SPATIO-TEMPORAL SLOWFAST SELF-ATTENTION NETWORK FOR ACTION RECOGNITION

*Myeongjun Kim** *Taehun Kim** *Daijin Kim**

* Department of Computer Science and Engineering, POSTECH, Pohang, Korea

ABSTRACT

We propose Spatio-Temporal SlowFast Self-Attention network for action recognition. Conventional Convolutional Neural Networks have the advantage of capturing the local area of the data. However, to understand a human action, it is appropriate to consider both human and the overall context of given scene. Therefore, we repurpose a self-attention mechanism from Self-Attention GAN (SAGAN) to our model for retrieving global semantic context when making action recognition. Using the self-attention mechanism, we propose a module that can extract four features in video information: spatial information, temporal information, slow action information, and fast action information. We train and test our network on the Atomic Visual Actions (AVA) dataset and show significant frame-AP improvements on 28 categories.

Index Terms— Action Recognition, Self-Attention Mechanism, Atomic Visual Actions

1. INTRODUCTION

Deep Convolutional Neural Networks (CNN) [1] have achieved great performances in image classification [2], object detection [3], and semantic segmentation [4]. In addition, the video classification and action recognition [5] as well as image task make significant progress. However, action recognition is not easy to solve using each local features. Because, human action has a characteristic that are related to other people or objects. Therefore we should be able to consider not only local features but also global features. As it passes through CNN layers, the receptive fields are gradually expanded. And they can capture a larger areas of context. but it's not enough to maintain the context of the long-term dependency. Human action is divided into three categories: person movement, object manipulation and person interaction. When we observed the characteristics of the human action, we have to contemplate objects or human interactions to solve the action recognition. We solved using a self-attention mechanism that allows us to detect long-range interactions and focus on important regions.

We propose a novel the Spatio-Temporal SlowFast Self-Attention Network. The proposed module induced training to focus on the outstanding regions of spatial and temporal information in video features. Spatial attention mechanism more focused on objects or people that are related to each other, and temporal attention mechanism more focused on when the action occurs in the video clip. Also, human action includes long-acting behavior and short-acting behavior. For example, jogging can be long. But boxing's fist is a short act. Therefore, we considered features of the behavior occurring for a short time and for a long time as different features. Finally, we propose a network that can consider all four features that are thought necessary for action recognition.

2. RELATED WORK

2.1. Action Recognition

Research to analyze and localize human behavior in video data has recently been accelerated. The most commonly used datasets are Kinetics [6], UCF-101 [7]. A dataset consists of a person movement, human-to-human interaction, and human-object interaction. As new data come out, understanding the relationships between people and the association between people and objects has become a critical factor in action recognition [8], and it is also important to be aware of the situation appropriately. There were several approaches for action recognition. They found human joints information through human pose estimation, and there was a network for judging human action by capturing how each joint moves with temporal axis [9]. The other network uses more abundant information by fusion of video and optical flow features [10]. However, the recent trend is solving a action recognition using only video clips.

2.2. Self-Attention

The Attention module [11] is designed to focus on meaningful regions where networks are thought to be a significant to target. The module can guarantee a long-term dependency that remain a challenge on CNN.

The Attention module also have a good effect on the image field. Self-attention module was introduced in Self-Attention GAN (SAGAN) [12] and improved results in image generation. We applied a self-attention module to the video

