Foreground Enhanced Network for Weakly Supervised Temporal Language Grounding

Hongzhou Wu[a], Xuechen Zhao[a], Yifan Lyu[b], Xiang Zhang[a][1]
[a] College of Computer, National University of Defense and Technology, Changsha, China
[b]Institute of Software, Chinese academy of sciences, Beijinng, China
[a]{whz, zhaoxuechen, zhangxiang08}@nudt.edu.cn, [b]yifan2018@iscas.ac.cn

Abstract

Temporal language grounding (TLG) aims to localize query-related events in videos, which explores how to cognize relationships of video content with language descriptions. According to selective visual attention mechanism in cognitive science, people's cognition and understanding of what happens often rely on dynamic foreground information in the video. Nonetheless, background usually predominates the scenes so that query-related visual features and irrelevant ones are confused. Thus, we propose a Foreground Enhanced Network (FEN) to diminish the background effect from two aspects. FEN at first in spatial dimension explicitly models the evolving foreground in video features by removing relatively unchanged background content. Besides, we propose a progressive contrastive sample generation module to gradually learn the differences between the predicted proposal and its elongated proposals that include the former as a portion, thereby distinguishing similar neighborhood frames. Experiments on two common-used datasets show the efficacy of our model.

Keywords: selective visual attention mechanism, temporal language grounding, multi modal, visual perception

Introduction

With the aim of enabling machines to cognize and understand video content and language like the humans, temporal language grounding (TLG) is a multi-modal task that focuses on predicting the start and end time of the specific video segment in an untrimmed video that matches a given natural language query. This needs a model to perceive visual information as well as to understand natural language. In this field, several approaches (G. J., C., Z., & R., 2017; A. et al., 2017; G. S., A., Z., & A., 2019; W. J., L., & W., 2019; L. Z., Z., Z., Z., & D., 2020; K., D., & M., 2021; X. et al., 2022) focus on the fully-supervised setting, where the start time and end time of the matching video segment are annotated manually. Annotating the temporal boundaries to match video segments is expensive and time-consuming. Alternatively, some works start to consider weakly supervised setting in TLG, where only video-sentence pairs are used for training without any temporal annotations. For lack of temporal supervision, most existing respective approaches (C., S., & K., 2019; Z. Z., Z., Z., J., & X., 2020; G. M., S., R., & C., 2019; T. R., H., K., & A., 2021; W., T., Y., & F., 2021; W. Y., J., W., & H., 2021) treat this task as a multiple instance learning (MIL) problem. They consider each video as a bag of segments, where at least one segment within the bag may match the query. The similarity between video segments and the query then acts as a measure signal to identify the matched segment. Recently, another cousin is born by using video segments to reconstruct the masked queries. The typical works like (L. Z., Z., Z., Q., & H., 2020; Z. M., Y., Q., & Y., 2022) compare the reconstructed queries with the original queries to identify the matched segments. However, they mostly ignore the role of foreground information in temporal language grounding.

Figure 1: (a) The room (background) is stable in all video frames while the human (foreground) burdens appearance variations. (b) The relative activation scores of the background and foreground have the foresaid same analogy. So background subtraction for foreground enhancement is feasible.

According to the selective visual attention mechanism in cognitive science, humans will selectively attend certain information and cope with it in complex visual scenes. When people understand a video and recognize different events, the human eye often has a higher sensitivity to the dynamic foreground information, such as human actions, object interactions and changes. By contrast, the background contains relatively static information with little variation over evolving time, which is improper for action recognition. Thus, the foreground is often closely related to queries and thus plays a crucial role in TGL task. Recall that video data can be decomposed into evolving foreground part and the relatively un-

