wMAN: WEAKLY-SUPERVISED MOMENT ALIGNMENT NETWORK FOR TEXT-BASED VIDEO SEGMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Given a video and a sentence, the goal of weakly-supervised video moment retrieval is to locate the video segment which is described by the sentence without having access to temporal annotations during training. Instead, a model must learn how to identify the correct segment (*i.e.* moment) when only being provided with video-sentence pairs. Thus, an inherent challenge is automatically inferring the latent correspondence between visual and language representations. To facilitate this alignment, we propose our Weakly-supervised Moment Alignment Network (wMAN) which exploits a multi-level co-attention mechanism to learn richer multimodal representations. The aforementioned mechanism is comprised of a Frame-By-Word interaction module as well as a novel Word-Conditioned Visual Graph (WCVG). Our approach also incorporates a novel application of positional encodings, commonly used in Transformers, to learn visual-semantic representations that contain contextual information of their relative positions in the temporal sequence through iterative message-passing. Comprehensive experiments on the DiDeMo and Charades-STA datasets demonstrate the effectiveness of our learned representations: our combined wMAN model not only outperforms the state-of-the-art weakly-supervised method by a significant margin but also does better than strongly-supervised state-of-the-art methods on some metrics.

1 INTRODUCTION

Video understanding has been a mainstay of artificial intelligence research. Recent work has sought to better reason about videos by learning more effective spatio-temporal representations (Tran et al., 2015; Carreira & Zisserman, 2017). The *video moment retrieval* task, also known as *text-to-clip retrieval*, combines language and video understanding to find activities described by a natural language sentence. The main objective of the task is to identify the video segment within a longer video that is most relevant to a sentence. This requires a model to learn the mapping of correspondences (alignment) between the visual and natural language modalities.

In the strongly-supervised setting, existing methods (Hendricks et al., 2017; Chen et al., 2018; Ghosh et al., 2019) generally learn joint visual-semantic representations by projecting video and language representations into a common embedding space and leverage provided temporal annotations to learn regressive functions (Gao et al., 2017) for localization. However, such temporal annotations are often ambiguous and expensive to collect. Mithun et al. (2019) seeks to circumvent these problems by proposing to address this task in the weakly-supervised setting where only full video-sentence pairs are provided as weak labels. However, the lack of temporal annotations renders the aforementioned approaches infeasible. In their approach (Figure 1a), Mithun et al. (2019) proposes a Text-Guided Attention (TGA) mechanism to attend on segment-level features w.r.t. the sentence-level representations. However, such an approach treats the segment-level visual representations as independent inputs and ignores the contextual information derived from other segments in the video. More importantly, it does not exploit the fine-grained semantics of each word in the sentence. Consequently, existing methods are not able to reason about the latent alignment between the visual and language representations comprehensively.

