IEICE TRANS. INF. & SYST., VOL.E104–D, NO.1 JANUARY 2021

LETTER

# Spatio-Temporal Self-Attention Weighted VLAD Neural Network for Action Recognition

**Shilei CHENG**[†a)], **Mei XIE**[†], **Zheng MA**[†], **Siqi LI**[†], **Song GU**[††], ***Nonmembers***, ***and*** **Feng YANG**[†], ***Member***

**SUMMARY** As characterizing videos simultaneously from spatial and temporal cues have been shown crucial for video processing, with the shortage of temporal information of soft assignment, the vector of locally aggregated descriptor (VLAD) should be considered as a suboptimal framework for learning the spatio-temporal video representation. With the development of attention mechanisms in natural language processing, in this work, we present a novel model with VLAD following spatio-temporal self-attention operations, named spatio-temporal self-attention weighted VLAD (ST-SAWVLAD). In particular, sequential convolutional feature maps extracted from two modalities *i.e.*, RGB and Flow are receptively fed into the self-attention module to learn soft spatio-temporal assignments parameters, which enabling aggregate not only detailed spatial information but also fine motion information from successive video frames. In experiments, we evaluate ST-SAWVLAD by using competitive action recognition datasets, UCF101 and HMDB51, the results show our proposed approach achieves outstanding performance. The source code is available at:*https://github.com/badstones/st-sawvlad*.

***key words:*** *human action recognition, video representation, VLAD, self-attention module*

## 1. Introduction

Human action recognition is one of the fundamental problems in computer vision with applications ranging from video understanding to Human-Computer interaction. The methods of incorporating the spatio-temporal information have been shown crucial in different tasks of video analysis [1]. Early typical methods such as Improved Dense Trajectories (iDT) [2] and HOG3D [3] are both dependent on hand-crafted spatio-temporal descriptors and then encode them through Vector of Locally Aggregated Descriptors (VLAD) [4] to form the final video representations.

With promising success in image classification, deep learning methods also show excellent performance in action recognition tasks. Typically, Two-Stream [1] and Temporal Segment Networks (TSN) [5] both of which decompose the video into RGB and Flow streams to respectively model the appearance changing and dynamic motions are superior to iDT over several challenging datasets. Along with the cue of space and time, another primary CNN architecture, named C3D [6], which aims to learn the spatio-temporal features by using 3D CNNs is capable of modeling appearance and motion simultaneously. While both approaches have been made rapid progress, Two-Stream architectures have generally outperformed the spatio-temporal convolution which requires the pre-trained model on large video datasets such as Kinects [7] and Sports1M [8].

However, both architectures largely disregard the long-term temporal structure of the video and essentially learn a classifier that operates on an individual frame or short blocks of frames (16 for C3D), it is quite difficult to model the complex spatio-temporal structure of human actions. To address this issue, ActionVLAD [9] was proposed to build a trainable video-level representation architecture, which aggregates simultaneously appearance and motion features. In detail, each video descriptor is assigned to one of the action words, while the method computes their sum inside each of the visual centers according to the assignment, which is computed individually for each frame and loses much temporal information in the successive frames. One possible solution to address this drawback is to utilize Long Short-Term Memory (LSTM) units [10], [11] to capture long-range temporal dependencies, however, this method requires repeating local operations, which may cause huge memory cost and optimization difficulties.

Attention mechanism plays a significant role in the field of natural language processing and image recognition [12]. But it is still an ascendant research topic in human action recognition. Inspired by the recent works [12], [13], in this work, the spatio-temporal weighted VLAD is proposed, to further boost the performance of action recognition by introducing a self-attention module used as soft assignments between video snippets. One important distinction between non-local neural networks [13] and ours is that our method computes the response as a weighted sum of residual vectors between features and cluster centers instead of using the response as video representation directly. The characteristics of our ST-SAWVLAD are summarized as follows:

- Compared with Two-Stream and TSN methods, our approach not only utilizes self-attention module to learn the temporal contents from successive video frames, but also fully takes advantage of both spatial and temporal information to aggregate the discriminative video representations.
- Compared with ActionVLAD, the ST-SAWVLAD is capable of modeling the spatio-temporal relationships

Manuscript received January 7, 2020.
Manuscript revised February 12, 2020.
Manuscript publicized October 1, 2020.

[†]The authors are with School of Information and Communication, University of Electronic Science and Technology of China, P.R.China.
[††]The author is with Department of Aircraft Maintenance Engineering, Chengdu Aeronautic Polytechnic, Chengdu, China.
a) E-mail: slcheng1986@foxmail.com

DOI: 10.1587/transinf.2020EDL0002

between local descriptors and action words by our proposed self-attention module, rather than simply computing the sum of the aggregation results of each frame, which ignores the temporal dependencies of sequential frames.

