Temporal Modeling Approaches for Large-scale Youtube-8M Video Understanding

Fu Li, Chuang Gan, Xiao Liu, Yunlong Bian, Xiang Long, Yandong Li, Zhichao Li, Jie Zhou, Shilei Wen

Baidu IDL & Tsinghua University

Abstract

This paper describes our solution for the video recognition task of the Google Cloud & YouTube-8M Video Understanding Challenge that ranked the 3rd place. Because the challenge provides pre-extracted visual and audio features instead of the raw videos, we mainly investigate various temporal modeling approaches to aggregate the frame-level features for multi-label video recognition. Our system contains three major components: two-stream sequence model, fast-forward sequence model and temporal residual neural networks. Experiment results on the challenging Youtube-8M dataset demonstrate that our proposed temporal modeling approaches can significantly improve existing temporal modeling approaches in the large-scale video recognition tasks. To be noted, our fast-forward LSTM with a depth of 7 layers achieves 82.75% in term of GAP@20 on the Kaggle Public test set.

1. Introduction

Video understanding is a challenging task which has received significant research attention in computer vision and machine learning. The ubiquitous video capture devices have created videos far surpassing what we can watch. Therefore, it has been a pressing need to develop automatic video understanding algorithms for various applications.

To recognize actions and events in videos, existing approaches based on deep convolutional neural networks (CNNs) [11, 15, 6, 20] and/or recurrent networks [9, 18, 10, 3] have achieved state-of-the-art results. However, due to the lack of publicly available datasets, existing video recognition approaches are restricted to small-scale data, while large-scale video understanding remains an under-addressed problem. To remedy this issue, Google releases a new web crawled large-scale video dataset, named as YouTube-8M, which contains over 7 million YouTube videos with a vocabulary of 4716 classes. A video may have multiple tag classes and the average number of tag classes per video is 1.8. Prior to this, Gan *et. al* [5, 7] also investigated to learn video recognition models using Web videos

and images.

Another appealing point of the Youtube-8M dataset is that this competition only provides the pre-extracted visual and audio features from every second of video instead of raw videos. We can neither train different CNNs architectures nor learn as optical flow features from the raw videos. Therefore, we focus on temporal modeling approaches to aggregate the frame-level features that yield robust and discriminative video representation for further multi-label recognition. Particularly, we propose three novel temporal modeling approaches, namely two-stream sequence model, fast-forward sequence model and temporal residual neural networks. Experiment results verity the effectiveness of the three models over the traditional temporal modeling approaches. We also find that these three 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 our temporal modeling approach to learn robust and discriminative video feature representation for recognition. Section 3 reports empirical results, followed by discussion and conclusion in Section 4.

2. Approach

In this section, we describe our three families of temporal approaches respectively.

2.1. Two-stream Sequence Models

Our two stream sequence models build upon the bidirectional LSTM [10] and GRU [3], since they have shown strong temporal modeling abilities for video recognition. The challenge here is how to incorporate the visual and audio information contained in the videos. In order to best take the advantage of multi-modal clues, we propose several sequence architectures to fuse these two modality features.

