AsyNCE: Disentangling False-Positives for Weakly-Supervised Video Grounding

Cheng Da*
Machine Intelligence Technology Lab, Alibaba Group
dacheng.dc@alibaba-inc.com

Yanhao Zhang*
Machine Intelligence Technology Lab, Alibaba Group
yanhao.zyh@alibaba-inc.com

Yun Zheng
Machine Intelligence Technology Lab, Alibaba Group
zhengyun.zy@alibaba-inc.com

Pan Pan
Machine Intelligence Technology Lab, Alibaba Group
panpan.pp@alibaba-inc.com

Yinghui Xu
Machine Intelligence Technology Lab, Alibaba Group
renji.xyh@taobao.com

Chunhong Pan
Institute of Automation, Chinese Academy of Sciences
chpan@nlpr.ia.ac.cn

ABSTRACT

Weakly-supervised video grounding has been investigated to ground textual phases in video content with only video-sentence pairs provided during training, for the lack of prohibitively costly bounding box annotations. Existing methods cast this task into a frame-level multiple instance learning (MIL) problem with the ranking loss. While an object might appear sparsely across multiple frames, causing uncertain false-positive frames. Thus, directly computing the average loss of all frames is inadequate in video domain. Moreover, the positive and negative pairs are equally coupling in ranking loss, so that it is impossible to handle false-positive frames individually. Additionally, naive inner production is suboptimal for the similarity measure of cross domains. To solve these issues, we propose a novel AsyNCE loss to flexibly disentangle the positive pairs from negative ones in frame-level MIL, which allows for mitigating the uncertainty of false-positive frames effectively. Besides, a cross-modal transformer block is introduced to purify the text feature by image frame context, generating a visual-guided text feature for better similarity measure. Extensive experiments on YouCook2, RoboWatch and WAB datasets demonstrate the superiority and robustness of our method over state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**.

KEYWORDS

AsyNCE, weakly-supervised learning, cross-modal transformer

ACM Reference Format:
Cheng Da, Yanhao Zhang, Yun Zheng, Pan Pan, Yinghui Xu, and Chunhong Pan. 2021. AsyNCE: Disentangling False-Positives for Weakly-Supervised Video Grounding. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3481539

*Both authors contributed equally to this research.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3481539

1 INTRODUCTION

Visual grounding [2, 8, 31–33, 38] investigates the alignment between vision and language modalities, and has emerged as a prominent interdisciplinary research problem in the computer vision and natural language processing communities. Specifically, image grounding [10, 12, 16, 35] aims to locate corresponding image regions associated with sentence description. Recently, there have been numerous studies that attempt to address video grounding, where the related objects are required to be localized in each frame. Owing to the prohibitively costly object region annotations in image, weakly-supervised visual grounding [5, 9, 11, 21, 33, 34] has been researched widely. Notably, temporal localization aims to ground video segment from untrimmed videos [2, 8, 14, 17, 30, 36]. And spatio-temporal grounding targets at predicting the temporal and spatial location of query [1, 5, 33]. In this paper, we mainly focus on the weakly-supervised spatial video grounding [21, 40] on trimmed videos.

Technically, DVSA [11] introduces a multiple instance learning (MIL) framework to solve the weakly-supervised image grounding problem, where only the alignment of image and sentence is available for training. Specifically, given one image and the corresponding sentence, DVSA assumes that at least one region of the image is related to one entity in sentence reasonably. Thus, if the image is defined as a "bag" and the regions of image are denoted as instances, we then denote this image-sentence pair as a "positive bag" in MIL problem. And the "negative bag" is defined as one image with an uncorrelated query sentence. Furthermore, the image-sentence scores can be formulated as one aggregated function of the individual region-entity scores of positive or negative bags. Moreover, ranking loss is employed to guarantee a good matching, in which the image-sentence score of aligned positive bag is far beyond the unaligned negative one. Finally, the correspondence between region and entity can be uncovered by the generated region-entity score.

Following weakly-supervised image grounding, most weakly-supervised video grounding methods [21, 33, 40] adapt the mechanism of ranking-based MIL of image to video domain. However, one video contains multiple frames. If the all regions of all frames

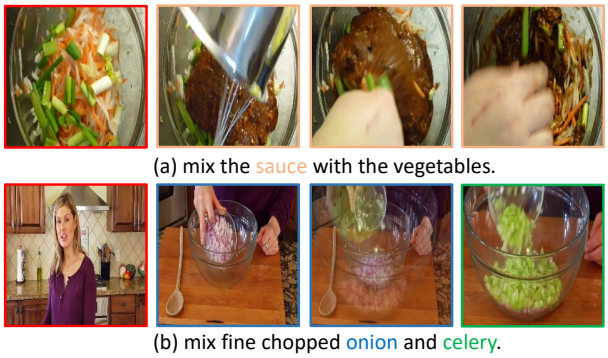

Figure 1: The illustration of false-positive frames of two video sequences in YouCook2.

are regarded as one bag, the volume of bag is too huge to control. This video-level MIL problem is cumbersome, as the video is long. Alternatively, frame-level MIL problem is constructed for video grounding, in which the sentence of the video is assigned as the annotation of each frame and each frame is defined as a bag in MIL [21, 40].

