Neural Decision Forests for Semantic Image Labelling

Samuel Rota Bulò
Fondazione Bruno Kessler
Trento, Italy
rotabulo@fbk.eu

Peter Kontschieder
Microsoft Research
Cambridge, UK
pekontsc@microsoft.com

Machine learning is one of the strongest driving forces behind modern computer vision systems, giving rise to impressive results on practical tasks like image classification and body part recognition. One of the prerequisites for making such systems work is the availability of large and accurately labelled training data sets. Of similar importance is how to find an optimal data representation for the task at hand. In the computer vision community, much effort has gone into the careful design of appropriate representations (*aka* features), *i.e*. SIFT, HOG or Shape Context are examples of ingeniously engineered representations, exploiting prior knowledge about the tasks to be accomplished. However, in the recent years increasing efforts have been spent on addressing also the data representation problem by means of machine learning techniques, in order to reduce the dependency on properly designed, hand-crafted features by jointly learning of classifiers *and* the representations they operate on. This trend has mostly been stimulated by the successful application of deep architectures to *representation learning*. In such architectures, multiple sequences of non-linear processing stages generate data representation hierarchies, that are ultimately able to describe highly complex compositions in the data. However, it is known that deep learning architectures require substantial experience for hyper-parameter tuning.

In our work we investigate how to deploy representation learning within the conceptually simple framework of decision forests [3, 7]. Decision forests are ensembles of binary decision trees that have become very popular in computer vision due to their efficiency, flexibility, good generalization capability and inherent ability to handle multi-class problems. They have been applied to various tasks including semantic segmentation [6, 9], object detection [5, 6], and edge detection [4]. However, decision forests have not yet been adopted to account for representation learning in the previously introduced sense. The learning approach we introduce is called Neural Decision Forests (NDF), which provides a novel perspective on classification trees for joint learning of data representations and the decisions taken upon them. The core of our method is a novel, *randomized Multi-Layer Perceptron* (rMLP) that we deploy as substitute for the conventional *split* (or interior) nodes of the trees. The rMLP allows us to jointly learn i) new (and possibly non-linear) data representations by means of its hidden layers, based on the discriminatively routed data it is reached by and ii) optimal predictions for the emerging left and right child nodes to which the output of our rMLP routes the data of the parent in a soft way, respectively. An illustration in Fig. 1 shows an input RGB image with heat map visualizations of four obtained representations, automatically learned by our rMLP.

In connection to the rMLP we provide a probabilistic model to describe the splitting process and design a new split function quality measure, which also guides the optimization of the network parameters. Additionally, we prove that our quality measure is substantially equivalent to the usual information gain criterion used to train standard decision forests *only if* the routing function is binary. However, the rMLP-based routing function is non-binary, thus rendering the information gain a vehicle of *non-optimal* decisions in terms of log-loss. For this reason, we propose to adopt a different quality measure.

Standard MLPs show a strong tendency to overfit to data [2], consequently giving poor generalization accuracy and therefore low performance on unseen test data. We prevent this effect by taking a number of countermeasures. First, the topology of our rMLP is determined by the distribution of the labels arriving at a node: we impose a higher complexity when many different classes are present. Second, we apply a randomized selection step for choosing the signals to the input layer of the rMLP. Therefore, we have no fixed 'wiring' of image pixel positions to the network, which drastically reduces the number of parameters to be learned but also allows us to interpret our rMLP as representation generator that learns weights for non-local kernels - A desideratum that was highlighted in [1] to exploit the principle of non-local generalization. Finally, we introduce an ℓ_1-norm based regularization strategy to further escape the risk of overfitting and to obtain more concise networks with sparsified connections. The regularization is naturally

Figure 1. Example input RGB image and learned representations of our rMLP taken from a hidden layer, visualized using heat-maps. Please note the diverse responses in different areas of the image.

