Attention to Scale: Scale-aware Semantic Image Segmentation

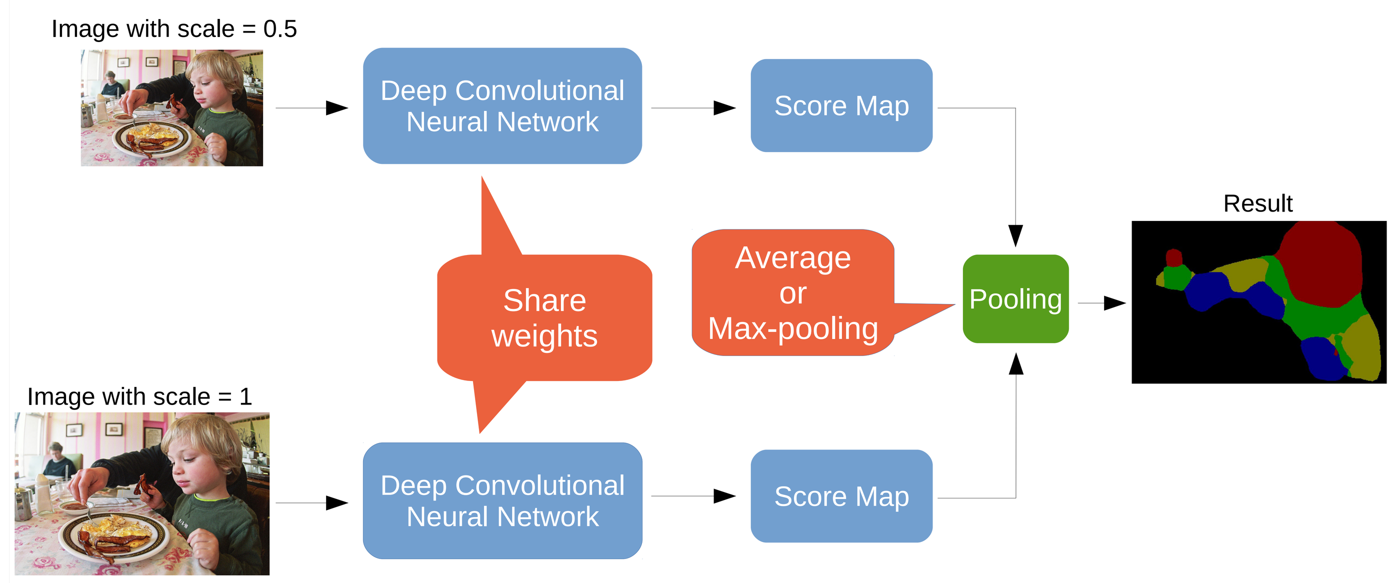
Liang-Chieh Chen[1] Yi Yang[2] Jiang Wang[2] Wei Xu[2] Alan L. Yuille[1,3]
[1]UCLA [2]Baidu USA [3] JHU