But this straightforward adaptation of frame-level MIL is inadequate in videos. Concretely, a video segment usually lasts a few seconds or even minutes, and thus one object might appear sparsely across frames. Since the video annotation is weakly aligned to each frame in the frame-level fashion, it is easy to cause false-positive frames. As illustrated in Fig. 1 (a), the sauce does not appear in the first frame, so that the first frame with sauce is a false-positive bag. Additionally, the onion only appears in the second and third frame, but the celery only appears in the last frame, as in Fig. 1 (b). It is uncertain to measure how positive each frame is, and thus dealing with all frames equally and computing the average loss of each frame are not meaningful in frame-level video grounding. To overcome this limitation, a weighted loss function is proposed [40] to reduce the effect of false-positive frames. Then, an augmented similarity is defined [21] to handle false-positive frames. However, this ranking-based MIL framework does not solve false-positive frames in essence. In short, we summary two main issues of these ranking-based weakly-supervised video grounding methods as follows:

Inequality. Since a video segment usually lasts a few seconds or even minutes, and one entity might appear sparsely across frames, causing false-positive frames. Due to the unequal certainties of frames on video domain, directly computing the average loss of all frames is not meaningful in video grounding problem. Meanwhile, ranking loss directly measures the discrepancy between positive and negative pairs, both of which are definitely convinced and equally important. This assumption is guaranteed in image grounding. However, the false-positive frames might impair the computation of the ranking loss in videos. In addition, since the positive and negative pairs are coupling in ranking-based formulation, it is difficult to handle false-positive frames individually and retain the importance of the negative pairs concurrently.

Incompatibility. The similarity score is often directly computed by naive inner production between vision and language domains. However, due to the semantic gap, the linguistic semantic information might be more abstract than image one. For instance, a word "egg" can represent a whole egg or yolk. Thus, inner production could be suboptimal to measure the features between two incompatible domains.

In this paper, we propose a novel asymmetric noise contrastive estimation (AsyNCE) loss and a cross-modal transformer block to address above mentioned issues. The NCE loss is commonly used in self-supervised feature learning. Natively, since all the independent samples are exponential summed in the softmax manner, NCE enables access to eliminating the coupling of positive and negative pairs. Therefore, we propose AsyNCE to control positive pairs individually. Specifically, soft-frame similarity is devised as the aggregation over all frames on positive pairs to reduce the uncertainty of false-positive frames. This aggregation can be implemented by the average form or weighted one with softmax. While the similarity scores of negative pairs are still computed for each frame. By this asymmetric manner, the impact of false-positive frames can be reduced greatly, and the effect of negative pairs can also be exploited sufficiently in AsyNCE. Additionally, we propose soft-entity similarity to alleviate sparsity further. Concretely, the entity-frame similarities are weighted over all frames by softmax, referring to each entity. By leveraging soft-entity similarity on positive and negative pairs, the positive frames could be emphasized and the hard negative frames might attract more attention.

Visual-language BERT [13, 26] aims to utilize cross-modal transformer to exchange information between different modalities, yielding good pre-trained feature for down-stream tasks. Thus, a cross-modal transformer block is introduced to measure the features between two incompatible domains in our framework. Concretely, we set the modal of Query as text feature, Key and Value as image feature. The output of this cross-modal transformer can be regarded as an attention weighted image feature. In this way, the query feature is adapted into the image feature space. And this newly generated visual-guided text feature can correlate with the specific image context sufficiently. Then, this refined text feature is further utilized to compute the cross-modal similarity by inner production form for a promising similarity measure.

Extensive experiments on YouCook2 demonstrate the superiority and robustness of our method. Then the generalization is also evaluated on RoboWatch dataset. Typically, we carry out experiments on Watch and Buy (WAB) dataset by grounding the commodity title to the clothing bounding box to show the transportability, which has been used on the live-stream video commodity localization in E-commerce.

The main contributions are outlined as follows:

- Beyond the traditional ranking loss, a novel AsyNCE loss with soft-frame and soft-entity similarity is proposed to solve the essential problem of false-positive frames in weakly-supervised video grounding, resulting in a flexibly decoupling mechanism for alleviating the uncertainty of false-positive frames in MIL.
- Cross-modal transformer is introduced to purify the text feature according to the specific image context, generating a visual-guided text feature. By resorting to this refined text

feature, the similarity between two incompatible domains is more convinced.
- Experiments on video grounding dataset YouCook2, RoboWatch and commodity video dataset WAB demonstrate the superiority, robustness, transportability and generalization of our approach.

2 RELATED WORK

2.1 Weakly-supervised Spatial Video Grounding

In the literature, supervised image grounding has been investigated intensively [12, 16, 35]. However, the elaborate annotations of bounding box are expensive to obtain [20]. Recently, weakly-supervised image grounding only requires descriptive phrases, rather than explicit grounding boxes for training, attracting extensive attention from the community. [19] proposes a phrase reconstruction approach, where the intermediate latent text feature is represented by the most related image regions with cross-modal attention. [11] casts this task into a MIL framework, where image-sentence score is formulated as the function of the individual region-entity scores. Moreover, a ranking loss function can be optimized based on these aligned and unaligned image-sentence pairs. Then, the correspondence between region and entity can be identified by the generated region-entity scores.

Weakly-supervised spatial video grounding attempts to adapt the mechanism of image grounding to video domain [34]. Following [11], [40] extends the segment-level annotations of video to the frame-level ones, and presents frame-wise weighting loss to ground text in each frame. [9] optimizes the visually grounded action graph that explicitly encompasses the latent dependencies between visual grounding and reference resolution, resulting in a new reference-aware multiple instance learning (RA-MIL) method. Then, contextual similarity [21] is proposed to deal with sparse objects association across frames in frame-level MIL learning. Furthermore, ranking loss is also utilized in [5] to align spatio-temporal tube and natural sentence, making spatio-temporally grounding accessible. In brief, most of weakly-supervised video grounding methods rely on the coupling ranking loss and MIL framework. Whereas, we propose a novel AsyNCE loss with soft similarity to tackle this problem in this paper.

2.2 Cross-modal Transformer

Transformer architecture is first proposed in [28] for neural machine translation tasks, and further used in BERT [6] for large-scale self-supervised representation learning in NLP domain. Recently, there has been a lot of progress in vision-language BERT to learn visual-linguistic representations [23–25, 39]. Typically, cross-modal transformer is employed to exchange information between different modalities, where the modal of Query is different to Key and Value. For instance, [13] presents co-attentional transformer to further fuse the image and text feature, and [26] proposes cross-modality encoder to better learn the cross-modal alignment between vision and language. Moreover, cross-modal transformer is utilized to model the interactions between multi-modal sequences over distinct time steps and latently align audio, vision and language streams to each other [27]. Beyond two modalities in one transformer, [42] devises a tangled transformer block for cross-modal feature learning from three sources, *i.e.* global actions, local regional objects, and linguistic tokens. As mentioned above, cross-modal transformer with distinct modalities between query and key-value pair is capable of facilitating communication of cross-modal information, leading to promising refined feature.

