Large-scale Video Classification

2016. 9. 26

Tensorflow study group

Hyungjoo Cho

Deeplearning for gesture recognition

CNN method

Large-scale Video Classification with Convolutional Neural Networks

Andrej Karpathy[1,2] karpathy@cs.stanford.edu
George Toderici[1] gtoderici@google.com
Sanketh Shetty[1] sanketh@google.com
Thomas Leung[1] leungt@google.com
Rahul Sukthankar[1] sukthankar@google.com
Li Fei-Fei[2] feifeili@cs.stanford.edu
[1]Google Research [2]Computer Science Department, Stanford University
http://cs.stanford.edu/people/karpathy/deepvideo

Abstract

Convolutional Neural Networks (CNNs) have been established as a powerful class of models for image recognition problems. Encouraged by these results, we provide an extensive empirical evaluation of CNNs on large-scale video classification using a new dataset of 1 million YouTube videos belonging to 487 classes. We study multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information and suggest a multiresolution, foveated architecture as a promising way of speeding up the training. Our best spatio-temporal networks display significant performance improvements compared to strong feature-based baselines (55.3% to 63.9%), but only a surprisingly modest improvement compared to single-frame models (59.3% to 60.9%). We further study the generalization performance of our best model by retraining the top layers on the UCF-101 Action Recognition dataset and observe significant performance improvements compared to the UCF-101 baseline model (63.3% up from 43.9%).

1. Introduction

Images and videos have become ubiquitous on the internet, which has encouraged the development of algorithms that can analyze their semantic content for various applications, including search and summarization. Recently, Convolutional Neural Networks (CNNs) [15] have been demonstrated as an effective class of models for understanding image content, giving state-of-the-art results on image recognition, segmentation, detection and retrieval [11, 3, 2, 20, 9, 18]. The key enabling factors behind these results were techniques for scaling up the networks to tens of millions of parameters and massive labeled datasets that can support the learning process. Under these conditions, CNNs have been shown to learn powerful and interpretable image features [28]. Encouraged by positive results in domain of images, we study the performance of CNNs in large-scale video classification, where the networks have access to not only the appearance information present in single, static images, but also their complex temporal evolution. There are several challenges to extending and applying CNNs in this setting.

From a practical standpoint, there are currently no video classification benchmarks that match the scale and variety of existing image datasets because videos are significantly more difficult to collect, annotate and store. To obtain sufficient amount of data needed to train our CNN architectures, we collected a new Sports-1M dataset, which consists of 1 million YouTube videos belonging to a taxonomy of 487 classes of sports. We make Sports-1M available to the research community to support future work in this area.

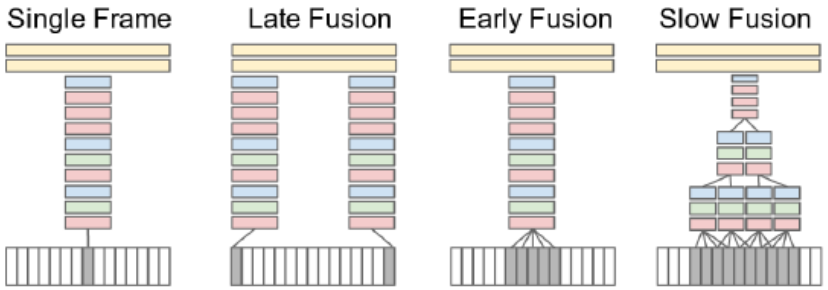
From a modeling perspective, we are interested in answering the following questions: what temporal connectivity pattern in a CNN architecture is best at taking advantage of local motion information present in the video? How does the additional motion information influence the predictions of a CNN and how much does it improve performance overall? We examine these questions empirically by evaluating multiple CNN architectures that each take a different approach to combining information across the time domain.

From a computational perspective, CNNs require extensively long periods of training time to effectively optimize the millions of parameters that parametrize the model. This difficulty is further compounded when extending the connectivity of the architecture in time because the network must process not just one image but several frames of video at a time. To mitigate this issue, we show that an effective approach to speeding up the runtime performance of CNNs is to modify the architecture to contain two separate streams of processing: a *context* stream that learns features on low-resolution frames and a high-resolution *fovea* stream that only operates on the middle portion of the frame. We

Deeplearning for gesture recognition

Large-scale Video Classification with Convolutional Neural Networks

- Purpose
  - Classifying 1 million videos to 487 classes. (Sports-1M dataset)
  - Applying this methods to the UCF-101 dataset to compare with state-of-the-art results (13,320 videos and 101 classes)
- Models

