Multi-modal Content Localization in Videos Using Weak Supervision

Gourab Kundu [1] **Prahal Arora** [1] **Ferdi Adeputra** [1] **Polina Kuznetsova** [1] **Daniel McKinnon** [1] **Michelle Cheung** [1] **Larry Anazia** [1] **Geoffrey Zweig** [1]

Abstract

Identifying the temporal segments in a video that contain content relevant to a category or task is a difficult but interesting problem. This has applications in fine-grained video indexing and retrieval. Part of the difficulty in this problem comes from the lack of supervision since large-scale annotation of localized segments containing the content of interest is very expensive. In this paper, we propose to use the category assigned to an entire video as weak supervision to our model. Using such weak supervision, our model learns to do joint video level categorization and localization of content relevant to the category of the video. This can be thought of as providing both a classification label and an explanation in the form of the relevant regions of the video. Extensive experiments on a large scale data set show our model can achieve good localization performance without any direct supervision and can combine signals from multiple modalities like speech and vision.

1. Introduction

The task of video categorization is to assign a category for a video at the level of the entire video. Videos can be long and even if a long video contains very short segments that exhibit content indicative of that category, that category will be assigned to that video. This lends directly into the motivation for the task of video localization where the aim is to identify those segments in a video with content relevant to the category of that video. Video categorization has applications in video indexing and retrieval. Video localization can further aid these applications by providing more precise and fine-grained indexing and retrieval.

Both video categorization and localization have been extensively studied in the computer vision literature (Brezeale & Cook, 2008; Weinland et al., 2011). However, localization methods typically require annotation that specify the temporal segments containing relevant content for the category of interest. Obtaining such fine-grained annotation is expensive and difficult to collect. However, annotation for the category of a video is relatively cheap and is readily available in some domains. Therefore, learning models for video localization task using the categories for the video categorization task as weak supervision signals is a practical solution.

In this paper, we propose a model that can be trained using the categories at the video level but can be used for predicting both categories at the video level and for predicting the temporal segments inside the video that contain relevant content for that category. We use a bidirectional recurrent neural network for aggregating information from each time step and then an attention network that provides a weight for each time step. The weighted average across all time steps is then sent to a feed-forward network followed by a softmax layer to output the category of the video. This network is trained using the categories of the videos, but at test time, the attention weights can indicate the hot spots inside the video. Thus, our model can be used for both classification and localization.

Another advantage of our model is that it is a multi-modal model that combines the speech transcripts and visual features in a simple but effective way. In most real-word applications, solving both the video categorization and localization tasks require a multi-modal understanding. As we show experimentally, both modalities bring significant gains for our model, thus validating our hypothesis that multi-modality is crucial.

We conduct experiments on a large scale data set with roughly 200K videos, across a wide range of topics. Our extensive experiments show that our model learns to localize using weak supervision at video level and different modalities bring cumulative gains for our model.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 contains the in-depth description of our model. A brief description of our data set is presented in Section 4. We present our experimental

[1] Facebook Inc.. Correspondence to: Gourab Kundu <gkfacebookny@fb.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

results in Section 5. Finally, section 6 ends the paper with some concluding remarks.

2. Related Work

Temporal action localization has been an active research topic recently. However, most of the works have focused on fully supervised setting, where there is a sufficiently large set of videos annotated with localization for the actions of interest(Caba Heilbron et al., 2016; Escorcia et al., 2016; Richard & Gall, 2016; Yeung et al., 2016; Yuan et al., 2016; Shou et al., 2016; 2017; Lin et al., 2017; Heilbron et al., 2017; Zhao et al., 2017; Gao et al., 2017; Xu et al., 2017). Unfortunately, in most practical applications, such a large set of annotations for videos with temporal segmentation is not readily available. It is also very expensive to collect these annotations for a sufficiently large number of videos. Therefore, there is a need to resort to weakly supervised approaches for training localization models.

Several works in the literature have proposed using the category at the video level as a weak supervision for training models for localization(Singh & Lee, 2017; Wang et al., 2017; Shou et al., 2018). However, these approaches rely on visual features only. As shown in our experiments, for many applications, visual features alone may not be sufficient since the action or category of interest may depend on the speech in that video. Different from these localization models, our model is multi-modal one, that can use both speech and visual features in videos for joint action localization and classification.

Our task is a restricted version of the task of the spatio-temporal localization (Bhoi, 2019; Tian et al., 2013; Ma et al., 2013). In multi-modal setting, spatial localization for categories that depend on speech may not be informative. In such multi-modal cases, temporal localization is more appropriate.

