COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning

Simon Ging[1*], **Mohammadreza Zolfaghari**[1*], **Hamed Pirsiavash**[2], **Thomas Brox**[1]
[1]University of Freiburg, [2]University of Maryland Baltimore County
[1]{gings, zolfagha, brox}@cs.uni-freiburg.de, [2] hpirsiav@umbc.edu

Abstract

Many real-world video-text tasks involve different levels of granularity, such as frames and words, clip and sentences or videos and paragraphs, each with distinct semantics. In this paper, we propose a **Coo**perative hierarchical **T**ransformer (**COOT**) to leverage this hierarchy information and model the interactions between different levels of granularity and different modalities. The method consists of three major components: an attention-aware feature aggregation layer, which leverages the local temporal context (intra-level, e.g., within a clip), a contextual transformer to learn the interactions between low-level and high-level semantics (inter-level, e.g. clip-video, sentence-paragraph), and a cross-modal cycle-consistency loss to connect video and text. The resulting method compares favorably to the state of the art on several benchmarks while having few parameters. All code is available open-source at https://github.com/gingsi/coot-videotext

1 Introduction

Representation learning based on both vision and language has many potential benefits: visual grounding[1, 2, 3, 4]; visual learning with a more natural, almost self-supervised annotation process; and direct applicability to cross-modal tasks, such as video retrieval by text[5, 6, 7, 8, 9], video summarization [10], and automated indexing. This research direction has recently boomed [11, 12, 13, 14, 15, 8, 16, 17] also due to the success of self-attention in text analysis [18, 19] with its almost immediate applicability in the cross-modal context. Many different research foci are currently developing in this area, where some are concerned with large-scale pretraining to leverage the abundant data available [17, 8, 16, 11] to learn a joint embedding space, and others to bring in more explicit structure [20, 21, 22] or new losses [17, 23] into the learning process.

In this paper, we focus on long-range temporal dependencies and propose a hierarchical model that can exploit long-range temporal context both in videos and text when learning the joint cross-modal embedding. For instance, the action of "making tea" involves boiling water, pouring it into a cup, and then adding a tea bag. This action can take a long time and may have lots of details that distinguish a particular style of making tea from other styles. To capture the whole temporal context, we leverage the idea of a hierarchical model with losses that enforce the interaction within and between different hierarchy levels. The idea of such a hierarchy is generic and has been explored by several works [21, 24, 22] in the context of video-text learning. In addition, we use alignment losses from Zhang et al. [21] and extend our baseline model with a new feature aggregation method for the intra-level interactions between features and a new transformer-based module for inter-level interactions (between local and global semantics). We consider three different levels of hierarchy: frame/word, clip/sentence and video/paragraph, visualized by the three blocks in Figure 1.

[*] Equal contribution

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

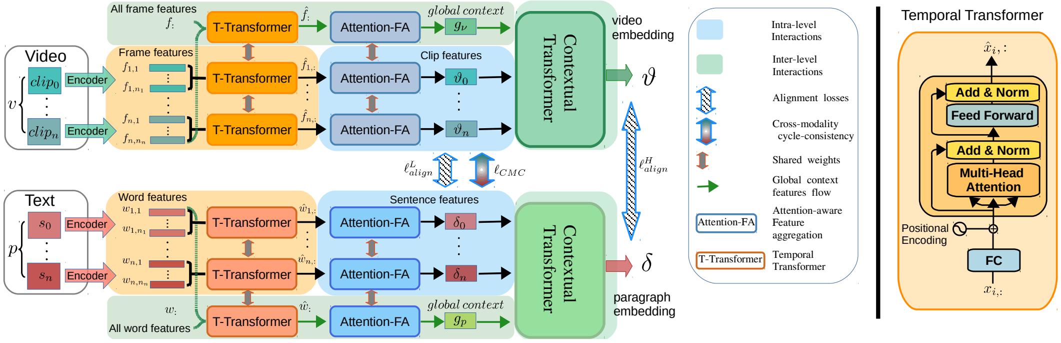


Figure 1: **Overview of COOT model** (best viewed in color). The model consist of two branches: one for video input (top) and one for text input (bottom). Given a video and a corresponding text, we encode them to frame-level/word-level features. Features belonging to each segment (clip/sentence) are fed to a standard temporal transformer (T-Transformer) followed by the proposed feature aggregation module (Attention-FA) to obtain clip/sentence-level features. Finally, a new contextual transformer produces the final video/paragraph embedding based on interactions between local context (clip/sentence features) and global context (all frames/words features). ℓ^L_{align}, ℓ^H_{align}, ℓ^g_{align} and ℓ_{CMC} enforce the model to align the representations at different levels.

To model *intra-level cooperation*, we introduce an attention-aware feature aggregation layer to focus on temporal interactions between low-level entities (Figure 1-Attention-FA).

This component replaces traditional sequence representation aggregation methods in transformers such as using a `[CLS]` token [19, 11, 14, 15] or mean pooling [25] with an attention-aware fusion. It leverages temporal context to encourage important entities to contribute more to the final representation of a sequence of frames or words.

For the *inter-level cooperation*, we introduce a contextual attention module, which enforces the network to highlight semantics relevant to the general context of the video and to suppress the irrelevant semantics. This is done by modeling the interaction between low-level (clips-sentences) and high-level entities (global contexts), as shown in Figure 1-green region.

In addition to this architectural contributions, we introduce a new *cross-modal cycle-consistency loss* to enforce interaction between modalities and encourage the semantic alignment between them in the learned common space. We show that enforcing two domains to produce consistent representations leads to substantially improved semantic alignment.

