A Temporal Concept Localization Method Integrating Attention Mechanism

Jupan Li, Borui Li, Zhengdong Li, Jinchao Xia, Guanglei Zhang
Beijing eHualu
{lijupan01, liborui, lizhengdong, xiajch, zhangguanglei}@ehualu.com

Abstract

Temporal concept localization within video is a new challenging task in the field of video understanding, which aims to associate specific concepts with the corresponding video segments. In practice, it is very hard to collect segment-level annotations for massive datasets, making this problem very difficult to solve. To help solve this problem, Google Research introduces it as a Kaggle competition task, whose main focus is to better localize concepts of interest based on video-level labels and a small subset of segment-level labels. Motivated by the competition, we try our best to explore novel and effective approaches. Firstly, we tried the models of excellent performance during the 1st and 2nd competition. However, all these approaches do not grasp the key points of the problem. Considering that the encoder of the Transformer with mask is suitable for the situation in which training dataset and testing dataset are of different distribution, we introduced it to our solution for the competition task. Experiments shows that the average loss of validation is very close to that of training. Further more, UntrimmedNet, one of weakly supervised approaches for temporal action localization, was also introduced to our solution. Rank average is adopted by us to ensemble different models. Finally, we rank 27 on the private leaderboard.

1. Introduction

Temporal concept localization(TCL) aims at finding the start time and end time of some concept within a video, which helps easily discover and share memorable moments within long videos. In the past few years, many approaches have been proposed to address the problem[1]-[6]. These approaches require not only the ground truth of start time and end time but also class labels. However, manually annotating start time and end time for each concept for a new massive dataset is of high cost, which might hinder the application of these fully-supervised approaches to new areas that are lack of enough data with complete annotations. Currently, these fully-supervised approaches are applied on two datasets which only consists of thousands of videos, namely THUMOS14 [7] and ActivityNet [8].

To avoid expensive annotation cost, weakly supervised TCL approaches are proposed. They only require ground truth class labels for the whole video. UntrimmedNet [9] and AutoLoc [10] achieve the state-of-the-art performance. UntrimmedNet couples the classification module and the temporal module to simultaneously learn the class labels and time boundaries. Given a video, several video segments are sampled by uniform sampling or shot-based sampling and these samples are combined together to form a whole. Then the combined video is fed into a deep neural network to predict video-level class labels. During test, the trained network slides over time dimension to produce the classification score sequence of being each concept. Finally a simple thresholding method is applied to the score sequence to localize each concept instance in terms of the start time and the end time. AutoLoc decouples the classification module and the temporal module. A novel Outer-Inner-Contrastive loss is proposed to training the temporal boundary model, which dramatically improves the performance of TCL.

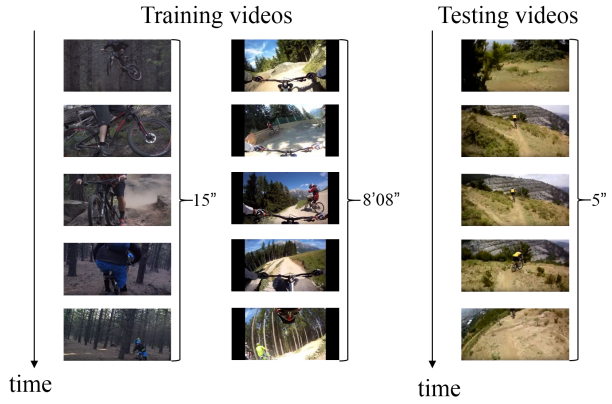

Figure 1: The training videos and testing videos of the 3rd YouTube-8M video understanding challenge.

However, almost all existing TCL approaches are performed on relatively small datasets, which are probably not suitable for the practice applications. Nowadays, Google AI releases a large-scale video dataset named YouTube-8M [11], which contains about 8 million YouTube videos with

multiple class labels.

The 1st YouTube-8M video understanding challenge requires participants to predict class labels of the whole video. Several approaches have been proposed, such as multi-stage training [12], context gating [13], temporal modeling [14], and feature aggregation [15]. However, the outstanding performance of these approaches is mainly attributed to the ensemble of multiple models, which is not suitable for practical applications due to the enormous computation. To promote more video understanding approaches performing well in practice applications, the 2nd YouTube-8M video understanding challenge concentrates on compact video understanding models. Specifically, the size of submission model is restricted to be under 1GB. During the competition, Model distillation [16], label denoising [17], NeXtVLAD [18], Non-local NetVLAD [19] and Temporal attention [20] are proposed. Neither the 1st nor 2nd YouTube-8M video understanding challenge focuses on the start time and end time of some concept within video. Therefore, the 3rd YouTube-8M video understanding challenge encourages participants to train models based on videos of various length and to predict class labels of short video segments. As illustrated in Figure. 1, videos in training dataset can be in any length, such as 15 seconds, 25 seconds, etc, while videos in testing dataset last for 5 seconds. Obviously, the 3rd YouTube-8M video understanding challenge focuses on temporal concept localization within videos.