3. Model Architecture

Assume a video has length of M seconds and the transcript in the video contains N words $w_1, w_2, \ldots, w_N$. Let the word embeddings of these N words be given by $u_1, u_2, \ldots, u_N$. We extract dense visual features for every integer time stamp of the video. Let these visual feature vectors be given by $v_1, v_2, \ldots, v_M$. Let t_i be the time stamp of the utterance of the word w_i rounded to the nearest integer. The multi-modal embedding in our work is a concatenation of a word embedding and the visual embedding at the time stamp of the word. For word w_i, it is given by $[u_i, v_{t_i}]$. If no word is spoken at an integer time stamp in the video, we assume the token "NO_ASR" is spoken at that time stamp. Thus each time stamp has associated with it both a visual feature vector and a real or notional word.

Using the alignment outlined above, we convert the sequence of spoken word vectors and visual vectors into a sequence of multi-modal vectors. Let the sequence of vectors be $x_1, x_2, \ldots, x_L$.

The architecture of our model is shown in Figure 1. We use a bidirectional LSTM (Hochreiter & Schmidhuber, 1997) as the recurrent network in our model. We feed the multi-modal vectors $x_1, x_2, \ldots, x_L$ to the LSTM network and output a sequence of hidden vectors $h_1, h_2, \ldots, h_L$.

$$\overrightarrow{h_i} = LSTM_{forward}(x_i, h_{i-1}) \tag{1}$$

$$\overleftarrow{h_i} = LSTM_{backward}(x_i, h_{i+1}) \tag{2}$$

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \tag{3}$$

The LSTM output vectors $h_1, h_2, \ldots, h_L$ are then sent to an attention layer (Denil et al., 2012; Bahdanau et al., 2014) to produce a normalized weight for every output $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_L$.

$$a_i = w_{att} \cdot h_i \tag{4}$$

$$\hat{a}_i = \frac{a_i}{\sqrt{\sum_{i=1}^{L} a_i^2}} \tag{5}$$

Here, w_{att} is the attention layer vector that is used for producing the attention scores. Then a weighted sum of the LSTM output vectors is sent to a feed-forward network followed by a softmax layer. The feed-forward network f consists of two layers with ReLU activation function.

$$g_i = \sum_{i=1}^{L} \hat{a}_i h_i \tag{6}$$

$$\hat{y}_i = \text{softmax}(f(g_i)) \tag{7}$$

Training
During training, we minimize the cross entropy loss measured over a the training data D. Assume there are C categories in the data set. The loss is given by:

$$l = -\frac{1}{|D|} \sum_{i=1}^{|D|} \sum_{c=1}^{C} y_{ic} \log \hat{y}_{ic} \tag{8}$$

Training proceeds via gradient descent at the minibatch level.

Prediction of Video Category
During test time, the category of a video is predicted by:

$$l^* = \text{argmax}_l\, \hat{y}_l \tag{9}$$

Prediction of Temporal Segment
For a fixed duration D, we find non-overlapping temporal segments of duration D that have the highest sum of attention scores in an iterative manner. Let H be the list of hidden vectors corresponding to multi-modal vectors that

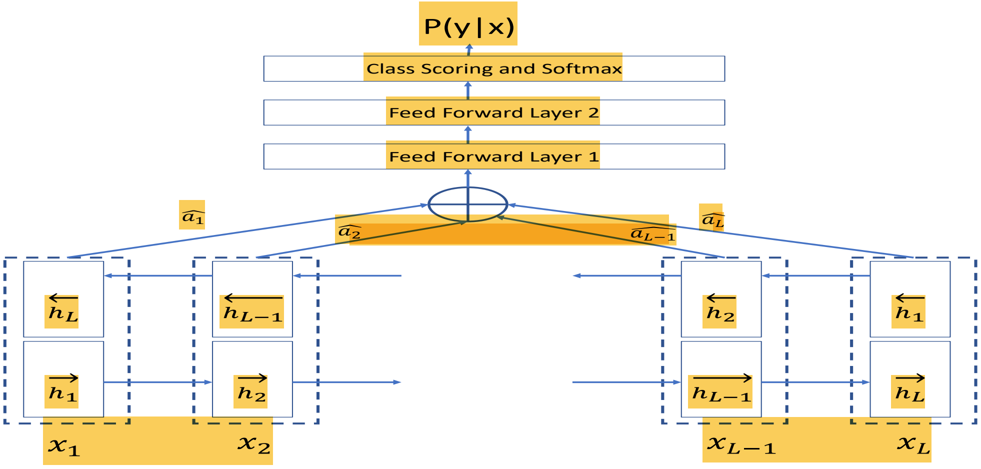


Figure 1. Architecture of our Model

are derived from words spoken in the time interval (b, e) and let $\hat{A}$ be the corresponding list of normalized attention scores. The score for the interval (b, e) is given by:

