingual model under the zero-shot setting is still 5% better. We suspect the improvement comes from the in-domain pre-training objective: we use image as the grounding media in ITM to learn cross-modal mapping from one language to another. With strong transfer capability, our model could potentially generalize the learned knowledge from a high-resource language to downstream tasks in low-resource languages. When we finetune UC² model on all-language data, our model still demonstrates a consistent advantage over M³P on all languages, with 5% improvement on Meta-Ave. Our model has a noticeable advantage over M3P in French, Czech, and Chinese with much less finetuning data than the other three languages. This again proves that tasks in lower-resource languages can be significantly improved with our pre-trained model. Our best model is also better than the strong translate-test baseline UNITER_{CC} on all languages except English in MSCOCO. The slightly worse performance on COCO English is potentially due to lack of pre-training in English data, given that our pre-training time is evenly splitted to multiple languages. However, this does not overshadow the fact that we achieve overall better performance across all languages. Thanks to the cross-lingual pre-training and finetuning, our model can leverage the complementary information captured in different languages to improve the performance on each language.

Evaluation on Multilingual VQA

For multilingual VQA, our pre-trained model is finetuned and evaluated on the target language for each dataset. Unlike image-text retrieval where the same output layer is shared across different languages, multilingual VQA has different classes of answers for each language, which makes joint training across different languages impossible. We compare our model with state-of-the-art methods without pre-training as well as V+L pre-training

method	VQA v2.0	VG VQA JA	
	Test-Dev Acc	Acc	BLEU
MCAN [107]	70.63	-	-
PCATT [108]	-	19.2	-
Vil-BERT [74]	70.55	-	-
VL-BERT [76]	71.16	-	-
UNITER _{CC} [78]	71.22	22.7	11.8
UC ²	71.48	34.2	26.8

Table 3.4: Evaluation results on multilingual VQA task over VQA v2.0 and VG VQA Japanese datasets. We highlight the results for PCATT in blue as they are using different dev/test splits.

methods that use the same pre-training corpus. When evaluating the translate-test baseline UNITER_{CC} on VG VQA Japanese dataset, we first finetune it on VQA v2.0 [109] with english answer candidates translated from VQA VG Japanese to ensure the same reference is used during evaluation as in UC². We then use machine translation model to translate the test dataset of VG VQA Japanese to English, and evaluate the finetuned translate-test model using classification accuracy and BLEU. Results are summarized in Table 3.4.

On VQA v2.0, our model achieves significant improvement over SOTA task-specific method, and also outperforms existing monolingual models pre-trained on Conceptual Conception [74, 76, 78] by an obvious margin. On VG VQA Japanese, we finetune our model with a different data split from the original baseline method PCATT proposed in VG VQA Japanese, where we have much less training data than their split (ours: 61K images vs. PCATT: 91K images). Even under this disadvantage in an unfair comparison, our pre-trained model still achieves more than 10% improvement on both accuracy and BLEU over baselines. Although achieving better performance compared to the task-specific method, the translate test baseline (UNITER_{CC}) performs much worse than UC² on the translated VQA VG Japanese dataset. Despite strong performance on the VQA English dataset, the noisiness from the machine translated language would lead to un-

Objectives	Flickr30K				MSCOCO				VQA v2.0	VG VQA	JA
	EN	DE	FR	CS	EN	ZH	JA	Meta	Acc	Acc	BLEU
UC ² (full model)	88.2	84.5	83.9	81.2	88.1	89.8	87.5	86.2	71.48	34.2	26.8
-VTLM	87.5	83.6	82.4	79.6	87.7	89.2	87.2	85.3	71.45	34.1	26.7
-VTLM-MRTM	86.8	82.9	81.3	79.3	87.5	88.9	86.7	84.8	69.94	33.4	26.4

Table 3.5: Ablation study on pre-training objectives.

Topology	Flickr30K				MSCOCO				
	EN	DE	FR	CS	EN	ZH	JA	Meta-Ave	
UC ² (Image pivoting)	87.5	83.6	82.4	79.6	87.7	89.2	87.2	85.3	
UC ² (English pivoting)	86.2	81.9	80.7	77.4	88.1	88.5	87.3	84.2	

Table 3.6: Comparison between the pre-training topology of pivoting on image against pivoting on English.

avoidable degradation especially for tasks like VQA that requires fine-grained level understanding and interpretation on multi-modal context. Hence, building unified cross-lingual cross-modal pre-training model like UC² is a better solution to directly work on tasks in target languages than a translate-test method.

Ablation Study

Effect of Training Objectives To validate the effectiveness of the proposed pre-training objective MRTM and VTLM, we conduct ablation study to verify their contributions to the model performance. We gradually remove the two proposed training objectives and evaluate these ablated models on our two downstream tasks. When finetuning the pre-trained model on the image-text retrieval task, we follow the best experimental setting to train the model on all language data. On VQA task, the model is directly finetuned on the target language data.

From Table 3.5, we observe that MRTM has led to significant performance boost on multilingual VQA tasks over the two languages while gaining some incremental improvement on image-text retrieval tasks. VQA requires more fine-grained understanding about connections between language and visual context, therefore benefits more from the cross-modal local alignment captured by MRTM. After adding VTLM, we can further improve

image-to-text retrieval tasks by 1% via leveraging cross-modal cross-lingual parallel corpus to learn more substantial joint alignment.

Effect of Pivoting on Image To validate the effectiveness of image pivoting, we conduct a controlled experiment where the model variant only pivots on English. In this setting, we train UC2 with all the pre-training objectives on English Conceptual Caption data, except for VTLM which involves image as one of the pivoting points. To capture the alignment between English and other languages, we train UC² on pairs of captions in two different languages with one language fixed as English. The training objective is translated language modeling adopted from XLM [110]. From Table 3.6, we can see that UC² pre-trained by pivoting on image achieves overall better performance in multilingual image-text retrieval task. The advantage is particularly sound when the target language has limited training data. This indicates that the cross-lingual cross-modal representation learned by pivoting on images imbues stronger cross-modal mapping transfer across different languages.



Figure 3.5: Visualization of Text-to-Image Attention on aligned words across English, German and Czech (Flickr30K).

Visualization To visualize the cross-lingual cross-modal alignment learned by UC², we provide examples of text-to-image attention from salient words in multilingual captions to the shared image context. As shown in figure 3.5, words from different languages that share the same semantic meaning can attend to similar corresponding regions in the

image. This shows that while our model can effectively capture cross-modal alignments between regions and words, it also connects different languages by grounding them to similar image regions.

Pre-training tasks	ITR Meta-Ave	VQA EN	VQA JA
ITM+MLM+MRC	85.1	70.60	33.4
ITM+MLM+MRTM	85.3	71.45	34.1

Table 3.7: Direct ablation on comparison between the proposed MRTM and the MRC. The presentation of the result is simplified to only include the Meta-Average for the multilingual image-text retrieval over both Multi30K and MSCOCO, the accuracy on VQA v2.0 test-dev split (referred as VQA EN), and the accuracy on VQA VG Japanese (referred as VQA JA).

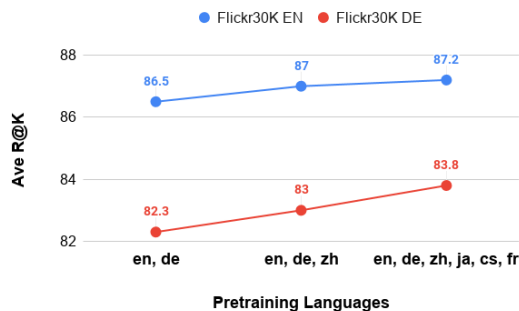


Figure 3.6: Comparison of image-text retrieval performance when pre-trained with different groups of languages (average R@K on Flickr30K English and German).

MRTM vs MRC For this ablation, we pre-train UC² with ITM, MLM and MRC and compare the results to the pre-trained UC² optimized with ITM, MLM and MRTM. The results is summarized in Table. 3.7. As shown in table 3.7, compare the pretrained UC² that employs the traditional task MRC and the one that employs our proposed MRTM, we can see that the performance on the image-text retrieval task are similar, but MRTM leads to marginal improvement on the multilingual VQA tasks. This observation is consistent with our hypothesis that the proposed MRTM augments the local alignment between image regions and the words in different languages which benefits downstream tasks that rely on region-level recognition and reasoning.

Effect of Pre-training languages As we use machine translation models to expand the pre-training corpus, theoretically, we can have as many languages as needed. We conduct further experiments to verify the impact of number of languages included in pre-training data. We create three variants of pre-training corpus, where the number of languages are 2, 3, and 6, respectively.² Every corpus contains English and German. We add Chinese to construct the corpus with 3 languages, and the corpus with 6 languages contains all the languages used to pre-train our full model. The pre-trained models are evaluated on image-text retrieval task in English and German, by finetuning on target language.

Figure 3.6 shows that when the number of pre-training languages increases, the performance on image-text retrieval on different languages (English and German) slightly improves. This result demonstrates that cross-lingual cross-modal pre-training can effectively leverage different languages to learn stronger vision-to-monolingual-sentence alignment. Meanwhile, as we maintain the same pre-training epochs for all three experiments, we also observe that the benefit of multilingual V+L pre-training is compensating for the reduced training time allocated to each language. Although more comprehensive analysis in future study can help us better understand the trade-off between language capacity and performance on downstream tasks, our observation to some extent still suggests that our model is scalable to pre-training on a large corpus with many languages within a reasonable time frame.

Effect of Pivoting Language in VTLM We also conducted a controlled experiment to learn the effect of different pivoting languages in VTLM for the multi-lingual multi-modal pre-training. In this controlled experiment, we pre-train UC2 with all the objectives but change the pivoting language in VTLM from English to Chinese. When we evaluate the pre-trained model on the multilingual image-text retrieval task, the meta-ave score for the pre-trained model with VTLM pivoted on Chinese is dropped from 86.2 to 85.5. This to some extent suggest that English is a more optimal pivoting language to learn the cross-lingual cross-modal shared representation space. Another potential reason for the

²For fair comparison, we constraint the training time to be the same with different pre-training corpus.

limited performance is due to the noisiness in the pre-trained Chinese captions gained via automatic machine translation. To gain more solid conclusion to determine the optimal pivoting language, more comprehensive experiments need to be conducted in the future work.

