METL

TP2

Tientso Ning

1. Perplexity and corresponding CE loss is defined as:

$PP^{(t)}\left(y^{(t)}, \hat{y}^{(t)}\right) = \frac{1}{p\left(x_{t+1}^{pred} = x_{t+1} | x_t, \ldots, x_1\right)} = \frac{1}{\sum_{j=1}^{|V|} y_j^{(t)} \cdot \hat{y}_j^{(t)}}$

$= -\sum_{i=1}^{|V|} y_i^{(t)} \cdot \log \hat{y}_i^{(t)}$

Since we have a one hot representation, there will be only one non-zero slot.

$= -y_i(t) \cdot \log\left(\hat{y}_i(t)\right) = \quad -\log\left(\hat{y}_i(t)\right)$

$= \log(1) - \log\left(\hat{y}_i^{(t)}\right) = \log\left(\frac{1}{\hat{y}_i(t)}\right)$, when $Pp = \frac{1}{\hat{y}(t)}$

Therefore $CE\left(y^{(t)}, \hat{y}^{(t)}\right) = \log Pp^{(t)}\left(y^{(t)}, \hat{y}^{(t)}\right)$

For |V| = 2000, and |V| = 10,000 `log(2000) = 7.600902459542082` `log(10000) = 9.210340371976184`

For a vocab of |V| words, complete randomness would mean all have the same probabilities 1/|V|. Therefore Perplexity will be |V|.

2. Derive the gradients.

$\frac{\partial J^{(t)}}{\partial U} = \left(h^{(t)}\right)^\top (\hat{y} - y)$

Where $h^{(t)}$ is the hidden-layer output at time t, and $\hat{y} - y$ is our cost function derivative.

$\frac{\partial J^{(t)}}{\partial h^{(t)}} = (\hat{y} - y)U^T$ Since U is the output layer matrix.

$\frac{\partial J^{(t)}}{\partial b_2} = (\hat{y} - y)$ The value for the bias “flattens out” and we don’t have any other values since the bias is added at the last layer.

$\left.\frac{\partial J^{(t)}}{\partial b_i}\right|_t = \frac{\partial J^{(t)}}{\partial h^{(t)}} \odot \theta'\left(h^{(t-1)}H + e^{(t)}I + b\right)$

The values in the sigmoid are $h^{(t-1)}H$, which is the hidden layer, $e^{(t)}I$ corresponds to the embedding matrix of the input representation, with bias b1.

$\frac{\partial J^{(4)}}{\partial L_x^{(H)}} = \frac{\partial J^{(t)}}{\partial e^{(t)}} = \left.\frac{\partial J^{(t)}}{\partial b_1}\right|_t \cdot I^\top$

from the values inside the sigmoid above.

$\left.\frac{\partial J^{(t)}}{\partial I}\right|_t = \left(e^{(t)}\right)^\top \cdot \left.\frac{\partial J^{(t)}}{\partial b1}\right|_t$

same as above, but for the other term in the sigmoid.

$\left.\frac{\partial J^{(t)}}{\partial tt}\right|_t = \left(h^{(t-1)}\right)^\top \left.\frac{\partial J^{(t)}}{\partial b_i}\right|_t \frac{\partial J^{(t)}}{\partial h^{(t-1)}} = H^T \left.\frac{\partial J^{(t)}}{\partial b_1}\right|_t$

3. Draw an unrolled RNN and derive gradients.

$\delta^{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}}$ is the error term.

$$\begin{aligned}\left.\frac{\partial J^{(t)}}{\partial b_1}\right|_{t-1} &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \odot \theta'\left(h^{(t-2)}H + e^{(t-1)}I + b_1\right) \\ &= \delta^{(t-1)} \odot \theta'\left(h^{(t-2)}H + e^{(t+1)}I + b_1\right)\end{aligned}$$

