UNIVERSITY OF CALIFORNIA

Los Angeles

Advancing Vision-Language and Language Models in Low-Resource Settings

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Masoud Monajatipoor

2024

ABSTRACT OF THE DISSERTATION

Advancing Vision-Language and Language Models

in Low-Resource Settings

by

Masoud Monajatipoor

Doctor of Philosophy in Electrical and Computer Engineering

University of California, Los Angeles, 2024

Professor Kai-Wei Chang, Co-Chair

Professor Lin Yang, Co-Chair

Vision-language modeling is a crucial subfield of AI that focuses on jointly learning and representing image and text data, often using one modality to enhance understanding of the other. In cognitive science, humans use their visual system to grasp deep aspects of a concept, such as shape and size, while language helps them understand its semantics. Similarly, a machine can gain a better understanding of the world by utilizing multiple modalities, providing deeper insights compared to learning from a single modality. VL modeling is widely explored in the general domain, thanks to the vast image-text data available online and extensive annotated VL datasets. There are several strong VL models in the general domain, such as CLIP, which perform well on various tasks. However, in low-resource domains with limited data or dense knowledge areas, like the medical field, the data shortage hinders the development of robust multimodal models with reliable performance,

especially where model reliability is critical. My research goal is to study the underlying capability of Vision-Language and Language Models and to develop innovative approaches to enhance their usage for low-resource domains such as the medical domain.

The dissertation of Masoud Monajatipoor is approved.

Aditya Grover

Ali Mosleh

Lin Yang, Committee Co-Chair

Kai-Wei Chang, Committee Co-Chair

University of California, Los Angeles

2024

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to begin by expressing my profound gratitude to both my advisors, Dr. Kai-Wei Chang and Dr. Lin Yang, for their invaluable guidance and unwavering support throughout my Ph.D. journey. Coming into this program without any background in NLP, I was fortunate to have Dr. Chang and Dr. Yang, who patiently answered countless questions and nurtured my growth from a novice to an independent researcher. Their encouragement and dedication have been instrumental in shaping my professional and personal development, and I am deeply thankful for their mentorship.

Special thanks to the esteemed members of my committee, Professors Ali Mosleh and Aditya Grover. Their insightful critiques and suggestions challenged me to think more broadly and deeply, significantly enhancing the quality of my research.

I am also immensely grateful for the opportunity to collaborate with brilliant minds such as Mozhdeh Rouhsedaghat, Liunian Li, Da Yin, Hritik bansal, Sunipa Dev, Arjun Subramonian, Jeff M Phillips, Anaelia Ovalle, Joel Stremmel, Jiaxin Yang, Ehsan Mohaghegh, Melika Emami, and many others. Your expertise and insights have left a lasting impact on my work.

Finally, my heartfelt appreciation goes to my family for their love, support, and sacrifice, which have been my anchor. Above all, my deepest gratitude is reserved for my wife, whose unwavering love, immense patience, and constant encouragement have been the true foundation of my success. Her support throughout this journey has been nothing short of inspirational.

VITA

2012–2017 B.E. (Electrical Engineering), Sharif University of Technology

2017–2019 M.S. (Electrical and Computer Engineering), University of California, Los Angeles

PUBLICATIONS

Masoud Monajatipoor, Mozhdeh Rouhsedaghat, Liunian Harold Li, Aichi Chien, C.C. Jay Kou, Kai-Wei Chang. BERTHop: An Effective Vision-and-Language Model for Chest X-Ray Disease Diagnosis. ICCV Workshop 2021

Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin F. Yang, Kai-Wei Chang. MetaVL: Transferring In-Context Learning Ability From Language Models to Vision-Language Models. ACL 2023

Masoud Monajatipoor, Zi-Yi Dou, Nanyun Peng, Kai-Wei Chang. Medical Vision-Language Pre-Training for Brain Abnormalities. LREC-COLING 2024

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Ehsan Mohaghegh, Melika Emami, Mozhdeh Rouhsedaghat, Kai-Wei Chang. LLMs in Biomedical: A Study on Clinical Named Entity Recognition. Under review

CHAPTER 1

Introduction

In the field of Artificial Intelligence (AI), it's become clear that understanding multimodal data - such as text, image, and video - is a crucial yet challenging task. While researchers have made significant strides in understanding text data in natural language understanding and image data in computer vision, combining the two has become a focus of attention in recent years. The benefits of developing an AI model that can understand both images and text are numerous, as one modality can help improve the understanding of the other. This has spurred interest in building multimodal models that can process information from multiple sources simultaneously.

The field of visual-language (VL) involves numerous tasks, including but not limited to classification and generation. In classification tasks, a model learns to predict the probability of multimodal data belonging to various classes. In generation tasks, the model learns to take either vision or language data as input and generate an output in the other modality format. Visual question answering (VQA) is an example of a multimodal task, in which a model is presented with an image and a question related to the image, and is expected to generate an appropriate answer. Such tasks have received considerable attention in the research community due to their potential for enabling intelligent systems to process and understand multimodal information.

The advent of deep learning techniques, coupled with the availability of large annotated datasets and computational power, has enabled AI models to achieve high accuracy and performance. However, in domains where a significant amount of annotated data is not available, such models tend

to perform poorly. This makes the development of data-efficient AI models a major challenge that requires further attention.

During my PhD, I have focused on learning-based low-resource VL models and tasks, with a primary focus on scenarios where the limited availability of annotated data is the main bottleneck. Specifically, I explored building VL and language models for tasks such as few-shot visual question answering (VQA), disease diagnosis, and information extraction. By addressing these challenges, we aim to contribute to the development of more efficient and effective AI models that can operate in low-resource environments.

In Chapter 2, we review Computer-Aided Diagnosis (CADx) systems [GS08]. CADx offers significant benefits for disease diagnosis, including improving the quality and consistency of predictions and reducing medical errors, as they are not prone to human error. While most research focuses on the medical image-based diagnosis, such as chest X-ray (CXR) images [AU19, AA19, AM18], radiology reports contain crucial information, such as patient history and previous studies, which are difficult to detect from the image alone. Hence, VL models that take both images and text as input have the potential to improve the accuracy of CADx, and several attempts have been made in this direction [WPL18, ZCS19, LWL20]. However, training V&L models with limited annotated data remains a significant challenge.

In Chapter 3 we go through the Few-shot learning in Visual question answering (VQA). VQA is a vision-language task that requires a model to answer a question based on an image. Various methods have been proposed for this task [LYL20, LYY19, LBP19], but they all rely on a large amount of training data. These models are typically large-scale and require extensive training to perform well. When there is a scarcity of data, the models can become ineffective, and their performance can suffer greatly. While some models have been developed for the few-shot scenario, they may require a large capacity or a massive corpus in a specific data format.

In Chapter 4, we revisited foundational Vision-language models. In the context of multimodal

clinical AI, there is a growing need for models that possess domain-specific knowledge, as existing models often lack the expertise required for medical applications. We take brain abnormalities as an example to demonstrate how to automatically collect medical image-text aligned data for pre-training from public resources such as PubMed. In particular, we present a pipeline that streamlines the pre-training process by initially collecting a large brain image-text dataset from case reports and published journals and subsequently constructing a high-performance vision-language model tailored to specific medical tasks.

Finally, in Chapter 5, we Explored large language models (LLMs) for medical application. LLMs demonstrate remarkable versatility in various NLP tasks but encounter distinct challenges in biomedical due to the complexities of language and data scarcity. We investigates LLMs application in the biomedical domain by exploring strategies to enhance their performance for the Named Entity Recognition (NER) task.

CHAPTER 2

Multimodal Representation for Disease Diagnosis

Systems for Computer-Aided Diagnosis (CADx) provide significant advantages in disease detection by enhancing the accuracy and consistency of diagnostic outcomes and offering an additional review to minimize errors in medical judgments [GS08]. While the majority of research has concentrated on image-based diagnostics, including chest X-ray (CXR) imaging [AA19, AM18], it is important to note that radiological reports frequently include critical textual information, such as patient history and previous examinations, which are not immediately apparent from images alone. Additionally, integrating both textual and visual data in diagnostics aligns more closely with the methods used by human medical experts. Thus, V&L models that utilize both image and text inputs could potentially enhance the precision of CADx systems.

The scarcity of labeled data in healthcare complicates the application of V&L models. Labeling medical data demands the involvement of trained professionals, making it a costly endeavor. While recent initiatives have introduced large-scale, automatically labeled datasets for specific medical tasks like chest X-rays [WPL17, JPG19], these datasets tend to be of lower quality and can negatively impact model performance. Moreover, datasets of this nature are not broadly available for many medical applications, making the training of V&L models with minimal annotated data a significant obstacle.

Recent advancements in the field have introduced pre-trained V&L models as a strategy to minimize the necessity for labeled data in training precise models for specific tasks [LYY19, TB19a,

CLY20]. These models initially undergo training on extensive datasets of image captions, employing self-supervision techniques such as the masked language model loss, where parts of the input are concealed and the model is tasked with predicting the hidden words or image sections from the available context [1]. The learned parameters from these pre-trained V&L models are subsequently transferred to initialize and fine-tune models for particular downstream applications. In many V&L tasks, it has been observed that pre-training in V&L significantly enhances performance. However, a significant issue arises when these standard pre-trained V&L models are employed in the medical field: there is a substantial gap between the general (source) and medical (target) domains. This discrepancy renders the typical pre-training and fine-tuning approach far less effective within the medical sector. Thus, there is a necessity for domain-specific adaptations. Particularly, V&L models often depend on object-focused feature extraction techniques like Faster R-CNN [RHG16], which are trained on non-medical data to recognize common objects such as cats and dogs. Unfortunately, this means that unusual patterns found in medical imagery, such as anomalies in X-ray images, may not be detected by these models optimized for everyday visuals.

To address this challenge, we introduce BERTHop, a transformer-based V&L model tailored for healthcare settings. BERTHop enhances the V&L model's visual encoder through the integration of PixelHop++ [CRY20], employing a fully unsupervised approach that drastically minimizes the dependency on annotated data [RMA21]. PixelHop++ is adept at deriving image features across various frequency bands, which is particularly advantageous for detecting anomalies at multiple levels, allowing the transformer to better correlate these findings with the textual input. Additionally, BERTHop mitigates issues related to domain disparity by incorporating a pre-trained language encoder, BlueBERT [PYL19], a variant of BERT [DCL18] specifically trained on biomedical and clinical text corpora.

[1] part of the input is masked and the objective is to predict the masked words or image regions based on the remaining contexts

2.1 Related Work

Transformer-based V&L models Drawing from the achievements of BERT in NLP, a range of transformer-based V&L models have emerged [LYY19, CLY20, TB19a]. Typically, these models employ an object detector, initially trained on the Visual Genome [KZG17], to capture visual elements from images. Subsequently, a transformer processes these visual features alongside textual input. Like BERT, these models undergo pre-training on extensive datasets of image-text pairs, using a similar mask-and-predict strategy.

Models of this nature are utilized across various V&L tasks[ZPZck, LGR20, CCL20]. Nonetheless, the effectiveness of knowledge transfer from these pre-trained models hinges on the similarity in data distributions between the source and target domains, or alternatively, the availability of substantial data for the target domain to ensure successful knowledge transfer.

V&L models in the medical domain Several V&L frameworks have been developed for analyzing chest X-rays (CXR) to diagnose medical conditions. Among these, TieNet integrates a CNN-RNN structure with multi-layer attention mechanisms, creating a unified model for V&L tasks specifically aimed at disease diagnosis. It employs a ResNet-50 base, initially trained on generic visual data, combined with an RNN to merge V&L data effectively. This model relies heavily on a substantial dataset, ChestX-ray14, for domain-specific adaptation, which restricts its utility in broader applications. In contrast, Li *et al.* [LWL20] explored the adaptability of renowned pre-trained V&L models by adjusting them using datasets like MIMIC-CXR [JPG19] and OpenI. These models, originally crafted for general use and trained with extensive data, exhibit decreased effectiveness when fine-tuned with limited specialized data. This approach is referred to as VB w/ BUTD in the subsection 2.3.2).

PixelHop++ for visual feature learning PixelHop++ was initially developed as a substitute for deep convolutional neural networks, specifically tailored for feature extraction in image and video

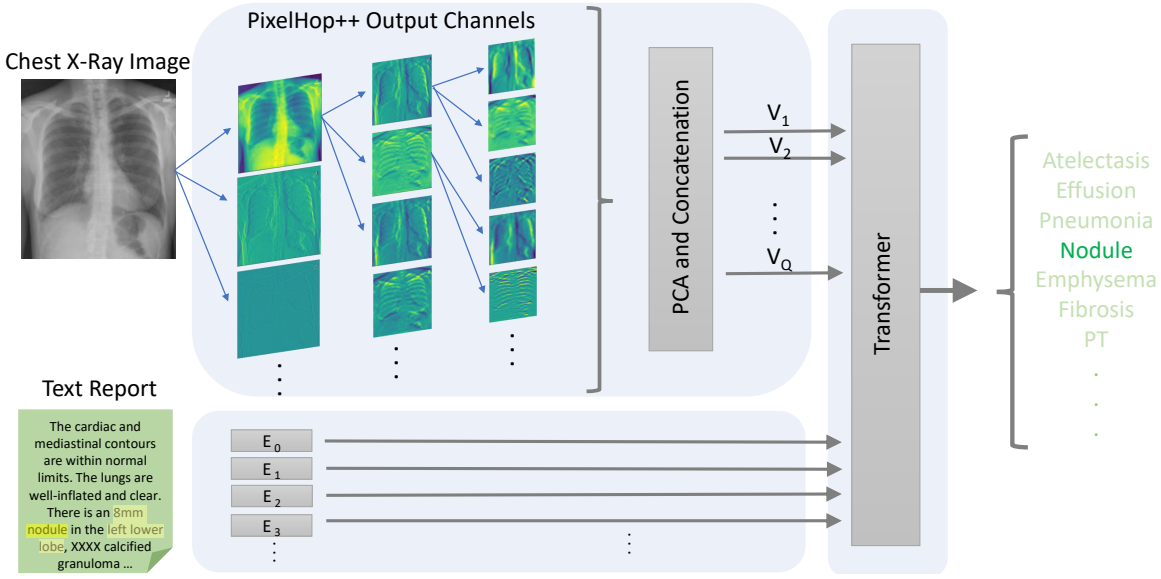


Figure 2.1: Our model architecture is introduced for diagnosing diseases from chest X-rays (CXR). Initially, a PixelHop++ model paired with a "PCA and concatenation" module is employed to create Q feature vectors. Subsequently, these vectors, integrated with language embeddings, are input into a transformer pre-initialized with BlueBERT.

processing within environments with limited resources. This model operates on multiple levels, producing output channels that encapsulate varying frequencies of an image. Demonstrated to be particularly effective with small datasets, PixelHop++ has been employed in diverse scenarios such as distinguishing gender through facial analysis [RWG20], recognizing faces [RWH20], identifying deep fake content [CRG21], and applications in healthcare [LXY21]. To our knowledge, this research marks the inaugural integration of PixelHop++ with DNN architectures. While using only PixelHop++ features as inputs to a transformer model (without textual input) tends to lag behind other vision-centric models like ChexNet [RIZ17], our model leverages the attention mechanism to merge visual features extracted from PixelHop++ with linguistic embeddings, enhancing the correlation between these two modalities.