3.3 Unsupervised Cross-modal Representation Learning

3.3.1 Introduction

Vision-and-Language pre-trained (VLP) models [111, 112, 113, 75, 114, 115, 116, 117, 118, 119, 120, 121, 122] that learn the joint cross-modal representation have revolutionized the research on various vision-and-language tasks in recent years. However, the success of VLP models relies on the availability of a large-scale aligned image-text corpora. The widely used crowd-sourced pre-training datasets such as MS COCO [123, 124] and Visual Genome [125] require expensive human annotations which are hard to scale up. Recently, the web crawled image-text datasets like Conceptual Captions 3M [126] and CC12M [127], and SBU Captions [128], have dramatically reduced the need for massive human annotation but still require heavy post-cleaning procedures to get aligned image-text pairs. In comparison, the language corpora and image collection are readily available from the web. The convenience of getting a large-scale single-modality data has benefited the self-supervised learning of vision [129, 130, 131] and language [132, 133] domains respectively. This raises a question: Can we take advantage of easily-accessible large-scale single-modality data to perform unsupervised V+L pre-training without parallel text and images (UVLP)?

We define UVLP as follows: given the crawled image collection $\mathbf{I} = \{\mathbf{i}_1, \dots, \mathbf{i}_{n_I}\}$ and text corpus $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{n_T}\}$, we aim to pre-train a multi-modal model from such data. U-VisualBERT[134] is the first UVLP work, where the authors have trained their model on un-aligned text and image data in a round-robin fashion and simply use object tags as an anchor point to bridge the gap between the two modalities. Their research demonstrates that a shared multi-modal embedding can be learned by just presenting a single modality

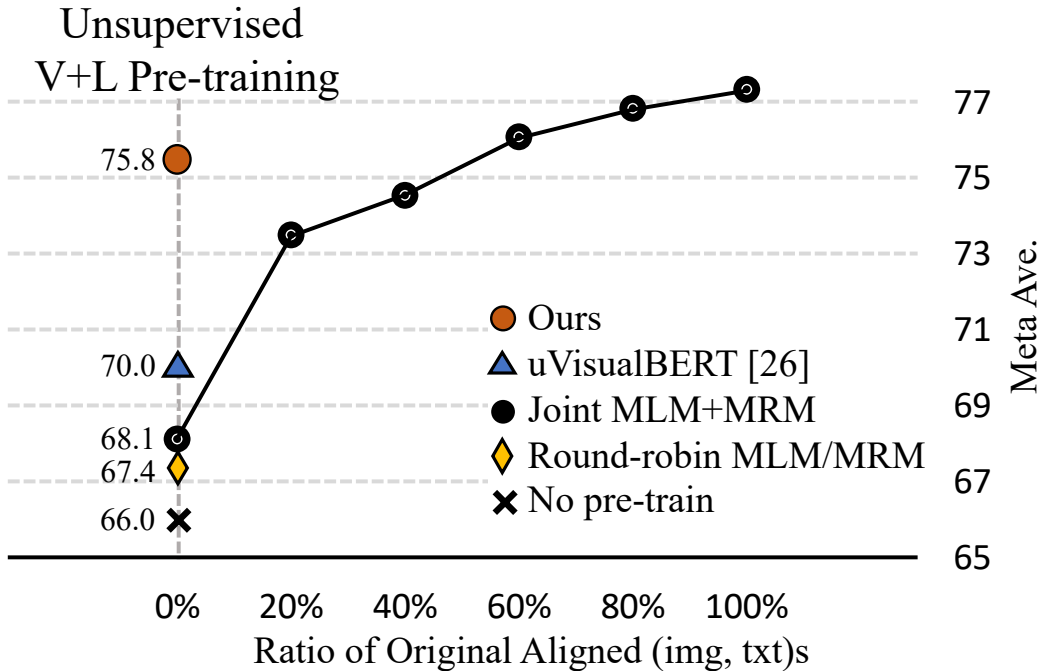


Figure 3.7: Meta average scores of VQA, NLVR2, VE, and RefCOCO+ fine-tuned from different pre-trained models. All pre-training are conducted on Conceptual Captions (CC) with different ratio of parallel data, i.e., a fixed amount of data is originally aligned while the rest is randomly shuffled. 0% refers to the case of unsupervised V+L pre-training. We also plot the performance of our proposed approach against U-VisualBERT[134]. Breakdown of the accuracy of each task is listed in the supplementary file.

at a time. This however introduces an input discrepancy between pre-training and fine-tuning stages as each downstream V+L task requires both modalities (image, text) as the input. In this work, we investigate (i) whether presenting a joint image-text data from non-parallel data would improve the learned joint embedding space. Furthermore, (ii) if joint image-text data is fed into the model, how does its latent alignment affect the cross-modal representation learning?

To explore these two questions, we simply use the images and captions from Conceptual Captions (CC) dataset [126] as independently collected uni-modal corpus and perform the following analysis. First, we compare the pre-trained model’s performance between the two data input strategies: one is presenting one image or text at a time (round-robin) and the other is presenting a concatenation of a pair of randomly sampled image and text

(0% alignment ratio). Second, we prepare five sets of image-text pairs from Conceptual Captions with different levels of pairwise alignment by controlling the ratio of original aligned image-text data from 20% to 100% (while the remaining is randomly sampled from each modality). A single-stream transformer is used for all experiments with the standard pre-training objectives: masked language modeling (MLM) on language input and masked region modeling (MRM) on vision input. After pre-training, we adapt the model to a series of four downstream V+L tasks, including VQA [135], NLVR2 [136], Visual Entailment (VE) [137], and RefCOCO+ [138]. The performance is measured as the meta average of all tasks after fine-tuning. The results are summarized in Fig. 3.7. From Fig. 3.7, it is clear that joint MLM+MRM learning outperforms round-robin MLM/MRM. Such gains show that **joint image-and-text input is necessary for UVLP** even when the input is un-aligned. We also observe a strong positive correlation between the alignment of image-text pairs and the meta average of the fine-tuned downstream tasks of the resulting model. This conveys a seemingly intuitive but quite important message that **the more aligned the image-text data is the better the pre-trained model performs**.

Inspired by these analyses, we propose Unsupervised Vision-and-Language Pre-training via Retrieval-based Multi-Granular Alignment (μ -VLA), which uses our novel unsupervised V+L pre-training curriculum for non-parallel data. We first construct a weakly-aligned image-text dataset via retrieval. Given an image, we take its detected object tags as the reference sentence and retrieve the closest sentences from the text corpus via sentence BERT embedding [139] similarity. Though the constructed pairs are noisy, the mere weak alignment of concepts is key to learning the latent alignment. We propose to let the model gradually learn a multi-granular alignment, i.e., region-to-object tag level, region-to-noun phrase level, and image-to-sentence level to more effectively bridge the gap between the two modalities. We show how each granularity learned from the weakly-aligned pairs contributes to the final pre-trained model’s performance. Experiments show our approach achieves the state-of-art performance (in Fig. 3.7), with a clear gain over [134] on the 4 downstream tasks.

Towards practical applications, we also validate the effectiveness of our approach un-

der a more realistic setting, where the images are from CC and the captions are from BookCorpus (BC) [140]. Similar performance gains are achieved in this harder setting, showing the robustness of our approach.

To summarize, our contributions are three-fold: (i) We analyze what leads to a good unsupervised V+L pre-training and found two key factors: joint image-and-text input, and overall alignment between image-text pairs. (ii) Accordingly, we propose a novel retrieval-based pre-training curriculum, which applies multi-granular alignment pre-training tasks between weakly aligned image-text pairs to bridge the gap between the two modalities. (iii) We provide comprehensive experiments and analyses showing the robustness of our approach when compared to SOTA supervised and unsupervised V+L pre-training methods.

3.3.2 Related Work

Vision-and-Language Pre-training Inspired by the success of natural language processing [132, 141], there is a recent surge of interest in pre-training for vision and language. For example, there are different architectures (*e.g.*two-stream models [113, 75, 111, 114, 121, 142] vs. single-stream models [143, 115, 116, 112, 120]), features (*e.g.*regions [144] grids [118]), backbones (*e.g.*ConvNets [118] Transformers [117]). All these works aim to exploit the large-scale aligned image-text corpora [123, 125, 126, 128, 142] to pre-train a powerful multi-modal model, which is then adapted to various downstream V+L tasks, such as VQA [135], NLVR2 [136], Visual Entailment (VE) [137], Referring Expression Comprehension [138], and Image-Text Retrieval.

Various pre-training tasks have been introduced to achieve this, including the most notable Masked Language Modeling (MLM), Masked Region Modeling (MRM), and Image-Text Matching (ITM). Several other variants have also been explored, such as predicting the object tags [145, 146], sequence generation [147, 119], word-region alignment [112]. In this paper, we propose learning a multi-granular alignment between word and region, phrase and region, and image and sentence to better bridge the gap between vision and language.

Unsupervised Vision-and-Language Pre-training without Parallel Data Inspired by the works on multi-lingual contextual language modeling [148, 149, 150, 151], U-VisualBERT [134] first propose the *unsupervised* vision-and-language pre-training without parallel data (UVLP). U-VisualBERT [134] conducts the masked prediction on the text-only and image-only corpora and introduce the object tags as anchor points to bridge the two modalities. The authors treat the tags as a sentence when performing MLM, where tags provide alignment with the regions in a picture and implicitly learn a tag-region-level alignment. However, the anchor tags are still quite different from the text input, missing the sentence completeness and naturalness. Besides, the latent cross-modal alignment is shown to be important in our analysis (from Fig. 3.7). As comparison, our pre-training involves a retrieval-based weakly aligned V+L data construction and learns a more comprehensive multi-granular cross-modal alignment. With same data as U-VisualBERT, our approach achieves a clear and consistent gain across all the downstream tasks in our experiments.

