International Journal of Science Academic Research
Vol. 03, Issue 09, pp.4347-4348, September, 2022
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

Research Article

TPE-CPL: TRAINABLE POSITIONAL EMBEDDING-BASED CONTRASTIVE PROPOSAL LEARNING FOR WEAKLY SUPERVISED VIDEO GROUNDING

Richard Hua, Yantao Wang, Jason Zou, Allen Jiang, Sunny Kim and *Jong Wook Lee

Lexington high school, United States

Received 25th July 2022; **Accepted** 27th August 2022; **Published online** 20th September 2022

Abstract

As videos on the internet become more common, we need to understand the contents of the videos for recognizing important human actions or highlights. Moreover, videos with texts which depict the key points in the videos have encouraged research on video grounding (Gao, Jiyang, *et al,* 2017). Video grounding is an important task with many applications in video surveillance. Video grounding aims to find a grounding location, which is a video segment semantically corresponding to a query sentence in a long and untrimmed video. Recently, weakly supervised video grounding (Zheng, Minghang *et al.,* 2022) has drawn more attention because it requires little annotation cost. In weakly supervised video grounding, the ground-truth grounding location is not available for training, and only matched pairs of video and query sentence are available. In this paper, we propose Trainable Positional Embedding (TPE)-based contrastive proposal learning for weakly supervised video grounding. The previous method for contrastive proposal learning (Zheng, Minghang *et al.,* 2022) leverages several Gaussian masks which can be positive proposals for finding grounding locations. However, the predefined Sinusoidal positional embedding is used in that method, which is not efficient because it ignores varying information of word positions in the query sentence. To solve this problem, we leverage trainable positional embedding for contrastive proposal learning. We verify that the proposed method improves performance through quantitative experiments, outperforming the previous state-of-the-art methods.

Keywords: Video, Trainable Positional Embedding, Weakly Supervised, Training

INTRODUCTION

As videos on the internet become more common, we need to understand the contents of the videos for recognizing important human actions or highlights. Moreover, videos with texts which depict the key points in the videos have encouraged research on video grounding (Gao, Jiyang, *et al,* 2017). Video grounding is an important task with many applications in video surveillance. Video grounding aims to find a grounding location, which is a video segment semantically corresponding to a query sentence in a long and untrimmed video. Recently, weakly supervised video grounding (Zheng, Minghang *et al.,* 2022) has drawn more attention because it requires little annotation cost. In weakly supervised video grounding, the ground-truth grounding location is not available for training, and only matched pairs of video and query sentence are available. In this paper, we propose Trainable Positional Embedding (TPE)-based contrastive proposal learning for weakly supervised video grounding. The previous method for contrastive proposal learning (Zheng, Minghang *et al.,* 2022) leverages several Gaussian masks which can be positive proposals for finding grounding locations. However, the predefined Sinusoidal positional embedding is used in that method, which is not efficient because it ignores varying information of word positions in the query sentence. To solve this problem, we leverage trainable positional embedding for contrastive proposal learning. We verify that the proposed method improves performance through quantitative experiments, outperforming the previous state-of-the-art methods.

*Corresponding Author: *Jong Wook Lee*
Lexington high school, United States.

METHODS

Given a long and untrimmed video and a sentence query, the goal of video grounding is to find a grounding location that is defined as a video segment corresponding to the query. There is one ground-truth grounding location in one video-query pair, which can be represented as starting time and ending time(g_s, g_e). The network of the proposed method is depicted in Fig. 1. Based on the Contrastive Proposal Learning (CPL) network [2], we added Trainable Positional Embedding (TPE) network before CPL network. TPE network makes position-aware word features from the sentence query for the input of CPL network. Then CPL network finally predicts a video segment corresponding to the query for solving weakly supervised video grounding.

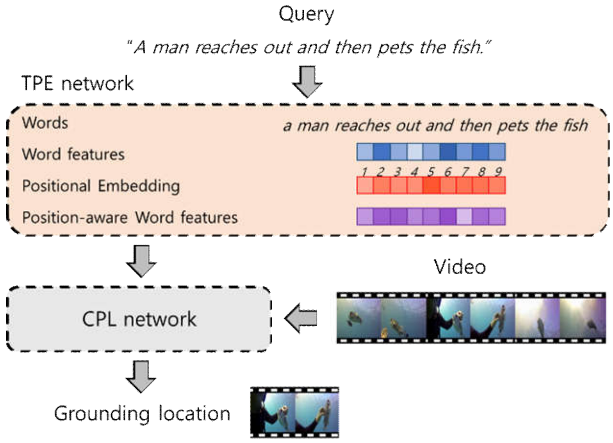
Query
"A man reaches out and then pets the fish."
TPE network
Words
a man reaches out and then pets the fish
Word features
Positional Embedding
1 2 3 4 5 6 7 8 9
Position-aware Word features
Video
CPL network
Grounding location

Figure 1. the overall architecture of Trainable Positional Embedding-based Contrastive Proposal Learning (TPE-CPL) for Weakly Supervised Video Grounding

