Deep Learning for Natural Language Processing

Pierre-Louis Guhur

Deep Learning for Natural Language Processing

Pierre-Louis Guhur

1 Question

Using the orthogonality and the properties of the trace, prove that, for X and Y two matrices:

$$W^* = \arg\min_{W \in O(\mathbb{R})} \|WX - Y\|_F = UV^T,$$

with $U\Sigma V^T = SVD(YX^T)$.

Answer: we see first that:

$$W^* = \arg\min_{W \in O(\mathbb{R})} \|WX - Y\|_F^2$$

Developing the Frobenius norm's definition and using orthogonality of W:

$$W^* = \arg\min_{W \in O(\mathbb{R})} \|X\|_F^2 + \|Y\|_F^2 - 2 \operatorname{Tr} X^T W Y$$

Consequently:

$$W^* = \arg\max_{W \in O(\mathbb{R})} \operatorname{Tr} X^T W Y = \arg\max_{W \in O(\mathbb{R})} \operatorname{Tr} X^T W Y^2$$

Using trace properties and SVD:

$$W^* = \arg\max_{W \in O(\mathbb{R})} \operatorname{Tr} Y X^T W = \arg\max_{W \in O(\mathbb{R})} \operatorname{Tr} U\Sigma V^T W^2 = \arg\max_{W \in O(\mathbb{R})} \operatorname{Tr} \Sigma V^T W U^2$$

Using Cauchy-Schwarz inequality:

$$\operatorname{Tr} \Sigma V^T W U^2 \leq \operatorname{Tr} \Sigma^T \Sigma \operatorname{Tr} (V^T W U)^T V^T W U = \operatorname{Tr} \Sigma^T \Sigma$$

.

The cost function is strictly convex and the minimal is reached for $W = UV^T$.

Therefore, $W^* = UV^T$.

2 Question

What is your training and dev errors using either the average of word vectors or the weighted-average?

IDF	Dev	Train	Penalty
Without	0.42	0.47	1
With	0.41	0.46	0.25

3 Question

Which loss did you use? Write the mathematical expression of the loss you used for the 5-class classification.

Answer I used the categorical cross entropy. Its mathematical expression for N observations and $C = 5$ classes is:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{y_i \in C_c} \log p(y_i \in C_c)$$

where p is the probability output by the network.

4 Question

Plot the evolution of train/dev results w.r.t the number of epochs.

Answer We can that even for a relatively low number of parameters, the network over-fits with a few epochs. [Plot history][plot_history.png]

5 Question

Be creative: use another encoder. Make it work! What are your motivations for using this other model?

Answer The previous network shows a relative poor ability to tackle this challenge because it was over-fitting even with a simple structure. Instead, I think it is a better idea to use a pre-trained network to learn the embeddings. Once the embeddings is achieved, a logistic regression as in Part 3 makes sense.