In this paper we are committed to design approaches of excellent performance in the 3rd YouTube-8M video understanding challenge. Taking into account that training dataset and testing dataset are in different distribution, this paper introduces self-attention mechanism with mask to avoid zero padding, in which we call the approach TransformerEncode. Although MAP@100k on the public leaderboard is hardly improved after introducing the encoder of the Transformer with mask, the average loss of validation is much closer to that of training than excellent models of the 1st and 2nd competition.

2. Approach

Firstly, we tried the models of superior performance during the 1st and 2nd YouTube-8m video understanding competition, including NetFV, NetVLAD, NeXtVLAD, Dbof. However, these models based on the original training dataset have similar performance to the baseline given by the competition host. To bridge the gap between the training and testing dataset, we then randomly sample frames of 5 seconds in training dataset to train the models, which brings about a 0.02 performance boost. Despite the the improvement, the random sample approach does not catch the key to the problem. The mask mechanism used in the encoder of the Transformer is able to alleviate the problem that the training dataset and testing dataset have different distribution to some extent. Therefore, we introduced the Transformer's encoder into the classification model, which will be described in detail later. In addition, the 3rd competition task is essentially a weakly supervised learning problem. To try out weakly supervised approach for temporal concept localization is a natural idea. Based on this idea, we introduced the UntrimmedNet to our solution which is also described in the following part.

2.1. TransformerEncode

Transformer [21] is the first sequence transduction architecture based entirely on attention mechanism, which achieves state-of-the-art performance in multiple sequence to sequence tasks, such as machine translation. Like other sequence transduction model, the Transformer also follows an encoder-decoder structure. *N* identical layers compose the encoder. Each layer is composed of two sub-layers. The first is a multi-head self-attention module, and the second is a fully connected network. A residual connection is applied around each of the two sub-layers, followed by layer normalization.

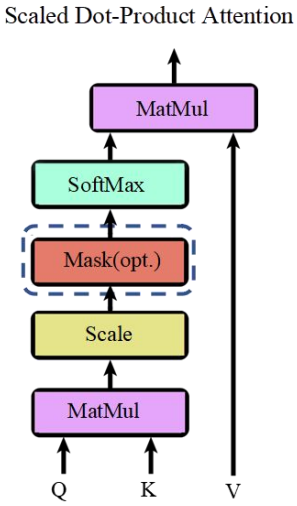

Figure 2. Scaled Dot-Product Attention.

The particular attention mechanism used in the Transformer is called "Scaled Dot-Product Attention", which is illustrated in Figure 2. The input is composed of queries, keys and values. We compute the dot products of the query with all keys, divide each other by the scale transformation parameter, and apply a softmax function to obtain the weights on the values.

In practice, sequences in the same dataset are usually of different length. When using such a dataset to train deep neural networks, it is necessary to perform length normalization. Zero-padding is one of commonly used transformation approaches. However, zero-padding is likely to have a bad influence on model training. The mask mechanism in "Scaled Dot-Product Attention" is introduced to reduce the side effects of zero-padding. Considering the training sample and the testing sample have significantly

different length, the encoder of the Transformer is quite suitable for the 3rd YouTube-8m video understanding challenge. Therefore, we introduce the encoder of the Transformer to our solution, which is called TransformerEncode. Figure 3 shows the framework of TransformerEncode, in which a classification module is introduced after the encoder of the Transformer. Different from sequence transduction tasks, the decoder is not necessary in the video understanding task.

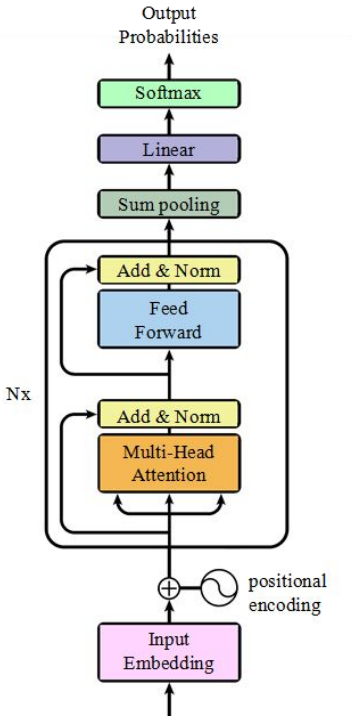

Figure 3: The TransformerEncode used for video classification task.