2.3 NCE for Weakly-supervised Alignment

Self-supervised learning has emerged as a prominent problem for visual feature representation from unlabelled data. Recent methods of contrastive learning [3, 4] conventionally employ NCE loss to distinguish positive and negative pairs for feature training, resulting in superior feature expression. Typically, [24] directly utilizes real-valued features, rather than a fixed discrete vocabulary, for video BERT training by replacing the softmax loss with NCE one. [29] improves weakly-supervised image grounding by contrastive knowledge distillation, which is also formulated as NCE framework. MIL-NCE [14] learns a strong video representation to address visually misaligned narrations from uncurated instructional videos. [37] presents counterfactual contrastive learning paradigm for weakly-supervised temporal video grounding by well-designed counterfactual positive and negative results. In short, the mechanism of NCE is intuitively suitable for the learning of weakly-supervised alignment. Thus, we go a step further and present AsyNCE with soft similarity to handle the characteristic of weakly-supervised video grounding.

3 METHOD

3.1 MIL Ranking-based Formulation

Suppose that the given video segment is represented by $\mathcal{V} = \{\mathbf{V}_t\}_{t=1}^{T}$, where T denotes the total number of frames in one video and each frame $\mathbf{V}_t$ includes a set of N region proposals $\{\mathbf{v}_t^n\}_{n=1}^{N} \in \mathbb{R}^{N\times d}$. And its corresponding natural sentence description is denoted as $\mathbf{Q} = \{\mathbf{q}_k\}_{k=1}^{K} \in \mathbb{R}^{K\times d}$, where each query entity $\mathbf{q}_k$ represents one word or one phase in the sentence, K is the total number of words and d is the dimension. The goal of video grounding task is to localize the query entity in every frame $\mathbf{V}_t$, referring to each query $\mathbf{q}_k$.

However, as the task is established in a weakly-supervised manner, the learning only has access to the ground-truth correlation of video-sentence pairs $(\mathcal{V}, \mathbf{Q})$. The fine-grained region-entity annotations $(\mathbf{v}_t^n, \mathbf{q}_k)$ are not available during training. To conquer the challenge of weakly-supervised setting, this video grounding task is recast into a frame-level multiple instance learning (MIL) problem. Specifically, the frame-sentence score is formulated as a function of the individual region-entity scores. Intuitively, the maximum of similarities between an entity and regions in one frame can be interpreted as the matching score of frame-entity pair. Thus, the frame-sentence similarity is formulated as:

$$S(\mathbf{V}_t, \mathbf{Q}) = \frac{1}{K}\sum_{k=1}^{K} S(\mathbf{V}_t, \mathbf{q}_k) = \frac{1}{K}\sum_{k=1}^{K} \max_{n} s(\mathbf{v}_t^n, \mathbf{q}_k). \quad (1)$$

Here, $s(\cdot)$ measures the similarity between text entity $\mathbf{q}_k$ and image region $\mathbf{v}_t^n$. Concretely, this similarity is commonly realized by

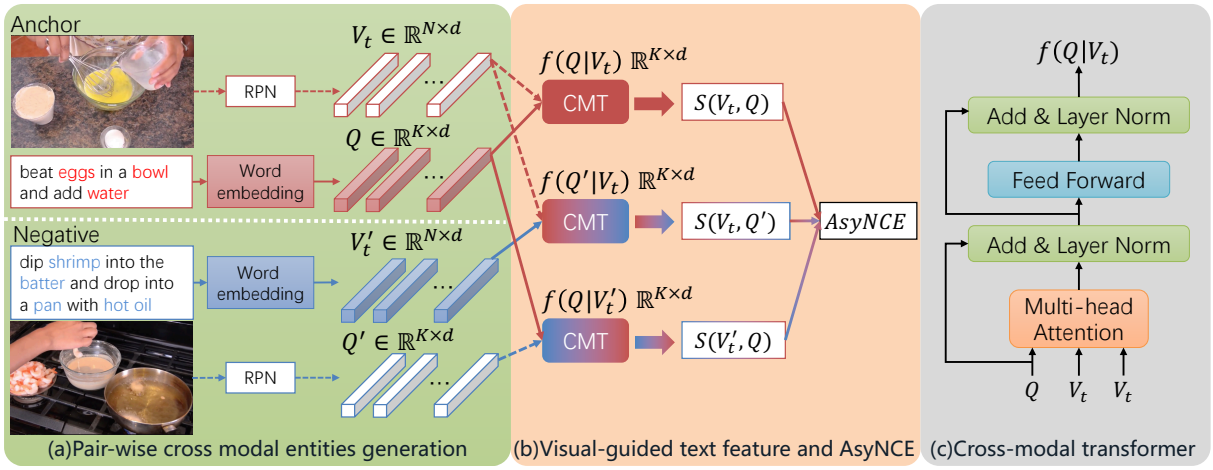

Figure 2: The overview architecture of our method. (a) Region proposal features are produced by RPN and the query entities are encoded into the text features. (b) Visual-guided text features are generated by cross-modal transformer, and AsyNCE loss is employed for training. (c) The structure of cross-modal transformer.

inner-product, multi-layer perceptron (MLP) or cosine similarity for simplicity.

