Spatial-Temporal Bottom-Up Top-Down Attention Model for Action Recognition

Jinpeng Wang and Andy J. Ma(✉)

Sun Yat-sen University, Guangzhou, China
majh8@mail.sysu.edu.cn

Abstract. Driven by the importance of capturing non-local information in video understanding, we propose Spatial-temporal Bottom-up Top-down Attention Module (STBTA). Features are processed across in multiple scales and then combined to best capture the spatial relationships associated with the region of interest and the surrounding environment in a complicated scene. Attention maps are used for adaptive feature refinement. STBTA can be plugged into any feedforward network architectures and is end-to-end trainable along with CNN. Extensive experiments on UCF101, HMDB51, Kinetics-400 datasets demonstrate that the proposed method can improve the performance for action recognition.

Keywords: Attention mechanism · Bottom-up top-down · Action recognition

1 Introduction

Non-local information is found to be of central importance for video understanding and image recognition [3,25]. By stacking a series of convolutional layers, CNN is capable of capturing non-local information [25]. However, each of the learned filters in a special layer operates in a local receptive field and consequently, each corresponding unit of the transformation output is unable to exploit global information outside of this local receptive field. This problem becomes more severe in the lower layers of the network [8].

Stacked Hourglass Networks (SHN) [14] repeats bottom-up, top-down processing with intermediate supervision to improve the performance of human pose estimation. A single pipeline with skip layers is used to preserve spatial information on each scale. Bottom-up top-down mechanism combines multi-scale information and filters operate in a non-local receptive field, can be considered as another way to capture non-local information. But videos/images own much irrelevant and background information [3]. Nevertheless, SHN considers multi-scale feature maps as the same without adaptive feature refinement.

Attention mechanism has been proven to be an efficient way to help the network see important parts and diminishes background responses [29]. On cognition theory, people focus sequentially on different parts of the scene to extract

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11901, pp. 81–92, 2019.
https://doi.org/10.1007/978-3-030-34120-6_7

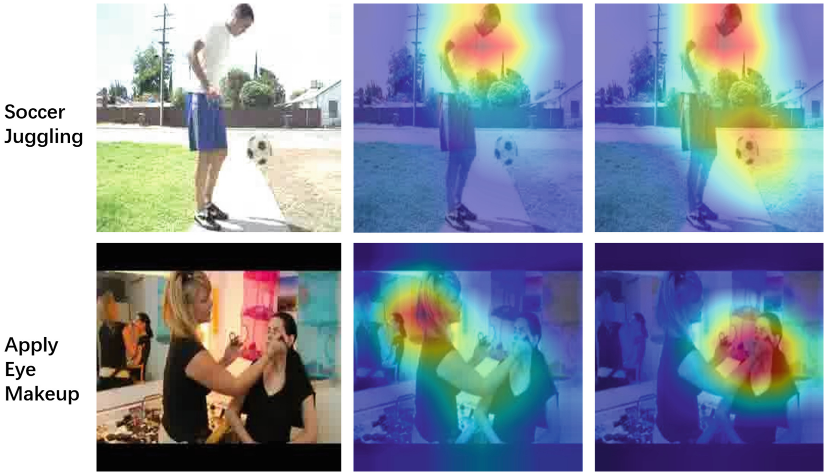


Fig. 1. Visualization of some samples on UCF101 using Grad-CAM [16]. The ground-truth label is shown on the left of each input image. We compare the visualization results of the STBTA network(STBTA + Inception v3) with baseline(Inception v3). The Grad-CAM visualization highlight the class-specific discriminative regions, which is calculated for the last convolutional outputs. These visualizations show STBTA network focus on target objects more properly

relevant information [13]. Attention mechanism has been shown to achieve promising results of image caption generation, machine translation, image recognition [22,23,29].