[1]Corresponding Author

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

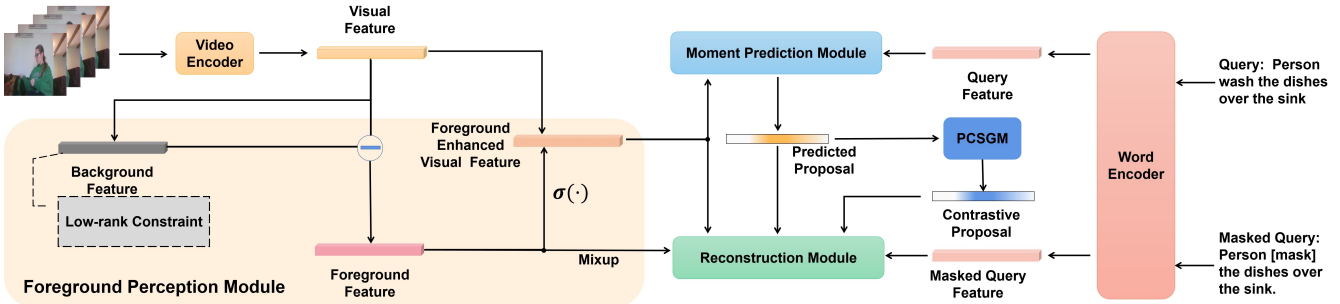

Figure 2: The overall architecture of our approach, including four components: a foreground perception module, a moment prediction module, a progressive contrastive sample generation module (PCSG) and a reconstruction module.

changed background part (See Figure 1). That is, the background is low-rank. This property helps us to remove the background from visual features because the background always predominates visual features in the whole representation and is almost harmful.

Inspired by the human selective visual attention mechanism, we develop a Foreground Enhanced Network (FEN) to lessen the negative impact of the background on visual features in both spatial and temporal dimensions. We at first explicitly induce foreground information of events by removing the background features with the nuclear norm for the low-rankness. Then we employ the mixup-based reconstruction strategy on the separated foreground features to achieve better cross-modal alignment between the foreground features and queries. Since the clear foreground features are more easily aligned with the query, foreground enhanced video features can greatly improve the cross-modal interaction, which benefits to the following reconstruction goal. We also note that RTBPN (Z. Z. et al., 2020) models semantic enhancement and semantic suppression in the temporal dimension. However, they lack explicit semantic guidance. In contrast, ours directly separates foreground and background in the *spatial* dimension for all video frames. This is still unused by previous siblings.

Besides, no temporal annotations makes the cross-modal alignment more challenging. To distinguish similar neighborhood frames, we further improve temporal representation of the foreground enforced visual features via a progressive contrastive sample generation module. This enlogates the moment of the predicted proposal to gradually improve the discriminability of foreground features in the *temporal* dimension. Our model follows CNM (Z. M. et al., 2022) but improves it by perceiving the foreground information in both spatial and temporal dimensions.

In summary, the contributions of our work are:

- We propose a novel weakly supervised TLG model called Foreground Enhanced Network (FEN) to boost the cross-modal interaction via explicit foreground modeling.
- We devise a progressive contrastive sample generation module to generates contrastive samples which gradually approach the predicted temporal proposal from the longer size of the moment.
- Experiments on two widely-used datasets demonstrate the effectiveness of our approach.

Related Work

Temporal language grounding

TLG aims to identify and localize the event, which is closely related to the nature language query, in a given video. Previous fully supervised works relies on temporal annotations during training. In contrast, weakly supervised TLG exploit video-sentence pairs for training. Most approaches in this respect are based on multi-instance learning (MIL) paradigm to learn cross-modal alignment. Recently, reconstruction-based methods are introduced for this field. SCN generates temporal proposals by sliding windows and choose the best one on reconstructing the masked words in the given query. In the CNM approach, gaussian masks are utilized to formulate temporal proposals, enabling an end-to-end training manner. However, the aforementioned methods are unaware of the importance of foreground, which represents the dynamic elements and closely related to the query.

Compared to previous methods, our approach has special advantages in that we explicitly model the dynamic foreground features in video features that are more relevant to the query by drawing inspiration from human visual attention mechanisms. We enhance the foreground in both spatial and temporal dimensions, thereby obtaining more discriminative cross-modal features, which is advantageous for achieving more accurate temporal localization.