2.2 Approach

Our model draws inspiration from VisualBERT's design, employing a unified transformer to amalgamate visual and textual embeddings. As depicted in Fig. 2.1, our methodology begins with the application of PixelHop++ to derive visual features from the X-ray images. Following this, the associated radiology report is transformed into subword embeddings. Subsequently, a combined transformer is utilized to analyze the interplay between these two modalities and identify underlying alignments.

2.2.1 Visual encoder

We contend that utilizing visual feature extraction methods from general-domain object detectors, specifically the widely-used BUTD [AHB18] model in various V&L tasks, proves inadequate for the medical field. This method underperforms in identifying medical anomalies within X-ray imagery. This shortcoming stems from the fact that such anomalies, crucial for accurate diagnoses, do not conform to typical object characteristics and are often overlooked by detectors designed for general usage. Additionally, the absence of a comprehensive annotated dataset for training disease anomaly detectors exacerbates this issue [SRG16].

We suggest utilizing PixelHop++ [CRY20] for learning visual features without supervision in the healthcare sector, given its proven effectiveness on limited datasets. PixelHop++ determines its model parameters using a closed-form solution, avoiding the need for back-propagation [RMA21]. By employing PCA to derive parameters, PixelHop++ efficiently captures image features across different frequencies autonomously. Drawing inspiration from deep neural network architectures, PixelHop++ is structured in multiple layers, with each layer composed of one or more PixelHop++ units, followed by a max-pooling layer.

Consider a scenario where we possess N training images, each with dimensions $s_1 \times s_2 \times$

d. Here, d equals 1 for grayscale images and 3 for color ones. These images are input into a PixelHop++ unit at the initial stage of the framework. The primary aim of this stage is to devise orthogonal projection vectors, or kernels, which are adept at isolating robust features from the input data. Multiple PixelHop++ units might be present at each tier of a PixelHop++ structure.

During the initial phase in a PixelHop++ unit, a window of size $w \times w \times d$ slides across the image with a stride of s. This action extracts and flattens patches from each image, represented as $x_{i1}, x_{i2}, ..., x_{iM}$, where x_{ij} denotes the jth flattened patch of the ith image, and M signifies the total number of patches derived from each image.

Subsequently, these gathered patches are employed to derive the kernels for the PixelHop++ unit. The kernels are developed as follows:

- The primary kernel, termed the DC kernel, functions as an average filter: $\frac{1}{\sqrt{n}} \times (1, 1, ..., 1)$, where n is the length of the input vector. This kernel calculates the average of each input vector.

- Subsequent to the mean calculation, PCA kernels are applied to the residual data to produce AC kernels. The foremost k kernels are those that optimally represent the residual variation.

Following this, each patch projection onto the kernels adds a constant bias to the result. In this transformation using a PixelHop++ units kernel, each series of patches, $x_{i1}, x_{i2}, ..., x_{iM}$, produces a singular output channel. For instance, at the model's first level, the PixelHop++ unit outputs one DC channel and $w \times w \times d - 1$ AC channels.

The final procedure involves pruning the model to eliminate underperforming channels. This decision is based on the "energy ratio", which compares each kernel's explained variance to the overall training data variance. This ratio serves as a pruning criterion. An energy threshold, E, is set, and channels are pruned according to the following criteria:

- Channels with an energy ratio below E are removed due to their negligible data variation along the respective kernel.
- Channels exceeding the E threshold are retained and passed to the next model level for further compression.

At every subsequent model level beyond the first, each intermediate channel output from a PixelHop++ unit is directed into another separate unit within the same level.

2.2.2 In-domain text pre-training

In BERTHop, the textual input from diagnostic reports significantly influences the transformer model's focus on relevant visual cues within its attention mechanism. Authored by a specialist radiologist, the report details both typical and atypical findings under the "finding" section, and incorporates critical patient data like medical history, affected body areas, and previous diagnostics in the "impression" section. This report's linguistic style diverges considerably from the usual pre-training datasets used for BERT, such as Wikipedia and BookCorpus, or visual-linguistic (V&L) models trained on datasets like MSCOCO and Conceptual Captions. Consequently, we suggest employing BlueBERT [PYL19] as the foundational architecture in BERTHop to more effectively assimilate the nuances of the report text. While text-only pre-training has often shown limited benefits [TB19a], in the context of the medical field, leveraging a transformer initially trained on specialized text corpora is a straightforward yet potent strategy. This method addresses the significant gap often overlooked by previous approaches, which typically initialize transformers with models trained on broad image-text datasets for V&L tasks [LWL20].

2.3 Experiments

We conduct an assessment of BERTHop using the OpenI dataset and contrast its performance with various established models. Additionally, comprehensive analyses are undertaken to ascertain the

contributions of the visual encoder and the transformer initialization to the model's effectiveness.

2.3.1 Experiment setup

Our research utilizes the OpenI dataset, which includes 3,996 individual patient reports paired with 8,121 images. This dataset features 14 prevalent thoracic conditions. Initially, we standardize the dimensions of all images to 206×206 and employ a three-tier PixelHop++ system for extracting features without supervision. Subsequently, Principal Component Analysis (PCA) is conducted on the outputs from PixelHop++, and the resultant vectors are amalgamated to construct a feature set, Q, where each feature vector, v_i, belongs to a D-dimensional space, thus forming $V = [v_1, v_2, ..., v_Q]$ where $v_i \in \mathbb{R}^D$.

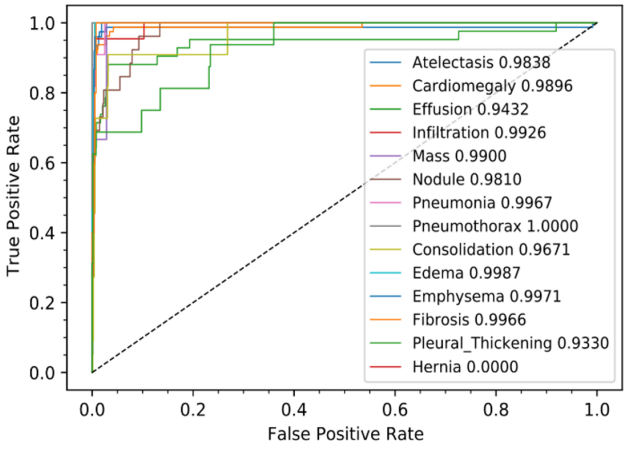
In the case of BERTHop, the dimension D is established at 2048. For our experimental configuration, we set Q to 15, though this can be adjusted based on the PixelHop++ models output channel breadth and the PCA components used. The transformer architecture is supported by BlueBERT-Base (Uncased, PubMed + MIMIC-III) from Huggingface [WDS19], a prominent transformer framework. With the amalgamated visual features and textual embeddings, the transformer is trained using a subset of 2,912 image-text pairs from the OpenI dataset. Our training parameters include a batch size of 18, a learning rate of $1e-5$, a maximum sequence length of 128, and the employment of Stochastic Gradient Descent (SGD) with a momentum of 0.9 over 240 epochs.

2.3.2 Main results

BERTHop was developed using the OpenI training dataset and assessed on its test set. The evaluation of BERTHop and other baseline models was conducted using the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metrics, utilizing the AUC function from scikit-learn [BLB13] as illustrated in Fig. 2.2 b). A comparison of BERTHop's performance with other existing approaches is detailed in Fig. 2.2 a). The findings reveal that BERTHop surpasses the

	TNNT	TieNet	VB w/ BUTD	BERTHop
Atelectasis	-	0.976	0.9247	**0.9838**
Cardiomegaly	-	0.962	0.9665	**0.9896**
Effusion	-	**0.977**	0.9049	0.9432
Infiltration	-	0.984	0.8867	**0.9926**
Mass	-	0.903	0.6428	**0.9900**
Nodule	-	0.960	0.8480	**0.9810**
Pneumonia	-	0.994	0.8537	**0.9967**
Pneumothorax	-	0.960	0.8931	**1.0000**
Consolidation	-	**0.989**	0.7870	0.9671
Edema	-	0.995	0.9500	**0.9987**
Emphysema	-	0.868	0.8565	**0.9971**
Fibrosis	-	0.960	0.6274	**0.9966**
PT	-	**0.953**	0.7612	0.9330
Hernia	-	-	-	-
AVG	0.854	0.965	0.8386	**0.9823**

a)



b)

Figure 2.2: *a)* Our model demonstrates superior diagnostic performance for thoracic diseases on OpenI when compared to three other techniques. Specifically, BERTHop exceeds the performance of models like VB w/ BUTD trained using comparable datasets. *(b)* The ROC curve for BERTHop illustrates its effectiveness across all 14 thoracic diseases.

previous state-of-the-art model, TieNet, in diagnosing 11 out of 14 thoracic diseases, recording an average AUC of 98.23%. This score exceeds the performance of VB w/ BUTD, TNNT, and TieNet by 14.37%, 12.83%, and 1.73%, respectively. It is important to note that despite TieNet being trained on the significantly larger ChestX-ray14 dataset, which includes 108,948 training examples, BERTHop has been trained on a dataset that is 9 times smaller.

2.3.3 In-domain text pre-training

We further investigate the influence of different transformer backbone initialization on model performance by pairing it with different visual encoders. The results are listed in Table 2.1. First, we find that the proposed initialization with a model pre-trained on in-domain text corpora (BlueBERT) brings significant performance boosts when paired with PixelHop++. Initializing with BlueBERT gives a 6.46% performance increase compared to initializing with BERT. Second, when using BUTD, the model is less sensitive to the transformer initialization and the performance is generally low (from 83.09% to 85.64%). In contrast to other V&L tasks [LYY19], general-domain V&L pre-training is not instrumental. The above findings suggest that for medical V&L appli-

Visual Encoder	BUTD			PixelHop++	
Transformer Backbone	VB	BERT	BlueBERT	BERT	BlueBERT
Atelectasis	0.9247	0.8677	0.8866	**0.9890**	0.9838
Cardiomegaly	0.9665	0.8877	0.8875	0.9772	**0.9896**
Effusion	0.9049	0.8940	0.9120	0.9013	**0.9432**
Mass	0.6428	0.7365	0.7373	0.8886	**0.9900**
Consolidation	0.7870	0.8766	0.8906	0.8949	**0.9671**
Emphysema	0.8565	0.7313	0.8261	0.9641	**0.9971**
AVG	0.8386	0.8309	0.8564	0.9177	**0.9823**

Table 2.1: Effect of the transformer backbones when paired with different visual encoders. When using BUTD features, the model becomes insensitive to the transformer initialization and the expensive V&L pre-training brings little benefit compared to BERT. When using PixelHop++, the model benefits significantly from BlueBERT, which is pre-trained on in-domain text corpora.

cations, in-domain single modality pre-training can bring larger performance improvement than using pre-trained V&L models from the general domain, even though the latter is trained on a larger corpus. The relation and trade-off between single-modality pre-training and cross-modality pre-training are overlooked by previous works [LYY19] and we advocate for future research on this.

2.3.4 Effectiveness of BERTHop with different dataset scales

To demonstrate the effectiveness of BERTHop on datasets of different scales and justify our designs, we experiment to compare BERTHop with its two variants: (1) In PH_BERT, we replace BlueBERT with BERT. We compare BERTHop with PH_BERT to show how a domain-specific BERT model helps to improve performance in medical applications. (2) In BUTD_BlueBERT, we replace the visual encoder PixelHop++ with the general visual encoder of BUTD.

We randomly select fractions of the training set of OpenI to train these three models and compare their performance on the entire test set of OpenI. Figure 2.3 illustrates that the performance of BERTHop is consistently better than the other two settings.

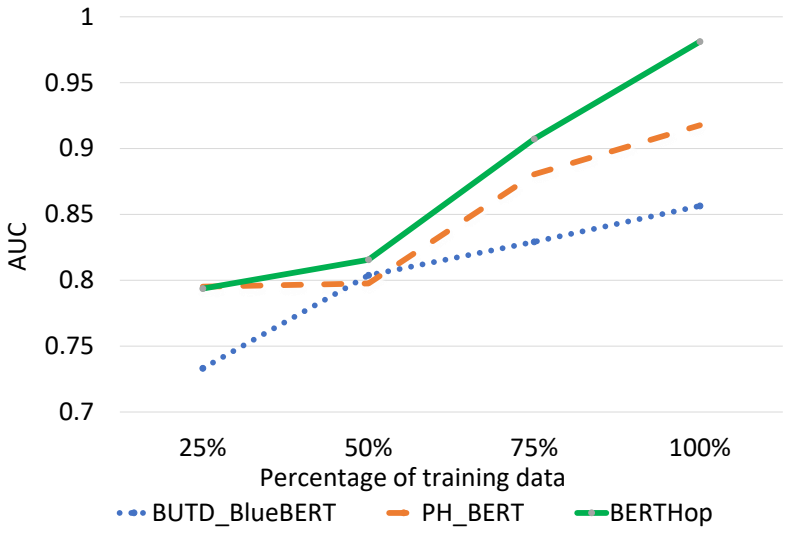


Figure 2.3: Avg AUC of three settings with different percentages of training data. BERTHop remains effective with different dataset scales.

2.3.5 Visualize abnormal regions identified by BERTHop

We visualize PixelHop++ output channels of BERTHop to probe whether it can effectively capture abnormal regions in CXR images. In this study, we asked two radiologists to annotate pathology regions of a few examples related to different diseases. As shown in Figure 2.4, some output channels can successfully highlight the abnormalities in CXR images. This is since PixelHop++ extracts image representations at different frequencies which is beneficial for abnormality detection.