$$s(b, e) = \sum_{\hat{a} \in \hat{A}} \hat{a} \tag{10}$$

We calculate the top 3 non-overlapping temporal segments of duration D with the highest scores in a greedy manner. At every step, we find the temporal segment with the highest score and remove all time stamps that belong to that segment and find the next temporal segment with the highest score and continue. By design, the selected temporal segments are non-overlapping.

4. Dataset

Our data set consists of a set of videos grouped into five broad categories: “Pornography”, “Graphic Disturbing”, “Hate-speech”, “Bullying”, and benign. We list the number of examples and average length of videos per category in Table 1. The data set is highly imbalanced, with categories like “Pornography” and “Graphic Disturbing” are relatively dominant in our data set. The remaining categories are much less frequent. However, videos from the category of “Hate-speech” are the longest. Table 1 shows that our data set is a large scale data set with class imbalance in terms of both frequencies and length of videos.

Although there exist several data sets for studying video categorization and localization(Soomro et al., 2012; Jiang et al., 2014; Kay et al., 2017; Gu et al., 2018; Karpathy et al., 2014), most of them focus on identifying and localizing human actions that are atomic and simplistic. For example, the dominant categories in the widely used data set UCF-101(Soomro et al., 2012) include “Shaving Beard”, “Playing Tennis” and “Playing Dhol”. These actions tend to be atomic in nature and do not require understanding of other modalities except vision. The data set we consider in our work consists of categories of complex human actions that are compositions of many atomic actions and that require both speech and visual understanding. One objective of our research is to explore interesting research regarding combination of signals from multiple modalities like speech and vision.

Table 1. Data set Statistics

Label	Number of Examples	Average Length in Sec
Pornography	37059	98
Graphic Disturbing	17792	82
Hate-speech	8217	1270
Bullying	9490	28
Benign	135314	344
Total	208782	305

We use a state-of-the-art automatic speech recognition system (ASR) to extract the speech transcript along with the time stamps where the words are spoken. We also use a state-of-the-art video categorization system to extract the penultimate embedding (before the softmax layer) once per second for each video. Details of the ASR and video categorization models are provided below.

4.1. ASR system

Each video in our data set is processed through an automatic speech recognition (ASR) system to produce a speech transcript. The speech recognition system was built using a hybrid BLSTM-HMM framework. The lexicon is a graphemic lexicon with character sequences as pronunciations. The acoustic model is an LC-BLSTM (latency-controlled bi-directional LSTM (Zhang et al., 2016; Xue & Yan, 2017), trained with cross-entropy and lattice-free MMI criterion (Povey et al., 2016). It takes 80-dimensional log mel-filterbank coefficients as the input, and posterior over around 9000 tied context-dependent graphemes (clustered by a decision tree) as the output. The language model is a 5-gram model.

4.2. Video Categorization System

The model used by the video categoriztion system is ResNeXt3D which is an efficient clip-based model. It takes a short clip of frames as input to the model. Compared to the frame based model, or image model, a clip-based model can capture motion information by 3D convolution, and performs better on video datasets. For details, readers can refer to the paper (Tran et al., 2018). ResNeXt3D achieves state-of-the-art result on benchmark video datasets.

5. Experimental Results

Hyperparameters
For word vectors, we use 300 dimensional fast-text embeddings (Bojanowski et al., 2017). These embeddings are tuned as part of the training process. The dimension of video embedding obtained from ResNext3D model is 2048. We project the video embedding to a lower dimension of 500 using a linear layer learned as part of the training process. The subsequent LSTM network consists of a single layer, with input dimension of 800 and dimension of 1024 for each direction for the output hidden vectors. The hidden vectors from the backward & forward LSTMs are concatenated, resulting in 2048 dimensionality for each hidden vector. The attention layer is a 2048 dimensional vector that produces a single score for each of these hidden vectors. The subsequent feed-forward MLP consists of two feed-forward layers each with 2048 neurons with ReLU activation function. We use a drop-out of 0.2 on these two layers and a drop-out of 0.15 for LSTM weights. We use the ADAM optimizer (Kingma & Ba, 2014) with a scheduler for reducing learning rate by 0.25 every time the validation loss plateaus. Our initial learning rate is 0.00008 and we train for 32 epochs. We have tuned all the hyper-parameters on the development set. We use the distributed training framework of pytorch using 4 nodes and 8 gpus per node.