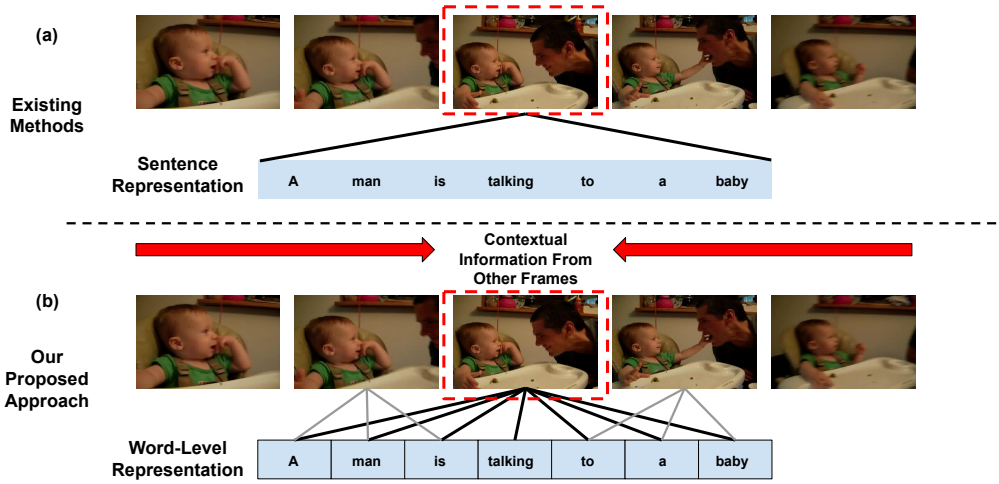

Figure 1: Given a video and a sentence, our aim is to retrieve the most relevant segment (the red bounding box in this example). Existing methods consider video frames as independent inputs and ignore the contextual information derived from other frames in the video. They compute a similarity score between the segment and the entire sentence to determine their relevance to each other. In contrast, our proposed approach aggregates contextual information from all the frames using graph propagation and leverages fine-grained frame-by-word interactions for more accurate retrieval. (Only some interactions are shown to prevent overcrowding the figure.)

In this paper, we take another step towards addressing the limitations of current weakly-supervised video moment retrieval methods by exploiting the fine-grained temporal and visual relevance of each video frame to each word (Figure 1b). Our approach is built on two core insights: 1) The temporal occurrence of frames or segments in a video provides vital visual information required to reason about the presence of an event; 2) The semantics of the query are integral to reasoning about the relationships between entities in the video. With this in mind, we propose our Weakly-Supervised Moment Alignment Network (wMAN). An illustrative overview of our model is shown in Figure 2. The key component of wMAN is a multi-level co-attention mechanism that is encapsulated by a Frame-by-Word (FBW) interaction module as well as a Word-Conditioned Visual Graph (WCVG). To begin, we exploit the similarity scores of all possible pairs of visual frame and word features to create frame-specific sentence representations and word-specific video representations. The intuition is that frames relevant to a word should have a higher measure of similarity as compared to the rest. The word representations are updated by their word-specific video representations to create visual-semantic representations. Then a graph (WCVG) is built upon the frame and visual-semantic representations as nodes and introduces another level of attention between them. During the message-passing process, the frame nodes are iteratively updated with relational information from the visual-semantic nodes to create the final temporally-aware multimodal representations. The contribution of each visual-semantic node to a frame node is dynamically weighted based on their similarity. To learn such representations, wMAN also incorporates positional encodings (Vaswani et al., 2017) into the visual representations to integrate contextual information about their relative positions. Such contextual information encourage the learning of temporally-aware multimodal representations.

To learn these representations, we use a Multiple Instance Learning (MIL) framework that is similar in nature to the Stacked Cross Attention Network (SCAN) model (Lee et al., 2018). The SCAN model leverages image region-by-word interactions to learn better representations for image-text matching. In addition, the WCVG module draws inspiration from the Language-Conditioned Graph Network (LCGN) by Hu et al. (2019) which seeks to create context-aware object features in an image. However, the LCGN model works with sentence-level representations, which does not account for the semantics of each word to each visual node comprehensively. wMAN also distinguishes itself from the above-mentioned models by extracting temporally-aware multimodal representations from videos and their corresponding descriptions, whereas SCAN and LCGN only work on images.

