Action Modifiers: Learning from Adverbs in Instructional Videos

Hazel Doughty[1] Ivan Laptev[2] Walterio Mayol-Cuevas[1,3] Dima Damen[1]

[1]University of Bristol [2]Inria, École Normale Supérieure [3]Amazon

Abstract

We present a method to learn a representation for adverbs from instructional videos using weak supervision from the accompanying narrations. Key to our method is the fact that the visual representation of the adverb is highly dependant on the action to which it applies, although the same adverb will modify multiple actions in a similar way. For instance, while 'spread quickly' and 'mix quickly' will look dissimilar, we can learn a common representation that allows us to recognize both, among other actions.

We formulate this as an embedding problem, and use scaled dot-product attention to learn from weakly-supervised video narrations. We jointly learn adverbs as invertible transformations operating on the embedding space, so as to add or remove the effect of the adverb. As there is no prior work on weakly supervised learning of adverbs, we gather paired action-adverb annotations from a subset of the HowTo100M dataset for 6 adverbs: quickly/slowly, finely/coarsely, and partially/completely. Our method outperforms all baselines for video-to-adverb retrieval with a performance of 0.719 mAP. We also demonstrate our model's ability to attend to the relevant video parts in order to determine the adverb for a given action.

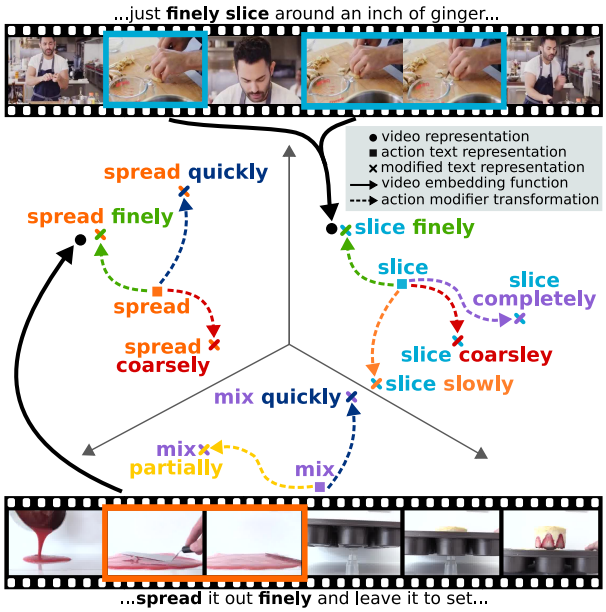

Figure 1. We learn a joint video-text embedding space from instructional videos and accompanying action-adverb pairs in the narration. Within this space, we learn adverbs as action modifiers — that is transformations which modify the action's embedding.

1. Introduction

Instructional videos are a popular type of media watched by millions of people around the world to learn new skills. Several previous works aimed to learn the key steps necessary to complete the task from these videos [1, 30, 45, 62]. However, identifying the steps, or their order, is not all one needs to perform the task well; some steps need to be performed in a certain way to achieve the desired outcome. Take for example the task of making a meringue. An expert would assure you it is critical to add the sugar *gradually* and avoid over-beating by folding the mixture *gently*.

This is related to recent efforts on assessing the performance of daily tasks [10, 11, 26], however, these works do not assess individual actions or identify whether they have been performed as recommended by, say, a recipe. As in the example before, steps with such caveats are often indicated by adverbs describing how actions should be performed. These adverbs (*e.g. quickly, gently, ...*) generalize to different actions and modify the manner of an action. We thus learn these as **action modifiers** (Fig. 1).

To learn action modifiers for a variety of tasks and actions, we utilize the online resource of instructional videos with accompanying narrations. However, this form of supervision is weak and noisy. Not only are the narrations just roughly aligned with the actions in the video, but often the narrated actions may not be captured in the video altogether. For example, a YouTube instructional video might be narrated as "pour in the cream quickly" but the visuals only show the cream already added. In this case the video would not be useful to learn the adverb 'quickly'.

As the main contribution of this paper, we propose the

first method for weakly supervised learning from adverbs, in which we embed relevant video segments in a latent space and learn adverbs as transformations in this space. We collect action-adverb labels from narrations of a subset of tasks in the HowTo100M dataset [33]. The method is evaluated for video-to-adverb retrieval, as well as adverb-to-video retrieval and shows significant improvements over baselines. Additionally, we present a comprehensive ablation study demonstrating that jointly learning a good action embedding is key to learning action modifiers.

2. Related Work

We review works which learn from instructional videos, followed by works using parts-of-speech in video. We then review the related task of object attributes in images and methods which learn embeddings under weak supervision.

Instructional Videos. Movies accompanied by subtitles and scripts have been used for learning from video [12, 13, 25, 47]. However, movies typically focus on talking heads with few object interactions. More recently, instructional videos are a popular source of datasets [1, 33, 44, 60] with hundreds of online videos of the same task. Narrations are used to learn steps of complex tasks [1, 18, 30, 42, 45, 62], and more recently for video retrieval [33], visual grounding [17, 19], action segmentation [60] and learning actions through object state changes [2, 14].

In this work, we offer a novel insight into how these instructional videos can be used beyond step identification. Our work utilizes videos from the recently released HowTo100M dataset [33], learning adverbs and their relevance to critical steps in these tasks.

Learning from Parts-of-Speech in Video. Several problems are at the intersection between language and video: captioning [24, 38, 55, 59], retrieval [9, 16, 21, 31, 33, 52, 54] and visual question answering [15, 56, 57, 61]. The majority of these works use LSTMs or GRUs to combine words into sentence-level features. While some works use learned pooling [32] or attention [55, 56, 57], they do not use knowledge of the words' parts-of-speech (PoS).