Our goal is to increase representation power by using Bottom-up Top-down mechanism and attention mechanism: capturing non-local information both in space and temporal and focusing on important features. In this paper, we design two efficient module: Spatial Attention Module (SAM) and Temporal Attention Module (TAM), which is different from existing attention module. Based on these modules, we propose Spatial Bottom-up Top-down Attention Module (STBTA) as an efficient and a general component for capturing non-local spatial dependencies and to obtain more discriminative attentional maps. As shown in Fig. 1, an STBTA-integrated network focus on class-discriminative objects more properly compared with baseline. There are several advantages of using STBTA. (a) STBTA can generate temporal-wise statistics and spatial grids statistics, which increases the sensitivity to informative features and choose useful information. (b) Our method can be considered as a general module which is feedforward fashion and can be inserted into any CNNs directly. (c) STBTA can improve the visual recognition performance efficiently.

2 Related Work

Attention Mechanism. Human perception does not tend to process the whole scene at once and focus selectively on parts of the visual space to acquire information when and where it is needed [13]. Soft attention developed in recent work can be trained end-to-end for convolutional neural network [23]. CBAM [27] emphasizes meaningful features along two principal dimensions: channel and spatial axes. In our model, we first propose a Spatial Attention Module I(SAM I) based on SE Net [8], then we design a new grid-wise spatial attention module II(SAM II) with depthwise convolution. Otherwise, driven by the intuition that different frame play different role for action recognition, we design a fully new temporal attention model.

Residual Network. Deep residual learning [7] is designed to learn residual of identity mapping. This method has proved to be an efficient way to prevent overfitting and increase the depth of the feedforward neuron network. Inception-Resnet architectures [18] showed that the network can achieve competitive accuracy by embedding multi-scale processes in the deep residual network. In our work, we use the residual connection to add different scale feature maps with origin feature maps together.

Multi-scale Fusion. The work in [20] uses multiple resolution banks in parallel and capture features at a variety of scales. Based on this method, bottom-up (from high resolutions to low resolutions) and top-down (from low resolutions to high resolutions) [14] is proposed to capture information at every scale. This approach uses a single pipeline with skip layers has the capacity to capture full body information and bring it to the next layer. Residual attention network [23] uses bottom-up top-down mechanism as attention mask. Our network design partly builds off of their work, exploring how to capture information across scales and adapting their method of combing features across different resolutions. Instead, we don't use intermediate supervision process and introduce attention mechanism which is different from previous work.

To the best of our knowledge, this is the first single-pipeline end-to-end feedforward attention module that encoding non-local information with bottom-up top-down mechanism about action recognition.

3 Proposed Method

STBTA: A STBTA net based on Inception-v3 [19] and TSN [24] for action recognition is illustrated in Fig. 2. All of these submodules in STBTA are residual modules and STBTA performs like a big residual block. For each STBTA, max pooling layer with stride 2 is used to process features down to a very low resolution. We use t to denote the number of downsample and upsample times of this paper, which is 1 default.

There exists a residual submodule between any adjacent layer during downsampling and upsampling (We have not visualized the residual submodule in

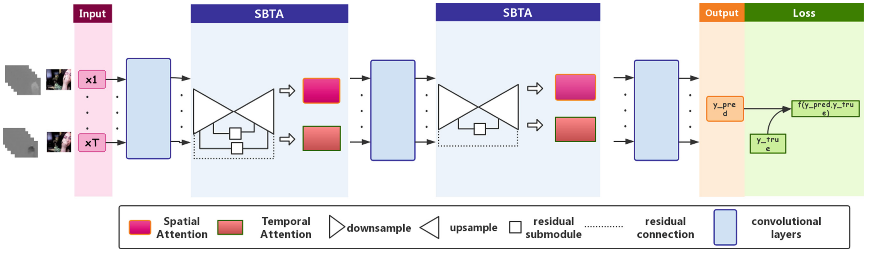
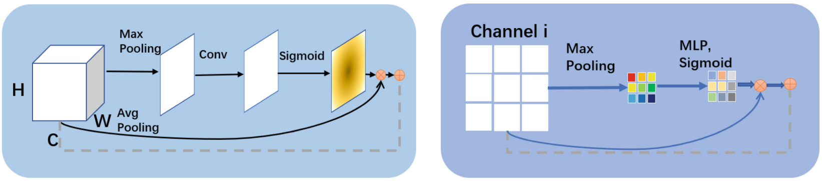


Fig. 2. A STBTA network based on Inception-v3 and TSN. The first STBTA with $t = 2$ is add after inception 3a. The second STBTA with $t = 1$ is add after inception 4e . T is the number of temporal segments, 3 in our experiment. t is the number of downsample and upsample times in STBTA

Fig. 2 for simplicity). The design of residual submodule is the same as SHN [14]. We downsample the input feature map several times in this module. After reaching the lowest resolution, the module begins the sequences of bilinear upsample and combination features across scales by a symmetrical top-down architecture. Furthermore, we add spatial and channel attention module to emphasize the features of key local regions and further improve the performance of the network. The output size is the same as the input feature map.

