Unrolling loopy top-down semantic feedback in convolutional deep networks

Carlo Gatta
Centre de Visió per Computador
Bellaterra, Barcelona, Spain, 08193.
cgatta@cvc.uab.es

Adriana Romero
Dept. MAiA
Universitat de Barcelona, Spain
adriana.romero@ub.edu

Joost van de Weijer
Centre de Visió per Computador
Bellaterra, Barcelona, Spain, 08193.
joost@cvc.uab.es

Abstract

In this paper, we propose a novel way to perform top-down semantic feedback in convolutional deep networks for efficient and accurate image parsing. We also show how to add global appearance/semantic features, which have shown to improve image parsing performance in state-of-the-art methods, and was not present in previous convolutional approaches. The proposed method is characterised by an efficient training and a sufficiently fast testing. We use the well known SIFTflow dataset to numerically show the advantages provided by our contributions, and to compare with state-of-the-art image parsing convolutional based approaches.

1. Introduction

Deep convolutional architectures have been shown to perform as well or better than methods based on handcrafted features in a large set of computer vision problems. Probably, the most interesting property of convolutional deep networks is that they can learn a set of increasingly complex features from raw pixels. In the case of end-to-end supervised learning, the deep convolutional architecture is at the same time "feature extractor" and classifier. Unsupervised techniques have shown to be able to learn relevant representations of data. Also, they are usually more efficient than supervised ones and less prone to fall into local minima [1, 7].

Recently, in the field of image parsing, convolutional deep networks have reached state-of-the-art performance on several datasets [10, 8, 23] without the need of handcrafted features, and being faster than previous methods at test time. These methods show that convolutional deep networks can learn local appearance features that are effective for image parsing. The work in [23] demonstrates the relevance of semantic feedback in solving image parsing problems. Authors implement top-down semantic feedback by means of recurrent convolutional networks.

In this paper, we propose an alternative strategy to perform top-down semantic feedback that is easier and faster to train than previous convolutional approaches. Instead of using a recurrent network, we learn a large set of features in an unsupervised way and propagate the categories posterior probability to the next deep convolutional network. In this case, subsequent deep architectures **do not share** the parameters, thus decoupling feature learning and top-down semantic feedback, thereby simplifying the training process. Within this approach, we also add global appearance and semantic features, which have been shown to improve the performance of image parsing methods [28], and that are not present in current convolutional approaches.

The paper is organised as follows. Section 2 presents related works; section 3 provides the paper's motivation and contributions; section 4 outlines the contributions by building on a convolutional deep network; section 5 shows experimental results supporting our claims both quantitatively and qualitatively; section 6 discusses relevant issues regarding the proposed method in comparison with previous state-of-the-art approaches; section 7 concludes the paper.

2. Related work

For the sake of clearness, from now on, we divide the features into local and global, where global means that the feature provides information on the whole image, and local refers to a given receptive field (even a very large one). We also divide features into appearance and semantic, where semantic refers to features based on posterior probabilities or labelling related to the (supervised) categories.

Table 1. Characteristics of our method compared to state-of-the-art convolutional-based image parsing approaches. Training and testing times are excerpted from respective papers for the configurations that gave the best results as reported in Table 2. Training time for the method in [23] has been provided by the author in a personal communication. To provide a testing time independent on the image size used by different methods, we provide the testing speed in kilo pixels per second.

Method	Local features		Global features		Top-down feedback	Training Time	Testing Speed
	Appearance	Semantic	Apperance	Semantic			
Multi-scale [8]	Learned	No	No	No	No	2∼5 days	1.3 kpixels/s
Recurrent [23]	Learned	Learned	No	No	Yes	$>$ 1 week	8.9 kpixels/s
Our method	Learned	Learned	Learned	Learned	Yes	3h50'	3.7 kpixels/s

2.1. Image Parsing

Image parsing (also known as scene parsing, scene labeling, or semantic segmentation) has been tackled by several methods. These methods can be categorized into four main groups.

The first large group of methods is based on handcrafted appearance features coupled with a conditional random field (CRF) approach (of increasing complexity), which imposes spatial and semantic consistency in the final classification [15, 14, 16, 11, 31, 2]. The interaction in the CRF approach is partially designed and partially learned. Some recent proposals aim at fully learnable contextual interaction [21, 26, 3]. Since CRFs have an expensive learning/inference, the great majority of these methods require an initial over-segmentation in super pixels to reduce their computational complexity. Although this group of methods has shown to perform well on several datasets, the handcrafted features and the computational complexity of the CRFs can hinder their successful applicability to other datasets.

