Papers about Two-stream ConvNet reading

Miaoran Chen

2019/5

Two-Stream CNN for Action Recognition in Videos（2014）

Main points

① Proposed a deep video classification model incorporating separate spatial and temporal recognition streams based on ConvNets.
② Employed multi-task learning on HMDB-51 and UCF-101 to increase the effective training set.
③ Demonstrated that a ConvNet trained on multi-frame dense optical flow is able to achieve very good performance in spite of limited data.

The structure of Two-stream

Video includes spatial and temporal information. The spatial part, in the form of individual frame appearance, carries information about scenes and objects depicted in video. The temporal part, in the form of motion across the frames, conveys the movement of the observer(camera) and the objects.[Sect.2 of the paper]. This is the theoretical basis of the Two-stream structure. The input videos are divided into 2 streams, each of them is implemented using a deep ConvNet, softmax of which are combined by late fusion. The structure of Two-stream is shown below.

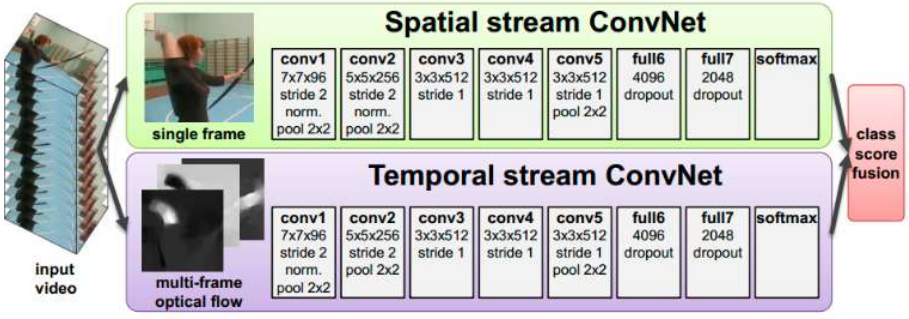

Figure 1: **Two-stream architecture for video classification.**

Spatial stream ConvNet

This ConvNet operates on individual video frames. Actually there are lots of representative classification networks could deal with this problem. In this paper, the researchers pre-trained on ImageNet ILSVRC-2012 and migrated the parameters.

Optical flow ConvNets

Optical flow stacking

Calculating the optical flows(the displacement vector fields) between the pairs of consecutive frames and simply stack them. And divide the optical flow into horizontal and vertical components, which can be seen as 2 image channels.

Trajectory stacking

Sampling at the same locations across several frames with the flow. Also, the optical flows are divided into vertical and horizontal components.

Bi-directional optical flow

The former methods deal with the forward optical flow. Bi-directional optical flow deals with both forward flows and backward flows.

Mean flow subtraction

To compensate the effect caused by camera movement, subtract the mean vector of each displacement field.

Multi-task learning

The temporal ConvNet needs to be trained on video data and the size of available dataset is not big enough.

Multi-task learning is used to deal with this problem. The researchers used 2 datasets: HMDB-51 and UCF-101. And there are 2 softmax classification layers on top of the last fully-connected layer to compute the scores of 2 datasets respectively. The ultimate loss is the summation of 2 losses.

Training

Dataset

HMDB-51, UCF-101

Spatial ConvNets

3 methods were considered.

① Training from scratch on UCF-101(×).

② Pre-training on ILSVRC-2012 followed by fine-tuning on UCF-101（√）.

③ Keeping the pre-trained network fixed and only training the last layer（√）.

Temporal ConvNets

Optical flow stacking +mean subtraction is best

Multi-task learning

Multi-task learning on HMDB-51 and UCF-101 is best. (Compared to fine-tuning a ConvNet or adding classes to UCF-101).

Possible improvement

① The spatial pooling in Two-stream does not take the trajectories into account.
② A more effective way to handle camera motion.

Towards Good Practices for Very Deep Two-Stream ConvNets(2015)

The author of this paper held confidence in Two-stream ConvNet has 2 major shortcomings. First, Two-stream ConvNet is relatively shallow(5 convolutional layers, 3 fully-connected layers) compared with those very deep models in image domain and the modeling capacity is constrained by the depth. Second, the training dataset of action recognition is limited(Multi-task learning is a method to deal with this problem proposed in the former paper). Thus over-fitting is easy to happen.

This paper focus on address these problems.

Structure of Very Deep Two-stream ConvNet

The researchers chose 2 network structures to design very deep two-stream ConvNet: GoogLeNet and VGGNet.

GoogLeNet

A 22-layer with occasional max-pooling layers with stride 2.

Inception module: composed of multiple convolutional filters with different sizes alongside each other.

1×1 convolutional operation: used for dimension reduction and computational efficiency.

VGGNet

A (up to)19-layer network with 3×3 convolutional kernel, 1×1 convolutional stride, 2×2 pooling window.

