Noise Learning for Weakly Supervised Segment Classification in Video

Zhaoyu Zhang[1], Xiang Wu[1], Jianfeng Dong[2,3], Yuan He[1], Hui Xue[1] and Feng Mao[1]
[1]Alibaba Group, Hangzhou, China
[2]Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou, China
[3]Zhejiang Gongshang University, Hangzhou, China
{zhaoyu.zzy, weiyi.wx, heyuan.hy, hui.xueh, maofeng.mf}@alibaba-inc.com
dongjf24@gmail.com

Abstract

This paper describes our solution for the 3rd YouTube-8M video understanding challenge. The challenge of this year is different from the previous challenge. Given a large scale video dataset with video-level labels and a small scale video dataset with segment-level labels, we are asked to recognize segments in videos this year. It can be regarded as a weakly supervised learning problem. To answer the challenge, we propose a solution consists of three different models, i.e., segment-level classifier, self-attention mechanism, noise learning classifier. Among them, the noise learning classifier performs the best. By noise learning, it can reduce the noise of label and sample for training, and improve the performance. Moreover, we achieve the MAP of 0.78878 in the private leaderboard by model ensemble based on introduced models, ranking the 8th place on the challenge.

1. Introduction

With the development of the internet, video becomes a popular way for people to share their life. Now, we can understand the content of video by video classification [1] and audio recognition [2]. As for knowing when events occurred, we can utilize the temporal localization method [3] by supervised learning.

However, the cost of supervised learning is very expensive because of labeling accurate category and time stamp for temporal localization in large scale of videos. Therefore, weakly supervised learning approach is an alternative solution to solve the problem.

Last year, video classification is still the only task in the 2^{nd} YouTube-8M video understanding challenge [4] [5] [6]. Fortunately, the 3^{rd} YouTube-8M video understanding challenge noticed this potential research field. This year, YouTube-8M Segments Dataset was provided. In the dataset, each segment indicates a topic in video. Normally, there is more than one topic in most videos.

1.1. Dataset

The YouTube-8M Dataset and YouTube-8M Segments Dataset are available this year. For the YouTube-8M Dataset, samples are split into 3 partitions: training, validation and test set, following nearly 70%, 20%, 10% split. The videos in the YouTube-8M Dataset are labeled with 3862 class and each video has an average of 3 video-level tags. The video-level data from training set and validation set can be used for pre-train model. In the YouTube-8M Segments Dataset, there are about 237K 5-second segments from the validation set of the YouTube-8M dataset. These segments are labeled with 1000 different classes. Segments in video are not all labeled because of expensive cost, only 5 segments per video are labeled on average. The dataset does not provide original videos, while providing 1024-dimensional Inception-v3 [7] feature and 128-dimensional audio feature per frame.

1.2. Evaluation

Following the evaluation protocol of the challenge, we report the score of Mean Average Precision@K (MAP@K). Specifically, we first predict the segment classification score, then sort the score in descending order. The MAP@K score based on the ranking order is computed as:

$$MAP@K = \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{k=1}^{n} P(k) \times rel(k)}{N_c}, \tag{1}$$

where K=100,000, C is the number of classes, $P(k)$ is the precision at cutoff k, n is the number of segments predicted per class, $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant class, or zero otherwise, and N_c is the number of positively-labeled segments for the each class.

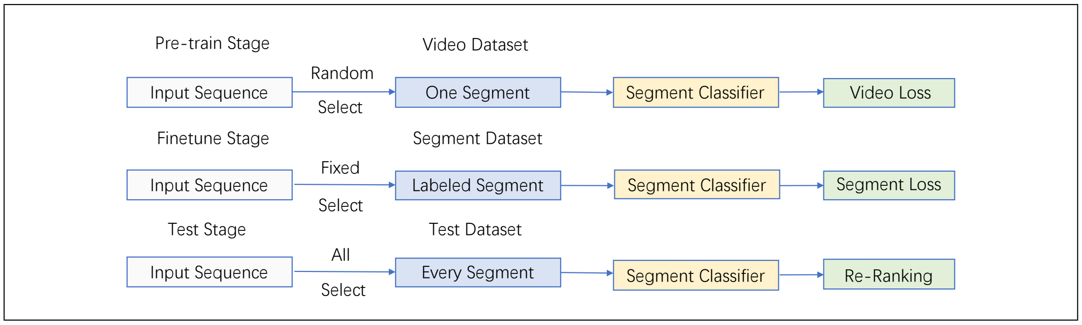

Figure 1. Overview of Baseline Solution. We divide solution into three stages. In the first stage, we pre-train on the video-level dataset. In the second stage, we finetune on the segment-level dataset. In the third stage, we test on the test dataset.