A few recent works differentiate words by their PoS tags. Xu et al. [54] learn a joint video-text embedding space after detecting (subject, verb, object) triplets in the input caption. Wray et al. [52] perform fine-grained action retrieval by learning a separate embedding for each PoS before combining these embeddings. Both works focus on verb and noun PoS, as they target action recognition. Alayrac et al. [1] also use verb-noun pairs; the authors use direct object relations to learn unsupervised clusterings of key steps in instructional videos.

While some adverbs are contained in video captioning datasets [24, 59], no prior captioning work models or recognizes these adverbs. The only prior work to utilize adverbs is that of Pang et al. [39] where many adverbs in the ADHA dataset model moods and facial expressions (*e.g.* 'happily', 'proudly'). The work uses full supervision including action bounding boxes. Instead, in this work we target adverbs that represent the manner in which an action is performed, using only weak supervision from narrations.

Object Attributes in Images. Adverbs of actions are analogous with adjectives of objects. Learning adjectives for nouns has been investigated in the context of recognizing object-attribute pairs [4, 7, 20, 34, 36, 37, 50, 51] from images. Both [7, 34] tackle the problem of contextuality of attributes, where the appearance of an attribute can vastly differ depending on the object it applies to. Chen and Grauman [7] formulate this as transfer learning to recognize unseen object-attribute compositions. Misra et al. [34] learn how to compose separate object and attributes classifiers for novel combinations. Instead of using classifiers to recognize attributes, Nagarajan and Grauman [36] model attributes as a transformation of an object's embedding. Our work is inspired by this approach.

While some works learn attributes for actions [28, 43, 58], these detect combinations of specific attributes (*e.g.* 'outdoor', 'uses toothbrush') to perform zero shot recognition and do not consider adverbs as attributes.

Weakly Supervised Embedding. Learned embeddings are commonly used for retrieval tasks, however few works have attempted to learn embeddings under weak supervision [3, 35, 46, 53]. In [3], weak supervision is overcome using a triplet loss that only optimizes distances to the definite negatives and identifies the best matching positive. Two works [35, 46] perform video moment retrieval from text queries without temporal bounds in training. Similar to our approach, both use a form of text-guided attention to find the relevant portion of the video, however these use the full sentence. In our work, we simultaneously embed the relevant portion of the video while learning how adverbs modify actions. We detail our method next.

3. Learning Action Modifiers

The inputs to our model are action-adverb narrations and the accompanying instructional videos. Fig. 2(a) shows a sample instructional video, narrated with "...start by quickly rolling our lemons...", from which we identify the action roll and the adverb quickly (see Sec. 4 for NLP details). After training, our model is able to assess whether videos in the test set, of the same or different action, have been achieved quickly, among other learned adverbs.

We present an overview of our method in Fig. 2. We learn a joint video-text embedding shown in Fig. 2(b), where the relevant video parts are embedded (blue dot) close to the text representation of the adverb-modified action 'roll quickly' (yellow dot). We review how joint video-

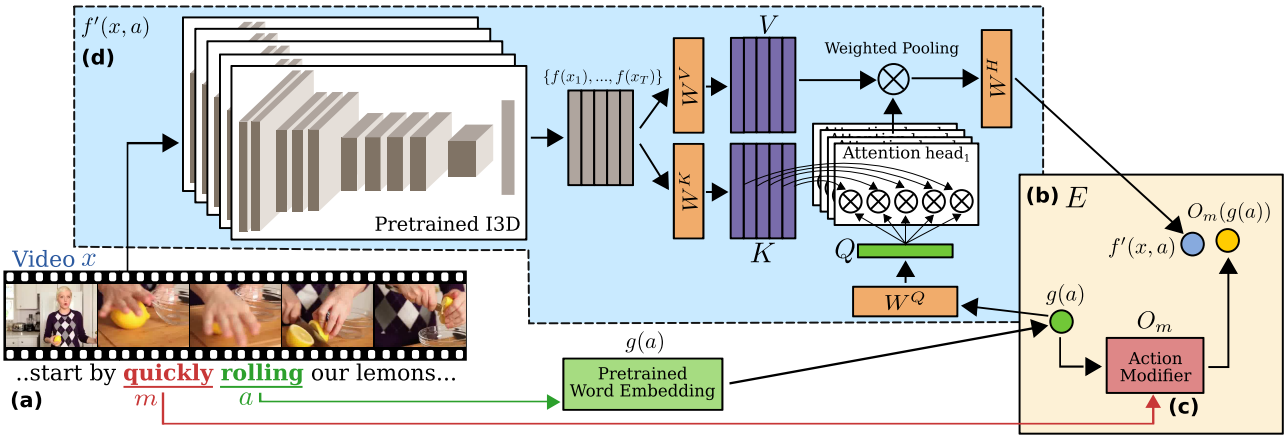

Figure 2. (a) Our input is a video x with the weak label (a, m) for the action and adverb respectively. (b) We aim to learn a joint video-text embedding space for adverb and video retrieval where the embedded video (blue) and action-adverb text representation (yellow) are close. (c) We learn adverbs as action modifiers which are transformations in the embedding space. (d) We embed $f'(x, a)$, a visual representation of the relevant video parts using multi-head scaled dot-product attention where the query is a projection of the action embedding $g(a)$.

text embeddings are typically learned in Sec. 3.1. This section also introduces the notations for the rest of the paper.