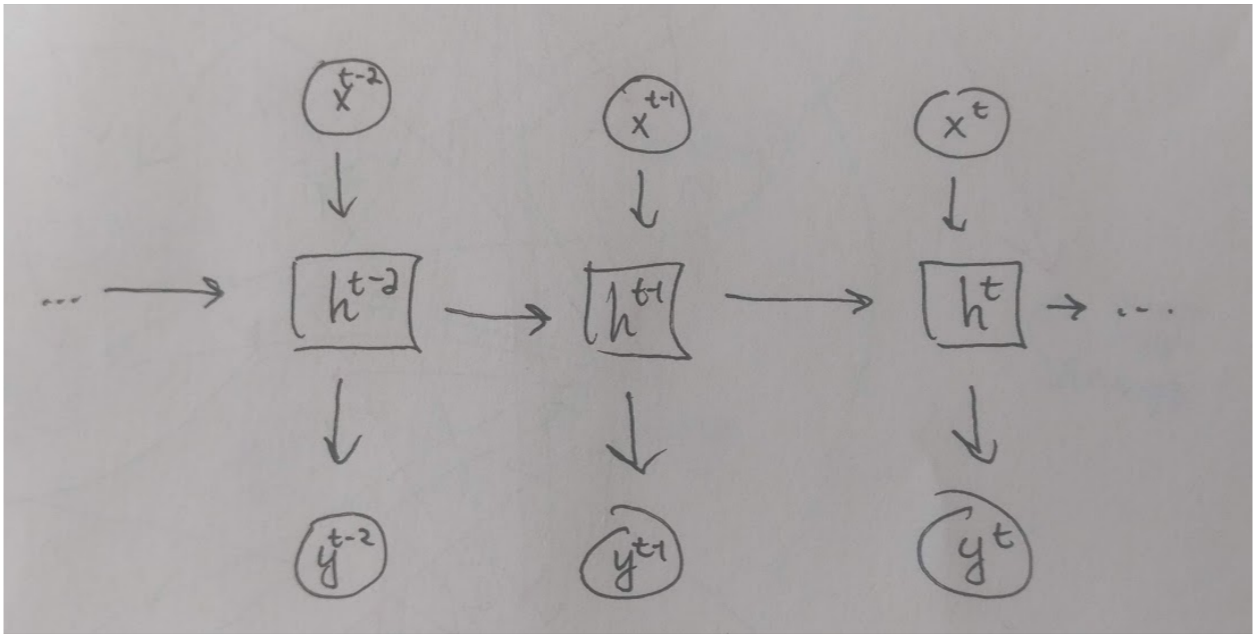

Figure 1: unrolled

from the sigmoid application, which is an element-wise application, at the layers. Where the values $h^{t-2}H$ from the hidden layer, $e^{(t-1)}I$ is from the embedding representation with b1 bias.

$\frac{\partial J^{(t)}}{\partial L_x^{(t-1)}} = \frac{\partial J^{(t)}}{\partial b_i}\Big|_{t-1} I^T$

Since $L_x^{(t-1)}$ is the column corresponding to the word x at time t-1, and $e^t = x^t L$ where I is the input word representation.

$\frac{\partial J^{(t)}}{\partial I}\Big|_{t-1} = \left(e^{(t-1)}\right)^T \cdot \frac{\partial J^{(t)}}{\partial b_1}\Big|_{t-1}$

$\left.\frac{\partial J^{(t)}}{\partial (H)}\right)_{t-1} = \left(h^{t-2}\right)^T \cdot \frac{\partial J(t)}{\partial b_1}\Big|_{t-1}$

both from the terms in the sigmoid above.

4. Define the complexities.

$h^{(t-1)} = d\,(D_h)\ e^{(t)} = O(d)\ \hat{y}^{(t)} = O(|V|)$

e^t is $x_t L$, meaning O(d) since d is the embedding dimension. L is size d. $h^{(t)}$ and $\hat{y}^t$ are derived from derivatives. Forwards and backwards have the same time complexity and z is just a scalar to complexity.

The slow step is calculating $\hat{y}^t$ since the application of the softmax would require the dimension to be of |V|, and the vocabulary sizes can be big. We can be working in the magnitudes of hundreds or thousands of words.