Supplemental Material

Shengxin Zha[1]
szha@u.northwestern.edu
Florian Luisier[2]
fluisier@bbn.com
Walter Andrews[2]
wandrews@bbn.com
Nitish Srivastava[3]
nitish@cs.toronto.edu
Ruslan Salakhutdinov[3]
rsalakhu@cs.toronto.edu

[1] Northwestern University
Evanston IL USA
[2] Raytheon BBN Technologies
Cambridge, MA USA
[3] University of Toronto
Toronto, Ontario, Canada

1 Alternative CNN architecture

Starting with the winning CNN architecture of the ImageNet 2012 challenge [3], we made some modifications that lead to better classification performance on both the ImageNet dataset and video classification datasets, validating the importance of adopting small receptive field and increasing depths. Our proposed convolutional neural network is described in Table 1. This architecture differs from the architecture in [3] in three ways:

- In place of a two-column architecture, we use a densely connected one with the same number of total maps at each layer.
- Inspired by [5], we added 1×1 convolutions. These 1×1 convolutional layers can be seen as a cheap way of adding depth into a convolution net. They are cheap because they have few parameters and can be implemented using dense matrix multiplies which are much faster than convolutional kernels, especially on a GPU. This means that they do not contribute to overfitting and also do not increase computation time by much.
- We used dropout in the convolutional layers besides the fully connected ones.

We found that these modifications allow us to get a 5% improvement over the model in [3] on the ILSVRC-2012 validation set (averaged over 5 patch positions and 2 flips per position for each image) and consistent improvement in video classification. The trained model, as well as the code for training the model, are publicly available[1].

We applied the proposed video classification approach to this CNN architecture on the TRECVID MED'14 dataset. The result is given in Table 2. Its improvement over the classification result obtained with CNN architecture [3] shows the effectiveness of the 1×1 receptive fields. Yet, the 19-layer CNN architecture [8] yields superior performance than the proposed CNN architecture, which also demonstrates the importance of small receptive fields and depths.

[1] Link to the proposed pre-trained CNN architecture (not yet displayed to preserve anonymity).

Type	Patch size - stride - pad	Output size
Input	-	$224 \times 224 \times 3$
Convolutional 1	$7 \times 7 - 2 - 1$	$110 \times 110 \times 96$
Max Pool	$3 \times 3 - 2 - 1$	$55 \times 55 \times 96$
Response Norm	-	$55 \times 55 \times 96$
Convolutional 2a	$5 \times 5 - 2 - 1$	$27 \times 27 \times 256$
Convolutional 2b	$1 \times 1 - 1 - 0$	$27 \times 27 \times 256$
Max Pool	$3 \times 3 - 2 - 1$	$14 \times 14 \times 96$
Response Norm	-	$14 \times 14 \times 96$
Convolutional 3a	$3 \times 3 - 1 - 1$	$14 \times 14 \times 256$
Convolutional 3b	$1 \times 1 - 1 - 0$	$14 \times 14 \times 256$
Convolutional 4a	$3 \times 3 - 1 - 1$	$14 \times 14 \times 384$
Convolutional 4b	$1 \times 1 - 1 - 0$	$14 \times 14 \times 768$
Convolutional 4c	$1 \times 1 - 1 - 0$	$14 \times 14 \times 384$
Convolutional 5a	$3 \times 3 - 1 - 0$	$12 \times 12 \times 512$
Convolutional 5b	$1 \times 1 - 1 - 0$	$12 \times 12 \times 1024$
Convolutional 5c	$1 \times 1 - 1 - 0$	$12 \times 12 \times 512$
Max Pool	$3 \times 3 - 2 - 1$	$6 \times 6 \times 512$
FC (hidden6)	-	$1 \times 1 \times 4096$
FC (hidden7)	-	$1 \times 1 \times 4096$
FC (output)	-	$1 \times 1 \times 1000$

Table 1: Proposed CNN architecture. Response Norm refers to Cross-map response normalization.