The main contributions of this paper are summarized in two aspects. Firstly, an effective spatio-temporal self-attention weighted VLAD neural network is proposed to capture spatio-temporal characteristics, instead of stacking a certain number of recurrent operations. Secondly, each stream of proposed ST-SAWVLAD can be optimized with end-to-end manner and excellent performance is achieved on UCF-101 and HMDB-51 datasets.

## 2. Spatio-Temporal Self-Attention Weighted VLAD (ST-SAWVLAD)

As VLAD/NetVLAD encoding only aggregates local information of spatial context of images, it is not straightforward to extend these methods to the tasks of video processing. We introduce the ST-SAWVLAD network which specifically contains a self-attention module (the green box diagram of Fig. 1). The self-attention module is similar with the non-local block [13], given the feature maps $\mathbf{x} = [x_1, x_2, \cdots, x_C] \in \mathbb{R}^{T\times C\times H\times W}$, they are first convolved with kernel $W_\theta$ and $W_\phi$ to generate attention features, $\theta(x_i) = W_\theta * x_i$, $\phi(x_j) = W_\phi * x_j$, where $*$ denotes convolution operation, $x_i$, $x_j \in \mathbb{R}^{T\times H\times W}$. $W_\theta$ and $W_\phi$ are the weight matrices with $K$ channels to be learned. Both $\theta(x_i)$ and $\phi(x_j)$ represent attention features which integrate different information from spacetime. Then the attention map can be calculated as $\sigma(\mathbf{x}) = \frac{exp(s)}{\sum_{i=1}^{K} exp(s)}$, where $s = \theta(x_i)^T\phi(x_j)$, and $\sigma(\mathbf{x}) \in \mathbb{R}^{THW\times THW}$ is the attention map, which indicates the weights of all positions in the attention features. We define the soft-assignment from self-attention module as $\alpha = \sigma(\mathbf{x})g(\mathbf{x})$ with the shape of $TK\times H\times W$, $g(x_j) = W_g * x_j$,

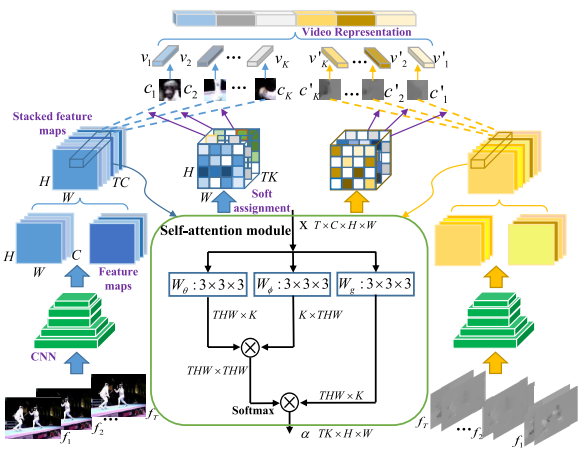



**Fig. 1** The flowchart of the proposed ST-SAWVLAD framework, which is essentially based on Two-Stream architecture. Our self-attention is described in the green box diagram, where "⊗" denotes matrix multiplication, the softmax operation is employed to generate the attention map of every position, and the blue boxes denotes $3 \times 3 \times 3$ convolutions.

where $W_g$ is also a 3D trainable kernel with the same size as $W_\theta$ and $W_\phi$. Note that $\sigma(\mathbf{x})$ has the form of $softmax$ function, thus the formulation of soft-assignment can be rewrite as:

$$\alpha = softmax(\mathbf{x}^T W_\theta^T W_\phi \mathbf{x})g(\mathbf{x}) \tag{1}$$

As Fig. 1 shown, we first utilize the segment-based strategy [5] to divide the video into several segments with equal duration, and then pick up one snippet randomly from its corresponding segment to form a snippet group which retains the sequentiality of original video frames. We take this snippet group as a batch to feed the CNN, thus it ensure the temporal structure even with shuffle process. As Fig. 1 shown, $T$ snippets for one iteration are prepared and put into a CNN model to extract feature maps which will be stacked along the temporal dimension to form the input of self-attention module, whose output, the trainable spatio-temporal soft assignment are utilized as the weights for aggregating local descriptors to certain visual centers. The final representations of our ST-SAWVLAD are formulated as