The second group of methods substitute the CRF approach by a multi-scale Stacked Sequential Learning (SSL) strategy [12, 30, 19, 9]. This allows an easier training procedure, since semantic features are fed to the classifier together with appearance features, in order to learn complex appearance/semantic interactions. These methods do not usually require super-pixel segmentation. While surpassing the previous group of methods on some datasets [9], they still have the problem that appearance and semantic features are hand-crafted.

Both CRF and SSL methods provide strategies to account for global semantic coherence, by means of handcrafted features. Some of them also account for global appearance features.

The third group includes non-parametric methods, which have also shown to perform very well in image parsing problems [18, 6, 28, 29]. Non-parametric methods are based on a k-nearest neighbor (k-NN) classifier on superpixels and introduce the idea that extensive training of powerful classifiers (and graphical models) is not necessary if a careful selection of the training images can be done at test time. This selection allows to discard training images that are not relevant for a given test image. In [28], authors add the statistics of classes to improve the image selection, providing a form of global semantic information. Experiments conducted in [28] show that global appearance and semantic features have a significant impact on the final result. However, as the previous two groups, non-parametric methods also require handcrafted local and global features.

Finally, methods based on supervised convolutional deep learning architectures are proposed in [10, 8, 23]. In [8], authors present a multi-scale convolutional deep architecture, which requires little pre-processing and performs well on several datasets. Nonetheless, this method requires post-processing based on CRF or the use of an optimal purity cover criterion to achieve state-of-the-art performance. In [23], authors present a recurrent version of convolutional networks, which obtains good results without any pre-processing nor post-processing and, to the best of our knowledge, is the first introducing a form of semantic feedback in convolutional deep networks.

2.2. Unsupervised dictionary learning

Unsupervised learning strategies have revealed to be helpful in tasks such as image classification [24, 4] and greedy layerwise pre-training of deep architectures [1].

Methods such as [17, 24, 22, 13, 5, 20] have been successfully used in the literature to extract sparse feature representations. However, the great majority require a certain number of meta-parameters to be tuned to achieve good performance and/or are computationally expensive. To overcome such limitations, we recently introduced a method [25] to pre-train deep architectures in a greedy layerwise fashion. The algorithm focuses on the sparsity properties of the output distribution. Given the output of a layer, the algorithm generates a sparse target with one "hot code" selected ensuring the same activation frequency among all outputs and optimizes for that specific target to learn the dictionary. The norm of the dictionary bases is limited to 1 in order to avoid degenerate solutions. The method is fast and meta-parameter free, highly simplifying the pre-training of deep architectures, and allows to learn a large set of complementary features thanks to the "over-regularization" imposed by

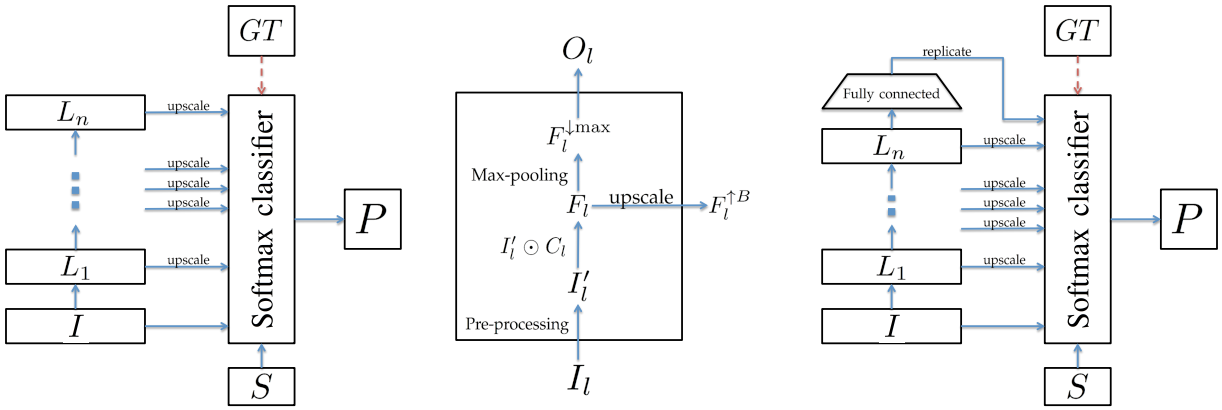


Figure 1. (a) The basic deep architecture coupled with a softmax classifier. (b) A detailed view of the computation within convolutional layers of the basic deep architecture. (3) The addition of a top fully connected layer, which provides a compact global appearance descriptor.

both the target construction and the limitation of the bases norm to 1.

