Semantic Adversarial Network with Multi-Scale Pyramid Attention for Video Classification

De Xie,[1] **Cheng Deng,**[1*] **Hao Wang,**[1] **Chao Li,**[1] **Dapeng Tao**[2]

[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]School of Information Science and Engineering, Yunnan University, Kunming 650091, China
{xiede.xd, chdeng.xd, haowang.xidian,}@gmail.com, li_chao@stu.xidian.edu.cn, dapeng.tao@gmail.com

Abstract

Two-stream architecture have shown strong performance in video classification task. The key idea is to learn spatio-temporal features by fusing convolutional networks spatially and temporally. However, there are some problems within such architecture. First, it relies on optical flow to model temporal information, which are often expensive to compute and store. Second, it has limited ability to capture details and local context information for video data. Third, it lacks explicit semantic guidance that greatly decrease the classification performance. In this paper, we proposed a new two-stream based deep framework for video classification to discover spatial and temporal information only from RGB frames, moreover, the multi-scale pyramid attention (MPA) layer and the semantic adversarial learning (SAL) module is introduced and integrated in our framework. The MPA enables the network capturing global and local feature to generate a comprehensive representation for video, and the SAL can make this representation gradually approximate to the real video semantics in an adversarial manner. Experimental results on two public benchmarks demonstrate our proposed methods achieves state-of-the-art results on standard video datasets.

Introduction

Video classification is a fundamental task in computer vision community, and it serves as an important basis for high-level tasks, such as video caption (Wang et al. 2018), action detection (René and Hager 2017), and video tracking (Li et al. 2018b). Significant progress on video classification has been made by deep learning on account of the powerful modeling capability of deep convolutional neural networks that obtain superior performance than those hand-crafted representation based methods. However, compared with other visual tasks (Li et al. 2018a; Fan et al. 2018; Deng et al. 2018; Yang et al. 2018), video classification should consider not only static spatial information in each frame but also dynamic temporal information between frames. Although deep convolutional neural networks can model spatial information well, it is limited ability to capture temporal information only from frame sequence. Therefore, how to model spatial and temporal information effectively with deep learning framework is still a challenging problem.
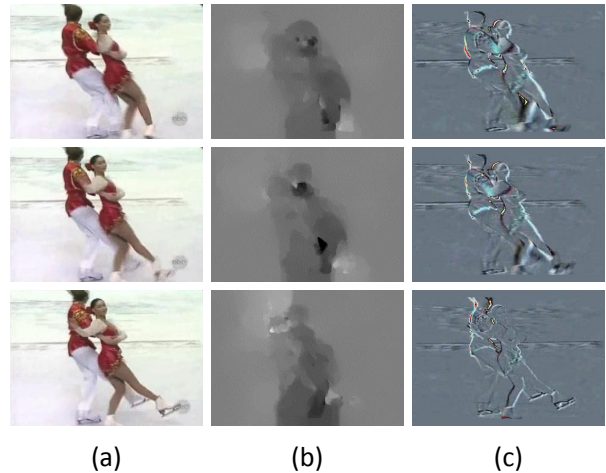
---

*Corresponding author.

Figure 1: Modeling temporal information with images. (a) input frames; (b) The optical flows between these frames, (c) The differential images between multiple video frames.

Video classification methods based on deep learning can be divided into three different categories. The first category relies on a combination of multiple input modalities, which models spatial and temporal information, respectively. The two-stream CNN (Simonyan and Zisserman 2014) is a groundbreaking work of this category, which captures static spatial information and dynamic temporal information with different streams from multi-modality input, usually RGB images and optical flow. Due to its prominent performance, many state-of-the-art methods can be considered as variants and improvements of this paradigm. However, this method suffers from the heavy reliance on optical flow to model temporal information, which are often expensive to compute and store. To overcome this limitation, the second category takes 2D CNN with temporal models on top such as LSTM (Donahue et al. 2015), temporal convolution (Yue-Hei Ng et al. 2015) and sparse sampling and aggregation (Wang et al. 2016). This category usually extracts features from different frames with 2DCNN, then captures the relationship between these features using temporal models. Such type of method more intuitive but lacks capacity to obtain local dynamic information and global context information.