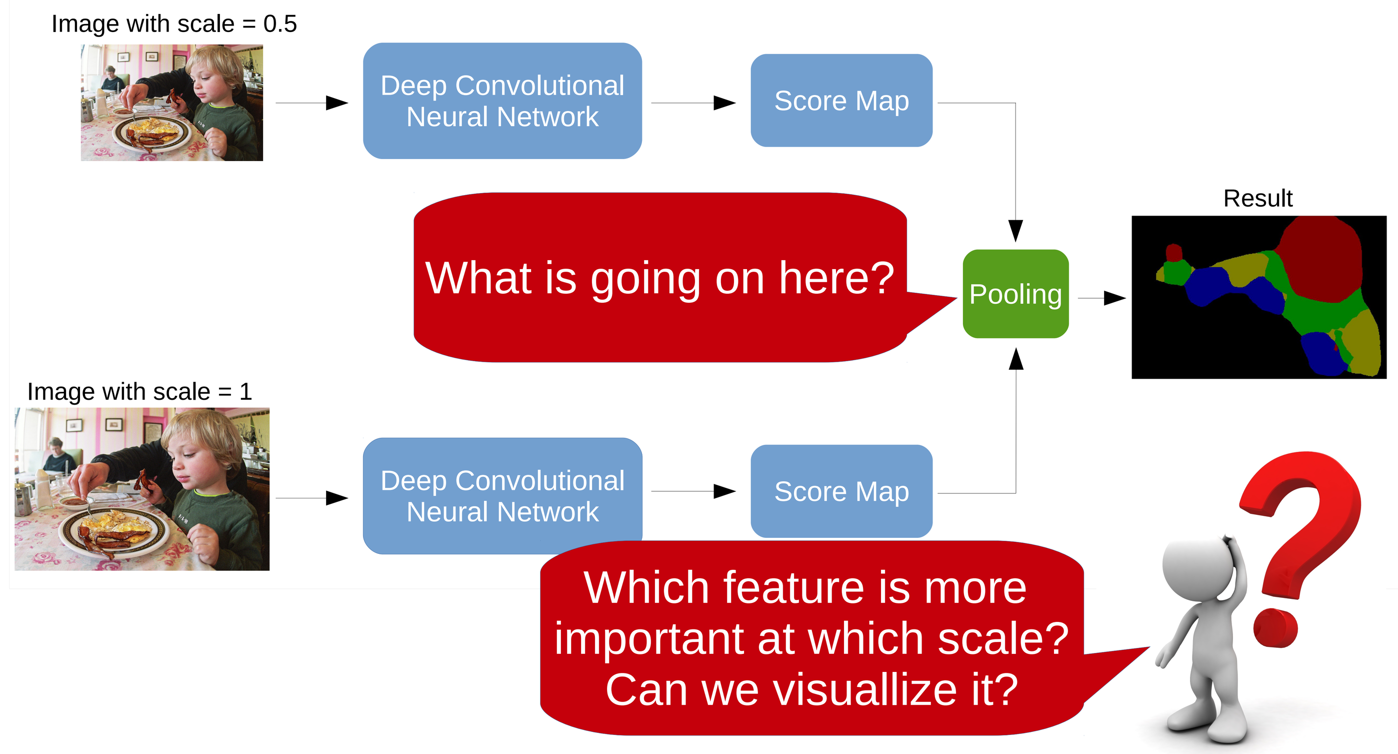
Motivation

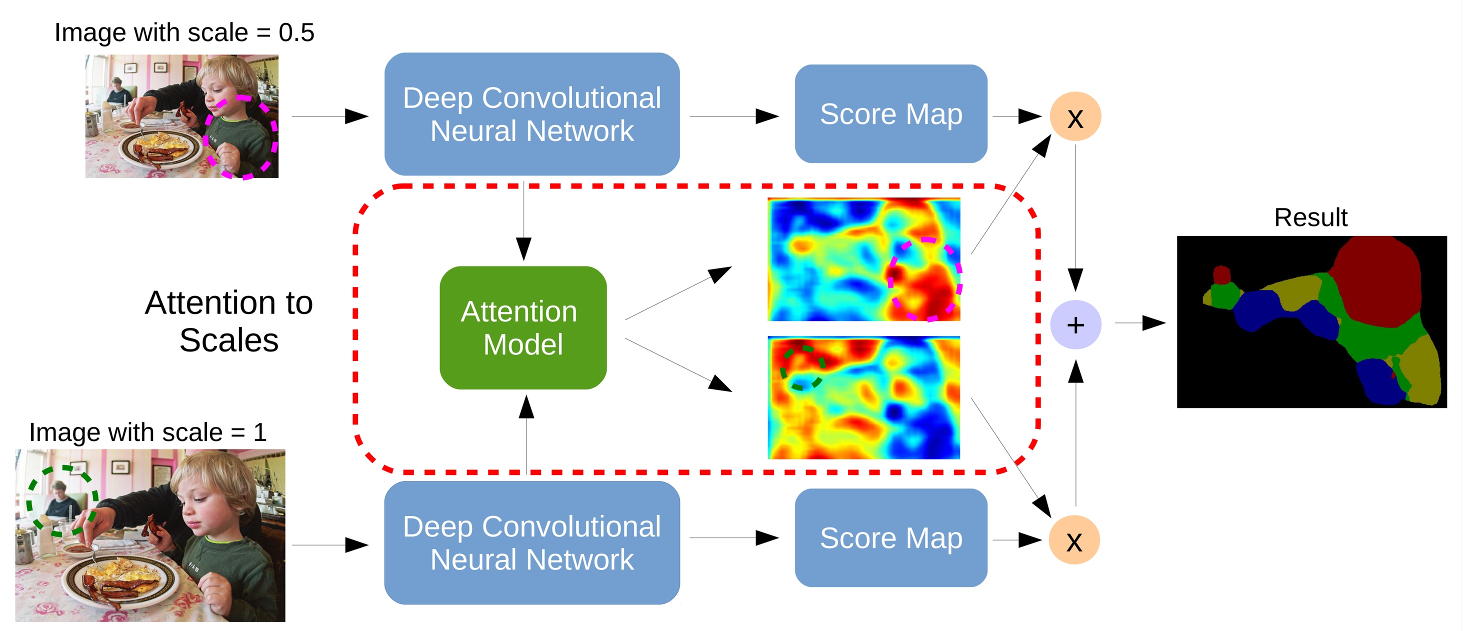
Multi-scale features key of s-o-a semantic segmentation models.

