Recurrent Neural Networks for Semantic Instance Segmentation

Amaia Salvador[1], Míriam Bellver[2], Manel Baradad[1], Víctor Campos[2]
Ferran Marqués[1], Jordi Torres[2], Xavier Giro-i-Nieto[1]
[1]Universitat Politècnica de Catalunya [2]Barcelona Supercomputing Center
{amaia.salvador,xavier.giro,ferran.marques}@upc.edu, {miriam.bellver,victor.campos,jordi.torres}@bsc.es

1. Introduction

Semantic instance segmentation is defined as the task of assigning a binary mask and a categorical label to each object in an image. It is often understood as an extension of object detection where, instead of bounding boxes, accurate binary masks must be predicted. Current state of the art methods for semantic instance segmentation [7, 8, 11, 10, 4, 9] extend object detection pipelines based on object proposals [16] by incorporating an additional module that is trained to generate a binary mask for each object proposal. Such architectures follow a two-stage procedure, i.e. a set of object-prominent proposal locations are selected first, and then each of them is given a score, a categorical label and a binary mask. Typically, the number of selected locations is much greater than the actual number of objects that appear in the image, meaning that post-processing is needed to select the subset of predictions that better covers all the objects.Although in most recent works the two different stages (i.e. proposal generation and scoring) are optimized jointly [11, 10, 4, 9], the objective function still does not model the target task, but a surrogate one which is easier to handle at the cost of an additional filtering step.

While most systems analyze images in a single step, the human exploration of static visual inputs is actually a sequential process [14, 1] that involves reasoning about objects that compose the scene and their relationships. Inspired by this behavior, we design a model that performs a sequential analysis of the scene to deal with complex object distributions.

Recent works [17, 15] have also proposed sequential solutions for instance segmentation. Romera-Paredes & Torr [17] train a model composed of Convolutional LSTMs [21] that receives convolutional features from a pretrained FCN [12] and outputs the separate object segments for the image. A post-processing based on CRFs is applied to their final masks. Ren & Zemel [15] propose a complex multi-task recurrent pipeline for instance segmentation that predicts the box coordinates for a different object at each time step. These object coordinates are then used to extract a sub-region of the image from which a binary mask is predicted. Both [17, 15] are class-agnostic methods and, while [15] reports results for semantic instance segmentation benchmarks, class probabilities for their predicted segments are obtained from the output of a FCN [12] trained for semantic segmentation. To the best of our knowledge, our method is the first to directly tackle semantic instance segmentation with a fully end-to-end recurrent approach.

2. Model

Given an input image x, the goal of semantic instance segmentation is to provide a set of masks and their corresponding class labels, $y = \{y_1, \dots, y_n\}$. The cardinality of the output set, i.e. the number of instances, depends on the input image and thus the model needs to be able to handle variable length outputs. This poses a challenge for feedforward architectures, which emit outputs of fixed size. Similarly to previous works involving sets [20, 19, 17], we propose a recurrent architecture that outputs a sequence of masks and labels, $\hat{y} = (\hat{y}_1, \dots, \hat{y}_{\hat{n}})$. At any given time step $t \in \{1, \dots, \hat{n}\}$, the prediction is of the form $\hat{y}_t = \{\hat{y}_m, \hat{y}_b, \hat{y}_c, \hat{y}_s\}$, where $\hat{y}_m \in [0,1]^{h \times w}$ is the binary mask, $\hat{y}_b \in [0,1]^4$ are the bounding box coordinates normalized by the image dimensions, $\hat{y}_c \in [0,1]^C$ are the probabilities for the C different categories, and $\hat{y}_s \in [0,1]$ represents the objectness score, which is the stopping criterion at test time. Obtaining bounding box annotations from the segmentation masks is straightforward and it adds an additional training signal, which resulted in better performing models in our experiments.

We design an encoder-decoder architecture that resembles typical ones from semantic segmentation works [12, 18], where skip connections from the layers in the encoder are used to recover low level features that are helpful to obtain accurate segmentation outputs. The main difference between these works and ours is that our decoder is recurrent, enabling the prediction of one instance at a time instead of a single semantic segmentation map where all objects are present, thus allowing to naturally handle variable length outputs. Figure 1 shows the details of the recurrent decoder for a single time step.

