Submission to Moments in Time Challenge 2018

Yunkai Li[1], Ziyao Xu[1], Qian Wu[2,†], Yu Cao[3,†], Shiwei Zhang[4,†], Lin Song[5,†], Jianwen Jiang[6,†], Chuang Gan[6*], Gang Yu[1*], Chi Zhang[1*]
[1]Megvii Inc. (Face++), {liyunkai, xuziyao, yugang, zhangchi}@megvii.com
[2]Zhejiang University, wq1601@zju.edu.cn
[3]Beihang University, cqcy1208@buaa.edu.cn
[4]Huazhong University of Science and Technology, swzhang@hust.edu.cn
[5]Xian Jiaotong University, stevengrove@xtu.xjtu.edu.cn
[6]Tsinghua University, jjw17@mails.tsinghua.edu.cn, ganchuang1990@gmail.com

***Abstract*—This paper introduces our solution for the full track of the Moments in Time 2018 video event recognition challenge. Our system is built on spatial networks and 3D convolutional neural networks to extract spatial and temporal features from the videos. We also take advantage of multi-modality cues, including optical flow and audio information to further improve the performances. Our final submission is an ensemble of 5 models: three based on RGB frames as well as one optical flow model and one audio model, achieving top1 38.1%, top5 65.3% on the validation set.**

I. Introduction

Video recognition is one of the most fundamental research topics in the computer vision. With development of computation and release of large video classification dataset such as Kinetics [8] and Moments in Time [13], it has therefore been an urgent need to develop more efficient automatic video understanding and analysis algorithms.

Currently, there are three kinds of successful frameworks that dominate the video recognition (1) two-stream CNNs [15], [17], (2) 3D CNNs [16] and its variant [14], [18], and (3) 2D CNNs with temporal models on top such as LSTM [3], [9], temporal convolution [1] and attention modeling [10], [11]. The winner of Kinetics challenges last year [1] proposed a novel solution by first extracting the multi-modality features from the learned networks and then fed them into the off-shelf multi-modality temporal models to conduct video classification. However, these approaches are not applicable to large-scale video datasets, such as Moments in Time [13], since they rely on extracting features from all videos beforehand, which is extremely time-consuming and expensive.

To address these challenge, we mainly adopt end-to-end training architectures with three modalities, namely appearance, motion and acoustic information. We compared the performance of different models and finally chose Inflated 3D and Non-local module for appearance modality and 2D CNN model for the motion and acoustic modalities.

The remaining sections are organized as follows. Section II presents some details of our method. In section III, we compare different approaches, followed by the conclusion of this report in section IV.

† Work down while interning at Megvii

II. The proposed Method

In this section, we will introduce the applied multiple models and modality, including observation and obtained score for each model.

A. Appearance clues

We have experimented different methods including 2D CNNs, Temporal Segment Networks and inflated 3D neural network to extract the video feature. We extract RGB frames from the videos at 25 fps as original resolution and applied random crop as augmentation.

Spatial Network. We used Xception network [2] pre-trained on the Kinetics dataset [8], as well as SENet and SEResNeXt [6] initialized on ImageNet.

In training, one single RGB image is randomly selected from the video as the input. In validation, we followed the testing method in TSN [17] with 25 segmentation and average fusion.

Temporal Segment Networks. We also explored the Temporal Segment Networks with 5 segments. Surprisingly, it is not as effectively as in other activity recognition tasks. The performance of TSN is even lower than single image performance described above. We speculate that it is due to the large intra-classes variances and short video duration of the dataset.

Inflated 3D Network. We combine the Inflated ResNet50 network with non-local modules as the base model.

We apply spatial and temporal convolution separately in the ResNet block [5], which improves accuracy while reducing the calculations. We pre-trained the model in Kinetics, and fine-tuned the network as the base model of TSN. In validation inference, we crop the input lager than 224, with spatially average the predictions after the Softmax layer as described in [4].

Word2Vec Network. Considering the large intra-class variance of the dataset, We tried to transfer labels into vectors and minimize the distance between feature and vectors instead of classification loss.

B. Motion clues

We used an OpenCV implementation of TV-L1 [19] algorithm for computing dense optical flow and converted 2-

