SPATIO-TEMPORAL SLOWFAST NETWORK FOR ACTION RECOGNITION

Author: Myeongjun Kim, Taehun Kim, Daijin Kim

Organization: POSTECH

Session Title: ARS-18, Machine Learning for Recognition in Images and Videos II (ARS-18.2)

Contents

Introduction

- What is Action Recognition?
  - Action Classification & Bounding Box Regression
    - Action recognition is localizing the location of a person and recognizing the behavior of target person.
    - Each target person can have a multi-label actions.
    - Ex, Atomic Visual Actions (AVA) [1], etc.

(A) GT: Sit, Answer Phone

(B) GT: Bend/Bow (at the waist), Carry/Hold (an object)

Fig. 1.1, The example of Atomic Visual Actions (AVA) dataset.

[1] 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, pages 6047–6056, 2018.

Related Work

- Conventional Action Recognition
  - To localizing human information, action recognition follows Faster-RCNN [1] algorithm. However, the RoIPool module performs RoIPooling across the entire temporal axis.

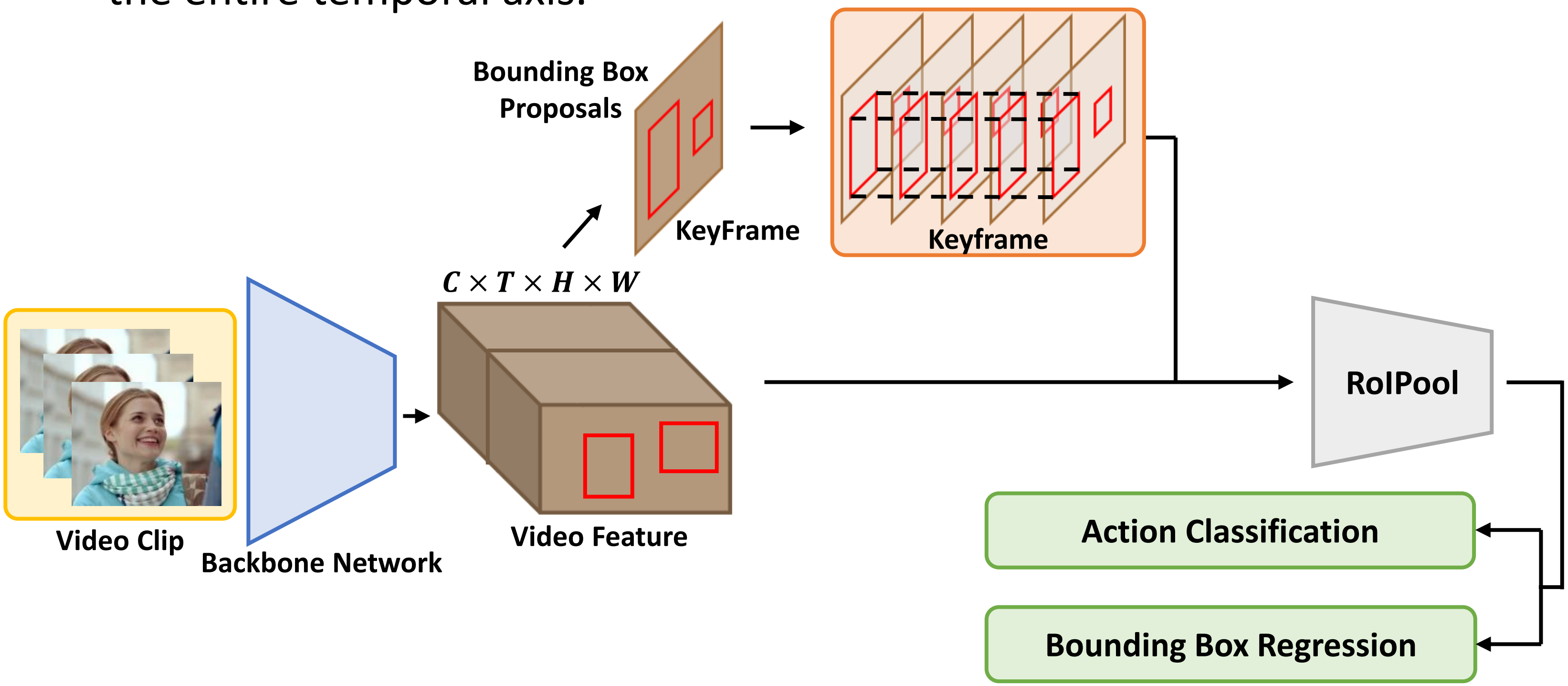

Fig. 2.2, The overall architecture of conventional action recognition network.