2.4 Discussion and Conclusion

Our research introduces a multimodal model that simultaneously processes X-ray imagery and clinical text with high efficiency and performance. Diverging from standard V&L models that employ object detectors for visual data extraction, we employ PixelHop++, a highly efficient unsupervised encoder. Our evaluations demonstrate PixelHop++'s robustness and emphasize the critical

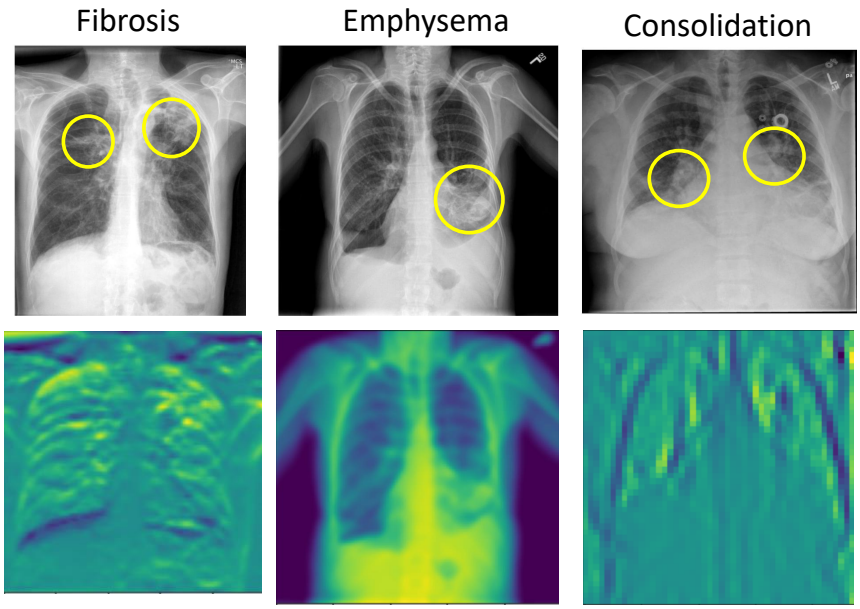


Figure 2.4: On the top, we mark the pathology regions annotated by two radiologists (the yellow circles and lines); on the bottom, we visualize the visual features from BERTHop (brighter colors means higher feature values). BERTHop can successfully highlight the abnormal regions identified by expert radiologists.

role of appropriate transformer pre-training, offering crucial perspectives for future model development. We advocate for the adaptation of this methodology in medical fields where annotated data is scarce. We contend that BERTHop significantly enhances accuracy in medical diagnosis, potentially reducing errors.

CHAPTER 3

Few-shot Visual Question Answering

In this chapter, we study Few-shot visual question answering (VQA). VQA is a task where the model is asked to answer a question given an image. There are several types of questions e.g. yes/no questions, color questions, quantity, and conceptual questions. The model has to have an up-to-the-mark and detailed representation of the image while understanding the question which is in text form to be able to answer the question properly. Therefore, training such a model which has several components is significantly tricky and challenging. For the answer, we expect the model to either select an answer from a list of candidate answers (multiple choice) or generate the answer as a sequence of text (open-ended). The latter is more challenging and has recently got more attention in the AI community. There are many proposed Vision-Language (VL) models for this task all requiring large datasets for high performance while their performances are close to random choice when few training data are available. We are interested in the scenario when few data is at hand and how large language models could ease the problem as they are few-shot learners [].

Pre-trained language models have shown impressive performance on a range of tasks by learning from large-scale text corpus [RNS18, RWC19, YDY19]. Recent studies find that some of these language models can be used to perform *in-context learning* out-of-the-box, i.e., adapting to a task by conditioning on a few demonstrations in context without any gradient update [BMR20, MLZ22], which is highly desirable.

In VL modeling, in-context learning is less explored and only a handful of models are proposed

to perform in-context learning mainly by limiting the amount of deviation of a pretrained large-scale language model from the language space and translating visual inputs to language embedding space. They either require a large capacity [TMC21, ADL22] or a giant corpus consisting of in-context learning examples [ADL22, LLW24, KSF23].

In this work, we explore whether we could enable in-context learning in VL tasks without resorting to extreme scale-up. We study an interesting hypothesis: can we transfer the in-context learning ability from the language domain to the VL domain? To elaborate, not every language model exhibits excellent *in-context* learning ability; recent studies [MLZ22] show that one could explicitly train language models to perform in-context learning, by training the model on multiple tasks with in-context few-shot examples, a process that resembles meta-learning. Thus, an intriguing query arises: when a language model is first meta-trained to perform in-context learning, can it be transferred to perform in-context learning for VL tasks better?

A remarkable observation in our study is the utilization of a meta-trained language model as the transformer encoder-decoder and the mapping of visual features to the language embedding space. This innovative approach led to the development of our proposed VL model (we name it MetaVL). Impressively, our experimental results demonstrate that MetaVL surpasses the baseline model's performance, even when MetaVL is designed to be 20 times smaller in size.

This study makes three main contributions: 1) To the best of our knowledge, this is the first attempt to transfer the meta-learning knowledge for in-context learning from single-modality to multimodality. 2) We propose a VL model, MetaVL, which outperforms the baseline in in-context learning while having a much smaller model size. 3) Through extensive experiments on VQA, GQA and OK-VQA, we demonstrate the in-context learning capability of MetaVL and analyze its components.

3.1 Related work

3.1.1 In-context learning in VL

Frozen [TMC21] is the first attempt for in-context learning in multimodality by leveraging a frozen GPT-like language model as the language backbone and mapping visual features to the language embedding space. Frozen sheds light on the feasibility of benefiting from the frozen LMs in VL modeling to learn a new task from a few examples in context. MAGMA [EBW21] is another encoder-decoder architecture for VL pre-training which showed that adding adaptor blocks between the frozen language model layers could further improve the performance for VL tasks in a few-shot scenario. Other recent works [YGW22, ADL22, ZAI22] follow the similar principle as the previous works to tackle in-context learning in VL modeling and achieve superior results by leveraging extremely large-scale models.

In this paper, we study a problem overlooked in prior work: we delve into the possibility of enabling in-context learning for VL tasks without relying on extensive scalability. Our focus lies in exploring the hypothesis: Is it feasible to transfer the in-context learning capability from the language domain to the VL domain?

3.1.2 Meta-learning in language modeling

Large-scale language models have shown the capability to be trained on a new task if properly prompted with in-context examples, i.e., in-context learning. In this learning strategy, the language model is asked to generate the desired output, e.g., an answer in the question-answering task, which is prompted by a few data examples along with their corresponding supervision sampled from the training split, and the language model learns the task in context without performing any gradient updates. Although such training is highly data-efficient, its performance is far behind supervised fine-tuning. Therefore, inspired by [VD02, EP04, FAL17, Rud17], MetaICL [MLZ22]

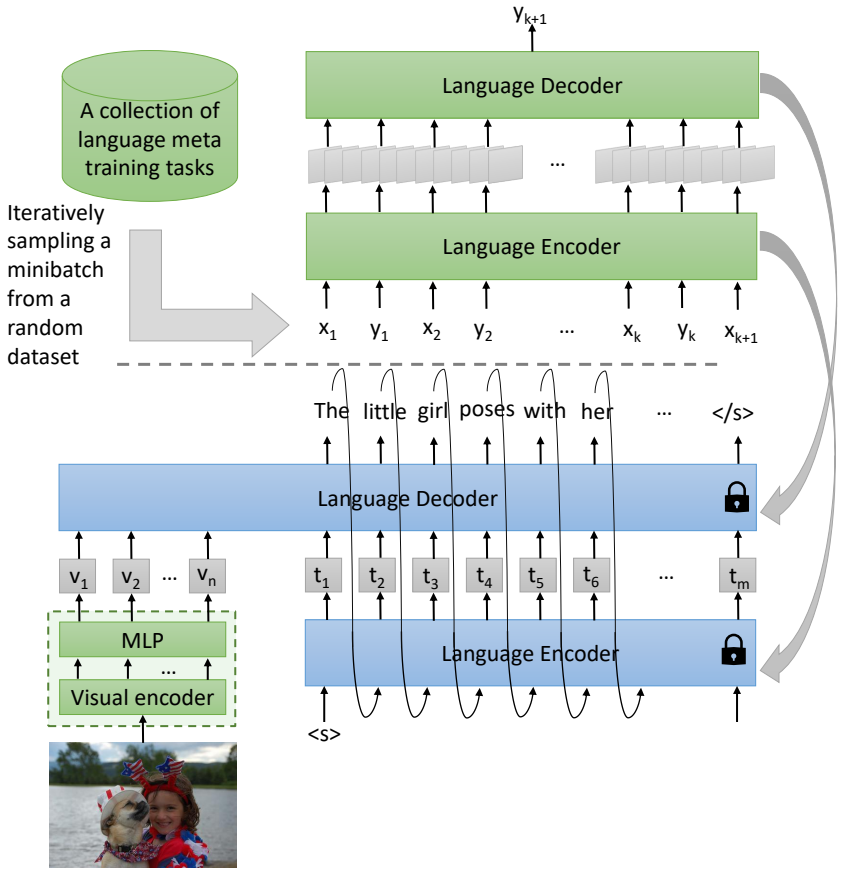


Figure 3.1: The training steps of MetaVL including meta-training the language encoder-decoder (above) and mapping the visual features into the language embedding space while keeping the meta-trained language encoder-decoder frozen (below).

proposes training the model for in-context learning as a kind of meta-learning. MetaICL meta-trained a gpt language model on a diverse set of natural language tasks and datasets and showed that meta-training a language model in an in-context learning manner could significantly improve the in-context learning capability of the language model for a new task.

3.2 Approach

In this section, we first explain the existing meta-training procedure for language modeling and then introduce our proposed method for in-context learning in VL.

3.2.1 Meta-training in language modeling

MetaICL has shown that a language model that is meta-trained on a diverse set of tasks in an in-context learning setup is a strong few-shot learner. To meta-train an auto-regressive language model, in each iteration, a meta-learning task is randomly chosen from a collection of diverse meta-training language tasks, and $k+1$ data-label examples are randomly sampled from its training split. Then, the model is supervised by the concatenation of $(x_1, y_1, x_2, y_2, ..., x_{k+1})$ which will be fed as a single input to the model for predicting the label (y_{k+1}) as the training objective, i.e., the meta-training step aims to maximize:

$$P(y_{k+1}|x_1, y_1, \cdots, x_k, y_k, x_{k+1}) \tag{3.1}$$

During inference, the same in-context setup (k examples from the training) are sampled from a target dataset to be used as the $(x_1, y_1)(x_2, y_2) \cdot \cdot, (x_k, y_k)(x)$ and given to the model to predict the label y. The meta-trained language model trained on a diverse set of natural language datasets has shown good performance for an unseen task when few data are given in context [MLZ22].

MetaVL has three main submodels including a meta-trained encoder-decoder and is being trained using Prefix Language Modeling (PrefixLM) [WYY21]. In the following, we discuss each submodel in detail.

3.2.2 Visual encoder and visual prefix.

The visual encoder is defined as a function $V_e(x)$ that takes an image of x and outputs visual features. We extract the feature grid before the pooling layer $n \times D_v$ where n is the number of feature maps and D_v is the feature size of the visual encoder. Then, the output features can be viewed as a sequence of n visual tokens representing the image.

The visual encoder is followed by the visual prefix module that is defined as $V_p(x) \in D_v \times D_l$

which maps the visual features to language embedding space. This module is seeking to properly project the visual tokens into language tokens.

During the VL training, the parameters of both of these modules are trainable and are learned with different learning rates by back-propagation guided by the frozen language model.

3.2.3 Language encoder-decoder

The meta-trained language encoder-decoder is used as the LM backbone and is frozen during the VL training process so the meta-trained language model preserves its few-shot capabilities. The language encoder encodes the text into text tokens represented by $t_1, t_2, ..., t_m$. Then, given the multimodal tokens (image and text) as $U = v_1, v_2, ..., v_n, t_1, t_2, ..., t_m$ the decoder is trained to reconstruct the corresponding text with a standard language modeling objective to maximize the following likelihood:

$$L(U) = \sum_{i=1}^{m} \log P(t_i | v_1, ..., v_n, t_1, ...t_{i-1}; \theta) \tag{3.2}$$

After the VL training, for learning a new VL task in-context, given a few examples from a new task with a new format, we concatenate k sampled data-label pairs from the training split along with one data from the val/test split to construct the prompt and feed it to the model for predicting the desired output. The entire process is visualized in Fig. 3.1.

3.3 Experiments

3.3.1 Datasets and Baseline

We use the dataset proposed in [MLZ22] as the meta-training dataset for the language model and the COCO dataset [LMB14] as the VL training dataset for MetaVL. The evaluation experiments are conducted on three datasets including VQA [AAL15], OK-VQA [MRF19], and GQA [HM19]. Frozen leveraged an internal GPT-like language model with 7 billion parameters as the backbone of their proposed model. As their model is not publicly available, we trained Frozen with GPT2-

Medium as the frozen language model and consider it as our main baseline (Frozen$_{\mathrm{A}}$) due to its model size. We also train a frozen with GPT-J 6B (The most similar GPT to Frozen) language model and obtained a close performance to the original Frozen model and use it as our second baseline denoted by Frozen$_{\mathrm{B}}$.

3.3.2 Training and evaluation setting

Initially, We meta-train a GPT2-Medium LM on a collection of 142 meta-training language datasets with a learning rate of 1e-5 and a batch size of 8 using the setting named as "HR→LR with instructions (all)" where datasets with equal or greater than 10,000 training examples are used as meta-training tasks and the rest of the datasets are used as target tasks. The training is done on 8 NVIDIA RTX A6000 for 80,000 steps which took $\sim$ 6 hours. Then, we train MetaVL on the training split of COCO where we use a learning rate of 5e-5 and 2e-6 for the visual prefix and visual encoder, respectively, while the rest of the model parameters are frozen. We use a batch size of 32 and trained MetaVL using 4 NVIDIA RTX A6000 for 8 epochs which take $\sim$ 48 hours. Inference time depends on the numebr of shots varies from 2-5 hours for 0-3 shots on 5000 test examples. Our visual encoder is CLIP-RN50x16 [RKH21] with a feature grid size of 144×3072 and our visual prefix is an MLP layer with a dimension of 3072×768. For in-context evaluation on VQA datasets, we randomly pick a specific number -n- of sampled data-label pairs, known as shots, from the training set and feed them to the model in-context followed by a single data from the val/test set. Fig. 3.4 provides some illustrative examples for the evaluation process.

To conduct the evaluation, we utilize a subset of 5,000 instances from the val/test dataset due to computational constraints. The generated output from the model is then compared against the expected answer, as established in previous studies. In cases where an exact match is not achieved, we employ a technique to identify the most closely related answer from a set of candidate answers (The set can be defined as a unique list of all answers in the training dataset). This involves comput-

ing the cosine similarity between the output's embedding and each candidate answer's embedding achieved by Sentence BERT [RG19].

We then compare the selected output with the corresponding answer to determine the match. The training datasets for VQA, OK-VQA, and GQA contain approximately 3,000, 4,200, and 3,000 distinct answers, respectively. Furthermore, we performed an additional round of human evaluation on model's output without matching, and the findings are summarized in the appendix (Table 2). The human evaluation on a separate test set of 2000 examples aimed to delve deeper into instances where the model's output, while accurate, didn't precisely match the provided answer. Three such examples are presented in Fig 3.2, where the initial evaluation did not consider the prediction as correct, but it was deemed correct in the subsequent evaluation setting.