3.3.3 Model

In this section, we introduce the two core components of our μ -VLA’s architecture for unsupervised V+L pre-training without parallel data: (1) construct a weakly aligned image-text corpus from independent vision and language data sources; (2) our novel pre-training curriculum to enable the model to capture the cross-modal alignment on three granularity including region-to-tag level alignment (RT), region-to-noun phrase level alignment (RN), and image-to-sentence level alignment (IS).

Model Overview

We use the well-known single-stream model architecture for our experiments as [143, 115, 116, 112, 120]. As shown in Fig. 3.8, our main backbone is a single transformer, where we feed the concatenation of visual embeddings of an image and the tokens of a caption as its input. Given an image \mathbf{i} , we first use an off-the-shelf Faster R-CNN (VinVL [152]) to detect the objects $\mathbf{v} = \{v_1, \dots, v_{k^v}\}$. The visual embedding of each region is then encoded as the

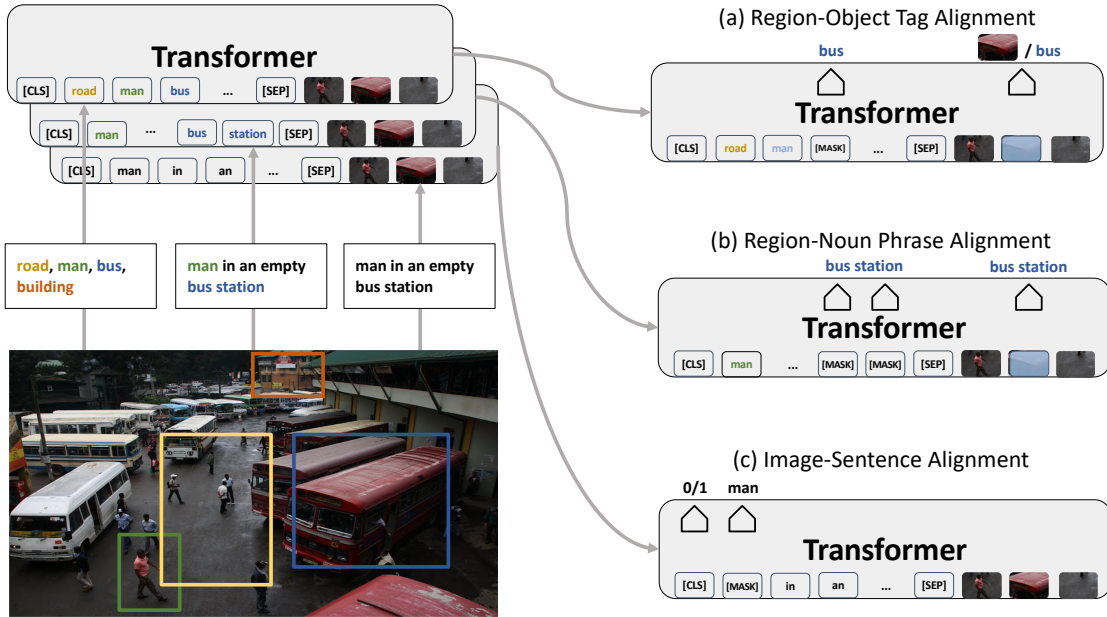


Figure 3.8: Overview of our method. On the left we form three types of image-text pairs as input data to learn cross-modal alignment on three different granularities: region-tag alignment, region-phrase alignment, and image-text alignment. The models is iteratively pre-trained on each granularity and the model parameters are shared. On the right-hand side, we demonstrate the details of the pre-training objectives for each granularity.

sum of its regional feature, its location embedding³, and the modality embedding. For a given caption \mathbf{t} , we denote its tokenized sequence as $\mathbf{t} = \{t_1, \dots, t_{k^t}\}$. After multiple layers of self-attention, the two modalities are fused together and the output hidden vectors can be used for various pre-training tasks.

Weakly-aligned Image-Text Corpus

As in the analysis of Sec 3.3.1, we observe a strong correlation between the degree of image-text alignment in the training data and the performance of the pre-trained model. Inspired by this finding, we believe it important to initialize some weak semantic alignment between the two modalities as the input source. Specifically, we retrieve k sentences that are semantically closed to a given I_i . Previous work [153] shows the visually grounded caption covers a good ratio of words that are naturally related to specific visual contents,

³The 5-dimensional vector $[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(y_2-y_1)(x_2-x_1)}{W.H}]$ is projected to the visual embedding space. $(x_1, y_1), (x_2, y_2)$ are the coordinates of the top left and bottom right point of the detected region, and W, H are the image width and height.

e.g. concrete nouns. Thus, we utilize the semantic association between the objects that appear in the image and a candidate sentence as the indicator to measure the alignment degree.

Specifically, we take the object tags $\mathbf{o} = \{o_1, \dots, o_{k^o}\}$ from the above detected \mathbf{v} and feed the sequence into an off-the-shelf sentence BERT embedding model [139] to obtain the query embedding \mathbf{e}_o . Similarly, we feed each candidate sentence into the same model getting the candidate embedding \mathbf{e}_t . We retrieve the top K candidates with the highest cosine similarity score to form an initial weakly-aligned image-text pairs for a given image \mathbf{i} . We denote the retrieved captions as $\{\mathbf{t}^r(\mathbf{i})\}_{r=1}^K$ and the overall weakly aligned corpus as \mathbf{R} .

Pre-training Tasks

In this subsection, we introduce a set of pre-training objectives that we designed to facilitate the model to capture the different levels of vision and language alignment. Fig. 3.8 shows the overview of our model and its pre-training tasks.

Region-Tag Alignment Learning We first propose to align the object tags onto the image regions. As shown in Fig. 3.8(a), We concatenate the object tags detected from each image with its source image to form an input pair $[\mathbf{o}, \mathbf{v}]$ fed into the model. We denote the mask indices as $\mathbf{m} \in \mathbb{N}^{M^4}$. We randomly mask out the object tags and regions, and apply masked language modeling (MLM) and masked region modeling (MRM) for the pre-training.

Specifically, MLM on the object tags is formulated as

$$\mathcal{L}_{\text{MLM}}^{\text{R-T}} = -\mathbb{E}_{(\mathbf{o}, \mathbf{v}) \sim \mathbf{I}} \log P(\mathbf{o}_{\mathbf{m}} | \mathbf{o}_{\setminus \mathbf{m}}, \mathbf{v}),$$

where the goal is to predict the masked object tags based on the observation of their surrounding tags $\mathbf{o}_{\setminus \mathbf{m}}$ and image regions \mathbf{v} . On the vision side, MRM includes both masked region classification loss (MRC) and masked region feature regression loss (MRFR):

$$\mathcal{L}_{\text{MRM}}^{\text{R-T}} = \mathbb{E}_{(\mathbf{o}, \mathbf{v}) \sim \mathbf{I}} [f_{\text{MRC}}(\mathbf{v}_{\mathbf{m}} | \mathbf{v}_{\setminus \mathbf{m}}, \mathbf{o}) + f_{\text{MRFR}}(\mathbf{v}_{\mathbf{m}} | \mathbf{v}_{\setminus \mathbf{m}}, \mathbf{o})].$$

⁴ \mathbb{N} is the natural numbers, M is the vocabulary size, and \mathbf{m} is the set of masked indices.

Between the two, MRC learns to predict the object semantic class for each masked region $c(\mathbf{v}_m)$. We feed the last hidden output of the masked region \mathbf{v}_m into a FC layer and softmax function to predict the classification probabilities $g_\theta(\mathbf{v}_m)$. The objective is to minimize the cross-entropy of $f_{\text{MRC}}(\mathbf{v}_m|\mathbf{v}_{\setminus m}, \mathbf{o}) = \text{CE}(c(\mathbf{v}_m), g_\theta(\mathbf{v}_m))$. MRFR learns to regress the transformer output of each masked region \mathbf{v}_m to its visual features. We apply a FC layer to convert its hidden output to a vector $h_\theta(\mathbf{v}_m)$ of the same dimension as the input regional feature $r(\mathbf{v}_m)$. We apply L2 regression: $f_{\text{MRFR}}(\mathbf{v}_m|\mathbf{v}_{\setminus m}, \mathbf{o}) = \|h_\theta(\mathbf{v}_m) - r(\mathbf{v}_m)\|_2^2$.

For region-tag alignment learning, we have our pretraining objective function as

$$\mathcal{L}^{\text{R-T}} = \mathcal{L}_{\text{MLM}}^{\text{R-T}} + \mathcal{L}_{\text{MRM}}^{\text{R-T}}$$

Region-Noun Phrase Alignment Learning Due to the small vocabulary size of object tags, the region-tag alignment learning can only capture a limited amount of localized concepts. To increase the diversity of concepts, we propose to align the noun phrases from the retrieved sentences to the corresponding regions as well. As in Fig. 3.8(b), given an image \mathbf{i} and its retrieved weakly aligned caption $\mathbf{t}^r(\mathbf{i})$, we first detect the noun phrases from the caption using spacy [154]. Note the detected noun phrases sometimes contain the attribute words, which further benefits this pre-training task. We link the noun phrase to its closest visual region by computing the word2vec similarity between the phrase and object tag (associated to each region). The pre-training still consists of MLM and MRM but are performed with different masking strategy and supervision signal.

Specifically, for both MRM and MLM, we only mask the linked noun phrases from the caption or the linked object regions. We make the masking probability proportional to the linked similarity score. Each time we only mask out one modality (phrase or region) to encourage it to be recovered by its linked content. The region-to-phrase MLM is then formulated as $\mathcal{L}_{\text{MLM}}^{\text{R-P}} = -\mathbb{E}_{(\mathbf{v}, \mathbf{t}^r) \sim \mathbf{R}} \log P(\mathbf{t}_m^r | \mathbf{t}_{\setminus m}^r, \mathbf{v})$.