[1] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015).

Related Work

- Self-Attention Mechanism [1, 2]
  - Self-Attention Mechanism was mainly used in the language model and was used to consider long-rage interaction.
  - However, it is used by extending it from language model to image data. In the image, Self-Attention Mechanism represents the effect of i^{th} pixel on j^{th} pixel in spatial axis.

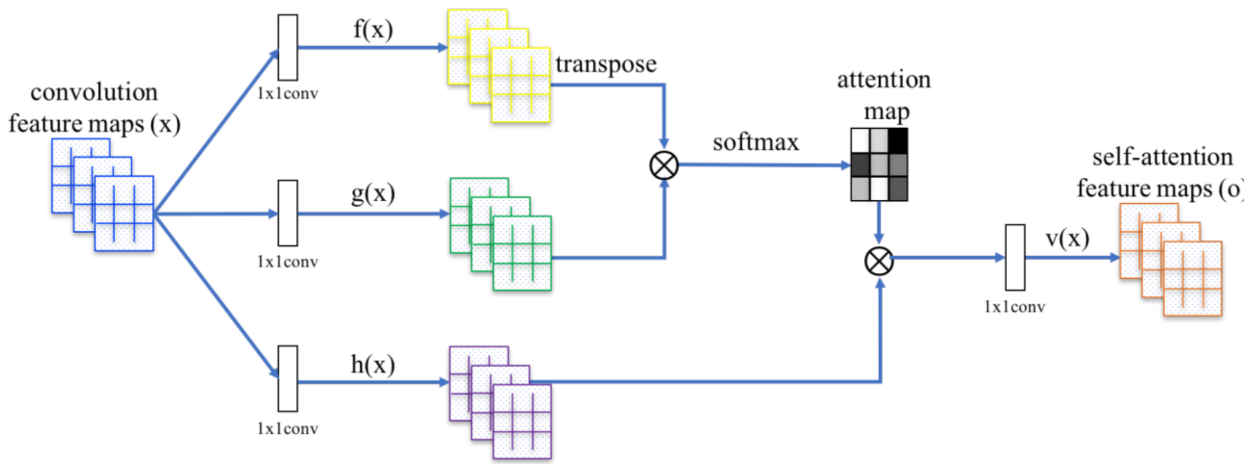

Fig. 2.1, The overall architecture of Self-Attention GAN [2] network.

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017).
[2] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning. pp. 7354–7363(2019).

Method

- Motivation
  - We need to capture a long-range interactions in the spatial axis and temporal axis.
    - When judging a person's behavior, important information is extracted from the features of hands, other objects, and other humans.

(A) Fight / hit (a person)

(B) Smoke

Fig. 3.1, The example of the ground truth bounding box of the “Fight/Hit (a person)” and “Smoke” classes.

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Method

- Spatio-Temporal SlowFast Self-Attention Network
  - We reconstruct the 3D self-attention module using a 2D self-attention mechanism.
  - In addition, the self-attention module was applied by dividing it into spatial information, temporal information, slow action information, and fast action information.