3. Motivation and contributions

Current approaches in image parsing show pretty heterogeneous characteristics and, up to now, seem to fail in providing an architecture, which encompasses all the relevant (and desirable) characteristics for image parsing.

CRF, SSL and non-parametric methods have achieved a high level of sophistication in the design of both appearance and semantic features, and provide elegant and effective strategies to exploit the contextual information. However, their biggest limitation is that they require handcrafted appearance and semantic features. In some cases, the computational cost is also a relevant issue, making the unsupervised super-pixel segmentation a mandatory pre-processing step.

Methods based on convolutional deep networks are relatively new in the area of image parsing [8, 23]. The main limitations of these methods can be summarized as follows: (1) the learning is based on back-propagation (or, back-propagation through time), which is a slow learning technique and, in the case of [23] is the main reason to limit the number of recursions; (2) they do not present a form of global appearance or semantic features, which have been shown to improve the image parsing performance in [28].

With the aim to overcome the main limitations of deep architectures, while keeping their desirable features, the main contributions of the proposed work are:

1. To include *global appearance and semantic features* in a simple and intuitive way, while not increasing substantially the train and test computational burden.

2. To propose a deep convolutional architecture, which performs *top-down semantic feedback* and can be trained in an efficient way.

Table 1 summarizes the main properties of state-of-the-art convolutional deep networks compared to our proposal.

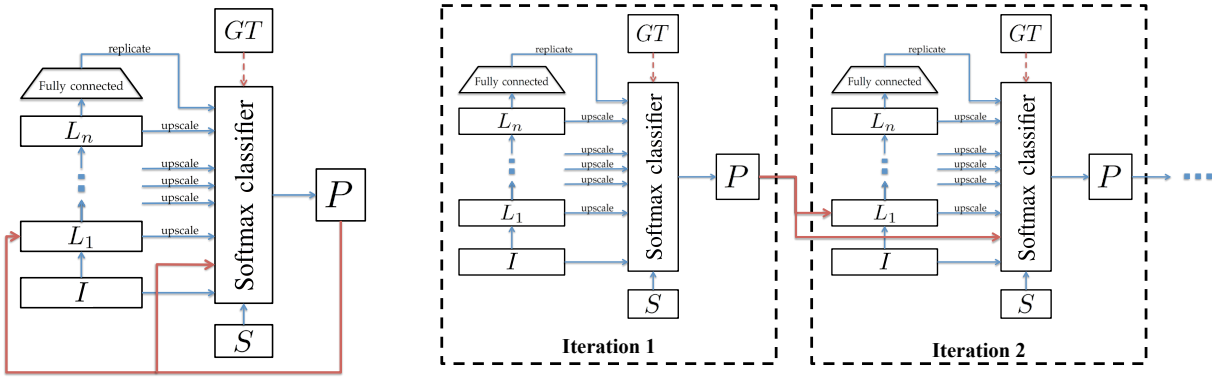
4. Method

We firstly explain how the basic architecture is defined and then we show the modifications we applied to it in the following subsections.

Figure 1(a) shows a schema of the basic architecture. Its main characteristics are the following: (1) convolutional filters are learned following the standard greedy pre-training strategy [1]; (2) unsupervised features are extracted from all the n layers of the deep architecture, providing features of different levels of complexity, from color and texture to higher level representations. Using the output of all layers of a deep hierarchy is not novel in deep convolutional networks (e.g. it has been used for pedestrian detection in [27]); however, to the best of our knowledge, it is the first time this strategy is used for image parsing.

The input image I is fed to the first layer L_1 and propagated until the topmost one L_n. Features at each layer F_l are extracted and upscaled, if necessary, to match the size of the input image. All features are concatenated into one vector; optionally a feature vector providing information on prior spatial classes distribution (S) can be added. The complete feature vector is then fed to a softmax classifier, which provides the posterior probability P of every class for all image pixels. During the training phase, the softmax classifier receives the ground truth GT information as input as well.

