Traffic Signs and Pedestrians Vision with Multi-Scale Convolutional Networks

Pierre Sermanet, Koray Kavukcuoglu and Yann LeCun

The Courant Institute of Mathematical Sciences - New York University

sermanet@cs.nyu.edu, koray@cs.nyu.edu, yann@cs.nyu.edu

Convolutional Networks (ConvNets) are biologically-inspired multi-stage architectures that automatically learn hierarchies of invariant features. While many popular vision approaches use hand-crafted features such as HOG or SIFT, ConvNets learn features at every levels from data that are tuned to the task at hand. The traditional ConvNet architecture was modified by feeding 1st stage features in addition to 2nd stage features to the classifier. We apply these multi-scale ConvNets to the tasks of traffic sign classification and pedestrian detection and establish new accuracy records, above human performance for road signs. We also show an significant accuracy gain on the pedestrian task when using unsupervised pre-training with Convolutional Predictive Sparse Coding [1] (ConvPSD). The ConvNet was implemented using the EBLearn C++ open-source package [1] [2].

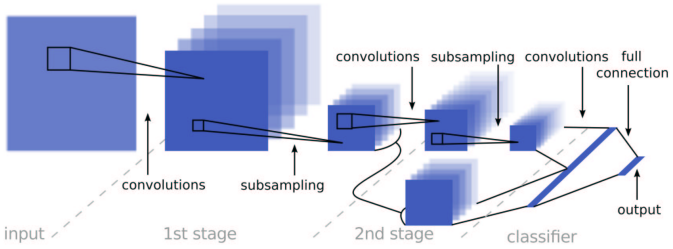

Figure 1: A 2-stage multi-scale ConvNet architecture. The input is processed in a feed-forward manner through two stage of convolutions and subsampling, and finally classified with a linear classifier. The output of the 1st stage is also fed directly to the classifier as higher-scale features.

Although traffic signs recognition is a relatively constrained problem because each sign is unique, rigid and have little variability in appearance, GTSRB [3] is a new realistic dataset challenged by real-world variabilities such as viewpoint variations, lighting conditions (saturations, low-contrast), motion-blur, occlusions, sun glare, physical damage, colors fading, graffiti, stickers and an input resolution as low as 15x15 (Fig. 2). The GTSRB traffic sign classification task held its first phase in January 2011, in which our system yielded the 2nd-best accuracy of 98.97% (the best entry obtained 98.98%), above the human performance of 98.81%, using 32x32 color input images. Experiments conducted after phase 1 produced a new record of 99.17% (Fig. 2) by increasing the network capacity, and by using greyscale images instead of color.

We also apply multi-scale ConvNets to the INRIA pedestrian detection task [4]. Additionnaly, we initialize the networks weights with unsupervised pre-training using the ConvPSD method, and establish a new record of 6.79% miss rate at 1 false positive per image (FPPI) and 12.67% of area under curve (AUC) in the [0,1] FPPI range. Fig. 3 compares algorithms published on the Caltech pedestrian website [2] . In a previous experiment [1], we show a miss rate improvement from 14.8% down to 11.5% at 1 FPPI when using unsupervised pre-training.

[1] http://eblearn.sf.net

[2] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians

#	Team	Method	Accuracy
	sermanet	**EBLearn 2LConvNet ms 108 feats + 100-feats CF classifier + No color**	**99.17%**
197	IDSIA	cnn_hog3	98.98%
196	IDSIA	cnn_cnn_hog3	98.98%
178	**sermanet**	**EBLearn 2LConvNet ms 108 feats**	**98.97%**
195	IDSIA	cnn_cnn_hog3_haar	98.97%
187	**sermanet**	**EBLearn 2LConvNet ms 108 + val**	**98.89%**
199	INI-RTCV	Human performance	98.81%
170	IDSIA	CNN(IMG)_MLP(HOG3)	98.79%

Figure 2: **Left:** Difficult road sign examples in GTSRB coming from real-word perturbations. **Right:** Official top 7 results and new accuracy record after GTSRB Phase 1 (99.17%).

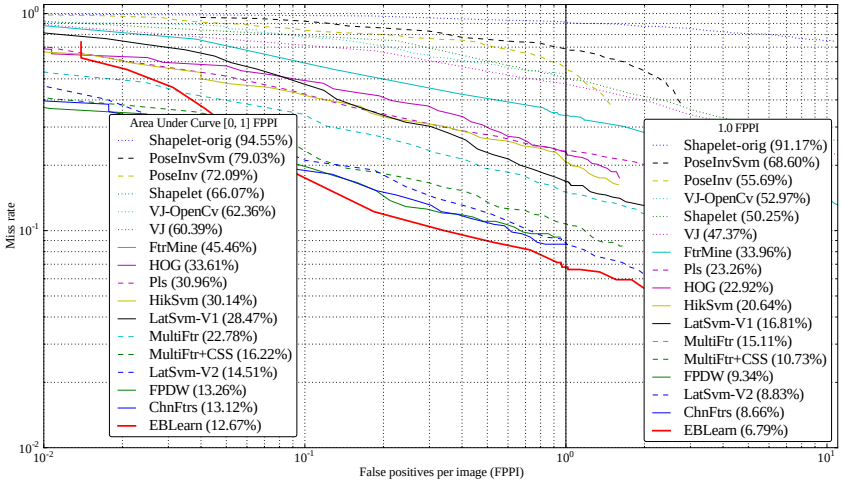

Figure 3: DET curves on the INRIA pedestrian test set, plotting false positives per image (FPPI) against miss rate. Algorithms are sorted from top to bottom using 2 metrics: on the left is the area under curve (AUC) between 0 and 1 FPPI, on the right is the miss rate at 1 FPPI. Our algorithm ("EBLearn") outperforms other algorithms published on the Caltech pedestrian website.

References

[1] Kavukcuoglu, K, Sermanet, P, Boureau, Y, Gregor, K, Mathieu, M, and LeCun, Y. Learning convolutional feature hierachies for visual recognition. In *Advances in Neural Information Processing Systems (NIPS 2010)*, 2010.

[2] Sermanet, P, Kavukcuoglu, K, and LeCun, Y. Eblearn: Open-source energy-based learning in c++. In *Proc. International Conference on Tools with Artificial Intelligence (ICTAI'09)*. IEEE, 2009.

[3] Stallkamp, J, Schlipsing, M, Salmen, J, and Igel, C. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *submitted to International Joint Conference on Neural Networks*, 2011.

[4] Dalal, N and Triggs, B. Histograms of oriented gradients for human detection. In Schmid, C, Soatto, S, and Tomasi, C, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005.