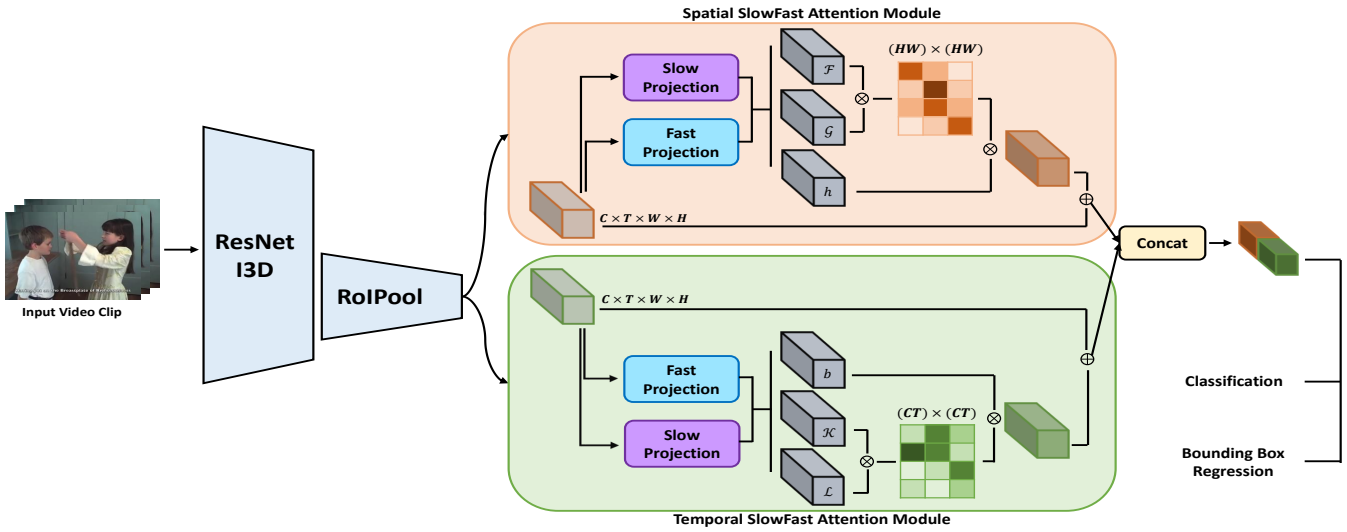

Fig. 1. An Overall Architecture of The Spatio-Temporal SlowFast Self-Attention Network

understanding task to ensure long-range interactions and focus on important features.

3. METHOD

In this section, we introduce the overall design of our network. The proposed network detect people in a given video clip and predicts what each person is doing. Our network is based on Faster-RCNN [3] object detection algorithm. As a video feature extraction network, the Kinetics-400 pretrained ResNet-I3D [8] network was used. The region proposal finds the person's bounding box. After performing the RoIPool [13] using the bounding box information obtained from the region proposal, the feature pass through the Spatio-Temporal SlowFast Self-Attention Module for classification of (action classes + 1) and bounding box regression.

3.1. Spatial Attention Module

The Spatial Attention Module can focus not only on RoIPool features, but also on other contextual information such as hands and faces to determine human action on features. The Spatial Attention Module reconstructs the self-attention module used in the image for video understanding. The existing self-attention module was used to determine the relationship between pixels in an image. But, the module can find spatially important parts of the entire video feature.

The video features have the shape of $C \times T \times H \times W$. The feature are transformed $C \times T$ and $H \times W$. The transformed video features $x \in \mathbb{R}^{(C\times T)\times(H\times W)}$ are projected into two new feature spaces $\mathcal{F}, \mathcal{G}$ to calculate the attention map,

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{H\times W} \exp(s_{ij})}, \text{ where } s_{ij} = \mathcal{F}(x_i)^T \mathcal{G}(x_j), \quad (1)$$

where $\mathcal{F}(x) = W_{\mathcal{F}} x$, $\mathcal{G}(x) = W_{\mathcal{G}} x$ and $\beta_{j,i}$ represents the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region. We define $C \times T$ is the number of feature channels and temporal spaces. Also, $H \times W$ is the number of spatial feature maps. The output of the attention layer is $o = (o_1, o_2, ..., o_j, ..., o_{H\times W}) \in \mathbb{R}^{(C\times T)\times(H\times W)}$, where,

$$o_j = \left(\sum_{i=1}^{H\times W} \beta_{j,i} h(x_i) \right), \; h(x_i) = W_h x_i. \quad (2)$$

In this formulation, $W_{\mathcal{F}}, W_{\mathcal{G}}, W_h$ are the learned weight parameters, which are implemented as $1 \times 1 \times 1$ convolutions. In addition, we further multiply the output of the attention layer by a scale parameter and add initial input feature map,

$$sa_i = \gamma o_i + x_i. \quad (3)$$

3.2. Temporal Attention Module

The Temporal Attention Module concentrates on the important areas of the temporal axis. Our network takes 32 frames of one video clip as input. All actions have a different length of time to represent each action. We assume that it is better to separate slow and fast action features because the amount of feature information is different. The Temporal Attention Module extracts the features needed when looking at the temporal axis to find human actions.

The transformed video features $x \in \mathbb{R}^{(C\times T)\times(H\times W)}$ are projected into two new feature spaces $\mathcal{K}, \mathcal{L}$ to calculate the attention map,

$$\alpha_{j,i} = \frac{\exp(t_{ij})}{\sum_{i=1}^{C\times T} \exp(t_{ij})}, \text{ where } t_{ij} = \mathcal{K}(x_i)^T \mathcal{L}(x_j), \quad (4)$$