(a) small-scale person large-scale train (b) large-scale person small-scale train (c) persons of several scales

Model Illustration







Attention Model: Multi-scale Features

- Suppose input image resized to several scales $s \in \{1, ..., S\}$.
- Input with scale s produces a score map $f^s_{i,c}$, (i over pixels, and c over object classes).
- Let $g_{i,c}$ be the weighted sum of score maps at (i, c) for all scales

$$g_{i,c} = \sum_{s=1}^{S} w_i^s \cdot f_{i,c}^s \tag{1}$$

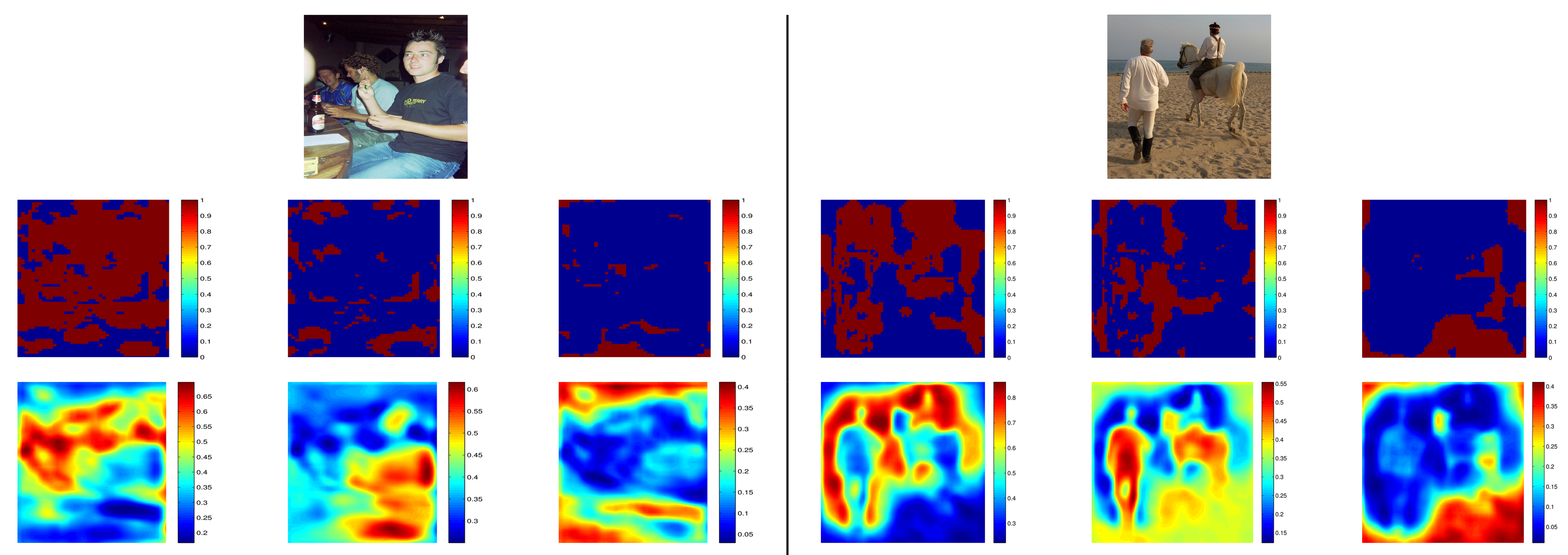
The weight w_i^s is computed by

$$w_i^s = \frac{\exp(h_i^s)}{\sum_{t=1}^{S} \exp(h_i^t)} \tag{2}$$

where h_i^s is score map by *attention* model.