The original two-stream CNN [15] framework trains CNNs with RGB and optical flow features separately, and then relies on a late score fusion strategy to leverage the complementary nature of the two modalities. Recently, Ma *et. al* [14] has proposed a temporal segment RNN network
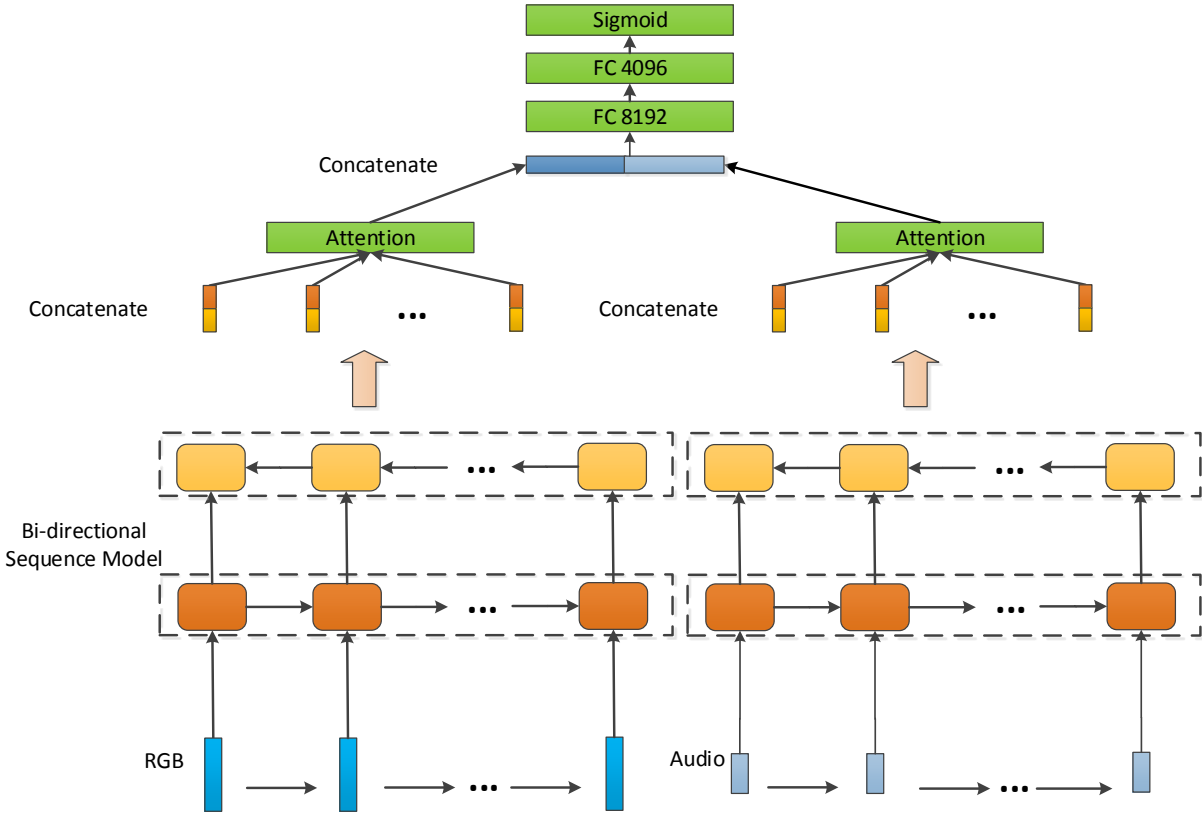
Figure 1. The architecture of our proposed two-stream LSTM model.

by first concatenating the two modality features and then fed them into one LSTM to achieve video recognition.

Different from them, we propose to train two bidirectional LSTMs or GRUs models (i.e. one for RGB features, and the other for audio features). Attention layers are inserted after the sequence models and attended feature vectors from two modalities are then concatenated. Finally, the concatenated feature vector is fed into two fully-connected layer and a sigmoid layer sequentially for multi-label classification. We outline the framework in Figure 1. Experiments results verity the effectiveness of the our proposed two-stream sequence model approaches over other alternative two-stream fusion approaches.

2.2. Fast-forward Sequence Models

Recently, we have witnessed the success of deep CNNs on large-scale image classification [16, 19, 8]. Typically, models with deeper convolution layers outperform shallow ones. However, this success has not been transferred to the sequence models that used in video recognition tasks. The best sequence models reported in literature are still shallow models. The phenomenon is caused by two reasons. First,

it is impossible to explore deeper sequence models in the pre-existing small-scale video recognition dataset [17, 12], which only contain around 10 thousands videos. Second, the optimization of deeper sequence model is much more challenging than training deeper CNNs because the existence of many more nonlinear activations and the recurrent computation results in smaller and instable gradient.