The third category is based on 3DCNN (Tran et al. 2015; Ji et al. 2013), which employs 3D convolutions and 3D pooling operations to directly learn spatio-temporal features from stacked RGB volumes. Such methods seem to having ability solve the problem of spatio-temporal modeling but the performance is still worse than two-steam CNN based methods. Meanwhile, 3DCNN based methods also suffer from a large number of parameters and huge computation burden. More important, all three categories methods ignore utilizing the semantic information embodied in video, which leads to limited generalization performance. In fact, RGB frames contain abundant of semantic information, which can greatly improve classification performance. In addition, Inspired by prior work (Wang et al. 2016), we find the RGB differential image between multiple video frames have sufficient ability to model temporal information, which is less computational cost than optical flows. As shown in Figure 1, the RGB differential images are sensitive to the part of the motion in the video, which means that the details of RGB differential images have ability to model the temporal information.

In this paper, we propose a new two-stream based architecture to address all mentioned problems. Specially, we design a *Spatial Network* to model spatial information which takes RGB frames as input and a *Temporal Network* to model temporal information which exploits differential images as input. In order to obtain more discriminative representation, we design a multi-scale pyramid attention (MPA) layer to capture multi-scale features from different stage of *Spatial Network* and *Temporal Network*, and then combine these multi-scale information into new representation. In addition, we devise semantic adversarial learning (SAL) module aiming to guide *Spatial Network* and *Temporal Network* to learn more discriminative and semantic video representation. Overall, the main contribution of the proposed method can be summarized as follows:

- We propose a new deep architecture for video classification, which contains *Spatial Network* and *Temporal Network* only taking RGB frames as input, which significantly reduced computational complexity without sacrificed.

- We devise a multi-scale pyramid attention (MPA) layer that conducts attention-driven multi-scale features extraction and it is pluggable that can be easily embedded to other CNNs based architecture.

- We introduce a semantic adversarial learning (SAL) module, which can make fully use of video semantic information and guide video representation learning in adversarial manner.

- Experimental results on two public benchmarks for action recognition, HMDB51 and UCF101, highlight the advantages of our method and obtain improved performance compared to state-of-the-art methods.

Related Work

Video classification has received sustained attention in recent years, and has spawned lots of excellent works(Yang et

al. 2017a; 2017b; 2016). Traditional methods rely on handcraft visual features such as Motion Boundary Histogram (MBH) (Dalal, Triggs, and Schmid 2006) and improved Dense Trajectory (iDT) (Wang and Schmid 2013) which lack the discriminative capacity to classify complex videos. Deeply learned features is proved more powerful than handcraft features which can achieve superior performance.

There are many works have been trying to design effective deep architecture for video classification. For example, Karpathy *et al.* (Karpathy et al. 2014) showed the first large-scale experiment on training deep convolutional neural networks from a large video dataset, Sports-1M. Two-Stream (Simonyan and Zisserman 2014), as a significant breakthrough method, containing spatial and temporal nets to model appearance and motion information respectively. Wang *et al.* (Wang et al. 2016) designed temporal segment network to perform sparse sampling and temporal fusion, which aims to learn from the entire video. Wang *et al.* (Wang et al. 2017) further improved this architecture by integrating appearance information, short-term and long-term motion information, which achieve outstanding classification performance. However, these methods used optical flows to caption motion which is time consuming. In order to capture the motion information directly from RGB frames, a set of methods have been proposed to use 3DCNN (Tran et al. 2015), containing 3D convolution filters and 3D pooling layers, to model spatial and temporal information simultaneously. Although it is intuitive, but in fact, spatial information and temporal information may interfere with each other during the modeling process. So, it is still unclear whether this pattern could efficiently model spatial and temporal relation. To explicitly modeling spatial and temporal information, CNN-LSTM (Shi et al. 2017) based methods is proposed to model spatial and temporal information in different stage. They use CNN to extract Spatial features firstly and then model temporal information by using Long Shot-Term Memory(LSTM) as an encoder to encode the relationship between the sequence-illustrating spatial features. The main problem of these methods is the neglecting of local temporal relationship.

As a solution to the above problems, our method uses only RGB frames as input and can obtain hybrid features from different level through the multi-scale pyramid attention layer. Moreover, our proposed semantic adversarial learning module can take fully use of the video semantic information. which can guide the whole framework to learn more discriminative and semantic video representations.

Proposed Method

In this section, we give detailed description about our method for video classification. Specifically, we first introduce the structure of our method as a whole. Then, we study the multi-scale pyramid attention layer for multiple level features fusion. Finally, we present the semantic adversarial learning module in detail.