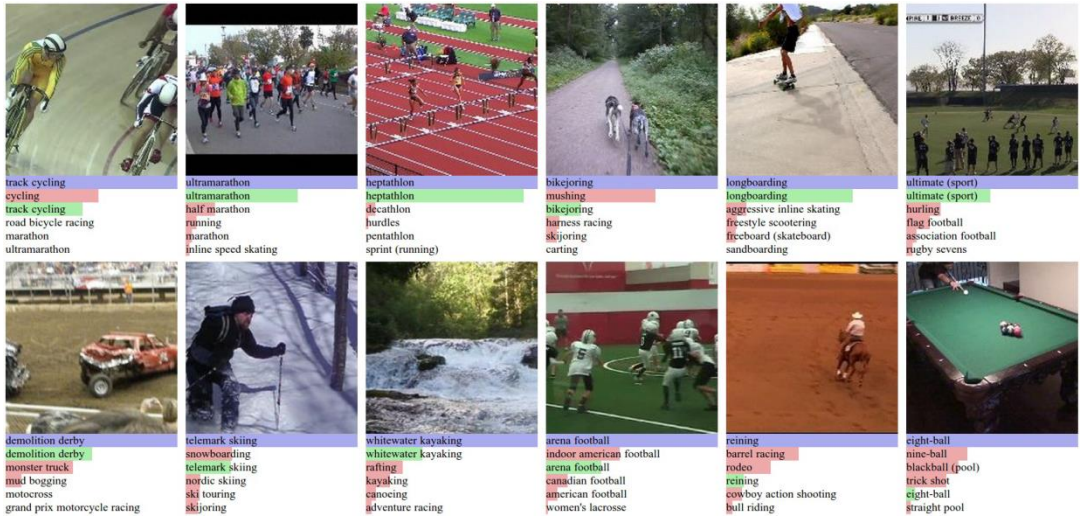


  - Red : Conv layer
  - Green : Normalization layer
  - Blue : Pooling layer
  - Yellow : Fully connected layer
- Condition
  - Fixed-sized clips
  - Alex-Net for CNNs
  - Downpour Stochastic Gradient Descent

Deeplearning for gesture recognition

Large-scale Video Classification with Convolutional Neural Networks

- Result



- Blue : Ground truth
- Green : correct prediction
- Red : Incorrect prediction

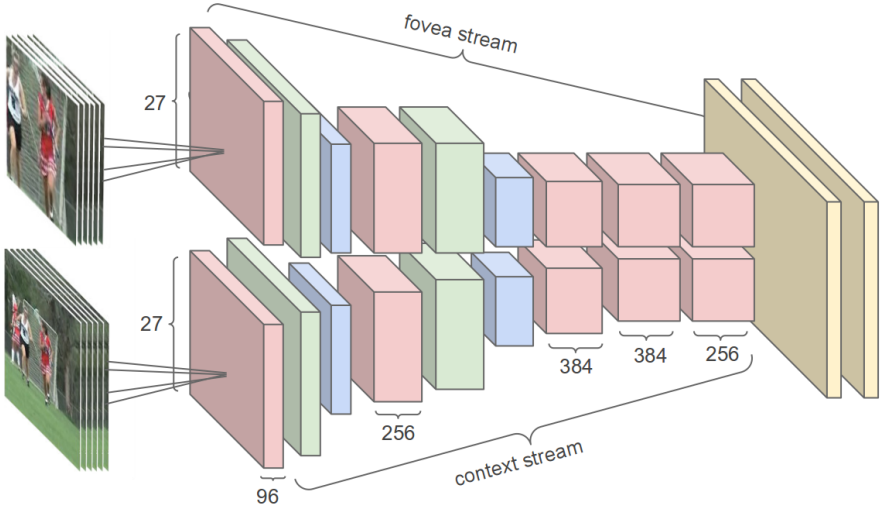
Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	**41.9**	**60.9**	**80.2**
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

→ **Slow Fusion is best choice??**

Deeplearning for gesture recognition

Multi-resolution CNNs

- Structure



- Result

Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multires	**42.4**	**60.0**	**78.5**
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	**41.9**	**60.9**	**80.2**
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

Applying UCF-101

- Result

Model	3-fold Accuracy
Soomro et al [22]	43.9%
Feature Histograms + Neural Net	59.0%
Train from scratch	41.3%
Fine-tune top layer	64.1%
Fine-tune top 3 layers	**65.4**%
Fine-tune all layers	62.2%

Group	mAP from scratch	mAP fine-tune top 3	mAP fine-tune top
Human-Object Interaction	0.26	0.55	0.52
Body-Motion Only	0.32	0.57	0.52
Human-Human Interaction	0.40	0.68	0.65
Playing Musical Instruments	0.42	0.65	0.46
Sports	**0.57**	**0.79**	**0.80**
All groups	0.44	0.68	0.66

Deeplearning for gesture recognition

Temporal CNN method

Beyond Short Snippets: Deep Networks for Video Classification

Joe Yue-Hei Ng[1] yhng@umiacs.umd.edu
Matthew Hausknecht[2] mhauskn@cs.utexas.edu
Sudheendra Vijayanarasimhan[3] svnaras@google.com
Oriol Vinyals[3] vinyals@google.com
Rajat Monga[3] rajatmonga@google.com
George Toderici[3] gtoderici@google.com

[1]University of Maryland, College Park [2]University of Texas at Austin [3]Google, Inc.