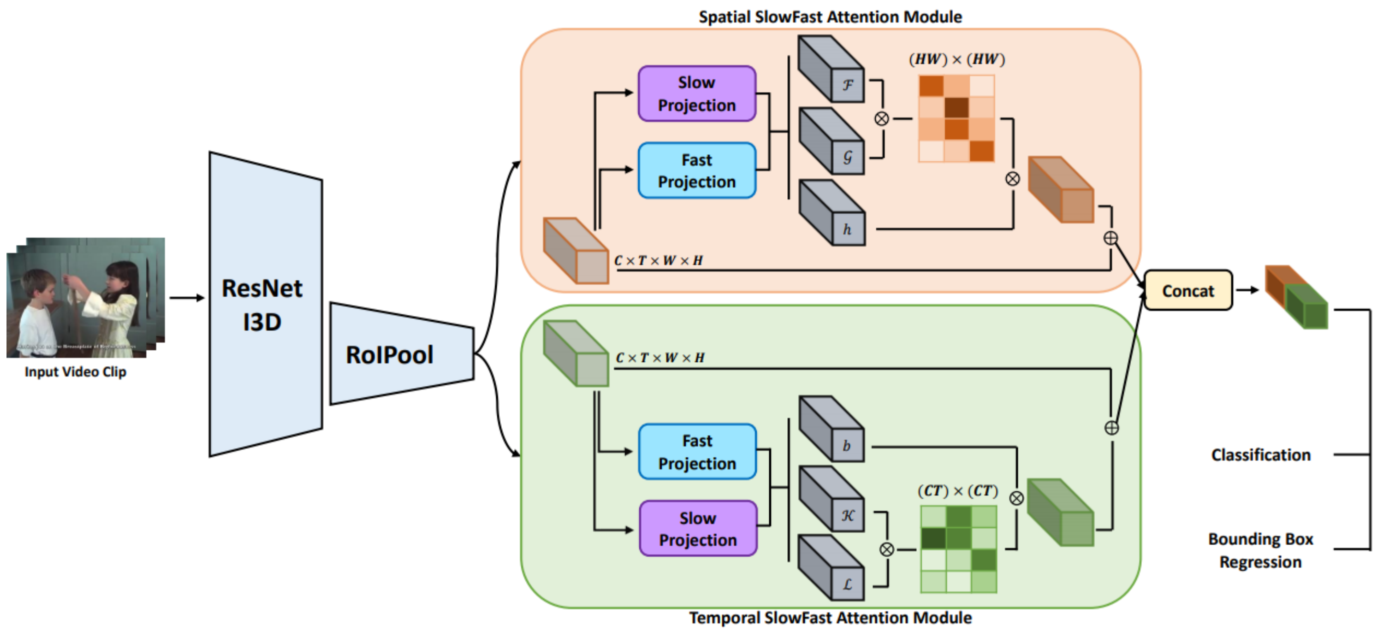

Fig. 3.2, The overall architecture of Spatio-Temporal SlowFast Self-Attention network.

Method

- Spatio-Temporal SlowFast Self-Attention Network
  - Spatio-Temporal Slow Self-Attention Module
    - This module extracts spatial and temporal information from slow actions.
    - In the Self-Attention module, there are linear projection parts of key, query, and value, and in the proposed 3D self-attention module, we project feature map using a 3D convolution layer. The slow action can capture using large temporal kernel size (7 x 1 x 1).

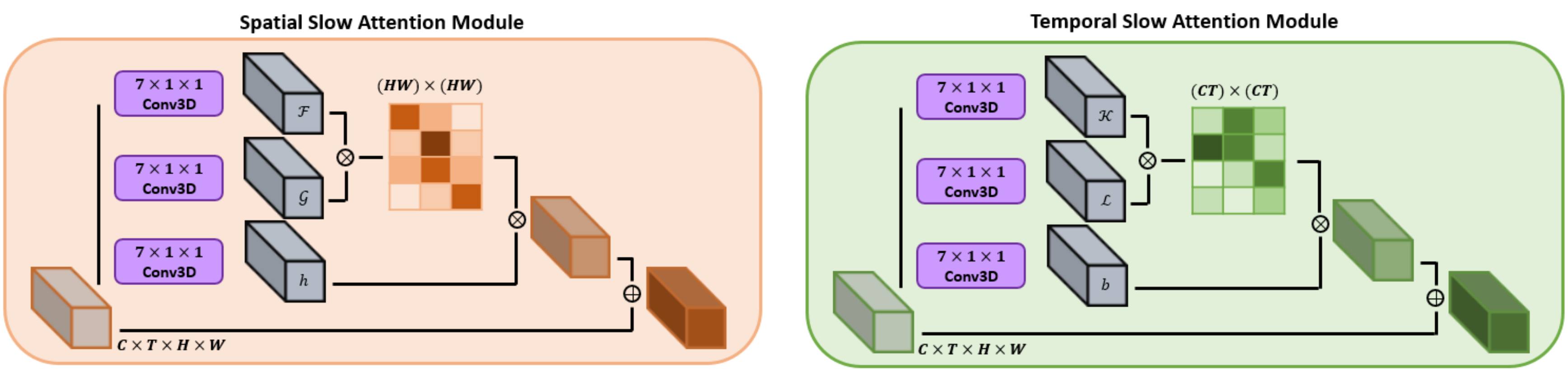

Fig. 3.3, The details of Spatio-Temporal Slow Self-Attention module.