Global content information and temporal information are both important for action recognition. Most simple actions can be recognitioned by a few frames or a still frame. But for complicated actions, recognition highly rely on temporal information. Based on this, we design two branch which are added after upsample. The first branch is spatial attention module, which focus on spatial information and process on feature maps which combined all scale information. Spatial attention module is added after upsample to control computing cost. Only one channel attention module is added into the last part of STBTA for simplifying and process on all channels which combined all scale information.



(a) Spatial Attention Module I (b) Spatial Attention Module II

Fig. 3. The design of SAM I and SAM II.

Spatial Attention: Inspired by the design of channel attention recently [8]. For action recognition, we care about 'where' is an informative part, which is symmetric with the channel attention branch. The design of spatial attention module

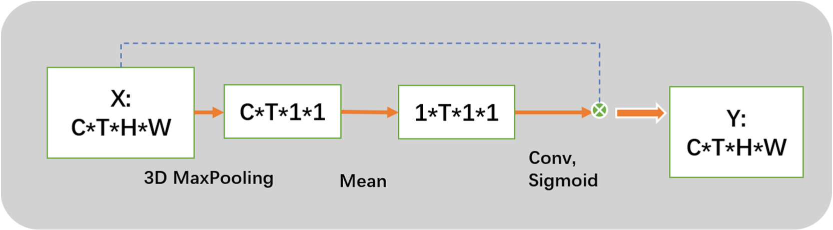


Fig. 4. Temporal attention module

has two ways. As shown in Fig. 3a. The first form is computed a 2D descriptor that encodes channel information at each pixel across the channel, which named Spatial Attention Module I (SAM-I). Formally, given an intermediate feature map $F \in \mathbb{R}^{C \times H \times W}$ as input, using channel max pooling and channel avg pooling, generate two 2D maps: $F^s_{avg} \in \mathbb{R}^{1 \times H \times W}$ and $F^s_{max} \in \mathbb{R}^{1 \times H \times W}$. Then do element-wise-addition between them and convolved by a standard convolution layer to produce 2D spatial attention map, sigmoid activate function is added in the last. Then we get spatial coefficients:

$$M_s(F) = \sigma(f^{conv}(F^s_{max})) \tag{1}$$

where f^{conv} represents a convolution operation and σ denotes the sigmoid function. Then $M_s(F)$ is multiplied with each channel and add with origin feature map to get the output.

Fusion channel may weaken distinguish information, so we design spatial attention module in a new way. In the second form, the spatial dimension is $W \times H$ for every channel. We divide every channel into $N \times N$ grids, N is chosen to be 3 in our experiments. Max pooling is performed with each grid, and then a conv layer and one softmax activation function are used to produce coefficients for these grids. We use depthwise separable convolution here to not change channel dependence. Which named Spatial Attention Module II(SAM-II). The details of SAM-II are in Fig. 3b.

Temporal Attention: Intuitively, every temporal information play different role for action, some temporal information may be key frame which has high distinction. Inspired by this intuitively, Temporal attention module(TAM) mainly consider relations along temporal dimensions. First, we reshape the input feature map as $B \times C \times T \times H \times W$. As shown in Fig. 4(batch size B is not shown for conveniently). Notice 2D CNNs without temporal sampling is a special situation, when $T = 1$. Firstly use 3D max pooling to get max response, then we calculate mean along channel dim, following with conv layer and sigmoid activation function too. Then we use the output to re-weight the input feature map. The benefit of this design is the computation overhead is negligible and strengthen the key information along temporal dim.

4 Experiments

4.1 Experiments Setup

We use the PyTorch framework for CNN implementation and all the networks are trained on 4 NVIDIA 1080Ti GPUs. Here, we describe the datasets and implementation details.