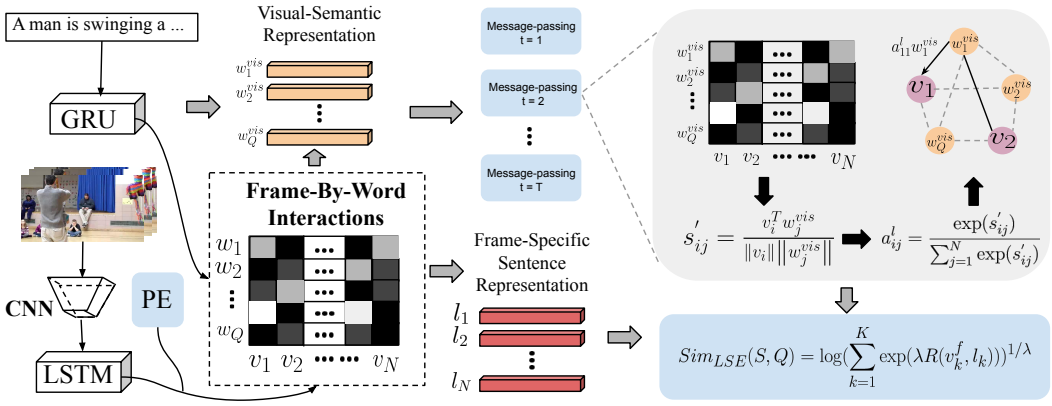

Figure 2: An overview of our combined wMAN model which is trained end-to-end. We use the outputs of the GRU as word representations where its inputs are word embeddings. The visual representations are the outputs of the LSTM unit where its inputs are the extracted features from a pretrained CNN. The visual representations are concatenated with positional encodings to integrate contextual information about their relative positions in the sequence. Our model consists of a two-stage multimodal interaction mechanism - Frame-By-Word Interactions and the WCVG.

The contributions of our paper are summarized below:

- We propose a simple yet intuitive MIL approach for weakly-supervised video moment retrieval from language queries by exploiting fine-grained frame-by-word alignment.
- Our novel Word-Conditioned Visual Graph learns richer visual-semantic context through a multi-level co-attention mechanism.
- We introduce a novel application of positional embeddings in video representations to learn temporally-aware multimodal representations.

To demonstrate the effectiveness of our learned temporally-aware multimodal representations, we perform extensive experiments over two datasets, Didemo (Hendricks et al., 2017) and Charades-STA (Gao et al., 2017), where we outperform the state-of-the-art weakly supervised model by a significant margin and strongly-supervised state-of-the-art models on some metrics.

2 Related Work

Representation Learning. Learning effective representations for application tasks of various modalities (Devlin et al., 2018; Yang et al., 2019; Dai et al., 2019; Tran et al., 2015; Carreira & Zisserman, 2017; He et al., 2016; Lu et al., 2019) has been shown to be integral to achieving good performance on any machine learning tasks.

Visual Phrase Grounding. The task of visual phrase grounding is to locate spatial and temporal bounding box regions described by the given phrase expression in images and videos respectively. A common approach for image-text retrieval and alignment is to learn joint visual-semantic representations and a distance metric (Chen et al., 2017; Liu et al., 2017; Faghri et al., 2018; Nam et al., 2017; Karpathy & Fei-Fei, 2015) to measure the similarity between the image region and language modalities. Plummer et al. (2018) also propose to differentiate sentences into semantically distinct subspaces to learn richer visual-semantic representations.

Video Moment Retrieval. Most of the recent works in video moment retrieval based on natural language queries (Hendricks et al., 2017; Ghosh et al., 2019; Xu et al., 2019; Zhang et al., 2019; Chen et al., 2018; Yuan et al., 2019; Chen & Jiang, 2019; Chen et al., 2019; Ge et al., 2019) are in the strongly-supervised setting, where the provided temporal annotations can be used to improve the alignment between the visual and language modalities. Among them, the Moment Alignment Network (MAN) introduced by Zhang et al. (2019) utilizes a structured graph network to model temporal relationships between candidate moments, but one of the distinguishing factors with our

wMAN is that our iterative message-passing process is conditioned on the multimodal interactions between frame and word representations. The TGN (Chen et al., 2018) model bears some resemblance to ours by leveraging frame-by-word interactions to improve performance. However, it only uses a single level of attention which is not able to infer the correspondence between the visual and language modalities comprehensively. To address the closely-related task of temporal activity detection in videos, a general pipeline of proposal and classification is adopted by various temporal activity detection models (Xu et al., 2017; Zhao et al., 2017; Shou et al., 2016) with the temporal proposals learnt by temporal coordinate regression.