$$v_k = \sum_{t=1}^{T}\sum_{j=1}^{W}\sum_{i=1}^{H} \alpha_t^k(i,j)(x_t(i,j) - c_k) \tag{2}$$

where $x_t(i,j)$ is a $D$-dimensional descriptor at location $(i,j,t)$, $\alpha_t^k(i,j)$ denote the assignment weight of aggregating the descriptor at location $(i,j)$ of $t$-th frame to the $k$-th visual word. Thus we have $\alpha = \{\alpha_t^k(i,j)\}^{T\times K\times H\times W}$. $c_k$ indicates $k$-th visual word. From Eq. (2), we can observe that the video representations have both spatial and temporal characteristics. Finally, we concatenates $v_k$ over $K$ visual words to form the video representation with shape $K \times D$.

## 3. Implementation Details

**Network architecture:** We incorporate the spatio-temporal self-attention module into general CNNs to form end-to-end attention networks for action recognition. We investigate VGGNet-16 and BN-Inception as backbone networks respectively.

**Training:** We train our networks with a single-layer linear classifier on top of the ST-SAWVLAD network. Throughout, we set a dropout of 0.5 over the representation to avoid overfitting. The number of cluster centers is empirically set to 64. Data argumentation is done as the same as previous work [5]. All the parameters of our whole model including the backbone network, ST-SAWVLAD model, and the classifier are optimized by SGD with a momentum of 0.9. We also adopt a two-stage optimization scheme [9], for the first stage, only the parameters in the ST-SAWVLAD model and the classifier could be trained, and in the next stage, all the parameters of the whole model are optimized. For the RGB stream, we set 90 and 120 epochs for the first and the second stage, respectively. The initial learning rate is set to 0.03 and decreased to its 0.1 on the 80th and 160th epochs. For the Flow stream, there are 70 epochs for the first stage, while

300 epochs for another. The initial learning rate is 0.01, and decreased to its 0.1 on the 70th and 220th epochs.

**Test:** In the testing stage, we report our performance using 10 crops which contains 4 corners and 1 center cropping, and their flips for every testing video.

## 4. Experiments and Analysis

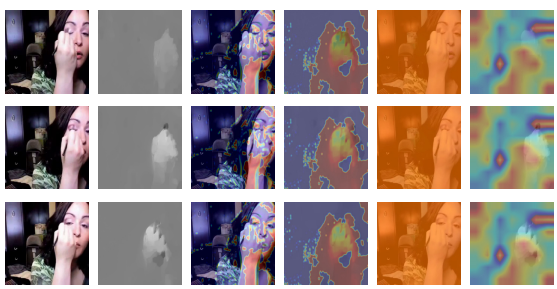
We evaluate the proposed ST-SAWVLAD with the various network architectures on standard action recognition benchmarks. We conduct experiments on two popular trimmed action recognition benchmarks, UCF101[14] and HMDB51[15] respectively. UCF101 contains 13320 sports video clips with 101 action categories, and HMDB51 consists of 6766 varied and realistic video clips from 51 action classes. The pre-defined three train/test splits are utilized for evaluation. We first visualize the soft assignment obtained by the self-attention module, and then evaluate the effect of varied input length on both UCF101 and HMDB51 dataset split1, finally, we compare our method with outstanding methods and evaluate the generality of the proposed network.

### 4.1 Visualization Analysis

We visualize what the proposed self-attention module pays attention to over the frames and from different spatial positions. We draw the soft assignment as a heat map and respectively weight to the RGB and Flow frames which are sampled by temporal segment strategy [5] ($T = 3$). From Fig. 2 (c)-(d), we can observe that the weight is largely distributed on the key positions of actions, such as arm and eyes. While as (e)-(f) of Fig. 2 shown, the weight of soft assignment obtained by a 2D convolution layer, *i.e.*, ActionVLAD, is almost uniformly distributed over frames.

### 4.2 Evaluation on Different Time Steps

We vary the number of time steps $T = [4, 6, 8, 10, 12, 14]$ and evaluate the recognition performance using the same test approaches. The results are illustrated in Fig. 3 from which we observe that it will lead to better performance with increasing the number of time steps. For instance, the performance of ST-SAWVLAD with $T = 10$ is remarkably outperformed than that with $T = 4$ for both two datasets. This improvement implies that using more input length will help to capture richer context information to better model long-range temporal structure. However, with the increasing of the input length, the greater memory size is required, considering the limited memory size and the trade-off between computational burden and recognition performance, we set $T = 12$ in the following experiments.



