Neural LMs

Feed forward

- Bengio et al.(2003, 2001) showed that NNLMs could be scaled reasonably well
- Feed forward based language model.
- p(w_t | w_{t-1}.. w_{t-n}). n was typically 4.
- Started with one-hot encodings of 4 words, w_{t-1} …, w_{t-4}
- Projection operator C for lower dimension embedding of words

Procedure

$w \in \mathbb{R}^{|V|}, C \in \mathbb{R}^d \times \mathbb{R}^{|V|} \qquad C(w) = C.w$

$[w_{t-1}, w_{t-2} ... w_{t-n}] \rightarrow C.[w_{t-1}, w_{t-2} ... w_{t-n}]$

$f(x) = C.[w_{t-1}, w_{t-2} ... w_{t-n}] \in \mathbb{R}^{nd}$

$A \in \mathbb{R}^{nd} \times \mathbb{R}^h \qquad \sigma = \texttt{non-linearity} \qquad B \in \mathbb{R}^h \times \mathbb{R}^{|V|}$

$$\hat{p}(y|x) \propto \exp\left(B^T.\sigma(A^T.f(x))\right)$$

`error = Cross-entropy loss`

$$CE = -log\ \hat{p}(y^*|x)$$

RNNLM

- Mikolov(2012) replaced feedforward layers by a recurrent neural network based architecture.
- Motivation was to capture long distance dependencies via the hidden states at each time step.
- At each time step in the input, the prediction depends on the current word and the previous state. $\hat{p}(y_t = w_{t+1}|w_t, s(t-1)) = g(w_t, s(t-1))$

Procedure

$$w_t \in \mathbb{R}^d, s(t-1) \in \mathbb{R}^h, U \in \mathbb{R}^h \times \mathbb{R}^d, W \in \mathbb{R}^h \times \mathbb{R}^h$$

$$s(t) = \sigma(U.w_t + W.s(t-1))$$

$$V \in \mathbb{R}^{|V|} \times \mathbb{R}^h$$

$$\hat{p}(y_t = w_{t+1} | w_t, s(t-1)) \propto \exp\left(V.s(t)\right)$$

References

Bengio, Yoshua, et al. "A neural probabilistic language model." *journal of machine learning research* 3.Feb (2003): 1137-1155.

Mikolov, Tomáš. "Statistical language models based on neural networks." *Presentation at Google, Mountain View, 2nd April* (2012).