How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs

Hazel Doughty  
University of Amsterdam  
hazel.doughty@uva.nl

Cees G.M. Snoek  
University of Amsterdam  
cgmsnoek@uva.nl

Abstract

We aim to understand how actions are performed and identify subtle differences, such as ‘fold firmly’ vs. ‘fold gently’. To this end, we propose a method which recognizes adverbs across different actions. However, such fine-grained annotations are difficult to obtain and their long-tailed nature makes it challenging to recognize adverbs in rare action-adverb compositions. Our approach therefore uses semi-supervised learning with multiple adverb pseudo-labels to leverage videos with only action labels. Combined with adaptive thresholding of these pseudo-adverbs we are able to make efficient use of the available data while tackling the long-tailed distribution. Additionally, we gather adverb annotations for three existing video retrieval datasets, which allows us to introduce the new tasks of recognizing adverbs in unseen action-adverb compositions and unseen domains. Experiments demonstrate the effectiveness of our method, which outperforms prior work in recognizing adverbs and semi-supervised works adapted for adverb recognition. We also show how adverbs can relate fine-grained actions.

1. Introduction

This paper aims to recognize fine-grained differences between actions such as whether a person is swimming slowly or quickly or cutting evenly or unevenly. Understanding how actions are performed is key to understanding the actions themselves and their outcomes. Improved perception of the action manner would allow both humans and robots to better imitate actions, as well as better discrimination between fine-grained action categories, where the difference can simply be how much an object moves [18]. Previous works can address the question of what is happening in a video [79], when an action is happening [67], who is performing an action [72] and where it is taking place [43]. However, very few works have looked at how actions happen, as we do in this paper.

In language, how an action is performed can be described with adverbs, thus we focus on recognizing such adverbs. Two works have previously investigated adverb recognition [12, 51]. However, these works either focus on adverbs describing facial expressions and moods [51] or only studied a handful of adverbs [12] limiting the ways to answer “how is the action being performed?”’. This highlights a key challenge in learning adverbs and more generally fine-grained video understanding: the time-consuming data collection.

As our main contribution, we propose to reformulate the adverb recognition problem as a semi-supervised learning problem. In Sec. 3, we propose the first method for semi-supervised learning of adverbs, in which we apply multiple adverb pseudo-labels to actions and use an adaptive threshold to cope with the long-tail distribution of adverbs. In Sec. 4, we create several new adverb recognition benchmarks by automatically mining action-adverb pairs from the captions in existing video retrieval datasets [29, 64, 69]. Alongside this we propose two new tasks for addressing how actions happen: first recognizing adverbs in unseen compositions and second in recognizing adverbs across domains. In Sec. 5, we demonstrate our multi-adverb pseudo-labeling approach obtains a considerable improvement over prior works in recognizing seen compositions of verbs and adverbs as well as improving generalization in these new tasks.
2. Related Work

We first review works focused on fine-grained understanding of actions followed by video retrieval. We then examine works which have focused specifically on adverbs. Finally, we look at semi-supervision for other vision tasks.

Fine-Grained Action Understanding. Recent datasets have focused on fine-grained actions [9, 18, 36, 59]. For instance, in FineGym [59] a model must distinguish between ‘salto forwards’ and ‘salto backwards’. While some actions are similar, the majority of works [6,14,30,37,40,63,65,77] model actions as distinct categories leaving the model to implicitly learn similarities. Some works instead explicitly model actions as compositions of components, either through sub-actions [52,53] or verbs and noun combinations with [26, 45] or without [3,9,24,44,60,80] the noun’s spatial location. We instead, look at fine-grained differences between actions by recognizing adverbs in combination with different verbs.

Other works recognize actions through combinations of specified attributes [39,56,57,73,75]. For instance, with Temporal Query Networks, Zhang et al. [75] propose to determine the correct attributes by first attending to the most relevant video parts with an attribute-focused query. The attributes studied in these works do not consider adverbs, instead they indicate the presence of an object, a person’s pose or the number of repetitions of an action.