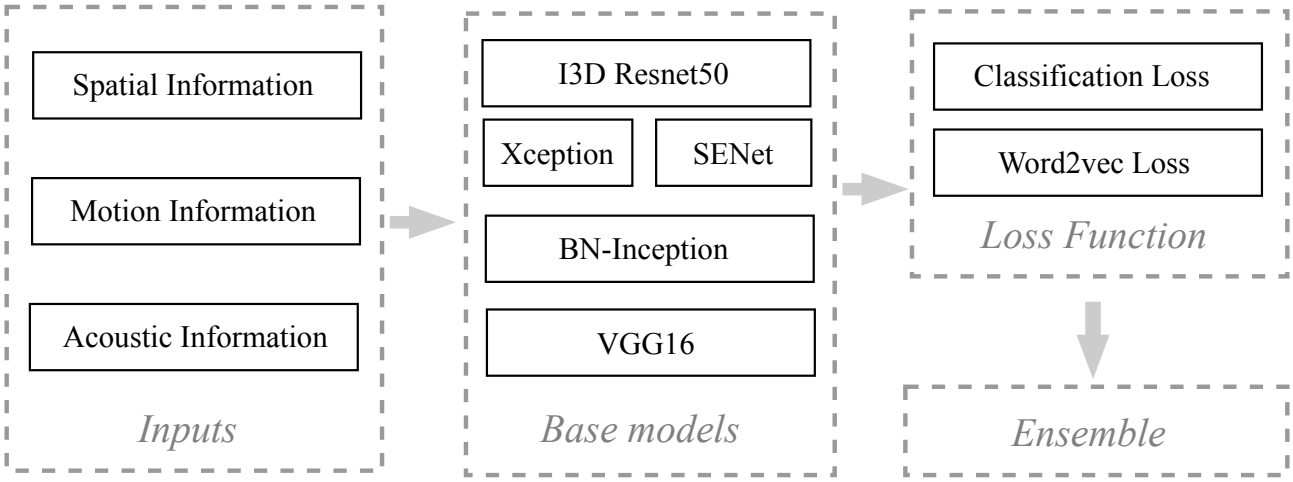

Fig. 1. The designed framework in our method. we apply 2D and 3D convolutional neural network to extract spatial and temporal feature from the video while take advantage of multi-modality cues, including optical flow and audio information. Our final submission is an ensemble of these models.

TABLE I
RESULTS ON VALIDATION SET.

Model	Modality	Top-1 Accuracy(%)	Top-1 Accuracy(%)
ResNet50	RGB	28.3	53.2
TSN	RGB	27.4	53.2
Xception	RGB	31.8	59.2
SENet152	RGB	33.7	61.3
SEResNeXt	RGB	33.0	60.2
Word2Vec ResNet50	RGB	29.9	56.2
I3d ResNet50	RGB	34.2	61.4
BN-Inception	Flow	19.1	41.2
VGG16	Audio	9.1	21.3
Ensemble		**38.1**	**65.2**

channel optical flow vectors (u, v) into its magnitude and direction and stored them as RGB images. We used these images in the BN-Inception [7] network and takes a stack of 5 consecutive optical flow fields as input. We employed SGDR [12] strategy in optical flow training, since we found that restart the learning rate is helpful to promote the accuracy. We obtained a validation accuracy of 19.09% (top-1,) 41.17% (top-5)

C. Acoustic clues

In compliance with the common practice to processing audio features, a convolutional network based audio classification system is used. With each video divided into 10 frames, its frequency domain information is extracted through Fourier Transformation, histogram integration and logarithm transformation. The vocal information of each video is shaped as 10x96x64 to a VGG classification net to generate a label probability distribution prediction.

The character that vocal information is hard to do augmentation makes it likely to overfit the training set. So generally less complex net leads to a better evaluation results.

D. Training

In this section, we present some details of our method during training stage. We train the our network end-to-end with 0.01 initial learning rate and reducing it by a factor of 10 at every 15 epoches. For each RGB and acoustic model, we train about 30 epoches and 60 epoches for flow model. We train our model on the 8 Titan GPUs for single image and TSN experiment, while 3D models are trained in distribution mode.

III. EXPERIMENT RESULTS

In this section, we present some experiments in our method in the Table II-A. In this table, we show the results with different 2D/3D mdoels. From the results, we can find that i3d resnet50 with model non local can achieve the best results. While the single image method accuracy is actually not much lower than i3d network, TSN performs not as efficient as other datasets. We found that spatial and temporal information are mutually complementary for final feature fusion. Meanwhile, motion and acoustic information are essential though the scores are low, showing the importance of different modality clues.

Finally, we ensemble all the models on the score after softmax function to obtain 38.1% top-1, and 65.2% top-5 accuracy on the validation set.

IV. Conclusion

In Moments in Time Challenge 2018, we design a new spatio-temporal action recognition framework. We make advantage of both 2D spatial network and 3D network, as well as multi modalities. By this means, we can better extract the feature of the video in more patterns. In the future, we will explore the fundamental difference between Moments in Time dataset with other datasets and find better general presentation under large variance in intra-class distribution.

References

[1] Y. Bian, C. Gan, X. Liu, F. Li, X. Long, Y. Li, H. Qi, J. Zhou, S. Wen, and Y. Lin. Revisiting the effectiveness of off-the-shelf temporal modeling approaches for large-scale video classification. *arXiv preprint arXiv:1708.03805*, 2017.

[2] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv preprint arXiv:1610.02357*, 2016.

[3] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. pages 2625–2634, 2015.

[4] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, R. Khrisna, V. Escorcia, K. Hata, and S. Buch. Activitynet challenge 2017 summary. *arXiv preprint arXiv:1710.08011*, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[8] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

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

[10] X. Long, C. Gan, G. de Melo, X. Liu, Y. Li, F. Li, and S. Wen. Multimodal keyless attention fusion for video classification. AAAI, 2018.

[11] X. Long, C. Gan, G. de Melo, J. Wu, X. Liu, and S. Wen. Attention clusters: Purely attention based local feature integration for video classification. *CVPR*, 2018.

[12] I. Loshchilov and F. Hutter. Sgdr: stochastic gradient descent with restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[13] M. Monfort, B. Zhou, S. Adel Bargal, A. Andonian, T. Yan, K. Ramakrishnan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, and A. Oliva. Moments in Time Dataset: one million videos for event understanding. *ArXiv e-prints arXiv:1801.03150*, 2018.

[14] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

[15] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497. IEEE, 2015.

[17] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. *ECCV*, 22(1):20–36, 2016.

[18] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. *arXiv preprint arXiv:1711.07971*, 2017.

[19] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007.