Two prime challenges exist in learning the embedding for our problem, *i.e.* learning from adverbs in instructional videos. The first is disentangling the representation of the action from the adverb, allowing us to learn how the same adverb applies across different actions. We propose to learn adverbs as action modifiers, one per adverb, as in Fig. 2(c). In Sec. 3.2 we introduce these action modifiers, which we represent as transformations in the embedding space.

The second challenge is learning the visual representation from the relevant parts of the video in a weakly-supervised manner, *i.e.* without annotations of temporal bounds. In Sec. 3.3, we propose a weakly-supervised embedding function that utilizes multi-head scaled dot-product attention. This uses the text embedding of the action as the query to attend to relevant video parts, as shown in Fig. 2(d).

3.1. Learning an Action Embedding

Our base model is a joint video-text embedding, as in [32, 52, 54]. Specifically, given a set of video clips $x \in X$ with corresponding action labels $a \in A$, our goal is to obtain two embedding functions, one visual and one textual, $f : X \to E$ and $g : A \to E$ such that $f(x)$ and $g(a)$ are close in the embedding space E and $f(x)$ is distant from other action embeddings $g(a')$. These functions f and g can be optimized with a standard cross-modal triplet loss:

$$\mathcal{L}_{triplet} = \max(0, d(f(x), g(a)) - d(f(x), g(a')) + \beta) \; s.t. \; a' \neq a \quad (1)$$

where a' is an action different to a, d is the Euclidean distance and β is the margin, set to 1 in all experiments. We use $g(a)$ as the GloVe [41] embedding of the action's verb, and explain how we replace $f(x)$ by $f'(x, a)$ in Sec. 3.3.

3.2. Modeling Adverbs as Action Modifiers

While actions exist without adverbs, adverbs are by definition tied to the action, and only gain visual representation when attached to one. Although adverbs have a similar effect on different actions, the visual representation is highly dependent on the action. Therefore, we follow prior work from [36] on object-attribute pairs and model adverbs as learned transformations in the video-text embedding space E (Sec. 3.1). As these transformations modify the embedding of the action, we call them **action modifiers**. We learn an action modifier O_m for each adverb $m \in M$, such that

$$O_m(z) = W_m z \quad (2)$$

where z is any point in the embedding space E and $O_m : E \to E$ is a learned linear transform by a weight matrix W_m. In Sec. 5, we test other geometric transformations: fixed translation, learned translation and nonlinear transformation. Each transformation O_m can be applied to any text representation $O_m(g(a))$ or video representation $O_m(f(x))$ in E to add the effect of the adverb m.

A video $x \in X$, labeled with action-adverb pair (a, m), contains a visual representation of the adverb-modified action. We thus aim to embed $f(x)$ close to $O_m(g(a))$. This is equivalent to embedding the inverse of the transformation $O_m^{-1}(f(x))$ near the action $g(a)$. We thus jointly learn our embedding, with the action modifiers O_m, using the sum of two triplet losses. The first focuses on the action:

$$\mathcal{L}_{act} = \max(0, d(f(x), O_m(g(a))) - d(f(x), O_m(g(a'))) + \beta) \; s.t. \; a' \neq a \quad (3)$$

where a' is a different action and d and β are the distance function and margin as in Sec. 3.1. Similarly, we have a

triplet loss that focuses on the adverb, such that:

$$\mathcal{L}_{adv} = \max(0, d(f(x), O_m(g(a))) - d(f(x), O_{\overline{m}}(g(a))) + \beta) \quad (4)$$

where $\overline{m}$ is the antonym of the labeled adverb m (*e.g.* when $m =$ 'quickly', the antonym $\overline{m} =$ 'slowly'). We restrict the negative in $\mathcal{L}_{adv}$ to only the antonym to deal with adverbs not being mutually exclusive. For instance, a video labeled 'slice quickly' does not preclude the slicing being also done 'finely'. However, it surely has not been done 'slowly'. We demonstrate the effect of this choice in Sec. 5.

3.3. Weakly Supervised Embedding

All prior works that learn attributes of objects from images [7, 20, 34, 36, 37] utilize fully annotated datasets, where the object the attributes relate to is the only object of interest in the image. In contrast, we aim to learn action modifiers from video in a weakly supervised manner. Our input is untrimmed videos containing multiple consecutive actions. To learn adverbs, we need the visual representation to be only from the video parts relevant to the action (*e.g.* 'roll' in our Fig. 2 example). We propose using scaled dot-product attention [49], where the embedded action of interest acts as a query to identify relevant video parts.

For each video x, we use a temporal window of size T, centered around the timestamp of the narrated action-adverb pair, containing video segments $\{x_1, x_2, ..., x_T\}$. We start from the visual representation of all segments $f(x) = \{f(x_1), ..., f(x_T)\}$, where $f(\cdot)$ is an I3D network. From this, we wish to learn an embedding of the visual features relevant to the action a, which we call $f'(x, a)$. Inspired by [49], we project $f(x)$ into keys K and values V:

$$K = W^K f(x); \qquad V = W^V f(x) \quad (5)$$

We then set the query $Q = W^Q g(a)$ to be the projection of the action embedding, to weight video segments by their relevance to that action. The attention weights are obtained from the dot product of the keys K and the action query Q. These then pool the values V. Specifically:

$$H(x, a) = \sigma\left(\frac{(W^Q g(a))^\top W^K f(x)}{\sqrt{T}}\right) W^V f(x) \quad (6)$$

where $H(x, a)$ is a single attention head and σ is the softmax function. We train multiple attention heads such that,

$$f'(x, a) = W^H [H_1(x, a), ..., H_h(x, a)] \quad (7)$$

where W^H projects the concatenation of the multiple attention heads $H_i(x, a)$ into the embedding space. We learn h attention head weights: W_i^Q, W_i^K, W_i^V as well as W^H parameters for our weakly-supervised embedding.

