architecture

Yu-Cheng Lin

December 2014

Two-Stream Convolutional Networks for Action Recognition in Videos

In this work the authors use multi-frame dense optical flow to train the temporal convolutional network (ConvNet). Leveraging on the temporal motion information, the temporal ConvNet performs much better than the spatial pre-trained ConvNet. In addition, they show that multi-task learning with different loss layers can boost the amount of training data and raise the performance on each task. Finally, they fuse the spatial and temporal ConvNet to a two-stream ConvNet model, which gets the best performance with SVM adopted as the fusion method.

First, the spatial ConvNet uses the video frames for action recognition. This is basically the same as image classification, on which the Alexnet architecture has already gotten great results. Hence the authors adopt the network pre-trained on the ImageNet challenge dataset. The input of the spatial ConvNet is WxHx3, with the image aspects W = H = 224. The authors compare the two methods: fine-tuning the whole network and training the last layer only. Since the performance difference is little, the authors opt for training the last layer only with the pre-trained ConvNet in the following experiments.

Second, the temporal ConvNet takes as input the multiple-frame dense optical flow. The displacement vector

$$d_t(u,v)$$

, for an arbitrary point (u, v) in frame t, represents the motion vector in the horizontal and vertical directions between two consecutive frames. The authors adopt the popular method of [2] to generate the displacement vectors by minimizing the energy with assumptions for

gradient constancy and displacement smoothness. The dense optical flow comprises of L consecutive frames. To get the best performance, the input of the temporal ConvNet is WxHx2L, with the image aspects W = H = 224 and L = 10. Since the temporal ConvNet is trained from scratch, multitask training is done on the UCF-101 and HMDB-51 datasets to avoid overfitting. Two softmax classification layers are added on top of the last fully connected layer. The overall training loss is calculated by summing up the both tasks' losses, which are active for the respective dataset.

Finally, the two-stream ConvNet is achieved by fusing the softmax scores from both the spatial and temporal ConvNets. There are mainly two fusion methods: averaging the softmax scores from both streams and training a multi-class linear SVM with the softmax scores as training features. From experimental results, we can see that the fusion ConvNet outperforms both the single spatial or temporal ConvNet, which proves that these two streams are complementary. In addition, the SVM fusion method gets the better performance than the averaging one.

Siamese Network

The use of the Siamese network is to learn an invariant mapping between a pair of images [6]. This invariant mapping is achieved by letting the similar pairs have small Euclidean Distance in the feature space [7]. The Contrastive Loss Function proposed in [6] takes this as a training target: not only pulling the similar pairs altogether, but also pushing apart the dissimilar pairs at least the margin distance m. The loss function is shown below:

$$L(x, y, l) = sD^2 + (1 - s)max^2(0, (m - ||D||_1))$$

where

$$s \in \{0, 1\}$$

is the similarity label. The label s is equal to 1 for similar pairs; s is equal to 0 for dissimilar pairs. The margin distance

$$m > 0$$

is used for the component of dissimilar pair loss calculation.

$$||D||_1 = ||f(x) - f(y)||_1$$

is the Euclidean Distance between the two input image pairs f(x) and f(y) in the feature space. The square of the max function makes it a differentiable function, which is more appropriate for back propagation in training.

Our Approach

In order to train the relationships between data of different views and action recognition at the same time, we train with contrastive loss and softmax loss layers, shown in Fig. 1. The both networks, CNN and CNNp, share the same parameters when training. The backbone CNN network is modified from the popular deep CNN model, Caffe's Imagenet pre-train model [5], which adopts the same architecture as AlexNet shown in Fig. 3 [1]. We fine-tune this pre-trained model by setting the learning rate as

$$10^{-4}$$

.

Our goal is to learn deep representations that capture the similarity between videos from different views. The input is composed of similar and dissimilar image pairs. Similar image pairs are two views of videos with the same objects shot at the same time; dissimilar image pairs are objects shot at different time with different actions. Since most frames are similar in

a second, to reduce computation, the trained/test frames are picked 1 from every 10 frames. For each input frame, dissimilar pairs are randomly selected from different class videos (the ratio of similar and dissimilar pairs is $1 : 20$).

While testing, only one CNN network is used, as shown in fig. 2. The video action recognition accuracy is calculated by averaging the estimated probabilities from input frames in a video.

[1] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. 2012. Imagenet classification with deep convolutional neural networks. In NIPS. [2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In Proc. ECCV, pages 25-36, 2004. [5] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the ACM International Conference on Multimedia (pp. 675-678). ACM. [6] Hadsell, R., Chopra, S., LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In Computer vision and pattern recognition, 2006 IEEE computer society conference on (Vol. 2, pp. 1735-1742). IEEE. [7] Lin, Tsung-Yi, et al. "Learning Deep Representations for Ground-to-Aerial Geolocalization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.