On the vision side, we propose using the phrase-guided masked region-to-token classification (p-MRTC) on the masked regions:

$$\mathcal{L}_{\text{MRM}}^{\text{R-P}} = \mathbb{E}_{(\mathbf{v}, \mathbf{t}^r) \sim \mathbf{R}} f_{\text{p-MRTC}}(\mathbf{v}_m | \mathbf{v}_{\setminus m}, \mathbf{t}^r),$$

where we directly classify the masked region to its linked noun phrase (sub-word tokens) in BERT vocabulary. Enlarging the vocabulary has shown to be beneficial to MRM [155]. Our proposed p-MRTC leverages the additional noun-phrase to encourage more diverse local region to language alignment.

For region-noun phrase alignment learning, we have our pretraining objective function as

$$\mathcal{L}^{\text{R-P}} = \mathcal{L}_{\text{MLM}}^{\text{R-P}} + \mathcal{L}_{\text{MRM}}^{\text{R-P}}$$

Image-Sentence Alignment Learning We apply image-text matching (ITM) objective as the previous supervised V+L pre-training research [112, 115] to learn the cross-modal sentence-level alignment. As in Fig. 3.8(c), given an input pair $[\mathbf{v}, \mathbf{t}^r]$, the final hidden vector of the special token [CLS] is fed through a FC layer to output a single score $s_\theta(\mathbf{v}, \mathbf{t}^r)$, which predicts if the given image-text input is a semantically matched pair or not. We use the label $y \in \{0, 1\}$ to indicate if a retrieved pair is a match. The training objective for the ITM task is to minimize the binary cross-entropy loss: $\mathcal{L}_{\text{ITM}} = \text{CE}(y, s_\theta(\mathbf{v}, \mathbf{t}^r))$. On the language side, we also apply standard MLM to help the model learn to align other language tokens besides noun phrases and object tags to the visual context. The objective function is then formulated as $\mathcal{L}_{\text{MLM}}^{\text{I-S}} = -\mathbb{E}_{(\mathbf{v}, \mathbf{t}^r) \sim \mathbf{R}} \log P(\mathbf{t}_m^r | \mathbf{t}_{\setminus m}^r, \mathbf{v})$. The image-sentence level alignment pretraining objective function is

$$\mathcal{L}^{\text{I-S}} = \mathcal{L}_{\text{MLM}}^{\text{I-S}} + \mathcal{L}_{\text{ITM}}$$

Multi-Granular Pre-training Curriculum We propose a multi-granular curriculum to iteratively pre-train the model on the region-to-tag, region-to-noun phrase, and image-to-sentence level. According to our findings in Sec. 3.3.1, learning from image-text pairs with higher degree of cross-modal alignment is beneficial to the performance of unsupervised V+L pre-trained model. Therefore, we propose using an estimated image-text alignment score to guide our multi-granular pre-training. Specifically, we have an ITM header defined in Sec. 3.3.3 to learn the image-text alignment. We also use it to predict matching score as a weight to modulate the input data for each of our retrieval-based pre-training tasks. This allows us to provide more importance to relatively more aligned

image-text pairs over time to help our model to learn better cross-modal alignment on multiple granularities.

To train the alignment model’s ITM classifier, we use our retrieved corpus \mathbf{R} as positive samples and randomly shuffled pairs as negative samples in the first m epochs. This warms up the models to make reasonable estimations on the alignment of image-text input pairs. After m epochs, we start to incorporate the alignment prediction score w_{ITM} in our training objective. To summarize, our multi-granular pre-training loss is

$$\mathcal{L} = \begin{cases} \mathcal{L}^{\text{R-T}} + \mathcal{L}^{\text{R-P}} + \mathcal{L}^{\text{I-S}} & \text{if epoch} < m \\ \mathcal{L}^{\text{R-T}} + w_{\text{ITM}}(\mathcal{L}^{\text{R-P}} + \mathcal{L}^{\text{I-S}}) & \text{if epoch} \geq m, \end{cases}$$

where $\mathcal{L}^{\text{R-T}}$, $\mathcal{L}^{\text{R-P}}$, and $\mathcal{L}^{\text{I-S}}$ are the loss functions for region-tag alignment pre-training, region-noun phrase alignment pre-training, and image-sentence alignment pre-training. We set m as 1 in our final implementation.

Model	VQA2	NLVR2	VE	RefCOCO+			Meta-Ave
	Test-Dev	Test-P	Test	Dev	TestA	TestB	
ViLBERT[113]	70.6	-	-	72.3	78.5	62.6	-
VL-BERT[116]	71.2	-	-	71.6	77.7	61.0	-
UNITER _{CC} [112]	71.2	-	-	72.5	79.4	63.7	-
VisualBERT [143, 134]	70.9	73.9	-	73.7	79.5	64.5	-
Aligned VLP	72.5	75.9	78.7	82.1	86.6	75.0	77.3
Base	70.1	51.2	73.2	69.4	74.8	60.3	65.9
U-VisualBERT[134]	71.8	53.2	76.8	78.2	83.6	69.9	70.0
μ -VLA _{CC}	72.1	73.4	77.3	80.3	85.5	73.7	75.8
μ -VLA _{BC}	71.2	67.1	77.1	79.7	85.0	72.7	73.8

Table 3.8: Evaluation results on four V+L downstream tasks. Our model trained with unaligned data (μ -VLA_{CC}, μ -VLA_{BC}) achieves comparable performance with the supervised model trained with aligned data (Aligned VLP). μ -VLA_{CC} and μ -VLA_{BC} also outperform U-VisualBERT on nearly all tasks.

V+L Alignment	VQA	NLVR2	VE	RefCOCO+			Meta-Ave
	Test-Dev	Test-P	Test	Dev	TestA	TestB	
μ -VLA _{CC} (R-T)	71.7	52.0	75.6	78.7	83.3	70.0	69.5
μ -VLA _{CC} (R-N)	71.4	69.4	76.5	77.4	81.5	68.7	73.7
μ -VLA _{CC} (I-S)	71.6	71.5	76.8	75.7	80.3	67.9	73.9
μ -VLA _{CC} (R-T + R-N)	71.9	72.4	76.4	79.3	84.5	71.7	75.0
μ -VLA _{CC} (R-T + R-N + I-S)	72.1	73.4	77.3	80.3	85.0	73.7	75.8

Table 3.9: Effect of cross-modal alignment on the three types of granularities: region-tag alignment(R-T), region-noun phrase alignment(R-N), and image-sentence alignment(I-S)

3.3.4 Experiment

In this section, we provide the detailed experimental set up to evaluate our proposed μ -VLA against previous supervised and unsupervised VLP models. More specifically, we introduce our pre-training dataset, baselines, and our pre-training setting.

Pre-training Datasets

We prepare the un-aligned data under two different settings: (1) We use images and text separately from Conceptual Captions (CC) [126] ignoring the alignment information; (2) We use images from Conceptual Captions (CC) [126] and text from BookCorpus (BC) [140]. Setting (1) sets up a fair comparison with previous supervised methods by keeping the domain and the quality of training data consistent. Our proposed model trained in this setting is called μ -VLA_{CC}. Setting (2) mimics a more realistic challenge where we have large-scale images and text data from different domains, in particular the text sources are not similar to captions of the images. μ -VLA_{BC} has been trained in this setting.

As introduced in section 3.3.3, for each image we retrieve 5 text data points (captions from CC or sentences from BC) from the text corpus that are semantically similar to the detected objects in the image. This creates weakly-aligned image-text pairs for our pre-training models.

Baselines

We compare the performance of our proposed μ -VLA to the following baselines:

Base Model VisualBERT that is initialized from BERT. It does not undergo any pre-training but is directly fine-tuned on the downstream tasks.

Supervised Pre-trained Models Supervised pre-trained VLP models that are trained only on CC, including VILBERT[113], VL-BERT[116], and UNITER[112]. We also report the numbers on the Supervised VisualBERT implemented in U-VisualBERT[134] that is trained on CC and an additional 2.5 Million text segments from BC. For fair comparison with our proposed method, we also introduce the aligned vision-language pre-training model (Aligned VLP) that is pre-trained on the 3M (image, caption) pairs from CC and 3M (image, object tag) pairs.

Unsupervised Pre-trained Models U-VisualBERT is pre-trained on individual image or text corpus in a round-robin fashion and captures the cross-modal alignment by using detected object tags as the anchor point. For fair comparison, we re-implemented this method to pre-train with the VinVL object features[156] and BC.

Training Setup

Our transformer architecture consists of 12 layers of transformer blocks, where each block has 768 hidden units and 12 self-attention heads. We initialize the model from BERT_{base} and pre-train for 20 epochs on their respective pre-training datasets with a batch size of 480. The region features for images are obtained from the pre-trained VinVL object detectors [156]. We use Adam optimizer [157] with a linear warm-up for the first 10% of training steps, and set the peak learning rate as 6e-5. After warm up, a linear-decayed learning-rate scheduler gradually drops the learning rate for the rest of training steps. All models were trained on 4 NVIDIA A100 GPUs, with 40GB of memory per GPU using MMF[158]. The pre-training takes 3 days. We evaluate our pre-trained models on four downstream tasks: Visual Question Answering (VQA 2.0)[144], Natural Language for Visual reasoning[136] (NLVR²), Visual Entailment[137] (VE), and Referring Expression[138] (RefCOCO+). Detailed training settings for each task can be found in our supplementary material.

Experimental Results

We first compare μ -VLA to various supervised models that are pre-trained on CC and to the state-of-the-art unsupervised V+L pre-training method, U-VisualBERT on the four downstream tasks. Besides reporting scores for each individual task, we also compute the meta-average score to reflect the overall performance across all tasks. The results are summarized in Table 3.8.