Based on this similarity definition, pair-wise ranking loss can be constructed to encourage aligned positive frame-sentence pairs have higher scores than misaligned ones. The objective function of per frame is formulated as:

$$\begin{aligned}\mathcal{L}^t_{rank} = &\max(0, S(\mathbf{V}_t^{'}, \mathbf{Q}) - S(\mathbf{V}_t, \mathbf{Q}) + \Delta) \\ &+ \max(0, S(\mathbf{V}_t, \mathbf{Q}^{'}) - S(\mathbf{V}_t, \mathbf{Q}) + \Delta),\end{aligned} \tag{2}$$

where Δ is the margin and the negative frame and query are defined as $\mathbf{V}_t^{'}$ and $\mathbf{Q}^{'}$ that are neither paired with $\mathbf{Q}$ nor $\mathbf{V}_t$, respectively. Furthermore, the final MIL ranking loss is the average over all frames, which is formulated as:

$$\mathcal{L}_{rank} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}^t_{rank}. \tag{3}$$

This formulation is commonly used in weakly-supervised video grounding methods [11, 21, 40].

However, since the query object sparsely appears across frames, the positive similarity $S(\mathbf{V}_t, \mathbf{Q})$ could be a false-positive one in Eq. (2). Thus, the loss function will be damaged when it runs into the false-positive frames. To handle this issue, contextual similarity is proposed to reduce the sparsity [21] by normalizing similarities across frames in Eq. (2). Besides, the certainty or importance of each frame is not equal in Eq. (3), making the average loss in Eq. (3) suboptimal. So the weighted loss formulation across frames [40] is employed to figure out the inequality in Eq. (3). Nevertheless, these methods still treat positive and negative pairs equally in coupling ranking loss and have no access to dealing with false-positive frames and negative ones concurrently.

3.2 AsyNCE Loss with Soft Similarity

We propose a novel and flexible AsyNCE loss with soft similarity to address the inequality on frame-level MIL framework in video domain. Concretely, we deal with the positive and negative pairs in a asymmetric way. For positive pairs, soft-frame similarity $S_p(\mathbf{V}, \mathbf{Q})$ is proposed to represent some kind of aggregation over all frames, which can reduce the uncertainty of the position of false-positive frames. While the negative pairs still employ the original frame-entity similarities. Obviously, $S_p(\mathbf{V}, \mathbf{Q})$ can simply formulated as:

$$S_p(\mathbf{V}, \mathbf{Q}) = \frac{1}{T}\sum_{t=1}^{T} S(\mathbf{V}_t, \mathbf{Q}). \tag{4}$$

Here, the uncertainty of positive pair can be softened evenly. Furthermore, considering the different importance among the different frames, a softmax function is employed to highlight the most corresponding frame, which is denoted as

$$S_p(\mathbf{V}, \mathbf{Q}) = \sum_{t=1}^{T}\left(\frac{\exp^{S(\mathbf{V}_t,\mathbf{Q})}}{\sum_{t=1}^{T}\exp^{S(\mathbf{V}_t,\mathbf{Q})}} S(\mathbf{V}_t, \mathbf{Q})\right). \tag{5}$$

Then, AsyNCE loss can be formulated as follows:

$$\mathcal{L}_{NCE} = -\log\frac{\exp^{S_p(\mathbf{V},\mathbf{Q})}}{\exp^{S_p(\mathbf{V},\mathbf{Q})} + \sum_{t=1}^{T}(\exp^{S(\mathbf{V}_t^{'},\mathbf{Q})} + \exp^{S(\mathbf{V}_t,\mathbf{Q}^{'})})}. \tag{6}$$

In addition, all query entities are averaged in Eq. (1). To further alleviate the sparsity that query may appear sparsely across frame, soft-entity similarity is proposed as follows:

$$\bar{S}(\mathbf{V}_t, \mathbf{Q}) = \sum_{k=1}^{K}\left(\frac{\exp^{S(\mathbf{V}_t,\mathbf{q}_k)/\tau}}{\sum_{t=1}^{T}\exp^{S(\mathbf{V}_t,\mathbf{q}_k)/\tau}} S(\mathbf{V}_t, \mathbf{q}_k)\right), \tag{7}$$

where τ is the temperature scale factor. Referring to one query entity $\mathbf{q}_k$, the weight over all frames is computed by softmax. Thus, the normalized similarity emphasizes the most positive frames of $\mathbf{q}_k$, and suppresses the false-positive ones.

Typically, soft-entity similarity is also utilized to locate the hardest negative frame referring to $\mathbf{q}_k$. This procedure of hard sample mining can enlarge the impact of difficult negative pairs. Notably, soft-frame similarity is not employed for the negative pairs, mainly

due to the following reasons. (1) For the NCE loss, many negative samples are needed for training; (2) There is no uncertainty on negative pair, since all the frames are definitely negative; (3) Abundant information of different frames can be explored sufficiently, while soft-frame similarity will (weighted) average the diversity, resulting in less loss for per negative frame.

Consequently, the final loss function of AsyNCE with soft similarity is formulated as follows:

$$\bar{\mathcal{L}}_{NCE} = - \log \frac{e^{(\bar{S}_p(\mathbf{V},\mathbf{Q})-\Delta)/\tau}}{e^{(\bar{S}_p(\mathbf{V},\mathbf{Q})-\Delta)/\tau} + \sum_{t=1}^{T}(e^{\bar{S}(\mathbf{V}'_t,\mathbf{Q})} + e^{\bar{S}(\mathbf{V}_t,\mathbf{Q}'))}}, \tag{8}$$

where Δ and τ represent margin and temperature, respectively. Compared to ranking loss in Eq. (3), AsyNCE loss can disentangle the positive pairs from negative ones. Thus, the contribution of positive and negative samples are independently for the loss function, leading to a flexible framework for adapting MIL mechanism. By leveraging soft-entity similarity, the most positive or negative frame can be located. With soft-frame similarity, the uncertainty of positive pairs is reduced, which can be regarded as a small margin in positive pairs. Therefore, Eq. (8) is more suitable for MIL problem.