Datasets. Three well-known benchmarks, UCF101 [17], HMDB-51 [11] and Kinetics-400 [10] are used in the evaluations of action recognition. UCF101 consists of 13,320 manually labeled videos from 101 action categories. It has three train/test splits, each split has around 9,500 videos for training and 3,700 video for testing. HDMB51 is a realistic and challenging dataset. It consists of 6,766 manually labeled clips from 51 categories. Kinetics-400 contains around 246K training videos and 20k validation videos from 400 categories.

Implement Details. For 2D networks, all of our network are based on TSN [24]. To conduct fair comparison, we keep most of the settings same as TSN. Random cropping and horizontal flipping are used for data augmentation. We train network by using the SGD optimizer with a mini-batch size of 64. The learning rate drops down by 10 every 30 epochs and we set the dropout radio at 0.7 to prevent over-fitting. We use a weight decay of 0.0005 with a momentum of 0.9 and set the initial learning to 0.001. The spatial size is 224 × 224 pixels. We train our module for 100 epochs. In the resting stage, 25 segments are sampled from RGB and optical flow. For 3D networks, we add our module on 3D Inception-v1 [1] and 3D ResNext-101 [6]. For 3D Inception-v1, we follow the design in [1]. What's different is in our practice we sample 10 clips randomly from a full-length video and compute the softmax scores, the final result is averaged of these scores. For 3D ResNext-101, we follow the implement details as [6] to conduct fair comparison. We choose ResNext as the back bone because the good performance. What's different is that we use fine-tune strategies which be describe in Sect. 4.3.

Table 1. Ablation study on our proposed module. We show RGB top-1 classification accuracy on split 1 of UCF-101.

Method	BNInception
baseline	84.30%
baseline + SAM-I	84.63%
baseline + SAM-II	85.02%
baseline + TAM	84.71%
basline + SAM-I + TAM	85.27%
baseline + SAM-II + TAM	85.66%

4.2 The Efficiencies of STBTA

First we add our proposed module on BNInception [9], the ablation study result is shown in Table 1. Both SAM and TAM can improve the recognize performance and combine them lead to better result.

Table 2. We compare 1, 2 STBTA be added to the BNInception(the first with $t = 2$ before inception (3c) and the second with $t = 1$ before inception (4d)), Inception-v3(the first with $t = 3$ before mixed_5b and the second with $t = 2$ before mixed_7a) and Inception-Resnet-v2(the first with $t = 3$ before mixed_5b and the second with $t = 2$ before mixed_7a). We show RGB top-1 classification accuracy on split 1 of UCF-101

Method	BNInception	Inception-v3	Inception-Resnet v2
baseline	84.30%	84.88%	86.49%
+ 1 STBTA	85.27%	85.93%	87.95%
+ 2 STBTA	85.76%	86.59%	88.44%

In order to show the efficiencies of STBTA, we use BNInception [9], Inception-v3 [19] and Inception-Resnet-v2 [18] as baseline and all pretrained on ImageNet. Table 2 shows the results of different number of STBTA be added to the baseline. A network with STBTA leads to a better result in general. It is noteworthy that add one STBTA lead to 1% improvement generally. Considering calculation overhead, we add 2 STBTA to baseline in this paper as default. Furthermore, to demonstrate our module's general applicability. We use our STBTA on Kinetics-400, which is two orders of magnitude larger than HMDB51 and UCF101 and is very time-consuming to train. Limited to the hardware resources, we only one STBTA with $t = 3$ on Inception-v3(before mixed_5b). The result is shown in Table 3. In Table 4, we list some recently comparable methods. Our result is based on Inception-Resnet-v2 baseline(the first STBTA with $t = 3$ is added before mixed_5b and the second STBTA with $t = 2$ is added before mixed_7a), we call this STBTA net. Only use RGB frame as input and pretrained on ImageNet, our method outperforms MiCT-Net by 1.4% on UCF101. In addition, use SAM-II, we can obtain an extra gain about 0.4% but time-consuming. We use SAM-I in STBTA as default in rest.

Table 3. We show video top-1 classification accuracy for RGB input on Kinetics-400. Report on the val sets.

Method	Inception V3
baseline	72.5%
+ 1 STBTA	73.7%

Table 4. Performance comparison to the state-of-the-arts methods on UCF-101 over three splits for RGB as input.