Method	ETRIMS8			CAMVID			LFW		
	Global	Class-Avg	Jaccard	Global	Class-Avg	Jaccard	Global	Class-Avg	Jaccard
RF Baseline (RGB only)	64.5 ±1.6	59.6 ±1.7	40.3 ±1.1	64.0	41.6	27.2	88.8	49.3	37.0
$\text{NDF}_{\texttt{MLPC}-\ell 1}$ (RGB only)	71.7 ±2.0 (**+7.2**)	65.3 ±2.3 (**+5.7**)	46.9 ±2.0 (**+6.6**)	69.0 (**+5.0**)	46.8 (**+5.2**)	31.7 (**+4.5**)	91.8	59.7 (**+10.4**)	46.5 (**+9.5**)
RF Baseline	72.2 ±1.9	68.0 ±0.8	47.5 ±1.0	68.5	50.3	32.4	89.2	55.6	41.6
$\text{NDF}_{\texttt{MLPC}-\ell 1}$	80.8 ±0.7 (**+8.6**)	74.6 ±0.7 (**+6.6**)	56.9 ±1.2 (**+9.4**)	82.1 (**+13.6**)	56.1 (**+5.8**)	43.3 (**+10.9**)	95.4 (**+6.2**)	74.1 (**+18.5**)	59.6 (**+18.0**)

Table 1. Experimental results comparing baseline random forests (RF) with our strongest model $\text{NDF}_{\texttt{MLPC}-\ell 1}$. Top: RGB inputs only. Bottom: Derived image representations as described in text. All numbers in [%].

encoded by the introduction of a Laplace prior for the network's parameters in the graphical model defining the splitting process.

As for the experimental evaluation, we assessed the quality of our Neural Decision Forests on three datasets for the task of semantic segmentation. We tested on Etrims8 (8 classes, 60 images for analysing behaviour with respect to overfitting, 5-fold cross validation), CamVid (11 classes, 600 images) and a subset of the Labelled Faces in the Wild (LFW, 8 classes, 601 images) datasets. The results we obtained (see, Tab. 1) show significant improvements over standard decision forests but also demonstrate comparable or better results to forests trained with more complex data representations. For instance, training $\text{NDF}_{\texttt{MLPC}-\ell 1}$ (a ℓ_1 regularized version of a randomized multi-layer perceptron) on RGB images directly performs similar to standard forests trained on derived representations (we used Lab raw intensities, first/second order derivatives of the luminance channel and HOG-like features) for Etrims8 and CamVid and even exceeds the performance on the LFW dataset. Whereas, if we train our neural decision forests on top of the derived representations, we get a considerable boost of all the scores.

Another impressive property of our NDF is that our trees consist only of a fraction of nodes compared to standard forests, *i.e.* we obtain high compression rates in our experiments (up to factor 50 for the LFW dataset). As an example, we show in Fig. 2 the average leaf entropy vs. average number of samples per leaf per forest for the Camvid dataset; we can clearly see that $\text{NDF}_{\texttt{MLPC}-\ell 1}$ yields trees with leaves of large cardinality and low entropy as opposed to the ones generated by the RF. These findings encourage and support our initial claim to exploit the hierarchical nature of decision trees to provide a principled and joint approach of representation and discriminative learning.

More details of our novel Neural Decision Forest model

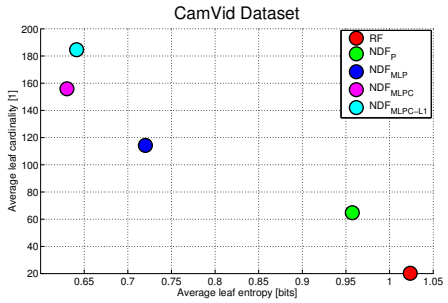

Figure 2. Average entropy vs. average cardinality of leaves for the CamVid dataset (best viewed in colour).

can be found in [8], where we explain the learning procedure for both, the rMLP weights and the child node posteriors. Also, we show how to apply our model for semantic image labelling and provide more experimental evaluations and comparisons to state-of-the-art approaches.

References

[1] Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[3] L. Breiman. Random forests. In *Machine Learning*, volume 45, pages 5–32, 2001.

[4] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *(ICCV)*, 2013.

[5] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *(CVPR)*, pages 1022–1029, 2009.

[6] P. Kontschieder, S. Rota Bulò, M. Pelillo, and H. Bischof. Structured labels in random forests for semantic labelling and object detection. *(PAMI)*, to appear, 2014.

[7] J. R. Quinlan. Induction of decision trees. *(ML)*, pages 81–106, 1986.

[8] S. Rota Bulò and P. Kontschieder. Neural decision forests for semantic image labelling. In *(CVPR)*, 2014.

[9] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *(CVPR)*, pages 1–8, 2008.