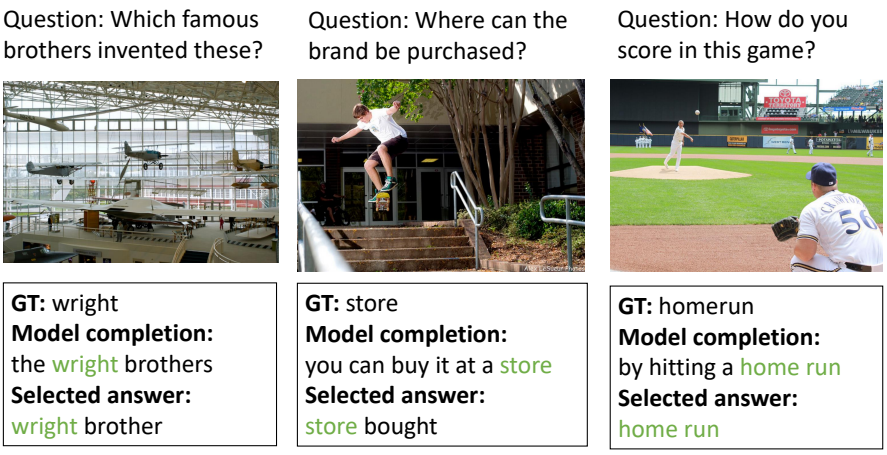


Figure 3.2: Three examples of VQA cases which The model's output, although correct, slightly differs from the ground-truth and selected answer from the candidate set.

		Frozen_A	Frozen_B	MetaVL
	LM size	375M	7B	375M
Automatic evaluation	VQA	18.63	**34.07**	33.12
	OK-VQA	3.17	**11.97**	9.60
	GQA	13.86	25.76	**31.96**
Human evaluation	VQA	16.68	-	**35.09**
	OK-VQA	6.41	-	**19.22**
	GQA	19.96	-	**38.29**

Table 3.1: The performance of MetaVL compared with two baselines on 3-shot in-context learning. We report the performance of our re-implemented Frozen models.

3.3.3 Results and analysis

3.3.4 Quantitative analysis

To evaluate MetaVL, we consider three common visual question-answering datasets including VQA, OK-VQA, and GQA. We compare MetaVL results with the mentioned two baselines in Table 3.1 for 3-shot in-context learning based on both automatic and human evaluation. According to the results, the performance of Frozen improves as its model size increases while MetaVL achieved competitive results in all three tasks. To further analyze how many image-text pairs are required to enable In-context learning for the VL task, we have trained MetaVl with 50 percent of training data and the results show that the performance slightly dropped but the model preserved its capability to learn from in-context data (Table 3.3).

3.3.5 The effect of the number of in-context shots

According to Figure 3.3, in almost all settings, the performance of MetaVL is improving by increasing the number of shots which shows the model is gaining knowledge from the data in context. This result further gives us an illustration of the model's capability to learn from the in-context examples supporting that MetaVL is benefiting from the meta-learning knowledge for in-context learning. The numbers on the graph are summarized in Table 3.2.

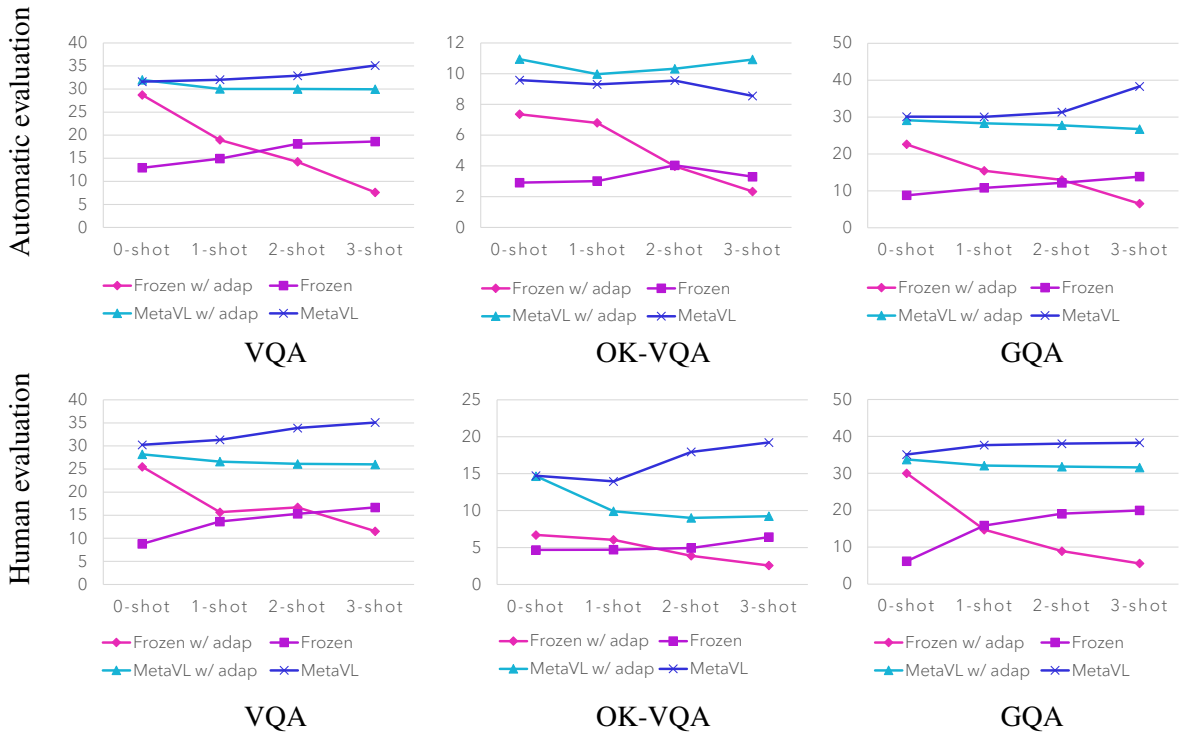


Figure 3.3: Automatic and human evaluation Accuracy of MetaVL and Frozen, w/ and w/o adaptors with 0-3 shots of in-context data.

3.3.6 The effect of having adaptor layers in LM

MAGMA claims that adding trainable adaptor layers and letting the LM slightly be trained during the VL training process is beneficial for in-context learning. Compared with Frozen, in addition to being trained on an x8 larger set of VL datasets, MAGMA also includes the training splits of the target datasets to its training set, while Frozen is adapted to an unseen new task in-context (in-context learning). We evaluated this method by adding adaptor layers to both Frozen and MetaVL and denoted the corresponding models by Frozen w/adap and MetaVL w/adap, respectively, in Fig. 3.3. Our results demonstrate that having a fully frozen language model in MetaVL could better preserve the in-context learning ability of the language model. It is also noticeable that adding adaptor layers improves the zero-shot performance of Frozen. We hypothesize that this improvement is due to getting a better vision and language alignment by letting both vision and

model	Frozen$_A$ w/ adap				Frozen$_A$				MetaVL w/ adap				MetaVL			
n-shot	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Automatic evaluation																
VQA	28.72	18.98	14.23	7.60	12.94	14.92	18.11	18.63	31.98	30.03	30.01	29.96	31.6	32.01	32.89	**33.12**
OK-VQA	7.36	6.30	3.98	2.34	2.91	3.02	4.04	3.30	**10.94**	9.97	10.32	10.92	9.58	9.30	9.55	9.60
GQA	22.62	15.44	12.96	6.54	8.80	10.81	12.17	13.86	29.12	28.31	27.78	26.74	30.10	30.05	31.32	**31.96**
Human evaluation																
VQA	25.49	15.66	16.70	11.53	8.79	13.62	15.31	16.68	28.20	26.61	26.12	26.01	30.24	31.33	33.89	**35.09**
OK-VQA	6.70	6.04	3.88	2.56	4.67	4.71	4.94	6.41	14.67	9.97	9.01	9.24	14.72	13.95	17.95	**19.22**
GQA	30.01	14.72	8.92	5.59	6.18	15.85	19.07	19.96	33.74	32.09	31.81	31.58	35.08	37.65	38.03	**38.29**

Table 3.2: Accuracy of MetaVL and Frozen, w/ and w/o adaptors with 0-3 shots of in-context data.

		MetaVL	MetaVL$_{50\%}$
Automatic evaluation	VQA	**33.12**	30.32
	OK-VQA	**9.60**	7.56
	GQA	**31.96**	27.77
Human evaluation	VQA	**35.09**	34.02
	OK-VQA	**19.22**	18.19
	GQA	**38.29**	35.66

Table 3.3: The performance of MetaVL was evaluated using the complete CoCo training dataset as well as a subset containing 50 percent of the CoCo training data. The experimental results indicate that even with the reduced training data, MetaVL maintains its capacity for in-context learning, albeit with a slight decrease in performance.

language submodels be involved in the alignment process.

3.3.7 Qualitative analysis

We provide some qualitative examples to better illustrate the performance of MetaVL for in-context learning in different VQA tasks. Some positive and negative examples are provided in Fig. 3.4 and Fig. 3.5 respectively, which show the output of MetaVL for 3-shot in-context learning.

3.4 Limitations and Conclusion

While we have shown the potential of transferring in-context learning ability from a language model to VL tasks, the experiments in this paper are limited in two aspects. (1) We considered only

the VQA task, which is limited in scope. It is unclear whether our method generalizes to other VL tasks. In fact, as most tasks in the VL domain take the form of visual question answering, it is less well-defined what would "cross-task generalization" entail in VL, compared to in NLP where (2) Due to computational limitations, we experiment with only a moderate-sized LM. It is unclear the performance of our method after scaling up.

We investigate the feasibility of transferring meta-learning knowledge for in-context learning from resource-rich single modality to multimodality. We have shown that by leveraging a meta-trained language submodel in a vision-language model, we can transfer the ability of "learning to learn" in-context to VL and it results in a strong VL few-shot leaner. With extensive experiments on three common VL datasets, we have shown that the in-context learning performance of our model, MetaVL, is superior compared with the baseline even when the size of our model is 20 times smaller.

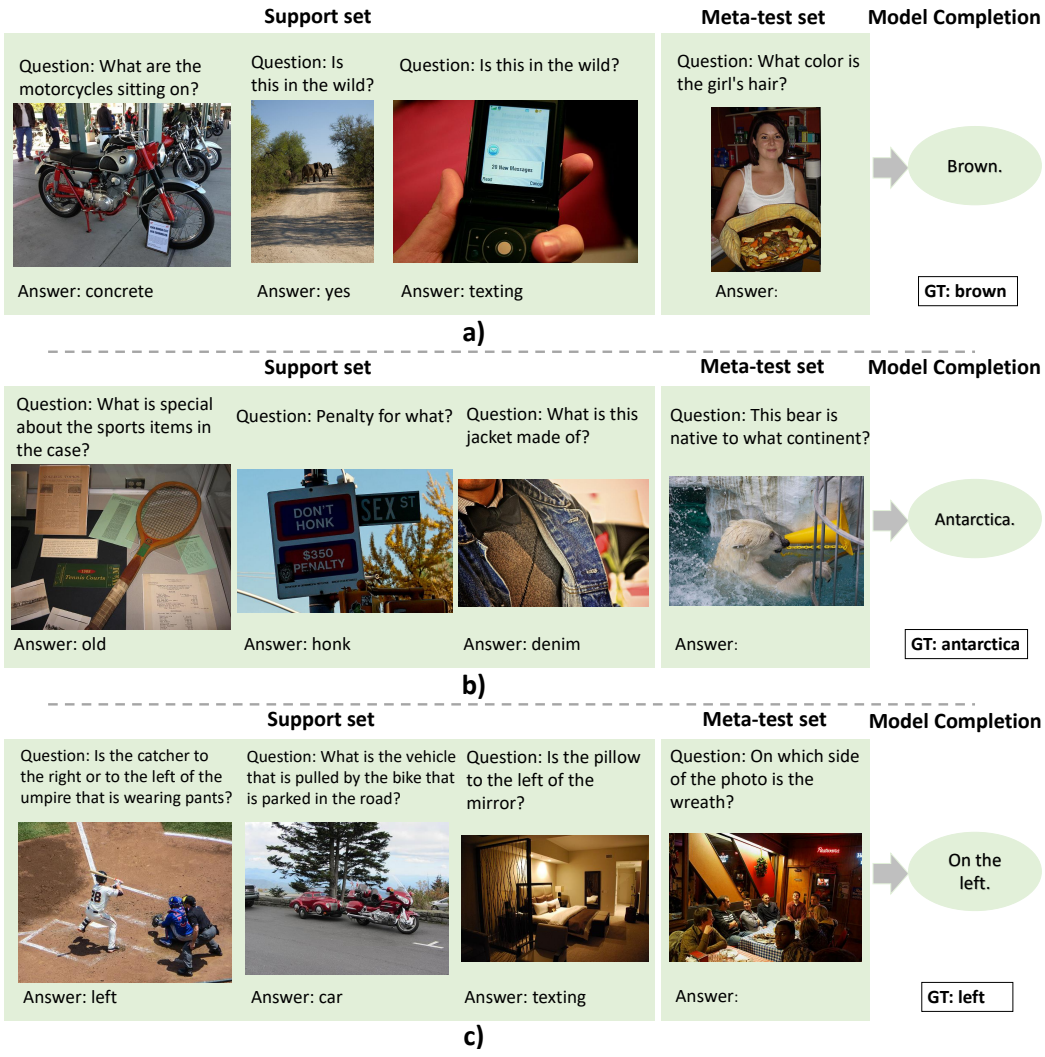


Figure 3.4: Qualitative examples of in-context learning from three datasets: a) VQA, b) OK-VQA, and c) GQA. For each example, there is also a task induction sentence of "please answer the question.".