In summary, this paper contributes:

- a hierarchical transformer architecture with a new attention-aware feature aggregation layer and a new contextual attention module;
- a cross-modal cycle-consistency loss that encourages semantic alignment between vision and text features in the joint embedding space;
- state-of-the-art results on video-text retrieval.

2 Cooperative Hierarchical Transformer

Videos and text descriptions naturally involve different levels of granularity. Every paragraph contains multiple sentences, and each sentence is composed of several words. Similarly, videos have a hierarchical semantic structure, even if it is not as exactly defined as for text. Figure 1 illustrates the overview of the COOT model which consists of three levels: 1) A temporal transformer that captures the relationships between frame/word features, 2) attention-aware feature aggregation to produce clip/sentence features (Sec. 2.2), and 3) a contextual transformer to produce final video and text embeddings (Sec. 2.3). We use alignment losses from Zhang et al. [21] to align representations at different granularity levels. In addition, we introduce a new cross-model cycle-consistency loss to connect video and text (Sec. 3). In this section, we briefly summarize the alignment losses from Zhang et al. [21] and the standard transformer.

2.1 Preliminaries

Semantic Alignment Losses. For the video-text alignment, Zhang et al. [21] leverage a contrastive loss to enforce the positive samples to stay in a close neighborhood and negative samples far apart [26, 27, 28, 29]. Assuming the positive pair $\mathcal{P} = (x, y)$, two negative pairs (x, y') and (x', y) expressed as $\mathcal{N} = \{(x, y'), (x', y)\}$, and a margin α, they define the following loss:

$$L(\mathcal{P}, \mathcal{N}, \alpha) = max(0, \alpha + D(x, y) - D(x', y)) + max(0, \alpha + D(x, y) - D(x, y')) \tag{1}$$

where $D(x, y) = 1 - x^{\intercal} y / (\|x\| \|y\|)$ is the cosine distance of two vectors.

To align representations at clip-sentence (ϑ_i^k, δ_i^k), video-paragraph (ϑ^k, δ^k) and global context (g_v, g_p) levels, Zhang et al. [21] use the following losses:

$$\begin{aligned}
\ell_{align}^{L} &= \sum_{k \in \mathcal{D}, i, k' \neq k, i' \neq i} L((\vartheta_i^k, \delta_i^k), \{(\vartheta_i^k, \delta_{i'}^{k'}), (\vartheta_{i'}^{k'}, \delta_i^k)\}, \beta) \\
\ell_{align}^{H} &= \sum_{k \in \mathcal{D}, k' \neq k} L((\vartheta^k, \delta^k), \{(\vartheta^k, \delta^{k'}), (\vartheta^{k'}, \delta^k)\}, \alpha) \\
\ell_{align}^{g} &= \sum_{k \in \mathcal{D}, k' \neq k} L((g_v^k, g_p^k), \{(g_v^k, g_p^{k'}), (g_v^{k'}, g_p^k)\}, \alpha_g)
\end{aligned} \tag{2}$$

Here, ϑ_i^k denotes the embedding for the i-th clip of the k-th video and similarly δ_i^k is the embedding of the i-th sentence of the k-th paragraph. α, α_g and β are constant margins, and $\mathcal{D}$ is a dataset of videos with corresponding text descriptions. Zhang et al. [21] employed an additional loss to model the clustering of low-level and high-level semantics in the joint embedding space:

$$\begin{aligned}
\ell_{cluster} = &\sum_{k \in \mathcal{D}, i, k' \neq k, i' \neq i} L((1, 1), \{(\vartheta_i^k, \vartheta_{i'}^{k'}), (\delta_{i'}^{k'}, \delta_i^k)\}, \gamma) \\
&+ \sum_{k \in \mathcal{D}, k' \neq k} L((1, 1), \{(\vartheta^k, \vartheta^{k'}), (\delta^{k'}, \delta^k)\}, \eta)
\end{aligned} \tag{3}$$

where γ and η both are constant margins. The $(1, 1)$ pairs denote that positive samples are not changed. In short, the goal of this loss is to push apart embeddings for negative samples.

Note. Due to the symmetrical design of the video and text branches in our model, from now on, we explain only the video branch. For simplicity, we assume a single head in transformer formulations. All transformers use residual connections.

Temporal Transformer We use standard attention-blocks [18] to learn frame and word representations, as shown in Fig 1-Right. We learn two temporal transformers (T-Transformer); one for the video branch and another one for the text branch. Both have the same architecture. All T-Transformers in each branch share their weights. This module draws the relationship between temporal features and yields improved representations as output. Given a video v^k, we first encode all its frames to obtain the frame-level features $\{f_{i,:}^k\}_{i=1}^n$, where $f_{i,:}^k$ are all frame-level features of the i-th clip for video v^k (orange parts in Figure 1). We also consider all frame features ($f_{:}^k$) of a video as extra input for the global context computation (green parts in Figure 1). This yields $\{\hat{f}_{i,:}^k\}_{i=1}^n$ and $\hat{f}_{:}^k$.

2.2 Intra-Level Cooperation

Standard feature fusion methods consider each feature independently by average pooling or max pooling. Hence, they miss the relationship between features to highlight the relevant features. Recent transformers use a `[CLS]` token [11, 19, 14, 15] or average pooling [25] to obtain the aggregated features. For example, when a person is cooking, objects on the table are more relevant than objects on the wall or in the background. Therefore, we need to attend to specific features depending on the context. There have been some attempts in other domains to design a context-aware feature fusion method [30, 31, 32, 33]. However, we introduce an attention-aware feature aggregation module (Attention-FA in Fig. 1) for video-text transformers.