Selective visual attention mechanism

Selective visual attention mechanism is one of the most fundamental cognitive functions in humans. It describes the tendency of visual processing to be primarily focused on stimuli that are relevant to behavior. When individuals engage in cognitive activities, the brain's selective attention mechanism amplifies behaviorally relevant sensory information

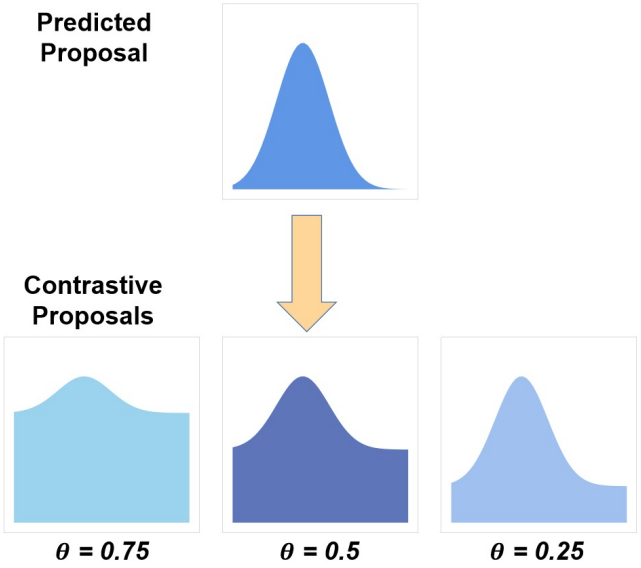

Figure 3: The weight mask of the generated contrast samples progressively aligns with the predicted proposal, promoting the model to learn the subtle differences between event boundaries and neighboring frames. α denotes the ratio of current training process to the total epochs.

while suppressing interference from distractors, thereby aiding in the rapid and accurate cognition . It is worth noting that the "attention" in Transformer (Ashish et al., 2017) is different from the selective attention mechanism mentioned here. The former is mainly used for processing sequential data and can capture long-range dependencies and global contextual information. In contrast, the human visual selective attention mechanism refers to the brain's conscious selection and processing of specific visual information during the perception and cognition process. Inspired by the selective visual attention mechanism, we explicitly model and process the foreground that is more relevant to the queried events, and reduce interference from backgrounds.

APPROACH

Overall Architecture

Given an untrimmed video and a sentence of query, our task aims to predict the moment T of the event matching to the query. In our task, we focus on weakly supervised setting for TLG, so we do not use temporal annotations in training. Figure 2 shows the entire architecture of our approach. We first use a video encoder to extract visual features $V = \{v_1, v_2, ..., v_{N_v}\} \in \mathbb{R}^{N_v \times D}$, where N_v is the number of frames, and D denotes feature size. For the query, we extract word embedding with Glove and feed them to a fully-connected layer to get sentence features $S = \{s_1, s_2, ..., s_{N_w}\}$ where N_w is the number of words. Then visual features will be divided into background and foreground features in a foreground perception module (FP). Visual features will be enhanced by foreground features and interact with sentence features to generate the Gaussian mask of the predicted proposal (moment). Both foreground and visual features serve to reconstruct the masked query to measure our proposal. Besides, a progressive contrastive sample generation module (PCSG) is devised to gradually produce the masks of different moment sizes of contrastive samples.

Feature Extractor

For a given video, a pre-trained visual model are adopted for extracting visual features $V = \{v_1, v_2, ..., v_N\} \in \mathbb{R}^{N_v \times D}$, where N_v is the number of frames and$D$ denotes to the feature dimension. For a input query, we adopt Glove to extract word embedding which is fed into a full connected layer to obtain the sentence feature $S = \{s_1, s_2, ..., s_N\} \in \mathbb{R}^{N_w \times D}$ where N_w is the number of words.