It is important to highlight that these weights are jointly trained with the embedding space E, so that $f'(x, a)$ is used instead of $f(x)$ in Equations 3 and 4. We opted to explain our embedding space before detailing how it can be learned in a weakly-supervised manner, to simplify the explanation.

3.4. Weakly Supervised Inference

Once trained, our model can be used to evaluate cross-modal retrieval of videos and adverbs. For video-to-adverb retrieval, we consider a video query x and the narrated action a, and we wish to estimate the adverb m. For example, we have a video and wish to find the manner in which the action 'slice' was performed. We use the learned function $f'(x, a)$ to embed the relevant visual representation for action a in E. We then rank adverbs by the distance from this embedding to all modified actions $\forall m : O_m(g(a))$.

For adverb-to-video retrieval, we consider an action-adverb pair (a, m) as a query, embed $O_m(g(a))$, *e.g.* 'slice finely', and calculate the distance from this text representation to all relevant video segments $\forall x : f'(x, a)$. For both cases, this allows us to use a to query to the weakly supervised embedding, so as to attend to the relevant video parts.

4. Dataset

HowTo100M [33] is a large scale dataset of instructional videos collected from YouTube. Each video has a corresponding narration from manually-entered subtitles or Automatic Speech Recognition (ASR). No ground-truth is available in terms of correct actions or temporal extents.

To test cross-task generalization, we use the same 83 tasks previously used in [62]. These come from cooking, DIY and car maintenance, and are divided into 65 tasks for training and a disjoint set of 18 tasks for testing. However, in [62], only 30 videos per task were used in training. Instead, we use all videos available for these 65 training tasks, where each task consists of 100-500 videos. In total, we have 24,558 videos in training and 1,280 videos in the test set. For these we find action-adverb pairs as follows.

We use the accompanying narrations to discover action-adverb pairs, for both training and testing. First we employ T-BRNN [48] to punctuate the subtitles[1], then perform Part-of-Speech (POS) tagging with SpaCy's English core web model. We search for verb→adverb relationships with the *advmod* dependency, indicating the adverb modifies the verb. We exclude verbs with VBD (past tense) and VBZ (third person singular) tags as these correlate with actions not being shown in the video. For example, in 'sprinkle some finely chopped coriander', 'chopped' is tagged with VBD. Similarly, in 'everything fits together neatly', the verb 'fits' is tagged as VBZ. Examples of the (action, adverb) pairs obtained from the pipeline with the correspond-

[1]Note: YouTube ASR lacks punctuation

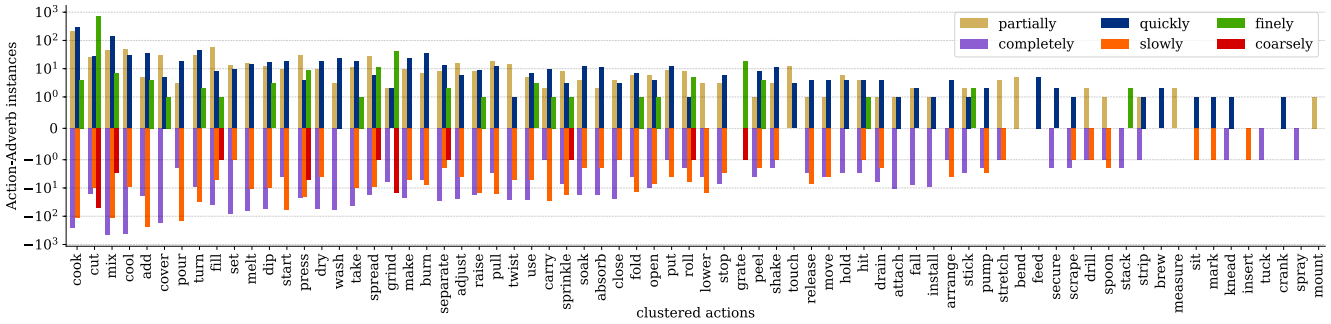

Figure 3. Log-scaled y-axis shows instances of each adverb plotted per action. We display adverbs against their paired antonym (+/- axis).

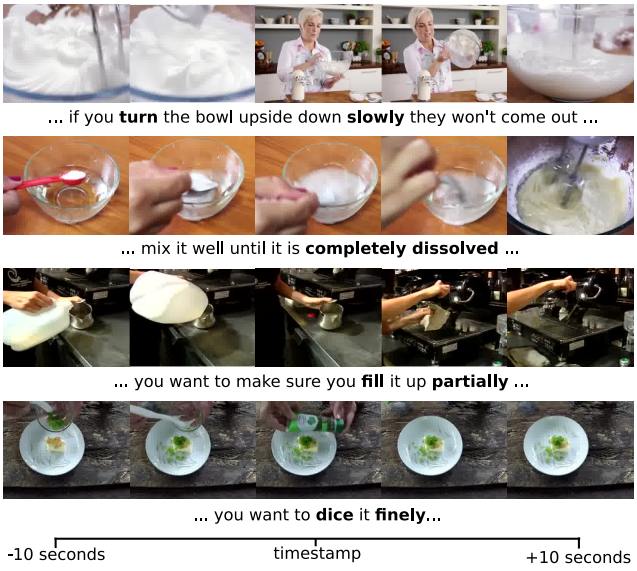

Figure 4. Example videos and narrations, highlighting the action and adverb discovered with our NLP pipeline. In some cases the weak timestamp is a good localization of the action (top), however in others the action is long (second), the timestamp is a poor match (third), or the action is not captured in the video (bottom).

ing video snippets are shown in Fig. 4. Additionally, we manually filter actions and adverbs that are not visual, *e.g.* 'recommend' and 'normally', respectively. We explored automatic approaches such as word concreteness scores [5], but found these approaches to be unreliable. We also group verbs into clusters to avoid synonyms as in [8], *i.e.* we consider 'put' and 'place' as the same action. From this process, we obtain 15,266 instances of action-adverb pairs.