Considering a simple formulation of ranking loss, the similarities of the negative and positive pairs are denoted as S_n^t and S_p^t, respectively. This ranking loss function is denoted as

$$\begin{aligned}\bar{\mathcal{L}}_{rank} &= \frac{1}{T}\sum_{t=1}^{T}\max(0, S_n^t - S_p^t + m) \\ &\approx \frac{1}{T}\sum_{t=1}^{T}\log(1 + e^{S_n^t - S_p^t + m}) \\ &= -\frac{1}{T}\log(\prod_{t=1}^{T}\frac{e^{S_p^t - m}}{e^{S_p^t - m} + e^{S_n^t}}),\end{aligned} \tag{9}$$

where the ranking loss function is approximate to a softmax fashion. When $t = 1$ (image grounding), $\bar{\mathcal{L}}_{rank}$ is equivalent to AsyNCE as in Eq. (8). When $t > 1$ (video grounding), the logits are multiplied with each other, resulting in higher cumulative error by false-positive pairs in the ranking loss. While the loss of negative pair is linearly accumulated and the soft similarity (linear) of positive one is used to handle false- positive frames in AsyNCE. Thus, AsyNCE can alleviate the uncertainty of false-positive frames effectively, and the ranking loss can be regarded as a particular case of AsyNCE.

3.3 Visual-guided Text Feature

In this section, we propose a cross-modal transformer block to transfer text feature into image space, leading to the visual-guided text feature. Specifically, the weights in the transformer are denoted as $\mathbf{W}_q \in \mathbb{R}^{d\times d}$, $\mathbf{W}_k \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_v \in \mathbb{R}^{d\times d}$ for Query, Key and Value, respectively. Thus, the cross modal attention is formulated as follows:

$$\mathbf{Z}(\mathbf{Q}|\mathbf{V}_t) = softmax(\frac{\mathbf{Q}\mathbf{W}_q\mathbf{W}_k^T\mathbf{V}_t^T}{\sqrt{d}})\mathbf{V}_t\mathbf{W}_v, \tag{10}$$

where d is the dimension of feature. The attention values between query $\mathbf{Q}$ and image regions $\mathbf{V}_t$ are computed by softmax function. Furthermore, $\mathbf{Z}(\mathbf{Q}|\mathbf{V}_t)$ is the weighted summary of $\mathbf{V}_t\mathbf{W}_v$. By this cross-modal attention, text feature is adapted to image feature space with the same shape of $\mathbf{Q}$.

Then, as shown in Fig. 2(c), the whole cross-modal transformer is computed as follows:

$$\begin{aligned}\mathbf{G}_0 &= \mathbf{Z}(\mathbf{Q}|\mathbf{V}_t) \\ \mathbf{G}_1 &= LN(\mathbf{G}_0 + \mathbf{Q}\mathbf{W}_q) \\ f(\mathbf{Q}|\mathbf{V}_t) &= LN(g(\mathbf{G}_1) + \mathbf{G}_1).\end{aligned} \tag{11}$$

Here, $LN(\cdot)$ is $LayerNorm$ and $g(\cdot)$ is a feed-forward layer. Since query information $\mathbf{Q}\mathbf{W}_q$ is injected to transformer again as in Eq. (11), the final result $f(\mathbf{Q}|\mathbf{V}_t) \in \mathbb{R}^{K\times d}$ can be regarded as purified text features from both two modality, resulting in the visual-guided text feature.

Contrast to the common method of which the similarity $s(\mathbf{v}_t^n, \mathbf{q}_k)$ between text feature and image region is realized by a simple inner product form $\mathbf{q}_k^\mathbf{T}\mathbf{v}_t^n$, the purified text feature $f(\mathbf{Q}|\mathbf{V}_t)$ is employed in similarity computation in our approach, denoted as $s(\mathbf{v}_t^n, \mathbf{q}_k) = f(\mathbf{q}_k|\mathbf{V}_t)^T\mathbf{v}_t^n$. In this way, the similarity with cross-modal transformer is amenable to a meaningful metric of cross-modal features. Finally, the overview architecture of our method is shown in Fig. 2.

4 EXPERIMENTS

In this section, we first compare our approach with some baselines on the YouCook2 dataset. Then, ablation studies are conducted. Finally, we show how well our model generalizes on the commodity WAB dataset and RoboWatch dataset.

4.1 Datasets

YouCook2 [41] is a large-scale instructional video dataset, which contains 2K YouTube cooking videos from 89 recipes. Each recipe includes multiple steps that are temporally annotated as different segments. And each segment is only described by a natural language sentence, without fine-grained bounding box annotations. For video grounding task, [40] provides bounding box annotations for evaluation, where the most frequently appearing objects are annotated as categories at 1 fps. In weakly-supervised setting, only segment-description pairs are available for training, and the annotations in [40] are utilized for evaluation in our experiments.

RoboWatch consists of many instructional videos crawled from YouTube. Each video is annotated with ground-truth temporal steps, and the description of each step is given. Box annotations of the test set of RoboWatch are presented in [9]. Following [21], we evaluate the generalizability on the 225 videos of RoboWatch test set.

Watch and Buy dataset (**WAB**) [1] is a recent video dataset for commodity identification. We conduct a commodity grounding task on the subset of WAB. Specifically, each clip of WAB contains corresponding commodity title and 5 key frames with instance-level bounding boxes. In order to produce the training pairs in the weakly-supervised grounding problem, we ground the title to the correct bounding boxes as video-title pairs for each the video clip. Finally, we collect 5K or 10K aligned video-title pairs without boxes for weakly-supervised training and 5K with box annotations for validation.

[1] https://tianchi.aliyun.com/dataset/dataDetail?dataId=75730

Table 1: Results on YouCook2 with NAFAE features.

Method	Macro		Micro	
	Val	Test	Val	Test
Upper Bound	62.42	62.41	-	-
$DVSA_v$	36.67	36.30	43.62	42.87
$DVSA_f$	36.90	37.55	44.26	44.16
WSVOG	35.69	35.08	43.04	42.42
NAFAE	38.14	38.29	45.04	44.89
AsyNCE	38.43	38.41	45.88	45.29
AsyNCE-CMT	**38.52**	**38.86**	**46.44**	**46.02**

Table 2: Results on YouCook2 with YCBB features.

