Short note on recurrent neural networks

Grzegorz Chrupała

1 Simple Recurrent Nets

Simple recurrent nets (SRN) were introduced by Elman (1990, 1991):

$$h_t = \sigma(W x_t + U h_{t-1}) \tag{1}$$

where σ is an elementwise nonlinearity such as the inverse logit:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

2 Long Short-term Memory

Long Short-Term Memory networks were introduced by Hochreiter and Schmidhuber (1997) in order to remedy problems with the backpropagation in SRN. The formulation is the following:

$$\begin{aligned}
i_t =& \sigma(W_i x_t + U_i h_{t-1}) \\
f_t =& \sigma(W_f x_t + U_f h_{t-1}) \\
o_t =& \sigma(W_o x_t + U_o h_{t-1}) \\
c_t =& f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1}) \\
h_t =& o_t \odot c_t
\end{aligned}$$

where $\odot$ is elementwise multiplication. As can be seen, at each time step, LSTM has two separate states: the hidden state h_t, and the memory cell c_t. Three gates control whether to forget the current cell value (f), whether it should read its input (i) and whether to output the new cell value (o). This video lecture by Hinton has a good explanation of LSTMs: `https://www.youtube.com/watch?v=lsV5rFbs-K0&list=PLnnr1O8OWc6YM16tj9pdhBZOS9tDktNrx&index=5`

3 Gated Recurrent Units

Gated Recurrent Units (GRU) were first introduced by Cho et al. (2014) and Chung et al. (2014) as a simpler alternative to LSTMs. In a GRU, activation at

time t is the linear combination of previous activation, and candidate activation:

$$\begin{aligned} h_t =& (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \\ z_t =& \sigma_s(W_z x_t + U_z h_{t-1}) \\ \tilde{h}_t =& \sigma(W x_t + U(r_t \odot h_{t-1})) \\ r_t =& \sigma_s(W_r x_t + U_r h_{t-1}) \end{aligned}$$

The partially linear dependence between current and previous state makes it easier for these networks to be trained via backpropagation.

References

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Deep Learning and Representation Learning Workshop*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2-3):195–225.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.