Overall Architecture

We design a unified convolutional network that can be divided into three components, *Spatial Network*, *Temporal*
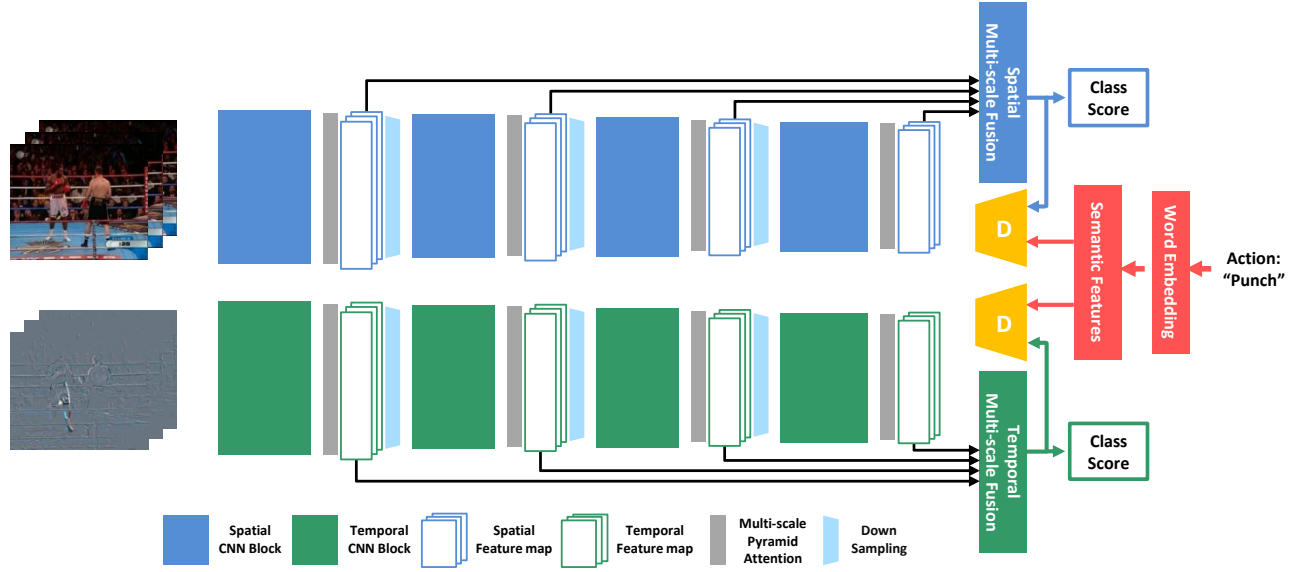
Figure 2: Our proposed framework. It can be divided into three components. The top branch is *Spatial Network*, which can model spatial information. The middle branch is *Semantic encoder*, which provides video semantic features. The bottom branch is *Temporal Network*, which can model temporal information. The multi-scale pyramid attention and semantic adversarial learning is introduced in our framework to learn discriminative and semantically rich representation from videos.

Network and *semantic autoencoder*. Figure 2 shows the overall architecture of our method. Specifically, we divide a deep convolutional neural network into four *stages*. The output of each stage represents multi-scale features of different visual levels. Then we insert multi-scale pyramid attention (MPA) layer at each stage in order to obtain refined multi-scale features. Given a video clip in the form of n frames sequence $\{\theta_1, \theta_2, \ldots, \theta_n\}$, we can obtain four different levels of features $\{V_1^i, V_2^i, \ldots, V_n^i\}, i \in \{1, 2, 3, 4\}$, after MPA layer of each stage. V_j^i can be rewritten as:

$$V_j^i = \text{Attn}(\theta_j : W_i), \qquad (1)$$

where $\text{Attn}(\theta_j : W_i)$ is a function representing MPA layer after the i-th stage with its parameters W_i operating on the frame θ_j. The multi-scale features can be formulated as:

$$\bar{V} = \sum_{i=1}^{4} \alpha_i \cdot \mathcal{G}(V_1^i, V_2^i, \ldots, V_n^i), \qquad (2)$$

where $\mathcal{G}$ is the frames consensus function, which is able to combine the features from multiple frames to obtain a consensus of representations among them. The α_i is a weight parameter about of i-th level, which can be learned automatically. In order to obtain video semantics, we design *semantic autoencoder*, taking video labels as input. Then semantic adversarial learning module is adopted to guide the architecture to learn semantic representations. We can formulize the adversarial loss as:

$$\mathcal{L}_{adv}(S, \bar{V}) = \text{Adv}(\bar{V}, S), \qquad (3)$$

where S is video semantic extracted from *semantic autoencoder*. After that, the proposed framework can thus generate semantically rich representation $\widetilde{V}$. We note that $\widetilde{V_k}$ is

the representation of the k-th video, and the class score C_k is the output of classification layer with $\widetilde{V_k}$ as input. Combining with standard categorical cross-entropy loss, the final loss function is formed as:

$$\mathcal{L}_{cls}(y, C) = -\sum_{i=1}^{K} y_i (C_i - \log \sum_{j=1}^{K} \exp C_j), \qquad (4)$$

where K is the number of video classes and y_i is the groundtruth label concerning class i. Finally, by combining Eq. 3 and Eq. 4, we can get the final loss function as:

$$\mathcal{L} = \mathcal{L}_{cls}(y, C) + \mathcal{L}_{adv}(S, \bar{V}). \qquad (5)$$

