Deep Hierarchical Parsing for Semantic Segmentation

Abhishek Sharma[1], Oncel Tuzel[2], David W. Jacobs[1]
[1]Computer Science Department, University of Maryland at College Park. [2]Mitsubishi Electric Research Lab, Cambridge.

A popular approach for semantic segmentation is labeling each super-pixel with one of the required semantic categories. The rich diversity in the appearance of even simple concepts (sky, water, grass) due to the variation in lighting and view-point makes semantic segmentation very challenging. Contextual information from the entire image has been shown to be useful in resolving the ambiguity in super-pixel labeling [1, 2, 4]. The popular approaches for encoding context, MRF or CRF based image models, often make use of simple human-designed interaction potentials that limit the possible complexity of interactions between different parts of the image. This is done to avoid an intractable and computationally intensive inference step.

Recently, an elegant deep recursive neural network approach for semantic segmentation was proposed in [3], referred to as RCPN, Fig. 1. The main idea was to facilitate the propagation of contextual information from each super-pixel to every other super-pixel in a feed-forward manner through random binary parse trees $\mathcal{T}$ on super-pixels. The leaf nodes of $\mathcal{T}$ correspond to super-pixel features and the internal nodes correspond to the features of contiguous merged-regions that result from mergers, as per $\mathcal{T}$, of multiple super-pixel regions. RCPN consists of an assembly of four neural networks - *semantic mapper* (F_{sem}), *combiner* (F_{com}), *decombiner* (F_{dec}) and *categorizer* (F_{cat}). First, F_{sem} mapped visual features of the super-pixels $\mathbf{v}_i$ into semantic space features $\mathbf{x}_i$. This was followed by a recursive combination of semantic features of two adjacent image regions ($\mathbf{x}_i$ and $\mathbf{x}_j$), using F_{com} to yield the holistic feature vector of the entire image, termed the *root feature*. Next, the global information contained in the root feature was disseminated to every super-pixel in the image, using F_{dec}, followed by classification of the enhanced super-pixel features $\tilde{\mathbf{x}}_i$ by F_{cat}. RCPN has the potential to learn complex non-linear interaction between different parts of the image that resulted in impressive *real-time* performances on standard semantic segmentation datasets.

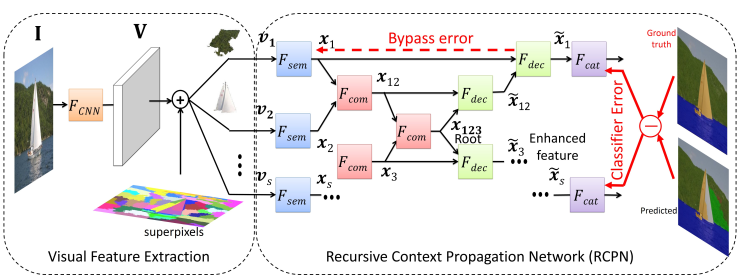
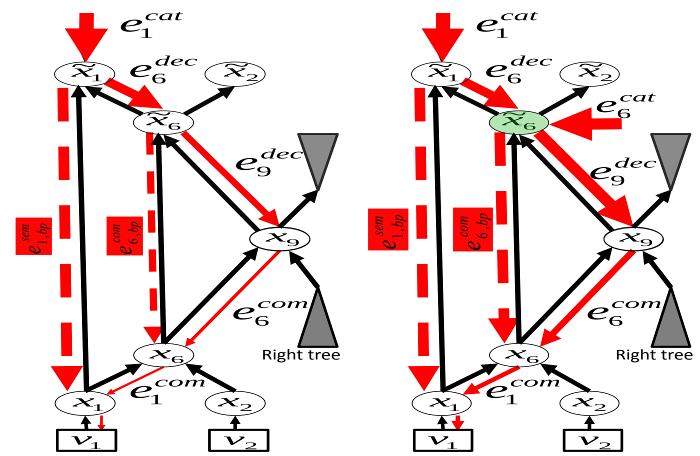


Figure 1: Flow diagram of RCPN with bypass-error path.