However, these have a long tail of adverbs that are mentioned only a handful of times. We restrict our learning to 6 commonly used adverbs, that come in 3 pairs of antonyms: 'partially'/'completely', 'quickly'/'slowly' and 'finely'/'coarsely'. These adverbs appear in 263 unique action-adverb pairs with 72 different actions. We show the distribution of adverbs per action in Fig. 3. While our training is noisy, *i.e.* actions might not appear in the video (refer to Fig. 4 bottom), we clean the test set for accurate evaluation of the method. We only consider test set videos where the action-adverb is present in the video and appears within the 20 seconds around the narration timestamp. These correspond to 44% of the original test set, which is comparable to the 50% level of noise reported by the authors in [33].

This results in 5,475 action-adverb pairs in training and 349 in testing. We consider the mean timestamp between the verb and adverb narrations as the weak supervision for the action's location. These action-adverb weak timestamp annotations and accompanying code are publicly available[2].

5. Experiments

We first describe the implementation details of our method, followed by the metrics we use for evaluation. We then present our results against those of baselines and evaluate the contribution of the different components.

Implementation Details. We sample all videos at 25fps and scale to a height of 256 pixels. We use I3D [6] with 16 frame segments, pre-trained on Kinetics [22], for both RGB and optical flow. We concatenate these to create 2048D features, extracted once per second as in [62], for $T = 20$ seconds around the narration timestamp.

In all experiments, our embedding space E is 300D, the same as the GloVe word representation [41]. We initialize the action embeddings with the verb's GloVe vector, pre-trained on the Wikipedia and Gigaword corpora. The action modifiers O_m are initialized with the identity matrix such that they have no effect at first. For our scaled dot-product attention, Q is of size 75×1 and K and V are of size $75\times T$. We use 4 attention heads in $f'(x, a)$.

All our models are trained with the Adam optimizer [23] for 1000 epochs with a batch size of 512 and a learning rate of 10^{-4}. To aid disentangling the actions and adverbs, we first let the model learn only actions (optimized by $\mathcal{L}_{triplet}$) for 200 epochs before introducing the action modifiers. The weights of the action modifiers W_m (Eq. 2) are then learned at a slower rate of 10^{-5}.

Evaluation Metric. We report mean Average Precision (mAP) for video-to-adverb and adverb-to-video retrieval. For **video-to-adverb** given a video and the narrated

[2] https://github.com/hazeld/action-modifiers

action we rank the 6 adverbs' relevance. For **adverb-to-video** given an adverb query (*e.g.* 'slowly'), we rank videos by the distance of each video labelled with its associated action (*e.g.* 'put') to the text embedding of the verb-adverb (*e.g.* 'put slowly') and calculate mAP across the 6 adverbs.

We also report mAP where we restrict the retrieval to the adverb and its antonym, which we refer to as the **Antonym** setting. This 'Antonym' metric better represents the given labels, therefore we use it for the ablation study. To clarify, we may have a video narrated 'cut coarsely'. We are thus confident the cut was not performed 'finely', however we cannot judge the speed of ('quickly' or 'slowly'). In Antonym video-to-adverb, there are only two possible adverbs to retrieve, thus we report Precision@1 (P@1) which is the same as binary classification accuracy. Similarly, we report mAP Antonym for adverb-to-video retrieval, where we only rank videos labeled with the adverb or its antonym.

5.1. Comparative Results

We first compare our work to baselines. Since ours is the first work to learn from adverbs in videos, we adapt methods that learn attributes of objects in images for comparison, as this is the most similar existing task to ours. In this adaptation, actions replace objects, and adverbs replace attributes/adjectives.

We compare to RedWine [34] and AttributeOp [36] as well as the LabelEmbed baseline proposed in [34] which uses GloVe features in place of SVM classifier weights. We replace the image representation by a uniformly weighted visual representation of video segments. Similar to our evaluation, we report results when the action is given in testing, referred to as the 'oracle' evaluation in [36]. Furthermore, for a fair comparison, we use only the antonym as the negative in each method's loss, as we do in Eq. 4. AttributeOp proposes several linguistic inspired regularizers; we report the best combination of regularizers for our dataset — the auxiliary and commutative regularizers. We also compare to random chance and a naive binary classifier per adverb pair. This classifier is analogous to the Visual Product baseline used in [34, 36]. We report on both versions of this baseline, a Linear SVM which trains a binary one-vs-all classifier per adverb (Classifier-SVM) and a 6-way MLP of two fully connected layers (Classifier-MLP). In video-to-adverb, we rank adverbs by classifiers' confidence scores, as in [36]. In adverb-to-video, we use the confidence of the corresponding classifier or MLP output to rank videos.