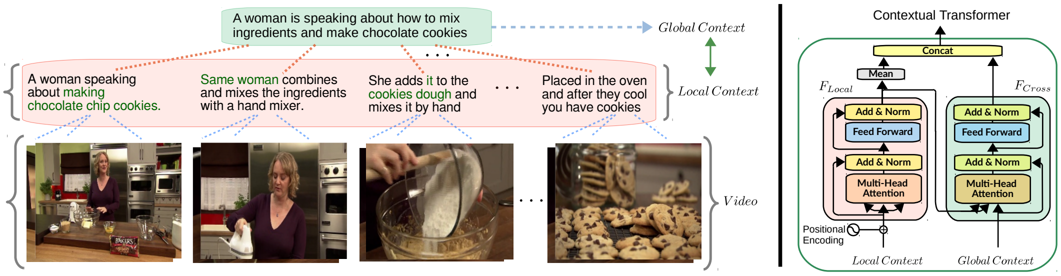

A woman is speaking about how to mix ingredients and make chocolate cookies
$Global\ Context$
A woman speaking about making chocolate chip cookies.
Same woman combines and mixes the ingredients with a hand mixer.
She adds it to the cookies dough and mixes it by hand
. . .
Placed in the oven and after they cool you have cookies
$Local\ Context$
$Video$
Contextual Transformer
Concat
Mean
F_{Local}
F_{Cross}
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
$Local\ Context$
$Global\ Context$

Figure 2: **Contextual Transformer (CoT).** This module (right) encourages the model to optimize the representations with respect to interactions between local and global context. In the third sentence, to know the type of dough (*cookie*) the model should have information about the general context of the video (*making chocolate cookies*). Likewise, in the second sentence, to know that she is the *"same woman"*, the model must be aware of the person's identity throughout the video.

Suppose we have a sequence with T feature vectors, denoted by $X = \{x_1, \ldots, x_T\}$ (e.g. $\hat{f}^k_{i,:} = \{\hat{f}^k_{i,1}, \ldots, \hat{f}^k_{i,T}\}$). We set **key** $K = X$ and utilize two learnable transformation weights W_2 and W_1 together with two biases b_1 and b_2. The attention matrix A is computed as:

$$A = \text{softmax}(W_2 Q + b_2)^T, \qquad Q = GELU(W_1 K^T + b_1),\ K = X \tag{4}$$

We compute the final feature as $\hat{x} = \sum_{i=1}^{T} a_i \odot x_i$, where $\odot$ denotes element-wise multiplication and a_i is the i-th attention vector of A for the i-th feature. This module differs from attention [18] in two aspects: (1) we use only two learnable weights for query (Q) and key (K) and then aggregate the values based on calculated scores; (2) the query equals to transformed keys (K) and then we apply the activation function GELU [34, 35]. We feed $\{\hat{f}^k_{i,:}\}_{i=1}^n$ and $\hat{f}^k_{:}$ to this component and obtain the clip-level ($\{\vartheta^k_i\}_{i=1}^n$) features and the global context for the video (g_ν).

2.3 Inter-Level Cooperation

By modeling the interactions between local and global context, the network learns to highlight semantics relevant to the general context of the video and to suppress the irrelevant ones: interactions between clip embeddings and the general context of the video; interactions between sentence embeddings and the general context of the text. As shown in Figure 2-Left, without knowing the global context, just from observing the frame in the third clip, there is no information about what type of *"dough"* is involved. Also the *"same woman"* in the second clip could not be related to the woman seen in the first clip.

Thus, we propose a Contextual Transformer (CoT) in Figure 2-Right to model the interactions between low-level and high-level semantics. More formally, we build the Contextual Transformer with two modules F_{Local} and F_{Global}. We append the positional embedding to the inputs of F_{Local}. The goal of F_{Local} is to model the short-term interactions between low-level semantics ($\{\vartheta^k_i\}_{i=1}^n$), whereas F_{Global} models the interactions between local and global context (g_ν) to highlight the important semantics.

Given local representations $\{\vartheta^k_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$, where n is the number of clips and d indicates the feature dimension, F_{Local} applies multi-head attention followed by a feed-forward layer and a normalization layer on top of both layers and produces embeddings $\{h_i\}_{i=1}^n$.

We compute **key (K)-value(V)** pairs based on these embeddings $\{h_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$ and **query(Q)** based on the global context g_v. F_{Global} produces the attention output as follows,

$$H_{attn} = \text{softmax}(\frac{\mathbf{QK}^T}{\sqrt{d}})\mathbf{V}, \qquad Q = \mathcal{W}_q g_v,\ K = \mathcal{W}_k \{h_i\}_{i=1}^n,\ V = \mathcal{W}_v \{h_i\}_{i=1}^n \tag{5}$$

where $\mathcal{W}_q$, $\mathcal{W}_k$, and $\mathcal{W}_v$ are the embedding weights. H_{attn} is a weighted sum of values (local semantics), where the weight of each value is calculated based on its interaction with the global context query $\mathbf{Q}$. H_{attn} is further encoded by a feed-forward layer to produce the contextual embedding $H_{context}$. We calculate the mean of $\{h_i\}_{i=1}^n$ and concatenate it with $H_{context}$ to obtain the final video embedding $\vartheta^k = \text{concat}(mean(\{h_i\}_{i=1}^n), H_{context})$; see Figure 2.