By minimizing $\mathcal{L}$, the objective of video classification can be achieved. Next, we will illustrate the proposed multi-scale pyramid attention layer and semantic adversarial learning module in detail.

Multi-Scale Pyramid Attention

The deep convolutional neural network extraction feature is a process from low-level visual features to high-level semantic features. Although the higher network layer is able to extract the global information, it will inevitably lose the details. Therefore, we intend to collect different levels features from a unified convolutional neural network. Specifically, we divide a network (such as ResNet101) into four *stages*, making each halve the resolution of the previous. Each stage contains multiple convolutional layers operating on feature maps of the same resolution. Then we can obtain four sets of features containing high-level semantic information and low level detailed information. However, these multi-scale features are too redundant for the classification
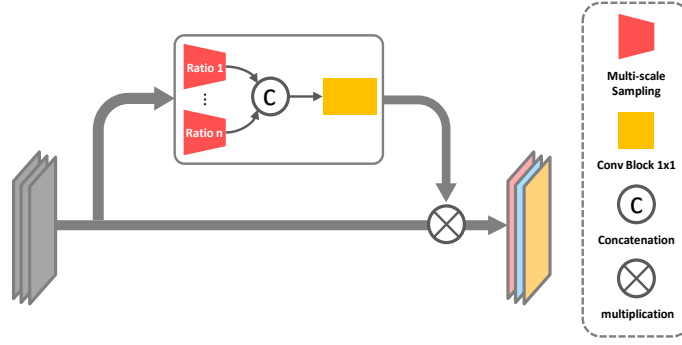
Figure 3: Multi-scale pyramid attention unit. We first put feature maps to multi-scale sampling branch to obtain multi-scale information. Then we concatenate them and feed them into a 1×1 convolutional layer to merge the information as well as generate the attention weights of feature maps. Finally we obtain the refined video representation.

task and may degrade the performance. So it is necessary to refine them for classification, and meanwhile, maintaining the multi-scale properties. Motivated by recent progress on residual learning, we introduce a novel multi-scale pyramid attention (MPA) layer that enables the network to consider the importance of each stage feature maps comprehensively with the informations of different receptive fields, so as to obtain reasonable attention weights. The structure of MPA is shown in Figure 3. It is a pluggable architecture and we put it to the end of each stage, as the Figure 2 shows.

Considering the importance of each stage feature maps from T scales, the attention weights h^i of the i-th stage feature maps f^i can be formulated as:

$$h^i = \tau(\sum_{t=1}^{T} \beta_t l_t(f^i : W_{att}^t)), \quad (6)$$

where τ is the function corresponding to the 1×1 convolutional layer. β_t is a weight that can be learned automatically. l_t represents the extractor of the t-th scale and W_{att}^t is its parameters. Therefore, the function of MPA layer can be rewritten as:

$$
\begin{aligned}
\bar{V} &= \sum_{i=1}^{4} \alpha_i \cdot \mathcal{G}(V_1^i, V_2^i, \ldots, V_n^i) \\
&= \sum_{i=1}^{4} \alpha_i \cdot \mathcal{G}(\text{Attn}(\theta_1 : W_i), \ldots, \text{Attn}(\theta_n : W_i)) \\
&= \sum_{i=1}^{4} \alpha_i \cdot \mathcal{G}(h_1^i \cdot f_1^i, \ldots, h_n^i \cdot f_n^i).
\end{aligned}
\quad (7)
$$

Then we can easily obtain the video representation with hybrid multi-scale information.

Semantic Adversarial Learning

Although multi-scale features have the ability to model video information, it still requires the guidance of explicit semantics. So, it is necessary to explore the exact semantic of videos. To this end, the *semantic autoencoder* is introduced in the proposed method. The structure of *semantic autoencoder* contains three fully connected layers supervised by video labels. After training, we freeze trained encoder

to generate exact semantic information. Specifically, the semantic S of videos can be written as:

$$S = \text{encoder}(y), \quad (8)$$

where y is the groundtruth label of videos.

