CNN+BERT FOR VIDEO RECOGNITION

Kritesh Garg
AI Labs,
Course5 Intelligence
Bengaluru, India
kritesh.garg@course5i.com

Karthikeyan Murugan
AI Labs,
Course5 Intelligence
Bengaluru, India
karthikeyan.murugan@course5i.com

January 25, 2021

ABSTRACT

Traditionally, for video-based Action Recognition task, 3D CNN are used to extract temporal information and Temporal Global Average (TGAP) layer to summarize this information. In this work, we replace the TGAP layer with the attention mechanism of BERT as it has been state-of-the-art for many sequence-based tasks. BERT's bidirectional attention mechanism gets a better representation of Temporal information with respect to TGAP. 3D CNN+BERT showed improvement over popular 3D CNN models such as R(2+1)D and ResNext for Action Recognition over our custom Action Recognition Dataset. We compare the performances of 3D CNN+TGAP and 3D CNN+BERT on our dataset, where 3D CNN+BERT resulted an accuracy improvement of up to 3%.

Keywords CNN · Transformer · BERT · Video Recognition

1 Introduction

Action Recognition (AR) refers to classify a video based on the activity/action performed in the video. AR can provide insights related to the activity/action in a Video. AR is essential in many domains such as Video Retrieval, Surveillance, Robotics.

A video contains multiple scene/action sequences known as clips. These clips can be extracted from the video based on the use-case using scene change, rule-based extraction or manual extraction. Each clip contains critical information in the spatial domain as well as in temporal domain. Spatial domain contains information related to the spatial entities present in the clip, such as Objects, Context, etc. whereas, the information of the interaction between these entities known as action/activity is present in Temporal domain.

Temporal information is extracted either using 3D convolutions with TGAP layer, which extracts spatio-temporal information using multiple 3D convolutional and then this information is integrated using TGAP layer or by using 2D convolution layers to extract spatial information, which is then passed to a recurrent architecture, such as LSTM or RNN.

The spatio-temporal features extracted using 3D CNN can be considered as features from different temporal regions. Traditional temporal global average pooling (TGAP) layer averages these features leading to loss of important temporal information.

In this work, we are going to replace TGAP with BERT's attention mechanism. 3D convolution blocks are used to extract spatio-temporal information, which is processed by BERT using transformer-based encoder, which encodes the temporal information in to a feature vector. Single feed forward layer uses this feature vector to classify the video clip in to one of the action/activity class.

BERT's attention mechanism apply weights to these temporal feature as per their importance which leads to better representation of the temporal information.

2 Related Work

Action Recognition is performed in the following ways:

2.1 Temporal feature summarization using Pooling, Fusion and Recurrent Networks

Pooling is performed to summarize Temporal features using concatenation, averaging, maximum, minimum, ROI, feature aggregation techniques and time-domain convolutions [1], [2].

Fusion is used to fuse different modalities or stream of networks. Late fusion, Early fusion and Slow fusion are used to integrate temporal information along the channel dimension at various points in CNN architectures. Fusion of RGB and optical flow with extra 3D convolution layer inserted towards the end of the architecture is used in the two-stream fusion architecture in [3] to generate spatio-temporal relationship.

Recurrent networks such as LSTMs are utilized for extracting Temporal (Sequential) information on the spatial features extracted using 2D CNN from the frames of a video [1], [4]. E.g., VideoLSTM [5] integrates temporal information by using convolutional LSTM with spatial attention.

Using above feature summarization techniques can lead to loss temporal information.

2.2 3D CNN Architectures

3D CNN Architectures utilize 3D convolution layers which extract spatial as well as temporal information from sequence of images (clip). These layers use the sequence of images as 3rd dimension (temporal dimension). Throughout the 3D CNN network, the temporal information is processed hierarchically. Before 3D CNN network, temporal information was extracted at a later stage of the network resulting in loss of temporal information. 3D CNN displayed improvement in accuracy with a huge increase in computation cost and memory demands with respect to their counterpart 2D CNN.

Traditionally for AR, first 3D CNN was used in the C3D model [6]. Inception 3D model (I3D) [7] which uses a deeper 3D CNN architecture then C3D. The ResNet version of 3D convolution is introduced in [8]. Then, R(2+1)D [9] and S3D [10] architectures came up with a new convolutional block with separate 3D CNN

layers working on temporal as well as spatial information proved effective than the regular stack of 3D convolution layers. Another important 3D CNN architecture is Channel-Separated Convolutional Networks (CSN) [11] which separates the channel interactions and spatio-temporal interactions which can be thought of as the 3D CNN version of depth-wise separable convolution [12].

