UTS Submission to ActivityNet Challenge 2017

Linchao Zhu Yi Yang
University of Technology Sydney
{zhulinchao7, yee.i.yang}@gmail.com

Abstract

We present our submission to the trimmed action recognition task in ActivityNet Challenge 2017. We mainly train the two-stream ConvNets and propose a multi-scale attention model to learn a compact video representation. We found that the multi-scale attention model outperforms average pooling on the Kinetics dataset.

1. Our Approach

We followed the basic two-stream ConvNets for video classification [2]. The ResNet-101 and ResNet-152 [1] architectures are used.

In our multi-scale attention model, we introduce a memory block $\mathbf{C}$ with shape (m, n), where m is the number of memory slots and n is the memory size. $\mathbf{C}$ will be updated with activations from different convolutional layers. The update procedures are as follows. We denote $\mathbf{X}_i$ as the activations of the ith convolutional layer for the given video, $\mathbf{C}_i$ is the center at iteration i. $\mathbf{X}_i$ is first transformed to $\mathbf{X}_i{}'$ by

$$\mathbf{X}_i{}' = \text{ReLU}(\text{layer\_norm}(\mathbf{W}_0\mathbf{X}_i)). \quad (1)$$

$\mathbf{X}_i{}'$ is then been blended into memory $\mathbf{C}_i$ through attention mechanism. The updated memory $\mathbf{C}_{i+1}{}'$ is obtained by

$$\begin{aligned}\text{Attend}(\mathbf{X}, \mathbf{C}) &= \text{normalize}(\sum \text{softmax}(\mathbf{X}\mathbf{C}^T)(\mathbf{X} - \mathbf{C})),\\ \mathbf{C}_{i+1}{}' &= \text{Attend}(\mathbf{X}_i{}', \mathbf{C}_i),\end{aligned} \quad (2)$$

where the normalization function is SSR normalization followed intra normalization and ℓ_2 normalization. $\mathbf{C}_{i+1}{}'$ is then transformed to the next memory $\mathbf{C}_{i+1}$ with

$$\mathbf{C}_{i+1} = \text{ReLU}(\text{layer\_norm}(\mathbf{W}_1\mathbf{C}_{i+1}{}')). \quad (3)$$

After iterations N, $\mathbf{C}_N$ is flattened and used for classification.

2. Experiments

To train the RGB network, we initialize the weights from the pre-trained ImageNet models. The flow net is initialized with the trained RGB model [3]. We used SGD with momentum 0.9, and the initial learning rate is 0.01. The batch size is set to 256 for ResNet-101 and 128 for ResNet-152. The learning rate decays 0.1 every 200,000 iterations.

Model	RGB	Flow	RGB+Flow
ResNet V1 101	69.9 / 88.2	60.5 / 81.1	72.6 / 89.8
ResNet V2 152	70.2 / 88.5	–	–
Multi-scale Attention	**71.1 / 90.0**	–	–

Table 1. Single checkpoint, single scale performance.

Model	Validation	Test
RGB models	72.4 / 90.4	–
RGB+Flow	74.2 / 91.0	–

Table 2.

To train the multi-scale attention network, we randomly sample 16 frames from a video and 1 / k positions are sampled from the feature map with size (k, k). We used 3 feature maps from ResNet-101, which are the activations from block 2, block 3, and block 4. The results are shown in Table 1.

We fused all RGB models by average fusion. The RGB scores and flow scores are also averaged to obtain the final scores. The fusion results are shown in Table 2.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[2] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[3] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.