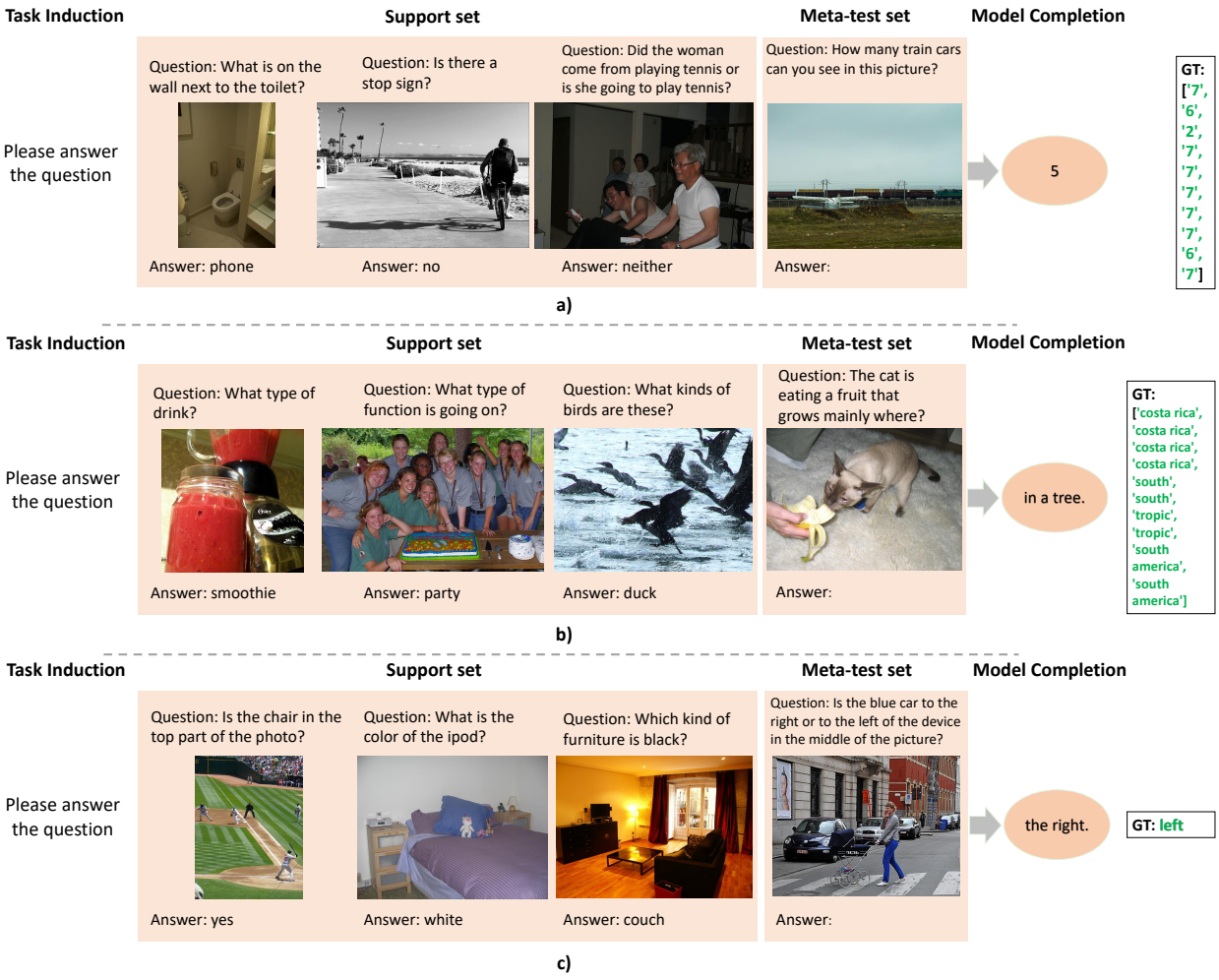


Figure 3.5: MetaVL failure examples from a) VQA, b) OK-VQA, and c) GQA.

CHAPTER 4

Foundational Biomedical Multimodal modeling

Readily available open-access datasets in the broad domain [LMB14, SDG18] have enabled the development of vision-language (VL) models. Typically, researchers pre-train VL models on large image-text data and then fine-tune them for specific downstream tasks. Such a pre-training/fine-tuning paradigm is highly effective for tasks with limited data and therefore is a standard approach for various downstream applications.

On the other hand, the medical domain presents unique challenges due to the scarcity and complexity of its data [BI21]. Pre-trained models, trained on general domain datasets, often exhibit reduced effectiveness when applied to a medical domain with limited data, as observed in the cases of chest x-ray analysis [WZZ23b, MRL22] and Alzheimer's disease detection [CHW23]. Furthermore, domain-specific data suitable for pre-training is notably limited. Although several domain-specific medical Vision-Language (VL) datasets do exist, such as MIMIC [JPG19], CheXNet [RIZ17], and datasets for Alzheimer's disease [PAB10], their creation requires substantial manual curation, involving significant labor and time. Besides, they are presented for specific domains and their knowledge is not transferable to other medical domains.

In this paper, we take the brain disease domain as an example and propose an automatic pipeline for extracting image-text pairs for pre-training a VL model for specific medical domains. The pipeline first collects raw image-text pairs from medical sources like PubMed and prepares aligned image-text pairs which can be leveraged for VL pre-training. There are two main chal-

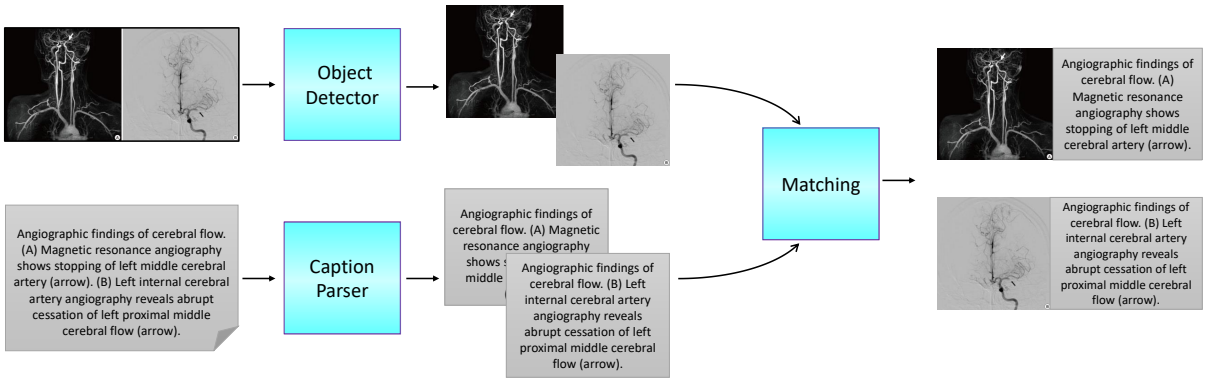


Figure 4.1: Our main pipeline for matching subfigures/subcaptions. The object detector outputs subfigures while the caption parser parses the caption into subcaptions simultaneously. Then, the module called 'matching' provides us the subfigure/subcaption pairs for the pre-training

lenges with the data we collected. First, medical data in both image and text form inherently possess a complex nature [HYK18]. Moreover, image-caption pairs sourced from PubMed literature and journals often incorporate subfigures and subscaptions, introducing additional intricacies that can pose formidable challenges for the pre-training of VL models. Therefore, there is a unique challenge in medical VL pre-training where we care about the fine-grained alignment between subfigures/subcaptions. Our pipeline is equipped with subfigure/subcaption aligning designed to enhance multimodal learning. The letters in subfigures and subcaptions are often referred to as "subfigure labels" or "subcaption labels." are used to identify the alignment between subfigures and subcaptions.

Getting the aligned image-text pairs from our pipeline, we pre-train a VL model, BLIP [LLX22], on both the original collected data and processed data to analyze the effectiveness of our pipeline. We consider quantitative and qualitative intrinsic evaluations including image-text retrieval and attention visualization. We have observed that the model pre-trained on the processed data has a better multimodal understanding. We believe that our pipeline can be used for other domain-specific medical applications such as Prostate Cancer Diagnosis or Alzheimer's Disease Prediction.

Our work has three main contributions. We identified the challenge in medical VL pre-training by taking brain disease as an example. We built a pipeline that collects domain-specific medical image-text pairs followed by a module for matching subfigures/subcaptions. We release the dataset which is suitable for pre-training a VL model for brain diseases upon the acceptance of our work. Lastly, we pre-trained a VL model (BLIP) [LLX22] on our dataset, and with quantitative and qualitative analysis, we demonstrate that our pre-trained dataset and model are useful for building VL models in the medical domain.

4.1 Related Work

4.1.1 Vision-Language Learning.

Large-scale VL pretraining has demonstrates impressive performance for various VL tasks [LBP19, TB19b, SZC19, LYY19]. Early works in this direction use off-the-shelf object detectors to extract objects [AHB18] and feed the region features into a modality fusion module [SZC19, TB19b, CLY20, LYL20, ZLH21]. End-to-end training methods have been proposed to be efficient and effective alternatives to object detector-based models [KSK21, LSG21, DXG22] because of their ability to unleash the power of vision encoders. While state-of-the-art VL models have achieved impressive performance, many of them are mainly focused on general-domain VL tasks such as visual question answering and image captioning, and it is unclear whether the pre-trained representations can be helpful for medical tasks.

4.1.2 Medical-Domain VL Learning.

Multi-modal learning in the medical domain is of great significance due to the presence of medical images, text notes, and electronic health records. While there are a few medical-domain datasets and models being proposed [EMD23, LYW21], many of them are focused on CLIP-style models [RKH21] trained with contrastive objectives, neglecting to investigate more effective VL

training methods in a large-scale setting. In addition to this line of work, **(author?)** [ZHY22] introduce the mmFormer, a novel Transformer-based approach, for accurate brain tumor segmentation from incomplete MRI modalities, achieving significant improvements in segmentation performance compared to state-of-the-art methods on the BraTS 2018 dataset. **(author?)** [LZZ23] present PMC-OA, a large-scale biomedical dataset with 1.6M image-caption pairs from PubMedCentral's OpenAccess subset. Using this dataset, the PMC-CLIP model achieves state-of-the-art results in image-text retrieval and classification tasks, addressing data scarcity issues in the biomedical domain. In this work, we adapt the advanced techniques in the general domain VL learning to the medical domain and demonstrate its effectiveness.

4.2 Data collection and processing

In this section, we explain the details of our pipeline for processing the data collected from PubMed and pre-training the vision-language (VL) model.

4.2.1 Image-caption pairs.

Out of 9,371 journal papers from 1,021 various journal titles like "Brain tumor research and treatment" and "BMC Medical Imaging" dated from 1937-2018, We collected image-caption pairs from PubMed and filtered them by only scraping "case reports" with both types of image and text data. Pairs are additionally filtered by the keyword "brain" that appears in the corresponding caption assuming that both the image and caption are brain-related data. We collected 22,795 data initially and pre-trained the BLIP model on image-caption pairs as our baseline. Other data from the medical domain can be collected following our procedure to build another domain-specific VL model.

4.2.2 Fine-grained alignment.

Due to the presence of subfigures and subcaptions in a significant portion of collected images (43 percent of data), there is a unique challenge in their usage for the pre-training process and following the pre-training schema of existing VL models (taking the entire image and entire caption as image-text pairs) could lead to a low-performance model. This is because the model may struggle to match subcaptions and subfigures, which is not the issue in general domain VL datasets. Additionally, in numerous images within the dataset, subfigures exhibit significant distribution variations, possibly stemming from diverse sources such as MRI scans, surgical cameras, or simulation visualizations (as exemplified in Fig. 4.2). Presenting the complete figure and caption to the model could potentially lead to misguidance, as it may focus on image portions less pertinent to the caption and subcaptions. Moreover, the complexity of medical terms in clinical notes and abnormal regions in images poses a significant challenge to the model's understanding. To overcome these challenges, we developed a pipeline shown in Figure 1 where we used an object detection model [TC17] to identify subfigures in the images while simultaneously leveraging an NLP tool [TSP21] to parse captions into subcaptions. With these two steps, we were able to convert the collected image-caption pairs into 39,535 subfigure/subcaption pairs. However, matching subfigures and subcaptions was another challenge in training the VL model. We utilized an OCR tool[1] to detect the small characters in the subfigures commonly referred to as "subfigures labels" that are most likely the letters we can use to match with subcaptions. If the detected text does not match with any of the "subcaption labels" or does match with two or more of them, we pair the subfigure with the entire caption. Then aligned subfigures/subcaptions pairs are used for pre-training the model.

[1]google OCR: `https://cloud.google.com/vision/docs/ocr`

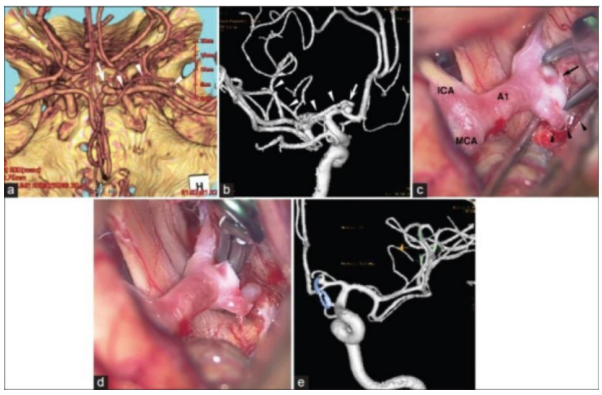


Figure 4.2: An example data from the dataset where each subfigure comes from a different source and Providing the model with the entire figure and its accompanying caption could potentially result in misdirection, causing it to emphasize image areas that are less relevant to both the caption and subcaptions.

4.2.3 Pre-Training and Evaluation

We trained the BLIP model on the processed data with a learning rate of 1e-5 for 60 epochs, using Adam as the optimization algorithm and Masked Language Modeling (MLM) [DCL18], Image-text matching (ITM) [CLY20], and Image-text contrastive (ITC) [RKH21] as the pre-training objectives. We evaluated the pre-trained model on qualitative and quantitative intrinsic tasks including image-text retrieval. Image-text retrieval evaluates the model's ability to retrieve the corresponding text/image from a pool of data given an image/text. We also conducted qualitative analysis by visualizing the attention map of some important medical terms and their relationship with abnormal regions.

4.3 Experiments

4.3.1 Quantitaive evaluation

Our pre-trained model for image-text matching is evaluated to assess its ability to match an image to its corresponding caption. To perform this evaluation, we evaluate the model performance for

		Val				Test			
		i2t@1	i2t@10	t2i@1	t2i@10	i2t@1	i2t@10	t2i@1	t2i@10
Only PT	Raw data	18.04	48.36	21.51	51.98	17.37	47.65	19.58	39.86
	Processed data	**36.9**	**69.88**	**38.4**	**69.04**	**36.94**	**69.37**	**37.40**	**69.58**
PT + FT	Raw data	24.6	60.47	27.13	59.14	24.45	59.4	26.1	58.7
	Processed data	**38.22**	**72.09**	**39.98**	**69.56**	**36.52**	**72.62**	**36.38**	**70.1**

Table 4.1: Image-text retrieval results for the BLIP model pre-trained/fine-tuned (PT for only pre-training and PT + FT for pre-training plus fine-tuning) under two different schemes: following existing VL models and our pipeline.

image-text retrieval tasks as an intrinsic evaluation. In this task, we test how well our model can retrieve the corresponding caption given an image and a pool of all captions, and vice versa. We evaluate our model's performance in two settings: '@1', where we determine the percentage of val/test dataset where the corresponding image or caption is the top 1 retrieved data; and '@10', where we determine the percentage of data where the corresponding image or caption is among the top 10 retrieved data. Our results which are summarized in Table 4.1 validate our pre-trained model's capability for VL modeling. Additionally, we conducted pre-training experiments with varying proportions of the available data, ranging from 25 to 100 percent, and assessed the impact on our image-text retrieval performance. As depicted in Figure 4.3, our model exhibited a steeper performance curve, indicating its increased learning capacity as more data became available. In contrast, the baseline performance appeared to saturate and reach a plateau. This observation underscores the effectiveness of our pre-training pipeline in enhancing a model's learning capabilities.