2.2. UntrimmedNet

UntrimmedNet is a weakly supervised architecture for temporal concept localization, which only requires class labels of the whole video. The classification module and the selection module are coupled together to simultaneously learn the models and reason about the temporal duration of interested concept respectively. These two components are implemented with feed-forward networks. UntrimmedNet is an end-to-end trainable architecture. Due to time constraints, the hard selection module is dropped in our solution, as shown in Figure 4.

Different from typical multi-class classification module, a fully connected layer of K nodes is used for classifying each feature vector into K classes and another fully connected layer with just one node is used for predicting the attention weight for each feature vector. The sum of the element-wise product of the attention weight and the output of the fully connected layer of K nodes indicates the classification distribution.

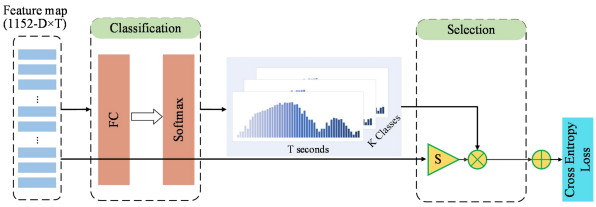

Figure 4. UntrimmedNet without the hard selection module.

3. Experiments

3.1. YouTube-8M Segments Dataset

The 3rd YouTube-8M Video Understanding Challenge provides the YouTube-8M Segments Dataset, which is an extension of the YouTube-8M dataset with human-verified segment annotations. About 237K segments on 1000 classes are collected for the validation set of the YouTube-8M dataset. In order to make predictions at segment-level granularity, each sample in the YouTube-8M Segments dataset comes with time-localized frame-level features. Compared to the YouTube-8M dataset, the size of the YouTube-8M segments dataset is much smaller.

3.2 Implementation Details

Since the YouTube-8M Segments Dataset is very small, we use the YouTube-8M Dataset for model training. The YouTube-8M Segments Dataset is used for model validation. Each single model is trained independently on Tensorflow [23]. The training procedure converges around 100,000 steps. Rank average is introduced to combine the advantages of different models after training, which is one of model ensemble approaches. Finally, we ensemble 5 models as the final submission.

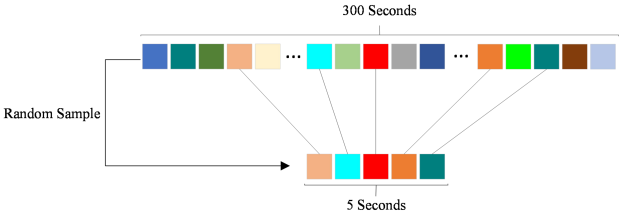

Figure 5. Illustration of random sample 5 seconds from training sample.

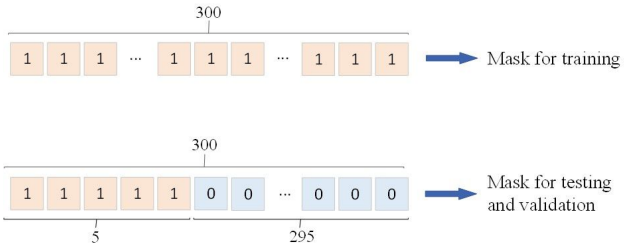

Figure 6. The training mask and the testing mask for the encoder of the Transformer.

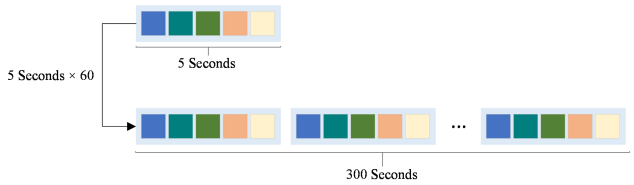

Figure 7. Illustration of tiling YouTube-8M segments dataset 60 times to acquire the same length as the training sample.

3.3 Single Model Evaluation

The six models adopted by us will be evaluated in this section. For the Dbof model, the same parameters as the baseline1 given by the competition host are used. Since NetFV and NetVLAD are both of big size, we random sample 5 seconds from each sample during training procedure for the NetFV and NetVLAD model, which is illustrated in Figure 5. For the Self-attention model, each training sample shares the same mask whose elements are all one, and each validating or testing sample shares the same mask whose first five elements are one and the other elements are all zero, as Figure 6 shows. Different from standard UntrimmedNet, the hard selection module based on the principle of multiple instance learning was dropped in our experiment. The performance of these six models on the Youtube-8M segments validation dataset is shown in Table 1. Among the six models, UntrimmedNet has the most parameters, and Logistic is of the smallest size. It is the soft selection module that brings more parameters for UntrimmedNet. In terms of average loss, Self-attention performs best, which may be due to the mask mechanism. The Logistic model used here consists of only one fully connected layer, which has the highest average loss. The performance of these six models on the public-test dataset is shown in Table 2, in which NetFV performs best. We ensemble all models shown in Table 1 as the final submission, which performance is shown in Table 3. In addition, we tried to tile validation and testing segments 60 times, as shown in Figure 7, which did not improve the performance yet.

