Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition

Liangjie Cao

Aug. 7, 2018

1. Using Optical Flow Guided Feature in Convolutional Neural Network

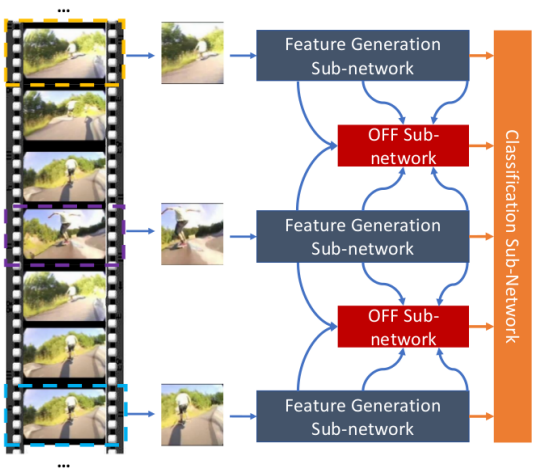

Figure 1. Network architecture overview.

Figure 1 shows an overview of the whole network architecture. The network consists of three sub-networks for different purposes: feature generation sub-network, OFF sub-network and classification sub-network. The feature generation sub-network generates basic features using common CNN structures. In the OFF sub-network, the OFF features are extracted using the features from the feature generation sub-network, and then several residual blocks are stacked for obtaining the refined features. The features from the previous two subnetworks are then used by the classification sub-network for obtaining the action recognition results. The Figure 2 exhibits the more detailed network structure with the inputs of two segments. As shown in Figure 2, we extract features from multiple layers on a specific level with the same resolution by concatenating them together and feed them into one OFF unit. The whole network has 3 OFF units with different scales. The details about the structure of each subnetwork is discussed as follows.

The OFF unit can be applied for CNN layers on different

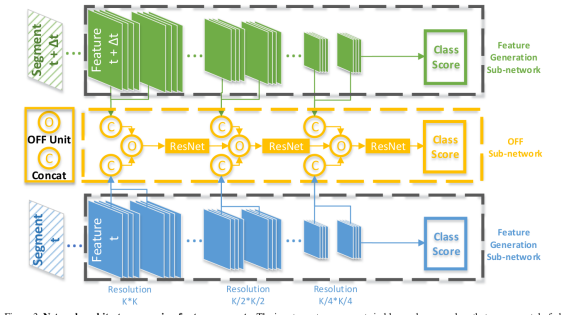

Figure 2. Network architecture overview for two segments

levels. The inputs of one OFF unit include the basic deep features from two segments, and the feature from the OFF unit on the previous feature level if it exists. In this way, the OFF at the previous semantic level can be used for refining the OFF at the current semantic level. Figure 3 shows the detailed implementation the OFF layer.

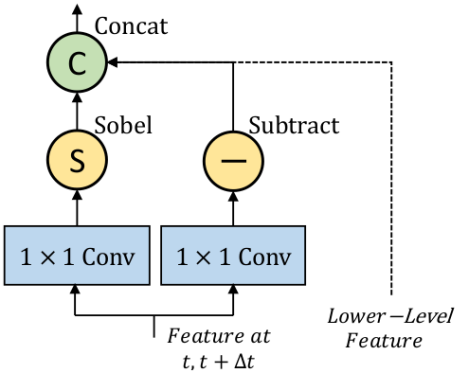

Figure 3. Detailed architecture of OFF unit.

2. Network Training

Action recognition is treated as a multi-class classification problem. Followed by the settings in TSN, as there are multiple classification scores produced by each segment,

they need to fuse them all in each sub-network separately to generate a video-level score for loss calculation. Here, for the OFF sub-networks, the features produced by the output of OFF sub-network for the tth segment on level l is denoted by $F_{t,l}$. The classification score for segment t on the level l using $F_{t,l}$ is denoted by $G_{t,l}$. The aggregated video-level score at level l is denoted by G_l . The video-level action classification score G_l is obtained by:

$$G_l = G(G_{0,l}, ...G_{N_{t-1}-1,l}) \quad (1)$$

where N_t denotes the number of frames for extracting features. The aggregation function denoted by G is used for summarizing the scores predicted from different segments along time. Following the investigations in TSN, G is imple- mented by average pooling for better performance [1]. As for the feature generation sub-network, the above equations are also applicable. While as they does not need intermediate supervision for feature generation sub-network, the feature $F_{t,l}$ at level l for segment t is simply equivalent to the final feature output of the sub-network.

Reducing the Memory Cost. As their framework consists of several sub-networks, it costs more memory than the original TSN framework, which extracts and stores motion frames before training CNNs, and trains several networks independently. In order to reduce the computational and memorial cost, they samples less frames in the training phase than in the testing phase, and still obtain satisfactory results.

3. Experiment

In this section, datasets and implementation details used in experiments will be first introduced. Then they will explore the OFF and compare it with other modalities under current state-of-the-art frameworks. Moreover, as their method can be extended to other modalities such as RGB difference and optical flow, we will show how such a simple operation could improve the performance for input with different modalities. Finally, we will discuss the meaning and difference between the OFF and other motion modalities such as optical flow and RGB difference.

	RGB	Hyp-Net + RGB	OFF(RGB) + RGB
Acc.	85.5%	86.0%	90.0%

Table 1. Experimental results of accuracy for hypercolumn network and the comparison with OFF on UCF-101 Split1. The denotation Hyp-Net indicates the output of hypercolumn network.

From the experimental results shown in Table 3, it is clear that, despite the hypercolumn network could get a slight 0.5% improvement on UCF-101 split 1, its final accuracy is still apparently less than the one obtained by OFF(RGB). Therefore, a conclusion could be drawn that it is the OFF calculation rather than the hypercolumn structure that plays the key role in achieving the significant gain.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2