Method	Macro		Micro	
	Val	Test	Val	Test
Upper Bound	57.77	58.56	-	-
GroundR	19.63	19.94	-	-
$DVSA_f$	30.51	30.80	-	-
WSVOG	30.09	31.99	36.09	35.37
NAFAE	31.12	32.60	37.70	37.14
AsyNCE	33.12	34.27	40.31	39.39
AsyNCE-CMT	**33.41**	**34.43**	**40.57**	**40.28**

4.2 Compared Methods and Metrics

We employ the following weakly-supervised video grounding methods as baselines for comparison. DVSA [11] is a classical ranking-based MIL image grounding method, and we adapt DVSA to a frame-level $DVSA_f$ and video-level $DVSA_v$ for video grounding as in [21]. GroundR [19] is also a image-based baseline by phrase reconstruction. WSVOG [40] presents a weighted frame ranking loss to reduce the uncertainty of false-positive frames for frame-level MIL problem. NAFAE [21] proposes contextual similarity and visual clustering to improve the performance of ranking-based MIL method. Upper bound is computed by regarding all box proposals as the grounded boxes of each query.

Box accuracy is adopted as criteria to evaluate grounding performance [21]. Specifically, top-1 box is chosen as grounded box for each query. And the box accuracy is the ratio of correctly grounded boxes to all ground-truth boxes, where the grounded box that has over 0.5 IoU with the ground-truth annotation is positive. The average of each class accuracy is denoted as macro-accuracy, and micro-accuracy is denoted as the class-agnostic accuracy.

4.3 Experimental Details

Feature. These video grounding methods mainly rely on pretrained features and the quality of region proposals. Thus, two kinds of feature are utilized for evaluation in YouCook2. (1) NAFAE features: Following [21], region proposal features are produced by RPN [18] pre-trained on Visual Gnome, and 4096-d features are extracted from the last FC layer of VGG [22] followed by the RoI pooling. Then, the feature dimension is reduced to 512 by FC layer and TanH activation function. For text feature, each word is embedded with 200-d GloVe [15] feature, and further encoded as a 512-d feature. (2)YCBB features: As in WSVOG [40], 2048-d region proposal features are extracted from Faster-RCNN with ResNet-101 [7] pre-trained on MSCOCO, and random word embedding features are utilized for text embedding. Finally, 512-d features are generated by FC layer for fair comparison.

Implementation. For each video segment, only 5 frames are uniformly sampled for training and top-20 proposals are selected. Stochastic gradient descent (SGD) with Nesterov momentum is used as optimizer. We set learning rate to 0.001 and momentum to 0.9. The hyper-parameter τ is set to 0.07 as in [4] and margin Δ is set to 0.05. One layer and 4 heads are utilized for the construction of multi-head transformer. Moreover, 2 aligned segment-sentence pairs are included in one batch, and thus 1 positive and 10 negative samples are generated for AsyNCE loss. For testing, all frames with ground-truth annotations are grounded for evaluation.

Table 3: The effect of different components on YouCook2 with NAFAE and YCBB features.

Components			NAFAE / Test		YCBB / Test	
CMT	Entity	Frame	Macro	Micro	Macro	Micro
×	mean	mean	38.28	44.76	30.03	35.88
×	mean	soft	38.08	44.73	31.14	37.12
×	soft	mean	38.01	45.38	33.82	38.45
×	soft	soft	38.41	45.29	34.27	39.39
✓	mean	mean	38.32	45.03	32.83	37.14
✓	mean	soft	38.77	45.36	33.40	37.91
✓	soft	mean	38.74	45.69	33.73	38.58
✓	soft	soft	**38.86**	**46.02**	**34.43**	**40.28**

4.4 Grounding Performance

The exhaustive grounding results on YouCook2 with NAFAE features are illustrated in Table 1. Since there is a certain fluctuation on the training performance of NAFAE, it is reimplemented with the same random seed with our method in the following experiments for fair comparison. And soft similarities are employed to train our models in Table 1. In general, ranking-based methods (DVSA, WSVOG and NAFAE) have achieved comparable performance. Typically, WSVOG obtains worse results than DVSA and NAFAE, due to the sensitivity of the hyper-parameter and features. NAFAE utilizes contextual similarity and visual clustering to obtain promising results on their own features. Our AsyNCE method can achieve a little better performance than NAFAE, which verifies the superiority of AsyNCE. By adding CMT, AsyNCE-CMT outperform NAFAE 2.5% on micro-test accuracy, indicating the effectiveness of AsyNCE-CMT.

4.5 The Robustness of Features

To investigate the robustness of different proposals and features, we conduct experiments on YCBB features. As reported in Table 2, the upper bound and total results of YCBB features are lower than those of NAFAE in Table 1. These results verify that grounding