3 Cross-Modal Cycle Consistency

We introduce a cross-modal cycle-consistency loss to enforce the semantic alignment between clips and sentences, as illustrated in Figure 3. It replaces the cross-modal attention units used in [14, 8]. A pair of clip and sentence will be identified as semantically aligned if they are nearest neighbors in the learned common spaces. Consider as input a sequence of clip embeddings $\{\vartheta_i\}_{i=1}^{n} = \{\vartheta_1, \ldots, \vartheta_n\}$ and sentence embeddings $\{\delta_i\}_{i=1}^{m} = \{\delta_1, \ldots, \delta_m\}$. As the sentences of a paragraph have a temporal order, given a sentence embedding δ_i on this sequence, we first find its soft nearest neighbor [36, 37, 38]

$$\bar{\vartheta}_{\delta_i} = \sum_{j=1}^{n} \alpha_j \vartheta_j \quad \text{where} \quad \alpha_j = \frac{\exp(-\|\delta_i - \vartheta_j\|^2)}{\sum_{k=1}^{n} \exp(-\|\delta_i - \vartheta_k\|^2)} \tag{6}$$

in the clip sequence $\{\vartheta_i\}_{i=1}^{n}$. α_j is the similarity score of clip ϑ_j to sentence δ_i. We then cycle back from $\bar{\vartheta}_{\delta_i}$ to the sentence sequence $\{\delta_i\}_{i=1}^{m}$ and calculate the soft location

$$\mu = \sum_{j=1}^{m} \beta_j j \quad \text{where} \quad \beta_j = \frac{\exp(-\|\bar{\vartheta} - \delta_j\|^2)}{\sum_{k=1}^{m} \exp(-\|\bar{\vartheta} - \delta_k\|^2)}. \tag{7}$$

The sentence embedding δ_i is semantically cycle consistent if and only if it cycles back to the original location, i.e., $i = \mu$. We penalize deviations from cycle-consistency for sampled sets of clips and sentences, which encourages the model to learn semantically consistent representations.

Our objective is the distance between the source location i and the soft destination location μ.

$$\ell_{CMC} = \|i - \mu\|^2 \tag{8}$$

Computing nearest neighbors as soft nearest neighbors makes the loss differentiable [36, 37, 38]. We can use this loss in both supervised and self-supervised scenarios. In the self-supervised case, we split each video uniformly into several clips and each paragraph into sentences. Beside the cycle-consistency from text to video, we also calculate ℓ_{CMC} from video to text. Therefore, the final ℓ_{CMC} loss includes both cycles.

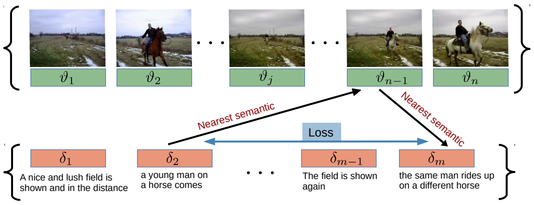


Figure 3: **Cross-Modality Cycle-Consistency.** Starting from a sentence s_i, we find its nearest neighbor in the clip sequence and again its neighbor in the sentence sequence. Deviations from the start index are penalized as alignment error.

The final training loss for the overall model is:

$$\ell_{final} = \ell_{align}^{L} + \ell_{align}^{H} + \ell_{align}^{g} + \ell_{cluster} + \lambda \ell_{CMC} \tag{9}$$

4 Experimental Setup

Datasets. We evaluate our method on the datasets **ActivityNet-captions** [39] and **Youcook2** [40]. ActivityNet-captions consists of 20k YouTube videos with an average length of 2 minutes, with 72k clip-sentence pairs. There are ∼10k, ∼5k and ∼5k videos in train, val1 and val2, respectively. Youcook2 contains 2000 videos with a total number of 14k clips. This dataset is collected from YouTube and covers 89 types of recipes. There are ∼9.6k clips for training and ∼3.2k clips for validation. For each clip there is a manually annotated textual description.

Evaluation Metrics. We measure the performance on the retrieval task with standard retrieval metrics, i.e., recall at K (R@K e.g. R@1, R@5, R@10) and Median Rank (MR).

Text encoding. We feed paragraphs consisting of several sentences into a pretrained "BERT-Base, Uncased" model [19] and use the per-token outputs of the last 2 layers, resulting in 1536-d features.

Video encoding. For Activitynet-Captions, we use the 2048-d features provided by Zhang et al. [21] (at 3.8 FPS). For Youcook2, we test two approaches: (*A*) We follow Miech et al., 2019 [16] and concatenate 2D (Resnet-152 pretrained on ImageNet [41]) and 3D (ResNext-101 model [42] pretrained on Kinetics [43]) outputs to obtain 4096-d features at 6.2 FPS; (*B*) We use the video embedding network provided by Miech et al., 2020 [17] pretrained on video-text learning on the Howto100m dataset to obtain 512-d features at 0.9 FPS.

Table 1: **Ablation study on ActivityNet-captions (val1).** We quantify the individual contributions of the attention-aware feature aggregation (AF), the Contextual Transformer (CoT), and the cross-modal cycle-consistency loss (CMC). HSE results are reproduced by us. Disabling CoT means removing the cross-attention layer between local and global context.