Figure 1(b) provides more details of the computation within a layer. The input data at layer l, I_l is firstly pre-processed, obtaining I'_l. The pre-processing is formed by local contrast and intensity normalization only for photo-



(a) (b)

Figure 2. (a) A possible schema providing loopy top-down semantic feedback. (b) The unrolled architecture for a two iterations case.

metric RGB input. Subsequently, we normalize data such that the average of each input over the training set is zero and is scaled to have unitary standard deviation, in order to accelerate the stochastic gradient descend optimization convergence during dictionary learning. A set of $N_{c,l}$ convolutional filters C_l are applied linearly to I'_l, $F_l = I'_l \odot C_l$, where $\odot$ denotes the application of a set of multidimensional convolutional filters. The features F_l take two different paths: (1) they are propagated as output O_l to the next layer by a max pooling operation, obtaining $F_l^{\downarrow \max}$; (2) F_l is upscaled, if necessary, to match the size of the input image I, obtaining $F_l^{\uparrow B}$; where B denotes that the resizing is performed by bicubic interpolation.

4.1. Top fully connected layer: a global appearance descriptor

Here we present our first contribution. Without loss of generality, let us assume that the top-most convolutional layer of the architecture in Figure 1(a) has 2×2 pixels spatial support; thus, the top-most output is a $2\times2\times N_{c,n}$ vector. This vector summarizes the content of the whole input image; it is however pretty high-dimensional, especially if $N_{c,n}$ is large, and possible correlations between the 4 spatial regions are not explicitly encoded.

Here we propose to add a top fully connected unsupervised layer that maps the $2\times2\times N_{c,n}$ elements into a smaller feature vector (see Figure 1(c)) that captures correlations between all the features within quadrants. The resulting feature is a global representation of the whole image. To feed this information to the softmax classifier, the vector must be replicated for all image pixels. The behaviour and the contribution of this additional layer to the deep architecture performance is quantitatively and qualitatively shown in section 5.2.

4.2. Unrolling loopy top-down semantic feedback

Our second contribution is detailed in this subsection. Figure 2(a) shows a possible way to introduce top-down semantic feedback in the proposed architecture. Since we use the output of all layers as features, the unfolding approach proposed in [23] cannot be employed in a straightforward way. Instead, we can replicate the architecture as many times as we want and feed all the deep networks (except the first one) with the posterior probability generated by the previous softmax classifier. Figure 2(b) shows our approach for the two iterations case. The parameters of different deep architectures and classifiers cannot be shared since the deep architectures are trained in an unsupervised way and their input data depends on the output of previous classifier. While this seems a disadvantage with respect to [23], it in fact allows to train the whole system without the need of an expensive training algorithm as the backpropagation through time (BPTT) used in [23]. The BPTT algorithm is the main limitation that does not allow to train the method in [23] with more than 3 compositions of the basic net. Another advantage of our method is that subsequent deep architectures can learn different features depending on the input data, thus being able to blend information from RGB data and posterior probability in an implicit way.

5. Experimental results

The experimental section is divided in three parts. The first one presents the dataset and the method setting. The second shows the advantage of using a fully connected unsupervised layer on top of the convolutional layers. The third part shows the advantage of the unrolled top-down feedback providing a quantitative and qualitative analysis of results.

5.1. Experiment setting

We tested our method on the "SIFT Flow dataset" [18], which is composed of 2688 images and presents 33 different categories. We use the standard training/test split as proposed in [18]. Differently than [8], we do not apply any kind of distortion on the images. As done in [23], we re-scale the input image by 1/2 to speed-up both training and testing phases. Nonetheless, for a fair comparison to other methods, the evaluation is performed by upscaling by a factor 2 the posterior probability P before comparing the maximum a posteriori labelling with the ground truth.

The spatial prior S is computed by accumulating the occurrences of each class in 33 separate maps (at full resolution) for all the training images; then the resulting maps are normalised and blurred with a Gaussian filter with $\sigma = 32$ pixels.

The basic architecture is composed of 6 convolutional layers with receptive field of 3×3 pixels and a spatial max pooling of non-overlapping 2×2 pixel regions. The size of the pooling region has been set to its minimal possible value so that the convolutional architecture could be as deep as possible. The receptive field is set to the minimal symmetric size, so to minimise the computational cost of convolutions and to delegate the learning of complex spatial configuration as much as possible to higher layers. Each convolutional layer has $N_{c,l} = 100$ output features. When employed, the top fully connected layer has 33 outputs. We set the number of outputs to these values for two practical reasons: (1) having 33 outputs for the fully connected layers allows a fair comparison with the spatial prior contribution in section 5.2; (2) when using the unrolled architecture the total number of features is 3 (RGB input) + 6×100 (6 conv layers) + 33 (spatial prior, S) + 33 (fully connected top layer) + 33 (posterior of previous iteration, P) = 702. This has been specifically done to have less features that the method in [8] (768 features).