Video Retrieval. Potentially more fine-grained than action recognition is video-text retrieval, which aims to retrieve the correct caption describing the video. The majority of such works create sentence-level features with recurrent networks [10,21,48], learned pooling [46] or transformers [15, 41, 71, 78]. While retrieval datasets [21,29,33,50,64,69] do contain adverbs, models use verbs and nouns to distinguish videos as they are more frequent [66]. Rather than relying on a sentence encoding to indicate the distinctive elements of a caption, some prior works focus on certain parts of speech [8,66,70]. Again, the focus is on verbs and nouns, with Wray et al. [66] learning separate embeddings for each and Chen et al. [8] learning a hierarchical text encoding from verbs, nouns and the semantic relation between them. We instead focus on understanding adverbs and how these apply to different verbs. We obtain new, more varied, action-adverb annotations from three video retrieval datasets.

Adverbs. Some works have studied individual adverbs. For instance, Benaim et al. [4] identify whether videos are played quickly, Epstein et al. [13] recognize whether an event occurred accidentally and Heidarivincheh et al. [20] pinpoint when an action has finished completely. There are two prior works which look at recognizing adverbs more generally. Pang et al. [51] propose a fully-supervised method using video, pose and expression features. The adverbs in this work focus primarily on moods and expressions such as solemnly and excitedly. Doughty et al. [12] learn adverbs from weak supervision with attention locating the video segments relevant to the action. Adverbs are then learned as transformations in an embedding. This approach is still label-hungry, requiring sufficient adverb-labeled actions for all action-adverb compositions. We instead utilize action-only labeled videos to recognize adverbs in both seen and unseen compositions. For this we introduce three new adverb datasets, significantly increasing the number of adverbs from 6 to 34 and the number of compositions from 263 to 1,550.

Semi-Supervision. Many strategies have been explored for semi-supervised learning such as pseudo-labeling [1,19,31], consistency regularization [2,5,61,62], generative models [49,54] and fine-tuning self-supervised models [74]. For instance, Lee [31] proposes an efficient method for pseudo-labeling where one-hot labels are obtained for an unlabeled sample by taking the highest confidence prediction. Sohn et al. propose the consistency regularization approach FixMatch [62], where the loss aims to make the label predicted for two augmented versions of an image consistent.

Several works focus on semi-supervised learning for video [17,27,61,68]. TCL by Singh et al. [61] maximizes the prediction similarities between different speeds of a video. Xiong et al. [68] target consistency in the pseudo-labels predicted by RGB, optical flow and temporal gradient streams. Gavrilyuk et al. [17] also propagate pseudo-labels between modalities, but instead aim to distill motion information so downstream tasks only need the RGB modality in training.

Since these works target image, object or action recognition, they are unsuitable for adverbs. Adverbs are compositional both with actions and other adverbs and these compositions have a long-tailed distribution. We propose a semi-supervised approach to learn adverbs via multi-adverb pseudo-labeling and adaptive thresholding to address these challenges. We also demonstrate how our approach can improve generalization to unseen action-adverb compositions.

3. Semi-supervised Learning of Adverbs

Our work aims to understand how an action is being performed in a video by predicting the adverb(s) applicable to that action. An overview of our approach can be seen in Fig. 2. Labeled data can be used to learn to recognize adverbs in composition with different actions (Sec. 3.1). However, a key challenge in understanding subtle differences, such as adverbs, is lack of labeled data. In this work we propose to better learn adverbs with semi-supervised learning by creating pseudo-adverb labels for video clips with action labels only (Sec. 3.2). We observe that multiple adverbs can apply to the same action, therefore we propose to better utilize the available data with our multi-adverb pseudo-labeling (Sec. 3.3). Another challenge is the long-tailed nature of adverbs. We use adaptive thresholding on the adverb pseudo-labels so that our approach is effective on all adverbs, not only the most frequent (Sec. 3.4).
3.1. Learning Adverbs with Labeled Data

Given a video clip \( x \in X \) and a label for the action of interest \( a \in A \), the goal of adverb recognition is to correctly predict the adverb \( \tilde{m} \) which applies to action \( a \). Since many adverbs are not mutually exclusive and can simultaneously apply to the same action, we particularly focus on distinguishing between the labeled adverb \( m \) and its antonym \( \text{ant}(m) \). In the supervised case we learn to recognize adverbs with a labeled set of videos \( S = \{(x, a, m)\} \).

