**Implementation of Squeeze-Extension And Self-Attention Modules For 3D Action Recognition**

**Fazliddin Anvarov, Dae Ha Kim, and Byung Cheol Song\***

**Inha University**

**email**: fazliddinanvarov19950611@gmail.com, kdhht5022@gmail.com bcsong@inha.ac.kr\*

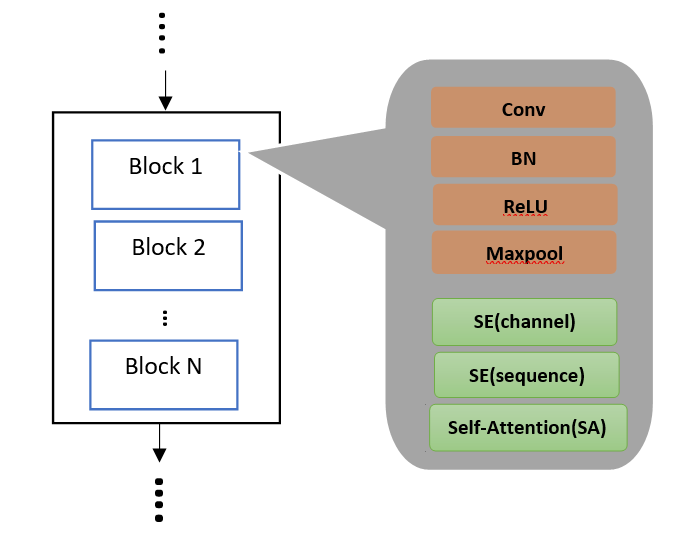
**Abstract**

Human action recognition is the task of detecting and defining the behavior and objectives of individuals from a sequence of human’s actions.Recently, there has been a significant improvement by using the 3D convolutional neural network (CNN) in action recognition. However, to date, only very naïve 3D CNN architectures have been used. In this paper, we propose an upgrade version of 3D CNN by using squeeze-and-excitation (SE) and self-attention (SA) modules. The suggested models perform better than prior work, enhancing the SOTA result for ResNext architecture by 1.83%(top-1) and 0.25%(top-5) for UCF-101 dataset, 3.03%(top-1) and 0.86%(top-5) for HMDB-51 dataset.

**1.Introduction**

Human action recognition is an essential and well-studied issue in computer vision. Traditional action recognition methods were based on object detection, articulated pose, thick trajectories and part-based /structured models [2][3]. We solve the action recognition task based on 2D CNN architecture. Modality such as optical flow and RBG video input frame data are used in this process [4][5].

Video-based action recognition has significantly benefited from advances in image-based models. With the exception of a few 3D-conv-based methods, most approaches, including the current state-of-the-art , use a variant of discriminatively trained 2D CNN over the appearance(frames) and in some cases movement(optical flow) modalities of the input video. Noticeable progress has been produced with CNNs for image-based tasks such as classification, segmentation and object detection. Their impact on issues in the video domain was not significant to the absence of big video datasets. Recently, K. Hara *et al*. [1] proposed to use a larger dataset to train 3D CNN architectures and have achieved a magnificent result. As long as their purpose was only training using bigger dataset, they used fundamental structures of 3D CNN, which are a rather simple approach to learning 3D information. In this paper, we propose squeeze-and-excitation(SE) module to improve the quality of images generated by a network by explicitly modeling the interdependencies of its convolutional characteristics between the channels. To this end, we suggest a mechanism that enables the network to recalibrate characteristics, through which it can learn to use global data to selectively distinguish informative characteristics and eliminate less helpful characteristics. Besides the SE module, we also suggest the self-attention (SA) module. The SA module calculates the response at all positions as a weighted sum of the features. Moreover, the SA is complementary to convolution and helps to model long-range, multi-level dependencies across an image area. Our main contribution is to apply SE and SA modules to 3D-based CNN architecture and to enhance action recognition performance only using RGB information. Overall network structure illustrated in Figure 1.

Figure 1. The overall framework of our network. SE and SA followed after each layer. Total we have 4 layers.

**2.Approach**

**2.1 Squeeze-and-Excitation (SE) module**

The reason for applying squeeze-and-excitation module is that we expect to improve the learning of convolutional features by explicitly designing inter-connections of channels so that the network can enhance its sensitivity to informative attributes that can be utilized through subsequent transformations. Consequently, we would like to give access to global data and readjust filter responses in two steps, *i.e*., squeeze and excitation, before they are transformed into the next transformation. The basic SE module can be found in [6].

In the first step, we suggest squeezing global data into a descriptor channel. This task is accomplished through the use of global average pooling. To use the aggregated data in the squeeze operation, we follow it with a second operation aimed at completely capturing channel-specific dependencies.

Unlike SE for 2D CNN [6], we regard not only channel but also number of frames(sequence). After each layer in our network, we consider the concatenation of channels and sequences that make our study even efficient and make it possible to achieve even higher performance. Our module is demonstrated in Figure 2.