Although 3D CNNs are powerful, they still lack an effective temporal fusion strategy at the end of the architecture to integrate the regional temporal information.

3 Methodology

In this section, we will discuss about the 3DCNN + BERT based action recognition method we chose to implement using C5AR dataset. Bidirectional Encoder Representations for Transformers (BERT) [13] is most prevalent and successful approaches on most NLP tasks. Unlike simple transformers or RNNs or other self-attention mechanism, BERT being bi-directional it can comprehend information from both directions. The chosen approach uses BERT to get benefit of bidirectional temporal information in a video feature i.e., contextual information of both past as well future frames for better action recognition. (See Figure 1)

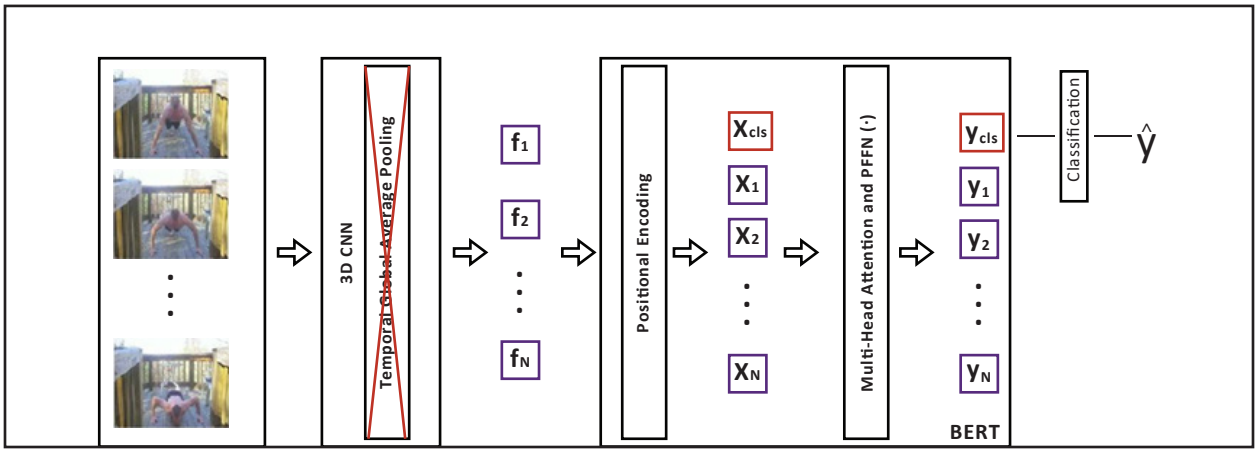

Figure 1: BERT-based late temporal Modeling. Adapted from [14] pg.4

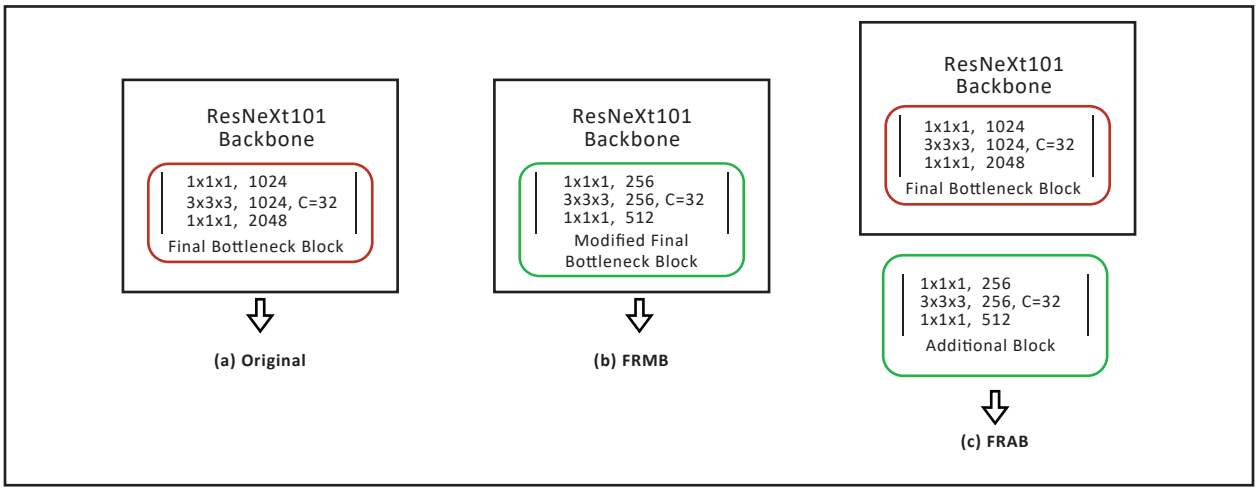

Figure 2: Explaining FRAB and FRMB implementation on ResNext101 backbone. Adapted from [14] pg.6

First, Video frames pass through 3DCNN to get features extracted and BERT-based temporal pooling is fused after 3DCNN feature extractor replacing temporal global average pooling. Also, the chosen method added a learned positional encoder to the extracted features for preserving order information and multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. TGAP just uses pooling layer but the chosen BERT based method uses learnable classification tokens.