VGG-16 and VGG-19 are most representative ones.

Very deep Two-stream ConvNet

Input of spatial net: singe frame image(224×224×3)

Input of temporal net: 10-frame stacking of optical flow fields(224×224×20)

Advancement

Pre-trained for two-stream ConvNet

① Extracting optical flow fields into interval of [0,255] by a linear transformation.
② Averaging the ImageNet model filters of first layer across the channel and copying the average results 20 times as the initialization of temporal nets.

Smaller learning rate

For temporal net: learning rate starts with 0.005, decreases to its 1/10 every 10,000 iterations, stops at 30,000 iterations.
For spatial net: learning rate starts with 0.001, decreases to its 1/10 every 4,000 iterations, stops at 10.000 iterations.

Data Augmentation

① Only cropping 4 corners and 1 center of the images to reduce the effect of over-fitting.
② Employing multi-scale cropping method for training very deep two-stream.

High Dropout Ratio

For temporal net: 0.9 and 0.8 drop out ratios for the fully-connected layers
For spatial net: 0.9 and 0.9 drop out ratios for the fully-connected layers.

Result

① The deeper the network, the better the performance. Very deep two-stream ConvNet outperforms Two-stream ConvNet and is better than the best result by 3.4%
② Very deep Two-stream works well on dealing with overfitting because it uses ImageNet to pre-train the temporal net and uses more data augmentation techniques.

Temporal Segment Networks: Towards Good Practices for Deep Action Recognition（2016）

The author of this paper proposed that Two-stream ConvNet has 2 major shortcomings. First, mainstream ConvNets frameworks usually focus on appearances and short-term motions, thus lacking the capacity to incorporate long-range temporal structure. Second, training datasets are limited in diversity and size.

This paper focus on dealing with 2 problems: 1) how to design an effective and efficient video-level framework for learning video representation that is able to capture long-range temporal structure 2) how to learn the ConvNet models given limited training samples.

TSN framework

Theoretical basis:

① Long-range temporal structure plays an important role in understanding the dynamics in action videos. However, in terms of temporal structure modeling, consecutive frames are highly redundant. Therefore, dense temporal sampling, which usually results in highly similar sampled frames is unnecessary. Instead a sparse temporal sampling strategy will be more favorable in this case.

② Dense temporal sampling would incur excessive computational cost when applied to long video clips.

Basic idea

TSN extracts short snippets over a long video sequence with a sparse sampling scheme, where the samples distribute uniformly along the temporal dimension. A segmental structure is employed to aggregate information from the sampled snippets.

(The former paper demonstrate that the performance of two-stream ConvNet would be better if the depth of the network be larger).TSN adopts very deep ConvNet and uses a number of good practices to overcome the difficulties caused by limited training samples.

Structure

TSN is also composed of spatial net and temporal net. Instead of working on single frames or frame stacks, temporal segment networks operate on a sequence of short snippets sparsely sampled from the entire video.

One input video is divided into K segments and a short snippet is randomly selected from each segment. The class scores of different snippets are fused by a segmental consensus

function.

Inputs

In addition to RGB images and stacked optical flow fields, TSN uses RGB difference and warped optical flow fields as input to enhance the discriminative power.

RGB difference could present the contextual information about previous and next frames of a frame of video at a specific time point.

Warped optical flow fields could suppress the background motion and makes motion concentrate on the actor.

Advancement

Cross modality pre-training

Utilizing RGB models to initialize the temporal networks(mentioned in *Towards Good Practices for Very Deep Two-Stream ConvNets*).

Regularition techniques

Re-estimating the mean and variance of the activation of first convolutional layer. Adding a extra dropout layer after the global pooling layer.

Data augmentation

Same as *Towards Good Practices for Very Deep Two-Stream ConvNets*).

Deep Local Video Feature for Action Recognition (2017)

This paper points out one of shortcomings of TSN: the global video labels might not be suitable for all of the temporally local samples as the videos often contain content besides the action of interest. The researchers proposed to instead treat the deep networks trained on local inputs as local feature extractors. The local features are then aggregated to form global features which are used to assign video-level labels through a second classification stage. The researchers firstly use snippet-level analysis for local feature extraction, then add a second stage which maps the aggregated features to the video-level labels.

This paper proposed a possible improvement of TSN: DOVF. DOVF is a class of local video features that are extracted from deep neural networks trained on local video clips using global video labels.

Problems

① Videos would occupy a large amount of memory and need lots of computational resource to train and apply CNN on video level.
② Imprecise frame/clip-level labels populated from video labels are too noisy to guide precise mapping from local video snippets to labels.

Main idea

Training the feature extraction networks with local data and with very noisy labels.
Step1 : Pre-trained TSN(with video-level labels) is used to snipped-level classification
Step2: Local features are aggregated to form global features. Another classifier is used to perform video-level classification.