Model	Pooling	CMC	CoT	Paragraph $\Longrightarrow$ Video			Video $\Longrightarrow$ Paragraph			Param (M)
	Lowlvl			R@1	R@5	R@50	R@1	R@5	R@50	
HSE	Max	✗	✗	45.6±0.3	76.1±0.7	96.0±0.3	44.9±0.5	75.8±1.2	95.8±0.4	26.1
HSE	Max	✓	✗	46.6±0.4	78.1±0.3	97.3±0.1	46.4±0.3	77.6±0.3	97.1±0.3	26.1
COOT	CLS	✗	✗	49.4±1.4	77.7±1.3	95.7±0.2	49.7±1.9	77.8±0.9	95.8±0.3	4.9
COOT	AVG	✗	✗	52.6±0.6	80.6±0.4	97.0±0.2	52.1±0.4	80.8±0.2	97.0±0.2	4.9
COOT	Max	✗	✗	58.2±0.5	84.9±0.2	98.1±0.1	58.7±0.5	86.0±0.2	98.2±0.1	4.9
COOT	AFA	✗	✗	59.0±0.5	85.4±0.2	98.2±0.0	59.8±0.6	85.8±0.8	98.2±0.1	5.8
COOT	Max	✓	✓	59.4±0.9	86.1±0.6	98.3±0.0	60.5±0.1	87.1±0.2	98.5±0.1	6.7
COOT	AFA	✗	✓	59.8±1.1	86.3±0.3	98.5±0.1	60.1±0.1	87.1±0.4	98.5±0.1	7.6
COOT	AFA	✓	✗	59.5±0.5	85.5±0.4	98.1±0.0	60.5±0.7	86.2±0.5	98.2±0.1	5.8
COOT	AFA	✓	✓	**60.8**±0.6	**86.6**±0.4	**98.6**±0.1	**60.9**±0.3	**87.4**±0.5	**98.6**±0.0	7.6

For each clip as well as for the entire video, we sample up to 80 frame features. If needed, we split the frames into 80 equal length intervals and uniformly sample a frame from each interval during training or take the center frame during validation.

Training. Similar to [21] we set all margins $\alpha = \alpha_g = \beta = \gamma = \mu = 0.2$. We use a mini-batch size of 64 video/paragraph pairs and sample all corresponding clips and sentences. All possible combinations of embeddings with non-matching indices in a batch are used as negative samples for the contrastive loss. To apply the cycle-consistency loss, we found that sampling 1 clip per video and 1 sentence per paragraph works best. The optimal loss weight λ depends on architecture and dataset. As activation function, we found GELU [34] to perform best. We set the hidden size to 384 and use a pointwise linear layer to reduce the input feature dimension. We use one self-attention layer for the T-Transformer and one self-attention and one cross-attention layer for CoT. For further details on optimization and hyperparameters we refer the interested reader to the supplementary material.

Video-Language Retrieval. For video-text retrieval, the query is a paragraph and the task is to find the most relevant video from a database. Alternatively, the query can be a video and the task is to retrieve the most relevant paragraph. We follow the experimental protocol from Zhang et al. [21] to evaluate the models. We use the final embedding output of our model (ϑ^k, δ^k) to do the retrieval.

Clip-sentence retrieval. For Youcook2, we also evaluate the quality of our model when retrieving a short video clip given a single sentence. For this experiment, we use the intermediate low-level embeddings produced by our model $(\vartheta^k_i, \delta^k_i)$ to do the retrieval.

5 Results

Influence of each component. We show results of a model ablation study in Table 1. First, to validate the general effectiveness of the proposed cross-modal cycle consistency loss (CMC), we apply it to the HSE architecture [21]. The ℓ_{CMC} loss provides a significant boost in performance for both HSE and COOT, which indicates that it will be beneficial if plugged into other video-text representation learning methods. Second, the Attention-FA module shows better performance (7.2% average improvement on R@1 for paragraph $\Longrightarrow$ video and video $\Longrightarrow$ paragraph tasks) than common average pooling. Third, we observe that integrating the Contextual Transformer into the overall model improves the performance. This confirms that interactions between local and global context help the model to highlight the relevant semantics (more in supp. material).

Comparison to the state of the art Table 2 summarizes the results of paragraph to video and video to paragraph retrieval tasks on the ActivityNet-captions dataset. For a fair comparison, our model utilizes the same video features as HSE [21]. Our method significantly outperforms all previous methods across different evaluation metrics. COOT obtains on average 16.6% better R@1 in comparison to HSE [21] while having fewer parameters. We believe the major gain comes from our attention-aware feature aggregation component and the ℓ_{CMC} loss.

Table 2: **Video-paragraph retrieval results on AcitvityNet-captions dataset (val1).**

	Paragraph $\Longrightarrow$ Video				Video $\Longrightarrow$ Paragraph			
Method	R@1	R@5	R@50	MR	R@1	R@5	R@50	MR
LSTM-YT [52]	0.0	4.0	24.0	102.0	0.0	7.0	38.0	98.0
No Context [53]	5.0	14.0	32.0	78.0	7.0	18.0	45.0	56.0
DENSE [39]	14.0	32.0	65.0	34.0	18.0	36.0	74.0	32.0
VSE [54]([5])	11.7	34.7	85.7	10	-	-	-	-
FSE [21]	18.2	44.8	89.1	7	16.7	43.1	88.4	7
HSE [21]	$44.4_{\pm 0.5}$	$76.7_{\pm 0.3}$	$97.1_{\pm 0.1}$	2	$44.2_{\pm 0.6}$	$76.7_{\pm 0.3}$	$97.0_{\pm 0.3}$	2
COOT	$\mathbf{60.8}_{\pm 0.6}$	$\mathbf{86.6}_{\pm 0.4}$	$\mathbf{98.6}_{\pm 0.1}$	**1**	$\mathbf{60.9}_{\pm 0.3}$	$\mathbf{87.4}_{\pm 0.5}$	$\mathbf{98.6}_{\pm 0.0}$	**1**

We further provide retrieval results on the Youcook2 [40] dataset in Table 3. We compare our model under two settings: (1) with features pretrained on classification (2) with features from a pretrained SOTA video-text model.