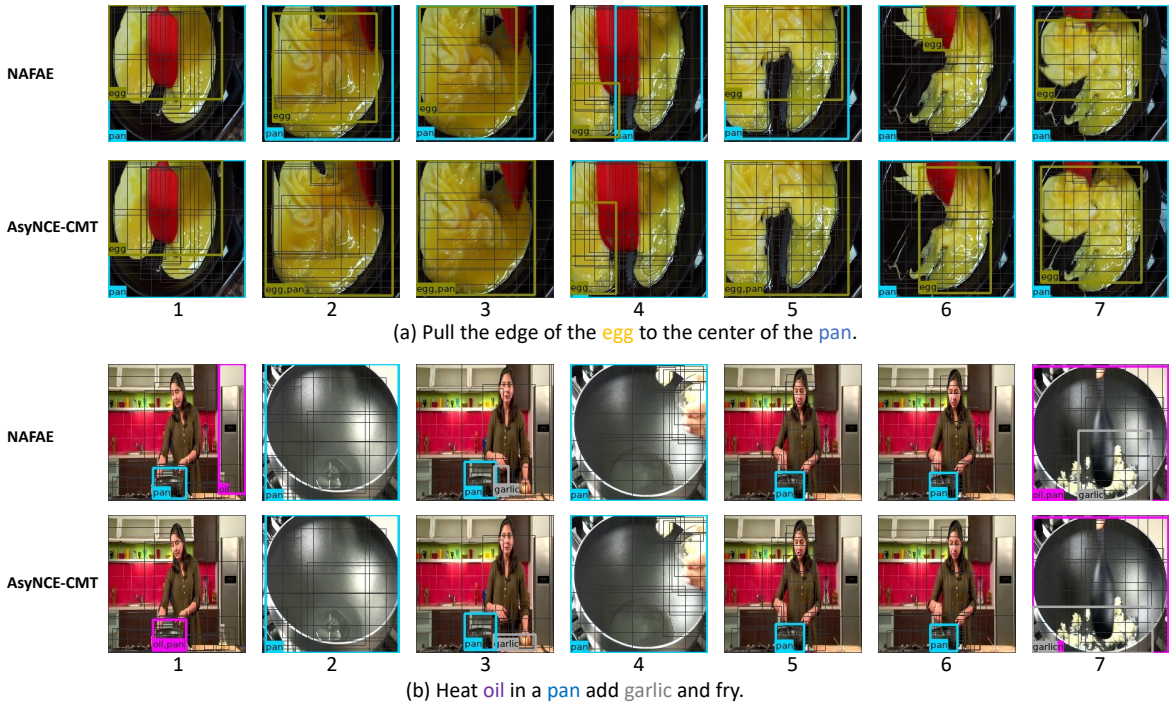

Figure 3: Visualization of location results of NAFAE and AsyNCE-CMT on two videos of YouCook2. Black boxes represent region proposals. And Boxes with different colours represent the grounding results.

performance mainly rely on features and region proposals. Meanwhile, WSVOG is not robust and obtains 3.9% improvements over DVSA$_f$ on macro-test accuracy, opposite to Table 1. While our approach with the same hyper-parameters still surpasses the compared methods on YCBB features greatly. Notably, AsyNCE obtains 6% improvements on micro-test accuracy over NAFAE, and AsyNCE-CMT gets 8.5% improvements further, demonstrating the robustness and superiority of our method.

4.6 The Effect of Soft Similarity

The effects of CMT and different components of similarities in AsyNCE are fully analysed, and the accuracy results on YouCook2 with two features are listed in Table 3. Generally, the models with CMT is mostly superior to the ones without CMT, especially for the ones without soft-entity or soft-frame similarity. These results demonstrate that CMT enables effective communication between two modalities and meaningful similarity construction. Additionally, soft-entity and soft-frame can obtain a little better results on NAFAE features, while greatly improvements on YCBB features, indicating that our method is flexible to similarity plug-in.

4.7 The Effect of Margin

We conduct the parameter sensitivity analysis on margin Δ, and the marco accuracy with different margins on validation set is illustrated in Fig.4. In order to comprehensive analysis the effect of margin, we evaluate our method with different similarity plug-ins. The method is denoted as "CMT/Only-entity-frame", where

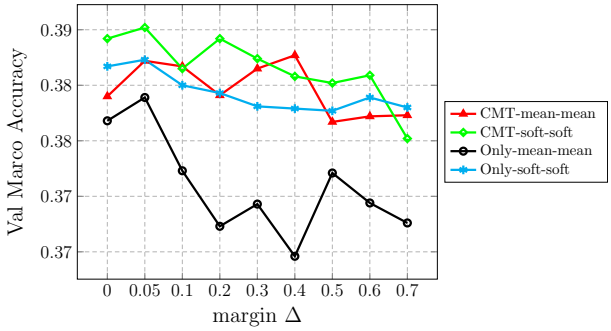

Figure 4: The parameter sensitivity on margin Δ in our method on YouCook2 with NAFAE features.

"only" means method without cross-modal transformer. It is shown that the accuracies get little worse as margin increase, but they are not much sensitive to margin. Thus, we set margin to 0.05 in our method. Moreover, the method with CMT or soft-similarity is better than the one without any of them, which is consistence to Table 3.

4.8 Performance on RoboWatch

We also evaluate the generalization on RoboWatch following [21]. Different from [21] that employs LSTM to encode query, we still use word embedding for the query representation. Notably, the model is trained only on YouCook2 and directly evaluated on RoboWatch.

(a) Pink short sleeve T-shirt;solid color short skirt;clashing color pants;children pants.

(b) Short sleeve cardigan sweater;green short sleeve T-shirt;pink pants;straight pants.

Figure 5: Visualization of location results of AsyNCE-CMT on WAB. Red and green boxes represent the ground-truth and grounding results, respectively. (a) and (b) represent the successful cases and the failed ones.

Table 4: The macro accuracy on test set of RoboWatch with NAFAE features.

Method	RoboWatch
$DVSA_f$	28.25
RA-MIL	19.80
NAFAE	28.72
AsyNCE	28.79
AsyNCE-CMT	**29.20**

Table 5: Results on WAB with clothing features.

Method	WAB-5K	WAB-10K
Supervised CMS	85.30	88.10
Random	48.40	48.40
NAFAE	69.10	74.52
Ranking	79.87	83.51
AsyNCE	81.90	86.05
AsyNCE-CMT	**82.04**	**86.25**

The grounding results are reported in Table 4. We observe that AsyNCE is comparable with NAFAE and the accuracy of AsyNCE-CMT gets a little better. These results imply that AsyNCE and cross-modal transformer have good generalization as ranking loss.