- w_i^s reflects importance of feature at position i and scale s.
- Visualize attention for each scale by visualizing w_i^s.
- Average- or max-pooling over scales are two special cases.

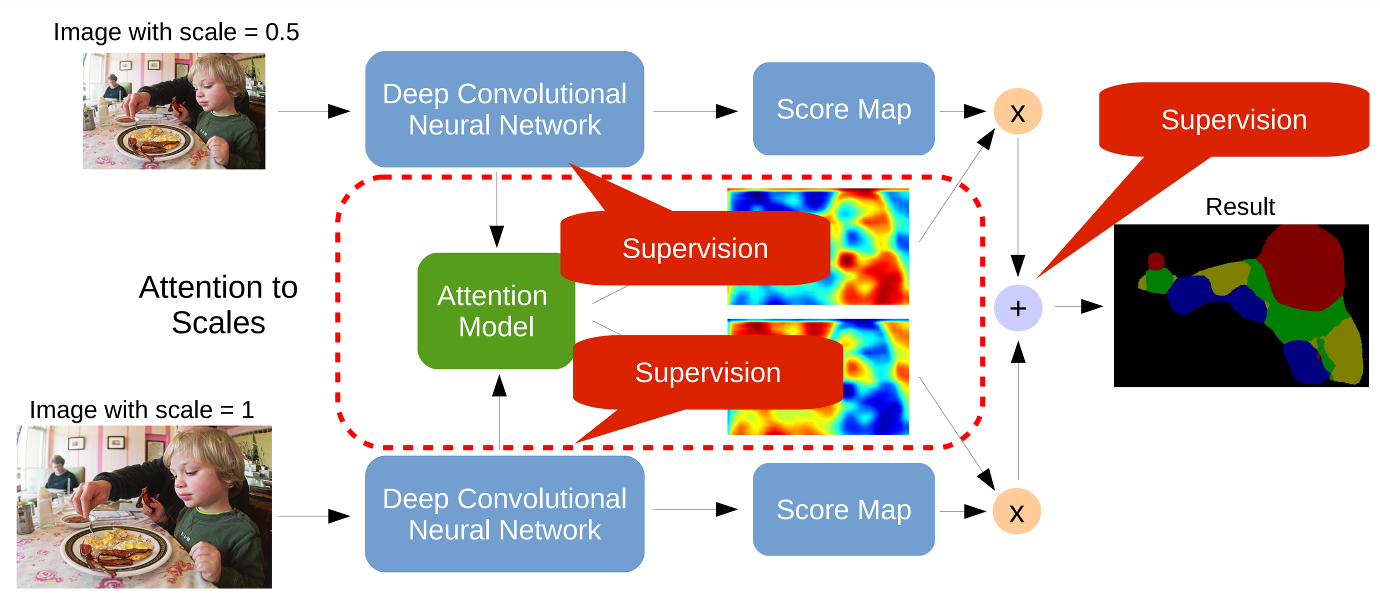
Learned Attention: Max vs. Attention



(a) Scale-1 Att. (b) Scale-0.75 Att. (c) Scale-0.5 Att. (a) Scale-1 Att. (b) Scale-0.75 Att. (c) Scale-0.5 Att.

- Scale-1 attention → small-scale objects.
- Scale-0.75 attention → middle-scale objects.
- Scale-0.5 attention → large-scale objects or background.