Foreground Perception Module

Visual features across frames can be thought of being primarily influenced by the foreground, because the background features remains relatively stable. But the background usually intervenes foreground information. Inspired by background subtraction, we use the low-rankness of the background to model the background, then subtract it to induce the foreground. We apply this insight to obtain background features V_{bg} by feeding visual features to a fully-connected layer:

$$V_{bg} = \mathbf{W}^{bg} V + \mathbf{b}^{bg} \tag{1}$$

where $\mathbf{W}^{bg}$ and $\mathbf{b}^{bg}$ are learnable parameters. To ensure the low-rank property of background features, we introduce nuclear norm to constrain the background features:

$$\mathcal{L}_{bg} = tr(\sqrt{{V_{bg}}^T V_{bg}}), \tag{2}$$

where $tr(\cdot)$ denotes trace of feature maps. Since the nuclear norm can serve as a convex approximation to the rank of feature maps, we can separate the background features by minimizing $\mathcal{L}_{bg}$. Thus, we can obtain the foreground explicitly by

$$V_{fg} = V - V_{bg}. \tag{3}$$

Moment Prediction Module

Since foreground features are relatively sparse and easily aligned with the query, we utilize them to enhance visual features in the spatial dimension:

$$V' = V \otimes \sigma(\mathbf{W}^{\mathbf{fg}} V_{fg} + \mathbf{b}^{fg}), \tag{4}$$

where $\mathbf{W}^{fg}$, $\mathbf{b}^{fg}$ are learnable parameters, $\sigma(\cdot)$ denotes the sigmoid function, and $\otimes$ means the Hadamard product. To conduct cross-modal interaction, the foreground enhanced visual features will be fed to a standard transformer with a localization head to predict the center c and the width w of the

matched moment. The architecture with a transformer encoder $Enc(\cdot)$ and a transformer decoder $Dec(\cdot)$ is:

$$H = Dec(V', Enc(S)) \tag{5}$$

where H is the hidden features produced by cross-modal fusion, which denotes the semantic-specific visual tokens. Using the foreground enhanced visual features, we follow CNM to formulate the matched temporal proposal by a Gaussian weight mask vector $p = N(c, w^2)$ to train the reconstruction model in an end-to-end manner.

Progressive Contrastive Sample Generation Module

To further improve temporal representation of foreground enhanced features, we design a progressive contrastive sample generation module (PCSG) to generate different sizes of contrastive temporal proposals. It initially generates a mask covering the entire video, and as the training iterations proceed, the generated mask gradually approaches the predicted temporal proposal by the smooth strategy:

$$p^* = \theta * p + (1-\theta) * \mathbf{1}, \tag{6}$$

where θ represents the ratio of completed training iterations to the total planned training iterations, and $\mathbf{1}$ denotes a vector with all the ones. As shown in Figure 3, the contrastive samples gradually approach the predicted proposal, which forces the model put its attention on from the entire video to the neighbor segments of the predicted segment. This allows for a better understanding of the differences between the predicted segment and other segments, thereby improving the discriminative ability of foreground features in different moments.

Reconstruction Module

To evaluate the predicted temporal proposal, we adopt the mask conditioned transformer in CNM to reconstruct the masked query Q^*, and calculate the cross-entropy $\mathcal{L}_{ce}^{p}$, $\mathcal{L}_{ce}^{p^*}$, and $\mathcal{L}_{ce}^{ev}$ with the reconstruction results using the foreground-enhanced visual features V', the mask of the predicted moment p, the mask of contrastive sample p^*, and the weight for entire video $\mathbf{1}$. The reconstruction loss $\mathcal{L}_r$ is the sum of such three cross-entropy losses. Considering the relevance of the event in predicted proposal is larger than that in the query, we compared the reconstruction results of different proposals:

$$\mathcal{L}_c = max(\mathcal{L}_{ce}^{p} + \mathcal{L}_{ce}^{p^*} + \beta_1, 0) + max(\mathcal{L}_{ce}^{p} + \mathcal{L}_{ce}^{ev} + \beta_2, 0), \tag{7}$$

where β_1 and β_2 are two hyperparameters to control the margins. To encourage better cross-modal interaction and generalization, a mixup based foreground feature V_{mix} is obtained by a linear interpolation operation on two foreground features V_{fg}^i and V_{fg}^k in a batch:

$$V_{mix} = \lambda V_{fg}^i + (1-\lambda) V_{fg}^k, \tag{8}$$

where λ is a random weight from 0 to 1. Then V_{mix} will be used to reconstruct the query S_i and S_k, which are related to V_{fg}^i and V_{fg}^k, respectively, and calculate the cross-entropy $\mathcal{L}_{ce}^i$, $\mathcal{L}_{ce}^k$, separately. The loss of the mixup based reconstruction is obtained by the weighted summation between $\mathcal{L}_{ce}^i$ and $\mathcal{L}_{ce}^k$. Additionally, $\mathcal{L}_{cmix}$ is designed to ensure the reconstruction results using the original video feature to perform better than that using the mixup features. They have the following forms:

$$\mathcal{L}_{mix} = \lambda \mathcal{L}_{ce}^i + (1-\lambda)\mathcal{L}_{ce}^k, \tag{9}$$

$$\mathcal{L}_{cmix} = max(\mathcal{L}_{ce}^{p} - \mathcal{L}_{mix} + \beta_3, 0), \tag{10}$$

where β_3 is a marginal parameter. Then, the total loss is calculated as follows:

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_c + \mathcal{L}_{mix} + \mathcal{L}_{cmix} + \alpha \mathcal{L}_{bg}, \tag{11}$$

where α is a weight to balance the impact of different losses.

Table 1: Performance comparison on ActivityNet.The best and second best results in all the tables are highlighted in **bold** and underlined

Method	IoU = 0.1	IoU = 0.3	IoU = 0.5
MARN	-	47.01	29.95
SCN	71.48	47.23	29.22
RTBPN	73.73	49.77	29.63
WSLLN	75.4	42.8	22.7
LCNet	78.58	48.49	26.33
WSTAN	79.78	52.45	30.01
CRM	**81.61**	55.26	32.19
CNM	78.13	55.68	33.33
Ours	80.86	**57.64**	**33.75**

Table 2: Performance comparison on Charades-STA. The best results are in **bold**, and the second best are underlined.

Method	IoU = 0.3	IoU = 0.5	IoU = 0.7
TGA	32.14	19.94	8.84
SCN	42.96	23.58	9.97
WSTAN	43.39	29.35	12.28
LoGAN	48.04	31.74	13.71
MARN	48.55	31.94	14.81
CRM	53.66	34.76	16.37
LCNet	59.60	**39.19**	**18.87**
RTBPN	60.04	32.36	13.24
CNM	60.04	35.15	14.95
Ours	**61.43**	38.44	18.37

Experiment

Experimental Settings

To evaluate our method, we conduct the experiments of temporal language grounding on two benchmark datasets. An indoor activity dataset **Charades-STA** contains 12,408 moment-sentence pairs in the training set and 3,720 pairs in the training set. We report the results on the test set.

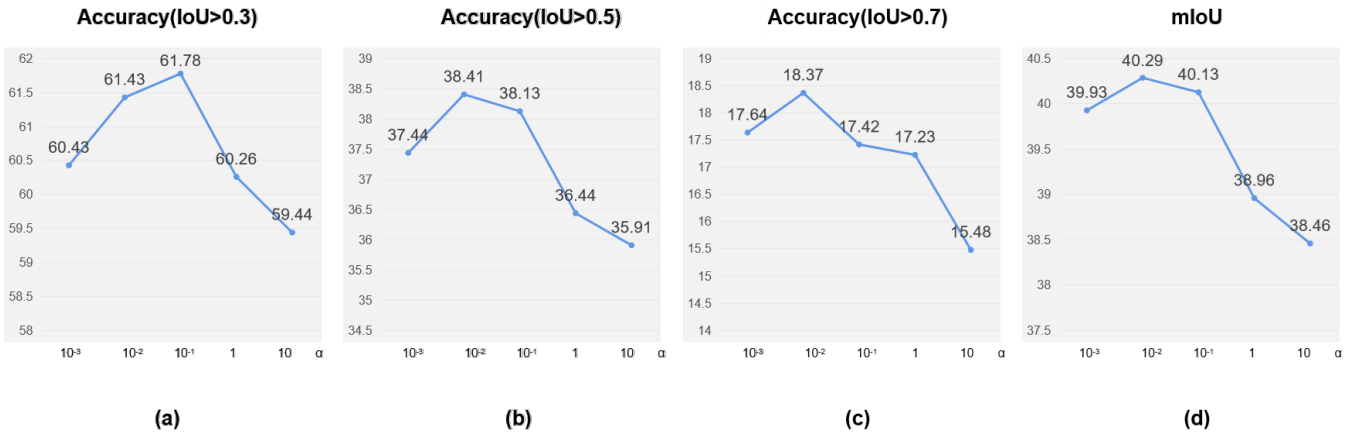

Figure 4: The results with different values of α.