In order to eliminate the difference of "real" semantic S and "fake" semantic $\bar{V}$, we design a semantic adversarial learning module because of its excellent ability of perfectly model the data distribution (Goodfellow et al. 2014; Li et al. 2018a). The adversarial loss function is used to encourage $\bar{V}$ close to S_i on the manifold to preserves semantics, by "fooling" a discriminator network D that outputs the probabilities to ensure $\bar{V}$ is as "real" as S. The adversarial loss function is formulated as:

$$
\begin{aligned}
\mathcal{L}_{adv}(S, \bar{V}) &= \text{Adv}(\bar{V}, S) \\
&= \mathbb{E}_S(\log D(S)) + \mathbb{E}_{\bar{V}}(\log[1 - D(\bar{V})]),
\end{aligned}
\quad (9)
$$

where $\bar{V}$ can be regarded as a transformation G of frames sequence θ. So the loss function can be rewritten as:

$$\mathcal{L}_{adv}(S, \bar{V}) = \mathbb{E}_S(\log D(S)) + \mathbb{E}_\theta(\log[1 - D(G(\theta))]), \quad (10)$$

where G tries to minimize $\mathcal{L}_{adv}$ against D that tries to maximize it, *i.e.,* $D^* = \text{argmin}_G \text{max}_D \mathcal{L}_{adv}(S, \bar{V})$ For better gradient of learning G, we actually minimize $\mathbb{E}_\theta(\log[-D(G(\theta))])$ instead of $\mathbb{E}_\theta(\log[1 - D(G(\theta))])$. Therefore, the final adversarial loss function is defined as:

$$\mathcal{L}_{adv}(S, \bar{V}) = \mathbb{E}_S(\log D(S)) + \mathbb{E}_\theta(\log[-D(G(\theta))]), \quad (11)$$

Combining with Eq. (5), we can obtain the optimization of our proposed method. For *Spatial Network*, the whole network can be divided into three parts: semantic generator G, classification layer C and semantic discriminator D. We adopt an alternating optimization to train all three parts mentioned above to avoid gradient vanishing problem caused by the minmax loss.

Firstly, semantic discriminator D is trained by minimizing Eq. (5). Update the parameters of D with G and C fixed:

$$D^* = \text{argmin}_D \mathcal{L}, \quad (12)$$

Algorithm 1 The optimization algorithm of SAL

Input: video dataset $\mathcal{V} = \{(\theta_k, y_k), k = 1, \ldots, K\}$
Output: semantic representation V, class score C
1: Initialized D, G, C; pretrained senmantic encoder
2: **repeat**
3:     Extract S by Eq. 8.
4:     Fixing parameters of G and D
5:     Update D by minimizing Eq. 5. ($S \to 1, \bar{V} \to 0$)
6:     Fixing parameters of G and D
7:     Update G, C by minimizing Eq. 5. ($\bar{V} \to 1$)
8:     Extract $\bar{V}$ and C
9: **until** convergence

Then, semantic generator G and classification layer C are trained by minimizing Eq. (5). Update the parameters of G and C with D fixed:

$$G^*, C^* = \operatorname*{argmin}_{G,C} \mathcal{L}. \tag{13}$$

The whole semantic adversarial learning module (SAL) is summarized in Algorithm 1. The *Temporal Network* has the same setting as *Spatial Network*.

Experimental Evaluations

In this section, evaluation datasets and implementation details used in experiments will be first introduced. Then we will study different aspects of our proposed modal to verify the effectiveness, respectively. Finally, we will make a comparison between our model with other RGB based state-of-the-art methods and provide a visualization of our experimental results.

Datasets and Implementation Details

Evaluation Datasets. In order to evaluate our proposed model, we conduct action recognition experiments on two popular video benchmark datasets: UCF101 (Soomro, Zamir, and Shah 2012) and HMDB51 (Kuehne et al. 2011). The UCF101 dataset are collected from the Internet, containing 13,320 videos which are divided into 101 classes. While the HMDB51 dataset are collected from the realistic videos, including movies and web videos, containing 6,766 videos which are divided into 51 action categories. We follow the officially offered scheme which divides dataset into 3 training and testing splits and finally report the average accuracy over the three splits. For *Spatial network*, we directly utilize RGB frames extracted from videos. For *Temporal Network*, the difference between adjacent frames is used to model temporal information of videos.
Implementation details. In generation procedure, we use stochastic gradient descent algorithm to train our *Spatial Network* and *Temporal Network*. The mini-batch size is set to 64 and the momentum is set to 0.9. The initial learning rate is set to 0.001 for *Spatial Network* and *Temporal Network* and decreases by 0.1 every 40 epochs.

