On Evaluating Weakly Supervised Action Segmentation Methods

Yaser Souri[†], Alexander Richard[†], Luca Minciullo[§], Juergen Gall[†]
[†] University of Bonn, [§] Toyota Motor Europe

{souri, richard, gall}@iai.uni-bonn.de, luca.minciullo@toyota-europe.com

Abstract

Action segmentation is the task of temporally segmenting every frame of an untrimmed video. Weakly supervised approaches to action segmentation, especially from transcripts have been of considerable interest to the computer vision community. In this work, we focus on two aspects of the use and evaluation of weakly supervised action segmentation approaches that are often overlooked: the performance variance over multiple training runs and the impact of selecting feature extractors for this task. To tackle the first problem, we train each method on the Breakfast dataset 5 times and provide average and standard deviation of the results. Our experiments show that the standard deviation over these repetitions is between 1 and 2.51%, and significantly affects the comparison between different approaches. Furthermore, our investigation on feature extraction shows that, for the studied weakly-supervised action segmentation methods, higher-level I3D features perform worse than classical IDT features.

1. Introduction

Action segmentation is the task of temporally segmenting every frame of an untrimmed video. It is usually studied in the context of instructional activities performed by a single person [6, 14, 16] and fully supervised approaches for action segmentation achieve already very good results [7, 10, 9, 1]. As providing frame-level supervision for action segmentation is an expensive task, weakly supervised approaches have received a lot of attention from the community [8, 12, 5, 13, 4, 3, 11]. Due to their cheap and fast acquisition, *video transcripts* are a popular form of weak supervision [8, 12, 13, 3, 11]. In this case, only the order of occurring actions is given for training instead of frame-level labels.

In this technical report, we take a closer look at the evaluation protocol for weakly supervised action segmentation using transcripts. In particular, we consider the recent approaches NNV [13], ISBA [4], and CDFL [11] for which

source code is publicly available[1].

Furthermore, we study the features used in weakly supervised action segmentation approaches. For the Breakfast dataset [6], most approaches use IDT features [15] whereas I3D features [2] are extracted for the more recent CrossTask dataset [16].

From the experiments, we make the following observations:

- The variance in the performance of weakly supervised approaches is high when running the training and testing multiple times with different random seeds. It is therefore necessary to report the average and standard deviation over multiple runs to evaluate the performance of an approach.

- Unexpectedly, higher-level I3D features do not perform better than low-level IDT features.

2. Reproducibility Experiments

For the experiments, we used the official source code for NNV [13], ISBA [4], and CDFL [11]. We run the training and testing on the Breakfast dataset without any modification to the respective source codes. As the parameters of the models published by NNV, ISBA, and CDFL were only provided for the Breakfast [6] dataset, we only perform experiments on this dataset. We perform 5 different iterations for each of the 4 splits of the Breakfast dataset and report the results in Table 1. We use the standard mean accuracy over frames (MoF) as the accuracy measure.

We observe that the average over 5 runs is a little bit lower than the reported results for all three methods, but the reported results are within the standard deviation. We also observe that the standard deviation is relatively high. For NNV [13] and CDFL [11], it is about 2.5. This shows that the methods for weakly supervised learning are more sensitive to the random model initialization and the random sampling of mini-batches compared to fully supervised ap-

---

[1]https://github.com/Zephyr-D/TCFPN-ISBA
https://github.com/alexanderrichard/NeuralNetwork-Viterbi
https://github.com/JunLi-Galios/CDFL

Model	MoF Reported	MoF Avg ($\pm$ Std)	MoF Max	MoF Min
ISBA [4]	38.4	36.4 ($\pm$ 1.0)	37.6	35.1
NNV [13]	43.0	39.7 ($\pm$ 2.4)	43.5	37.5
CDFL [11]	50.2	48.1 ($\pm$ 2.5)	50.9	44.6

Table 1. Weakly supervised action segmentation performance on the Breakfast dataset. The mean over frame (MoF) is averaged over all dataset splits and 5 different runs. 'MoF Reported' denotes the reported MoF value from the corresponding paper.