As in prior work [12], we learn adverbs in a video-text embedding space as this allows actions and adverbs to be compositional. The goal is to embed the parts of the video relevant to the action close to a text representation of that action modified by the adverb. Specifically, we learn a visual embedder \( f : X, A \rightarrow E \) and a textual embedder \( g : A, M \rightarrow E \). We aim for \( f(x, a) \) and \( g(a, m) \) to be close in the embedding space \( E \) and \( f(x, a) \) to be far from embeddings with other actions \( g(a', m) \) and with the antonym adverb \( g(a, \text{ant}(m)) \). We do this with two triplet losses:

\[
\mathcal{L}_{\text{act}}(S) = \sum_{(x,a,m) \in S} \max\{0, \text{dist}(f(x,a), g(a,m)) - \text{dist}(f(x,a), g(a', m)) + \gamma_1\} \quad \text{s.t.} \quad a \neq a', \tag{1}
\]

\[
\mathcal{L}_{\text{adv}}(S) = \sum_{(x,a,m) \in S} \max\{0, \text{dist}(f(x,a), g(a,m)) - \text{dist}(f(x,a), g(a, \text{ant}(m))) + \gamma_2\}, \tag{2}
\]

where \( \text{dist} \) is a distance metric and \( \gamma_1 \) and \( \gamma_2 \) are margins.

3.2. Pseudo-Labeling Adverbs

Now we consider how we can improve the learning of adverbs by utilizing video clips with only action labels. To make use of this type of data we propose to label the action clips with adverbs. Formally, we have a video set without adverb labels \( U = \{(x, a)\} \). For each video clip \( x \) with action label \( a \), we can create a single adverb pseudo-label \( \tilde{m} \) by selecting the adverb in the closest text representation as the pseudo-label. Let \( d(x, a, m) = \text{dist}(g(a, m), f(x, a)) \):

\[
\tilde{m} = m_n \quad \text{where} \quad n = \arg\min_i d(x, a, m_i), \tag{3}
\]

where \( m_n \) is a label indicating adverb \( n \). For the action-only videos we can then use \( \tilde{m} \) in place of \( m \) in \( \mathcal{L}_{\text{adv}} \) (Equation 2). This gives us the overall loss:

\[
\mathcal{L} = \mathcal{L}_{\text{act}}(S) + \mathcal{L}_{\text{adv}}(S) + \mathcal{L}_{\text{act}}(U) + \mathcal{L}_{\text{adv}}(U). \tag{4}
\]

3.3. Multi-Adverb Pseudo-Labeling

While actions in the supervised set \( S \) are labeled with a single adverb, the majority of adverbs are not mutually exclusive, meaning multiple adverbs can apply to a single action. We thus propose multi-adverb pseudo labeling. To do this we take the top \( k \) most confident adverbs and let the adverb pseudo-label \( \tilde{m} \) to be a set of pseudo-labels:

\[
\tilde{m} = \{m_n\} \quad \text{s.t.} \quad n \in \text{topk}_i(\text{conf}(x, a, m_i)), \tag{5}
\]

where

\[
\text{conf}(x, a, m) = \frac{e^{d(x,a,m)}}{e^{d(x,a,m)}} \tag{6}
\]

With this definition of \( \text{conf}(x, a, m) \) we take the most confident to mean the greatest relative difference between the adverb and its antonym rather than the closest adverbs.

Now we have multiple adverb pseudo labels for each of the action-only labeled videos in \( U \). We optimize for each pseudo-labeled adverb, meaning the overall loss becomes:

\[
\mathcal{L} = \mathcal{L}_{\text{act}}(S) + \mathcal{L}_{\text{adv}}(S) + \sum_{\tilde{m}} (\mathcal{L}_{\text{act}}(U) + \mathcal{L}_{\text{adv}}(U)). \tag{7}
\]
3.4. Adaptive Adverb Thresholding

The problem of recognizing adverbs is naturally long-tailed. Not only are some adverbs much more common than others, but certain compositions of actions and adverb are also more frequent. Using our multi-adverb pseudo-labeling we are able to make better use of the available data. However, it has a tendency to only select the most frequent adverbs, as these are the adverbs it is most confident in predicting.

