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 $C$ with shape $(m, n)$, where $m$ is the number of memory slots and $n$ is the memory size. $C$ will be updated with activations from different convolutional layers. The update procedures are as follows. We denote $X_i$ as the activations of the $i$th convolutional layer for the given video, $C_i$ is the center at iteration $i$. $X_i$ is first transformed to $X_i'$ by

$$X_i' = \text{ReLU}(\text{layer\_norm}(W_0 X_i)).$$

(1)

$X_i'$ is then been blended into memory $C_i$ through attention mechanism. The updated memory $C_{i+1}'$ is obtained by

$$C_{i+1}' = \text{Attend}(X_i', C_i),$$

(2)

where the normalization function is SSR normalization followed intra normalization and $\ell_2$ normalization. $C_{i+1}'$ is then transformed to the next memory $C_{i+1}$ with

$$C_{i+1} = \text{ReLU}(\text{layer\_norm}(W_1 C_{i+1}')).$$

(3)

After iterations $N$, $C_N$ 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.

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