It is clear from Table 3.8 that both μ -VLA_{CC} and μ -VLA_{BC} outperform the Base model by a large margin on all benchmarks. It also achieves better performance than existing supervised models like ViBERT[113], which is potentially due to the usage of better visual regional features of VinVL [152]. When compared to Aligned VLP, which is trained with the same architecture and visual features, our model is only slightly worse. This shows the effectiveness of our proposed pre-training curriculum which can learn comparable universal representation across vision and language as the supervised models without any parallel image-text corpus. Our μ -VLA also achieves consistently better performance than the previous UVLP method: U-VisualBERT. This improvement shows how our proposed cross-modal alignment pre-training curriculum effectively bridges the gap across the two modalities. In particular, our model outperforms U-VisualBERT in the task of NLVR2 by more than 20%. As NLVR2 is known to benefit more from image-sentence cross-modal alignment from previous supervised V+L pre-training research [112], this observation indicates that our model is able to capture the instance-level cross-modal alignment without parallel data. When μ -VLA is trained on BC text and CC images μ -VLA_{BC}, it still achieves comparable or better performance than U-VisualBERT except for VQA. The slight advantage U-VisualBERT has over μ -VLA_{BC} in VQA is potentially due the similar style between the VQA text and the pre-trained CC captions. However, this does not overshadow the overall better performance of μ -VLA. It shows that our proposed method is more robust than U-VisualBERT training on the uni-modal datasets collected from separate domains, which makes it more useful in practical settings.

Ablation Study on Multi-Granular Alignment We conduct ablation study to verify the effectiveness of the three types of visual-language alignment for unsupervised V+L

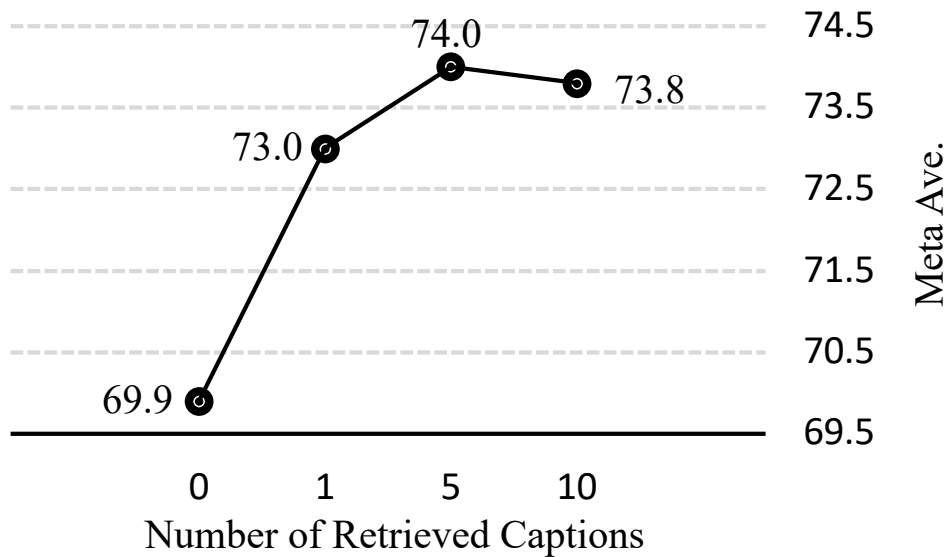


Figure 3.9: Meta average scores of non-parallel V+L pre-training with different number of retrieved candidate sentences.



	Caption: short haired woman is coughing .
	Objects: wall couch sofa woman girl pant shirt arm leg hair sleeve face
	Retrieved from <i>Conceptual Caption</i>: <ol style="list-style-type: none"> 1. portrait of beautiful young woman looking at the camera while sitting on the sofa in the living room 2. young woman sitting on a couch looking over her shoulder 3. close - up shot of a girl in headscarf sitting on the sofa and looking to the camera .
	Caption: this living room has a lot of energy and excitement with the combination of bright orange and blue .
	Objects: room wall pillow blanket table couch coffee table cushion vase toy picture
	Retrieved from <i>Book Corpus</i>: <ol style="list-style-type: none"> 1. in front of the sofa was a rug that was a cyan blue and matched the throw pillows on the sofa and vases on the dining table and entrance table. 2. there was a lovely patterned rug under the coffee table and little pillows on the couch. 3. there is also a small table at the end of the sofa.

Figure 3.10: Examples of retrieved text from both CC and BC. The covered grounded noun phrases in retrieved sentences are highlighted in green bar for positive examples.

pre-training, namely region-tag alignment (R-T), region-noun Phrase alignment (R-N), and image-sentence alignment (I-S). We first evaluate each individual type of alignment to measure its usefulness for different downstream tasks. Then, we gradually add each type of alignment into the UVLP. For this ablation study we pre-train μ -VLA on CC

images and text, and the results are summarized in Table 3.9.

From Table 3.9, we can see that aligning local regions to object tags (R-T) and noun phrases (R-N) are especially helpful for the task of RefCOCO+, which requires the model to understand specific objects that natural expressions describe. Meanwhile, aligning the image and sentence at instance-level (I-S) benefits NLVR2 and VE. Especially on NLVR2, the model that captures the global vision and language alignment μ -VLA_{CC} (I-S) obtains 19.5% gain over the model that only learns the local alignments between regions and object tags μ -VLA_{CC} (R-T). This observation is consistent with previous research [112], where the performance of model on NLVR2 is boosted after introducing pre-training objectives that capture the cross-modal alignment in the image-text pairs. Our results demonstrate that even with just weakly-aligned sentences, we can still effectively learn the instance-level cross-modal alignment. Combining the region-tag and region-noun phrase alignment (R-T+R-N) for UVLP, we observe that these two types of grounding and matching compensate each other. μ -VLA_{CC} (R-T+R-N) shows a marginal but consistent improvement over models that only learn a single type of local region-to-language alignment (R-T, R-N). After adding object-phrase level alignment we can further improve the performance on NLVR2 and VE, which gives us our best performing model μ -VLA_{CC} (R-T + R-N + I-S).

Ablation Study on Number of Retrieved Candidates We conduct experiments to verify the impact the number of retrieved candidate text for each image has on the performance. We create three variants of pre-training corpus, where the number of retrieved candidate are 1, 5, and 10 based on the rank of the similarity of each candidate text to the query image’s detected object tags. The candidate text is sampled from CC. We pre-train our μ -VLA model with only the pre-training objectives to capture the sentence-image alignment (I-S). For each variant of pre-training corpus, we train the model for the same number of steps. We compute the meta average score for the three resulting pre-trained models and visualize them in Fig. 3.9.

Fig. 3.9 shows that retrieving more than one candidate text for an image greatly benefits the pre-trained model to learn a better joint representation between vision and

language, demonstrated by stronger performance in the downstream tasks. We suspect this is because the closeness between the candidate caption and the detected object tags in language embedding space does not always mean a better alignment between the candidate caption and the image. A better and more semantically similar caption candidate for the image could be found in the other caption candidates. However, when we increase the number of candidate captions to 10, we observe a slight drop on the overall performance compared to the model that is pre-trained on corpus with 5 candidate captions. This indicates that having too many candidate captions to form the weakly-aligned pairs with the query image for V+L pre-training may also introduce unnecessary noise. Hence, we set the number of retrieved captions in our experiments to 5.

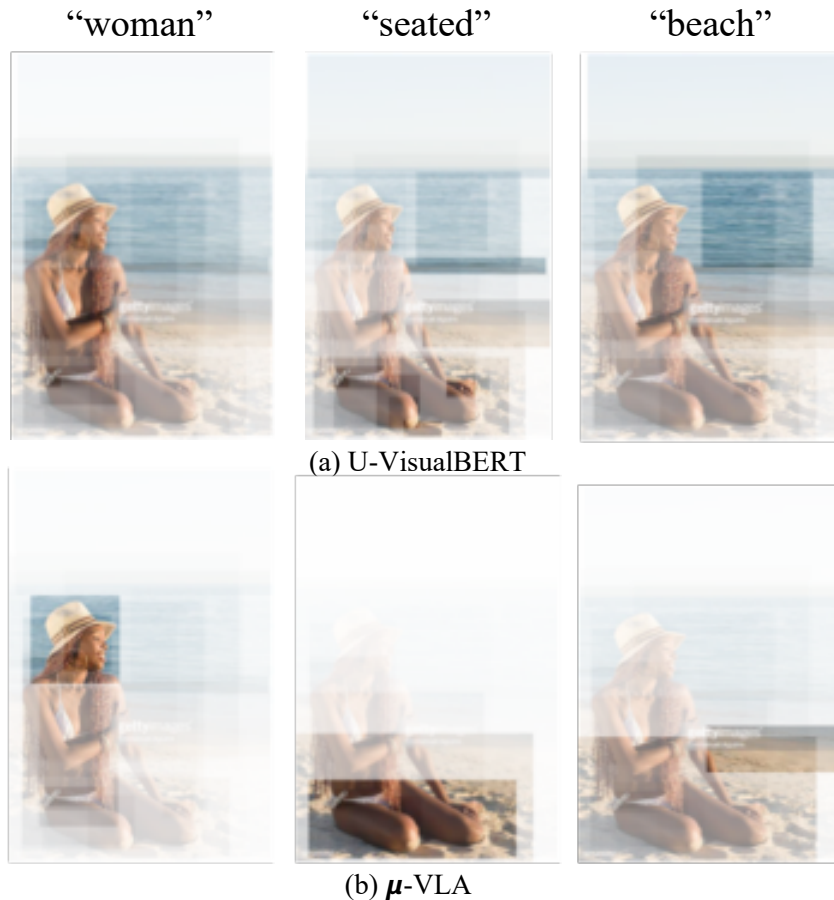


Figure 3.11: Text-to-image attention given the aligned pair whose caption is “young woman seated on the beach”.

Visualization To get a sense of the quality of the retrieved sentences, we show some examples of retrieved text from both CC and BC in Fig. 3.10. The first row demonstrates a positive case of retrieved captions from CC, where we observe a good coverage of the objects in the image such as “young woman”, “sofa”, and “couch” in the top retrieved sentences. Similarly, our retrieval method can also retrieve good candidates from BC that describe many visual objects from the image as depicted in row 2. This observation demonstrate the effectiveness of picking candidates based on their closeness to the object list in the language embedding space.