We take inspiration from semi-supervised object detection where the long-tail is also present [35] and propose to use adaptive thresholding. The threshold is dynamically adjusted for each adverb \( m \). Not only does this mean that the threshold is increased for the more confident adverbs so that fewer noisy pseudo-labels are used, but importantly the threshold is lowered for the adverbs with fewer confident predictions, meaning they are no longer unrepresented in the pseudo-labels. We adapt an initial threshold \( \tau \) to an adverb-specific threshold \( \tau_m \) as follows:

\[
\tau_m = \left( \frac{\sum_{U:x\in\tilde{m}} \text{conf}(x, a, m)}{N \sum_{U} \left| \tilde{m} \right|} \right)^{\lambda} \tau,
\]

where \( N \) is the number of adverbs. The sum of confidence scores for an adverb \( m \), \( \sum_{U:x\in\tilde{m}} \text{conf}(x, a, m) \), acts as an approximation of the model’s overall confidence for predicting this adverb over its antonym. We then divide this by the average number of pseudo-labels per adverb. \( \lambda \) is a smoothing factor which controls the amount the model focuses on underrepresented adverbs. With \( \lambda=0 \), all adverbs use the original threshold \( \tau \). The adverb-specific threshold \( \tau_m \) is applied to filter the available pseudo-labels, so that only the pseudo labels with \( \text{conf}(x, a, m) > \tau_m \) for \( m \in \tilde{m} \) are used.

4. Adverb Datasets and Tasks

We evaluate our approach on HowTo100M Adverbs [12] which mined adverbs from 83 tasks in HowTo100M [47]. Since the annotations were obtained from automatically transcribed narrations of instructional videos, they are noisy; \( \sim 44\% \) of the annotated action-adverb pairs are not visible in the video clip. The dataset contains 5,824 clips annotated with action-adverb pairs from 72 verbs and 6 adverbs. A clear limitation of this dataset is the small number of adverbs it contains, we thus create three new adverb datasets from existing video retrieval datasets [29, 64, 69]: VATEX Adverbs, MSR-VTT Adverbs and ActivityNet Adverbs. These contain less noise and a greater variety of adverbs.

4.1. Adverb Annotations from Video Captions

We extract verb-adverb annotations for videos in existing video-text datasets to obtain three new adverb datasets. From available datasets [16, 21, 23, 25, 29, 32–34, 38, 50, 55, 58, 64, 69, 76] we find VATEX [64], ActivityNet Captions [29] and MSR-VTT [69] contain the best variety of adverbs with sufficient instances. VATEX consists of 35k 10 second video clips, each with 10 English captions, resulting in a total of 260k captions. In MSR-VTT each clip is 10-30 seconds and has 20 captions giving a total of 10k clips and 200k captions. ActivityNet Captions contains 20k videos with an average of 3.65 temporally localized sentences per video, resulting in a total of 100k clips and matching captions. Each dataset uses YouTube videos, thus some videos are no longer available. At the time of collection we obtained: 32,161 video clips for VATEX, 7,511 for ActivityNet and 5,197 for MSR-VTT.

Extracting Adverb Annotations. To extract adverb annotations from the captions in these datasets we search for adverbs and their corresponding verbs. We use SpaCy’s English core web model where RoBERTa [42] performs Part-of-Speech tagging and dependency parsing on each caption. We search for verbs which have adverbs as children, excluding any verbs with a negative dependency to another word. We filter out non-visual adverbs, adverbs whose antonym doesn’t appear and adverbs which appear less than 10 times or only appear in combination with a single action. The resulting verbs and adverbs from the three datasets are manually clustered, starting with the clusters from [12]. This process forms 137 verb clusters and 34 adverb clusters in 17 adverb-antonym pairs. Fig. 3 shows examples of the video clips alongside the discovered action-adverb pairs.

Adverb Datasets. This results in three adverb datasets: VATEX Adverbs, ActivityNet Adverbs and MSR-VTT Adverbs which we make publicly available alongside the code\(^1\). Table 1 shows statistics of each. VATEX Adverbs is the largest with 34 adverbs appearing across 135 actions to form 1,550 unique action-adverb pairs. The distribution of actions, adverbs and their compositions are heavily long-tailed (see Fig. 4). Each dataset considers many more adverbs than the existing HowTo100M Adverbs which contains only 6. We measure the quality of each dataset’s annotations with a 200 video sample. Since the new datasets come from human written captions, where a person has explicitly chosen the adverb to describe the action, the annotations are much less noisy than HowTo100M Adverbs.