The new Youtube8M dataset sheds light on opportunities to explore sequence models with deep architectures. Since large-scale video recognition is a very difficult and challenging problem, we believe that deeper sequence models with more complex architecture is necessary for capturing the temporal relationship between frames. In the competition, we focus on enhancing the complexity of the sequence model by increasing the model depth. However, we observe that naively increasing the depth of the LSTM and GRU still entails to overfitting and optimization difficulties, and thus always have negative results for the video recognition. This phenomenon is consistent with the results reported by the original Youtube8M technique report [1].

To address these challenges, we explore a novel deep LSTM/GRU architecture by adding the fast-forward con-
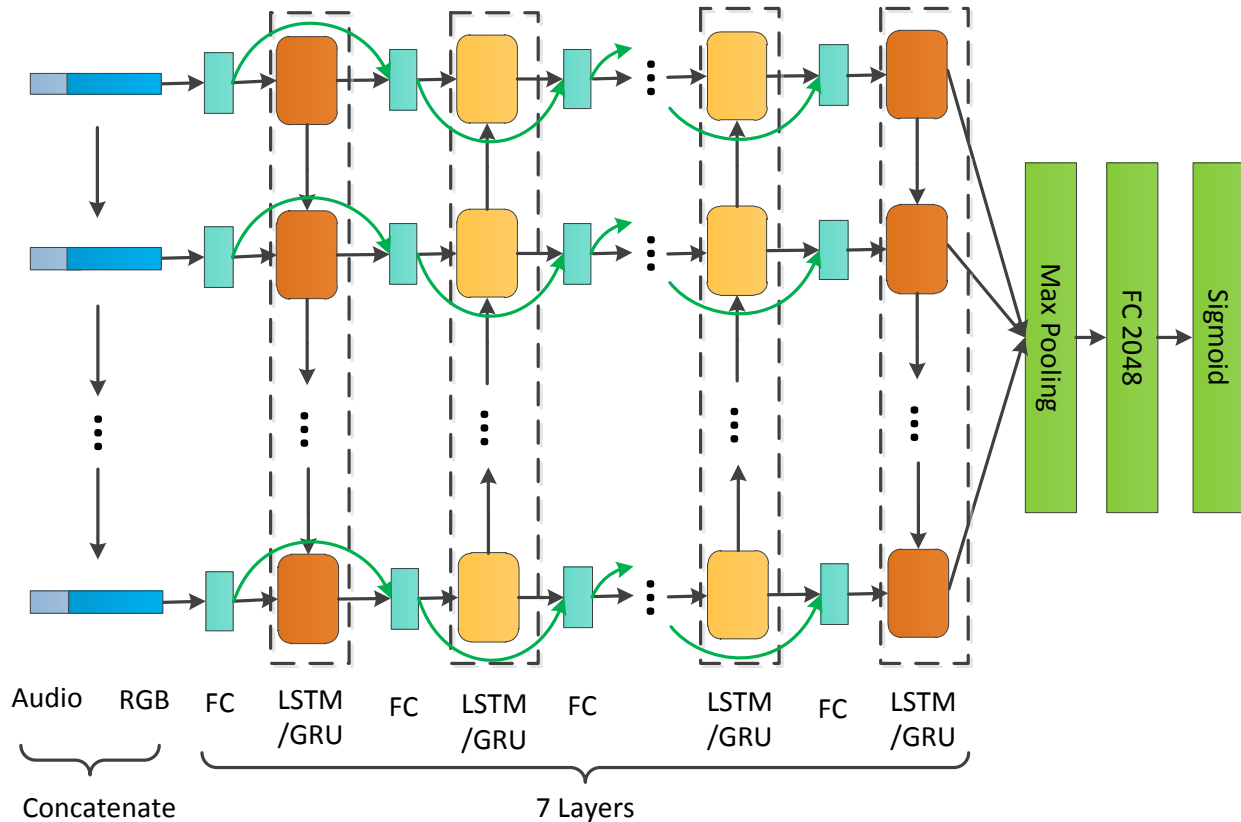
2