2. Related Work

2.1. Video Classification

The YouTube-8M video understanding challenge in the last two years mainly cares about video classification task. One of keypoints for video classification is how to efficiently aggregate the frames in video. Recurrent neural networks, such as LSTM (Long Short-Term Memory networks) [8] and GRU (Gated recurrent units) [9] are general solutions. These sequence modeling approaches can capture the temporal information in videos. However, not all videos need to be regarded orderly [10]. Hence, a number of methods that do not explicitly model the video order are proposed, such as DBoF (Deep Bag of Frames networks) [11], NetFV [12] and NetVLAD [13]. The methods are also obtained excellent performance for video classification. For instance, Lin et al. proposed NextVLAD and ensemble method of MixNextVLAD to improve the original NetVLAD by group convolution in the 2^{nd} YouTube-8M video understanding challenge. Following these good practices, we also utilize NextVLAD variants for feature aggregation.

2.2. Noise Learning

Noise learning is one of the weakly supervised learning method which has been widely used in image field. Most of noise learning solves the problem of label noise from raw data. For instance, Hu et al. [14] create a residual structure to reflect the difference between noise data and clean data. Guo et al. [15] propose CurriculumNet to learn process with designed curriculum and handle a massive amount of noise labels and data imbalance effectively. Both two manners serve as noise regularization and increase the generalization of model.

Method	MAP
Nextvlad	0.762
Nextvlad + Early Attention	0.771
MixNextvlad + Early Attention	0.779
MixNextvlad + Late Attention	0.782

Table 1. Self-Attention Mechanism Results

3. Weakly Supervised Classification Method

3.1. Problem Definition

Given a video $X = \{x_n\}_{n=1}^{N}$ consists of N segments, where x_n indicates a 5-second segment in X, we are asked to predict the label for each segment. Now, we have large scale of video-level label and small scale of segment-level label. And we would like to get segment classification result. It is obviously a weakly supervised learning problem.

3.2. Segment-Level Classifier

As shown in Figure 1, segment-level classifier is the baseline solution for weakly supervised segment classification. By learning the solution in past competition, we all know pooling approach works well, such as NetVLAD, NetFV, DBoF, NextVLAD and MixNextVLAD. So, we chose the best MixNextVLAD model for the backbone of segment-level classifier. First, we selected a segment from video randomly in the pre-train stage, then trained the segment-level classifier on the training dataset and the validation dataset in the YouTube-8M Dataset by treating video-level label as segment-level label. Second, we finetuned the segment-level classifier on all segment-level data in the YouTube-8M Segments Dataset. Finally, we predicted the result on the test data in the YouTube-8M Dataset by the segment-level classifier.

3.3. Self-Attention Mechanism

Self-attention mechanism approach is another solution on the pre-train stage. By using this method, we would like