\(^1\)https://github.com/hazeld/PseudoAdverbs
4.2. Adverb Recognition Tasks

With these datasets, as in prior work [12], we want to learn to recognize adverbs in previously seen action-adverb compositions. We additionally propose two new adverb recognition tasks: first in unseen compositions and second in unseen domains. We explain each below.

**Task I: Seen Compositions.** Adverbs and actions are compositional, an adverb \( m \in M \) can apply to many different actions \( a \in A \). Assume we have a set of action-adverb compositions \( (a, m) \in C \). When recognizing adverbs in seen compositions, all compositions in the test set are present in the labeled training set, *i.e.* \( C_{test} \subseteq C_{labeled} \). This tests whether the model can successfully compose and recognize adverbs across various actions. For this evaluation we use our newly proposed VATEX Adverbs as well as HowTo100M Adverbs [12]. We partition VATEX Adverbs into train and test following the original train and test split. This gives 11,782 video clips in training and 2,835 in testing over the 34 adverbs. HowTo100M-Adverbs contains 6 adverbs and consists of 5,475 video clips in training and 349 in testing.

**Task II: Unseen Compositions.** To fully capture the compositional nature of actions and adverbs it is necessary for a model to generalize beyond seen compositions. We thus propose to recognize adverbs in unseen compositions, *i.e.* \( C_{test} \cap C_{labeled} = \emptyset \). We focus on VATEX Adverbs for this since it has the most action-adverb pairs. We partition the pairs into two disjoint sets. For each action, both the pair with an adverb and its antonym are in the same set. Each set contains 50% of the pairs and every action and adverb is present in both sets. We take one split for training and further partition the second split, using half the instances of each pair as the test set and half as the action-only set.

**Task III: Unseen Domains.** Since a key challenge of fine-grained video understanding is the collection of labeled data, we cannot assume to have labels in every domain where we wish to recognize adverbs. We thus propose to test the transferability of learned adverbs to new domains. Here our labeled data \( S \) comes from a domain \( D_S \) while our test set and action-only labeled data \( U \) come from a distinct domain \( D_U \neq D_S \). We use VATEX Adverbs as the source and MSR-VTT Adverbs and ActivityNet Adverbs as targets. We partition both targets in two 50% splits, one for testing and the other as action-only labeled training data.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Adverbs</th>
<th>Actions</th>
<th>Pairs</th>
<th>Accuracy</th>
<th>Videos</th>
<th>Clips</th>
<th>Length (s)</th>
<th>Seen</th>
<th>Unseen</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>HowTo100M Adverbs [12]</td>
<td>6</td>
<td>72</td>
<td>263</td>
<td>44.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VATEX Adverbs</td>
<td>34</td>
<td>135</td>
<td>1,550</td>
<td>93.5%</td>
<td>14,617</td>
<td>10.0</td>
<td></td>
<td>✓</td>
<td></td>
<td>Source</td>
</tr>
<tr>
<td>MSR-VTT Adverbs</td>
<td>18</td>
<td>106</td>
<td>464</td>
<td>91.0%</td>
<td>1,824</td>
<td>15.7</td>
<td></td>
<td>✓</td>
<td></td>
<td>Target</td>
</tr>
<tr>
<td>ActivityNet Adverbs</td>
<td>20</td>
<td>114</td>
<td>643</td>
<td>89.0%</td>
<td>3,099</td>
<td>37.3</td>
<td></td>
<td>✓</td>
<td></td>
<td>Target</td>
</tr>
</tbody>
</table>

Table 1. Our three newly proposed adverb datasets have more adverbs, actions, unique pairs and higher annotation accuracy than HowTo100M Adverbs [12] and also allow us to study recognition of adverbs in unseen action-adverb compositions and unseen domains.

5. Experiments

We first describe the implementation details of our method and the evaluation metric used. We then analyze the contribution of our model’s components and compare to semi-supervised baselines for recognizing adverbs in seen compositions. Finally, we evaluate our approach for recognizing adverbs in unseen compositions and unseen domains.