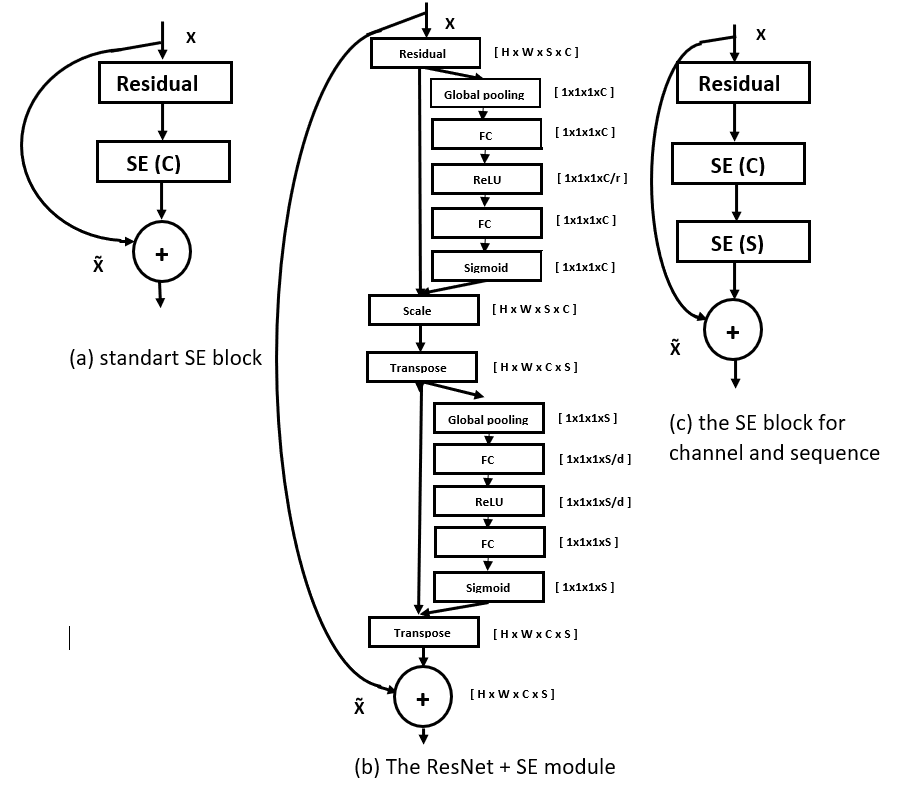
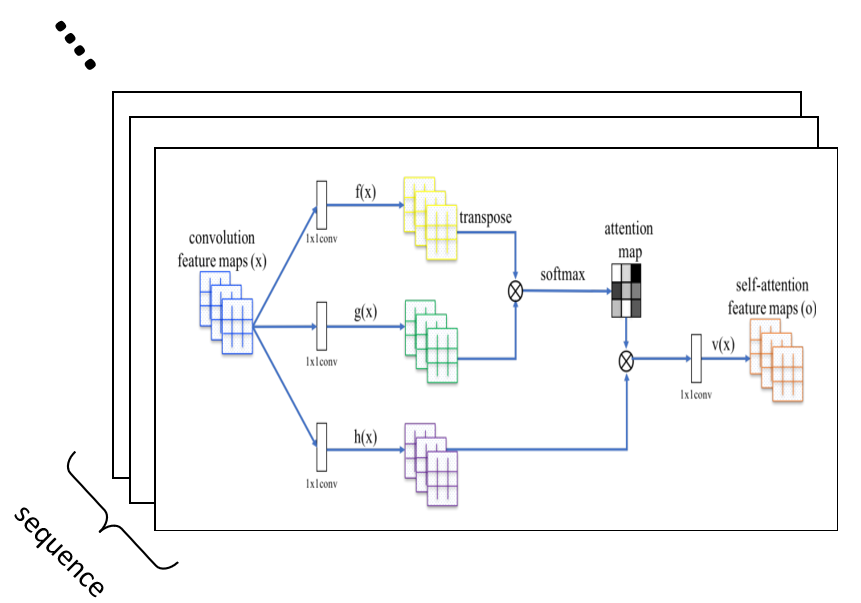


Figure 2. Module of squeeze-and-excitation. We expand SE into channel and sequence concatenation. Reduction value r=16, d =2.

**2.2 Self-Attention (SA) module**

Self-Attention, also known as intra-attention, is a mechanism of attention related to different positions of a single sequence to calculate representation of the same sequence. The primary concept of applying self-attention is helping convolutions to model long-range, full-level interconnections throughout the area of the image. The network armed with self-attention can draw pictures fine details are closely coordinated in remote parts of the image at each place with tiny details.

Some action recognition dataset involves interaction between human and item. Since we consider only a restricted amount of frames (16), not whole frames present this interaction precisely because the base network focuses only on the action portion. Moreover, some multi-action videos are available. SA module helps to properly define each region of action in the image. Because we consider 3D CNN with multiple frames, the concept of using SA is to capture the entire picture to create a better prediction based on the action part of the image and to use this understanding to logically connect all frames [7]. Self-attention module shown in Figure 3.