Table 3: The effectiveness of the progressive contrastive sample generation module on Charades dataset.

Method	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
Full Model	**61.43**	**38.44**	**18.37**	**40.29**
Fixed	60.20	37.59	17.13	39.12
None	58.90	36.73	16.37	38.48

Table 4: The importance of the foreground perception module (FP) on Charades dataset.

Method	IoU=0.1	IoU=0.3	IoU=0.5	mIoU
Full Model	**61.43**	**38.44**	**18.37**	**40.29**
w/o. FP	59.12	37.18	17.13	38.16

ActivityNet-Captions has more than 70k pairs about open world activities, which is currently the largest dataset in TLG. We report the results on val_2 split. All the results are measured by the evaluation metric '$IoU > m$', which means the percentage of the predicted moments when the Intersection over Union (IoU) is larger than the threshold m. We separately use m = {0.1,0.3,0.5} for ActivityNet-Captions, and m = {0.3,0.5,0.7} for Charades-STA.

Following CNM (Z. M. et al., 2022), we extract visual features by I3D for Charades-STA, and CLIP for ActivityNet-Captions, respectively. Word embeddings are extracted by Glove with 300 dimensions. For hyperparameters, we set $\alpha = 0.01$, $\beta_1 = 0.05$, $\beta_2 = 0.1$, and $\beta_3 = 0.25$ for all datasets.

Experiment Results

Comparison to the baseline. We choose CNM as the baseline to show the effectiveness of our method, since our reconstruction framework is similar to CNM. In contrast, our method has substantial improvements across all metrics on both datasets. Specially, on Charades-STA, when considering accuracy at IoU=0.3, 0.5, and 0.7, ours achieves large improvements of 1.41%, 3.29%, and 3.42%, respectively.

Comparison to SOTAs. Table 1 and Table 2 show the results of our method and previous state-of-the-art methods on ActivityNet-Captions and Charades-STA. From such tables, we observe that our model achieves the best performance on IoU=0.3 for Charades-STA, while LCNet surpasses ours on IoU=0.5 and IoU =0.7. However, ours outperforms LCNet on ActivityNet-Captions, which involve more complex scenes. Although CRM slightly performs better than our method on IoU=0.1, our method consistently outweighs previous methods on both IoU=0.3 and IoU=0.5. Importantly, ours behaves stably, unlike CRM and LCNet that have unstable results on two datasets, that is, behaves well on one dataset but poorly on the other. Ours always performs well on both datasets. This hints the potential of our approach.

Ablation Study

To analyze the effectiveness of the foreground perception module (FP), and progressive contrastive sample module, we conduct ablation study on Activity-Captions datasets. We also introduce the mean Intersection over Union (mIoU) to report the results.

Effectiveness of foreground perception module. We introduce a ablation model by removing FP. As shown in Table 4, the full model outweighs the ablation model on all the metrics. This can be attributed to FP which highlights the foreground in the spatial dimension, improving the cross-modal alignment.

Effectiveness of progressive contrastive sample generation module. Besides, we evaluate the different ways to produce contrastive temporal proposals to highlight the efficacy of our progressive proposal generation way. As the results shown in Table 3, where "Fixed" denotes that a fixed Gaussian mask are provided as a contrastive sample for all predicted proposals, and "None" means that there is no contrastive sample available, our progressive method exhibits overall superiority over all the other ways. It is worthy noting that the "Fixed" method surpasses ours on IoU = 0.1,