**Implementation Details.** All videos are sampled at 25fps and scaled to 256px in height. Each video is divided into 1-second segments with one 16-frame snippet extracted per segment. We use a frozen I3D network as the backbone, one for RGB and one for optical flow. The output of the global pooling layer for each modality is concatenated to create a \( T \times 2048D \) feature, where \( T \) is the length of the video clip in seconds. The video embedder \( f \) uses transformer-style attention to locate the relevant video parts with the \( T \) video features as the keys and the action as the query. The text embedder \( g \) uses GloVe embeddings to represent the actions and learns adverbs as linear transformations on action embeddings. See [12] for more details. Optimization is done with Adam [28]. Models are trained with a supervised batch size of 128 and learning rate of \( 10^{-4} \) for 1000 epochs. As in [12] we introduce the adverbs after the 200th epoch, until that moment we train \( g \) as an action embedder. In experiments without thresholding, we reduce noise by letting the adverb representations train for 100 epochs before introducing pseudo-labels. The ratio of adverb-labeled to action-only labeled samples in a batch is the same as the total ratio. Unless otherwise specified, we set the triplet loss margins to \( \gamma_1 = \gamma_2 = 1 \) (Eq. 1, 2), the maximum pseudo-labels per video to \( k = 5 \) (Eq. 5), the base threshold to \( \tau = 0.6 \) (Eq. 8) and the smoothing factor to \( \lambda = 0.1 \) (Eq. 8).
Evaluation Metric. We use adverb-antonym binary classification accuracy from [12]. That is the accuracy when considering the ground-truth adverb vs. its antonym. This suits the available ground-truth labels since they indicate a single adverb, while multiple adverbs may apply to an action. As the distributions of adverbs in our new datasets are long-tailed, we report the average accuracy over adverbs for these, rather than average over videos.

5.1. Ablation Study

We first perform several ablation studies evaluating the effect of each of the proposed model components. For these we recognize adverbs in seen compositions with VATEX Adverbs since this has the greatest variety of adverbs. Experiments are performed with 5% of the training set as the labeled set and the remainder as the action-only labeled set.

Multi-Adverb Pseudo-Labeling. Fig. 6 shows the effect of $k$, the maximum pseudo-labeled adverbs per video. Using multiple pseudo-labels ($k > 1$) offers a large advantage over supervised-only learning and semi-supervised learning with a single pseudo-label ($k = 1$). The best performance is with $k = 7$, although all values in $4 \leq k \leq 10$ are good. With many different adverbs applying to an action ($k \geq 13$) the performance drops, since this many adverbs rarely co-occur, although this is still better than supervised only learning.

Using multi-adverb pseudo-labeling lets us to make more efficient use of the data at our disposal as each video clip is used to learn multiple adverbs. It also encourages exploration of less frequently labeled adverbs, which we show in Fig. 5. With a single adverb pseudo-label (yellow), the overall distribution of pseudo-labels is highly imbalanced and mimics the long-tail distribution of the ground-truth. There are even 5 adverbs which do not have any pseudo-labels. Our multi-adverb pseudo-labeling (purple) reduces this bias, with pseudo-labels better distributed across possible adverbs.

Pseudo-Label Selection. A standard approach to pseudo-labeling in an embedding space would be to take the closest embeddings as the pseudo-label(s). We instead take the adverbs with the greatest difference between the embedded video’s proximity to the adverb modified action and the antonym modified action. We compare these approaches in Table 2a, which shows our approach improves the result.

Adaptive Thresholding. Table 2b compares the adaptive thresholding we use to no thresholding and a fixed threshold for all adverbs. The adaptive thresholding improves the result by 2.5% over fixed thresholding, which has little impact itself. With fixed thresholding, once the most common adverbs have exceeded this threshold the model will pseudo-label all actions with these adverbs, ignoring rarer adverbs. The adaptive thresholding allows the pseudo-label selection to be more balanced (as shown in Fig. 5).

Ratio of Action-Only Data. We test the effect of the ratio of adverb-labeled to action-only labeled videos in Fig. 7. This shows training with any amount of action-only data gives better performance. We observe two peaks in Fig. 7. With an action-only ratio $\geq 15$ the model is able to see all available action-only data in each epoch of the labeled data. This allows better learning of the rarer adverb-action compositions. With 2 times the amount of action-only data the model is less likely to overfit to any noisy pseudo-labels early in training.
5.2. Task I: Seen Compositions

We test our method’s suitability for recognizing adverbs in previously seen action-adverb compositions as in prior work [12]. Thus we use HowTo100M Adverbs [12] as well as VATEX Adverbs. For both datasets we test our approach with different amounts of labeled data used in training: 1%, 2%, 5%, 10% and 20%. The remaining training data is used in the action-only labeled set.

