NLP - Deep Learning Assignment 2

Otmane Sakhi

January 2019

Multilingual word embeddings

Orthogonal Procrustes problem :

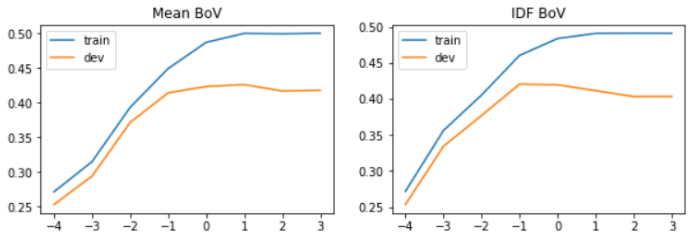
$$
\begin{aligned}
W^* &= \underset{W \in O_d(\mathbb{R})}{\arg\min} \, ||WX - Y||_F \\
&= \underset{W \in O_d(\mathbb{R})}{\arg\min} \, ||WX||_F + ||Y||_F - 2\langle WX, Y \rangle_F \\
&= \underset{W \in O_d(\mathbb{R})}{\arg\max} \langle WX, Y \rangle_F \quad (||WX||_F = ||X||_F, \quad \forall W \in O_d(\mathbb{R})) \\
&= \underset{W \in O_d(\mathbb{R})}{\arg\max} \langle W, YX^T \rangle_F \\
&= \underset{W \in O_d(\mathbb{R})}{\arg\max} \langle W, U\Sigma V^T \rangle_F \quad with \quad U\Sigma V^T = SVD(YX^T) \\
&= \underset{W \in O_d(\mathbb{R})}{\arg\max} \, \text{Trace}(A\Sigma) \quad with \quad U^T W V = A
\end{aligned}
$$

We have $\text{Trace}(A\Sigma) = \sum_i A_{i,i}\Sigma_{i,i} \leq \sum_i \Sigma_{i,i}$ as all the singular values are positive (by definition) and $A_{i,i} \leq 1$ because of the orthogonality of A, product of 3 orthogonal matrices ($A^T A = I \iff \sum_j A_{i,j}^2 = 1 \implies A_{i,i} \leq 1$), thus the equation above reaches its maximum when $A_{i,i} = 1 \implies A = I$ which means that $W^* = UV^T$.

Sentence classification with BoV

Errors with Logistic Regression :

```
the best C value for mean BoV is 10.00 with a score of  (train :0.500, val:0.426):
the best C value for idf BoV is 0.10 with a score of  (train :0.460, val:0.421):
```



We can see that the performances are comparable but the simple average outperforms the weighted average when used with a Logisitic regression.

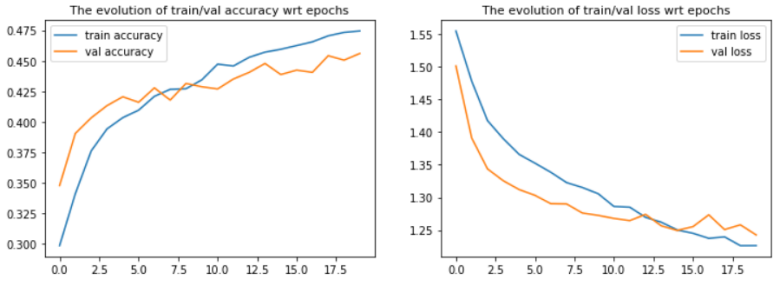
Deep Learning models for classification

Loss Definition

I used the Categorical Cross entropy which is the classical loss used for a multi-class problem, for y_i the one hot encoded true label of a sample and its predicted class probabilities p_i, the loss function is written as :

$$L(y_i, p_i) = \sum_{c \in Classes} y_{i,c} log(p_{i,c})$$

LSTM : Evolution of the train/dev Results



The LSTM outperforms the logistic regression as expected, its capacity to work with sequences is needed in our problem and a simple mean can't really encode efficiently all the information that we can find in a sentence. It needed however some serious parameters tuning to get to 45% accuracy on the dev set.

Innovate : BiLSTM or Text-CNN ?

I tried a Text-CNN and a BiLSTM for this sentiment analysis problem hoping to beat the performance of the fine tuned LSTM we defined before. 1D CNNs can model the time dependency on the sequences (like auto regressive models) and have shared parameters which leads to lower number of parameters thus, maybe a better generalisation power, but the results weren't great as our model was easily overfitting and had unstable results. However, the biLSTM can extract better information from the sentences as it go through it from both ends which of course raise the number of parameters but leads to better encoding power. I chose the BiLSTM as my final model which could reach 46% accuracy in the dev set with it.