For adversarial training procedure, we use adaptive moment estimation algorithm to train D Network and the initial learning rate is set to 0.0001. The training procedure of *Spatial Network* and *Temporal Network* stops after 80 epochs

Table 1: Performance comparison on HMDB51 split 1.

Method	Spatial	Temporal	Combine
TSN	52.0%	57.1%	62.5%
Ours	**54.4%**	**57.8%**	**63.7%**

Table 2: Performance comparison on UCF101 split 1.

Method	Spatial	Temporal	Combine
TSN	84.5%	85.4%	89.9%
Ours	**86.1%**	**86.8%**	**91.2%**

Table 3: Ablation study of our proposed method.

Dataset	Method	Network	Accuracy
HMDB51	Baseline	Spatial	52.0%
	MPA	Spatial	54.0%
	MPA + SAL	Spatial	**54.4%**
UCF101	Baseline	Spatial	84.5%
	MPA	Spatial	85.2%
	MPA + SAL	Spatial	**86.1%**

and 120 epochs respectively. We use gradient clipping of 20 and 40 for Spatial and Temporal training procedure to avoid gradient explosion. We train our model with 4 NVIDIA TITAN X GPUs and all the experiments are implemented under the Pytorch.

Results and Ablation Study

In this subsection, we will investigate the performance of our proposed method. The analysis for the performance of single and multiple modalities. All the results are trained with the same network backbone and strategies illustrated in previous sections for fair comparison.

We first evaluate the effectiveness of our proposed method. In this section, we compare the performance between ours and TSN (Wang et al. 2016) with the same experimental condition. The experiment is performed on HMDB51 split 1 and UCF101 split 1, the results are summarized in Table 1 and Table 2. As is shown in them, our method is superior than TSN both in spatial and temporal branch. In spatial branch, our method has about 2% improvement in performance compared to TSN, which proves the effectiveness of proposed MPA and SAL. but in temporal branch, it is a limited improvement because differential image has less visual elements than RGB.

In order to justify the effectiveness of our proposed MPA and SAL, we conduct an ablation study for them. All experiments in this ablation study are performed on the split 1 of HMDB51 and UCF101 by *Spatial Network*. The results are shown in Table 3. The *baseline* means the original network without multi-scale pyramid attention and semantic adversarial learning module. The *MPA* means add multi-scale pyramid attention layer to *baseline*. It has about 1%-2% improvement in performance compared to *baseline*, which can prove the effectiveness of proposed MPA. The *MPA+SAL* means add both multi-scale pyramid attention and semantic adversarial learning module. It has about 1% improvement

Table 4: Comparison with state-of-the-art methods on the UCF101 and HMDB51 datasets. The accuracy is reported as average over three splits. For fair comparison, we consider methods that use only RGB input. It can be seen our method obtains the best performance.

