CMU-AML Submission to Moments in Time Challenge 2018

Po Yao Huang
School of Computer Science
Carnegie Mellon University
poyaoh@andrew.cmu.edu

Xiaojun Chang
SCS, Carnegie Mellon University
Hangzhou Anmeilong Intelligence Co., Ltd.
cxj273@gmail.com

Alexander G. Hauptmann
School of Computer Science
Carnegie Melon University
alex@cs.cmu.edu

Abstract

*In this report, we describe our solution for Moments in Time Challenge 2018. We employed both visual and audio features in the submission. For visual features, we utilize the preprocessed RGB and optical flow data for training or fine-tuning 2D (*e.g. Temporal Segment Network (TSN) *and 3D (*e.g. *Inflated 3D ConvNets (I3D). For audio features, we use raw waveforms as the input modality and fine-tune the feature extracted from the last pooling layer of SoundNet. We achieve 31.56% in terms of Top-1 accuracy and 59.75% in terms of Top-5 accuracy on the validation set.*

1. Introduction

The last decodes have witnessed the success of deep learning in image understanding tasks, *i.e.* classification [8], segmentation [11], and *etc*. Researchers have demonstrated the superiority of state-of-the-art Convolutional Neural Networks (CNN) [9, 3] against traditional algorithms with hand-crafted features. Inspired by the progress, CNNs have been widely employed to improve the performance of video understanding tasks. Compared to image understanding tasks, temporal information of videos can boost the performance of video classification. Additionally, auditory soundtracks provides an additional clue for video analysis.

We cannot obtain discriminate models without large-scale labeled dataset, such as ImageNet [2] and ActivityNet [4]. Recently, the MIT-IBM Watson AI Lab has released a large-scale Moments dataset [7] to help AI systems recognize and understand actions and events in videos. This dataset contains a collection of one million labeled 3 second videos, involving people, animals, objects or natural phenomena, that capture the gist of a dynamic scene. The Moments in Time Challenge 2018 is based on this dataset.

2. Our Approach

In this section, we describe the features and models we used for the challenge. We use the standard split defined in the original paper where 802,244 training video and 37,800 validation video are available.

2.1. Features

Visual Features: All the videos are first resized to 340×256 under 30 fps. We rescale raw RGB values into $\big[-1, 1\big]$. We also computed optical flow with the TVL1 algorithm and rescale the value into $\big[-1, 1\big]$.

We utilize the preprocessed RGB and optical flow data for training or fine-tuning 2D(e.g. TSN) and 3D(e.g. I3D [1]) models. In order to fit the relatively shorter but constant period for the targeted Moments in Time dataset, we use dilated frames (with a fixed M network input size with step size $\lfloor N/(M-1) \rfloor$, where N is the frame size) as inputs instead of consecutive frames.

To leverage the knowledge from other dataset, we also use existing models pre-trained on external large datasets such as Kinetics. In practice, we use the RGB and Optical flow models pre-trained on ImageNet and as the feature extractors to extract the features reside in last pooling layer as new video representation. Specifically, we sample the center frames of a video as the input and store a $(7, 1024)$ vector for each video. This approach is equivalent to fine-tuning the layer above last pooling layer of a model.

Audio Features: We average the two channels and re-

sample the audio into 22,050 Hz .wav files. For videos without audio channel, we fill a 3-second silent audio for them. We extract the conv7 layer of the soundnet model, which is pretrained over 2,000,000 unlabeled videos. Then we feed the features into a 10-layer DenseNet [5] with the output layer changed to predict moment categories.

2.2. Models

Fine-tune all models For this challenge, we fine-tune 2D(spatial) and 3D(spatial-temporal) models with additional layers.

For 2D models, built upon TSN, we add an additional cutout layer. Each sampled (340×256) frame will be randomly cut out with a (90×90) region. We also tried other augmentation techniques such as mix-up but found cut-out is the most feasible one.

For 3D models, we use I3D models with ResNet 50 (R50) as its backbone. In addition to cutout augmentation layer we add an addition non-local layer to capture the interaction between spatial-temporal units. As in [10], we add 10 non-local blocks to R50.

Considering the size of the target dataset, we choose to use network with 8 frame inputs. Empirically we found that ImageNet pre-trained I3D model with non-local networks are prone to overfit for Moment in time dataset in comparison to 3D models without non-local networks. A better choice is to use ImageNet-Kinetics pre-trained models where we observed preferred behaviors.

Fine-tune last models As described before, we utilize the ImageNet-Kinectics pre-trained I3D models as the spatial-temporal feature extractor. Take (7,1024) features as the input, we randomly sample and average 2 frames then feed to the classification network.

For the classification, we choose the mixture-of-residual expert (MoRE) network as proposed in[6] with 4 experts with two-layer network (each layer with 2048 neurons) with residual links as the classification model for RGB and optical flow features. We found that with pre-extracted feature the network are prone to overfit and therefore apply a high dropout rate (0.8) and append an input batch-normalization later to train the model.