Comparative results are presented in Table 1. Our method outperforms all baselines for video-to-adverb retrieval, both when comparing against all adverbs and when restricting the evaluation to antonym pairs. We see that AttributeOp is the best baseline method, generally performing better than both RedWine and LabelEmbed. The two latter methods work on a fixed visual feature space, thus are prone to errors when the features are non-separable in that space. We can also see that LabelEmbed performs better than RedWine across all metrics, demonstrating that GloVe features are better representations than SVM classifier weights. While AttributeOp marginally outperforms our approach on adverb-to-video 'All', it underperforms on all other metrics, including our main objective, estimating the correct adverb over its antonym for a video query.

Method	video-to-adverb		adverb-to-video	
	Antonym	All	Antonym	All
Chance	0.500	0.408	0.511	0.170
Classifier-SVM	0.605	0.532	0.563	0.264
Classifier-MLP	0.685	0.602	0.603	0.304
RedWine [34]	0.693	0.594	0.595	0.290
LabelEmbed [34]	0.717	<u>0.621</u>	<u>0.618</u>	0.297
AttributeOp [36]	<u>0.728</u>	0.612	0.597	**0.350**
Ours	**0.808**	**0.719**	**0.657**	<u>0.329</u>

Table 1. Comparative Evaluation. Best performance in **bold** and second best <u>underlined</u>. We report results for both video-to-adverb and adverb-to-video retrieval with results restricted to the adverb and its antonym (Antonym) and when unrestricted (All).

5.2. Qualitative Results

Fig. 5 presents video examples. For each, we demonstrate attention weights for several action queries. Our method is able to successfully attend to segments relevant to various query actions. The figure also shows predicted actions, and predicted adverb when using the ground-truth action as the query. Our method is able to predict the correct adverb. In the last example, predicted actions are incorrect, but the method correctly identifies a relevant segment and that the action was done 'slowly'. We provide further insights into the learned embedding space in supplementary.

5.3. Ablation Study

We report 4 ablation studies on the various aspects of the method: the choice of action modifier transformation $O_m(\cdot)$, our scaled dot-product attention, the contributions of the loss functions, and the length of the video (T). We focus on video-to-adverb retrieval in the ablation using the Antonym P@1 metric, as this allows us to answer questions like: "was the 'cut' performed 'quickly' or 'slowly'?".

Action Modifier Representation. In Table 2 we examine different representations for the action modifiers $O_m(\cdot)$ (Eq. 2). We compare to a fixed translation from the GloVe representation of the adverb (m), which is not learned, to three learned representations. First, a learned translation

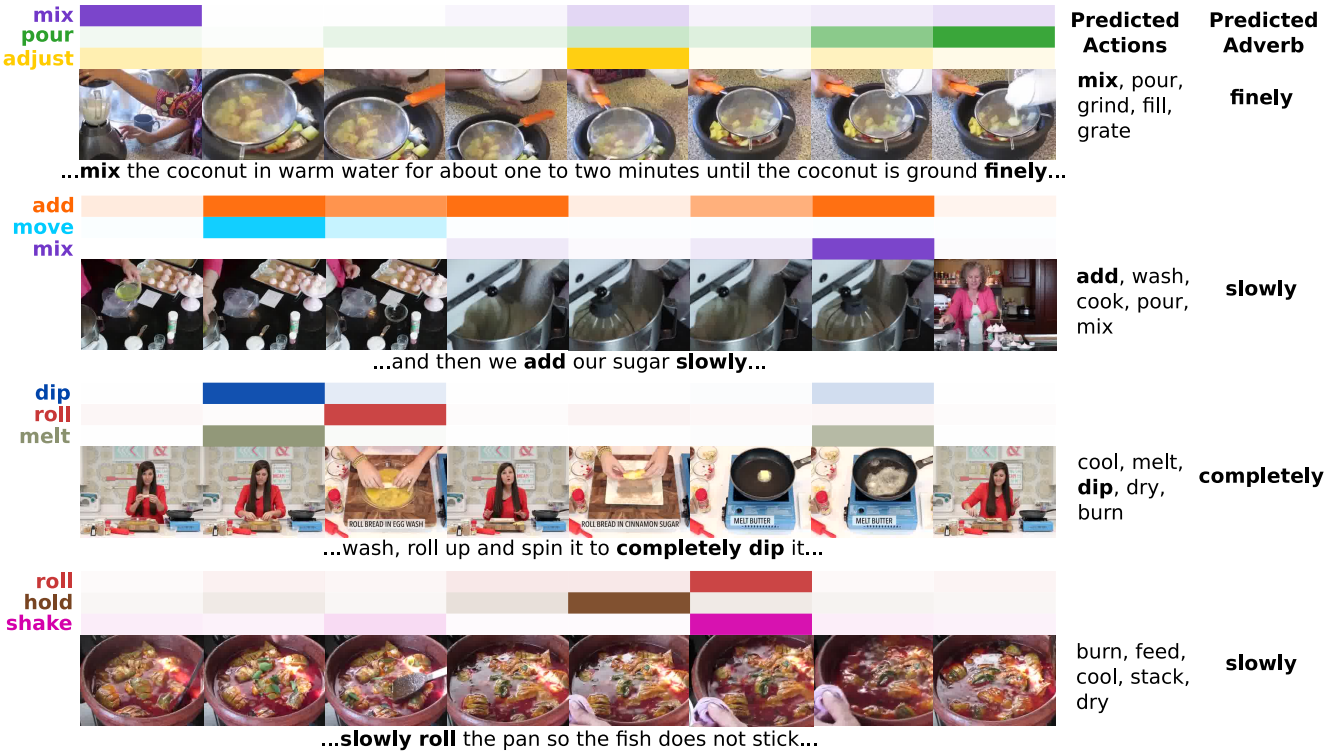

Figure 5. **Qualitative Results.** Temporal attention values from several action queries. The intensity of the color indicates the attention value. Recall that we use the narrated action to weight the relevance of video segments. Using that, we display the top-5 predicted actions, as well as the correctly predicted adverb for all cases.