Extra Supervision



PASCAL VOC 2012

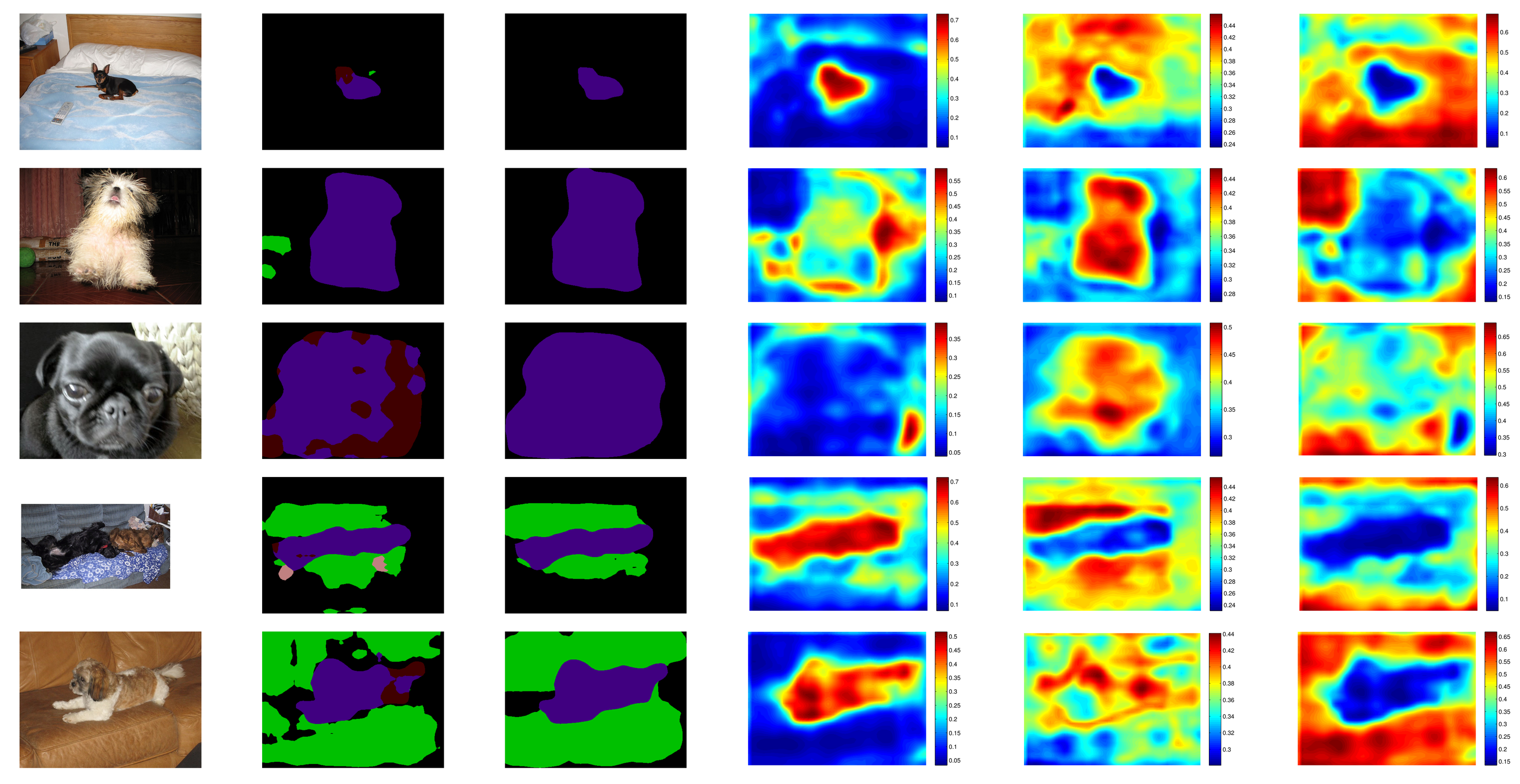
Baseline: DeepLab-LargeFOV	67.58	
Merging Method		w/ E-Supv
Scales = {1, 0.75, 0.5}		
Max-Pooling	69.70	70.06
Average-Pooling	68.82	70.55
Attention	69.47	**71.42**

(a) val set

Method	mIOU
DeepLab-CRF-COCO-LargeFOV	72.7
DeepLab-MSc-CRF-COCO-LargeFOV	73.6
DeepLab-CRF-COCO-LargeFOV-**Attention**	75.1
DeepLab-CRF-COCO-LargeFOV-**Attention**+	75.7

(b) test set

Segmentation Results



(a) Image (b) Baseline (c) Our model (d) Scale-1 Att. (e) Scale-0.75 Att. (f) Scale-0.5 Att.

Conclusion

- Using multi-scale inputs > single scale input.
- Attention model brings better performance and allows to visualize the importance of features.
- Adding extra supervision is essential for better performance.
- Try it out! Source code and trained models available at `http://liangchiehchen.com/projects/DeepLab.html`.

References

[1] C. Farabet et al. Learning hierarchical features for scene labeling. *PAMI*, 2013.
[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.
[3] G. Lin et al. Efficient piecewise training of deep structured models for semantic segmentation. *arXiv:1504.01013*, 2015.