Microsoft Wins ImageNet 2015 through Highway Net (or Feedforward LSTM) without Gates

Jürgen Schmidhuber

Microsoft Research dominated the ImageNet 2015 contest with a very deep neural network of 150 layers [1]. Congrats to Kaiming He & Xiangyu Zhang & Shaoqing Ren & Jian Sun on the great results [2]!

Their Residual Net or **ResNet** [1] of December 2015 is a special case of our **Highway Networks** [4] of May 2015, the first very deep feedforward networks with hundreds of layers. Highway nets are essentially feedforward versions of recurrent Long Short-Term Memory (LSTM) networks [3] **with** forget gates (or gated recurrent units) [5].

Let *g, t, h* denote non-linear differentiable functions. Each non-input layer of a Highway Net computes $g(x)x + t(x)h(x)$, where x is the data from the previous layer. (Like LSTM [3] with forget gates [5] for recurrent networks.)

The CNN layers of ResNets [1] do the same with $g(x)=1$ (a typical Highway Net initialisation) and $t(x)=1$, essentially like a Highway Net or a feedforward LSTM [3] **without** gates.

This is the basic ingredient required to overcome the fundamental deep learning problem of vanishing or exploding gradients. The authors mention it [1], but do not mention my very first student Sepp Hochreiter (now professor) who identified and analyzed it in 1991, years before anybody else did [6].

Apart from the quibbles above, I liked the paper [1] a lot. LSTM concepts keep invading CNN territory [e.g., 7a-e], also through GPU-friendly multi-dimensional LSTMs [8].

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. TR arxiv:1512.03385, Dec 2015.

[2] ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015): Results

[3] S. Hochreiter, J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735-1780, 1997. Based on TR FKI-207-95, TUM (1995). PDF. *Led to a lot of follow-up work, and is now heavily used by leading IT companies all over the world.*

[4] R. K. Srivastava, K. Greff, J. Schmidhuber. Highway networks. TR arxiv:1505.00387 (May 2015) and arXiv:1507.06228 (July 2015). Also at NIPS'2015.

[5] F. A. Gers, J. Schmidhuber, F. Cummins. Learning to Forget: Continual Prediction with LSTM. Neural Computation, 12(10):2451-2471, 2000. PDF.

[6] S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, TU Munich, 1991. Advisor: J. Schmidhuber. Overview.

[7a] 2011: First superhuman CNNs
[7b] 2011: First human-competitive CNNs for handwriting
[7c] 2012: First CNN to win segmentation contest
[7d] 2012: First CNN to win contest on object discovery in large images
[7e] Deep Learning. Scholarpedia, 10(11):32832, 2015

[8] M. Stollenga, W. Byeon, M. Liwicki, J. Schmidhuber. Parallel Multi-Dimensional LSTM, with Application to Fast Biomedical Volumetric Image Segmentation. NIPS 2015; arxiv:1506.07452.

Can you spot the Fibonacci pattern in the graphics above?