Trainable Positional Embedding Network

Given a sentence query, TPE network makes position-aware word features. Then, the features are used as input to CPL Network. Specifically, TPE network is composed of multiple steps as followings. First, we make every word in a sentence query to a lowercase letter. Second, we embed each word to produce a word feature matrix$Q$ using GloVe word embedding [6]:

$$Q = \{q_1, q_2, \dots, q_N\} \in \mathbb{R}^{N \times d},$$

where $N$is the number of words in the query, $d$is the dimension of the feature, and q_n is the n-th word feature vector. Third, inspired by the positional encoding method in BERT [3], we leverage trainable positional embedding to capture positional information of words. We embed the position n of the n-th word through a functionf_{emb}that multiplies the input by a trainable weight matrix. The output of the function is the matrix of position feature vectors$P$:

$$P = f_{emb}([1, 2, \dots, N]) \in \mathbb{R}^{N \times d}.$$

Finally, each position feature vector is added to the corresponding word feature vector (*i.e.*$f_{emb}(n)$is added to q_n) to reflect position information in the word features. As a result, we can produce a position-aware word feature matrixQ_{pos}, which can be defined as

$$Q_{pos} = Q + P = \{q'_1, q'_2, \dots, q'_N\},$$

where q'_n is the n-th position-aware word feature vector.

Contrastive Proposal Learning (Zheng, Minghang *et al.*, 2022)

Using the position-aware word features and video features, CPL (Zheng, Minghang *et al.,* 2022) yields a grounding location that represents A video segment corresponding to the query. CPL is composed of two stages: 1) the proposal generation stage and 2) the mask-conditioned reconstruction stage. The proposal generation stage generates trainable Gaussian masks to represent both positive and negative proposals within the given video. The positive proposal is used as not only a mask for the video but also predicted grounding locations. The mask-conditioned reconstruction stage reconstructs the sentence query from the masked query and video with the Gaussian masks. CPL assumes that the well-generated proposals can reconstruct the sentence query. There are two losses: 1) reconstruction loss which is the cross-entropy loss to better reconstruct the query using the generated proposals and 2) intra-video contrastive loss which makes the proposal generation stage more efficient for distinct positive and negative proposals.

Table 1. Comparisons on the Charades-STA dataset

Method	R1@ tIoU 0.3	R1@ tIoU 0.5	R1@ tIoU0.7	R1@ mIoU
CNM [1]	60.04	35.15	14.95	38.11
CPL [2]	66.40	49.24	22.39	43.48
Ours	68.27	50.63	23.37	44.29

EXPERIMENT

For the experiment, we evaluate the proposed method on the Charades-STA dataset, which is public and widely used for video grounding is a widely used dataset for temporal video grounding. In the previous method for video grounding (Gao, Jiyang *et al.,* 2017), this dataset is created from the Charades dataset (Sigurdsson *et al.,* 2016) which is used for action recognition and video captioning. We split the training data and testing data following the data splitting strategy in (Gao, Jiyang *et al.,* 2017). For training, there are 12,408 video-query pairs and for testing, there are 3,720. For evaluation, we adopt two metrics1) recall at predefined thresholds (0.3, 0.5, 0.7) of the temporal Intersection over Union~(R1@tIoU) and 2) mean of all tIoUs~(mIoU). We can calculate R@tIoU by computing the percentage of testing data with the tIoU larger than the predefined thresholds (0.3, 0.5, 0.7). As shown in Table 1, the results show that the proposed method makes the best performance in all evaluation metrics compared to the CNM (Zheng, Minghang *et al.,* 2022) and CPL (Zheng, Minghang *et al.,* 2022) .The proposed method outperforms the CPL by 1.87, 1.39, 0.98, and 0.81 in R1@tIoU 0.3, 0.5, 0.7, and mIoU, respectively. This result verifies the effectiveness of the proposed method and shows that the trainable positional embedding is effective for weakly supervised video grounding.

Conclusion

To solve the weakly supervised video grounding task, in this paper, we propose Trainable Positional Embedding-based Contrastive Proposal Learning (TPE-CPL). The previous contrastive proposal learning method uses the predefined Sinusoidal positional embedding and this technique ignores varying information of word positions in the query sentence. To fully understand the word position information, we leverage trainable positional embedding and employ this embedding technique to the previous contrastive proposal learning model. The quantitative experimental results show that the proposed method outperforms the previous state-of-the-art methods for weakly supervised video grounding.

REFERENCES

Devlin, Jacob and Chang, et al. 2018. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” arXiv preprint arXiv:1810.04805.

Gao, Jiyang, et al. 2017. "Tall: Temporal activity localization via language query." ICCV.

Pennington, J. et al. 2014. “Glove: Global Vectors for Word Representation.” Conference on Empirical Methods in Natural Language Processing.

Sigurdsson, Gunnar A., et al. 2016. "Hollywood in homes: Crowdsourcing data collection for activity understanding." ECCV.

Zheng, Minghang, et al. 2022. "Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning" CVPR.

Zheng, Minghang, et al. 2022. "Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining" AAAI.

****