We also compare the text-to-image attention between the pre-trained U-VisualBERT and μ -VLA without task-specific fine-tuning as [112, 155]. As shown in Fig. 3.11, we feed into the models an aligned pair whose caption is “young woman seated on the beach”, we visualize the local cross-modality alignment between regions and tokens. we found our full model μ -VLA can better attend on the described regions, showing higher-quality alignment is learned through the proposed pre-training.

Chapter 4

Conclusion and Future Work

4.1 Summary

This dissertation has demonstrated our exploration on building multi-lingual grounded dialogue systems that can integrate vision, language and actions to collaborate with human for real-world applications. Our work has mainly focused on improving the current dialogue systems from two aspects. First, we aim to situate dialogue systems in goal-oriented tasks that involves vision and language, where they learn how to carry out meaningful conversations and plan for actions to interact with vision context. It is a vital research area that would lead to agents to collaborate with humans in complex real-world scenarios. Second, we care about building dialogue agents to serve people regardless of the languages they speak. Our work focus on augmenting the knowledge transition across languages via leveraging vision as the common ground. Another exploration is on how to learn the alignment across modalities without parallel image-text corpus, which helps the learning between vision and low-resource languages. We expect our research work in these two directions would inspire further advanced work in multi-lingual multi-modal dialogue systems. Thus, I share the findings on the strengths, limitations and implications of our work below.

4.1.1 Ground Vision and Action

Learning to interact with vision via dialogue consists two important steps: carrying out a meaningful conversation and then acquiring knowledge from the conversation to take actions to interact with vision context. A task-oriented visual dialogue system often requires a joint learning of both steps. One contribution of our past research is that we propose a novel training curriculum where we alternatively train a reinforcement learning Policy for taking actions and apply supervised learning to improve dialogue generation. It disentangled the learning on the interaction with vision and that on the language generation, which leads to both high task success and high-quality dialogue utterance. While we specifically study this problem in an image guessing game, we believe our proposed training curriculum is generic and can be applied to other RL-based sequence to sequence visual dialogue system in task-oriented setting. However, the impact of RL on dialogue generation is not fully explored as we only apply RL on the interaction with vision. In reality the dialogue generated by the agent also matters on the success of the task, as it will determine how much knowledge we can gain from the environment. However, the traditional way that directly apply RL to language generation utilizes goal-related reward to guide the language generation on the word-level, which leads to degradation in language quality as discussed in our work. We believe it is crucially important to propose a new reward function that can both evaluate the task accomplishment as well as the language quality. Meanwhile, a better alternative action space to apply RL is also essential.

Another contribution from our research work is that we introduces a novel task for vision and language grounding research where we intend to build agent that can manipulate images via having conversation with human users. We highlight the importance of this new task as it could lead to many useful real-world application, such as understanding the surroundings of a robot in a human inaccessible region through natural language conversations or designing fashion products based on user’s language instructions. In addition to image synthesis from iterative conversation, there is other potential of this task. Our task also facilitates a good problem for multi-task learning on vision-to-language generation, where one agent is required to simultaneously handle image caption generation

and visual question answering. The interplay between the two agents with a common goal also encourages the study of applying multi-agent reinforcement learning to optimize the performance of each side.

4.1.2 Cross-Lingual Cross-modal Representation Learning

To build multi-lingual dialogue systems, we aim to jointly learn universal representation across languages and modalities by grounding various languages to vision. Our first contribution is augmenting the classic neural machine translation translation via connecting the visual-semantic words in the source language to its visual context. By grounding the visual-semantic words to vision, the correct sense of ambiguous words is identified, while the alignment between concepts that often appear in the same visual context is identified. (e.g *baseball* and *bat*). Taking it one step further, we leverage the powerful pre-training techniques to learn task-acoustic cross-modal cross-lingual joint representation. By performing weakly supervised learning on large-scale multilingual vision-text pairs, our model demonstrates superior advantage over task-specific model on different downstream tasks. Our model also demonstrates superior transition capability to transfer knowledge from high-resource language to low-resource language, via the mutual grounded visual context. While our model is still limited on the scale of languages that it can handle, we are confident that performing weakly supervised or self supervised on easy-to-get large scale corpus is the promising direction for multi-lingual multi-modal applications. However, our current method does not fully explore the dynamics of visual concept. The visual representation is often learned from pre-trained convolutional neural network, which limits the understanding of the visual information. The lack of understanding of visual context will inevitably introduce discrepancy to connect vision and language. Thus, reinforcing the visual representation by employing different pre-trained models or online learning is our critical future research direction. We also keep in mind that our models are still limited in the language scales. To deal with all the thousand of languages in our world, we will also focus on performing unsupervised multilingual language to vision grounding. Last but not least, while self-supervised learning lead to great success of cross-modal representation learning, the vision and language pre-training requires a large amount of parallel

image-text corpus which limits the scope of training dataset. We propose two core designs to learn robust vision and language representation from unpaired image and text corpus: (1) construct a retrieval-based weakly-aligned image-text corpus. (2) introduce multi-granular pre-training objectives to enable the model to capture the cross-modal alignment at different granularity levels. Our experiments show that our model can achieve similar performance as the fully-aligned pre-trained models.

4.2 Future Directions

My long-term research goal is to build multi-lingual grounded dialogue system that perceive and understand the world via input signals from all kinds of modalities(*i.e.* video, text, audio) and communicate with people with different backgrounds. With this goal in my mind, i have identified the following three direction that I want to pursue next.

4.2.1 Learning with Less Supervision

The amount of available annotated data limits the success of the majority of the current AI techniques. As annotated data is hard to scale up to address various real-world challenges, building an AI system that can learn efficiently with minimum supervision is crucial. I am excited to explore this direction with two paradigms: (1) learning useful representation from the signals contained in the raw data via self-supervised learning. (2) Learning efficiently from a much small set of annotated data via meta-learning and prompt tuning.

4.2.2 Unifying Modalities for Multi-modal Agent

My previous work proposes a conversational agent that can interact with human users in a multi-modal environment. However, a robust conversational agent should adapt to both single-modal (only text) and multi-modal (vision + language) conversational tasks. Visual and textual knowledge can often enhance and complement each other, so I am excited to improve the agent’s performance on language-only conversations with its unified knowledge from multi-modal data.

4.2.3 Lifelong Interactive Learning

Humans explore and learn about the world via interacting through different channels such as languages and vision. A conversational agent should also learn via interacting with real users instead of just static corpus so that it can keep update its knowledge base to adapt various users. Our chatbot that wins the 2018 Amazon Alexa Prize [159] is deployed on all Alexa devices and millions of users interact with it daily in real-world settings to social chat about any open topics. We construct the profile of the user in real-time based on the conversation, which we utilize to select their potentially interested topics. We also track their intent and feelings through the conversation via a set of built natural language understanding tools, including sentiment analysis and dialogue act detection, which assures users' engagement during the whole chat. Another good part about interacting with real users is that it can help the agent to improve over time. Our agent collects much information from users, which helps to answer unknown factual questions and extend the knowledge graph to cover more diverse topics. However, this dialogue system still requires heavy manual effort to keep track of updated information from various users. I am excited to continue explore this new paradigm with automatic information extraction approach to build robust grounded conversational AI via interacting with real users.

REFERENCES

- [1] P. Chattopadhyay, D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, and D. Parikh, “Evaluating visual conversational agents via cooperative human-ai games,” *CoRR*, vol. abs/1708.05122, 2017. [Online]. Available: <http://arxiv.org/abs/1708.05122>
- [2] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, “Visual Dialog,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [3] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville, “Guesswhat?! visual object discovery through multi-modal dialogue,” *CoRR*, vol. abs/1611.08481, 2016. [Online]. Available: <http://arxiv.org/abs/1611.08481>
- [4] J. Zhang, Q. Wu, C. Shen, J. Zhang, J. Lu, and A. van den Hengel, “Asking the difficult questions: Goal-oriented visual question generation via intermediate rewards,” *CoRR*, vol. abs/1711.07614, 2017. [Online]. Available: <http://arxiv.org/abs/1711.07614>
- [5] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra, “Learning cooperative visual dialog agents with deep reinforcement learning,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [6] J. Zhang, T. Zhao, and Z. Yu, “Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog,” *CoRR*, vol. abs/1805.03257, 2018. [Online]. Available: <http://arxiv.org/abs/1805.03257>
- [7] J. D. Williams and S. Young, “Partially observable markov decision processes for spoken dialog systems,” *Comput. Speech Lang.*, vol. 21, no. 2, pp. 393–422, Apr. 2007. [Online]. Available: <http://dx.doi.org/10.1016/j.csl.2006.06.008>
- [8] A. Raux, B. Langner, D. Bohus, A. W. Black, and M. Eskenazi, “Let’s go public! taking a spoken dialog system to the real world,” in *in Proc. of Interspeech 2005*, 2005.
- [9] W. Shi and Z. Yu, “Sentiment adaptive end-to-end dialog systems,” *CoRR*, vol. abs/1804.10731, 2018. [Online]. Available: <http://arxiv.org/abs/1804.10731>
- [10] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, “Deep reinforcement learning for dialogue generation,” *CoRR*, vol. abs/1606.01541, 2016. [Online]. Available: <http://arxiv.org/abs/1606.01541>
- [11] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra, “Deal or no deal? end-to-end learning for negotiation dialogues,” *CoRR*, vol. abs/1706.05125, 2017. [Online]. Available: <http://arxiv.org/abs/1706.05125>