Finally, for the unsupervised learning, we used 50k random samples per layer. The method in [25] does not require any meta-parameter. For the supervised training part of the method, we use 1% of ground truth data per iteration, corresponding to about 368k samples. The regularisation term in the softmax classifier is set to $\lambda = 10^{-3}$ in all experiments. Softmax parameters are learned using the LM-BFGS optimizer for a maximum of 500 iterations.

5.2. Unsupervised global image descriptor

In this sub-section, we provide experimental evidence that the top fully connected layer contributes in a substantial way to the performance of a deep unsupervised architecture for image parsing. Since the softmax is fed with the output of convolutional layers, the spatial prior S and the output of the fully connected layer, we separate these three components to analyse the system's performance.

Figure 3 shows the results in terms of global accuracy (left) and average per-class class accuracy (right) for four configurations. The first one is the basic 6 layer convolutional architecture (6L) using a total of 603 (3 + 100×6) features. The second configuration adds the spatial prior as an additional feature (6LS, 603 + 33 features). The third setting adds the top fully connected layer to the basic configuration (6LFC, 603 + 33 features). The last configuration adds both the top fully connected layer and the spatial prior (6LSFC, 603 + 33 + 33 features). As can be noticed, the spatial prior and the top fully connected layer allow to increase both performance measures in a significant way. However, the contribution of the top fully connected layer is clearly more important than the spatial prior. The increase in accuracy adding the top fully connected layer (+2%) is a clear sign of the advantage of using the proposed global image descriptor strategy. While the contribution of a spatial

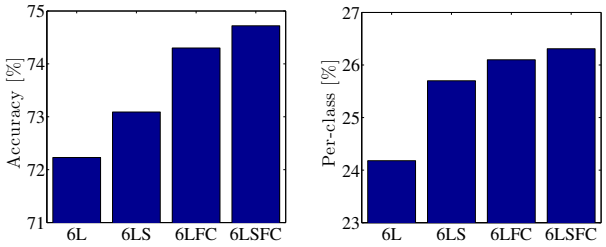


Figure 3. Quantitative comparison of 4 configurations with and without the top fully connected layer. Global accuracy (left) and per-class average accuracy (right). See text for details.

prior on the classes is easy to understand, the effect of the top fully connected layer is not trivial to grasp. The fully connected top layer summarizes the image information into one single (sparse) feature vector. Differently than [28], we do not perform a ranking nor we select a subset of images as a representative set of the "query" image. However, to show that we can obtain a similar effect, we use the output of the top fully connected layer as a global image descriptor and, given an input image, we perform a ranking based on the angle between the top layer output of the input image and all the outputs of the training set. The angle is a proper measure of (dis-)similarity since the softmax classifier is based on linear hyperplanes before the exponentiation. This procedure has been performed solely to show the representative power of the top fully connected layer.

Figure 4 shows some examples of the result of this procedure. The image on the left is the input image from the test set, and the four images in the same row are the retrieved and ranked images from the training set. Since the output of the top layer is replicated for all the image pixels, it is clear how this information acts as a global contextual priming for the classification. When used in conjunction with the top-down semantic feedback, this descriptor is able to blend appearance and semantic global features of the image

in a very compact way.

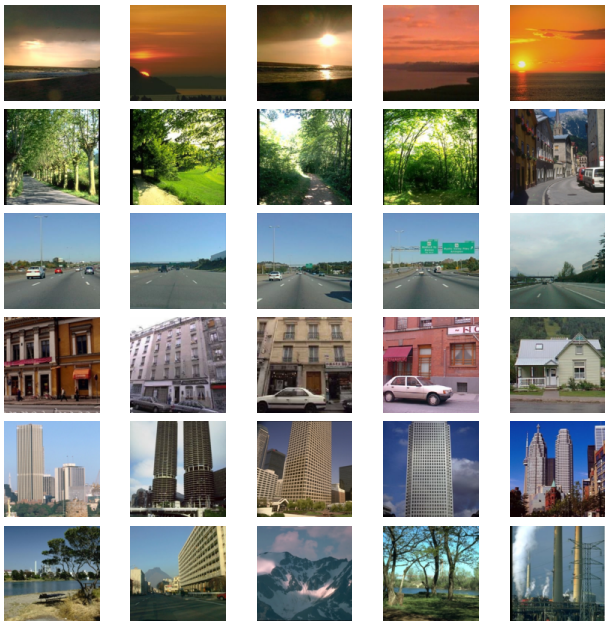


Figure 4. Six examples of ranking effect using the output of the top fully connected layer. See text for a detailed explanation.