3 WEAKLY-SUPERVISED MOMENT ALIGNMENT NETWORK

In the video moment retrieval task, given a ground truth video-sentence pair, the goal is to retrieve the most relevant video moment related to the description. The weakly-supervised version of this task we address can be formulated under the multiple instance learning (MIL) paradigm. When training using MIL, one receives a bag of items, where the bag is labeled as a positive if at least one item in the bag is a positive, and is labeled as a negative otherwise. In weakly-supervised moment retrieval, we are provided with a video-sentence pair (*i.e.*, a bag) and the video segments are the items that we must learn to correctly label as relevant to the sentence (*i.e.*, positive) or not. Following Mithun et al. (2019), we assume sentences are only associated with their ground truth video, and any other videos are negative examples. To build a good video-sentence representation, we introduce our Weakly-Supervised Moment Alignment Network (wMAN), which learns context-aware visual-semantic representations from fine-grained frame-by-word interactions. As seen in Figure 2, our network has two major components - (1) representation learning constructed from the Frame-By-Word attention and Positional Embeddings (Vaswani et al., 2017), described in Section 3.1, and (2) a Word-Conditioned Visual Graph where we update video segment representations based on context from the rest of the video, described in Section 3.2. These learned video segment representations are used to determine their relevance to their corresponding attended sentence representations using a LogSumExp (LSE) pooling similarity metric, described in Section 3.3.

3.1 LEARNING TIGHTLY COUPLED MULTIMODAL REPRESENTATIONS

In this section we discuss our initial video and sentence representations which are updated with contextual information in Section 3.2. Each word in an input sentence is encoded using GloVe embeddings (Pennington et al., 2014) and then fed into a Gated Recurrent Unit (GRU) (Cho et al., 2014). The output of this GRU is denoted as $W = \{w_1, w_2, \cdots, w_Q\}$ where Q is the number of words in the sentence. Each frame in the input video is encoded using a pretrained Convolutional Neural Network (CNN). In the case of a 3D CNN this actually corresponds to a small chunk of sequential frames, but we shall refer to this as a frame representation throughout this paper for simplicity. To capture long-range dependencies, we feed the frame features into a Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997). The latent hidden state output from the LSTM are concatenated with positional encodings (described below) to form the initial video representations, denoted as $V = \{v_1, v_2, ..., v_N\}$ where N is the number of frame features for video V.

Positional Encodings (PE). To provide some notion of the relative position of each frame we include the PE features which have been used in language tasks like learning language representations using BERT (Devlin et al., 2018; Vaswani et al., 2017). These PE features can be thought of similar to the temporal endpoint features (TEF) used in prior work for strongly supervised moment retrieval task (*e.g.*, Hendricks et al. (2017)), but the PE features provide information about the temporal position of each frame rather than the rough position in segment level. For the desired PE features of dimension d, let pos indicates the temporal position of each frame, i is the index of the feature being encoded, and M is a scalar constant, then the PE features are defined as:

$$PE_{pos,i} = \begin{cases} \sin(pos/M^{i/d}) & \text{if } i \text{ is even} \\ \cos(pos/M^{i/d}) & \text{otherwise.} \end{cases} \tag{1}$$

Through experiments, the hyper-parameter $M = 10{,}000$ works well for all videos. These PE features are concatenated with the LSTM encoded frame features at corresponding frame position before going to the cross-modal interaction layers.

3.1.1 FRAME-BY-WORD INTERACTION

Rather than relating a sentence-level representation with each frame as done in prior work (Mithun et al., 2019), we aggregate similarity scores between all frame and word combinations from the input video and sentence. These Frame-By-Word (FBW) similarity scores are used to compute attention weights to identify which frame and word combinations are important for retrieving the correct video segment. More formally, for N video frames and Q words in the input, we compute:

$$s_{ij} = \frac{v_i^T w_j}{\|v_i\| \, \|w_j\|} \text{ where } i \in [1, N] \text{ and } j \in [1, Q]. \tag{2}$$

Note that v now represents the concatenation of the video frame features and the PE features.

Frame-Specific Sentence Representations. We obtain the normalized relevance of each word w.r.t. to each frame from the FBW similarity matrix, and use it to compute attention for each word:

$$a_{ij} = \frac{\exp(s_{ij})}{\sum_{j=1}^{Q} \exp(s_{ij})}. \tag{3}$$

Using the above-mentioned attention weights, a weighted combination of all the words are created, with correlated words to the frame gaining high attention. Intuitively, a word-frame pair should have a high similarity score if the frame contains a reference to the word. Then the frame-specific sentence representation emphasizes words relevant to the frame and is defined as:

$$l_i = \sum_{j=1}^{Q} a_{ij} w_j. \tag{4}$$

Note that these frame-specific sentence representations don't participate in the iterative message-passing process (Section 3.2). Instead, they are used to infer the final similarity score between a video segment and the query (Section 3.3).