| Method	Pre-train	UCF101	HMDB51
HOG(Wang and Schmid 2013)	None	72.4%	40.2%
ConvNet+LSTM(Donahue et al. 2015)	ImageNet	68.2%	-
Two Stream Spatial Network(Simonyan and Zisserman 2014)	ImageNet	73.0%	40.5%
Conv Pooling Spatial Network(Feichtenhofer and Zisserman 2016)	ImageNet	82.6%	-
Spatial Stream ResNet	ImageNet	82.3%	43.4%
Spatial TDD(Wang, Qiao, and Tang 2015)	ImageNet	82.8%	50.0%
TSN Spatial Network(Wang et al. 2016)	ImageNet	86.4%	53.7%
TSN (RGB+RGB-Diff)(Wang et al. 2016)	ImageNet	91.0%	-
RGB-I3D(Carreira and Zisserman 2017)	ImageNet	84.5%	49.8%
CoViAR(Wu et al. 2018)	ImageNet	90.4%	59.1%
DCD(Zhao, Xiong, and Lin 2018)	ImageNet	91.8%	-
LTC(Varol, Laptev, and Schmid 2018)	Sports-1M	82.4%	48.7%
C3D(Tran et al. 2015)	Sports-1M	85.8%	54.9%
Pseudo-3D Resnet(Qiu, Yao, and Mei 2017)	ImageNet+Sports-1M	88.6%	-
C3D(Tran et al. 2015)	Kinetics	89.8%	62.1%
Ours	ImageNet	**92.7%**	**66.3%**
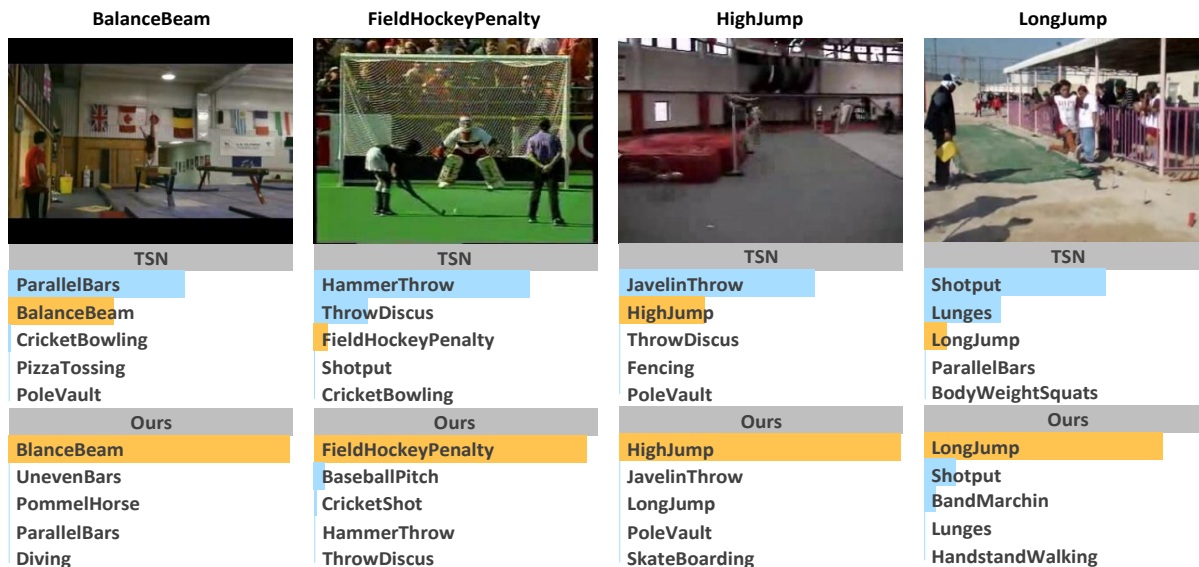
Figure 4: A comparison of top-5 predictions between TSN and our proposed method on UCF101 split 1. The tags on the top is the groundtruth labels. The yellow bars indicate correct classifications and the blue stand for incorrect cases. The length of each bar shows its confidence.

in performance compared to *MPA*, that can prove the effectiveness of the proposed SAL.

Comparison with the State-of-the-Arts

In this subsection we compare the classification performance of our approach with other state-of-the-art methods that take RGB frames as input. The experiment is conducted on two popular video action recognition benchmarks: UCF101 and HMDB51. The results are shown in Table 4, where we compare our method with traditional approach HOG and a series of deep learning based methods such as 3D convolutional

network, trajectory-pooled deep convolutional descriptors, temporal segment network and compressed video action recognition. It can be seen that our model achieves best results than other methods on these benchmark datasets, which can demonstrates the advantage of our proposed method and the effectiveness of multi-scale feature semantic modeling.

Visualization

In this subsection, we present a qualitative classification results. Figure 4 illustrate the comparison of top-5 predictions between TSN and our proposed method on UCF101 split 1.
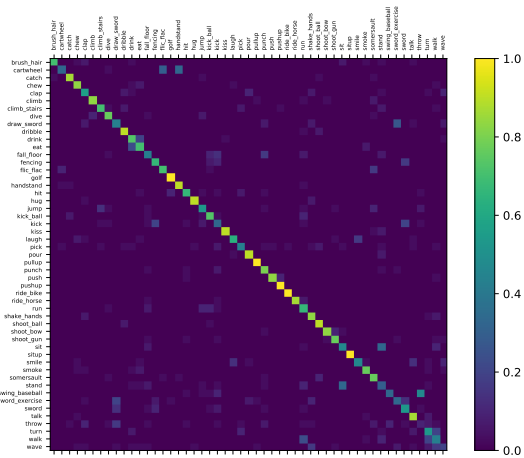
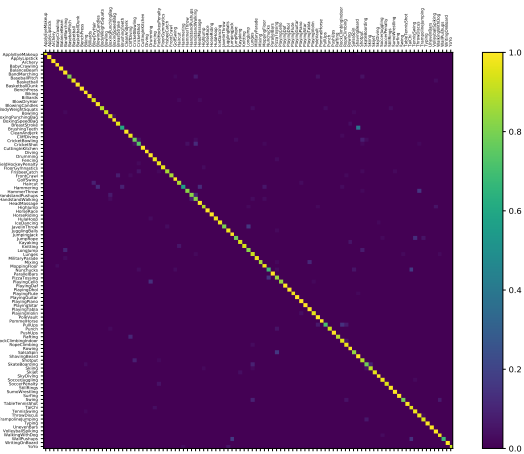Figure 5: Confusion matrix for HMDB51 dataset.

