Sharing spatial information in CNNs: depth, decoders, dilation, and multi-grids

Stephen Krewson

April 4, 2018

Abstract

Convolutional neural networks (CNNs) learn to map a two-dimensional input image (usually with three color channels) to a vector of class probabilities. The success of these models on the ImageNet task (1000 classes of real-world objects, one class per training image) dates back to AlexNet (2012) and is being extended to more difficult visions tasks such as semantic segmentation. This type of segmentation requires each pixel to be assigned to a class region; because normal CNN architectures destroy spatial information, a wide variety of new methods are being explored.

In this talk, I present two of the newest models: multigrids (Ke, Maire, and Yu 2016) and atrous or dilated networks. (Chen et al. 2018; Chen et al. 2017) Loosely descended from ResNets (He et al. 2014) and iterating on the idea of spatial pyramid pooling, these two architectures are closely related but differ in some key respects. I discuss these differences (encoder-decoder vs. learned routing for up/down-sampling) as well as why these architectures are suitable for tasks that are too difficult for normal CNNs such as dense prediction, recognition of rotated *and* transposed objects, and sharp object boundaries. Both the computational efficiency and required data augmentation of these models will be considered.

Spotlight Questions

- Do extensions of CNNs provide better expressivity or do they just make the training process more tractable for more difficult tasks? (thanks to Holly Rushmeier for this question)
- What is the relation between "attention" in multigrids and attention in an LSTM?

- Do the spatial metaphors of pyramid, multi-grid, and hypercolumn have any substantive connection to the visual cortex?

Helpful links

- TensorFlow DeepLabV3+ repo
- Independent replication of DeepLabV3+
- Original DeepLab website
- Andrew Ng's CNN video tutorial playlist
- Stanford machine learning course notes

References

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *ArXiv E-Prints*, February.

Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. "Rethinking Atrous Convolution for Semantic Image Segmentation." *CoRR* abs/1706.05587. http://arxiv.org/abs/1706.05587.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." *CoRR* abs/1406.4729. http://arxiv.org/abs/1406.4729.

Ke, Tsung-Wei, Michael Maire, and Stella X. Yu. 2016. "Neural Multigrid." *CoRR* abs/1611.07661. http://arxiv.org/abs/1611.07661.