4.3.2 Qualitative analysis

As part of our qualitative analysis, we employ attention maps to visualize the model's focus on some important medical terms, such as 'aneurysm'representing abnormal blood vessel outpouchingwithin the associated images and 'cerebral artery' a condition where there is a deviation from

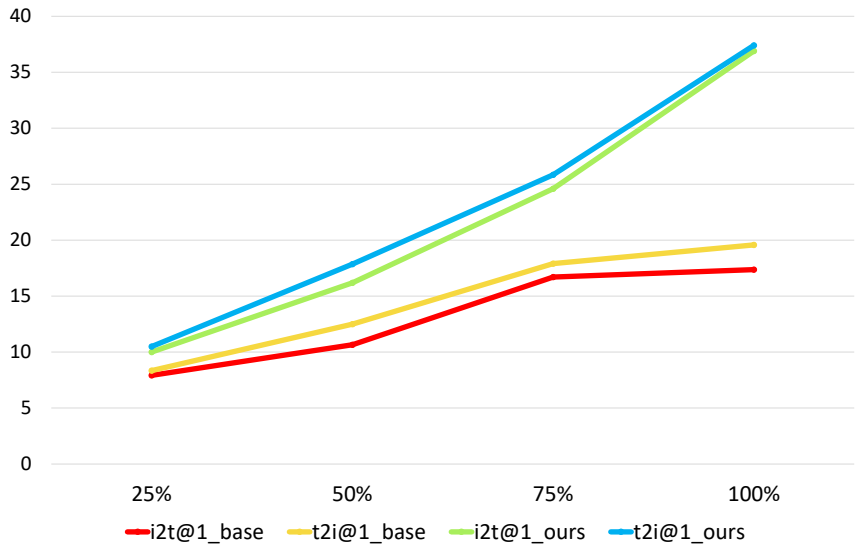


Figure 4.3: The plot visually demonstrates the dynamics of image-text retrieval performance across varying proportions of pre-training data. Our model results reveal a notable increase in learning capacity as more data becomes available, while the baseline exhibits characteristics that suggest a form of saturation.

the normal, healthy state of arteries. In Fig. 4.4 and 4.5, we present the model attention map for both the baseline and our pre-trained model concerning the terms 'cerebral artery' and 'aneurysm' respectively where 'aneurysm' is divided into subtokens. The upper heatmaps correspond to the baseline, while the lower ones depict the outputs of our pre-trained model. The attention maps demonstrate that our model emphasizes the relevant areas.

4.4 Discussion and Conclusion

Our study is centered on the creation of a pre-training dataset for domain-specific medical vision language models, utilizing a VL dataset sourced from PubMed and a pipeline for VL modeling. Our approach has been evaluated through intrinsic task assessments, demonstrating its effectiveness in constructing data-efficient VL models for the medical field. Our model has surpassed the performance of standard VL training methods, further validating the efficacy of our pipeline. Our proposed pipeline helps build powerful multimodal AI models which could be used for various

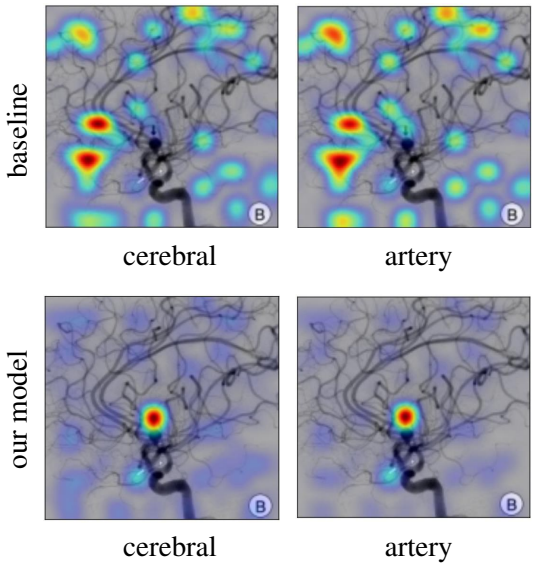


Figure 4.4: The visualization of the attention of the "cerebral artery" on the corresponding image. It is evident that ours highlighted the abnormal region more specifically.

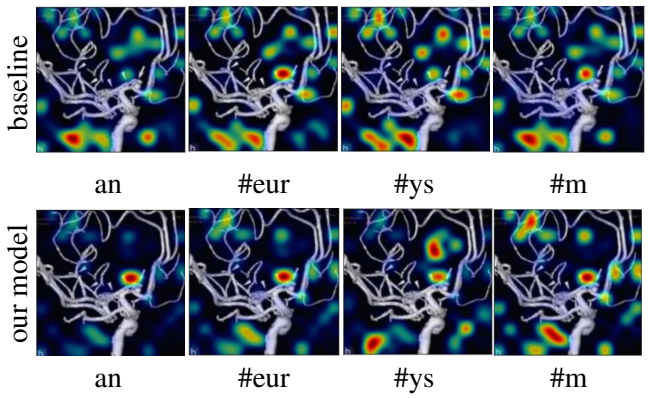


Figure 4.5: The visualization of the attention of the "aneurysm" on the corresponding image. It is evident that ours highlighted the abnormal region more specifically.

tasks like diagnosis, drug discovery, and information extraction.

CHAPTER 5

Large Language Models in Biomedical domain

LLMs such as GPT4 have demonstrated exceptional capabilities across diverse tasks and domains [EEA23, DSD23, DLD22]. These models could have a revolutionary impact on healthcare; however, their integration into medical research and practice has been slow [ZZG23, VMV23, NKM23] and it is crucial to examine the unique challenges presented by the biomedical field that contribute to this discrepancy. Specifically, LLMs encounter challenges in medical Information Extraction [GMW22, MBH21] due to the scarcity of high-quality biomedical data in their pretraining, and the need for a nuanced comprehension of the text for this task [GZU23]. Medical entities can have multiple synonyms and abbreviations, complicating their recognition by models [GGM21]. Furthermore, context sensitivity is even more critical in the biomedical compared to the general domain. The specificity of entity types and the complexity of their interrelations necessitate a level of background knowledge that standard prompts may fail to provide. LLMs are primarily exposed to vast amounts of generic text data limiting their effectiveness in managing the intricate nuances of medical language [KKS23, KM23].

In this paper, we concentrate on NER, a foundational task for various applications such as recruiting patients for clinical trials, searching biomedical literature, or building models that predict the progression of disease based on free-text notes.

In our initial analysis, we broaden the scope of TANL [PAK21] and DICE [MTW22], two text-to-text formats initially proposed for model training, adapting their use to prompt design specifi-

cally for biomedical NER. Our findings reveal that the relative effectiveness of the resulting prompt pattern varies based on specific dataset characteristics. Subsequently, we investigate the importance of example selection via In-Context Learning (ICL) and demonstrate the value of nearest neighbor example selection using pre-trained biomedical text encoders when performing biomedical NER. Our study reveals the importance of meticulously designed prompts in the biomedical. Strategic selection of in-context examples yields a marked improvement, offering $\sim 15 - 20\%$ increase in F1 score across all benchmark datasets for biomedical few-shot NER. A key question that arises in the deployment of LLMs concerns the comparative advantage of closed-source LLMs versus open-source ones. In our third study, we shed light on this question by presenting an assessment of performance and cost across various experiments. Furthermore, we explore the integration of external medical knowledge to refine LLM capabilities [GXG23, ZSC24]. Leveraging the insights gained from these techniques, we present a novel data augmentation method incorporating a medical knowledge base, e.g., UMLS [Bod04], which substantially improves zero-shot biomedical NER. Leveraging a medical knowledge base, our proposed method, DiRAG, inspired by Retrieval-Augmented Generation (RAG), can boost the zero-shot F1 score of LLMs for biomedical NER.

5.1 Background and Preliminaries

5.1.1 Prompt engineering

Prompt tuning [WFH23, LAC21, DHZ21] as its own research field shows that skillfully crafted prompts can significantly enhance LLM understanding for complex tasks [LBM21, KHM23, WP21]. Researchers have explored different prompt formats for IE tasks with LLMs [WZZ23a, GMW22, WSL23] including more work around knowledge insertion for prompt augmentation [SBT24, CJC23] Another type of prompting is ICL [BMR20], where LLMs use a limited set of "input-output" pairs within the prompt along with a query input as demonstrations of what the task output should be. In this realm, **(author?)** [LSZ21, MLH22, GWG23] demonstrated that choosing tar-

geted in-context examples over random sampling leads to more accurate model responses.

5.1.2 Named Entity Recognition

GPT-NER [WSL23] was one of the first methods to incorporate a unique symbol to transform the sequence tagging task into text generation via ICL with GPT-3 [BMR20], achieving performance on par with fully supervised baselines. Following this work, **(author?)** [GMW22, MBH21] showed that LLMs are not skilled few-shot learners in the biomedical domain. However, recent advancements, such as GPT-4, have increased LLM performance on many tasks [TJY24, HLZ24, NKM23] including in the biomedical domain [HCD24]. In the direction of knowledge distillation from LLMs [WZZ23a, GZU23], **(author?)** [ZZG23] presented UniNER, a targeted distillation technique coupled with instruction tuning to develop an efficient open-domain NER model. Our research draws from these works and explores the capabilities of LLMs for biomedical NER, employing prompt design, strategic ICL example selection, and data augmentation via an external knowledge base to enhance performance.

5.1.3 Problem definition

Assume data samples are represented as (X, Y) and the goal is to develop a model, denoted as $f : (X \times T) \rightarrow Y$, where X signifies the input set, T represents a predetermined set of entity types, and Y denotes the set of entity types. The task is to predict the entity type of each input word among the set T. We followed the standard practice of using the F1 score for evaluation purposes in both mention/token-level analyses.

5.1.4 Datasets

We used three biomedical NER datasets with different entity types: I2B2 [USS11] which includes test, treatment, and problem entities, NCBI-disease [DLL14] consisting of the disease entity, and BC2GM [STA08] containing the gene entity.

I2B2: I2B2 is a collection of annotated clinical records that are used primarily for Clinical NER.The task involves identifying clinical terms such as medical problems, treatments, and tests from patient records. The dataset typically includes a large number of annotated clinical narratives that are de-identified to protect patient confidentiality. This makes it a rich resource for training and testing NER models.

NCBI-disease: This dataset is specifically curated for disease name recognition and normalization in biomedical texts. It comprises abstracts from PubMed annotated for disease mentions and linked to the NCBI disease database. The corpus is relatively smaller compared to i2b2 but is densely annotated, providing high-quality, fine-grained annotations of disease entities, which are crucial for models aimed at medical literature.

BC2GM: This dataset focuses on the recognition of gene and gene product mentions in PubMed abstracts that is a suitable dataset for biological NER. The BC2GM dataset is extensively annotated to include a wide range of gene and gene product mentions, reflecting the complex and varied ways these entities are referred to in scientific literature.

5.2 Influence of Input-Output Format

Recent studies demonstrated the importance of prompt engineering for various tasks [WZO23, GLC23, NLZ23]. We studied the influence of input-output format by adapting TANL [PAK21] and DICE [MTW22] for biomedical NER. In TANL, the task is framed as a translation task which involves augmenting the text by tagging entity types for each word directly within the text. The method is exemplified in Fig 5.1, showcasing how the text incorporates entity types.

Then, the generated output is decoded into the BIO format [RM99] for the assessment. In the refined DICE format, the input-output format involves adding a description for each entity type in a template following DEGREE [HHB21]. Given an input text and corresponding labels, the desired output should be the input followed by the phrase "entity type is <entity_type>. <en-

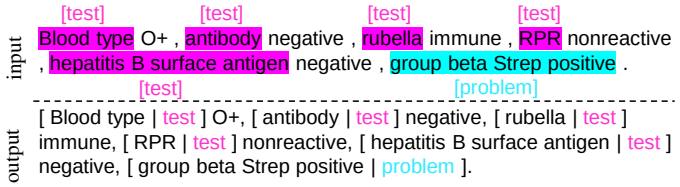


Figure 5.1: TANL input/output format for NER task.

tity_description>. entity is <entity>" for each class label, e.g., *test*, *treatment*, and *problem* in the I2B2 dataset. Then, we expect the model to output the same template filling out the <entity> with the corresponding entities in the given text as demonstrated in Fig 5.2. For the entity types with no matched entities in the sentence, the output returns <entity> token in the output. Examples for the NCBI-disease and BC2GM datasets are presented in Appendix **??**.

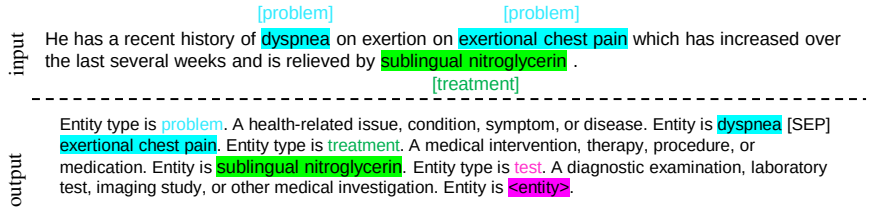


Figure 5.2: DICE input/output format for NER task.

Our experiments in Table 5.1 reveal that neither format consistently outperforms the other; rather, the effectiveness of each format varies depending on the complexity of the dataset and model size. To maintain consistency in the rest of our experiments, we opted for the TANL format, in which the input-output relationship exhibits a more straightforward pattern.

5.3 In-Context Examples Selection: A Key to Improving ICL Outcomes

In-context examples can be randomly chosen from the training set; however, researchers have demonstrated that the performance of ICL depends on the order and similarity of ICL examples to the test samples [LSZ21, MLH22, GWG23]. **(author?)** [LSZ21] presented Knn-Augmented in-conText Example selection (KATE). KATE identifies in-context examples selectively using nearest

Model	input-output format	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
GPT-3.5-turbo	DICE	41.2 /50.0	45.3 /**62.0**	**43.3 /55.6**
	TANL	**52.9/59.7**	**46.5**/51.3	39.1/50.8
GPT-4	DICE	58.8/70.1	**68.1/77.8**	**57.1**/67.9
	TANL	**61.9/73.5**	67.5/70.0	56.4/**69.6**

Table 5.1: TANL vs. DICE format with GPT-3.5-turbo/GPT-4 . The superiority of any single format varies with the complexity of the dataset and model size.

neighbor search on example embeddings, leading to better performance than random example selection. We tested KATE on TANL formatted examples with 16-shot ICL using four different LM encoders (w/o fine-tuning) to produce example embeddings. We used MPNET [STQ20] for its popularity and performance on sentence embedding benchmarks [RG19] , SimCSE [GYC21] for its documented performance as an alternative to standard sentence transformers, and BioClinicalBERT [AMB19] and BioClinicalRoBERTa [GMS20] for their dominance on clinical data tasks [LHM23].

Our results summarized in Table 5.2 show that strategic in-context example selection via KATE outperforms random selection. BioClinicalRoBERTa achieved the best results among all example encoders tested. The strong performance of BioClinicalBERT and BioClinicalRoBERTa underscores the importance of using LM encoders pretrained on biomedical text when applying KATE for biomedical NER.