Method	RGB
TSN [24]	86.01%
I3D [1]	84.5%
MiCT-Net [30]	87.3%
STBTA net (SAM-I)	88.70%
STBTA net (SAM-II)	**89.10%**

4.3 Fine-Tune Strategy

Due to the large number of 3D ConvNets's parameters, small datasets can be easily over-fitting. One would fine-tune existing networks that are trained on Kinetics or Sports1M. There are three general guidelines for fine-tuning if new dataset is similar to the original dataset. The first common practice is to truncate the last layer. The second common practice is to use a smaller learning rate to train all the network. The third method is to freeze the weights of the first few layers and train others later.

A general solution is the first few layers capture universal features like curves and edges. But ignore data imbalanced totally. In this paper, we propose a new engineering strategy to fine-tune neural networks. Give different learning rates, according to the depth of neural networks, achieve an impressive performance advancing. Which be formulated with.

$$\beta_l = \sin\left(\frac{l}{L} * \frac{\pi}{2}\right) * \alpha \tag{2}$$

L is the network's depth, l is current layer's depth. α is the learning rate now. $\beta_l, l = 1, 2...L$ is the learning rate of the l layer.

Table 5 show the results of fine-tune strategy and a single STBTA added to 3D Inception-v1. We inflated a 2D Inception-v1 follow [1] and pretrained on Kinetics-400. Fine-tune strategy can lead to 1% improvement over the baseline. And with additional 1 STBTA can further lead to 0.8% improvement.

Table 5. We show top-1 result based on 3D Inception-v1. Report on the split1 of UCF101.

Method	3D Inception-v1
baseline	92.72%
+ fine-tune strategy	93.75%
+ 1 STBTA	94.55%

In order to show our method's effectiveness, we visualize several examples for the behavior of a SBTA be added to the baseline in Figs. 5 and 1. Our module

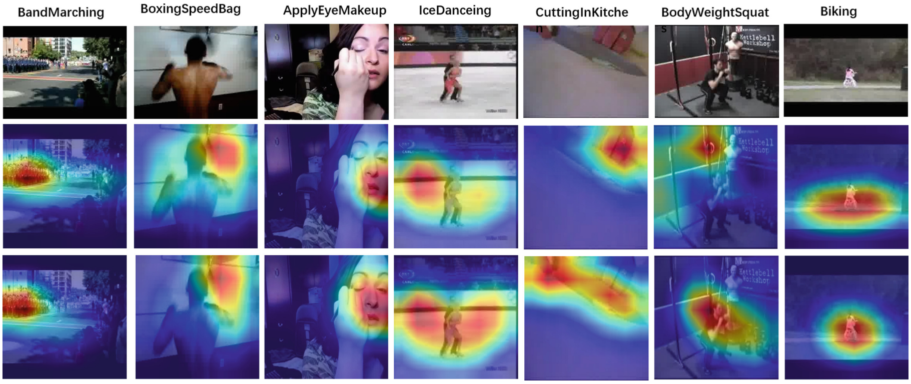


Fig. 5. We compare the visualization results of SBTA-integrated network(Inception-v3+SBTA) with baseline(Inception-v3). All the convnets are based on TSN. The grad-CAM visualization is calculated for the convolutional outputs after Mixed7D. The first row is input image, the second row is baseline's results and the third row is our SBTA-integrated network's results.

Table 6. Comparisons with state-of-the-art results on UCF101 and HMDB51 over 3 splits.

Method	UCF101	HMDB51
TSN [24]	94.0%	68.5%
ST-ResNet [4]	93.5%	66.4%
TLE [2]	**95.6%**	71.1%
Attention Cluster [12]	94.6%	69.2%
STP [26]	94.6%	68.9%
Two Stream MiCT-Net [30]	94.7%	70.5%
ActionVLAD [5]	92.7%	66.9%
CoViAR + optical flow [28]	94.9%	70.2%
ISPAN(30 frames) [3]	95.5%	70.7%
Two Stream STBTA Net	95.20%	**71.1%**

can learn to find meaningful relational clues in long distance and pas attention to more specific and accurate action regions in every frame.