Figure 6: Confusion matrix for UCF101 dataset.

The results show that the original two-stream based methods (such as TSN) are easily fooled by common background. For instance, it regards BalanceBeam as ParallelBars, since the similar backgroud of gym and equipments. The reasons is that those methods fail to capture details and local context information effectively. The results base on our method can classify this actions will, which demonstrate the effectiveness of proposed MPA and SAL module. Figure 5 and Figure 6 show the confusion matrix for HMDB51 and UCF101, respectively. All the results can demonstrate the superior performance of our method.

Conclusion

In this paper, we proposed a new deep network architecture for video classification. The proposed architecture can only need RGB frames as input to model spatial information and temporal information for videos, which greatly reduce the computational cost compared with those methods using optical flow. In order to obtain more discriminative video representation, we design multi-scale pyramid attention (MPA) to

refine and merge different level features. Then we introduce semantic adversarial learning (SAL) module to guide learning procedure and to generate more discriminative semantic representation. Comprehensive experiment results on two popular benchmark datasets show that our method yields state-of-the-art performance in video classification tasks.

Acknowledgments

References

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 4724–4733. IEEE.

Dalal, N.; Triggs, B.; and Schmid, C. 2006. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, 428–441. Springer.

Deng, C.; Chen, Z.; Liu, X.; Gao, X.; and Tao, D. 2018. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing* 27(8):3893–3903.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

Fan, X.; Yang, Y.; Deng, C.; Xu, J.; and Gao, X. 2018. Compressed multi-scale feature fusion network for single image super-resolution. *Signal Processing* 146:50–60.

Feichtenhofer, A., and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1933–1941.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35(1):221–231.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2556–2563. IEEE.

Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; and Tao, D. 2018a. Self-supervised adversarial hashing networks for cross-modal retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, F.; Tian, C.; Zuo, W.; Zhang, L.; and Yang, M.-H. 2018b. Learning spatial-temporal regularized correlation filters for visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 5534–5542. IEEE.

René, C. L. M. D. F., and Hager, V. A. R. G. D. 2017. Temporal convolutional networks for action segmentation and detection. In *IEEE International Conference on Computer Vision (ICCV)*.

Shi, Y.; Tian, Y.; Wang, Y.; Zeng, W.; and Huang, T. 2017. Learning long-term dependencies for action recognition with a biologically-inspired deep network. In *Proceedings of the International Conference on Computer Vision*, 716–725.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, 568–576.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Varol, G.; Laptev, I.; and Schmid, C. 2018. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence* 40(6):1510–1517.

Wang, H., and Schmid, C. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, 3551–3558.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, 20–36. Springer.

Wang, H.; Yang, Y.; Yang, E.; and Deng, C. 2017. Exploring hybrid spatio-temporal convolutional networks for human action recognition. *Multimedia Tools and Applications* 76(13):15065–15081.

Wang, J.; Jiang, W.; Ma, L.; Liu, W.; and Xu, Y. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, L.; Qiao, Y.; and Tang, X. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4305–4314.

Wu, C.-Y.; Zaheer, M.; Hu, H.; Manmatha, R.; Smola, A. J.; and Krähenbühl, P. 2018. Compressed video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6026–6035.

Yang, Y.; Liu, R.; Deng, C.; and Gao, X. 2016. Multi-task human action recognition via exploring super-category. *Signal Processing* 124:36–44.

Yang, Y.; Deng, C.; Gao, S.; Liu, W.; Tao, D.; and Gao, X. 2017a. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Transactions on Multimedia* 19(3):519–529.

Yang, Y.; Deng, C.; Tao, D.; Zhang, S.; Liu, W.; and Gao, X. 2017b. Latent max-margin multitask learning with skelets for 3-d action recognition. *IEEE Trans. Cybernetics* 47(2):439–448.

Yang, E.; Deng, C.; Li, C.; Liu, W.; Li, J.; and Tao, D. 2018. Shared predictive cross-modal deep quantization. *IEEE Transactions on Neural Networks and Learning Systems* (99):1–12.

Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4694–4702.

Zhao, Y.; Xiong, Y.; and Lin, D. 2018. Recognize actions by disentangling components of dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6566–6575.