$O_m(z) =$	Dimension	Learned	P@1
$z + GloVe(m)$	1D		0.735
$z + b_m$	1D	✓	0.749
$W_m z$	2D	✓	**0.808**
$W_{m_2}\text{ReLU}(W_{m_1} z + b_m)$	2D	✓	0.742

Table 2. Comparison of action modifier representation $O_m(\cdot)$. The linear transformation choice clearly improves results.

vector b_m initialized from the GloVe embedding is used. Second, our chosen representation - a 2D linear transformation with matrix W_m as in Eq. 2. Third, we learn a non-linear transformation implemented as two fully connected layers, the first with a ReLU activation.

Results show the linear transformation clearly outperforms a vector translation or the non-linear transformation. The translation vector does not having enough capacity to represent the complexity of the adverb, while the nonlinear transform is prone to over-fitting.

Temporal Attention. In Table 3, we compare our proposed multi-head scaled dot-product attention (Sec. 3.3) with alternative approaches to temporal aggregation and attention. In this comparison, we also report action retrieval results, with video-to-action mAP. That is, given the embedding of the video $f'(x, a)$ queried by the ground-truth action, we rank all actions in the embedding $\forall a : g(a)$ by their distances to the visual query and evaluate the rank of the correct action. Our method does not aim for action retrieval as it assumes knowledge of the ground-truth action, however this metric evaluates the quality of the weakly supervised embedding space. Results are compared to:

- **Single**: uses only a one-second clip at the timestamp.
- **Average**: uniformly weights the T features.
- **Attention from [29]**: widely used class agnostic attention, calculating attention with two fully connected layers, $f'(x, a) = \sigma(w_1 \tanh(W_2 f(x)))W_3 f(x)$.
- **Class-specific Attention**: a version of the above with one attention filter per action class.
- **Ours w/o two-stage optimization**: our attention without the first 200-epoch stage of learning action triplets without learning adverbs/modifiers.
- **Ours**: our attention as described in Sec. 3.3.

Table 3 demonstrates superior performance of our method for the learning of action embeddings and, as a consequence, better learning of action modifiers. These results also demonstrate the challenge of weak-supervision, with video-to-action only performing at 0.246 mAP when considering only one second surrounding the narrated action. This improves to 0.692 with our method.

Method	Action	Adverb
Single	0.246	0.705
Average	0.257	0.716
Attention from [29]	0.235	0.708
Class-specific Attention	0.401	0.728
Ours w/o two-stage optimization	0.586	0.774
Ours	**0.692**	**0.808**

Table 3. Comparison of temporal attention methods. We report video-to-action retrieval mAP and video-to-adverb retrieval P@1.

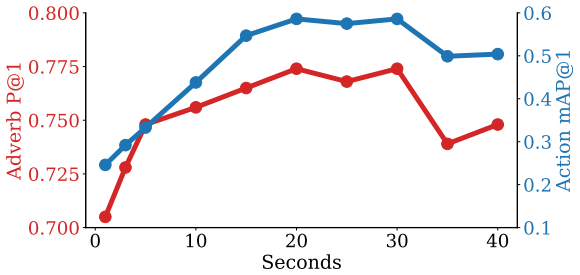

Figure 6. Performance as T increases. Blue (axis and plot) shows video-to-action retrieval mAP while red shows video-to-adverb retrieval with Antonym P@1.

Loss Functions. We also evaluate the need for two separate loss functions (Eqs. 3 and 4). As an alternative approach we use a single loss where the negative contains a different action, a different adverb or both. This performs worse by 0.03 P@1. Using both losses, but with another adverb as opposed to only the antonym $\overline{m}$ in Equation 4 also results in worse performance by 0.04 P@1.

Effect of T. In Fig. 6, we evaluate how the length of the video (T) extracted around the weak timestamp affects the model (Sec. 3.3). For larger T, videos are more likely to contain the relevant action, but also other actions. Our embedding function $f'(x, a)$ is able to ignore other actions in the video, up to a point, and successfully learn to attend to the relevant parts given the query action, resulting in better performance with $T \in \{20 \,..\, 30\}$.

Comparison with Action Localization. In this work, we perform weakly supervised embedding to learn action modifiers by attending to action relevant segments. Here, we test whether weakly supervised action localization can be used instead of our proposed attention, to locate key segments before learning action modifiers.

We use published code of two state-of-the-art weakly supervised action localization methods: W-TALC [40] and CMCS [27]. First, we test the output of these methods with a binary adverb-antonym classifier (Classifier-MLP as in Sec. 5.1). We also test these methods in combination with our embedding and action modifier transformations. For this, we use the methods' predicted action-relevant segments, and average their representation to replace $f'(x, a)$ (Avg). Finally, we combine these relevant segments with our scaled dot-product attention (SDP).

Method	Attention	Adverb Rep	P@1
W-TALC [40]	Avg	Classifier-MLP	0.705
	Avg	Action Modifiers	0.739
	SDP	Action Modifiers	0.768
CMCS [27]	Avg	Classifier-MLP	0.696
	Avg	Action Modifiers	0.699
	SDP	Action Modifiers	0.705
Ours	SDP	Action Modifiers	**0.808**

Table 4. Comparison of our method (Ours) to weakly supervised action localization methods, with and without our scaled dot-product (SDP) and action modifier representations.

