Revisiting the Effectiveness of Off-the-shelf Temporal Modeling Approaches for Large-scale Video Classification

Yunlong Bian, Chuang Gan* , Xiao Liu, Fu Li, Xiang Long
Yandong Li, Heng Qi, Jie Zhou, Shilei Wen, Yuanqing Lin
Baidu IDL & Tsinghua University

Abstract

*This paper describes our solution for the video recognition task of ActivityNet Kinetics challenge that ranked the 1st place. Most of existing state-of-the-art video recognition approaches are in favor of an end-to-end pipeline. One exception is the framework of DevNet [3]. The merit of DevNet is that they first use the video data to learn a network (*i.e. *fine-tuning or training from scratch). Instead of directly using the end-to-end classification scores (*e.g. *softmax scores), they extract the features from the learned network and then fed them into the off-the-shelf machine learning models to conduct video classification. However, the effectiveness of this line work has long-term been ignored and underestimated. In this submission, we extensively use this strategy. Particularly, we investigate four temporal modeling approaches using the learned features: Multi-group Shifting Attention Network, Temporal Xception Network, Multi-stream sequence Model and Fast-Forward Sequence Model. Experiment results on the challenging Kinetics dataset demonstrate that our proposed temporal modeling approaches can significantly improve existing approaches in the large-scale video recognition tasks. Most remarkably, our best single Multi-group Shifting Attention Network can achieve 77.7% in term of top-1 accuracy and 93.2% in term of top-5 accuracy on the validation set.*

1. Introduction

Video understanding is among one of the most fundamental research problems in computer vision and machine learning. The ubiquitous video acquisition devices (e.g., smart phones, surveillance cameras, etc.) have created videos far surpassing what we can watch. It has therefore been a pressing need to develop automatic video understanding and analysis algorithms for various applications.

To recognize actions and events in videos, recent approaches based on deep convolutional neural networks (CNNs) [9, 13, 3, 17, 4] and/or recurrent networks [7, 15, 1] have achieved state-of-the-art results. However, due to the lack of public available datasets, existing video recognition approaches are restricted to understand small-scale data, while large-scale video understanding remains an underaddressed problem. To remedy this issue, Google DeepMind releases a new large-scale video dataset, named as Kinetics dataset [10], which contains 300K video clips of 400 human action class.

To address this challenge, our solution follows the strategy of DevNet framework [3]. Particularly, we first learn the basic RGB, Flow and Audio neutral network models using the videos. Then we extract the multi modality feature and fed them into different off-shelf temporal models. We also design four novel temporal modeling approaches, namely Multi-group Shifting Attention Network, Temporal Xception Network, Multi-stream sequence Model and Fast-Forward Sequence Model. Experiment results verity the effectiveness of the four models over the traditional temporal modeling approaches. We also find that these four temporal modeling approaches are complementary with each others and lead to the state-of-the-arts performances after ensemble.

The remaining sections are organized as follows. Section 2 presents the basic multi modal feature extraction. Section 3 describe our proposed off-shelf temporal modeling approaches. Section 4 reports empirical results, followed by discussions and conclusions in Section 5.

2. Multimodal Feature Extraction

Videos are naturally multimodal because a video can be decomposed into visual and acoustic components, and the visual component can be further divided into spatial and temporal parts. We extracted multi modal features to best represent videos accordingly.

2.1. Visual Feature

As in [13], we used RGB images for spatial feature extraction and stacked optical flow fields for temporal fea-

*Corresponding author.

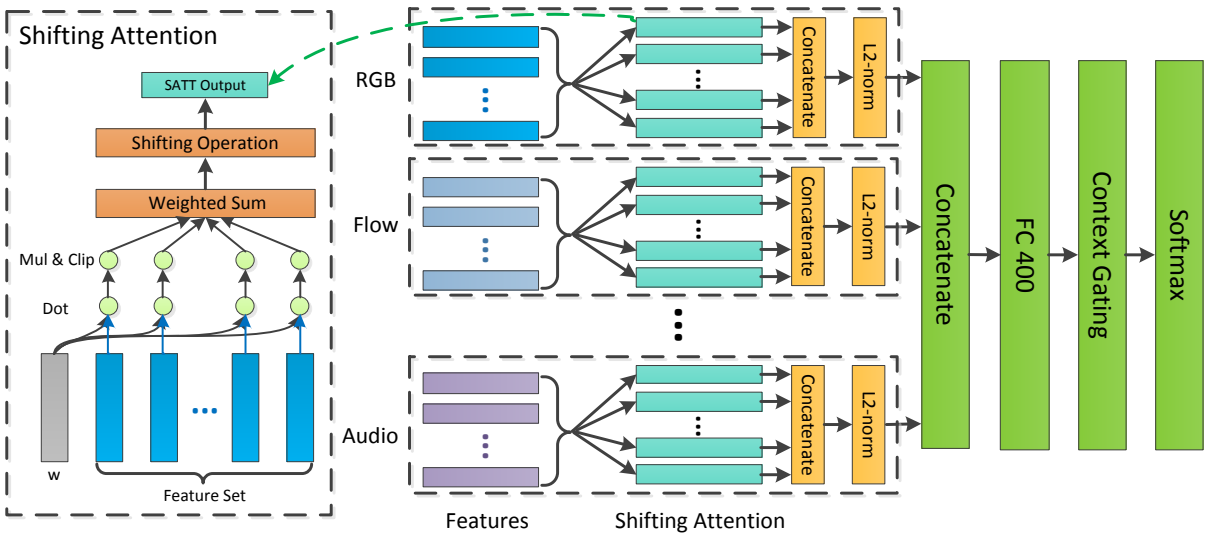

Figure 1. Multi-group Shifting Attention Network.