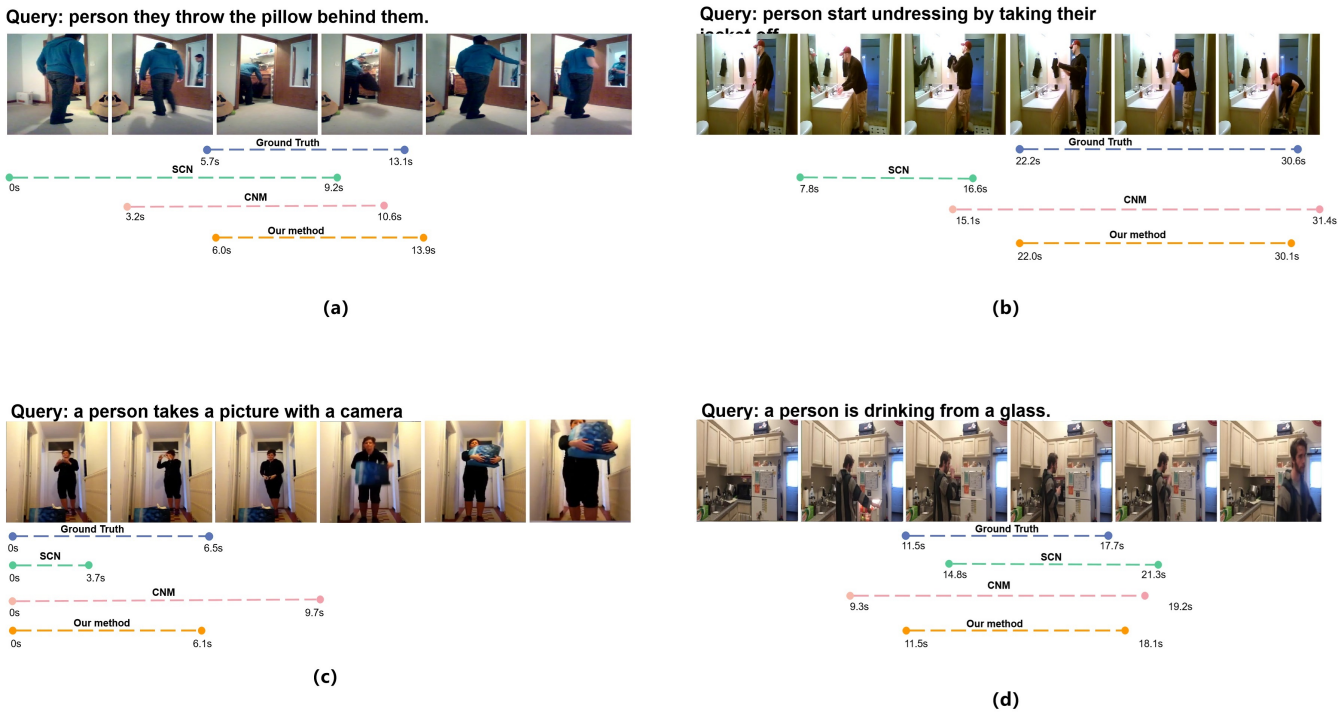

Figure 5: The qualitative results on Charades-STA.

while our method is better than it on both IoU = 0.3 and IoU =0.5, because progressive contrastive samples improves the discriminative ability of the model to distinguish similar neighborhood frames.

Sensitivity Analysis

We utilize different values of the weight of the low-rank constraint in the background loss α and observe the accuracy considering $IoU > (0.3, 0.5, 0.7)$ to analyze the sensitivity while other hyperparameters are unchanged.

As shown in Figure 4, the performance is stable when α is between 10^{-2} to 10^{-1}. This reason may be that the too high weight α will cause the model to mistakenly identify some foreground elements as background, consequently discarding useful dynamic information. Additionally, when this weight is too low, the results are also unsatisfactory, due to some background information is considered as foreground, still causing interference with the perception of the foreground.

Qualitative Result

Figure 5 shows some qualitative results of our approach. The subfigure (a) and (b) demonstrates that our method is capable of better cognition and understanding of video content and the events described in the natural language queries, thereby enabling more precise localization. This is attributed to our model's ability to capture dynamic elements in foreground by eliminating interference from useless background information. Moreover, from Figure 5(c) and (d), our approach localizes the more precisely moment boundaries than previous methods, benefiting from gradually learning the differences on foreground information between event boundaries and neighboring frames.