Layer	Dim.	SP	Norm	SVM	mAP
output	8,000	SP8	root	linear	22.86%
output	8,000	SP8	root	χ^2	28.88%
hidden6	32,768	SP8	ℓ_2	linear	24.98%
hidden6	32,768	SP8	ℓ_2	RBF	30.29%
hidden7	32,768	SP8	ℓ_2	linear	26.44%
hidden7	32,768	SP8	ℓ_2	RBF	31.97%

Table 2: Classification result on TRECVID MED'14 100Ex based on the proposed video classification approach and the new CNN architecture.

2 Fisher Vectors

The FV is a generalization of the bag-of-words approach that encodes the zero-order, the first- and the second-order statistics of the descriptors distribution. The FV encoding procedure is summarized below [7].

- First, learn a Gaussian mixture model (GMM) on low-level descriptors extracted from a generic set of unlabeled videos.
- Second, compute the gradients of the log-likelihood of the GMM (known as the score function) with respect to the GMM parameters. The gradient of the score function with respect to the mixture weight parameters encodes the zero-order statistics. The gradient with respect to the Gaussian means encodes the first-order statistics, while the gradient with respect to the Gaussian variance encodes the second-order statistics.
- Third, concatenate the score function gradients into a single vector and apply a signed square rooting on each FV dimension (power normalization) followed by a global ℓ_2

normalization.

As low-level features, we considered both the standard D-SIFT descriptors [6] and the more sophisticated motion-based improved dense trajectories (IDT) [10]. For both, we selected the parameters giving the best performance on the validation set. , in the same way as we did for the CNN-based features, to allow a fair comparison between the two approaches. For the SIFT descriptors, we opted for multiscale (5 scales) and dense (stride of 4 pixels in both spatial dimensions) sampling, root normalization and spatiotemporal pyramid pooling. For the IDT descriptors, we concatenated histogram of oriented gradients [2], histogram of flow [1], and motion boundary histograms [4] descriptors extracted along the estimated motion trajectory.

3 TRECVID MED 2014 dataset

The TRECVID MED'14 dataset is a realistic dataset of Youtube-like videos. This dataset consists of:

- a training set of 4,992 unlabeled *background* videos used as the negative examples;
- a training set of 2,991 *positive* and *near-miss* videos including 100 positive videos and about 50 near-miss videos (treated as negative examples) for each of the 20 pre-specified events (see Table 3);
- a test set of 23,953 videos contains positive and negative instances of the pre-specified events.

Some sample frames are given in Figure 1. Contrary to other popular video datasets, such as UCF-101 [9], the MED'14 dataset is not constrained to any class of videos (e.g., sports, human actions). It consists of a heterogeneous set of temporally untrimmed YouTube-like videos of various resolutions, quality, camera motions, and illumination conditions. This dataset is thus one of the largest and the most challenging dataset for video event detection.

As a retrieval performance metric, we considered the one used in the official MED'14 task, i.e., mean average precision (mAP) across all events. Let E denote the number of events, P_e the number of positive instances of event e, then mAP is computed as

$$\text{mAP} = \frac{1}{E}\sum_{e=1}^{E}\text{AP}(e), \tag{1}$$

where the average precision of an event e is defined as

$$\text{AP}(e) = \frac{1}{P_e}\sum_{tp=1}^{P_e}\frac{tp}{\text{rank}(tp)}.$$

mAP is thus normalized between 0 (low classification performance) and 1 (high classification performance). In this paper, we will report it in percentage value.

4 Confusion matrix on UCF-101

The confusion matrix of the fusion result of CNN-hidden6 and IDT+FV on UCF-101 split one is given in Figure 2. As shown in the confusion matrix, the worst case class is *Hammering*, which is resulted relatively poor classification performance of both the CNN and the IDT+FV features on this class.