We randomly split our data set into training, validation and test (70% training, 15% validation and 15% test). Since our model is capable of both video level classification and localization, we evaluate our model on both respects.

Video Categorization Results
Figure 2 shows the results for video categorization (F scores) for every category except the "benign" category in our data set.

For all categories except "Bullying", the multi-modal model outperforms the "ASR only" and "Visual Only" models. For the categories of "Pornography" and "Graphic Disturbing", the improvement is small but statistically significant. This makes sense because intuitively, we feel that identification of these two categories should be benefited by visual features most. For the category of "Hate-speech", multi-modal model gives significant improvements over using a model that uses only one modality (31% over "ASR only" and 11% over "Visual only"). This also seems intuitive since "Hate-speech" can be demonstrated either by speech or by imagery. Often, "Hate-speech" is also demonstrated by the combination of benign speeches and benign images where the combination of those specific images and speech make them Hate-speech. One exception to this trend is the category of videos marked as "Bullying". For this category of videos, the visual model performs the best. This is due to the fact that the videos for this category are relatively short and almost 80% videos do not contain significant amount of ASR transcripts.

If we compare the models of "Visual only" and "ASR only", we see that the "Visual only" model outperforms the "ASR only" model in all categories. Even for the category of "Hate-speech", "Visual only" model outperforms the "ASR only" model. In addition to having relatively high speech recognition error rates on out-of-domain data, many videos contain speech in foreign languages and our model was trained on videos containing English ASR only, thus resulting in some data without transcriptions. We leave the exploration using foreign language ASR and multi-lingual embeddings for text as a future work.

Video Localization Results
For evaluating localization results, we produced the top 3 temporal segments of duration $5s$ each for 271 videos. The segments were then evaluated to be either correct (contains evidence for the video level category) or incorrect (not relevant for categorization of the video). 83% of the segments were correct. Even better, for a given video, at least one of the three segments flagged the offensive content 93% of the time.

In addition, we also performed a qualitative evaluation of the topmost temporal segments predicted by our model. Figure 3 shows examples of temporal segments predicted by our model for each of the category except "benign" and "Pornography". We identify the three temporal segments with duration of $5s$ that have the highest sum of attention scores and highlight these segments on the time line of the video with red, orange and yellow colors. Since each segment is $5s$ long, we randomly select a time stamp inside that segment and show the image or thumbnail for that time

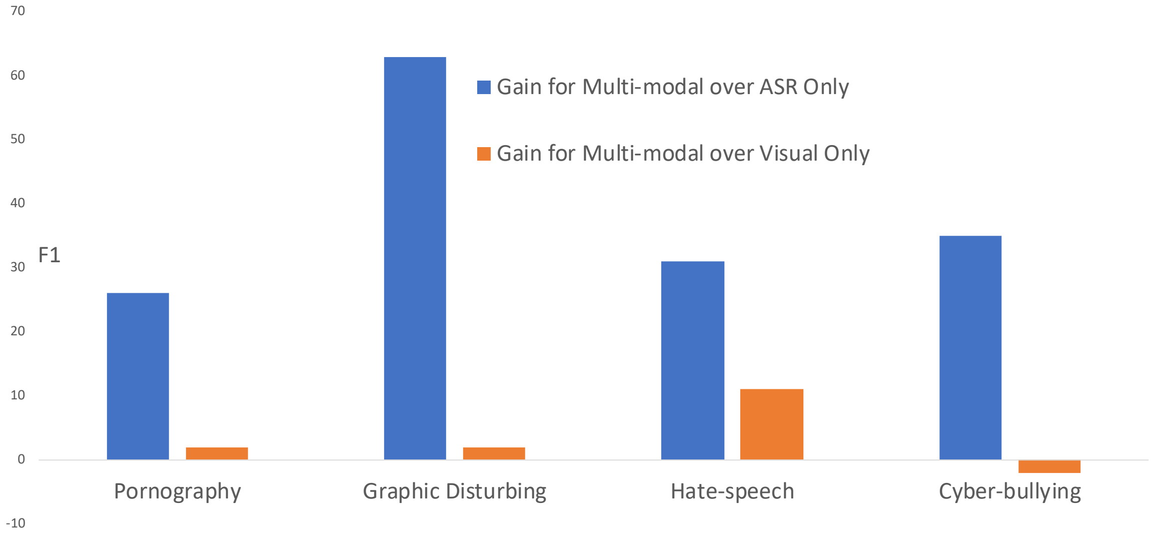


Figure 2. Relative Gain in video categorization results (F scores) for the multi-modal model over "ASR only" and "Visual Only" models, broken down across different categories. The multi-modal model outperforms single-modal model on 3 out of 4 categories.