Figure 2. The architecture of our proposed fast-forward sequence models.

nections [22] to sequence models, which plays an essential role in building a sequence model with 7 bidirectional LSTMs. We outline the framework in the figure 2. We first concatenate the RGB and audio features of each frame together and then fed them into the fast-forward sequence model. The fast-forward connections are added between two feed-forward computation blocks of adjacent recurrent layers. Each fast-forward connection takes the outputs of previous fast-forward and recurrent layer as input, and use a fully-connected layer to embed them. The fast-forward connect provides a fast path for information to propagate, so we call the path fast-forward connections. We will introduce more detail of our proposed fast-forward sequence model and implementation details in a following technique report.

2.3. Temporal Residual Neural Networks

Although the power of recurrent models (LSTMs and GRUs) have been widely acknowledged, recent sequential convolution architectures [13, 14] show strong potentials for various temporal modeling tasks. Li *et. al* [13] proposed a temporal ResCNN based neural speaker recognition system for speaker identification and verification. Ma

et. al [14] proposed a temporal-inception architecture for video recognition, and achieved state-of-the-art results on UCF101 and HMDB51 datasets.

In the competition, we investigate the usage of temporal convolution neural networks for temporal modeling on video recognition. In contrast with [14] that performs convolutions on frame-level features to learn global video-level representations, we combine convolution and recurrent neural networks to take the advantages of both models. The temporal convolution neural networks are utilized to transform the original frame-level features into a more discriminative feature sequence, and LSTMs are used for final classification.

The architecture of the proposed Temporal CNN is illustrated in Figure 3. RGB and audio features in each frame are concatenated and zero-valued features are padded to make fixed length data. The size of the resulted input data is $4000 \times 1152 \times 300$, where 4000, 1152, and 300 indicates mini-batch size, channel number, and length of frames, repsectively. We then propagate the batch data into a Temporal Resnet, which is a stack of 9 Temporal Resnet Blocks (TRB), and each TRB consists of two temporal convolutional layers (followed by batch norm and activation) and a
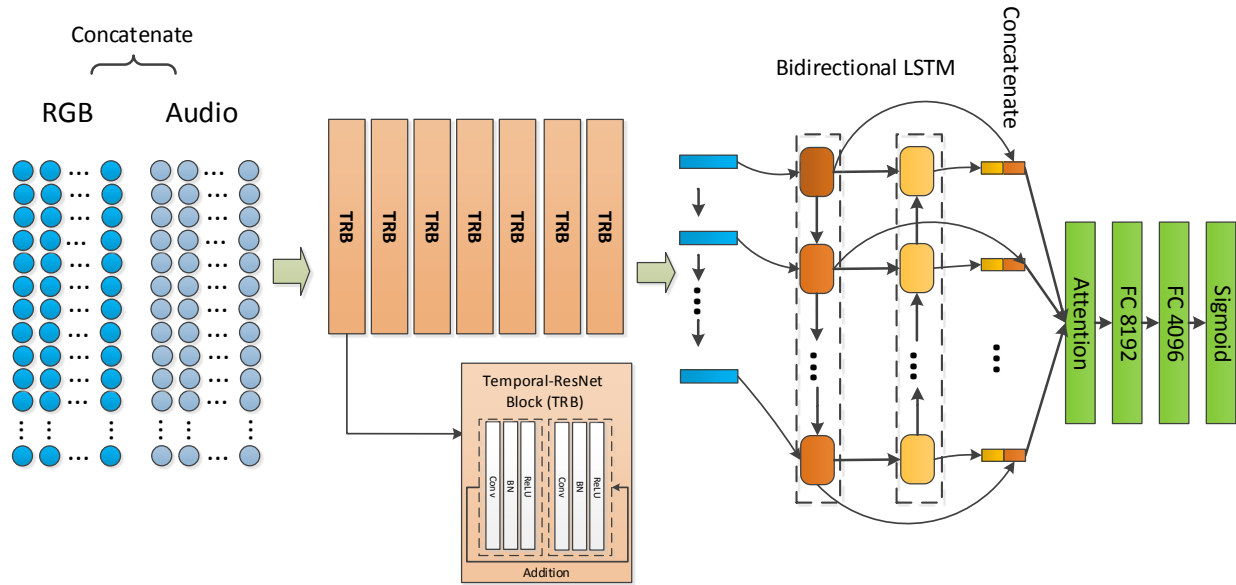
Figure 3. The architecture of our proposed temporal residual CNNs.