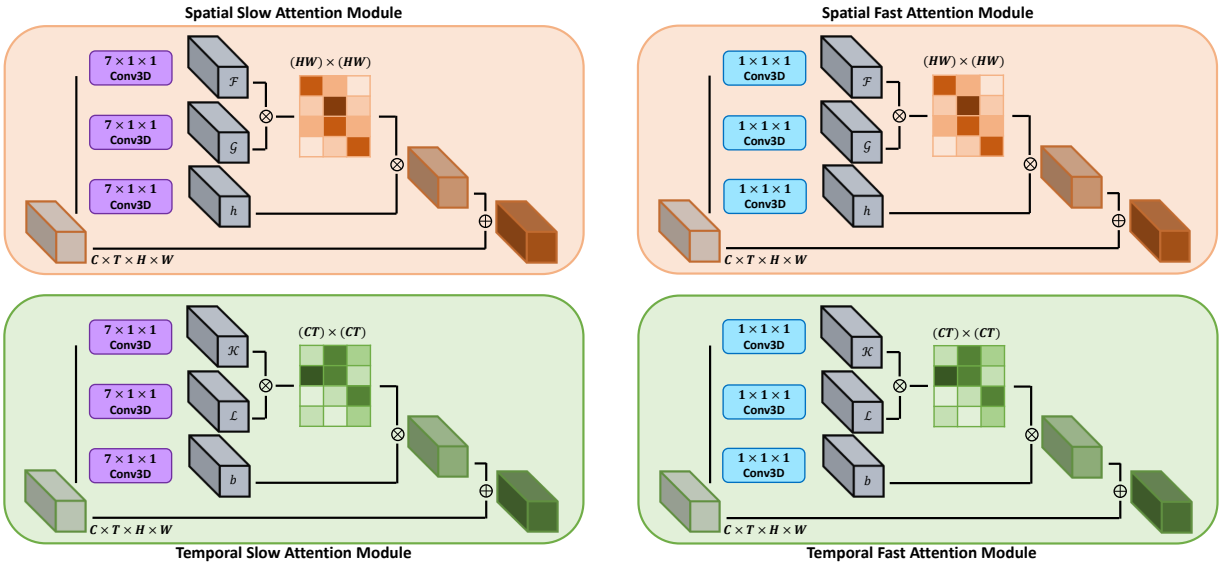

Fig. 2. The details of the Spatial SlowFast Attention Module and the Temporal SlowFast Attention Module

where $\mathcal{K}(x) = W_{\mathcal{K}}x$, $\mathcal{L}(x) = W_{\mathcal{L}}x$ and $\alpha_{j,i}$ represents the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region. We define $C \times T$ is the number of feature channels and temporal spaces. Also, $H \times W$ is the number of spatial feature maps. The output of the attention layer is $m = (m_1, m_2, ..., m_j, ..., m_{C\times T}) \in \mathbb{R}^{(C\times T)\times(H\times W)}$, where,

$$m_j = \left(\sum_{i=1}^{C\times T} \alpha_{j,i} b(x_i) \right), \; b(x_i) = W_b x_i. \tag{5}$$

In this formulation, $W_{\mathcal{K}}, W_{\mathcal{L}}, W_b$ are the learned weight parameters, which are implemented as $1 \times 1 \times 1$ convolutions. In addition, we further multiply the output of the attention layer by a scale parameter and add initial input feature map,

$$st_i = \gamma m_i + x_i. \tag{6}$$

3.3. SlowFast Attention Module

Human action has two characteristics: long-acting actions and short-acting actions. Most action recognition network considered slow action and fast action as a feature. However, short-acting behaviors will be an important region of every moment, and long-acting behaviors may be unnecessary features at the front and back. Therefore, we divided slow action feature and fast action feature separately.

To extract two features that distinguish slow and fast actions, we changed the kernel size of the convolution operation used in the Spatial Attention Module and the Temporal Attention Module. Therefore, the large size kernel is given for the slow action, and the small size kernel is given for the fast action.

4. EXPERIMENTS

4.1. Dataset

The AVA dataset [14] consists of 80 action classes, and each class is largely divided into three categories: individual behavior, behaviors related to people, and behaviors related to people. There are a total of 430 videos, training 235, validation 65, and test 131 videos. Each video is a 15 minute long video clip with one annotation per second. As in the previous evaluation, we evaluated 60 classes and used at least 25 instances for validation. Frame level average precision (frame-AP) [14] was used as the evaluation metric. The frame-AP reports the average precision (AP) using an IoU threshold of 0.5 in center frame of video clip.

4.2. Results

We compare the Spatio-Temporal SlowFast Self-Attention Module with the state-of-the-art approaches (Table. 1). When solving the first action recognition problem, both RGB image and optical flow feature were used. In contrast, since algorithms for extracting more abundant features such as Graph Convolutional Network (GCN) [19] and Attention Mechanism have emerged, only RGB images have been used to solve the action recognition. Existing networks require large amounts of video clips and use high image resolutions. However, we used a small frames and low resolution to obtain meaningful results.

Fig. 3 shows the comparison of frame-AP for each category according to the use of the Spatio-Temporal SlowFast Self-Attention Module. When using our attention module, there are performance improvements on 44 categories and over 1.0 frame-AP on 28 categories. We can see perfor-

Table 1. Performance and overall methods comparison with state of the art on AVA.