Without HowTo100M pretrained features. We use features (A) explained in Section 4 and train the COOT model on the YouCook2 dataset. Using the same training set, COOT outperforms Miech et al. [16] and HGLMM [44] on both paragraph-to-video and sentence-to-clip tasks. This supports our rationale that modeling interactions between different hierarchy levels is crucial for capturing long-term semantics.

With HowTo100M pretrained features. In Table 3, we compare our method with the recently proposed SOTA methods MIL-NCE [17], ActBERT [8], and Miech et al. [16], which utilize pretraining on the huge HowTo100M dataset. We use features (B) (Sec. 4) and train the model on the YouCook2 dataset. Note that the paragraph to video results of other methods are computed by us. Training our model with features of a model pretrained on the HowTo100M dataset clearly improves over training with features of a model pretrained on classification and over the state-of-the-art. We can see that our model outperforms MIL-NCE [17] 16.4% on R@1 score for paragraph-to-video task, which verifies COOT benefits from hierarchy interactions. This shows that the contributions of this paper are complementary to works that focus on large-scale pretraining.

Time complexity and number of parameters. The COOT model has 10.6M, parameters which is 60% less than the HSE method (Table 1). Training is fast and takes less than 3 hours on two GTX1080Ti GPUs (without data I/O).

5.1 Video Captioning

To show that the learned representations contain meaningful information for other tasks than retrieval, we use the learned representations for video captioning building upon the captioning model MART [45]. The original method uses appearance (RGB) and optical flow features extracted from ResNet-200 [41] and BN-Inception [46], respectively.

We use the clip (ϑ_i^k) and optionally the video (ϑ^k) representation generated with our COOT model. In comparison to MART, we input about 100 times less features per video into the captioning model. We use the standard language evaluation metrics BLEU@3/4 [47], RougeL [48], METEOR [49], CIDEr-D [50] and R@4 [51] which measures the degree of n-gram repetition. Our results in Table 4 and Table 5 show that the MART method using our representations improves over using appearance and optical flow video features. Generated captions in Table 6 show that our video representations encapsulate richer information about the video while being more compact.

6 Related Work

Image and Language. Many self-supervised visual-language representation learning methods have focused on improving one representation with the help of the other [44, 57, 58, 59, 60, 61, 25]. These approaches learn joint image-text embeddings or map images and sentences into a common space. Recently, there has been a surging interest in utilizing Transformers for image-text representation learning [62, 63, 64, 65, 66]. ViLBERT [13] and VisualBERT [12] pretrain a BERT-like architecture on an image-text dataset and then transfer learned representations to different downstream tasks. LXMERT [14] uses additional pretraining tasks and an object-relationship component. In contrast to [14, 13], VL-BERT [15] does not utilize the task of sentence-image relationship prediction and additionally pretrains the model on text-only datasets.

Table 3: **Retrieval results on YouCook2 dataset.** Results with * are computed by us. $^{\triangle}$ we use features of a video-text model [17] pretrained on the HowTo100m dataset.

		Paragraph ⟹ Video				Sentence ⟹ Clip			
Method	TrainSet	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
Random	-	0.21	1.09	2.19	229	0.03	0.15	0.3	1675
Miech et al. [16]	HowTo100M	43.1*	68.6*	79.1*	2*	6.1	17.3	24.8	46
ActBERT [8]	HowTo100M	-	-	-	-	9.6	26.7	38.0	19
MIL-NCE [17]	HowTo100M	61.9*	89.4*	98.9*	**1***	15.1	38.0	51.2	10
HGLMM [44]	YouCook2	-	-	-	-	4.6	14.3	21.6	75
Miech et al. [16]	YouCook2	32.3*	59.2*	70.9*	4*	4.2	13.7	21.5	65
COOT	YouCook2	50.4$_{\pm 2.6}$	79.4$_{\pm 0.6}$	87.4$_{\pm 0.8}$	1.3$_{\pm 0.6}$	5.9$_{\pm 0.7}$	16.7$_{\pm 0.6}$	24.8$_{\pm 0.8}$	49.7$_{\pm 2.9}$
Miech et al. [16]	HowTo100M+ YouCook2	59.6*	86.0*	93.6*	**1***	8.2	24.5	35.3	24
COOT	HowTo100M$^{\triangle}$+ YouCook2	**77.2**$_{\pm 1.0}$	**95.8**$_{\pm 0.8}$	**97.5**$_{\pm 0.3}$	**1.0**$_{\pm 0.0}$	**16.7**$_{\pm 0.4}$	**40.2**$_{\pm 0.3}$	**52.3**$_{\pm 0.5}$	**9.0**$_{\pm 0.0}$

Table 4: **Captioning results on the YouCook2 dataset (val split).** Results with * are computed by us. $^{\triangle}$ we use features of a video-text model [17] pretrained on the HowTo100m dataset. "MART w/o re" denotes a MART variant without recurrence.