Approach	Features	Average MoF
NNV	IDT	40.6
NNV	I3D	11.4
NNV	PCA-I3D	23.2
CDFL	IDT	48.9
CDFL	I3D	34.9
CDFL	PCA-I3D	38.0

Table 2. Performance of different feature types. Average MoF is computed by running the training and testing of each method 5 times on split 1 of the Breakfast dataset.

Approach	Feature Window Size	Average MoF
NNV	21	23.2
NNV	11	24.6
NNV	5	26.3
NNV	1	33.9

Table 3. Performance of NNV using PCA-I3D features with different feature window size. A smaller feature window size results in better accuracy. Average MoF is computed by running the training and testing 5 times on split 1 of the Breakfast dataset.

proaches. It is therefore important to report the accuracy and standard deviation over multiple runs.

3. Effect of Feature Type

The Breakfast dataset [6] is a standard benchmark for weakly supervised action segmentation and contains breakfast related preparation videos that were recorded in different kitchens. The more recent CrossTask dataset [16] is collected from YouTube videos and contains videos from different domains. While IDT features [15] are commonly used for the Breakfast dataset, [16] suggests to use I3D features [2]. In general, it is expected that I3D features perform better since they are pre-trained on YouTube videos, but the Breakfast dataset differs from YouTube videos and contains videos with much more challenging light conditions. We, therefore, evaluated whether weakly supervised approaches benefit from I3D features on the Breakfast dataset.

We extracted frame-wise I3D (RGB and Flow) features from the videos in the Breakfast dataset, obtaining a feature vector of size 2048 for each frame. We then re-run the experiments but only on split 1 of the Breakfast dataset. We use only the two best-performing methods NNV and CDFL. The results are reported in Table 2.

We observe that without modification of the hyperparameters NNV and CDFL cannot adapt to the new feature type and do not achieve a good performance. A major difference between the IDT features as they are used by NNV and CDFL compared to I3D is the feature dimensionality. We, therefore, reduced the dimensionality by PCA from 2048 to 64. We denote these features as PCA-I3D. Note that NNV and CDFL also reduce the dimensionality of IDT features to 64. The results in Table 2 show that the dimensionality of the features has indeed an impact on the

performance for both methods, but PCA-I3D is still worse than IDT.

Another difference between IDT and I3D is the number of neighboring frames that are used to compute a feature vector for a frame. We, therefore, modified the temporal window of the GRU input for NNV. The results of our experiments with different feature window sizes are reported in Table 3. They show that a smaller feature window size performs better when using PCA-I3D features. This indicates that current methods that perform well for weakly supervised action segmentation on Breakfast using IDT features cannot be directly applied to other features like I3D. Although the experiments do not fully reveal why this is the case, there are three possible explanations: a) I3D features do not work well on Breakfast in general, b) the existing methods need to be adapted to I3D, c) there is a need for different methods that benefit from I3D features.

4. Conclusion

In this report, we have taken a closer look at two crucial aspects of recent weakly-supervised action segmentation approaches performance stability over multiple training runs and the impact of feature extraction. By performing experiments using the officially published source code of the published works, we find that the variance of the accuracy is relatively high. It is therefore necessary to report average and standard deviation over multiple runs to assess the performance of an approach for weakly supervised action segmentation. Furthermore, we observed that current methods do not benefit from 'better' features. However, it remains an open research question why this is the case.

References

[1] Yazan Abu Farha and Juergen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. *CVPR*, 2019.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[3] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D^3TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. *CVPR*, 2019.

[4] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *CVPR*, 2018.

[5] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *ECCV*, 2016.

[6] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014.

[7] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *WACV*, 2016.

[8] Hilde Kuehne, Alexander Richard, and Juergen Gall. Weakly supervised learning of actions from transcripts. *CVIU*, 2017.

[9] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017.

[10] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *ECCV*, 2016.

[11] Jun Li, Peng Lei, and Sinisa Todorovic. Weakly supervised energy-based learning for action segmentation. In *ICCV*, 2019.

[12] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *CVPR*, 2017.

[13] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *CVPR*, 2018.

[14] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.

[15] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[16] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. 2019.