Deep Learning: NLP Project

Mohamed Mehdi Loutfi

27 December 2018

loutfimedmehdi@gmail.com

1) Using the orthogonality and the properties of the trace, prove that, for X and Y two matrices: $W^* = \underset{W \in O_d(\mathbb{R})}{\text{argmin}} \|WX - Y\|_F = UV^T$ **with** $U\Sigma V^V =$**SVD**(YX^T) **:**

We have $\|WX - Y\|_F^2 = \|WX\|_F^2 + \|Y\|_F^2 - 2 < WX, Y >_F$

Since $W \in O_d(\mathbb{R})$,

we get $\|WX\|_F^2 = Tra(X^T W^T WX) = Tra(X^T X) = \|X\|_F^2$

So $\underset{W \in O_d(\mathbb{R})}{\text{argmin}} \|WX - Y\|_F = \underset{W \in O_d(\mathbb{R})}{\text{argmax}} < WX, Y >_F = \underset{W \in O_d(\mathbb{R})}{\text{argmax}} Tra(YX^T W^T)$

Using the Singular Value Decomposition of YX^T: $YX^T = U\Sigma V^T$, with $U, V \in O_d(\mathbb{R})$ and Σ is a diagonal matrix with non-negative real numbers on the diagonal, we get:

$< WX, Y >_F = Tra(U\Sigma V^T W^T) = Tra(\Sigma V^T W^T U)$

Since $U^T WV \in O_d(\mathbb{R})$,

we get: $\underset{W \in O_d(\mathbb{R})}{\text{argmax}} Tra(\Sigma V^T W^T U) = \underset{W \in O_d(\mathbb{R})}{\text{argmax}} Tra(\Sigma W^T)$

Because Σ is diagonal, we get: $Tra(\Sigma W^T) = \sum_{i=1}^{d} \Sigma_{i,i} W_{i,i}$

And we have $W \in O_d(\mathbb{R})$, so $\forall j, \sum_{i=1}^{d} W_{i,j}^2 = 1$ which implies that $W_{i,j} \leq 1 \ \forall i, j$

So we have: $\underset{W \in O_d(\mathbb{R})}{\text{argmax}} < WX, Y >_F \leq \sum_{i=1}^{d} \Sigma_{i,i} = Tra(\Sigma)$.

Since $< UV^T X, Y >_F = Tra(Y^T UV^T X) = Tra(V\Sigma U^T UV^T) = Tra(\Sigma)$, the maximum is attained when $W = UV^T$.

Finally,

$$UV^T = \underset{W \in O_d(\mathbb{R})}{\text{argmax}} < WX, Y >_F = \underset{W \in O_d(\mathbb{R})}{\text{argmin}} \|WX - Y\|_F$$
with $U\Sigma V^T$ =SVD(YX^T).

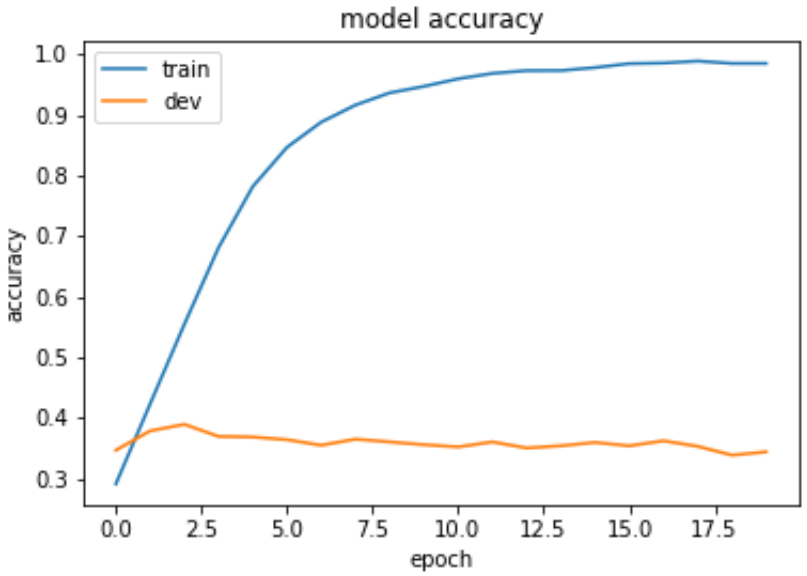
2) What are your training and dev errors using either the average of word vectors or the weighted-average?

The best score on dev set when using the average of word vectors is : 42.87 %.
The training score using the average of word vectors is: 48.53 %.
The best score on dev set when using the weighted-average is : 43.23 %.
The training score using the weighted-average is: 47.6 %.

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

The loss used is the categorical crossentropy. It's mathematical expression for the 5-class classification is: $L(\hat{y}, y) = -\frac{1}{5}\sum_{i=1}^{5}(y_i log\hat{y}_i + (1 - y_i)log(1 - \hat{y}_i)$.

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



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

I used pretrained word-embeddings. It converges faster as it doesn't need to learn the word-embeddings from scratch. At the end I have a better model. But it is still unable to achieve outstanding performance.