Table 1. Single model performance on the Youtube-8M segments validation dataset

Model	Dbof	NetFV	NetVLAD
GAP@20	0.769	0.772	0.764
Avg_loss	9.315	7.816	7.266
Model size	549M	96.52M	264M

Model	Self-attention	Untrimmed Net	Logistic
GAP@20	0.774	**0.776**	**0.776**
Avg_loss	**5.789**	13.430	21.089
Model size	628M	814M	**51M**

Table 2. Single model performance on the public-test.

Model	Dbof	NetFV	NetVLAD
MAP@100k	0.712	**0.723**	0.719

Model	Self-attention	Untrimmed Net	Logistic
MAP@100k	0.695	0.712	0.706

Table 3. Performances of ensemble models on the public-test.

Ensemble Model	Public-test MAP@100k
M1&M2&M3&M4&M5&M6	0.742

4. Conclusions

In this paper, a detailed description for our various attempts for the 3rd YouTube-8M Video Understanding Challenge is presented. Like most other participants, excellent models of the 1st and 2nd competition were first introduced. To grasp the key points of the competition, we then introduce the encoder of the Transformer with mask and the UntrimmedNet. However, the performance of UntrimmedNet is not good as expected, which may be caused by the lack of the hard selection module. In the future, one can implement the hard selection module based on Tensorflow and visualize the selection module to improve the performance.

References

[1] Heilbron, Fabian Caba , J. C. Niebles , and B. Ghanem . "Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition IEEE, 2016.

[2] Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: ACM MM (2017)

[3] Richard, Alexander , and J. Gall . "Temporal Action Detection Using a Statistical Language Model." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2016.

[4] Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: Cdc: Convolutionalde-convolutional networks for precise temporal action localization in untrimmed videos. In: CVPR (2017)

[5] Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: CVPR (2016)

[6] Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: ICCV (2017)

[7] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. In http://crcv.ucf.edu/THUMOS14/, 2014.

[1] https://www.kaggle.com/c/youtube8m-2019/discussion/105867#latest-645318

[8] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In CVPR, pages 961–970, 2015.

[9] Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR (2017)

[10] Shou Z , Gao H , Zhang L , et al. AutoLoc: Weakly-supervised Temporal Action Localization in Untrimmed Videos[J]. 2018.

[11] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classi fi cation benchmark. arXiv preprint arXiv:1609.08675 (2016)

[12] Wang, H.D., Zhang, T., Wu, J.: The monkeytyping solution to the youtube-8m video understanding challenge. arXiv preprint arXiv:1706.05150 (2017)

[13] Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classi fi cation. arXiv preprint arXiv:1706.06905 (2017)

[14] Li, F., Gan, C., Liu, X., Bian, Y., Long, X., Li, Y., Li, Z., Zhou, J., Wen, S.: Temporal modeling approaches for large-scale youtube-8m video understanding. arXiv preprint arXiv:1707.04555 (2017)

[15] Chen, S., Wang, X., Tang, Y., Chen, X., Wu, Z., Jiang, Y.G.: Aggregating framelevel features for large-scale video classi fi cation. arXiv preprint arXiv:1707.00803 (2017)

[16] Skalic, Miha, and David Austin. "Building a size constrained predictive model for video classification." Proceedings of the European Conference on Computer Vision (ECCV). 2018

[17] Aliev, Vladimir, et al. "Label denoising with large ensembles of heterogeneous neural networks." Proceedings of the 2nd Workshop on YouTube-8M Large-Scale Video Understanding. 2018.

[18] Lin, Rongcheng, Jing Xiao, and Jianping Fan. "Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[19] Tang, Yongyi, et al. "Non-local netVLAD encoding for video classification." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[20] Kim, Eun-Sol, et al. "Temporal attention mechanism with conditional inference for large-scale multi-label video classification." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[21] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[22] Wang, Limin, et al. "Untrimmednets for weakly supervised action recognition and detection." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017..

[23] Abadi, Martín, et al. "Tensorflow: A system for large-scale machine learning." 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016.