ture extraction. We tried different ConvNet architectures and found Inception-ResNet-v2 [16] outperforms others in both spatial and temporal components. The RGB model is initialized with pre-trained model from ImageNet and fine-tuned in the Kinetics dataset, while the flow model is initialized from the fine-tuned RGB model. Inspired by [19], the temporal segment network framework is used and three segments are sampled from each trimmed video for video-level training. During testing, we can densely extract features for each frames in the video.

2.2. Acoustic Feature

We use ConvNet-based audio classification system [6] to extract acoustic feature. The audio is divided into 960ms frames, and the frames are processed with Fourier transformation, histogram integration and logarithm transformation. The resulting frame can be seen as a 96×64 image that form the input of a VGG16 [14] image classification model. Similar with the visual feature, we trained the acoustic feature in the temporal segment network framework.

3. Off-shelf Temporal Modeling Approaches

In this section, we present a brief introduction of our proposed shifting attention network and temporal Xception network. More implementation details and analysis will be in a following technique report. We also refer [11] for the details of multi-stream sequence model and fast-forward sequence model.

3.1. Shifting Attention Network

Attention models have shown great potential in sequence modeling. For example, numerous pure attention architectures [18, 12] have been proposed and achieved promising results in natural language processing problems. In order to explore the capabilities of attention models in action recognition, a shifting attention network architecture is proposed, which is efficient, elegant and solely based on attention.

3.1.1 Shifting Attention

An attention function can be considered as mapping a set of input features to a single output, where the input and output are both matrices that concatenate feature vectors. The output of the shifting attention $\mathrm{SATT}(X)$ is calculated through a shifting operation based on a weighted sum of the features:

$$\mathrm{SATT}(X) = \frac{\lambda X \cdot a + b}{\|\lambda X \cdot a + b\|_2}, \tag{1}$$

where λ is a weight vector calculated as

$$\lambda = \mathrm{softmax}(\alpha \cdot w X^T), \tag{2}$$

w is learnable vector, a and b are learnable scalars, and α is a hyper-parameter to control the sharpness of the distribution. The shifting operation actually shifts the weighted sum and at the same time ensures scale-invariance. The shift operation efficiently enables different attention components to flexibly diverge from each other and have different distributions. This lays the foundation for Multi-SATT, which we describe next.

3.1.2 Multi-Group Shifting Attention Network

In order to collect multi modal information from videos, we extract a variety of different features, such as appearance