Word-Specific Video Representations. To determine the normalized relevance of each frame w.r.t. to each word, we compute the attention weights of each frame:

$$a_{ij}^{'} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}. \tag{5}$$

Similarly, we attend to the visual frame features with respect to each word by creating a weighted combination of visual frame features determined by the relevance of each frame to the word. The formulation of each word-specific video-representation is defined as:

$$f_j = \sum_{i=1}^{N} a_{ij}^{'} v_i. \tag{6}$$

These word-specific video representations are used in our Word-Conditioned Visual Graph, which we will discuss in the next section.

3.2 WORD-CONDITIONED VISUAL GRAPH NETWORK

Given the sets of visual representations, word representations and their corresponding word-specific video representations, the WCVG aims to learn temporally-aware multimodal representations by integrating visual-semantic and contextual information into the visual features. To begin, the word representations are updated with their corresponding video representations to create a new visual-semantic representation w_j^{vis} by concatenating each word w_j and the word-specific video representation f_j. Intuitively, the visual-semantic representations not only contain the semantic context of each word but also a summary of the video with respect to each word. A fully connected graph is then constructed with the visual features v_i and the embedded attention of visual-semantic representations w_j^{vis} as nodes.

Iterative Word-Conditioned Message-Passing The iterative message-passing process introduces a second round of FBW interaction similar to that in Section 3.1.1 to infer the latent temporal correspondence between each frame v_i and visual-semantic representation w_j^{vis}. The goal is to update the

representation of each frame v_i with the video context information from each word-specific video representation w_j^{vis} . To realize this, we first learn a projection W_1 followed by a ReLU of w_j^{vis} to obtain a new word representation to compute a new similarity matrix s'_{ij} on every message-passing iteration, namely, we obtain a replacement for w_j in Eq. (2) via $w'_j = ReLU(W_1(w_j^{vis}))$.

Updates of Visual Representations During the update process, each visual-semantic node sends its message (represented by its representation) to each visual node weighted by their edge weights. The representations of the visual nodes at the t-th iteration are updated by summing up the incoming messages as follows:

$$v_i^t = W_2(concat\{v_i^{t-1}; \sum_{j=1}^{Q} a_{ij}^l w'_j\}), \tag{7}$$

where a_{ij} is obtained by applying Eq. (3) to the newly computed FBW similarity matrix s'_{ij}, and W_2 is a learned projection to make v_i^t the same dimensions as the frame-specific sentence representation l_i (refer to Eq. (4)) which are finally used to compute a sentence-segment similarity score.

3.3 Multimodal Similarity Inference

The final updated visual representations $V^T = \{v_1^T, v_2^T, \cdots, v_V^T\})$ are used to compute the relevance of each frame to its attended sentence-representations. A segment is defined as any arbitrary continuous sequence of visual features. We denote a segment as $S = \{v_1^T, \cdots, v_K^T\}$ where K is the number of frame features contained within the segment S. We adopt the LogSumExp (LSE) pooling similarity metric used in SCAN, to determine the relevance each proposal segment has to the query:

$$Sim_{LSE}(S, Q) = \log(\sum_{k=1}^{K} \exp(\lambda R(v_k^f, l_k)))^{1/\lambda} \text{ where } R(v_k, l_k) = \frac{{v_k}^T l_k}{\|v_k\| \, \|l_k\|}. \tag{8}$$

λ is a hyperparameter that weighs the relevance of the most salient parts of the video segment to the corresponding frame-specific sentence representations. Finally, following Mithun et al. (2019), given a triplet (X^+, Y^+, Y^-), where (X^+, Y^+) is a positive pair and (X^+, Y^-) a negative pair, we use a margin-based ranking loss L_T to train our model which ensures the positive pair's similarity score is better than the negative pair's by at least a margin. Our model's loss is then defined as:

$$L_{total} = \sum_{(V^+, Q^+)} \{\sum_{Q^-} L_T(V^+, Q^+, Q^-) + \sum_{V^-} L_T(Q^+, V^+, V^-)\}. \tag{9}$$

Sim_{LSE} is used as the similarity metric between positive and negative pairs. During test-time, the pooled similarity scores Sim_{LSE} will also be used to rank the candidate temporal segments generated by sliding windows, and the top scoring segments will the localized segments corresponding to the input query sentence.