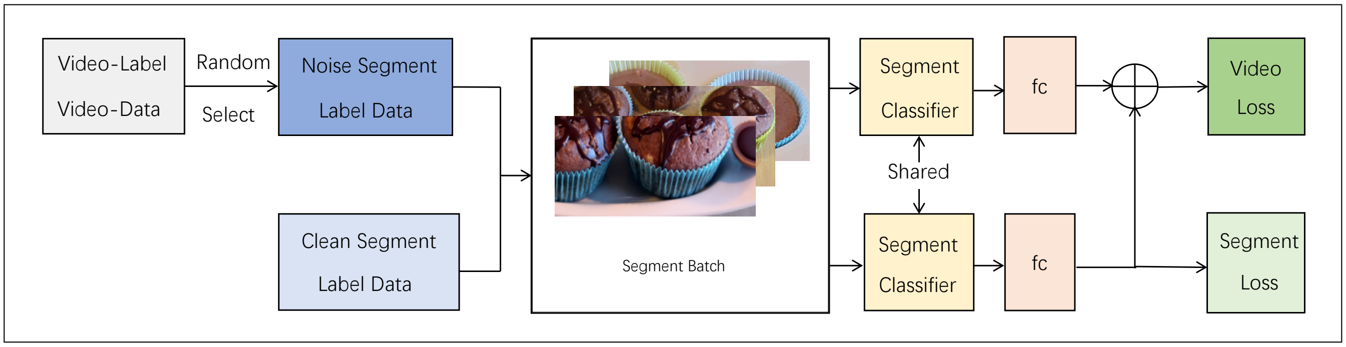

Figure 2. Overview of Noise Learning Classifier. We randomly select a segment from video in the YouTube-8M dataset as a noise segment. After stacking several of these noise segments and a clean segment in the YouTube-8M Segments Dataset, we send them to segment classifier separately. Then we utilize the residual structure to reflect the difference between the noise segment and clean segment. Finally, we calculate the video loss and segment loss at the same time.

Method	Segment	MAP
MixNextvlad	1	0.783
MixNextvlad + Attention2	2	0.781
MixNextvlad + Attention3	3	0.782
MixNextvlad + Attention4	4	0.777

Table 2. Segment of Different Amount Results

Method	MAP
Dbof + Transfer Learning	0.745
Dbof + Noise Learning	0.764
Netvlad + Transfer Learning	0.758
Netvlad + Noise Learning	0.768
Nextvlad + Transfer Learning	0.771
Nextvlad + Noise Learning	0.782

Table 3. Noise Learning Classifier Results

to figure out whether the relation between segments would have great influence on the recognition of video-level label. Therefore, we computed the similarity of segment feature or classification score among several segments, then use this similarity as weight to combine each segment classification score. In this way, we hope to get better pre-train model and it actually worked in the multi-segment classifier. But when we used multi-segment classifier at the pre-train stage, it's even a little worse than single segment classifier mentioned above. Maybe it is caused by the inconsistency when we train on the video-level label and then test on the segment-level label. Meanwhile, segment label is single label, but video label is multi-label. Also random selected sample will produce huge noise.

3.4. Noise Learning Classifier

For problems mentioned above, we try to train on the segment-level label and test on the segment-level label. Also, we would better to reduce the noise of label and sample. In order to solve the problem of noise, we proposed the noise leaning approach for videos by utilizing residual structure in the multi-task learning field. It was inspired by noise learning in images [14]. And this stage may be regarded as the improved version of finetune stage.

Figure 2 shows the overview solution of noise learning classifier in video based above general segment-level classifier. After the pre-train stage, we first randomly selected a segment data from video data (the YouTube-8M Dataset), then we marked the segment data by video label and apparently made the noise segment dataset. At the same time, we added the original clean segment dataset(the YouTube-8M Segments Dataset). Then we sampled from noise dataset and clean dataset to form the segment batch. The segment batch would be separately sent to two weight shared segment classifiers which are only different in fully connected layer. Finally, we added two classification scores to compute the video-level loss by segments from the noise dataset. However, the segment-level loss was computed purely by segment from clean dataset in the second classifier. This is the whole process of the multi-task learning we proposed.

In order to train the model, we denote the loss of video and segment as L_{vid} and L_{seg}. They can be formulated as follows

$$L_{vid} = -\frac{1}{N_n} \sum_{i \in D_n} p_i log l_i + (1 - p_i) log(1 - l_i), \qquad (2)$$

$$L_{seg} = -\frac{1}{N_c} \sum_{j \in D_c} p_j log l_j + (1 - p_j) log(1 - l_j), \qquad (3)$$

where N_n is the number of samples in noise dataset. N_c is the number of samples in clean dataset. i is a sample from noise dataset D_n. j is a sample from clean dataset D_c. p_i indicates predict of sample i. p_j indicates predict of sample j. l_i is a label of sample i. l_j is a label of sample j.