2.3. Training and Inference Details

For training finetune-all models, we use 3-Titan XP GPUs with batch size 24 and standard momentum SGD. With limited resource and time we train each 3D models for 20 epochs. The learning rate is set 0.005 and decayed by 0.1 at 10, 16, 18 epochs respectively. It take roughly 4 days to train a model. For inferencing, we sample 5 inputs (each with 8 frams) from a video then mean-pool the predictions as the video-level prediction.

Training finetune-last models are comparably cost-effective. We use one Titan XP GPU with batch size 512 and Adam optimizer and train for 80 epochs. The learning rate is set 0.001 and decayed by 0.1 at 30, 50, 70 epochs. At testing phase, we loop every frame of a video and generate frame-wise prediction then mean-pool the results.

2.4. Evaluation metric

Following the stand of the Moments challenge, we employ top-k accuracy as the evaluation metric. For each video, the system will generate k labels $l_j, j = 1 \dots k$. The ground truth label for the video is g. The error of the algorithm for that video would be:

$$e = \min_j d(l_j, g), \tag{1}$$

where $d(x, y) = 0$ if $x = y$ and 1 otherwise. The overall error score for an algorithm is the average error over all videos. We use $k = 1$ and $k = 5$.

2.5. Fusion

In this report, we fuse multiple features for video classification. We learn the optimal weights for different features on the validation set. Then we apply these weights on the testing set, and get the results for the final submission.

3. Results

In this section, we first evaluate the performance of the individual feature on the validation set. The performance are shown in Table 1. From the experimental results we can observe that I3D with non-local network have better performance than I3D without non-local network. For example, the performance of I3D with NLN improves the performance of I3D without NLN from 28.96 to 29.48 in terms of top-1 accuracy. However, we observe that I3D with NLN is prone to overfit. For example, when we use the ImageNet pretrained I3D with NLN, we get only 25.94 in terms of top-1 accuracy. This demonstrates the necessity of using ImageNet and Kinects pre-trained network to avoid overfitting.

After that, we learn the optimal weights for individual features on the validation set. The weights of utilized features are shown in Table 2. With these weights, we have obtained 31.56 in terms of Top-1 accuracy and 59.75 in terms of Top-5 accuracy on the validation set, respectively.

4. Conclusion

In this report, we have presented our solution to the Moments in Time Challenge 2018. We found that Inflated 3D ConvNets (I3D) with non-local networks has the best single model performance. However, we found that ImageNet pre-trained I3D model with non-local networks are prone to overfit for the challenge dataset. Hence, we choose to use ImageNet-Kinects pretrained models where we observed preferred performances.

Table 1. Performance evaluation of different features on the validation set.

Type	Model	Pre-Trained	val Top-1	val Top-5	Final Fusion
TSN-Spatial	Baseline	I+K	24.07	48.98	T
TRN-Multiscale	Baseline	I+K	21.02	43.27	T
Audio	SoundNet	U	6.83	15.41	T
2D-RGB	Last-RGB	I+K	19.84	41.75	T
2D-OF	Last-Optical Flow	I+K	18.49	39.64	T
3D-RGB	All-vanilla-I3D (8 frames)	I	28.12	56.04	F
3D-RGB	All-I3D (8 frames)	I	28.96	56.45	T
3D-RGB	All-I3D-NLN (8 frames)	I	25.94	52.60	F
3D-RGB	All-I3D-NLN (8 frames)	I+K	25.75	53.31	F
3D-RGB	All-I3D-NLN (16 frames)	I+K	29.48	57.37	T

Table 2. Weights of different features and the fusion results on the validation set.

Type	Model	Weights
TSN-Spatial	Baseline	2
TRN-Multiscale	Baseline	2
Audio	SoundNet	1
2D-RGB	Last-RGB	2
2D-OF	Last-Optical Flow	1
3D-RGB	All-I3D (8 frames)	8
3D-RGB	All-I3D-NLN (8 frames)	3
3D-RGB	All-I3D-NLN (16 frames)	12
Fusion Result	Top-1: 31.56	Top-5: 59.75

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733, 2017. 1

[2] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. 1

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. 1

[4] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 961–970, 2015. 1

[5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269, 2017. 2

[6] P.-Y. Huang, Y. Yuan, Z. Lan, L. Jiang, and A. G. Hauptmann. Video representation learning and latent concept mining for large-scale multi-label video classification. *CoRR*, abs/1707.01408, 2017. 2

[7] M. Monfort, B. Zhou, S. A. Bargal, A. Andonian, T. Yan, K. Ramakrishnan, L. M. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva. Moments in time dataset: one million videos for event understanding. *CoRR*, abs/1801.03150, 2018. 1

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826, 2016. 1

[10] X. Wang, R. B. Girshick, A. Gupta, and K. He. Non-local neural networks. *CoRR*, abs/1711.07971, 2017. 2

[11] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei. Fully convolutional adaptation networks for semantic segmentation. *CoRR*, abs/1804.08286, 2018. 1