Method

- Spatio-Temporal SlowFast Self-Attention Network
  - Spatio-Temporal Fast Self-Attention Module
    - This module extracts spatial and temporal information from fast actions.
    - we project feature map using a 3D convolution layer. The fast action can capture using (1 x 1 x 1) temporal kernel size 3D convolution layer.

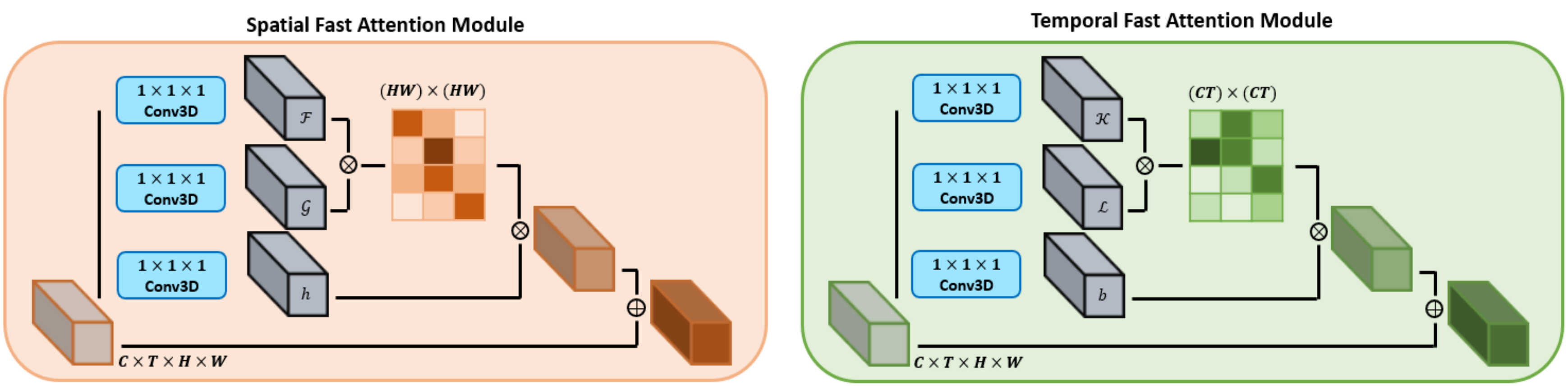

Fig. 3.4, The details of Spatio-Temporal Fast Self-Attention module.

Experiments

- Atomic Visual Actions (AVA) Dataset
  - The AVA dataset is more realistic compared to other datasets because the dataset crawls Youtube movies and has a multi-label for each person.
  - The AVA dataset is divided into Training 211K and 57K validation. Also, using the RTX Titan 8 gpus takes 3-5 days of training time.
  - Training is conducted on 80 classes, and evaluation is performed on 60 classes with 25 or more instances.
  - The evaluation metric uses Frame-AP and is Average Precision (AP) in the keyframe. Intersection of Union (IoU) threshold of 0.5 was used.

Left: Sit, Ride, Talk to; Right: Sit, Drive, Listen to

Left: Sit, Talk to, Watch; Right: Crouch/Kneel, Listen to, Watch

Fig. 4.1, The example of Atomic Visual Actions (AVA) Dataset.

[1] 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, pages 6047–6056, 2018.

Experiments

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

Table 4.1: Comparison of modalities, architecture, input size and Frame mAP with state-of-the-art methods on AVA.

[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, pages 6047–6056, 2018.
[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), pages 318–334,2018.
[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, pages 264–272, 2019.
[17] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 6299–6308.
[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.