Method	MAP
MixNextvlad + Nextvlad Noise + MixNextvlad Attention2	0.79757
MixNextvlad + Nextvlad Noise + MixNextvlad Attention2 + Netvlad Noise	0.79765
MixNextvlad + Nextvlad Noise + MixNextvlad Attention2 + Dbof Noise + MixNextvlad Attention3	0.79743
MixNextvlad + Nextvlad Noise + MixNextvlad Attention2 + Netvlad Noise + MixNextvlad Attention3	0.79756

Table 4. Final Ensemble method Results

And total loss can be formulated as follows

$$L_{total} = L_{seg} + \alpha L_{vid}, \tag{4}$$

where α is a trade-off parameter between losses.

4. Experiments

4.1. Implementation Details

For noise learning classifier, it is based on NextVLAD [6] backbone model. In the pre-train stage, we randomly selected 5-second segments from the original video training and validation data in the YouTube-8M Dataset for training. The learning rate is initialized as 0.0002 and the optimization would be complete after 5 epochs. In the noise learning stage, we choose samples from both clean dataset D_c and noise dataset D_n in a ratio of 1:20. Furthermore, we also set α 20. The learning rate is also initialized as 0.0002 and the optimization would be complete after 3 epochs. In the test stage, we only use one of the segment classifiers for clean dataset to predict the top100 classification score.

4.2. Results

It is worth noting that all the performance scores in the following tables are the public leaderboard which depends on the evaluation results on approximately 20% of the test data. The results of Section 3.3 are summarized in Table 1. These models are all trained with three segments like temporal segment networks [16]. And Nextvlad is the baseline method in pre-train stage. Early attention means attention weight is computed by feature after fusion. Late attention means attention weight is computed by score after fusion. Table 1 shows Late Attention is the best way to combine information from various segments. It achieves MAP score of 78.2%, making the improvement of 2% over the baseline. As shown in Table 1, self-attention mechanism works well in pre-train stage for video classification surely.

Additionally, we further the effect of segment number used in the self-attention method. As shown in Table 2, the classifier with the segment number of 1 performed best. It is even a little better than original methods. The potential reason we have analyzed above.

So, we proposed noise learning methods. And the result is shown in Table 3. We can see noise leaning works better than general transfer leaning in the finetune stage. Among these results, noise learning classifier based on Dbof model can make improvement over transfer learning by nearly 2%. Netvlad model was improved by 1%. When testing on Nextvlad model, MAP was also improved by 1%. It has confirmed that our noise learning classifier can learn the difference between the clean data and the noise data. Thus, the MAP result of single model has reached top.

4.3. Ensembling

For better performance, we combined various models from three base methods mentioned above. Ensemble results are reported in Table 4. We attempted to combine different amount of models. First, we computed the average score of 1000 class classification result of above models, then just chose top100 score to output the final result. In this way, we made the improvement over the single model by nearly 1.5%. Finally, we chose the last ensemble method in Table 4 as competition result and got score of 0.78878 in the private leaderboard which is calculated with approximately 80% of the test data.

5. Conclusions

Compared with segment-level classifier and self-attention mechanism methods, we can draw conclusion that our proposed noise learning classifier is effective for weakly supervised learning in videos. And by combining various excellent single models, we finally made further progress with ensemble models for our solution in the 3rd YouTube-8M video understanding challenge.

References

[1] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[2] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[3] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localiza-

tion in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5734–5743, 2017.

[4] Miha Skalic and David Austin. Building a size constrained predictive model for video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[5] Pavel Ostyakov, Elizaveta Logacheva, Roman Suvorov, Vladimir Aliev, Gleb Sterkin, Oleg Khomenko, and Sergey I Nikolenko. Label denoising with large ensembles of heterogeneous neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[6] Rongcheng Lin, Jing Xiao, and Jianping Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[10] Feng Mao, Xiang Wu, Hui Xue, and Rong Zhang. Hierarchical video frame sequence representation with deep convolutional graph network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[11] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

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

[13] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[14] Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. Weakly supervised image classification through noise regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11517–11525, 2019.

[15] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[16] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 2018.