Temporal Relational Reasoning in Videos(2017)

Temporal relational network is designed to learn and reason about temporal dependencies between video frames at multiple time scales.

Theoretical basis

① A single activity can consist of several temporal relations at both short-term and long-term timescales.
② It remains remarkably challenging for CNN to reason about temporal relations and to anticipate what transformations are happening to the observations.

TRN Structure

① Temporal Relations

$$T_2(V) = h_\phi\Big(\sum_{i<j} g_\theta(f_i, f_j)\Big)$$

Temporal relation is defined as above. *f* represents the feature of frames in the video. The functions *h* and *g* fuse features of different ordered frames.

② Multi-scale temporal relations

$$MT_N(V) = T_2(V) + T_3(V)... + T_N(V)$$

In order to accumulate temporal relations at different scales, the function above is proposed.

③ Overall structure

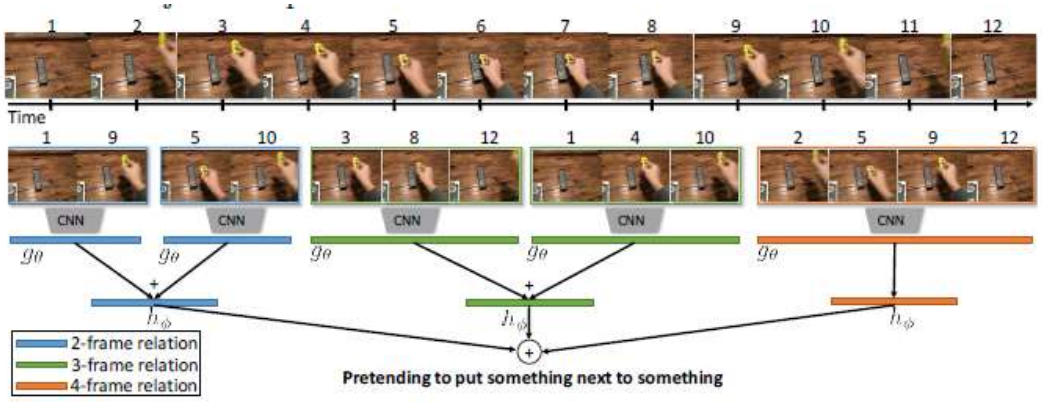

Fig. 2: The illustration of Temporal Relation Networks. Representative frames of a video (shown above) are sampled and fed into different frame relation modules. Only a subset of the 2-frame, 3-frame, and 4-frame relations are shown, as there are higher frame relations included.

The structure of TRN resembles the structure of TSN. TSN aims at possessing information

in long-time range, while TRN focus on temporal reasoning. Both of them take the consequence of action into account.

My review

I read the papers above and learn the evolution and basic ideas about two-stream networks. I find out there are 2 problems that have been raised repeatedly.

Firstly, the insufficient video dataset. Video datasets are limited in diversity and size. It is likely to over-fit if datasets have not been processed. Thus effective methods for data augmentation were discussed in nearly every paper I have read. Multi-task learning, multi-scale cropping and adding extra inputs data(RGB difference, warped optical flow fields) are main ones among them.

Secondly, the temporal information in the video. This problem includes 3 parts. 1) The temporal relation in the video. Actions in reality always follow some constant consequence. In order to learn this, TRN is proposed. Videos are cut into different segments. A snippet is produced by choosing samples randomly from the segment. Each snippet is trained with a two-stream network respectively. The ultimate classification of action in the video is based on the summation of classification results of video snippets. 2) The different interest in the different part of video. Each part of video plays different importance in classifying the action of video. DOVF is proposed to handle this problem. DOVF is a class of local video features that are extracted from deep neural networks trained on local video clips using global video labels. With the use of DOVF, local feature extraction is firstly processed with snippet-level analysis. Then the aggregated features to the video-level labels is produced based on local feature. 3) Networks should have capacity to process long-term motion. TSN is proposed to deal with this problem. Videos are cut into segments and snippets is produced. The class scores of different snippets are fused by a segmental consensus function.

Besides, there are lots of good practices proposed in these papers. Pre-training two-stream ConvNet, treatment of learning rate, high dropout ratio are just examples.

How to process temporal information properly remains an important problem in the field of action recognition. TSN is proposed firstly, it fuses short-term information at different time point. DOVF is then proposed to give advancement to TSN. Different snippets should have different weights. TRN is another advancement of TSN. TRN focus on temporal information reasoning. It constructs temporal relation between frames and fuses them to get classification results. These TSN-based papers aim to try different fusion methods to enhance the performance. While C3D change the CNN structure to process temporal information. And I plan to read papers about C3D in the next few weeks.