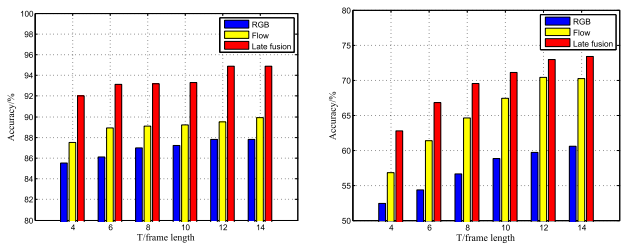
(a) (b) (c) (d) (e) (f)

**Fig. 2** Visualization of soft assignment in different frames from appearance (RGB) and motion (Flow) streams. (a) denotes RGB modality frames, (b) denotes the corresponding Flow modality frames, (c)-(d) respectively represent soft assignment obtained from self-attention module and weighted on RGB frames and Flow frames, while (e)-(f) represent soft assignment produced by a single 2D convolution and weighted on RGB frames and Flow frames respectively. The samples are the 23th, the 62th and the 102th frames respectively from action 'ApplyEyeMakeup'.

### 4.3 Comparison with the Outstanding Methods

In this subsection, we first investigate the generality of our self-attention module, we respectively plug it into VGGNet-16 and BN-Inception. The late fusion approach means that the prediction scores of the RGB and Flow stream are averaged as the final action classification. To compare with our self-attention module, we also use a single 2D convolution layer as a baseline to capture the soft assignment. We evaluate the performance on HMDB51 split1 and report the result in Table 1. For VGGNet-16, our proposed module respectively promotes 2.4%, 1.3%, 1.5% on RGB/Flow/Late fusion on the HMDB51 split1. For BN-Inception, our model respectively increases 3.0%, 1.3%, 2.2% on RGB/Flow/Late fusion. The improved results with our self-attention module demonstrate the generality of our layer for general deep networks. Furthermore, the evaluation results of VGGNet-16 show that our method outperforms with another similar framework like ActionVLAD.

Next, we compare the proposed approach to varieties




(a) Varied frame length on UCF101. (b) Varied frame length on HMDB51.

**Fig. 3** The performance of action recognition with varied frame length ($T = \{4, 6, 8, 10, 12, 14\}$) on UCF101 and HMDB51 split1 respectively.

**Table 1** Performance of the proposed self-attention module on popular networks. $l_{conv2D}$ denotes 2D convolution layer, $m_{attn}$ denotes self-attention module.

| Stream | VGGNet-16(%) | BN-Inception(%) |
|---|---|---|
| RGB + $l_{conv2D}$ | 51.2 | 51.9 |
| RGB + $m_{attn}$ | **53.6** | **54.9** |
| Flow + $l_{conv2D}$ | 58.4 | 60.2 |
| Flow + $m_{attn}$ | **59.7** | **61.5** |
| Late fusion + $l_{conv2D}$ | 66.9 | 68.1 |
| Late fusion + $m_{attn}$ | **68.4** | **70.3** |

of recent action recognition methods that use a comparable base architecture to ours. As Table 2 shows our model outperforms all previous approaches on both UCF101 and HMDB51 averaged over 3 splits. Since our ST-SAWVLAD takes both still images and stacked optical flow as inputs, we first compare ST-SAWVLAD with Two-Stream based methods (see the first block of Table 2) which also utilize the same modalities. In detail, our model with VGGNet-16 outperforms Two-Stream (VGGNet-16) about 1.5% on UCF101 and 3.0% on HMDB51, compared with TSN (BN-Inception, 2-modality) which also utilizes the same pre-trained BN-Inception networks to extract features of video frames, ST-SAWVLAD improves the performance about 0.7% and 1.8% on UCF101 and HMDB51 respectively. ActionVLAD which ignores the temporal information in computing soft assignment of aggregation, however, the evaluation results are respectively lower about 2.0% on UCF101 and 3.4% on HMDB51. As the sequential modeling ability of LSTM makes them appealing to capture long-range temporal dynamics in videos, we compare our method with two LSTM-based benchmark methods over all three split1 of UCF101 and HMDB51, as the second block of Table 2 shown, we observe that our method outperforms LSTM-based methods by a large margin. This experiment demonstrates that our devised spatio-temporal soft assignment which captures the long dependencies in spacetime has positive effects on performance improvement for action recognition.