The chosen approach has introduced two feature reduction blocks namely Feature Reduction with Modified Block (FRMB) and Feature Reduction with Additional Block (FRAB). In the backbone 3DCNN network, final block is replaced with a feature reduction block in FRMB and new feature reduction block is appended in the case of FRAB. FRMB has better computational complexity and parameter efficiency than FRAB as well as backbone 3DCNN network, but it has a drawback when using pretrained weight. If feature reduction block is in fine-tuning step, FRMB final block doesn't benefit from pretrained weights of standard datasets. (See Figure 2)

Backbone Architecture	Pooling Type	Feature Modification Block (FRMB/FRAB)	Top@1 Accuracy
ResNext101	TGAP	NA	79.58
ResNext101	BERT	FRMB	82.12
ResNext101	BERT	FRAB	82.56
R(2+1)D	TGAP	NA	79.93
R(2+1)D	BERT	NA	82.86

4 Experiments

In this section, we will discuss about the C5AR dataset, implementation details and ablation study based on ResNext and R(2+1)D architectures with BERT and with TGAP layer.

4.1 Dataset

We have used custom Course5 Action Recognition dataset which contains 13 classes with 2000 clips which we have manually curated for our use-case. Rule-based approach is used to extract the clips from these videos. These clips are then manually processed and labelled.

4.2 Experimental Setup

Firstly, we have extracted frames from each clip. An augmentation technique that is circulating over the frames is used for the clips where the frames are less than the desired number of frames. AdamW is selected as the optimizer with learning rate set to 10−4. Learning Rate scheduler that reduces learning rate by a factor of 10 if loss does not decrease for 5 epochs is used. Data normalization schemes are followed as per the pre-trained networks in order to best utilize the pre-training weights.

BERT is configured with 8 attention heads and a single transformer block. The dropout ratio was set to 0.8 and the attention masking was set to 0.2. Classification token(xcls) and the learned positional embedding are initialized using with zero mean normal weight with a standard deviation of 0.02. We have used Tesla V100 32*2 for all out experiments

4.3 Ablation Study

In this section, we will analyze two backbone architectures (ResNext and R(2+1)D) in our experiments step wise and see how BERT as pooling strategy behaves against the Temporal Global Average Pooling (TGAP). For the first experiment, we have selected ResNext101 architecture with 112*112 frame size and a fixed sequence length of 64 frames per clip. For the second experiment, we have selected R(2+1)D architecture with 112*112 frame size and a fixed sequence length of 32 frames per clip. In Table 1, used 3D CNN network, used pooling strategy, used feature reduction block and top 1 accuracy achieved are presented as columns of experiments

As we can see in Table 1. For ResNext based experiments FRAB provides an approx. 0.4% improvement over FRMB feature block. We can clearly note the effectiveness of BERT as a pooling strategy over the conventional TGAP pooling. BERT resulted in an approx. 3% improvement with R(2+1)D 3D CNN architecture over TGAP.

5 Discussion and Future work

In this work, we have seen the benefit of combining 3D CNN and BERT for AR. There are many Temporal feature summarization strategies, such as Pooling, Fusion and Recurrent Networks. To utilize these Temporal features, the Attention mechanism of BERT has shown an accuracy improvement with the most effective 3D CNN architectures, such as ResNext and R(2+1)D over custom Video Recognition Dataset.

This study has created the path for better pooling strategies on 3D CNN architecture over BERT. A possible research direction might be parameter efficient BERT that do not need feature reduction blocks such as FRMB or FRAB as they deteriorate the quality of the features by reducing their dimension. This proposed method also has the capability to improve similar tasks with AR, such as Temporal and Spatial action localization and Video Captioning.

References

[1] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3174, 2017.

[2] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702, 2015.

[3] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 1933–1941, 2016.

[4] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, 2017.

[5] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees G.M. Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding,* 166:41 – 50, 2018.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[7] 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)*, pages 4724–4733, 2017.

[8] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[9] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pages 6450–6459, 2018.

[10] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, 2018.

[11] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolu- tional networks, 2019.

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[14] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition, 2020.