- [12] T. Zhao, K. Xie, and M. Eskénazi, “Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models,” *CoRR*, vol. abs/1902.08858, 2019. [Online]. Available: <http://arxiv.org/abs/1902.08858>
- [13] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, “Hierarchical neural network generative models for movie dialogues,” *CoRR*, vol. abs/1507.04808, 2015. [Online]. Available: <http://arxiv.org/abs/1507.04808>
- [14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.
- [15] X. Guo, H. Wu, Y. Cheng, S. Rennie, and R. S. Feris, “Dialog-based interactive image retrieval,” *CoRR*, vol. abs/1805.00145, 2018. [Online]. Available: <http://arxiv.org/abs/1805.00145>
- [16] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.
- [17] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, “Visual dialog,” *CoRR*, vol. abs/1611.08669, 2016. [Online]. Available: <http://arxiv.org/abs/1611.08669>
- [18] A. El-Nouby, S. Sharma, H. Schulz, R. D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, and G. W. Taylor, “Keep drawing it: Iterative language-based image generation and editing,” *CoRR*, vol. abs/1811.09845, 2018. [Online]. Available: <http://arxiv.org/abs/1811.09845>
- [19] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao, “Storygan: A sequential conditional gan for story visualization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [20] A. Karpathy and F. Li, “Deep visual-semantic alignments for generating image descriptions,” *CoRR*, vol. abs/1412.2306, 2014. [Online]. Available: <http://arxiv.org/abs/1412.2306>
- [21] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” *CoRR*, vol. abs/1411.4555, 2014. [Online]. Available: <http://arxiv.org/abs/1411.4555>
- [22] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, “Translating videos to natural language using deep recurrent neural networks,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 1494–1504. [Online]. Available: <https://aclanthology.org/N15-1173>

- [23] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, “End-to-end dense video captioning with masked transformer,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [24] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell, “Visual storytelling,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), 2016*, June 2016, website: www.sind.ai. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/visual-storytelling/>
- [25] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, “Vqa: Visual question answering,” *Int. J. Comput. Vision*, vol. 123, no. 1, pp. 4–31, May 2017. [Online]. Available: <https://doi.org/10.1007/s11263-016-0966-6>
- [26] J.-H. Kim, N. Kitaev, X. Chen, M. Rohrbach, Y. Tian, D. Batra, and D. Parikh, “CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication,” *arXiv preprint arXiv:1712.05558*, 2019. [Online]. Available: <http://arxiv.org/abs/1712.05558>
- [27] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” *CoRR*, vol. abs/1605.05396, 2016. [Online]. Available: <http://arxiv.org/abs/1605.05396>
- [28] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. N. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” *CoRR*, vol. abs/1612.03242, 2016. [Online]. Available: <http://arxiv.org/abs/1612.03242>
- [29] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “Attngan: Fine-grained text to image generation with attentional generative adversarial networks,” *CoRR*, vol. abs/1711.10485, 2017. [Online]. Available: <http://arxiv.org/abs/1711.10485>
- [30] S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou, and Y. Bengio, “Chatpainter: Improving text to image generation using dialogue,” *CoRR*, vol. abs/1802.08216, 2018. [Online]. Available: <http://arxiv.org/abs/1802.08216>
- [31] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, “Manigan: Text-guided image manipulation,” 2020.
- [32] M. Günel, E. Erdem, and A. Erdem, “Language guided fashion image manipulation with feature-wise transformations,” *CoRR*, vol. abs/1808.04000, 2018. [Online]. Available: <http://arxiv.org/abs/1808.04000>
- [33] S. Nam, Y. Kim, and S. J. Kim, “Text-adaptive generative adversarial networks: Manipulating images with natural language,” 2018.

- [34] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” 2019.
- [35] T. Park, M. Liu, T. Wang, and J. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” *CoRR*, vol. abs/1903.07291, 2019. [Online]. Available: <http://arxiv.org/abs/1903.07291>
- [36] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: <http://arxiv.org/abs/1405.0312>
- [37] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: <http://arxiv.org/abs/1606.00915>
- [38] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston, “Parlai: A dialog research software platform,” *CoRR*, vol. abs/1705.06476, 2017. [Online]. Available: <http://arxiv.org/abs/1705.06476>
- [39] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: <http://arxiv.org/abs/1411.4038>
- [40] X. Guo, H. Wu, Y. Cheng, S. Rennie, and R. S. Feris, “Dialog-based interactive image retrieval,” *CoRR*, vol. abs/1805.00145, 2018. [Online]. Available: <http://arxiv.org/abs/1805.00145>
- [41] L. Shang, Z. Lu, and H. Li, “Neural responding machine for short-text conversation,” *CoRR*, vol. abs/1503.02364, 2015. [Online]. Available: <http://arxiv.org/abs/1503.02364>
- [42] D. Elliott, S. Frank, K. Sima’an, and L. Specia, “Multi30k: Multilingual english-german image descriptions,” *CoRR*, vol. abs/1605.00459, 2016. [Online]. Available: <http://arxiv.org/abs/1605.00459>
- [43] P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer, “Attention-based multimodal neural machine translation,” in *WMT*, 2016.
- [44] I. Calixto, Q. Liu, and N. Campbell, “Doubly-attentive decoder for multi-modal neural machine translation,” *CoRR*, vol. abs/1702.01287, 2017. [Online]. Available: <http://arxiv.org/abs/1702.01287>
- [45] D. Elliott and Á. Kádár, “Imagination improves multimodal translation,” *CoRR*, vol. abs/1705.04350, 2017. [Online]. Available: <http://arxiv.org/abs/1705.04350>

- [46] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, “Findings of the second shared task on multimodal machine translation and multilingual image description,” *CoRR*, vol. abs/1710.07177, 2017. [Online]. Available: <http://arxiv.org/abs/1710.07177>
- [47] J. Helcl and J. Libovický, “CUNI system for the WMT17 multimodal translation task,” *CoRR*, vol. abs/1707.04550, 2017. [Online]. Available: <http://arxiv.org/abs/1707.04550>
- [48] I. Calixto, Q. Liu, and N. Campbell, “Incorporating global visual features into attention-based neural machine translation,” *CoRR*, vol. abs/1701.06521, 2017. [Online]. Available: <http://arxiv.org/abs/1701.06521>
- [49] M. Ma, D. Li, K. Zhao, and L. Huang, “OSU multimodal machine translation system report,” *CoRR*, vol. abs/1710.02718, 2017. [Online]. Available: <http://arxiv.org/abs/1710.02718>
- [50] J. Zhang, M. Utiyama, E. Sumita, G. Neubig, and S. Nakamura, “Nict-naist system for wmt17 multimodal translation task,” in *WMT*, 2017.
- [51] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS’15. Cambridge, MA, USA: MIT Press, 2015, pp. 91–99. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2969239.2969250>
- [52] O. Caglayan, W. Aransa, A. Bardet, M. García-Martínez, F. Bougares, L. Barrault, M. Masana, L. Herranz, and J. van de Weijer, “LIUM-CVC submissions for WMT17 multimodal translation task,” *CoRR*, vol. abs/1707.04481, 2017. [Online]. Available: <http://arxiv.org/abs/1707.04481>
- [53] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *CoRR*, vol. abs/1411.2539, 2014. [Online]. Available: <http://arxiv.org/abs/1411.2539>
- [54] A. Karpathy and F. Li, “Deep visual-semantic alignments for generating image descriptions,” *CoRR*, vol. abs/1412.2306, 2014. [Online]. Available: <http://arxiv.org/abs/1412.2306>
- [55] I. Calixto, Q. Liu, and N. Campbell, “Multilingual multi-modal embeddings for natural language processing,” *CoRR*, vol. abs/1702.01101, 2017. [Online]. Available: <http://arxiv.org/abs/1702.01101>
- [56] S. Gella, R. Sennrich, F. Keller, and M. Lapata, “Image pivoting for learning multilingual multimodal representations,” *CoRR*, vol. abs/1707.07601, 2017. [Online]. Available: <http://arxiv.org/abs/1707.07601>

- [57] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [58] M. Schuster and K. Paliwal, “Bidirectional recurrent neural networks,” *Trans. Sig. Proc.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997. [Online]. Available: <http://dx.doi.org/10.1109/78.650093>
- [59] K. Cho, B. van Merriënboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *CoRR*, vol. abs/1406.1078, 2014. [Online]. Available: <http://arxiv.org/abs/1406.1078>
- [60] G. Chrupala, Á. Kádár, and A. Alishahi, “Learning language through pictures,” *CoRR*, vol. abs/1506.03694, 2015. [Online]. Available: <http://arxiv.org/abs/1506.03694>
- [61] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [62] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [63] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. M. Barone, J. Mokry, and M. Nadejde, “Nematus: a toolkit for neural machine translation,” *CoRR*, vol. abs/1703.04357, 2017. [Online]. Available: <http://arxiv.org/abs/1703.04357>
- [64] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: <http://arxiv.org/abs/1405.0312>
- [65] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” *CoRR*, vol. abs/1508.07909, 2015. [Online]. Available: <http://arxiv.org/abs/1508.07909>
- [66] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: <https://doi.org/10.3115/1073083.1073135>

- [67] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for any target language,” in *In Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.
- [68] D. Elliott, S. Frank, K. Sima’an, and L. Specia, “Multi30K: Multilingual English-German image descriptions,” in *Proceedings of the 5th Workshop on Vision and Language*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 70–74. [Online]. Available: <https://www.aclweb.org/anthology/W16-3210>
- [69] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, “Findings of the second shared task on multimodal machine translation and multilingual image description,” in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 215–233. [Online]. Available: <https://www.aclweb.org/anthology/W17-4718>
- [70] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, and S. Frank, “Findings of the third shared task on multimodal machine translation,” in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 304–323. [Online]. Available: <https://www.aclweb.org/anthology/W18-6402>
- [71] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, “STAIR captions: Constructing a large-scale Japanese image caption dataset,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 417–421. [Online]. Available: <https://www.aclweb.org/anthology/P17-2066>
- [72] X. Li, X. Wang, C. Xu, W. Lan, Q. Wei, G. Yang, and J. Xu, “COCO-CN for cross-lingual image tagging, captioning and retrieval,” *CoRR*, vol. abs/1805.08661, 2018. [Online]. Available: <http://arxiv.org/abs/1805.08661>
- [73] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423>
- [74] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” 2019.
- [75] H. Tan and M. Bansal, “Lxmert: Learning cross-modality encoder representations from transformers,” 2019.
- [76] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, “Vi-bert: Pre-training of generic visual-linguistic representations,” 2020.