This paper shows that the presence of *bypass-error* paths in RCPN can lead to sub-optimal parameter learning and proposes a simple modification to improve the learning. Specifically, we propose to include the classification loss of *pure-nodes* to the RCPN loss function that originally consisted of classification loss of the super-pixels only. Pure-nodes are the internal nodes of $\mathcal{T}$ that correspond to merged-regions consisting of pixels of a single semantic category only. Therefore, pure-nodes can be used as classification targets for learning RCPN parameters. The resulting model is termed Pure-node RCPN or PN-RCPN. It leads to these three immediate benefits - a) increased labels per image; around 65% of all internal-nodes are pure-nodes for three different datasets b) deeper and stronger gradients and c) explicitly forcing the combiner to learn meaningful combinations of two image-regions.

We use an example to understand the benefits of PN-RCPN over RCPN, depicted with the help of Fig. 2(a) and Fig. 2(b), respectively. The figures show the left-half of a random parse tree for an image I with 5 super-pixels. We denote, $\mathbf{e}_i^{cat} \in \Re^{d_s}$ as the error at enhanced super-pixel nodes;



((a)) RCPN ((b)) PN-RCPN

Figure 2: Back-propagated error tracking to visualize the effect of bypass error, please see text for the meaning of variables.

$\mathbf{e}_k^{dec} \in \Re^{2d_s}$ as the error at the decombiner; $\mathbf{e}_k^{com} \in \Re^{2d_s}$ as the error at the combiner and $\mathbf{e}_i^{sem} \in \Re^{d_s}$ as the error at the semantic mapper, d_s is the dimensionality of the semantic space features and subscript bp indicates the bypass-error at a node. We assume a non-zero categorizer error signal for the first super-pixel only, ie $\mathbf{e}_{i \neq 1}^{cat} = \mathbf{0}$. These assumptions facilitate easier back-propagation tracking through the parse tree, but the conclusions drawn will hold for general cases as well.

From Fig. 2(a) we can see that there are two possible paths for $\mathbf{e}_1^{cat}$ to reach the combiner. One of them requires 2 layers ($\tilde{\mathbf{x}}_1 \to \tilde{\mathbf{x}}_6 \to \mathbf{x}_6$) and the other requires 3 layers ($\tilde{\mathbf{x}}_1 \to \tilde{\mathbf{x}}_6 \to \mathbf{x}_9 \to \mathbf{x}_6$). Similarly, $\mathbf{e}_1^{cat}$ can reach $\mathbf{x}_1$ through a 1 layer bypass path ($\tilde{\mathbf{x}}_1 \to \mathbf{x}_1$) or a several layers path through the parse tree. Due to gradient attenuation, the smaller the number of layers the stronger the back-propagated signal, therefore, bypass errors lead to $\mathbf{g}_{sem} \geq \mathbf{g}_{com}$. This can potentially render the combiner network inoperative and guide the training towards a network that effectively consists of a $N_{sem} + N_{dec} + N_{cat}$ layer network from the visual feature ($\mathbf{v}_i$) to the super-pixel label ($\mathbf{y}_i$). This results in little or no contextual information exchange between the super-pixels. A comparison of the gradient-strengths for different modules ($\mathbf{g}^{sem}$, $\mathbf{g}^{com}$, $\mathbf{g}^{dec}$ and $\mathbf{g}^{cat}$) reveals that for RCPN $\mathbf{g}^{cat} > \mathbf{g}^{dec} \approx \mathbf{g}^{sem} \gg \mathbf{g}^{com}$ that leads to bypassing the combiner, causing loss of contextual information. On the other hand, PN-RCPN gradients follow the natural order, $\mathbf{g}^{cat} > \mathbf{g}^{dec} > \mathbf{g}^{com} > \mathbf{g}^{sem}$, based on the distance from the initial label error which leads to a healthy context propagation via combiner.

Furthermore, PN-RCPN also provides us with reliable estimates of the internal node label distributions. We utilize the label distribution of the internal nodes to define a tree-style MRF, termed TM-RCPN, on the parse tree to model the hierarchical dependency between the nodes that leads to spatially smooth segmentation masks. Both PN-RCPN and TM-RCPN lead to significant improvement in terms of per-pixel accuracy, mean-class accuracy and intersection over union over RCPN.

[1] Roozbeh Mottaghi, Sanja Fidler, Jian Yao, Raquel Urtasun, and Devi Parikh. Analyzing semantic segmentation using hybrid human-machine crfs. *IEEE CVPR*, 2013.

[2] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. *IEEE CVPR*, 2014.

[3] Abhishek Sharma, Oncel Tuzel, and Ming Yu Liu. Recursive context propagation network for semantic segmentation. *NIPS*, 2014.

[4] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. Context-based vision system for place and object recognition. *CVPR*, 2003.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.