Experiments

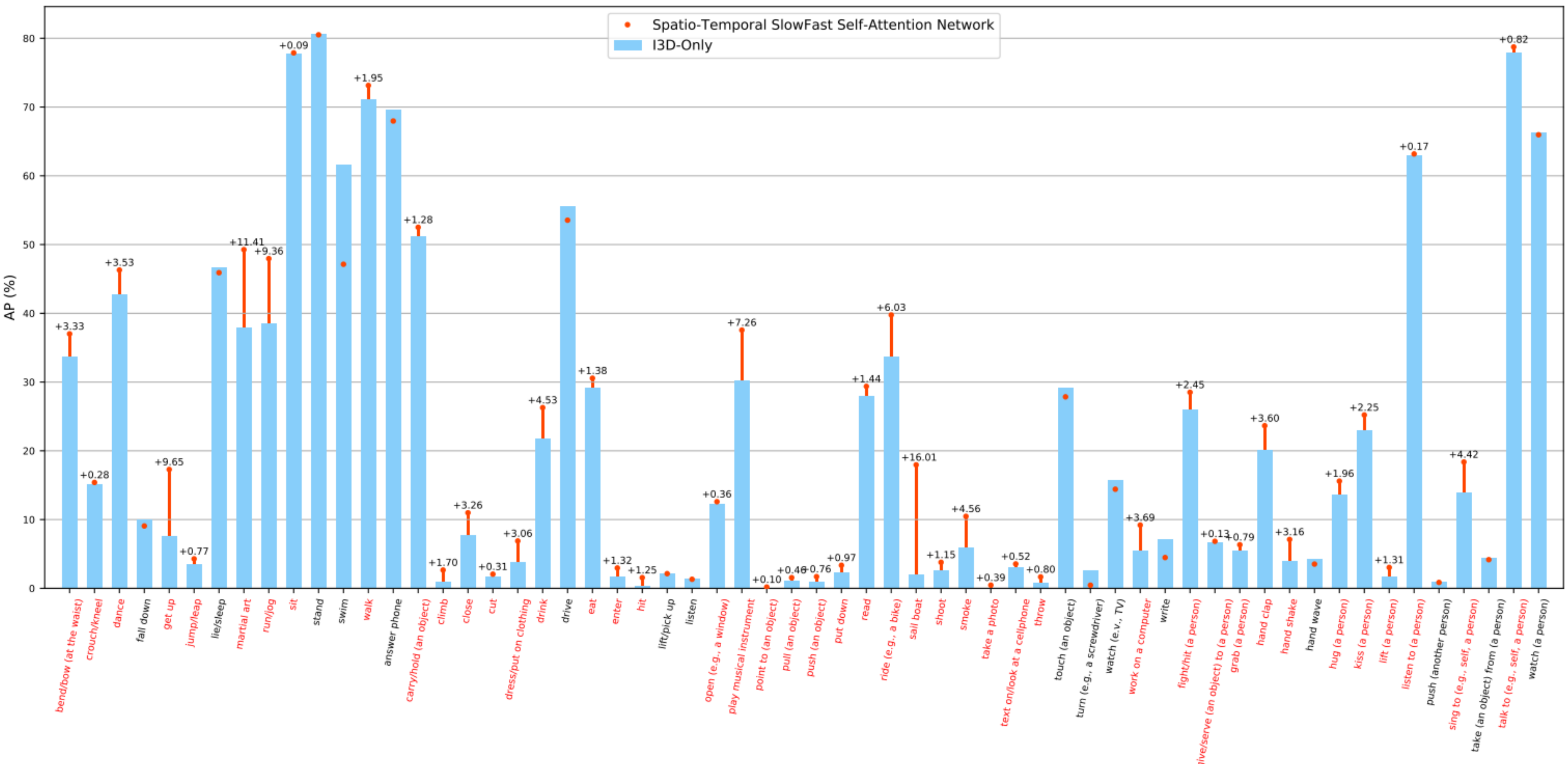

Fig. 4.2: Comparison of Spatio-Temporal SlowFast Self-Attention Network and baseline network on 60 classes.

Experiments

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**

Table 4.2: Comparison of module influence, pooling methods, number of layers, number of dimensions, and layer norm effects. SSFA: Spatial SlowFast Self-Attention, TSFA: Temporal SlowFast Self-Attention, GMP: Global Max Pooling, GAP: Global Average Pooling, LN: LayerNorm

Experiments

- Qualitative Results

GT: Sit, touch (an object)
Pred: Sit, touch (an object), carry/hold (an object)

GT: Stand, carry/hold (an object), talk to
Pred: Stand, carry/hold (an object), talk to, watch (a person)

GT L: Sit, Talk to
GT R: Sit, Listen to, Watch (a person)
Pred L: Sit
Pred R: Sit, Carry/Hold, Listen to, Watch (a person)

GT: bend/bow (at the waist), watch (a person)
Pred: bend/bow (at the waist), watch (a person)

Fig. 4.3, The Example of top predictions using Spatio-Temporal SlowFast Self-Attention Network.

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.

- Our network applied only the simple self-attention module and achieved 23.0 mAP compared the previous state-of-the-art network using less resources.

- Compared to the ResNet-I3D, 44 out of 60 evaluation classes represent performance improvement.