Model	Modalities	Input Size	Architecture	Frame mAP
SingleFrame [14]	RGB (1f), Flow (5)	320×400	R-50, FRCNN	13.7
AVA Baseline [14]	RGB (40f), Flow (40)	320×400	I3D, FRCNN, R-50	15.6
ARCN [15]	RGB, Flow	-	S3D-G, RN	17.4
STEP [16]	RGB (12f)	400×400	I3D, STEP	18.6
A Structued Model For Action Detection [17]	RGB (36f)	256×256	I3D, GCN	22.2
Action Transformer [18]	RGB (96f)	400×400	Tx, I3D Head	25.0
Ours	RGB (32f)	256×256	I3D, SSFA, TSFA	23.0

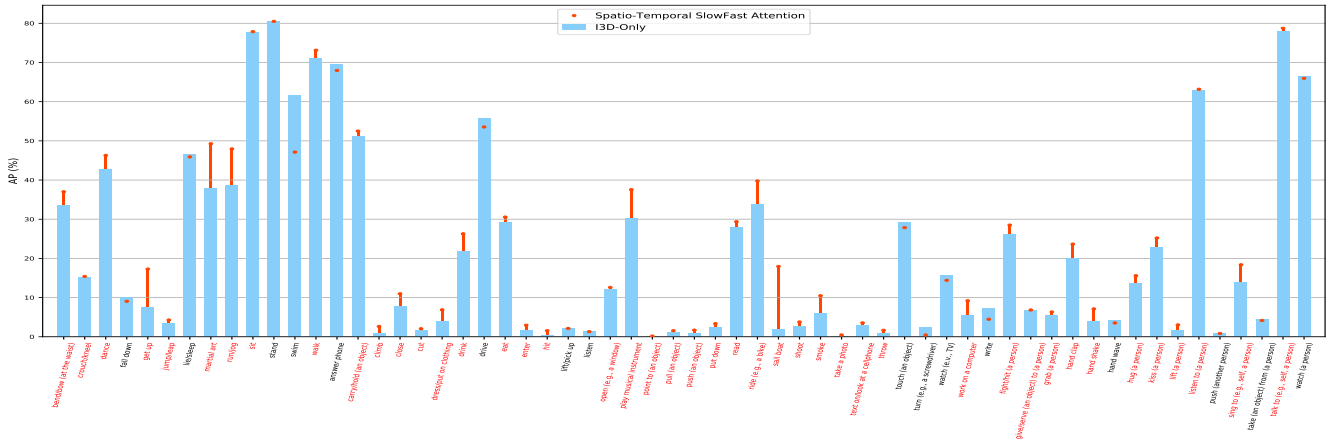

Fig. 3. Frame-AP for per-category on AVA.

Table 2. Ablation of diverse components of the self-attention module. Spatial SlowFast Attention (SSFA), Temporal SlowFast Attention (TSFA), Global Max Pooling (GMP), Global Average Pooling (GAP), Layer Norm (LN)

SSFA	TSFA	GMP	GAP	#layers	#dims	LN	mAP
	✓		✓	1	512		18.5
	✓	✓		1	512		19.2
✓		✓		1	512		20.8
✓	✓		✓	1	2048		21.3
✓	✓		✓	2	2048		21.7
✓	✓		✓	2	2048	✓	**23.0**

mance improvements in many categories that are associated with other objects or interactions with other humans. The reason is that the Spatio-Temporal SlowFast Self-Attention Module is applied to the features obtained through RoIPool, allowing network to focus more on objects or humans in the surrounding Pooled context.

4.3. Ablation Study

As we can see from the temporal axis, if GAP is performed, all frames are averaged regardless of the action, so there may be a loss of information. Therefore, we assume that GMP has achieved better frame-mAP. The results of SSFA and TSFA represent that it is more important to focus on the object or human associated with the action in spatial information than to find the important frame on the temporal axis.

5. CONCLUSION

We proposed the Spatio-Temporal SlowFast Self-Attention network which can extract important spatial information, temporal information, slow action information, and fast action information from video understanding. The four proposed features played a major role in distinguishing difficult action classes. Our network applied only the simple self-attention module and achieved remarkable performances compared the previous state-of-the-art network using less resources.

6. ACKNOWLEDGEMENT

This work was supported by Institute of Information Communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2018-0-01290). This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the SW Starlab support program (IITP-2017-0-00897) supervised by the IITP(Institute for Information communications Technology Promotion)

7. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[7] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[8] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[9] Hongsong Wang and Liang Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 499–508.

[10] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[12] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[13] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7310–7311.

[14] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al., "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[15] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid, "Actor-centric relation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 318–334.

[16] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz, "Step: Spatio-temporal progressive learning for video action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 264–272.

[17] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid, "A structured model for action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9975–9984.

[18] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman, "Video action transformer network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[19] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.