4 Experiments

We evaluate the capability of wMAN to accurately localize video moments based on natural language queries without temporal annotations on two datasets - DiDeMo and Charades-STA. On the DiDeMo dataset, we adopt the mean Intersection-Over-Union (IOU) and Recall@N at IOU threshold = θ. Recall@N represents the percentage of the test sliding window samples which have a overlap of at least θ with the ground-truth segments. mIOU is the average IOU with the ground-truth segments for the highest ranking segment to each query input. On the Charades-STA dataset, only the Recall@N metric is used for evaluation.

4.1 Datasets

Charades-STA The Charades-STA dataset is built upon the original Charades [Sigurdsson et al. (2016)] dataset which contains video-level paragraph descriptions and temporal annotations for activities. Charades-STA is created by breaking down the paragraphs to generate sentence-level annotations and aligning the sentences with corresponding video segments. In total, it contains 12,408

Table 1: Moment retrieval performance comparison on the Charades-STA test set. (a) contains representative results of strongly-supervised methods reported in prior works while (b) reports the performance of weakly-supervised methods including our approach.

	Training	iou = 0.3			iou = 0.5			iou = 0.7		
Method	Supervision	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
(a) CTRL (Gao et al., 2017)	Strong	-	-	-	23.63	58.92	-	8.89	29.52	-
MLVI (Xu et al., 2019)	Strong	54.7	95.6	99.2	35.6	79.4	93.9	15.8	45.4	62.2
MAN (Zhang et al., 2019)	Strong	-	-	-	46.53	86.23	-	22.72	53.72	-
(b) TGA (Mithun et al., 2019)	Weak	29.68	83.87	98.41	17.04	58.17	83.44	6.93	26.80	44.06
wMAN (ours)	Weak	**48.04**	**89.01**	**99.57**	**31.74**	**72.17**	**86.58**	**13.71**	**37.58**	**45.16**
Upper Bound	-	-	-	99.84	-	-	88.17	-	-	46.80

and 3,720 query-moment pairs in the training and test sets respectively. For fair comparison with the weakly model TGA (Mithun et al., 2019), we use the same non-overlapping sliding windows of sizes 128 and 256 frames to generate candidate temporal segments.

DiDeMo The videos in the Distinct Describable Moments (DiDeMo) dataset are collected from Flickr. The training, validation and test sets contain 8395, 1065 and 1004 videos respectively. Each query contains the temporal annotations from at least 4 different annotators. Each video is limited to a maximum duration of 30 seconds and equally divided into six segments with five seconds each. With the five-second segment as basic temporal unit, there are 21 possible candidate temporal segments for each video. These 21 segments will used to compute the similarities with the input query and the top scored segment will be returned as the localization result.

4.2 Implementation Details

For fair comparison, we utilize the same input features as the state-of-the-art method (Mithun et al., 2019). Specifically, the word representations are initialized with GloVe embeddings and fine-tuned during the training process. For the experiments on DiDeMo, we use the provided mean-pooled visual frame and optical flow features. The visual frame features are extracted from the fc7 layer of VGG-16 [Simonyan & Zisserman (2014)] pretrained on ImageNet [Deng et al. (2009)]. The input visual features for our experiments on Charades-STA are C3D [Tran et al. (2015)] features. We adopt an initial learning rate of $1e-5$ and a margin$= 0.5$ used in our model's triplet loss (Eq. 9). In addition, we use three iterations for the message-passing process. Our model is trained end-to-end using the ADAM optimizer.

4.3 Results

4.3.1 Charades-STA

The results in Table 1 show that our full model outperforms the TGA model by a significant margin on all metrics. In particular, the Recall@1 accuracy when IOU = 0.7 obtained by our model is almost doubled that of TGA. It is notable that we observe a consistent trend of the Recall@1 accuracies improving the most across all IOU values. This not only demonstrates the importance of richer joint visual-semantic representations for accurate localization but also the superior capability of our model to learn them. Our model also performs comparably to the strongly-supervised MAN model on several metrics.