shortcut connection. We use 1024 3×1 filters for all the temporal convolution layers. The output of the temporal CNN is then fed into a bidirectional LSTM with attention.

3. Experiment

In this section, we present the dataset, experiment setting and our experimental results.

3.1. Dataset

We conduct experiment on the challenging Youtube-8M dataset [1]. This dataset contains around 7 million Youtube videos. Each video is annotated with one or multiple tags. In the competition, visual and audio features are pre-extracted and provided with the dataset for each second of the video. Visual features are obtained by the Google Inception CNN pre-trained on the ImageNet [4], followed by the PCA-compression into a 1024 dimensional vector. The audio features are extracted from a pre-trained VGG [16] network. In the official split, the dataset is divided into three parts: 70% for training, 20% for validation, and 10% for testing. In practice, we only maintain 60K videos from the official validation set to cross validate the parameters. Other videos in the validation set are included into the training set. We observe that this strategy can slightly improve the classification performances. Results are evaluated using the Global Average Precision (GAP) metric at top 20 as used in the Youtube-8M Kaggle competition.

Table 1. Comparison results on Youtube8M test set.

Method	GAP@20
Video-level	0.80824
VLAD	0.80423
Temporal CNN	0.80889
Two-stream LSTM	0.82172
Two-stream GRU	0.82366
Fast-forward LSTM	0.81885
Fast-forward GRU	0.81970
Fast-forward LSTM (depth7)	**0.82750**
Ensemble	**0.84542**

3.2. Experiment Results

Table 1 reports the performance of individual models on the Youtube8M test set. For the video-level approach, we use the average pooling to aggregate the frame-level feature vector. For VLAD encoding based approaches, we use 256 cluster centers followed by signed square root and L2 normalizations as suggested in [2, 21]. We then fed these representations into a MLP classifier to obtain the final video classification scores.

From Table 1, we have three key observations. (1) Our proposed two-stream sequence models and fast forward sequence models achieve significantly better results compared to previous video pooling approaches. (2) The fast-forward LSTM model with depth 7 can boost the shallow sequence

model around 0.5% in term of GAP. (3) Different temporal modeling approaches are complementary to each other. Our final submission ensembles 57 models with different hidden cells and depths.

4. Conclusions

In this work, we have proposed three 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 the third place out of 650 teams in the challenge competition. Our PaddlePaddle video toolbox is available for download from `https://github.com/baidu/Youtube-8M` and includes implementations of three temporal modeling approaches.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] R. Arandjelovic and A. Zisserman. All about vlad. In *CVPR*, pages 1578–1585, 2013.

[3] 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.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] C. Gan, C. Sun, L. Duan, and B. Gong. Weblysupervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*, pages 849–866, 2016.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

[13] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

[14] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *arXiv preprint arXiv:1703.10667*, 2017.

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

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[17] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

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

[21] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. *CVPR*, 2015.

[22] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu. Deep recurrent models with fast-forward connections for neural machine translation. *arXiv preprint arXiv:1606.04199*, 2016.