5.4 In-Context Learning or Fine-Tuning?

Within the scope of LLMs for biomedical applications, an essential question is whether to prompt a closed-source LLM via ICL or fine-tune an open-source one. Comparing two different LLMs employing divergent strategies is not straightforward. To provide some insight into this dilemma, we examined two key factors, performance and cost, for biomedical NER, and present a detailed

Model	KATE vs RS	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
GPT-3.5-turbo (ICL)	RS	52.9/59.7	46.6/51.3	39.1/50.8
	BioClinicalRoBERTa	66.1/77.4	**68.0**/77.7	**61.6/72.5**
	BioClinicalBERT	**67.0/78.9**	67.6/**78.8**	60.9/72.0
	MPNET	65.3/76.7	63.7/76.7	59.1/70.0
	SimCSE	65.2/76.1	61.6/76.1	57.8/68.8
	[HCD24]	49.3/ -	-	-
GPT4 (ICL)	RS	67.7/73.5	62.6/70.0	59.2/69.6
	BioClinicalRoBERTa	81.2/**88.4**	**79.3/88.3**	**72.4/80.7**
	BioClinicalBERT	**81.7**/88.1	**79.3**/88.0	71.9/79.4
	MPNET	80.7/87.5	79.8/87.4	71.1/80.2
	SimCSE	79.6/86.6	77.3/86.5	69.9/77.9
	[HCD24]	59.3/ -	-	-
BioBERT	fully supervised	- /87.3	- /**89.1**	- /83.8
BioClinicBERT	fully supervised	- /87.7	- /89.0	- /81.7
BioClinicRoBERTa	fully supervised	- /**89.7**	- /89.0	- /**87.0**

Table 5.2: 16-shot ICL for Random example selection (RS) vs. KATE method Vs MLMs with Mention/Token-level (M/T) analysis. KATE significantly outperforms random sampling in all settings, and LMs pre-trained on biomedical text outperform general domain encoders.

analysis under various experiment settings. This comparison offers valuable perspective into the right strategy given the task and dataset attributes. For fine-tuning, we used LoRA [HSW21]. Details can be found in Appendix **??**. The cost of fine-tuning comes from training an LLM on a large labeled dataset while the cost of ICL mainly comes from calling an API for each input query. For 16-shot ICL experiments, we calculated the cost based on the number of processed and generated tokens considering the average text size based on current LLM API pricing.[1] The estimated cost for the entire test set of each benchmark dataset considering the input text, prompt, and generated text size using the TANL format is summarized in Table 5.3. Referring to the OpenAI API for fine-tuning pricing, we also estimated the cost for fine-tuning LLama2-7B which is summarized in Table 5.3. Interestingly, for the I2B2 dataset, GPT-3.5-turbo with a much cheaper cost outperforms fine-tuning Llama2-7B.

[1] `https://openai.com/pricing`

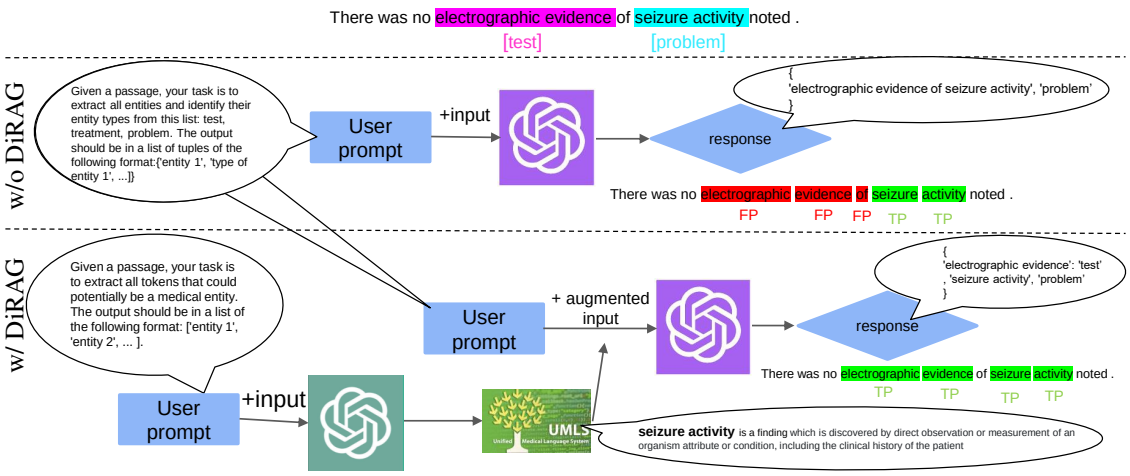


Figure 5.3: An overview of Dictionary-Infused RAG

	Model	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
Performance	GPT-3.5-turbo w/ KATE	67.0/78.9	68.0/78.8	61.6/72.5
	GPT4 w/ KATE	**81.7/88.4**	79.3/88.3	**72.4/80.7**
	Llama2-7B	61.2/76.2	**80.4/91.3**	68.1/75.1
Cost (T+I)	GPT3.5-turbo w/ KATE	($0.35)	($0.11)	($1.34)
	GPT4 w/ KATE	($10.42)	($3.12)	($40.13)
	Llama2-7B	($47.85+$7.4)	($23.5+$1.2)	($69.7+$12.9)

Table 5.3: Analysis of ICL vs fine-tuning LLMs: assessing performance and cost (Training + Inference) implications. Fine-tuning LLama2 exhibits superior outcomes on NCBI-disease, whereas GPT-4, enhanced by KATE using a biomedical encoder, achieves more favorable results on both the I2B2 and BC2GM datasets.

5.4.1 PEFT setting of Llama for fine-tuning

We fine-tuned Llama2-7B on the entire training set of each dataset for three epochs and maintained a batch size of 16, learning rate of 2e-4, and cap the maximum sequence length at 512, truncating any sequences that exceeded this limit. The LoRA dropout rate is adjusted to 0.1, and the LoRA α and rank parameters are also set at 16 and 32 respectively. The training was done on 4 NVIDIA Tesla V100 GPUs for approximately 24, 12, and 63 hours for I2B2, NCBI-disease, and BC2GM respectively.

5.5 Dictionary-Infused RAG - DiRAG

5.5.1 Few-shot Vs. Zero-shot

We introduced both few-shot and zero-shot settings to comprehensively evaluate the versatility and generalization capabilities of our study across different levels of data availability. While it's true that the performance in the zero-shot setting is generally lower compared to the few-shot setting, this approach offers valuable insights into the model's behavior when no training examples are provided. The zero-shot setting, leveraging techniques like Retrieval-Augmented Generation (RAG), demonstrates the model's potential to utilize pre-existing knowledge embedded in its parameters and external sources effectively. This is particularly important for scenarios where labeled data is scarce or unavailable, making zero-shot learning a critical area of study to ensure broader applicability of the model in real-world applications. Moreover, the inclusion of both methodologies allows us to highlight the performance trade-offs and strengths of the model under different instructional paradigms, contributing to a more robust and nuanced understanding of its capabilities.

5.5.2 DiRAG

Retrieval-Augmented Generation (RAG) [LPP20] is a technique to enhance the capabilities of LLMs by integrating external information or knowledge into the generation process. This method involves retrieving relevant documents from a large corpus and providing this external knowledge in the input context to improve the quality and relevance of the generated text. Inspired by RAG, we developed a new method, DiRAG, to utilize UMLS as an external resource to augment the input data for the biomedical NER task. The process with detailed prompts is visualized in Fig 5.3, while an expanded view of the UMLS component is depicted in Fig 5.4. Unlike traditional RAG techniques that rely on embedding similarities to retrieve relevant documents, our approach initially employs the LLM to tackle a more straightforward task: identifying all words that could

Model	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
UniversalNER [ZZG23]	40.4/ -	60.4/ -	47.2/ -
[RNC23] w/ GPT-3.5	-	33.4 / -	32.0 / -
[HCD24] w/ GPT-3.5-turbo	39.3/ -	-	-
[HCD24] w/ GPT-4	52.6/ -	-	-
GPT-3.5-turbo w/o DiRAG	41.9 /54.7	38.2 / 49.4	38.6 / 28.7
GPT-3.5-turbo w/ DiRAG	43.0 / 55.7	44.7 / 50.0	30.45 / 22.5
GPT-4 w/o DiRAG	46.3 / 59.1	55.7 /60.5	**52.1 / 58.4**
GPT-4 w/ DiRAG	**53.1 /62.8**	**61.0 /66.2**	51.1 / 55.0

Table 5.4: Zero-shot NER with GPT models w/ and w/o DiRAG vs. SOTA. DiRAG improved zero-shot NER significantly for I2B2 and NCBI-disease datasets for both GPT models. Results with confidence intervals are in the appendix.

potentially qualify as medical named entities. Then, we look up each selected word in an external knowledge base, e.g., UMLS to augment the input data with useful information such as term definition. Then, we call the LLM with augmented input text. The process is visualized in Fig 5.4. We tested the approach on zero-shot NER and compared it with SOTA in Table 5.4. Our proposed approach enhanced the performance of both GPT versions on the I2B2 and NCBI-disease datasets significantly. DiRAG with GPT-4 achieved SOTA for zero-shot NER. Our approach proved ineffective for the BC2GM dataset due to the nature of the UMLS knowledge base which is predominantly tailored to medical terminology rather than biogenetics. We expect our approach to outperform GPT-4 on BC2GM with a more relevant knowledge base.

5.5.3 UMLS detail

In Fig 5.4, we visualize the process by which potential words suggested by the LLM are searched within the UMLS and demonstrate how the input is augmented to enhance zero-shot prompting in LLMs.

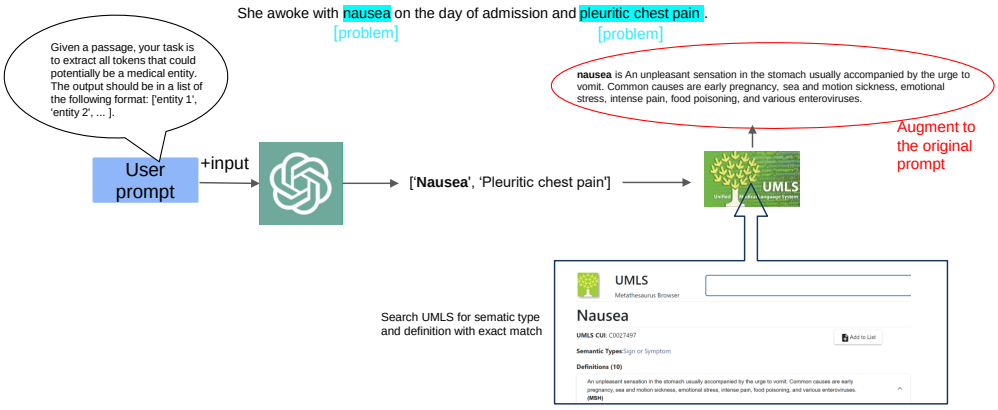


Figure 5.4: UMLS search. The GPT model is prompted for a simpler task of identifying all words that could potentially be a named entity. Then, the retrieved information from UMLS will augment the original input text for recalling the LLM

5.6 Conclusion

Our study is centered on the creation of a pre-training dataset for domain-specific medical vision language models, utilizing a VL dataset sourced from PubMed and a pipeline for VL modeling. Our approach has been evaluated through intrinsic task assessments, demonstrating its effectiveness in constructing data-efficient VL models for the medical field. Our model has surpassed the performance of standard VL training methods, further validating the efficacy of our pipeline. Our proposed pipeline helps build powerful multimodal AI models which could be used for various tasks like diagnosis, drug discovery, and information extraction.

CHAPTER 6

Conclusion

In the field of vision-language models, especially in low-resource settings such as the biomedical domain, this dissertation has tackled unique challenges and proposed innovative methodologies to enhance model performance and applicability. Recognizing the scarcity of labeled data and the specific needs of medical contexts, our research has been directed toward creating reliable models capable of functioning with limited inputs while maintaining high accuracy and reliability.

In Chapter 2, we investigate the application of VL models to the diagnosis of diseases from chest X-rays. Traditional VL models often underperform in the medical domain due to significant differences between general and medical imagery and text. In response, we developed BERTHop, a model that integrates enhanced visual representations with robust language understanding capabilities, demonstrating a notable improvement in diagnosis accuracy as evidenced by its performance on the OpenI benchmark dataset.

In Chapter 2, we extends the concept of in-context learning from large language models (LLMs) to vision-language models. This study addresses the challenge of enabling VL models to adapt through few-shot learning strategies, traditionally seen in LLMs. By transferring in-context learning abilities from language-only to multimodal models, we significantly enhanced the flexibility and reduced the model size necessary for effective performance on tasks like Visual Question Answering.

In Chapter 3, we focus on the pre-training of vision-language models specifically for identify-

ing brain abnormalities. By creating a tailored dataset from public medical texts and imagery, we pre-trained a model that not only understands medical visuals and texts in conjunction but also addresses unique challenges such as the correlation of subfigures to subcaptions. This approach not only improved model performance but also provided new insights into the automated processing of complex medical data.

In Chapter 4, we explore the application of LLMs to Named Entity Recognition (NER) in the biomedical field. Despite the complexity and specialized knowledge required in this domain, our strategically designed prompts and incorporation of external biomedical knowledge through DiRAG significantly enhanced the performance of LLMs on biomedical NER tasks, pushing the boundaries of what is possible with zero/few-shot learning in this critical area.

Throughout these chapters, we've seen that while advanced machine learning models offer substantial benefits to biomedical applications, challenges such as data scarcity and the need for domain-specific adaptations remain prevalent. This research contributes valuable methodologies and insights to the fields of machine learning and biomedical informatics, proving that with thoughtful adaptation, cutting-edge AI technologies can significantly enhance their performance across different applications from classification tasks such as disease diagnosis or biomedical NER to generation tasks such as question answering.

References

[AA19] Imane Allaouzi and Mohamed Ben Ahmed. "A novel approach for multi-label chest X-ray classification of common thorax diseases." *IEEE Access*, **7**:64279–64288, 2019.

[AAL15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. "Vqa: Visual question answering." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

[ADL22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. "Flamingo: a visual language model for few-shot learning." *arXiv preprint arXiv:2204.14198*, 2022.

[AHB18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[AM18] Rahib H Abiyev and Mohammad Khaleel Sallam Maaitah. "Deep convolutional neural networks for chest diseases detection." *Journal of healthcare engineering*, **2018**, 2018.

[AMB19] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. "Publicly Available Clinical BERT Embeddings." In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019.

[AU19] Enes Ayan and Halil Murat Ünver. "Diagnosis of pneumonia from chest x-ray images

using deep learning." In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1–5. Ieee, 2019.

[BI21] Lorenzo Brigato and Luca Iocchi. "A close look at deep learning with small data." In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2490–2497. IEEE, 2021.

[BLB13] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. "API design for machine learning software: experiences from the scikit-learn project." In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[BMR20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165*, 2020.

[Bod04] Olivier Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic acids research*, **32**(suppl_1):D267–D270, 2004.

[CCL20] Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. "Visual question answering on 360deg images." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1607–1616, 2020.

[CHW23] Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. "MedBLIP: Bootstrapping Language-Image Pre-training from 3D Medical Images and Texts." *arXiv preprint arXiv:2305.10799*, 2023.