stamp. Additionally, we also show the transcripts in a window around that time stamp with profanities replaced by '*'. To protect confidentiality, we have replaced the actual images with commercially available public images that exhibit similar semantics. Similarly, the actual transcripts have been replaced with paraphrased generic text with similar semantics, and without any personally identifiable information.

We notice that the temporal segments predicted by our model make intuitive sense. For example, temporal segments predicted by our model for videos belonging to the category of "Hate-speech" typically contain images of angry or disturbed persons. Those segments also frequently have offensive slurs inside the speeches(Figure 3-a). In Figure 3-b, we show an image and associated transcript for the temporal segment predicted by our model for one of the videos belonging to the category of "Graphic Disturbing". The image depicts surgery and the speech describes the conversation among the surgeons. Finally, Figure 3-(c & d) contains two example images and transcripts for the category of "Bullying". The images corresponding to the temporal segments predicted by the model for this category tend to show anger and fighting whereas the speech contain slurs directed to a specific person. In all cases, we can see that our model has produced very accurate temporal segments, without any direct supervision at the level of segments.

6. Conclusion

In this paper, we proposed a model architecture that can be trained with the category of a video as a weak supervision and the model can then be used for both categorization of the video and localization of the content that can explain the category. We evaluated our model on a large-scale data set and provided both quantitative and qualitative assessments of our model. In future, we want to extend our multi-modal framework to the audio modality by using the embeddings produced by an audio event detection model as additional input to our multi-modal model.

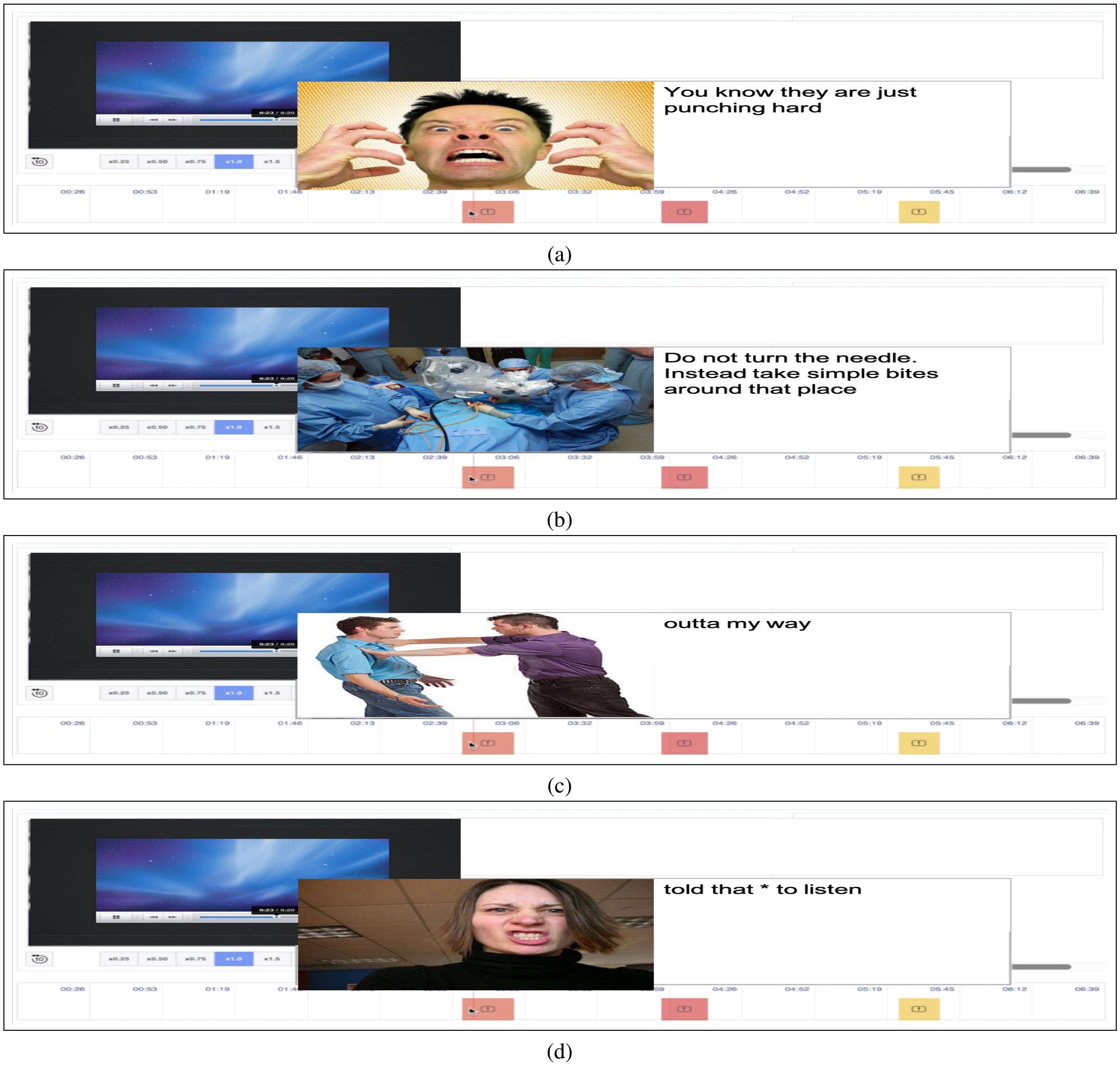
References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bhoi, A. Spatio-temporal action recognition: A survey, 2019.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