5.3. Effect of top-down semantic feedback

The top-down semantic feedback allows the system to build subsequent parsing hypothesis and refine them progressively. Since the deep architecture learns features from the input image and the previous posterior probability, the system is able to learn appearance-semantic configurations from the second iteration. Figure 5 shows this effect in terms of pixel accuracy (left) and per-class average accuracy (right) as a function of iterations. The red dashed line represents the performance of a single iteration when using 10% of the training data (twice the quantity used in the 5 iterations). This clearly shows that the top-down feedback is the relevant component to obtain state-of-the-art-performance.

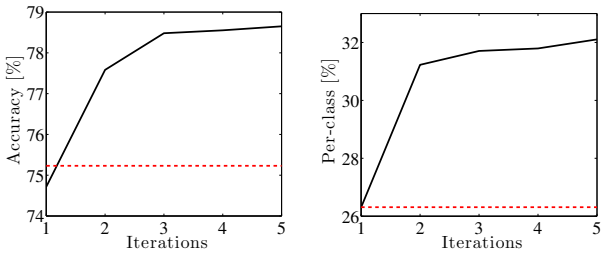


Figure 5. Global accuracy (left) and per-class average accuracy (right) as a function of the iterations when using the unrolled top-down semantic feedback architecture.

The per-class accuracy is also shown in Figure 6 separately for each class, ordered by decreasing accuracy. It can be seen that the top-down semantic feedback improves the per-class accuracy for almost all classes. The increment is particularly relevant when considering rare classes, showing that the semantic feedback contributes in learning a meaningful context. However, it can be noticed that some improvement is also present for the most frequent classes, showing that the algorithm is refining the parsing by better delineating boundaries and/or removing noisy classifications. Figure 7 shows some visual results, where the above mentioned effects can be observed. The result in the first row is especially interesting: the first iteration presents a lot of heterogeneous classes, as *road*, *mountain*, *car*, *sea*, *grass* and *field*; this can be explained by the poor global appearance image prior information (see last row of Figure 4). However, in subsequent iterations, both local and global information allow to reject inconsistent classes rapidly (first 3 iterations), while refining the boundaries of classification.

6. Discussion

Table 2 shows a comparison with the best performing convolutional methods on the SIFTflow dataset.

Table 2. Comparison with convolutional-based state-of-the-art methods in term of Global and average Per-class accuracy. The improvement from 1st to 5th iteration is provided in the last row.

	Global	Per-class
Farabet et al. [8]	78.5%	29.6%
Pinheiro et al. [23]	77.7%	29.8 %
1st iteration	74.7%	26.3%
5th iteration	**78.7%**	**32.1%**
Δ (5th - 1st)	+4.0%	+5.8%

Our method clearly outperforms previous convolutional approaches and gets really close to the best reported result on the SIFTflow dataset by the non-parametric method in [28], which obtains 79.2% and 33.8% for global and average per-class accuracy respectively. The contribution of the top-down semantic feedback is evident, allowing an improvement of 4% in accuracy and 5.8% in per-class average accuracy over 5 iterations.

An important characteristic of our method is that the training procedure is one order of magnitude faster than previous methods (see Table 1). We trained the 5 iterations system in less than 4 hours on a quad-core i7@2.3Ghz, using mildly-optimised Matlab code. Testing speed is comparable to previous convolutional methods and allows to compute a 128×128 pixels image in 4.4 seconds; which compares very well with the 20 seconds per image required by the method in [28]. We believe that implementing our method on GPUs

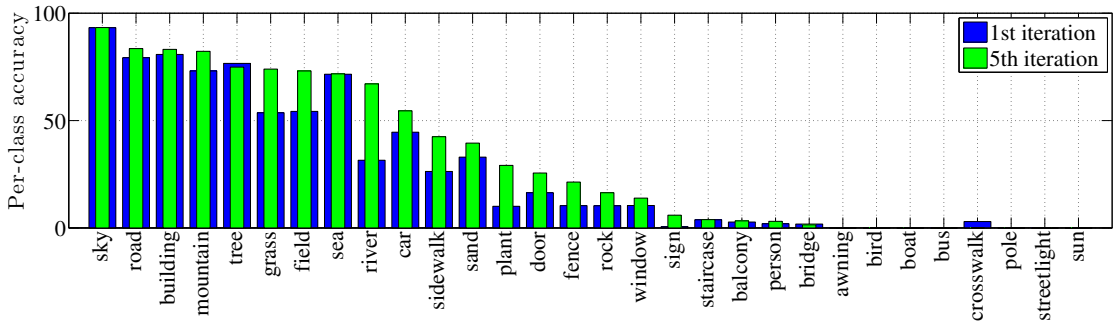


Figure 6. Per-class accuracy, in descending order for the 1st and 5th iteration.

