Recurrent Instance Segmentation

B. Romera-Paredes, P. H. S. Torr
University of Oxford, UK

Instance Segmentation Problem

Instance segmentation is the problem of detecting and delineating each distinct object of interest appearing in an image.

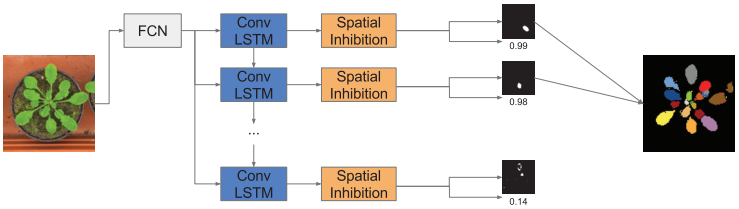
Most approaches proposed for instance level segmentation are based on a pipeline of modules whose learning process is carried out independently of each other.

Our approach

Humans count sequentially, using spatial memory in order to keep track of the accounted locations [4].

Driven by this insight, our purpose is to build a learning model capable of segmenting the instances of an object in an image sequentially, keeping the current state in an internal memory.

We rely on recurrent neural networks (RNNs), which exhibit both the ability to produce sequential output, and the ability to keep a state or memory along the sequence.

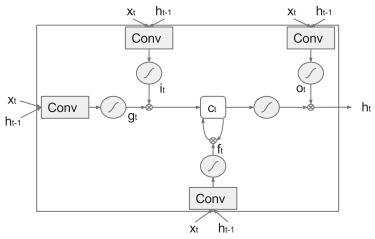


Our primary contribution is the development of an *end-to-end* approach for instance segmentation based on:

- RNNs containing convolutional layers.
- A principled loss function for instance segmentation.

ConvLSTM

We are interested in recurrent structures based on convolutions, in which the intermediate representations of the images preserve the spatial information.

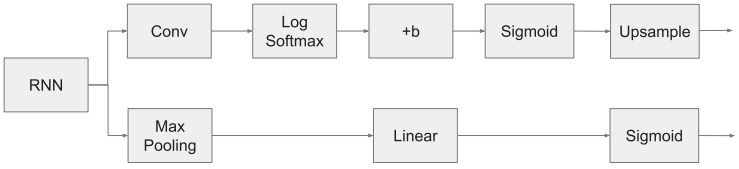


We use ConvLSTM: similar to an LSTM one, where the fully connected layers in each gate are replaced by convolutions.

Spatial inhibition module

Two outputs are produced by our model at each time:

1. A map that indicates which pixels compose the object that is segmented in the current iteration. Thus, the function learned for this stage has to be able to discriminate one, and only one instance, filtering out everything else.
2. The estimated probability that the current segmented candidate is an object.



Loss function

Our model predicts both a sequence of masks, $\hat{\mathbf{Y}} = \left\{\hat{\mathbf{Y}}_\mathbf{1}, \hat{\mathbf{Y}}_\mathbf{2}, \ldots, \hat{\mathbf{Y}}_{\hat{\mathbf{n}}}\right\}$, and a confidence score associated to those masks $\mathbf{s} = \{s_1, s_2, \ldots, s_{\hat{n}}\}$.

Given that the sequential order in the prediction does not matter, we find the optimal matching between predicted and ground truth masks. This can be found out efficiently by means of the Hungarian algorithm

$f_{\text{IoU}}(\hat{\mathbf{y}}, \mathbf{y})$			
	0	0.89	0.12
	0	0.01	0.99
	0.99	0.02	0

Our objective function is:

$$\ell(\hat{\mathbf{Y}}, \mathbf{s}, \mathbf{Y}) = \min_{\delta \in \mathcal{S}} - \sum_{\hat{t}=1}^{\hat{n}} \sum_{t=1}^{n} f_{\text{IoU}}\left(\hat{\mathbf{Y}}_{\hat{\mathbf{t}}}, \mathbf{Y}_\mathbf{t}\right) \delta_{\hat{t},t} + \lambda \sum_{t=1}^{\hat{n}} f_{\text{BCE}}\left([t \leq n], s_t\right)$$

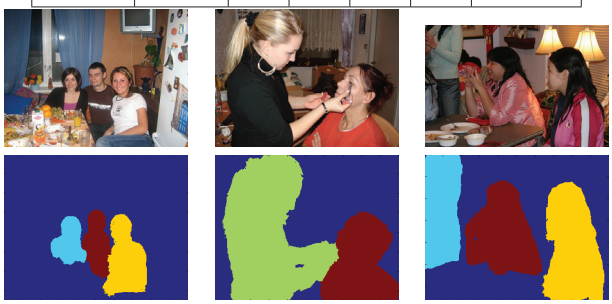
$\mathcal{S}$ is the set of all possible matchings between prediction and ground truth.

$f_{\text{IoU}}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\langle \hat{\mathbf{y}}, \mathbf{y} \rangle}{\|\hat{\mathbf{y}}\|_1 + \|\mathbf{y}\|_1 - \langle \hat{\mathbf{y}}, \mathbf{y} \rangle}$ is a relaxed version of the intersection over union (IoU).

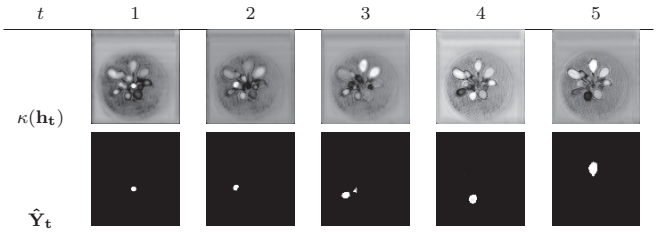
$f_{\text{BCE}}(a, b) = -\left(a \log(b) + (1-a) \log(1-b)\right)$ is the binary cross entropy.

Experiments

	Baseline	[1]	[2]	[3]	RIS	RIS+CRF
$AP^r(0.5)$	45.8	48.3	47.9	48.8	46.7	**50.1**
AP^r Ave	39.6			42.9	41.9	**43.7**



	IPK	Not	MSU	Wag	PRI	RIS	RIS+CRF
DiC	−1.9 (2.5)	−3.6 (2.4)	−2.3 (1.6)	−0.4 (3.0)	0.8 (1.5)	**0.2** **(1.4)**	**0.2** **(1.4)**
$\lvert DiC \rvert$	2.6 (1.8)	3.8 (2.0)	2.3 (1.5)	2.2 (2.0)	1.3 (2.0)	**1.1** **(0.9)**	**1.1** **(0.9)**
SBD	**74.4** **(4.3)**	68.3 (6.3)	66.7 (7.6)	71.1 (6.2)	–	56.8 (8.2)	66.6 (8.7)



Conclusion

Learning end-to-end instance segmentation is possible by means of a recurrent neural network.

We have shown that a recurrent structure is able to track visited areas in the image as well as to handle occlusion among instances.

References

[1] Yi-Ting Chen, Xiaokai Liu, and Ming-Hsuan Yang. Multi-instance object segmentation with occlusion handling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3470–3478, 2015.

[2] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, pages 297–312. Springer, 2014.

[3] Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Jianchao Yang, Liang Lin, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *arXiv preprint arXiv:1509.02636*, 2015.

[4] Gillian Porter, Tom Troscianko, and Iain D Gilchrist. Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology*, 60(2):211–229, 2007.