At last, we compare our method with the non-local based methods, since our self attention module is similar with the non-local block. Table 3 indicates the evaluation results, for fair comparison, the Resnet50 network is used as the backbone. For Resnet50 + 1 non-local block, we add 1 non-local block into res4 stage. For Resnet50 + 5 non-local blocks, we add 5 blocks (3 to $res_4$ and 2 to $res_3$, to every other residual block), then we respectively load the weights for these model, which have been pre-trained on the Kinetics dataset [7], finally we fine-tune the model on UCF101 and HMDB51 benchmark, the details of the implement are similar with the work descripted [13]. We report the results with late fusion approach, the first block of Table 3 shows the best performance is achieved when 5 non-local block are plugged into resnet50 with about 2% and 4% accuracy increasing on UCF101 and HMDB51 benchmarks compared with our method. The improvement may imply that multiple non-local blocks add depth to the baseline model and better perform long-range multi-hop communication. Note that our model also outperforms Resnet50 + 1 non-local block with Kinetics pre-training by 1.4% and 0.9% on UCF101 and HMDB51 respectively. Furthermore, we adjust our self-attention module to make it has the same shape between input and output, *i.e.*, $T \times K \times H \times W$, and thus the self-attention module evolves to the non-local block.

We further stack 5 non-local blocks as the ST-SAWVLAD module to evaluate the performance of long-range spatio-temporal characteristics. As the second block of Table 3 illustrated, this approach achieves the best performance among the comparison methods. We argue that there are two reasons, firstly, 5 stacked non-local blocks can get better spatio-temporal characteristics than single one, secondly, the trainable VLAD neural network precisely quantify the residual between feature vectors and cluster centers, which are distinguished to model the long temporal range sequences.

**Table 2** Comparison with the outstanding methods on UCF101 and HMDB51 averaged over three splits.

| Method | UCF101 | HMDB51 |
|---|---|---|
| Two-Stream Fusion (VGGNet-16) [1] | 92.5 | 65.4 |
| TSN (BN-Inception, 2-modality) [5] | 94.0 | 68.5 |
| ActionVLAD (LateFuse, VGGNet-16) [9] | 92.7 | 66.9 |
| Two-Stream+LSTM [11] | 88.6 | - |
| VideoLSTM [10] | 89.2 | 56.4 |
| Ours (VGGNet-16 + Late fusion) | 94.0 | 68.4 |
| Ours (BN-Inception + Late fusion) | **94.7** | **70.3** |

**Table 3** Comparison with the non-local (NL) based methods

| Methods | UCF101(%) | HMDB51(%) |
|---|---|---|
| Resnet50 + 1-NL-block (pretrained) | 89.4 | 65.3 |
| Resnet50 + 5-NL-block (pretrained) | 92.5 | 70.1 |
| Resnet50 + Ours | 90.8 | 66.2 |
| Ours + 5-NL-block | **94.8** | **72.7** |

## 5. Conclusion

In this paper, we propose a novel model with VLAD following spatio-temporal self-attention operations, named spatio-temporal self-attention weighted VLAD (ST-SAWVLAD). Our method is an end-to-end trainable network and can learn a video representation with long-temporal dependencies. Experimental results on benchmark datasets have shown the outstanding performances of our method. In future work, we devote to further reduce the number of parameters of our self-attention module and explore the deeper spatio-temporal information by a structured self-attention model.

## Acknowledgements

This research is supported by National Nature Science Foundation of China (Grant no. 61271288) and Science and technology program of Sichuan province (Grant no. 2018SZ0357).

### References

[1] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.1933–1941, 2016.

[2] H. Wang and C. Schmid, "Action recognition with improved trajectories," Proc. IEEE Int. Conf. Comput. Vis., pp.3551–3558, 2013.

[3] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," 2008.

[4] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," CVPR 2010-23rd, IEEE Computer Society Conf. Comput. Vis. Pattern Recognit.,

pp.3304–3311, 2010.

[5] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," European Conference on Comput. Vis., pp.20–36, Springer, 2016.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," Proc. IEEE Int. Conf. Comput. Vis., pp.4489–4497, 2015.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," IEEE Conf. Comput. Vis. Pattern Recognit., 2014.

[9] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.971–980, 2017.

[10] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C.G. Snoek, "Videolstm convolves, attends and flows for action recognition," Computer Vision and Image Understanding, vol.166, pp.41–50, Jan. 2018.

[11] J.Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.4694–4702, 2015.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems, pp.5998–6008, 2017.

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.7794–7803, 2018.

[14] 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.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," 2011 IEEE Int. Conf. Comput. Vis., pp.2556–2563, 2011.

[16] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: Cnn architecture for weakly supervised place recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.5297–5307, 2016.

[17] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Workshop on Statistical Learning in Computer Vision, ECCV, pp.1–2, Prague, 2004.