- [77] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, “Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training,” 2019.
- [78] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “{UNITER}: Learning {un}iversal image-{te}xt representations,” 2020. [Online]. Available: <https://openreview.net/forum?id=S1eL4kBYwr>
- [79] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and vqa,” in *AAAI*, 2020.
- [80] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, “Oscar: Object-semantics aligned pre-training for vision-language tasks,” 2020.
- [81] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, “Pixel-bert: Aligning image pixels with text by deep multi-modal transformers,” 2020.
- [82] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, “12-in-1: Multi-task vision and language representation learning,” 2020.
- [83] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran, “Bridge correlational neural networks for multilingual multimodal representation learning,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 171–181. [Online]. Available: <https://www.aclweb.org/anthology/N16-1021>
- [84] S. Gella, R. Sennrich, F. Keller, and M. Lapata, “Image pivoting for learning multilingual multimodal representations,” in *Empirical Methods in Natural Language Processing*, 2017.
- [85] J. Wehrmann, D. M. Souza, M. A. Lopes, and R. C. Barros, “Language-agnostic visual-semantic embeddings,” in *International Conference on Computer Vision*, 2019.
- [86] D. Kim, K. Saito, K. Saenko, S. Sclaroff, and B. A. Plummer, “MULE: Multimodal Universal Language Embedding,” in *AAAI Conference on Artificial Intelligence*, 2020.
- [87] A. Burns, D. Kim, D. Wijaya, K. Saenko, and B. A. Plummer, “Learning to scale multilingual representations for vision-language tasks,” 2020.
- [88] I. Calixto, Q. Liu, and N. Campbell, “Doubly-attentive decoder for multi-modal neural machine translation,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1913–1924. [Online]. Available: <https://www.aclweb.org/anthology/P17-1175>

- [89] J. Helcl and J. Libovický, “CUNI system for the WMT17 multimodal translation task,” in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 450–457. [Online]. Available: <https://www.aclweb.org/anthology/W17-4749>
- [90] I. Calixto and Q. Liu, “Incorporating global visual features into attention-based neural machine translation.” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 992–1003. [Online]. Available: <https://www.aclweb.org/anthology/D17-1105>
- [91] D. Elliott and Á. Kádár, “Imagination improves multimodal translation,” in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 130–141. [Online]. Available: <https://www.aclweb.org/anthology/I17-1014>
- [92] M. Zhou, R. Cheng, Y. J. Lee, and Z. Yu, “A visual attention grounding neural model for multimodal machine translation,” in *Empirical Methods in Natural Language Processing*, 2018.
- [93] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- [94] H. S. Arslan, M. Fishel, and G. Anbarjafari, “Doubly attentive transformer machine translation,” *CoRR*, vol. abs/1807.11605, 2018. [Online]. Available: <http://arxiv.org/abs/1807.11605>
- [95] S. Yao and X. Wan, “Multimodal transformer for multimodal machine translation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4346–4350. [Online]. Available: <https://www.aclweb.org/anthology/2020.acl-main.400>
- [96] P.-Y. Huang, J. Hu, X. Chang, and A. Hauptmann, “Unsupervised multimodal neural machine translation with pseudo visual pivoting,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8226–8237. [Online]. Available: <https://www.aclweb.org/anthology/2020.acl-main.731>
- [97] G. A. Sigurdsson, J.-B. Alayrac, A. Nematzadeh, L. Smaira, M. Malinowski, J. Carreira, P. Blunsom, and A. Zisserman, “Visual grounding in video for unsupervised

- word translation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [98] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, “Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization,” in *International Conference on Machine Learning*, 2020.
- [99] P. Sharma, N. Ding, S. Goodman, and R. Soiccut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565. [Online]. Available: <https://www.aclweb.org/anthology/P18-1238>
- [100] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” *arXiv preprint arXiv:1911.02116*, 2019.
- [101] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: <http://arxiv.org/abs/1506.01497>
- [102] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: <https://www.aclweb.org/anthology/D18-2012>
- [103] C. Dyer, V. Chahuneau, and N. A. Smith, “A simple, fast, and effective reparameterization of IBM model 2,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 644–648. [Online]. Available: <https://www.aclweb.org/anthology/N13-1073>
- [104] L. Wang, Y. Li, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” *CoRR*, vol. abs/1704.03470, 2017. [Online]. Available: <http://arxiv.org/abs/1704.03470>
- [105] S. Gella, R. Sennrich, F. Keller, and M. Lapata, “Image pivoting for learning multilingual multimodal representations,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2839–2845. [Online]. Available: <https://www.aclweb.org/anthology/D17-1303>

- [106] H. Huang, L. Su, D. Qi, N. Duan, E. Cui, T. Bharti, L. Zhang, L. Wang, J. Gao, B. Liu, J. Fu, D. Zhang, X. Liu, and M. Zhou, “M3p: Learning universal representations via multitask multilingual multimodal pre-training,” 2020.
- [107] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” *CoRR*, vol. abs/1906.10770, 2019. [Online]. Available: <http://arxiv.org/abs/1906.10770>
- [108] N. Shimizu, N. Rong, and T. Miyazaki, “Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps,” in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1918–1928. [Online]. Available: <https://www.aclweb.org/anthology/C18-1163>
- [109] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [110] G. Lample and A. Conneau, “Cross-lingual language model pretraining,” in *Advances in Neural Information Processing Systems*, 2019.
- [111] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, “12-in-1: Multi-task vision and language representation learning,” in *CVPR*, 2020.
- [112] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “Uniter: Universal image-text representation learning,” in *ECCV*, 2020.
- [113] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” *NeurIPS*, 2019.
- [114] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, “Ernie-vil: Knowledge enhanced vision-language representations through scene graph,” *arXiv preprint arXiv:2006.16934*, 2020.
- [115] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, “Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training,” in *AAAI*, 2020.
- [116] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, “Vl-bert: Pre-training of generic visual-linguistic representations,” *arXiv preprint arXiv:1908.08530*, 2019.
- [117] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *ICML*, 2021.
- [118] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, “Pixel-bert: Aligning image pixels with text by deep multi-modal transformers,” *arXiv preprint arXiv:2004.00849*, 2020.

- [119] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, “Simvlm: Simple visual language model pretraining with weak supervision,” *arXiv preprint arXiv:2108.10904*, 2021.
- [120] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, “Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning,” *arXiv preprint arXiv:2012.15409*, 2020.
- [121] J. Li, R. R. Selvaraju, A. D. Gotmare, S. R. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” *CoRR*, vol. abs/2107.07651, 2021. [Online]. Available: <https://arxiv.org/abs/2107.07651>
- [122] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, “Seeing out of the box: End-to-end pre-training for vision-language representation learning,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 12 976–12 985.
- [123] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Proceedings of ECCV*. Springer, 2014, pp. 740–755.
- [124] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft coco captions: Data collection and evaluation server,” *arXiv preprint arXiv:1504.00325*, 2015.
- [125] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
- [126] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *ACL*, 2018.
- [127] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts,” in *CVPR*, 2021.
- [128] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in *NIPS*, 2011.
- [129] H. Bao, L. Dong, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv preprint arXiv:2106.08254*, 2021.
- [130] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

- [131] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” *preprint arXiv:2002.05709*, 2020.
- [132] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2018.
- [133] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [134] L. H. Li, H. You, Z. Wang, A. Zareian, S.-F. Chang, and K.-W. Chang, “Unsupervised vision-and-language pre-training without parallel images and captions,” in *NACCL*, 2021.
- [135] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *ICCV*, 2015.
- [136] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, “A corpus for reasoning about natural language grounded in photographs,” *arXiv preprint arXiv:1811.00491*, 2018.
- [137] N. Xie, F. Lai, D. Doran, and A. Kadav, “Visual entailment: A novel task for fine-grained image understanding,” *arXiv preprint arXiv:1901.06706*, 2019.
- [138] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in *ECCV*, 2016.
- [139] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *EMNLP*, 2019.
- [140] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [141] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.
- [142] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “Mdetr-modulated detection for end-to-end multi-modal understanding,” in *ICCV*, 2021.
- [143] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visualbert: A simple and performant baseline for vision and language,” *arXiv preprint arXiv:1908.03557*, 2019.

- [144] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *CVPR*, 2018.
- [145] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in *ECCV*, 2020.
- [146] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu, “Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training,” in *AAAI*, 2021.
- [147] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and vqa,” in *AAAI*, 2020.
- [148] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” *arXiv preprint arXiv:1911.02116*, 2019.
- [149] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, “Phrase-based & neural unsupervised machine translation,” in *EMNLP*, 2018.
- [150] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, “Unsupervised machine translation using monolingual corpora only,” in *ICLR*, 2018.
- [151] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” in *ICLR*, 2018.
- [152] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, “Vinvl: Revisiting visual representations in vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5579–5588.
- [153] H. Tan and M. Bansal, “Vokenization: Improving language understanding with contextualized, visual-grounded supervision,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.
- [154] M. Honnibal and I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,” 2017, to appear.
- [155] M. Zhou, L. Zhou, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, “Uc2: Universal cross-lingual cross-modal vision-and-language pre-training,” in *CVPR*, 2021.
- [156] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, “Vinvl: Revisiting visual representations in vision-language models,” in *CVPR*, 2021.

- [157] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [158] A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh, “Mmf: A multimodal framework for vision and language research,” <https://github.com/facebookresearch/mmf>, 2020.
- [159] D. Yu, M. Cohn, Y. M. Yang, C. Y. Chen, W. Wen, J. Zhang, M. Zhou, K. Jesse, A. Chau, A. Bhowmick, S. Iyer, G. Sreenivasulu, S. Davidson, A. Bhandare, and Z. Yu, “Gunrock: A social bot for complex and engaging long conversations,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 79–84. [Online]. Available: <https://aclanthology.org/D19-3014>