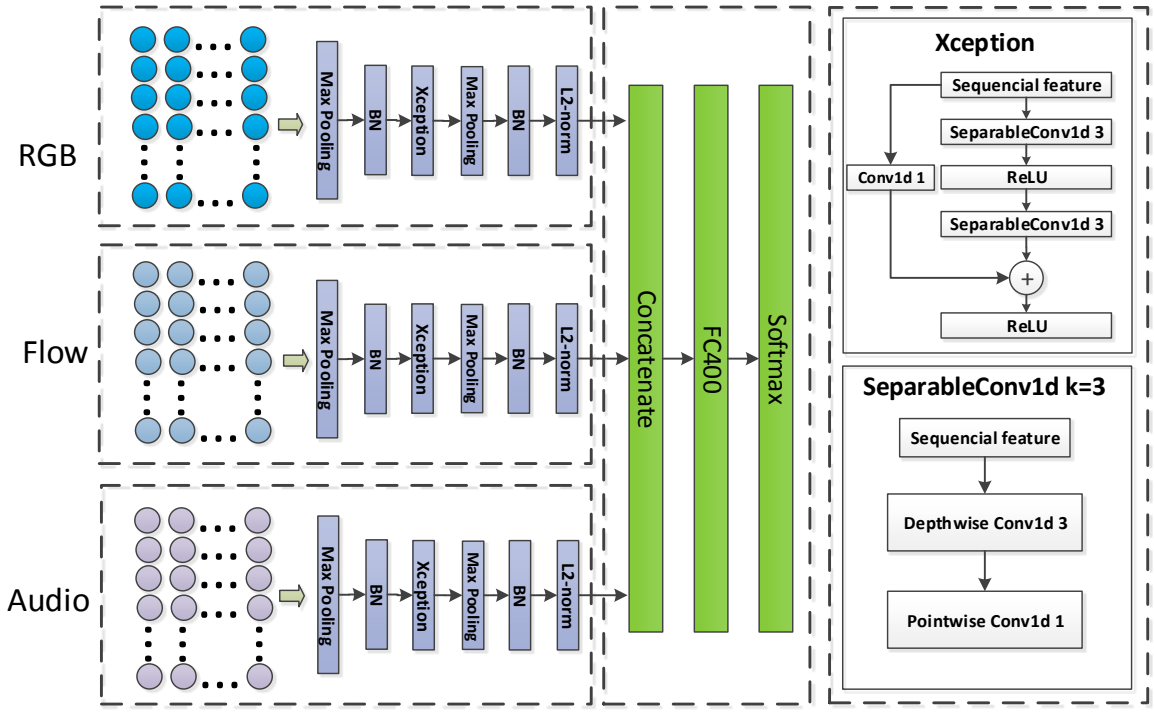

Figure 2. Temporal Xception Network.

(RGB), motion (flow) and audio signals. Although the attention model focuses on some specific features and effectively filters out irrelevant noise, it is unrealistic to merge all multi modal feature sets within one attention model, because features of different modality have different values, dimensions and scales. Instead, we propose Multi-Group Shifting Attention Networks for training multiple groups of attentions simultaneously. The architecture of the proposed Multi-SATT is illustrated in Figure 1.

First, we extract multiple feature sets from the video. For each feature set X_i, we apply N_i different shifting attentions, which we call one attention group, and then we concatenate the outputs. Next, the outputs of different attention groups are normalized separately and concatenated to form a global representation vector for the video. Finally, the representation vector is used for classification through a fully-connected layer.

3.2. Temporal Xception Network

Depthwise separable convolution architecture [2, 20] has shown its power in image classification by reducing the number of parameters and increasing classification accuracy simultaneously. Recently, convolutional sequence-to-sequence networks have been successfully applied to machine translation tasks [5, 8]. In this competition, we adopt the temporal Xception network for action recognition, which apply the depthwise separable convolution families to the temporal dimension and achieves promising performance. The proposed temporal Xception network architecture is shown in Figure 2. Zero-valued multi modal features were padded to make fixed length data for each stream. We applied adaptive temporal max pooling to obtain n segments for each video. We then feed the video segment features into a Temporal Convolutional block, which is consist of a stack of two separable convolutional layers followed by batch norm and activation with a shortcut connection. Finally, the outputs of three stream features are concatenated and fed into the fully-connected layer for classification.

4. Experiment Results

We conduct experiment on the challenging Kinetics dataset The dataset contains 246,535 training videos, 19,907 validation videos and 38,685 testing videos. Each video is in one of 400 categories.

Table 1 summarizes our results on the Kinetics validation dataset. From Table 1, we have three key observations. (1) Temporal modeling approaches with multi modal features are a more effective approach than naive combining the classification scores of different modality networks for

Model	Modality	Top-1 Accuracy (%)	Top-5 Accuracy (%)
Inception-ResNet-v2	RGB	73.0	90.9
Inception-ResNet-v2	Flow	54.5	75.9
VGG16	Audio	21.6	39.4
Late fusion	RGB + Flow + Audio	74.9	91.6
Multi-stream Sequence Model	RGB + Flow + Audio	77.0	93.2
Fast-forward LSTM	RGB + Flow + Audio	77.1	93.2
Temporal Xception Network	RGB + Flow + Audio	77.2	**93.4**
Shifting Attention Network	RGB + Flow + Audio	**77.7**	93.2
Ensemble	RGB + Flow + Audio	**81.5**	**95.6**

Table 1. Kinetics validation results.

the video classification. (2) The proposed Shifting Attention Network and Temporal Xception Network can achieve comparable or even better results than the traditional sequence models (e.g. LSTM), which indicates they might serve as alternative temporal modeling approaches in future. (3) Different temporal modeling approaches are complementary to each other.

5. Conclusions

In this work, we have proposed four temporal modeling approaches to address the challenging large-scale video recognition task. Experiment results verify that our approaches achieve significantly better results than the traditional temporal pooling approaches. The ensemble of our individual models has been shown to improve the performance further, enabling our method to rank first worldwide in the challenge competition. All the code and models will be released soon.

References

[1] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[2] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CVPR*, 2017.

[3] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015.

[4] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. CVPR, 2016.

[5] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017.

[6] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *arXiv preprint arXiv:1609.09430*, 2017.

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

[8] L. Kaiser, A. N. Gomez, and F. Chollet. Depthwise separable convolutions for neural machine translation. *arXiv preprint arXiv:1706.03059*, 2017.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[10] 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.

[11] 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:1707.04555*, 2017.

[12] Z. Lin, M. Feng, C. Nogueira dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A Structured Self-attentive Sentence Embedding. *ArXiv e-prints*, Mar. 2017.

[13] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

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

[15] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *ICML*, 2015.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *arXiv preprint arXiv:1602.07261*, 2016.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: Generic features for video analysis. *ICCV*, 2015.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *ArXiv e-prints*, June 2017.

[19] 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. In *ECCV*, 2016.

[20] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CVPR*, 2017.