Scalability and Future Work

Although the results in Figure 5 only demonstrate the effective localization of our model in videos containing single objects in the foreground, our model remains effective for videos containing multiple objects, as evidenced by its good performance on ActivityNet datasets that include complex scenes and multiple objects. In future work, we will investigate how to extend the foreground enhanced method to more complex scenarios, such as videos with frequent background changes due to a large number of scene transitions.

Conclusion

This paper introduces a novel weakly supervised TLG model named Foreground Enhanced Network (FEN), which is inspired by selective visual attention mechanism to explicitly capture foreground in spatial dimension, promoting better cross-modal interactions. Besides, a progressive contrastive sample generation module serves to force the model to gradually learning a more distinguishable foreground representation in the temporal dimension. Experiments on two popular datasets verify the effectiveness of our approach.

References

A., H. L., O., W., E., S., J., S., T., D., & B., R. (2017). Localizing moments in video with natural language. *Proceedings of the IEEE International Conference on Computer Vision*, 5804–5813.

Alec, R. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748-8763.

Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., N, G. A., ... Illia, P. (2017). Attention is all you need. *arXiv*.

C., M. N., S., P., & K., R.-C. A. (2019). Weakly supervised video moment retrieval from text queries. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11592-11601.

David, M. G. F. J. W. (2013). Motion-dependent representation of space in area mt. *NEURON*, *78*(3), 554-562.

J., C., & A., Z. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299-6308.

J., G., C., S., Z., Y., & R., N. (2017). Tall: Temporal activity localization via language query. *Proceedings of the IEEE International Conference on Computer Vision.*, 5267-5275.

J., H., Y., L., S., G., & H., J. (2021). Cross-sentence temporal and semantic relations in video activity localization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7199-7208.

J., P., R., S., & D., M. C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543.

J., R. R. A. J. K. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 368-373.

J., W., L., M., & W., J. (2019). Temporally grounding language queries in videos by contextual boundary-aware prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*.

J.M., S. T. P. (2012). The penny drops: Change blindness at fixation. *Perception*, 489-492.

K., L., D., G., & M., W. (2021). Proposal-free video grounding with contextual pyramid network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(3), 1902-1910.

M., G., S., D. L., R., S., & C., X. (2019). Wslln: Weakly supervised natural language localization networks. *Empirical Methods Natural Lang. Process.*, 1481–1487.

M., Z., Y., H., Q., C., & Y., L. (2022). Weakly supervised video moment localization with contrastive negative sample mining. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(3), 3517-3525.

R., K., K., H., F., R., F., L., & C., N. J. (2017). Dense-captioning events in videos. *Proceedings of the IEEE International Conference on Computer Vision*, 706-715.

R., T., H., X., K., S., & A., P. B. (2021). Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2083-2092.

S., G., A., A., Z., P., & A., H. (2019). Excl: Extractive clip localization using natural language descriptions. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

S., Z., H., P., J., F., & J., L. (2020). Learning 2d temporal adjacent networks for moment localization with natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(7), 12870-12877.

W., Y., T., Z., Y., Z., & F., W. (2021). Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 3252-3262.

X., S., Lan, L., H., T., X., Z., X., M., & Z., L. (2022). Joint modality synergy and spatio-temporal cue purification for moment localization. *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 369-379.

Y., S., J., W., L., M., Z., Y., & J., Y. (n.d.). Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048.*.

Y., W., J., D., W., Z., & H., L. (2021). Weakly supervised video moment retrieval from text queries. *IEEE Transactions on Multimedia*, 3276-3286.

Z., L., Z., Z., Z., Z., Q., W., & H., L. (2020). Weakly-supervised video moment retrieval via semantic completion network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(7), 11539-11546.

Z., L., Z., Z., Z., Z., Z., Z., & D., C. (2020). Moment retrieval via cross-modal interaction networks with query reconstruction. *IEEE Transactions on Image Process.*, 3750–3762.

Z., Z., Z., L., Z., Z., J., Z., & X., H. (2020). Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. *Proceedings of the 28th ACM International Conference on Multimedia*, 4098-4106.

Zirnsak;, T. M. (2017). Neural mechanisms of selective visual attention. *Annual Review of Psychology*, 47-72.