4.9 Performance on WAB

We conduct experiments on real commodity video dataset. Specific clothing detector and feature extractor are employed to produce region features, and LSTM is utilized to encode the title of commodity. Then, we choose top-5 regions and one title as query for grounding task. Since only one entity is evaluated, the results of AsyNCE without soft-entity similarity are reported in Table 5. Typically, NAFAE achieves poor performance on WAB due to the complex contextual similarity and visual clustering module with sensitive parameters. Thus, we conduct a naive ranking method as "Ranking" in Table 5. Our method outperforms ranking-based ones remarkably, verifying the good transportability. Meanwhile, weakly-supervised grounding method can achieve comparable results with supervised cross-modal search. With the scale of dataset increase, the accuracy further increases, demonstrating the great potential of weakly-supervised learning. Therefore, the weakly-supervised video title grounding with AsyNCE has been employed on the live-stream in E-commerce, resulting in a prominent performance of video commodity localization.

4.10 Visualization

The grounding visualization of NAFAE and our method on YouCook2 is shown in Fig. 3. In general, both two grounding methods can achieve comparable results on some frames, but it is clear that our proposed can obtain better location results than NAFAE. For example, as in Fig. 3(a), NAFAE obtains a half "pan" in the 4th frame and wrong "egg" in the 6th frame. While our method obtains more complete grounding results in the 4th, 6th and 7th frames. As in Fig. 3(b), "oil" is grounded at the refrigerator in the first frame by NAFAE, while our method can ground it into the pan.

Then, we also illustrate the grounding visualization of our method on WAB dataset. As shown in Fig. 5(a), even though there are multiple person and clothe candidates, the right clothes can be selected referring to the query. Moreover, some failure case are shown in Fig. 5(b). Notably, even if the grounding results are not correct, the wrong grounding objects are close to the ground-truth. For example, both ground-truth and grounded object are pants in the 4th frame. These results demonstrate that the weakly-supervised grounding method can coarsely locate the clothes in E-commerce, but the fine-grained grounding results are difficult to achieve.

5 CONCLUSION

In this paper, we explore AsyNCE loss and cross-modal transformer block for weakly-supervised grounding. By leveraging flexible AsyNCE loss, the impact of false-positive frame can be reduced, and the effect of negative pairs can be exploited sufficiently. In addition, cross-modal transformer enables effective communication between two modalities and meaningful similarity construction. Experimental results on two instructional datasets and one real commodity dataset demonstrate the superiority, transportability and generalizability of our methods to the conventional ranking-based methods. Moreover, the proposed method has been employed on the live-stream video commodity localization in E-commerce.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China under Grants 91838303.

REFERENCES

[1] Junwen Chen, Wentao Bao, and Yu Kong. 2020. Activity-driven Weakly-Supervised Spatio-Temporal Grounding from Untrimmed Videos. In *MM*. 3789–3797.

[2] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally Grounding Natural Sentence in Video. In *EMNLP*. 162–171.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, Vol. 119. 1597–1607.

[4] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. *CoRR* abs/2003.04297 (2020).

[5] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. 2019. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video. In *ACL*.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[8] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing Moments in Video with Natural Language. In *ICCV*. 5804–5813.

[9] De-An Huang, Shyamal Buch, Lucio M. Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding "It": Weakly-Supervised Reference-Aware Visual Grounding in Instructional Videos. In *CVPR*. 5948–5957.

[10] Chenchen Jing, Yuwei Wu, Mingtao Pei, Yao Hu, Yunde Jia, and Qi Wu. 2020. Visual-Semantic Graph Matching for Visual Grounding. In *MM*. 4041–4050.

[11] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 664–676.

[12] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring Expression Generation and Comprehension via Attributes. In *ICCV*. 4866–4874.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*. 13–23.

[14] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *CVPR*. 9876–9886.

[15] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. 1532–1543.

[16] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *Int. J. Comput. Vis.* 123, 1 (2017), 74–93.

[17] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained Iterative Attention Network for Temporal Language Localization in Videos. In *MM*. 4280–4288.

[18] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.

[19] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of Textual Phrases in Images by Reconstruction. In *ECCV*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). 817–834.

[20] Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video Object Grounding Using Semantic Roles in Language Description. In *CVPR*. 10414–10424.

[21] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. 2019. Not All Frames Are Equal: Weakly-Supervised Video Grounding With Contextual Similarity and Visual Clustering Losses. In *CVPR*. 10444–10452.

[22] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[23] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*. 1–12.

[24] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Learning Video Representations using Contrastive Bidirectional Transformer. *CoRR* abs/1906.05743 (2019).

[25] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *ICCV*. 7463–7472.

[26] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*. 5099–5110.

[27] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL*. 6558–6569.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[29] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. 2020. Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. *CoRR* abs/2007.01951 (2020). https://arxiv.org/abs/2007.01951

[30] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. In *MM*. 1283–1291.

[31] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual Relation Grounding in Videos. In *ECCV*, Vol. 12351. 447–464.

[32] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. In *SIGIR*. 1339–1348.

[33] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-Supervised Video Object Grounding by Exploring Spatio-Temporal Contexts. In *MM*. 1939–1947.

[34] Haonan Yu and Jeffrey Mark Siskind. 2017. Sentence Directed Video Object Codiscovery. *Int. J. Comput. Vis.* 124, 3 (2017), 312–334.

[35] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *CVPR*. 1307–1315.

[36] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*. 12870–12877.

[37] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jieming Zhu, and Xiuqiang He. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. In *NeurIPS*.

[38] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *CVPR*. 10665–10674.

[39] Chen Zheng, Quan Guo, and Parisa Kordjamshidi. 2020. Cross-Modality Relevance for Reasoning on Language and Vision. In *ACL*. 7642–7651.

[40] Luowei Zhou, Nathan Louis, and Jason J. Corso. 2018. Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction. In *BMVC*. 50.

[41] Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI*. 7590–7598.

[42] Linchao Zhu and Yi Yang. 2020. ActBERT: Learning Global-Local Video-Text Representations. In *CVPR*. 8743–8752.