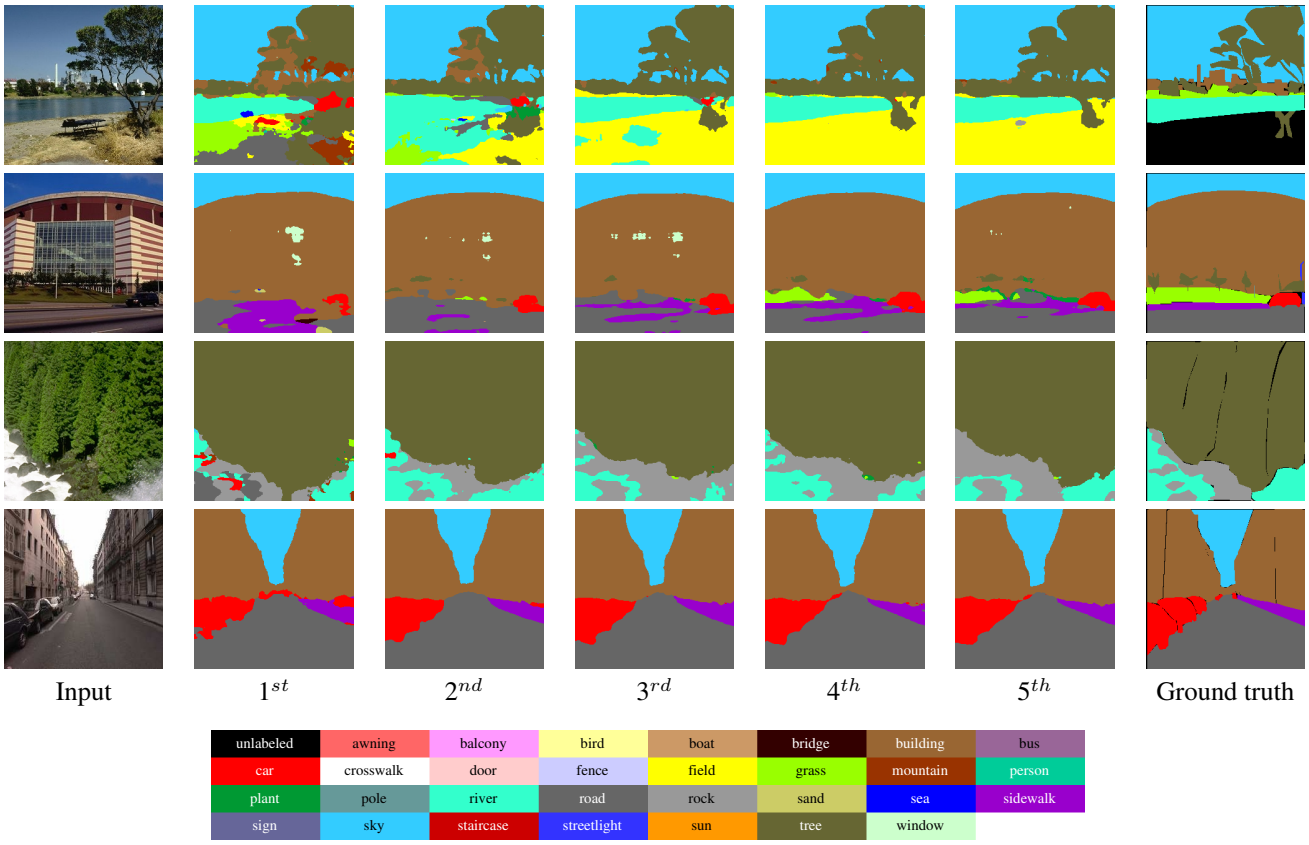


Figure 7. Results of image parsing for 4 test images (top). Color coded legend (bottom).

can achieve realtime performance.

All the experiments have been designed to focus on our method novelties and to compare in a fair way with previous convolutional deep networks. Nonetheless, the performance of our method can be improved in several ways: (1) the multi-scale approach proposed in [8] could be used in our approach by learning multiple deep architectures and feeding the result from different scales into the softmax classifier; (2) the softmax classifier regularisation term λ could be tuned to achieve the best possible accuracy; (3) the number of outputs at different layers could also be investigated to optimise the set of learned features for the given problem; (4) adding transformations on the training set, such as horizontal flipping, rotations, etc, could also help improve the system performance, as shown in [8].