We compare to the supervised-only adverb recognition approach Action Modifiers [12] on which our approach is based. This learns from only the adverb-labeled data, not the action-only labeled set. We also compare to several semi-supervised approaches: Pseudo-Label [31], FixMatch [62] and TCL [61], which we adapt for adverb recognition by combining them with Action Modifiers. This allows a fair comparison as the backbone and adverb representations are the same in all methods. Pseudo-Label simply takes the most confident prediction of a data sample to be the pseudo label. FixMatch obtains pseudo-labels from weak augmentations of the input data. Strongly augmented versions are then trained to predict the same pseudo-label. This also uses fixed thresholding. Instead of the image augmentations used in FixMatch, TCL uses the video speed. It also optimizes agreement between the predictions for all classes rather than a single pseudo-label. Full implementation details can be found in supplementary.

Results are presented in Table 3. For VATEX Adverbs our approach outperforms or obtains competitive results over all baselines for all percentages of labeled data used. On HowTo100M Adverbs our approach outperforms baselines for the 5%, 10% and 20% labeled data settings. Our multi-adverb pseudo-labeling has more impact on VATEX Adverbs since it contains more adverbs. The improvement is also greater when using ≥5% labeled data. With fewer labels each adverb is seen in fewer situations, meaning the pseudo-labels become more noisy. However, there is still room for further improvement over our method as using 100% adverb-labeled data obtains 73.9% on VATEX Adverbs and 80.8% on HowTo100M Adverbs. We observe that TCL often performs worse than other approaches, despite being designed for video. This is because TCL encourages invariance to speed which affects adverbs such as quickly and slowly. Each of the semi-supervised baselines are comparable overall to the supervised-only method, this highlights the importance of our proposed multi-label pseudo-labelling and adaptive thresholding. Without these elements models are more biased to the particular action-adverb compositions.

We show examples of our multi-adverb pseudo-labeling in Fig. 8. Our method provides multiple relevant adverb pseudo-labels for each video. The model is able to use the adaptive thresholding to exclude incorrect predictions (climb downwards) or frequent compositions (climb properly and sneeze loudly). There are still noisy pseudo-labels such as climb indoor and dip evenly. There are also cases where the adverbs make no sense in the context of the action, e.g. dip backwards. The incorrect prediction turn downwards highlights a challenge of adverb datasets, where there can be multiple actions occurring at the same time. Here downwards refers to the people falling, rather than boat turning.

Table 3. Seen Compositions. When using ≥5% of the labeled training data our method outperforms semi-supervised baselines for recognition of adverbs in previously seen action-adverb compositions.

<table>
<thead>
<tr>
<th>Method</th>
<th>VATEX Adverbs</th>
<th>HowTo100M Adverbs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
<td>2%</td>
</tr>
<tr>
<td>Supervised only</td>
<td>54.0</td>
<td>54.5</td>
</tr>
<tr>
<td>Pseudo-Label</td>
<td>55.1</td>
<td>54.4</td>
</tr>
<tr>
<td>FixMatch</td>
<td>55.4</td>
<td>52.3</td>
</tr>
<tr>
<td>TCL</td>
<td>51.6</td>
<td>56.6</td>
</tr>
<tr>
<td>Ours</td>
<td>55.0</td>
<td>56.6</td>
</tr>
</tbody>
</table>

Figure 8. Example pseudo-labels from our proposed multi-adverb pseudo-labeling for the indicated action. Pseudo-labels below their adverb threshold are shown in grey. Our method can successfully identify multiple relevant adverbs for each video (left column), can use the adaptive thresholding to ignore incorrect or unnecessary pseudo-labels (middle), but can struggle to be accurate when actions co-occur (downwards, top right) and has no notion of situations where an adverb is infeasible (backwards, bottom right).
5.3. Task II: Unseen Compositions

We investigate whether our method can improve recognition of adverbs in previously unseen action-adverb pairs. We compare to supervised only Action Modifiers [12]. Table 4 shows that our method improves performance by $\sim$4%. The adaptive thresholding is key. Without it the pseudo-labels primarily consist of previously seen adverbs compositions. However, there is much potential for future work as highlighted by the gap between our method and training with all compositions seen. Generalizing to unseen action-adverb composition is necessary since it is infeasible to acquire sufficient labeled data for every possible composition.