4.4 Comparison with 2D State-of-the-Arts

To prove the effectiveness, we further evaluate our STBTA net on all 3 splits of UCF-101 and HMDB-51 with only use ImageNet pre-trained in Table 6. We list recent state-of-the-art and comparable methods. Two stream STBTA net obtain the improved performance 95.2%/71.1%, which is on pair with TLE. It can be noticed that our proposed STBTA's performance is better on HMDB51 (a hard dataset). Note that the two-stream architecture numbers on individual RGB and Flow streams can be interpreted as a simple baseline, which applies a ConvNet independently on 25 uniformly sampled frames then average the predictions.

4.5 Comparison with 3D State-of-the-Arts

In Table 7, we compare 3D state-of-the-arts method on UCF101 and HMDB51 with only RGB as input. ResNext-101 are pre-trained on Kinetics-400. Our 3D STBTA obtain an extra gain about 1.3% on UCF101 and about 1.2% on HMDB51. The reason why STBTA's result on HMDB51 isn't competing with UCF101 may be HMDB51's samples is too small for 3D ConvNets.

Table 7. Comparisons with state-of-the-art results on UCF101 and HMDB51 over 3 splits.

Method	UCF101	HMDB51
C3D [21]	82.3%	–
RGB-I3D(64f) [1]	95.6%	74.8%
P3D Resnet + IDT [15]	93.7%	–
ResNext-101(64f) [6]	94.5%	70.2%
ResNext-101(64f) [6] + STBTA	95.8%	71.4%
+ fine-tune strategy	96.0%	72.2%
RGB-I3D(64f) [1] + STBTA	96.1%	75.4%
+ fine-tune strategy	**96.3%**	**75.8%**

5 Conclusions

We propose a novel Spatial Bottom-up Top-down Attention Module (STBTA), which can encoding non-local information and achieve adaptive feature refinement via Bottom-up Top-down and attention mechanism. Experimental results show that the proposed module can improve the recognition performance for the task of video classification. Even a simple addition of one STBTA in a baseline CNN can achieve significant improvement over the baseline.

For the future work, we will exploit different applications of our module such as action detection and image segmentation to better explore Bottom-up Top-down mechanism and attention mechanism for different tasks.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR, pp. 4724–4733. IEEE (2017)

2. Diba, A., Sharma, V., Van Gool, L.: Deep temporal linear encoding networks. In: CVPR, vol. 1 (2017)
3. Du, Y., Yuan, C., Li, B., Zhao, L., Li, Y., Hu, W.: Interaction-aware spatio-temporal pyramid attention networks for action classification. arXiv preprint arXiv:1808.01106 (2018)
4. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Advances in Neural Information Processing Systems, pp. 3468–3476 (2016)
5. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: learning spatio-temporal aggregation for action classification. In: CVPR, vol. 2, p. 3 (2017)
6. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 7 (2017)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
10. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
11. Kuehne, H., Jhuang, H., Stiefelhagen, R., Serre, T.: HMDB51: a large video database for human motion recognition. In: Nagel, W., Kröner, D., Resch, M. (eds.) High Performance Computing in Science and Engineering 2012, pp. 571–582. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-33374-3_41
12. Long, X., Gan, C., de Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: purely attention based local feature integration for video classification. In: CVPR, pp. 7834–7843 (2018)
13. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: NIPS, pp. 2204–2212 (2014)
14. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
15. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: ICCV (2017)
16. Selvaraju, R.R., et al.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)
17. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
18. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR, pp. 2818–2826 (2016)
20. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NIPS, pp. 1799–1807 (2014)
21. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks
22. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)

23. Wang, F., et al.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017)
24. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
25. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
26. Wang, Y., Long, M., Wang, J., Philip, S.Y.: Spatiotemporal pyramid network for video action recognition. In: CVPR, vol. 6, p. 7 (2017)
27. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
28. Wu, C.Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A.J., Krähenbühl, P.: Compressed video action recognition. In: CVPR, pp. 6026–6035 (2018)
29. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML, pp. 2048–2057 (2015)
30. Zhou, Y., Sun, X., Zha, Z.J., Zeng, W.: MiCT: mixed 3D/2D convolutional tube for human action recognition. In: CVPR, pp. 449–458 (2018)