Brezeale, D. and Cook, D. J. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):416–430, 2008.


You know they are just punching hard

(a)

Do not turn the needle. Instead take simple bites around that place

(b)

outta my way

(c)

told that * to listen

(d)

Figure 3. Localization Results. The examples correspond to the categories (a) - Hate-speech, (b) -Graphic Disturbing, (c) - Bullying, and (d) - Bullying. Note that the images have been replaced with commercially available public images with similar semantics, and the text is a generic paraphrase of actual transcript.

Caba Heilbron, F., Carlos Niebles, J., and Ghanem, B. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1914–1923, 2016.

Denil, M., Bazzani, L., Larochelle, H., and de Freitas, N. Learning where to attend with deep architectures for image tracking. *Neural computation*, 24(8):2151–2184, 2012.

Escorcia, V., Heilbron, F. C., Niebles, J. C., and Ghanem, B. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pp. 768–784. Springer, 2016.

Gao, J., Yang, Z., Chen, K., Sun, C., and Nevatia, R. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3628–3636, 2017.

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018.

Heilbron, F. C., Barrios, W., Escorcia, V., and Ghanem, B. Scc: Semantic context cascade for efficient action detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3175–3184. IEEE, 2017.

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

Jiang, Y., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., and Sukthankar, R. Thumos challenge: action recognition with a large number of classes (2014), 2014.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. The kinetics human action video dataset, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014.

Lin, T., Zhao, X., and Shou, Z. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 988–996. ACM, 2017.

Ma, S., Zhang, J., Ikizler-Cinbis, N., and Sclaroff, S. Action recognition and localization by hierarchical space-time segments. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pp. 2751–2755, 2016.

Richard, A. and Gall, J. Temporal action detection using a statistical language model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3131–3140, 2016.

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

Shou, Z., Chan, J., Zareian, A., Miyazawa, K., and Chang, S.-F. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5734–5743, 2017.

Shou, Z., Gao, H., Zhang, L., Miyazawa, K., and Chang, S.-F. Autoloc: Weakly-supervised temporal action localization. *arXiv preprint arXiv:1807.08333*, 2018.

Singh, K. K. and Lee, Y. J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3544–3553. IEEE, 2017.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

Tian, Y., Sukthankar, R., and Shah, M. Spatiotemporal deformable part models for action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2642–2649, 2013.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Wang, L., Xiong, Y., Lin, D., and Van Gool, L. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4325–4334, 2017.

Weinland, D., Ronfard, R., and Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011.

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

Xue, S. and Yan, Z. Improving latency-controlled blstm acoustic models for online speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5340–5344. IEEE, 2017.

Yeung, S., Russakovsky, O., Mori, G., and Fei-Fei, L. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2678–2687, 2016.

Yuan, J., Ni, B., Yang, X., and Kassim, A. A. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3093–3102, 2016.

Zhang, Y., Chen, G., Yu, D., Yaco, K., Khudanpur, S., and Glass, J. Highway long short-term memory rnns for distant speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5755–5759. IEEE, 2016.

Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., and Lin, D. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2914–2923, 2017.