Abstract

Convolutional neural networks (CNNs) have been extensively applied for image recognition problems giving state-of-the-art results on recognition, detection, segmentation and retrieval. In this work we propose and evaluate several deep neural network architectures to combine image information across a video over longer time periods than previously attempted. We propose two methods capable of handling full length videos. The first method explores various convolutional temporal feature pooling architectures, examining the various design choices which need to be made when adapting a CNN for this task. The second proposed method explicitly models the video as an ordered sequence of frames. For this purpose we employ a recurrent neural network that uses Long Short-Term Memory (LSTM) cells which are connected to the output of the underlying CNN. Our best networks exhibit significant performance improvements over previously published results on the Sports 1 million dataset (73.1% vs. 60.9%) and the UCF-101 datasets with (88.6% vs. 88.0%) and without additional optical flow information (82.6% vs. 73.0%).

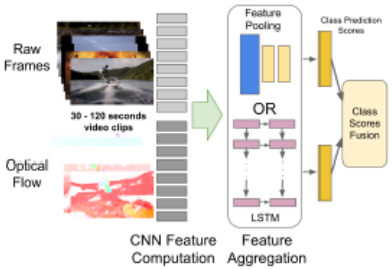


Figure 1: Overview of our approach.

1. Introduction

Convolutional Neural Networks have proven highly successful at static image recognition problems such as the MNIST, CIFAR, and ImageNet Large-Scale Visual Recognition Challenge [15, 21, 28]. By using a hierarchy of trainable filters and feature pooling operations, CNNs are capable of automatically learning complex features required for visual object recognition tasks achieving superior performance to hand-crafted features. Encouraged by these positive results several approaches have been proposed recently to apply CNNs to video and action classification tasks [2, 13, 14, 19].

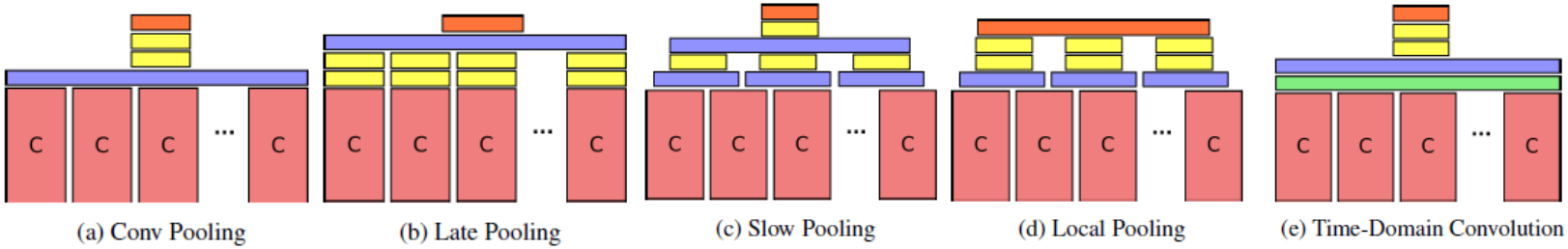
Video analysis provides more information to the recognition task by adding a temporal component through which motion and other information can be additionally used. At the same time, the task is much more computationally demanding even for processing short video clips since each video might contain hundreds to thousands of frames, not all of which are useful. A naïve approach would be to treat video frames as still images and apply CNNs to recognize each frame and average the predictions at the video level. However, since each individual video frame forms only a small part of the video's story, such an approach would be using incomplete information and could therefore easily confuse classes especially if there are fine-grained distinctions or portions of the video irrelevant to the action of interest.

Therefore, we hypothesize that learning a global description of the video's temporal evolution is important for accurate video classification. This is challenging from a modeling perspective as we have to model variable length videos with a fixed number of parameters. We evaluate two approaches capable of meeting this requirement: feature-pooling and recurrent neural networks. The feature pooling networks independently process each frame using a

Deeplearning for gesture recognition

Beyond Short Snippets: Deep Networks for Video Classification

- Purpose
  - Adapting CNNs to the time domain (Using both motion information and a global description of the video)
  - Long periods of a video
- Models



(a) Conv Pooling (b) Late Pooling (c) Slow Pooling (d) Local Pooling (e) Time-Domain Convolution