Features	Method	TrainSet	B@3	B@4	RougeL	METEOR	CIDEr-D	R@4↓
RGB+Flow	VTransformer [55]	YouCook2	13.08*	7.62	32.18*	15.65	32.26	7.83
RGB+Flow	TransformerXL [56]	YouCook2	11.46*	6.56	30.78*	14.76	26.35	6.30
RGB+Flow	MART [45]	YouCook2	12.83*	8.00	31.97*	15.90	35.74	**4.39**
COOT clip	MART	YouCook2	14.17	8.69	33.01	16.11	38.28	8.07
COOT video+clip	MART	YouCook2	15.75	9.44	34.32	18.17	46.06	6.30
COOT clip	MART	H100M$^{\triangle}$+YC2	17.12	10.91	37.59	18.85	54.07	5.11
COOT clip	MART w/o re.	H100M$^{\triangle}$+YC2	17.16	10.69	37.43	19.18	54.85	5.45
COOT clip	VTransformer	H100M$^{\triangle}$+YC2	17.62	11.09	37.63	19.34	54.67	4.57
COOT video+clip	VTransformer [55]	H100M$^{\triangle}$+YC2	17.79	11.05	37.51	19.79	55.57	5.69
COOT video+clip	MART	H100M$^{\triangle}$+YC2	**17.97**	**11.30**	**37.94**	**19.85**	**57.24**	6.69

Video and Language. The multi-modal nature of video is a great source of self-supervision. Modalities such as audio, text and motion provide strong cues to learn richer spatio-temporal features [67, 68, 16, 17, 9, 69, 11]. Aytar et al. [70] leverage natural synchronization to learn rich representations across vision, sound, and language. VideoBERT [11] learns joint video-text representations based on predicting whether the linguistic sentence is temporally aligned with the visual sentence. These approaches [8, 23, 11] focus on self-supervised pretraining and require a large set of paired video clips and texts to learn a good representation model [17].

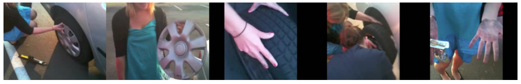
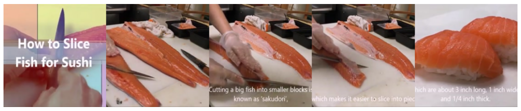
There has been growing interest in temporal localization of natural language in videos [7, 71, 72, 73, 2]. Moment localization identifies a time window given a text query [7, 71, 74]. Most related to our work are methods that focus on joint video-text embeddings and perform video-text retrieval or captioning [4, 75, 76, 21, 9, 69, 52, 6, 5]. Several works tried to utilize the temporal structure of video and text for the alignment task [77, 70, 78, 4, 79]. Miech et al. [68] proposed a mixture of experts approach to learn text-video representations. Likewise, CE [20] proposes a mixture-of-experts model to aggregate information from pretrained experts (e.g. object, action, audio) with a gating mechanism. In our work, we use a similar hierarchy as CMHSE [21, 24, 22] and extend their design by proposing three new components to learn the interactions between different levels of the hierarchy.

Cycle-Consistency uses transitivity as an objective for training [38, 80, 81, 82]. The assumption of cyclical structure has been used in various works [83, 84, 85]. Wang et al. [80] obtain the supervision for visual correspondence by tracking forward and backward. Shah et al. [81] enforce consistency between the generated and the original question in visual question answering. To prevent mode collapse, cycle-consistency is activated only after a certain number of training iterations [81]. TCC [38] employs cycle-consistency for temporal video alignment. In contrast to TCC, which works only in the video domain, we align video and text. To the best of our knowledge, this is the first work which introduces cycle-consistency to the video-text domain.

Table 5: **Captioning results on the ActivityNet-Captions dataset (ae-test split of MART [45]).** Results with * are computed by us. "MART w/o re" denotes a MART variant without recurrence.

Features	Method	TrainSet	B@3	B@4	RougeL	METEOR	CIDEr-D	R@4↓
RGB+Flow	VTransformer [55]	ActivityNet	16.27*	9.31	29.18*	15.54	21.33	7.45
RGB+Flow	TransformerXL [56]	ActivityNet	16.71*	10.25	30.53*	14.91	21.71	8.79
RGB+Flow	MART	ActivityNet	16.43*	9.78	30.63*	15.57	22.16	5.44
COOT video+clip	TransformerXL [56]	ActivityNet	16.94	10.57	30.93	14.76	22.04	15.85
COOT video+clip	VTransformer [55]	ActivityNet	16.80	10.47	30.37	15.76	25.90	19.14
COOT clip	MART w/o re.	ActivityNet	15.41	9.37	28.66	15.61	22.05	12.03
COOT video+clip	MART w/o re.	ActivityNet	16.59	10.33	29.93	15.64	25.41	17.03
COOT clip	MART	ActivityNet	16.53	10.22	30.68	15.91	23.98	**5.35**
COOT video+clip	MART	ActivityNet	**17.43**	**10.85**	**31.45**	**15.99**	**28.19**	6.64

Table 6: **Captioning samples, more accurate (left) and less accurate (right) cases.** First row: ActivityNet (ae-test split), second row: YouCook2 (val split). Red/bold indicates content errors, blue/italic indicates repetitive patterns.

	
MART: A person is **driving the car**. **A boy** is **holding a bottle of wood**. **COOT (Ours)**: A woman is seen kneeling down next to a car while others stand around her. The woman then pushes the tire **back and fourth**. **GT**: A girl is shown trying to change a tire. She successfully removes the tire, then replaces it with a spare, showing off their dirty hands afterward.	**MART**: A man *is kneeling down on* a floor. He *is kneeling down on* the ground. **COOT (Ours)**: A man is seen kneeling down on the ground and begins **putting shoes on**. The man continues to put on the shoes and ends by *putting his shoes on*. **GT**: A person is seen bending over a floor placing tiles down over the plaster. The person continues laying tiles down and pushing down on the floor to make sure it's sturdy.
	