To better understand the contributions of each component of our model, we present a comprehensive set of ablation experiments in Table 2. Note that our combined wMAN model is comprised of the FBW and WCVG components as well as the incorporation of PEs. The results obtained by our FBW variant demonstrate that capturing fine-grained frame-by-word interactions is essential to inferring the latent temporal alignment between these two modalities. More importantly, the results in the second row (FBW-WCVG) show that the second stage of multimodal attention, introduced by the WCVG module, encourages the augmented learning of intermodal relationships. Finally, we also observe that incorporating positional encodings into the visual representations (FBW-WCVG + PE)

Table 2: Charades-STA ablation experiment results on a held-out validation set.

	iou = 0.3			iou = 0.5			iou = 0.7		
Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
FBW	41.41	**93.79**	**99.23**	26.91	**72.19**	85.97	10.83	34.85	45.20
FBW-WCVG	43.99	90.85	99.19	28.24	70.70	86.14	11.64	34.85	45.20
FBW-WCVG + TEF	43.99	88.03	98.99	28.01	69.19	86.01	11.20	35.29	44.45
FBW-WCVG + PE (wMAN)	**46.05**	91.25	99.19	**29.00**	69.46	**86.26**	**13.30**	**36.99**	**45.32**

Table 3: Moment retrieval performance comparison on the DiDeMo test set. (a) contains representative results of strongly-supervised methods reported in prior works while (b) reports the performance of weakly-supervised methods including our approach.

	Method	Training Supervision	R@1	R@5	mIOU
(a)	MCN Hendricks et al. (2017)	Strong	28.10	78.21	41.08
	TGN (Chen et al., 2018)	Strong	28.23	79.26	42.97
(b)	TGA (Mithun et al., 2019)	Weak	12.19	39.74	24.92
	wMAN	Weak	**38.07**	**63.94**	**38.37**
	Upper Bound	-	74.75	100.00	96.05

are especially helpful in improving Recall@1 accuracies for all IOU values. We provide results for a model variant that include TEFs which encode the location of each video segment. In Table 2, our experiments show that TEFs actually hurt performance slightly. Our model variant with PEs (FBW-WCVG + PE) outperforms the model variant with TEFs (FBW-WCVG + TEF) on all of the metrics. We theorize that the positional encodings aid in integrating temporal context and relative positions into the learned visual-semantic representations. This makes it particularly useful for Charades-STA since its videos are generally much longer.

To gain insights into the fine-grained interactions between frames and words, we provide visualizations in Figure 3. Our model is able to determine the most salient frames with respect to each word relatively well. In both examples, we observe that the top three salient frames with respect to each word are generally distributed over the same subset of frames. This seems to be indicative of the fact that our model leverages contextual information from all video frames as well as words in determining the salience of each frame to a specific word.

4.3.2 DIDEMO

Table 3 reports the results on the DiDeMo dataset. In addition to reporting the state-of-the-art weakly supervised results, we also include the results obtained by strongly-supervised methods. It can be observed that our model outperforms the TGA model by a significant margin, even tripling the Recall@1 accuracy achieved by them. This demonstrates the effect of learning richer joint visual-semantic representations on the accurate localization of video moments. In fact, our full model outperforms the strongly-supervised TGN and MCN models on the Recall@1 metric by approximately 10%.

We observe a consistent trend in the ablation studies (Table 4) as with those of Charades-STA. In particular, through comparing the ablation models FBW and FBW-WCVG, we demonstrate the effectiveness of our multi-level co-attention mechanism in WCVG where it improves the Recall@1 accuracy by a significant margin. Similar to our observations in Table 2, PEs help to encourage accurate latent alignment between the visual and language modalities, while TEFs fail in this aspect.

5 CONCLUSION

In this work, we propose our weakly-supervised Moment Alignment Network with Word-Conditioned Visual Graph which exploits a multi-level co-attention mechanism to infer the latent alignment between visual and language representations at fine-grained word and frame level. Learning context-aware visual-semantic representations helps our model to reason about the temporal occurrence of an event as well as the relationships of entities described in the natural language query.

Table 4: DiDeMo ablation experiment results on the validation set.