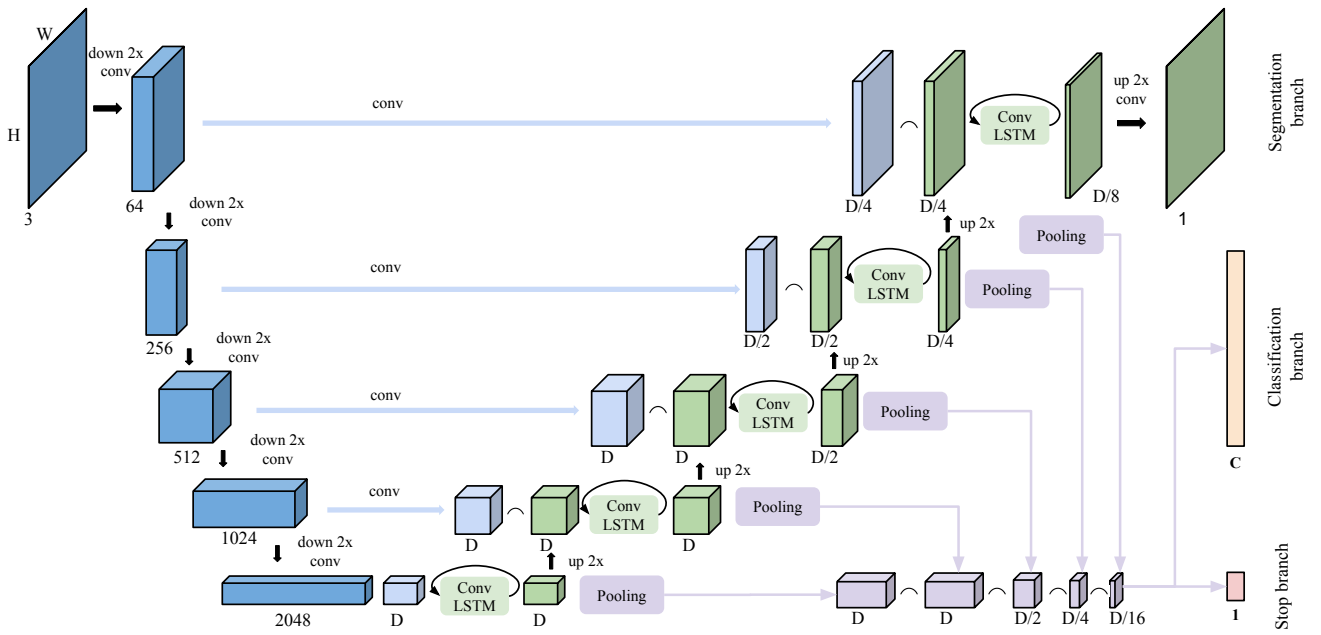

Figure 1: Our proposed recurrent architecture for semantic instance segmentation.

	Rec	Cls	Pascal VOC	CVPPP		Cityscapes			
			$AP_{person,50}$	SBD ↑	DiC ↓	AP	AP_{50}	AP_{car}	$AP_{car,50}$
Ren [15]	✗	✗	–	**84.9**(±**4.8**)	**0.8**(±**1.0**)	**9.5**	**18.9**	**27.5**	41.9
Romera-Paredes [17]	✓	✗	46.6	56.8(±8.2)	1.1(±0.9)	–	–	–	–
Romera-Paredes [17] + CRF	✓	✗	50.1	66.6(±8.7)	1.1(±0.9)	–	–	–	–
Ours	✓	✓	**60.7**	74.7(±5.9)	1.1(±0.9)	7.8	17.0	25.8	**45.7**

Table 1: Comparison against state of the art sequential methods for semantic instance segmentation. We specify whether the method is recurrent (Rec) and produces categorical probabilities (Cls).

3. Experiments

We evaluate our models on three benchmarks previously used for semantic instance segmentation (Pascal VOC 2012 [6], CVPPP Plant Leaf Segmentation [13] and Cityscapes [3]) that differ from each other in terms of the average amount of objects per image. Table 1 compares our results against Romera-Paredes & Torr [17], and Ren & Zemel [15].

We first train and evaluate our model with the Pascal VOC dataset. In Table 1 we compare our method with the recurrent model in [17], whose approach is the most similar to ours. However, since they train and evaluate their method on the person category only, we report the results for this category separately despite that our model is trained for all 20 categories. We outperform their results by a significant margin (AP_{50} of 46.6 vs. 60.7), even in the case in which they use a post processing based on CRFs, reaching an AP_{50} of 50.1.

In the case of the CVPPP dataset, our method also outperforms the one in [17] by a significant margin. However, the sequential model in [15] obtains better results in this benchmark. Their method incorporates an input preprocessing stage and involves multi-stage training with different levels of supervision. In contrast with [15], our method directly predicts binary masks from image pixels without imposing any constraints regarding the intermediate feature representation. Although the number of objects is much higher in this benchmark than in Pascal VOC, our model is able to accurately output one object at a time.

Our performance on Cityscapes is comparable to the results of the only sequential method previously evaluated on this dataset [15], but does not meet state of the art results obtained by non-sequential methods, which reach AP_{50} figures of 58.1 [9], 35.9 [5] and 35.3 [2]. It is also worth noting that the classification scores in [15] are provided by a separate module trained for the task of semantic segmentation.

Acknowledgements

This work was partially supported by the Spanish Ministry of Economy and Competitivity under contract TIN2012-34557 by the BSCCNS Severo Ochoa program (SEV-2011-00067), and contract TEC2016-75976-R. We acknowledge the support of NVIDIA Corporation for the donation of GPUs.

References

[1] T. A. Amor, S. D. Reis, D. Campos, H. J. Herrmann, and J. S. Andrade Jr. Persistence in eye movement during visual search. *Scientific reports*, 6:20815, 2016.

[2] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[4] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[5] B. De Brabandere, D. Neven, and L. Van Gool. Semantic instance segmentation with a discriminative loss function. In *CVPRW*, 2017.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge. *IJCV*, 2010.

[7] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.

[8] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[10] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017.

[11] X. Liang, Y. Wei, X. Shen, Z. Jie, J. Feng, L. Lin, and S. Yan. Reversible recursive instance-level object segmentation. In *CVPR*, 2016.

[12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[13] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89, 2016.

[14] G. Porter, T. Troscianko, and I. D. Gilchrist. Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology*, 2007.

[15] M. Ren and R. S. Zemel. End-to-end instance segmentation with recurrent attention. In *CVPR*, 2017.

[16] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[17] B. Romera-Paredes and P. H. S. Torr. Recurrent instance segmentation. In *ECCV*, 2016.

[18] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[19] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *NIPS*, 2016.

[20] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *NIPS*, 2015.

[21] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation. In *NIPS*, 2015.