5.4. Task III: Unseen Domains

In Table 5 we test whether our pseudo-labeling approach can improve transfer to new domains. We compare our approach to training with only the source data, i.e. VATEX Adverbs, as well as the Pseudo-Label [31] baseline. Our method outperforms the Pseudo-Label approach for MSR-VTT Adverbs and gives a $\sim$2% gain over using only source domain videos. On ActivityNet Adverbs all three approaches are comparable, as the gap to this dataset is larger both in terms of action and adverb appearance and action length. Table 5 also shows the upper bounds when target data is used in training. The gap between our model’s performance and source+target is relatively small, meaning adverb representations do transfer well between actions in different domains, however there is still much potential for improvement in the adverb representation itself. This is a more realistic setting to evaluate adverb representations, since labeled data is scarce. Transferring adverb representation to new domains is key to applications such as recognizing anomalous occurrences of an action or whether someone is following a recipe well.

5.5. Describing Relationships Between Actions

We foresee many applications of adverbs in video understanding, such as in dense video captioning [29], describing and detecting anomalies [22], determining a person’s skill [11] and procedure planning [7]. Here we demonstrate qualitatively how adverbs can be used to identify fine-grained zero-shot actions by describing the relationship between these unseen actions and those previously seen. Fig. 9 shows examples of such actions. In each case the zero-shot action can be described by applying an adverb to a known action.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSR-VTT Adverbs</th>
<th>ActivityNet Adverbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>62.9</td>
<td>67.2</td>
</tr>
<tr>
<td>Pseudo-Label</td>
<td>63.9</td>
<td>66.4</td>
</tr>
<tr>
<td>Ours</td>
<td>65.0</td>
<td>66.6</td>
</tr>
<tr>
<td>Source + Target</td>
<td>67.5</td>
<td>71.6</td>
</tr>
<tr>
<td>Target only</td>
<td>70.5</td>
<td>71.8</td>
</tr>
</tbody>
</table>

Table 5. Transfer to unseen domains from VATEX-Adverbs. Our method aids generalization to similar domains (MSR-VTT Adverbs), but struggles with larger shifts (ActivityNet Adverbs).

Figure 9. We use the learned video-text embedding to identify zero-shot actions by compositions of adverbs and seen actions. We show each zero-shot action in bold alongside the closest action-adverb pair in the embedding space and one of the closest videos.

6. Discussion

Limitations. Our method has several limitations. Firstly, our model has no concept of infeasible combinations of action and adverbs and can be confounded by co-occurring actions where different adverbs apply. It also struggles when an adverb is labeled in few contexts. While our method can aid generalization to unseen action-adverb compositions and unseen domains, there is still far to go in these areas.

Potential Negative Impact. All datasets in this paper are sourced from YouTube and therefore the subjects and activities contained within are not representative of the diversity in global society. Thus our trained models will contain biases.

Conclusions. This paper presents a semi-supervised method to recognize adverbs of actions. This allows us to understand how an action is being performed and understand fine-grained differences between actions. We propose multi-adverb pseudo-labeling to make use of videos with action-only labels. To cope with the long-tail distribution of adverbs and their action compositions our method also makes use of adaptive thresholding. We propose three new adverb datasets which allow us to evaluate how well our method recognizes adverbs in seen action-adverb compositions as well as unseen compositions and unseen domains. Results demonstrate our method improves performance in all three tasks.

Acknowledgements. Work is part of the Real-Time Video Surveillance Search project (18038), which is partly financed by the Dutch Research Council (NWO) domain Applied and Engineering/Sciences (TTW).
References


Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho.


Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captoring events in videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017. 1, 2, 4, 8


Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In European Conference on Computer Vision (ECCV), 2020. 2, 4


Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In European Conference on Computer Vision (ECCV), 2018. 2


Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In British Machine Vision Conference (BMVC), 2019. 2


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/CVF International Conference on Computer Vision (ICCV), 2019. 4


[59] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2


[64] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019. 1, 2, 4


[70] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In Proceedings of the AAAI Conference on Artificial Intelligence, 2015. 2


[77] Chen Zhu, Xiao Tan, Feng Zhou, Xiao Liu, Kaiyu Yue, Errui Ding, and Yi Ma. Fine-grained video categorization with redundancy reduction attention. In European Conference on Computer Vision (ECCV), 2018. 2