Events E021-E030	Events E031-E040
Attempting a bike trick	Beekeeping
Cleaning an appliance	Wedding shower
Dog show	Non-motorized vehicle repair
Giving directions	Fixing a musical instrument
Marriage proposal	Horse riding competition
Renovating a home	Felling a tree
Rock climbing	Parking a vehicle
Town hall meeting	Playing fetch
Winning a race w/o a vehicle	Tailgating
Working on a metal crafts project	Tuning a musical instrument

Table 3: TRECVID MED 2014 pre-specified events.

Figure 1: Sample frames from the TRECVID MED '14 dataset.

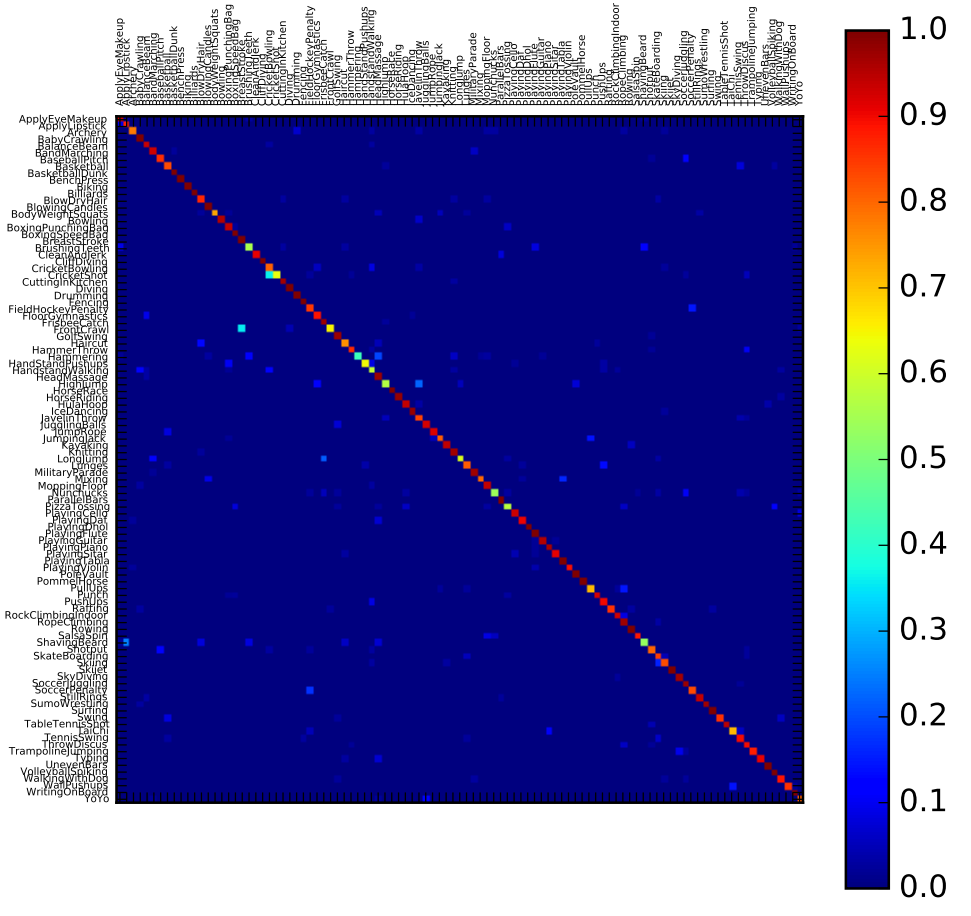

Figure 2: Confusion matrix of result obtained with fusion of CNN-hidden6 and IDT+FV on UCF-101 split 1. Horizontal axis: predicted classes. Vertical axis: true classes. The CNN features were extracted from the second last hidden layer of CNN architecture [8] with SP8, normalized by ℓ_2 norm and classified with linear SVM.

References

[1] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision (ECCV)*, 2006.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[4] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[5] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations (ICLR)*, 2014.

[6] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, pages 143–156, 2010.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012.

[10] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.