[CJC23] Xiusi Chen, Jyun-Yu Jiang, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Wei Wang. "MinPrompt: Graph-based Minimal Prompt Data Augmentation for Few-shot Question Answering." *arXiv preprint arXiv:2310.05007*, 2023.

[CLY20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. "Uniter: Universal image-text representation learning." In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

[CRG21] Hong-Shuo Chen, Mozhdeh Rouhsedaghat, Hamza Ghani, Shuowen Hu, Suya You, and C. C. Jay Kuo. "DefakeHop: A Light-Weight High-Performance Deepfake Detector.", 2021.

[CRY20] Yueru Chen, Mozhdeh Rouhsedaghat, Suya You, Raghuveer Rao, and C-C Jay Kuo. "Pixelhop++: A small successive-subspace-learning-based (ssl-based) model for image classification." In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3294–3298. IEEE, 2020.

[DCL18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.

[DHZ21] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. "Openprompt: An open-source framework for prompt-learning." *arXiv preprint arXiv:2111.01998*, 2021.

[DLD22] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. "A survey for in-context learning." *arXiv preprint arXiv:2301.00234*, 2022.

[DLL14] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. "NCBI disease corpus: a resource for disease name recognition and concept normalization." *Journal of biomedical informatics*, **47**:1–10, 2014.

[DSD23] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. "Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers." In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019, 2023.

[DXG22] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. "An Empirical Study of Training End-to-End Vision-and-Language Transformers." In *CVPR*, 2022.

[EBW21] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. "MAGMA–Multimodal Augmentation of Generative Models through Adapter-based Finetuning." *arXiv preprint arXiv:2112.05253*, 2021.

[EEA23] Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. "GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts." *Natural Language Processing Journal*, **5**:100032, 2023.

[EMD23] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. "PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?" In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1151–1163, 2023.

[EP04] Theodoros Evgeniou and Massimiliano Pontil. "Regularized multi–task learning." In

Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 109–117, 2004.

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

[GGM21] Lisa Grossman Liu, Raymond H Grossman, Elliot G Mitchell, Chunhua Weng, Karthik Natarajan, George Hripcsak, and David K Vawdrey. "A deep database of medical abbreviations and acronyms for natural language processing." *Scientific Data*, **8**(1):149, 2021.

[GLC23] Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. "Leveraging a medical knowledge graph into large language models for diagnosis prediction." *arXiv preprint arXiv:2308.14321*, 2023.

[GMS20] Suchin Gururangan, Ana Marasovi, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In *Proceedings of ACL*, 2020.

[GMW22] Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. "Thinking about gpt-3 in-context learning for biomedical ie? think again." *arXiv preprint arXiv:2203.08410*, 2022.

[GS08] Maryellen L Giger and Kenji Suzuki. "Computer-aided diagnosis." In *Biomedical information technology*, pp. 359–XXII. Elsevier, 2008.

[GWG23] Shuzheng Gao, Xin-Cheng Wen, Cuiyun Gao, Wenxuan Wang, Hongyu Zhang, and Michael R Lyu. "What makes good in-context demonstrations for code intelligence

tasks with llms?" In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 761–773. IEEE, 2023.

[GXG23] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997*, 2023.

[GYC21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." *arXiv preprint arXiv:2104.08821*, 2021.

[GZU23] Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, et al. "Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events." *arXiv preprint arXiv:2307.06439*, 2023.

[HCD24] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. "Improving large language models for clinical named entity recognition via prompt engineering." *Journal of the American Medical Informatics Association*, p. ocad259, 2024.

[HHB21] I Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, et al. "DEGREE: A data-efficient generation-based event extraction model." *arXiv preprint arXiv:2108.12724*, 2021.

[HLZ24] Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. "Zero-shot information extraction from radiological reports using ChatGPT." *International Journal of Medical Informatics*, **183**:105321, 2024.

[HM19] Drew A Hudson and Christopher D Manning. "GQA: A New Dataset for Real-World

Visual Reasoning and Compositional Question Answering." *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[HSW21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685*, 2021.

[HYK18] Yoseob Han, Jaejun Yoo, Hak Hee Kim, Hee Jung Shin, Kyunghyun Sung, and Jong Chul Ye. "Deep learning with domain adaptation for accelerated projection-reconstruction MR." *Magnetic resonance in medicine*, **80**(3):1189–1205, 2018.

[JPG19] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs." *arXiv preprint arXiv:1901.07042*, 2019.

[KHM23] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. "Challenges and applications of large language models." *arXiv preprint arXiv:2307.10169*, 2023.

[KKS23] Amita Kumari, Anita Kumari, Amita Singh, Sanjeet K Singh, Ayesha Juhi, Anup Kumar D Dhanvijay, Mohammed Jaffer Pinjar, Himel Mondal, and Anoop Kumar Dhanvijay. "Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing." *Cureus*, **15**(8), 2023.

[KM23] Mert Karabacak and Konstantinos Margetis. "Embracing Large Language Models for Medical Applications: Opportunities and Challenges." *Cureus*, **15**(5), 2023.

[KSF23] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. "Grounding language models to

images for multimodal inputs and outputs." In *International Conference on Machine Learning*, pp. 17283–17300. PMLR, 2023.

[KSK21] Wonjae Kim, Bokyung Son, and Ildoo Kim. "ViLT: Vision-and-language transformer without convolution or region supervision." In *ICML*, 2021.

[KZG17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *International journal of computer vision*, **123**(1):32–73, 2017.

[LAC21] Brian Lester, Rami Al-Rfou, and Noah Constant. "The power of scale for parameter-efficient prompt tuning." *arXiv preprint arXiv:2104.08691*, 2021.

[LBM21] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity." *arXiv preprint arXiv:2104.08786*, 2021.

[LBP19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *arXiv preprint arXiv:1908.02265*, 2019.

[LGR20] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. "12-in-1: Multi-task vision and language representation learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10437–10446, 2020.

[LHM23] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily

Alsentzer. "Do We Still Need Clinical Language Models?" *arXiv preprint arXiv:2302.08091*, 2023.

[LLW24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning." *Advances in neural information processing systems*, **36**, 2024.

[LLX22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

[LMB14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context." In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

[LPP20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems*, **33**:9459–9474, 2020.

[LSG21] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation." In *NeurIPS*, 2021.

[LSZ21] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. "What Makes Good In-Context Examples for GPT-3?" *arXiv preprint arXiv:2101.06804*, 2021.

[LWL20] Yikuan Li, Hanyin Wang, and Yuan Luo. "A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and

reports." In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1999–2004. IEEE, 2020.

[LXY21] Xiaofeng Liu, Fangxu Xing, Chao Yang, C-C Jay Kuo, Suma Babu, Georges El Fakhri, Thomas Jenkins, and Jonghye Woo. "VoxelHop: Successive Subspace Learning for ALS Disease Classification Using Structural MRI." *arXiv preprint arXiv:2101.05131*, 2021.

[LYL20] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

[LYW21] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. "Contrastive Attention for Automatic Chest X-ray Report Generation." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 269–280, 2021.

[LYY19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "Visualbert: A simple and performant baseline for vision and language." 2019.

[LZZ23] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. "Pmc-clip: Contrastive language-image pre-training using biomedical documents." *arXiv preprint arXiv:2303.07240*, 2023.

[MBH21] Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. "Gpt-3 models are poor few-shot learners in the biomedical domain." *arXiv preprint arXiv:2109.02555*, 2021.

[MLH22] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Rethinking the role of demonstrations: What makes in-context learning work?" *arXiv preprint arXiv:2202.12837*, 2022.

[MLZ22] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. "MetaICL: Learning to Learn In Context." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2791–2809, Seattle, United States, July 2022. Association for Computational Linguistics.

[MRF19] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. "Ok-vqa: A visual question answering benchmark requiring external knowledge." In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

[MRL22] Masoud Monajatipoor, Mozhdeh Rouhsedaghat, Liunian Harold Li, C-C Jay Kuo, Aichi Chien, and Kai-Wei Chang. "Berthop: An effective vision-and-language model for chest x-ray disease diagnosis." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 725–734. Springer, 2022.

[MTW22] Mingyu Derek Ma, Alexander K Taylor, Wei Wang, and Nanyun Peng. "DICE: data-efficient clinical event extraction with generative models." *arXiv preprint arXiv:2208.07989*, 2022.

[NKM23] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. "Capabilities of gpt-4 on medical challenge problems." *arXiv preprint arXiv:2303.13375*, 2023.

[NLZ23] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi,

Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. "Can generalist foundation models outcompete special-purpose tuning? case study in medicine." *arXiv preprint arXiv:2311.16452*, 2023.

[PAB10] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. "Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization." *Neurology*, **74**(3):201–209, 2010.

[PAK21] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. "Structured prediction as translation between augmented natural languages." *arXiv preprint arXiv:2101.05779*, 2021.

[PYL19] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets." In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 58–65, 2019.

[RG19] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[RHG16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: towards real-time object detection with region proposal networks." *IEEE transactions on pattern analysis and machine intelligence*, **39**(6):1137–1149, 2016.

[RIZ17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony

Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225*, 2017.

[RKH21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision." In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

[RM99] Lance A Ramshaw and Mitchell P Marcus. "Text chunking using transformation-based learning." In *Natural language processing using very large corpora*, pp. 157–176. Springer, 1999.

[RMA21] Mozhdeh Rouhsedaghat, Masoud Monajatipoor, Zohreh Azizi, and C-C Jay Kuo. "Successive Subspace Learning: An Overview." *arXiv preprint arXiv:2103.00121*, 2021.

[RNC23] Omid Rohanian, Mohammadmahdi Nouriborji, and David A Clifton. "Exploring the Effectiveness of Instruction Tuning in Biomedical Language Processing." *arXiv preprint arXiv:2401.00579*, 2023.

[RNS18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. "Improving language understanding by generative pre-training." 2018.

[Rud17] Sebastian Ruder. "An overview of multi-task learning in deep neural networks." *arXiv preprint arXiv:1706.05098*, 2017.

[RWC19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI Blog*, **1**(8):9, 2019.

[RWG20] Mozhdeh Rouhsedaghat, Yifan Wang, Xiou Ge, Shuowen Hu, Suya You, and C-C Jay Kuo. "Facehop: A light-weight low-resolution face gender classification method." *arXiv preprint arXiv:2007.09510*, 2020.

[RWH20] Mozhdeh Rouhsedaghat, Yifan Wang, Shuowen Hu, Suya You, and C-C Jay Kuo. "Low-Resolution Face Recognition In Resource-Constrained Environments." *arXiv preprint arXiv:2011.11674*, 2020.

[SBT24] Minju Seo, Jinheon Baek, James Thorne, and Sung Ju Hwang. "Retrieval-Augmented Data Augmentation for Low-Resource Domain Tasks." *arXiv preprint arXiv:2402.13482*, 2024.

[SDG18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. "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)*, pp. 2556–2565, 2018.

[SRG16] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." *IEEE transactions on medical imaging*, **35**(5):1285–1298, 2016.

[STA08] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. "Overview of BioCreative II gene mention recognition." *Genome biology*, **9**:1–19, 2008.

[STQ20] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. "Mpnet: Masked and

permuted pre-training for language understanding." *Advances in Neural Information Processing Systems*, **33**:16857–16867, 2020.

[SZC19] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. "VL-BERT: Pre-training of Generic Visual-Linguistic Representations." In *ICLR*, 2019.

[TB19a] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers." 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)*, pp. 5100–5111, 2019.

[TB19b] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers." In *EMNLP*, 2019.

[TC17] Satoshi Tsutsui and David J Crandall. "A data driven approach for compound figure separation using convolutional neural networks." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pp. 533–540. IEEE, 2017.

[TJY24] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. "Opportunities and challenges for ChatGPT and large language models in biomedicine and health." *Briefings in Bioinformatics*, **25**(1):bbad493, 2024.

[TMC21] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. "Multimodal few-shot learning with frozen language models." *Advances in Neural Information Processing Systems*, **34**:200–212, 2021.

[TSP21] Wu Te-Lin, Shikhar Singh, Sayan Paul, Gully Burns, and Nanyun Peng. "MELINDA:

A Multimodal Dataset for Biomedical Experiment Method Classification." In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

[USS11] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." *Journal of the American Medical Informatics Association*, **18**(5):552–556, 2011.

[VD02] Ricardo Vilalta and Youssef Drissi. "A perspective view and survey of meta-learning." *Artificial intelligence review*, **18**(2):77–95, 2002.

[VMV23] Raju Vaishya, Anoop Misra, and Abhishek Vaish. "ChatGPT: Is this version good for healthcare and research?" *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, **17**(4):102744, 2023.

[WDS19] Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. "Huggingfaces transformers: State-of-the-art natural language processing." *arXiv*, 2019.

[WFH23] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. "A prompt pattern catalog to enhance prompt engineering with chatgpt." *arXiv preprint arXiv:2302.11382*, 2023.

[WP21] Albert Webson and Ellie Pavlick. "Do prompt-based models really understand the meaning of their prompts?" *arXiv preprint arXiv:2109.01247*, 2021.

[WPL17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax dis-

eases." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

[WPL18] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058, 2018.

[WSL23] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. "Gpt-ner: Named entity recognition via large language models." *arXiv preprint arXiv:2304.10428*, 2023.

[WYY21] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. "Simvlm: Simple visual language model pretraining with weak supervision." *arXiv preprint arXiv:2108.10904*, 2021.

[WZO23] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. "Chatcad: Interactive computer-aided diagnosis on medical image using large language models." *arXiv preprint arXiv:2302.07257*, 2023.

[WZZ23a] Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. "InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction." *arXiv preprint arXiv:2304.08085*, 2023.

[WZZ23b] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. "MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21372–21383, 2023.

[YDY19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "Xlnet: Generalized autoregressive pretraining for language understanding." In *Advances in neural information processing systems*, pp. 5754–5764, 2019.

[YGW22] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. "An empirical study of gpt-3 for few-shot knowledge-based vqa." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3081–3089, 2022.

[ZAI22] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. "Socratic models: Composing zero-shot multimodal reasoning with language." *arXiv preprint arXiv:2204.00598*, 2022.

[ZCS19] Zizhao Zhang, Pingjun Chen, Xiaoshuang Shi, and Lin Yang. "Text-guided neural network training for image recognition in natural scenes and medicine." *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[ZHY22] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. "mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 107–117. Springer, 2022.

[ZLH21] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. "VinVL: Revisiting visual representations in vision-language models." In *CVPR*, 2021.

[ZPZck] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng

Gao. “Unified vision-language pre-training for image captioning and vqa.” In *AAAI*, 2020.

[ZSC24] Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. “AlmanacRetrieval-Augmented Language Models for Clinical Medicine.” *NEJM AI*, **1**(2):AIoa2300068, 2024.

[ZZG23] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. “Universalner: Targeted distillation from large language models for open named entity recognition.” *arXiv preprint arXiv:2308.03279*, 2023.