The scalability of our proposal with respect to the number of classes seems to be a problem, since the posterior probability map is used as input in the next iteration, potentially making the input data very high dimensional. However, as a natural possible solution, a classical dimension-

ality reduction method, such as an auto-encoder, could be employed.

7. Conclusion

In this paper we proposed a novel strategy for efficient training of a system that performs top-down semantic feedback. Within this architecture, global appearance/semantic features can be added easily. We have shown that both contributions help in improving the state-of-the-art achieved by deep convolutional networks in the field of image parsing. Further investigation will be devoted to the issues raised in the discussion and to test the proposed method on a wider spectrum of datasets.

8. Acknowledgements

The work of C. Gatta is supported by a Ramon y Cajal Fellowship. The work of A. Romero is supported by an APIF-UB grant.

References

[1] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, pages 153–160, 2006. 1, 2, 3

[2] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. Gonzàlez. Harmony potentials - fusing global and local scale for semantic image segmentation. *IJCV*, 96(1):83–102, 2012. 2

[3] F. Ciompi, O. Pujol, and P. Radeva. Ecoc-drf: Discriminative random fields based on error-correcting output codes. *Pattern Recognition*, 47:2193–2204, 2014. 2

[4] A. Coates, H. Lee, and A. Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 214–223, 2011. 2

[5] A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, pages 921–928, 2011. 2

[6] D. Eigen and R. Fergus. In *CVPR*, pages 2799–2806. 2

[7] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *JMLR*, 11:625–660, Mar. 2010. 1

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE TPAMI*, 35(8):1915–1929, 2013. 1, 2, 3, 5, 6, 7

[9] C. Gatta and F. Ciompi. Stacked sequential scale-space taylor context. *IEEE TPAMI*, 2014. 2

[10] D. Grangier, L. Bottou, and R. Collobert. Deep convolutional networks for scene parsing. ICML 2009 Workshop on Learning Feature Hierarchies, June 2009. 1, 2

[11] Q.-X. Huang, M. Han, B. Wu, and S. Ioffe. A hierarchical conditional random field model for labeling and segmenting images of street scenes. In *CVPR*, pages 1953–1960. IEEE, 2011. 2

[12] J. Jiang and Z. Tu. Efficient scale space auto-context for image segmentation and labeling. In *CVPR*, pages 1810–1817. IEEE, 2009. 2

[13] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierachies for visual recognition. In *NIPS*, 2010. 2

[14] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009. 2

[15] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, pages 739–746. IEEE, 2009. 2

[16] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV (5)*, volume 6315 of *LNCS*, pages 239–253. Springer, 2010. 2

[17] H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area v2. In *NIPS*, pages 873–880, 2008. 2

[18] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE TPAMI*, 33(12):2368–2382, 2011. 2, 5

[19] D. Munoz, J. A. D. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *ECCV*, pages 57–70, 2010. 2

[20] J. Ngiam, P. W. Koh, Z. Chen, S. Bhaskar, and A. Y. Ng. Sparse filtering. In *NIPS*, pages 1125–1133, 2011. 2

[21] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kholi. Decision tree fields. In *ICCV*, pages 1668–1675, 2011. 2

[22] B. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997. 2

[23] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. *JMLR*, 1(32):82–90, 2014. 1, 2, 3, 4, 5, 6

[24] M. A. Ranzato, C. Poultney, S. Chopra, and Y. Lecun. Efficient learning of sparse representations with an energy-based model. In *NIPS*, pages 1137–1144, 2006. 2

[25] A. Romero, P. Radeva, and C. Gatta. No more meta-parameter tuning in unsupervised sparse feature learning. arXiv:1402.5766, 2014. 2, 5

[26] S. Rota Bulò, P. Kontschieder, M. Pelillo, and H. Bischof. Structured local predictors for image labelling. In *CVPR*, pages 3530–3537. IEEE, 2012. 2

[27] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*. 3

[28] G. Singh and J. Kosecka. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *CVPR*, pages 3151–3157. IEEE, 2013. 1, 2, 3, 5, 6

[29] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *IJCV*, 101(2):329–349, 2013. 2

[30] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE TPAMI*, 32(10):1744–1757, 2010. 2

[31] M. Ying Yang and W. Forstner. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In *IEEE/ISPRS workshop on Computer Vision for Remote Sensing of the Environment*, pages 196–203, 2011. 2