  - Red : Conv layer
  - Green : Time-domain convolution
  - Blue : Pooling layer
  - Yellow : Fully connected layer
  - Orange : Softmax layer
- Condition
  - Sampling to reduce input nodes (1frame per 1 second)
  - Alex-Net for CNNs
  - Downpour Stochastic Gradient Descent

Beyond Short Snippets: Deep Networks for Video Classification

- Result

Method	Clip Hit@1	Hit@1	Hit@5
Conv Pooling	**68.7**	**71.1**	**89.3**
Late Pooling	65.1	67.5	87.2
Slow Pooling	67.1	69.7	88.4
Local Pooling	68.1	70.4	88.9
Time-Domain Convolution	64.2	67.2	87.2

→ **Just 1 temp-conv layer cannot represent high level features.**

Method	Hit@1	Hit@5
AlexNet single frame	63.6	84.7
GoogLeNet single frame	**64.9**	**86.6**
Conv pooling + AlexNet	70.4	89.0
Conv pooling + GoogLeNet	**71.7**	**90.4**

→ **More complex model is better.**

Method	Frames	Clip Hit@1	Hit@1	Hit@5
Conv pooling	30	66.0	71.7	90.4
	120	70.8	72.3	90.8

→ **More data is better.**

Deeplearning for gesture recognition

Temporal CNN method

Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video

Lionel Pigou, Aäron van den Oord*, Sander Dieleman*,
Mieke Van Herreweghe & Joni Dambre
{lionel.pigou,aaron.vandenoord,sander.dieleman,
mieke.vanherreweghe, joni.dambre}@ugent.be
Ghent University

February 11, 2016

Abstract

Recent studies have demonstrated the power of recurrent neural networks for machine translation, image captioning and speech recognition. For the task of capturing temporal structure in video, however, there still remain numerous open research questions. Current research suggests using a simple temporal feature pooling strategy to take into account the temporal aspect of video. We demonstrate that this method is not sufficient for gesture recognition, where temporal information is more discriminative compared to general video classification tasks. We explore deep architectures for gesture recognition in video and propose a new end-to-end trainable neural network architecture incorporating temporal convolutions and bidirectional recurrence. Our main contributions are twofold; first, we show that recurrence is crucial for this task; second, we show that adding temporal convolutions leads to significant improvements. We evaluate the different approaches on the Montalbano gesture recognition dataset, where we achieve state-of-the-art results.

1 Introduction

Gesture recognition is one of the core components in the thriving research field of human-computer interaction. The recognition of distinct hand and arm motions is becoming increasingly important, as it enables smart interactions with electronic devices. Furthermore, gesture identification in video can be seen as a first step towards sign language recognition, where even subtle differences in motion can play an important role. Some examples that complicate the identification of gestures are changes in background and lighting due to the varying environment, variations in the performance and speed of the gestures, different clothes worn by the performers and different positioning relative to the camera. Moreover, regular hand motion or out-of-vocabulary gestures should not to be confused with one of the target gestures.

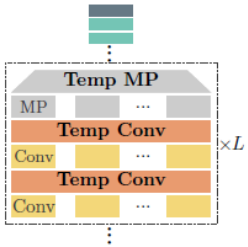
Convolutional neural networks (CNNs) (LeCun et al., 1998) are the de facto standard approach in computer vision. CNNs have the ability to learn complex hierarchies with increasing levels of abstraction while being end-to-end trainable. Their success has had a huge impact on vision based applications like image classification (Krizhevsky et al., 2012), object detection

*Now at Google DeepMind.

Deeplearning for gesture recognition

Beyond Temporal Pooling: Recurrence and Temporal Convolution for Gesture Recognition in Video

- Purpose
  - Deep architectures for temporal convolution
  - Montalbano Gesture Recognition Dataset (Gesture only)
- Models



  - Yellow : Conv layer
  - Red : Time-domain convolution
  - Gray : Max Pooling layer
  - Cyan : Fully connected layer
  - Dark green : Softmax layer
- Condition
  - Sampling to reduce input nodes (1frame per 1 second)
  - Alex-Net for CNNs
  - Adam optimizer

Deeplearning for gesture recognition

Beyond Temporal Pooling: Recurrence and Temporal Convolution for Gesture Recognition in Video

- Result

Architecture	Precision	Recall	Error Rate*
Single-Frame CNN (Figure 1a)	67.86%	57.57%	20.68%
Temp Max-Pooling (Figure 1b)	85.03%	82.92%	8.66%
Temp Mean-Pooling (Figure 1b)	85.93%	85.80%	8.55%
Temp Conv (Figure 1d)	89.36%	90.15%	4.67%

	Error rate (%)
Tanh units	18.90
ReLU	14.40
+ dropout	11.90
+ LCN (first 2 layers)	10.30
+ data augmentation	**8.30**

- Comparing with 3D-CNN for same dataset