Self-attention block

Figure 3. The proposed module of self-attention. The ⊗ denotes matrix multiplication. The softmax operation is performed on each row. Self-attention module is performed for each frame separately.

You can check additional properties oftheabove diagram in Self-Attention Generative Adversarial Networks [7].

**3. Experiments**

**3.1 Dataset and Training configurations**

We conduct experiments on the *UCF-101* and *HMDB-51* datasets to assess the influence of SE and SA approaches.

UCF-101 dataset comprises 13,320 images of action from 101 classes of human action. To remove non-action frames, the videos were temporarily cut. Each video’s average duration is about 7 seconds. The dataset includes three train/test splits(70% training and 30% testing). Most videos extracted from YouTube.

HMDB-51 dataset contains 6,766 videos from 51 human action classes. The videos were temporarily trimmed, similar to UCF-101. Each video’s average duration is about 3 seconds. This dataset provides three train/test splits( 70% training and 30% testing). Videos are obtained from movies.

The videos include a dynamic background and camera movement. The main difference between them is the number of action classes and instances. We resized the images to 240-pixel heights without altering

their aspect ratios and stored them afterward.

We use the stochastic gradient with momentum to train the network and randomly generate training samples in training data from videos to accomplish data augmentation. We use cross-entropy losses in our practice and backpropagate their gradients. The training parameters include a weight decay of 0.001 and 0.9 for momentum. For performing fine tuning, we start learning rate from 0.001 and weight decay of 1e-5. Models are trained for 200 epochs with 16 batch size and 32 resnext cardinality.

**3.2 Implementation details and results**

To evaluate the proposed methods, we performed experiments on the two different architectures: ResNext-101 and ResNet-152. Both networks contain 4 layers. Compare to ResNet-152, ResNext (using the setting of 32x4d) add extra computational construction(cardinality) block to the base network [1].

We extended the concept of SE module for image-based models into video-based models, taking into account the number of frames. Conducting SE module not only for channels but also sequence helped to boost the performance even greater. The default reduction is 16. The number of sequence frames after 1st, 2nd, 3rd, and 4th layers are 8, 4, 2, 1 respectively.

Likewise to the SE approach, the SA module expended from 2D CNN into 3D CNN too considering the number of frames rather than the single image as shown in Figure 3. Results for UCF-101 and HMDB-51 dataset are shown in Table 1 and Table 2 respectively.

Table 1: Accuracies on the UCF-101 test set. The average is averaged accuracy over Top-1 and Top-5.

Method Top-1 Top-5 Average

ResNext-101 87.13 95.48 91.31

ResNext-101 + SE 88.49 95.69 92.09

ResNext-101+SE+SA **88.96 95.73 95.35**

ResNet-152 84.54 95.07 89.81

ResNet-152 + SE 85.67 95.25 90.46

ResNet-152+SE+SA **86.76 95.30 91.03**

**3.3 Discussion**

From the acquired results we can conclude that the SE module performs better than the SA module. For each case, SE increased the average performance by 0.925% compared to 0.343% for SA. One explanation for this can be that we regarded both channel and sequence in the squeeze-and-excitation, however in the scenario of self-attention we concentrated only on channel itself, because concatenation of sequence and channel did not yield expected result.

Table 2: Accuracies on the HMDB-51 test set.

Method Top-1 Top-5 Average

ResNext-101 62.00 87.86 74.93

ResNext-101 + SE 64.77 88.63 76.70

ResNext-101+SE+SA **65.03 88.72 76.88**

ResNet-152 58.34 86.97 72.66

ResNet-152 + SE 58.69 87.62 73.16

ResNet-152+SE+SA **58.84 88.19 73.52**

**4.Conclusion**

In this study, we proposed the SE module incorporating with fundamental 3D CNN architectures, which is an architectural unit intended to enhance a network’s representational power by allowing it to conduct dynamic channel-wise feature recalibration. In addition to the SE module, we suggested self-attention mechanism too, which is used to model long-range dependencies. Conducted experiments demonstrate the effectiveness of our SE and SA modules, resulting in better performance than state-of-the-art.

**References**

[1] Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *CVPR 2018*

[2] Tanvi S. Motwani and Raymond J. Mooney. 2012. Improving video activity recognition using object recognition and text mining. In Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)

[3] Luvizon, D.C., Picard, D., Tabia, H.: 2D/3D pose estimation and action recognition using multitask deep learning. In: CVPR. (2018)

[4] Laura Sevilla-Lara, Yiyi Liao, Fatma Guney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. arXiv preprint arXiv:1712.08416, 2017.

[5] Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact generalized non-local network. arXiv preprint arXiv:1810.13125, 2018

[6] Squeeze-and-Excitation Networks IEEE Conference on Computer Vision and Pattern Recognition 2018

[7] Self-Attention Generative Adversarial Networks arXiv 2018