From Table 4 we can conclude that using the output of a weakly-supervised localization method is insufficient, and our joint optimization performs best. Worth noting, localizing the action using W-TALC followed by averaging relevant segments outperforms averaging all segments (0.739 vs. 0.716 from Table 3). This shows that W-TALC is capable of finding some relevant segments. This is further improved by our scaled dot-product attention.

6. Conclusion

This paper presents a weakly supervised method to learn from adverbs in instructional videos. Our method learns to obtain and embed the relevant part of the video with scaled dot product attention, using the narrated action as a query. The method then learns action modifiers as linear transformations on the embedded actions; shared between actions. We train and evaluate our method on parsed action-adverb pairs sourced from YouTube videos of 83 tasks. Results demonstrate that our method outperforms all baselines, achieving 0.808 mAP for video-to-adverb retrieval, when considering the adverb versus its antonym.

Future work will involve learning from few shot examples in order to represent a greater variety of adverbs as well as exploring applications to give feedback to people guided by instructional videos or written instructions.

Acknowledgements: Work is supported by an EPSRC DTP, EPSRC GLANCE (EP/N013964/1), Louis Vuitton ENS Chair on Artificial Intelligence, the MSR-Inria joint lab and the French government program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). Part of this work was conducted during H. Doughty's internship at INRIA Willow Team. Work uses publicly available dataset.

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4575–4583, 2016. 1, 2

[2] Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2127–2136, 2017. 2

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016. 2

[4] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, 2013. 2

[5] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911, 2014. 5

[6] 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 (CVPR)*, pages 6299–6308, 2017. 5

[7] Chao-Yeh Chen and Kristen Grauman. Inferring analogous attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 200–207, 2014. 2, 4

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[9] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9346–9355, 2019. 2

[10] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6057–6066, 2018. 1

[11] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7862–7871, 2019. 1

[12] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis R Bach, and Jean Ponce. Automatic annotation of human actions in video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 1491–1498, 2009. 2

[13] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! my name is... buffy"–automatic naming of characters in tv video. In *British Machine Vision Conference (BMVC)*, volume 2, page 6, 2006. 2

[14] Alireza Fathi and James M Rehg. Modeling actions through state changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2579–2586, 2013. 2

[15] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[16] 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 (CVPR)*, pages 5803–5812, 2017. 2

[17] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[18] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 137–153. Springer, 2016. 2

[19] De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2183–2192, 2017. 2

[20] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391, 2015. 2, 4

[21] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. Objects2action: Classifying and localizing actions without any video example. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4588–4596, 2015. 2

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. 5

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[25] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from

movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 2

[26] Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato. Manipulation-skill assessment from videos with spatial attention network. *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019. 1

[27] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1298–1307, 2019. 8

[28] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3344. IEEE, 2011. 2

[29] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7834–7843, 2018. 7, 8

[30] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015. 1, 2

[31] Pascal Mettes and Cees GM Snoek. Spatial-aware object embeddings for zero-shot localization and classification of actions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4443–4452, 2017. 2

[32] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2, 3

[33] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 4, 5

[34] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1792–1801, 2017. 2, 4, 6

[35] 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 (CVPR)*, pages 11592–11601, 2019. 2

[36] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018. 2, 3, 4, 6

[37] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019. 2, 4

[38] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6504–6512, 2017. 2

[39] Bo Pang, Kaiwen Zha, and Cewu Lu. Human action adverb recognition: Adha dataset and a three-stream hybrid model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2325–2334, 2018. 2

[40] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 8

[41] 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)*, pages 1532–1543, 2014. 3, 5

[42] Alexander Richard, Hilde Kuehne, and Juergen Gall. Action sets: Weakly supervised action segmentation without ordering constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5996, 2018. 2

[43] Amir Rosenfeld and Shimon Ullman. Action classification via concepts and attributes. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1499–1505. IEEE, 2018. 2

[44] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 862–871, 2019. 2

[45] Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Unsupervised semantic parsing of video collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4480–4488, 2015. 1, 2

[46] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. wman: Weakly-supervised moment alignment network for text-based video segment retrieval. *arXiv preprint arXiv:1909.13784*, 2019. 2

[47] Makarand Tapaswi, Martin Buml, and Rainer Stiefelhagen. Book2Movie: Aligning Video scenes with Book chapters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[48] Ottokar Tilk and Tanel Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*, 2016. 4

[49] 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 (NeurIPS)*, pages 5998–6008, 2017. 4

[50] Xiaoyang Wang and Qiang Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2120–2127, 2013. 2

[51] Yang Wang and Greg Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 155–168. Springer, 2010. 2

[52] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[53] Jian Xu, Chunheng Wang, Cunzhao Shi, and Baihua Xiao. Weakly supervised soft-detection-based aggregation method for image retrieval. *arXiv preprint arXiv:1811.07619*, 2018. 2

[54] 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 *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 2, 3

[55] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4507–4515, 2015. 2

[56] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4709–4717, 2017. 2

[57] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3173, 2017. 2

[58] Rowan Zellers and Yejin Choi. Zero-shot activity recognition with verb attribute induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 946–958, 2017. 2

[59] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Generation for user generated videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 609–625. Springer, 2016. 2

[60] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[61] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017. 2

[62] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3537–3545, 2019. 1, 2, 4, 5