MART: Heat up a pan and cook until golden brown. Add **onions** *to the pan*. Add flour salt and pepper *to the pan*. Add rice *to the pan* and stir. **COOT (Ours)**: Boil the potatoes in water. Add chopped *potatoes* to the pan. Add butter and mash. Add some milk and *mash*. **GT**: Boil some small pieces of potatoes in water. Mash the potato. Add some butter and salt and stir. Gradually add milk while stirring the potatoes.	**MART**: Cut the salmon into thin slices. *Cut the salmon into thin slices*. *Cut the salmon into thin slices*. *Cut the salmon into thin slices*. **COOT (Ours)**: Cut the salmon in half. *Cut the salmon in half*. *Cut the salmon* into thin slices. *Cut the salmon into thin slices*. **GT**: Slice the fish into smaller pieces. Chop the tail end off. Cut the fish at an angle. Cut the fish into thin pieces.

7 Conclusions

We have presented a cooperative hierarchical transformer architecture for learning a joint video and text embedding space where similar semantics are aligned. The architecture is designed to encourage the use of long-range temporal context in a cross-level manner. Our approach uses two new components to model the interactions within and between hierarchy levels; an attention-aware feature aggregation module to model the interactions between frames and words, a contextual transformer to model the interactions between local contexts and global context. In addition, we have introduced a new cross-modal cycle-consistency loss which enforces the semantic alignment of clips and sentences. We have shown that both components contribute – jointly and individually – to an improved retrieval performance. As a result, our approach achieves state-of-the-art retrieval and captioning performance on two challenging datasets.

Broader Impact

This work contributes fundamental research and does not present any foreseeable societal consequence. In the long run, this line of research can contribute to services on video search and video organisation.

Acknowledgments

We thank Ehsan Adeli for helpful comments, Antoine Miech for providing details on their retrieval evaluation, and Facebook for providing us a GPU server with Tesla P100 processors for this research work.

References

[1] Satwik Kottur, Ramakrishna Vedantam, José M. F. Moura, and Devi Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. *CoRR*, abs/1511.07067, 2015.

[2] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K. Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *CoRR*, abs/2001.09308, 2020.

[3] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[4] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[5] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[6] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Yi Yang Linchao Zhu. Actbert: Learning global-local video-text representations. In *CVPR*, 2020.

[9] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, page 19–27, New York, NY, USA, 2018. Association for Computing Machinery.

[10] B. A. Plummer, M. Brown, and S. Lazebnik. Enhancing video summarization via vision-language embedding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1052–1060, 2017.

[11] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.

[12] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[14] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[15] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[16] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

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

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *arXiv preprint arxiv:1907.13487*, 2019.

[21] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pages 385–401, 2018.

[22] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1029–1038, 2016.

[23] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019.

[24] Jiwei Li, Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, Beijing, China, July 2015. Association for Computational Linguistics.

[25] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. Mule: Multimodal universal language embedding, 2019.

[26] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

[27] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. *CoRR*, abs/1703.07464, 2017.

[28] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris. Near-duplicate video retrieval with deep metric learning. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 347–356, 2017.

[29] Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R. Scott. Deep metric learning with hierarchical triplet loss. *CoRR*, abs/1810.06951, 2018.

[30] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *CoRR*, abs/1706.06905, 2017.

[31] Mateusz Malinowski and Mario Fritz. Learnable pooling regions for image classification. In *ICLR workshop*, May 2013.

[32] Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. Enhancing sentence embedding with generalized pooling. *CoRR*, abs/1806.09828, 2018.

[33] Hayoung Eom and Heeyoul Choi. Alpha-pooling for convolutional neural networks. *CoRR*, abs/1811.03436, 2018.

[34] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[35] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[36] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018.

[37] Jacob Goldberger, Geoffrey E Hinton, Sam T. Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, 2005.

[38] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019.

[39] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, 2018.

[41] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[42] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[43] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[44] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4437–4446, 2015.

[45] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*, 2020.

[46] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[48] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[49] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[50] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[51] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, 2018.

[52] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[53] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[54] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc., 2013.

[55] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer, 2018.

[56] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

[57] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[58] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[59] Gil Sadeh, Lior Fritz, Gabi Shalev, and Eduard Oks. Joint visual-textual embedding for multimodal style search. *CoRR*, abs/1906.06620, 2019.

[60] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. *CoRR*, abs/1511.06078, 2015.

[61] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1928–1937, 2017.

[62] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2019.

[63] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data, 2020.

[64] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.

[65] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers, 2020.

[66] Kazuki Miyazawa, Tatsuya Aoki, Takato Horii, and Takayuki Nagai. lambert: Language and action learning using multimodal bert. *arXiv preprint arXiv:2004.07093*, 2020.

[67] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[68] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data. In *arXiv*, 2018.

[69] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. *CoRR*, abs/1505.01861, 2015.

[70] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017.

[71] Jonghwan Mun, Minsu Cho, , and Bohyung Han. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*, 2020.

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

[73] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, 2017.

[74] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. Wslln:weakly supervised natural language localization networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[75] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval, 2020.

[76] Ran Xu, Caiming Xiong, Wei Chen, and Jason J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2346–2352. AAAI Press, 2015.

[77] Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Weakly-Supervised Alignment of Video With Text. In *ICCV - IEEE International Conference on Computer Vision*, pages 4462–4470, Santiago, Chile, December 2015. IEEE.

[78] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1827–1835, 2015.

[79] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[80] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019.

[81] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

[82] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.

[83] Y. Chen, Y. Lin, M. Yang, and J. Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1791–1800, 2019.

[84] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 820–828, Red Hook, NY, USA, 2016. Curran Associates Inc.

[85] Duyu Tang, Nan Duan, Zhao Yan, Zhirui Zhang, Yibo Sun, Shujie Liu, Yuanhua Lv, and Ming Zhou. Learning to collaborate for question answering and asking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1564–1574, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.