Method	R@1	R@5	MIOU
FBW	30.19	**66.74**	39.06
FBW-WCVG	39.93	66.53	39.19
FBW-WCVG + TEF	37.55	66.36	39.11
FBW-WCVG + PE (wMAN)	**41.62**	66.57	**39.20**

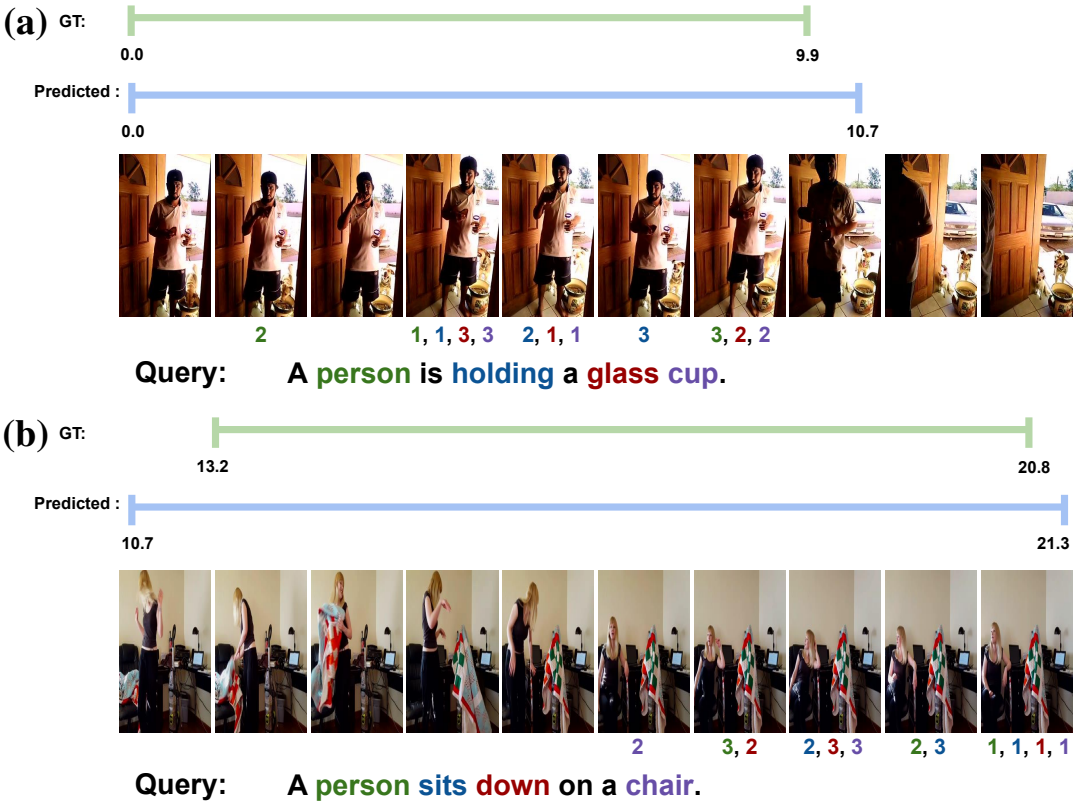

Figure 3: Visualization of the final relevance weights of each word in the query with respect to each frame. Here, we display the top three weights assigned to the frames for each phrase. The colors of the three numbers (1,2,3) indicate the correspondence to the words in the query sentence. We also show the ground truth (GT) temporal annotation as well as our predicted weakly localized temporal segments in seconds. The highly correlated frames to each query word generally fall into the GT temporal segment in both examples.

Finally, our experimental results empirically demonstrate the effectiveness of such representations on the accurate localization of video moments.

REFERENCES

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 162–171, 2018.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. AAAI, 2019.

Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 824–832, 2017.

Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. 2019.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. URL `https://github.com/fartashf/vsepp`.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017.

Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 245–253. IEEE, 2019.

Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *International Conference on Computer Vision (ICCV)*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 201–216, 2018.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4856–4864, 2017.

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.

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11592–11601, 2019.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 299–307, 2017.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 249–264, 2018.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1049–1058, 2016.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 5783–5792, 2017.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9062–9069, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9159–9166, 2019.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1247–1257, 2019.

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2914–2923, 2017.