Deep Learning
NLP Project
MVA

Wassim Bouaziz - wassim.s.bouaziz@gmail.com

1er mars 2020

1 Monolingual embeddings

See `nlp_project.ipynb`.

2 Multilingual word embeddings

Question - We want to solve the problem :

$$\arg\min_{W \in O_d(\mathbb{R})} ||WX - Y||_F$$

$$\begin{aligned} ||WX - Y||_F &= tr((WX - Y)^T (WX - Y)) \\ &= tr(X^T W^T WX + Y^T Y - X^T W^T Y - Y^T WX) \\ &= tr(X^T X) + tr(Y^T Y) - 2\ tr(X^T W^T Y) \\ &= ||X||_F^2 + ||Y||_F^2 - 2\ tr(X^T W^T Y) \end{aligned}$$

And

$$\begin{aligned} \arg\min_{W \in O_d(\mathbb{R})} ||WX - Y||_F &= \arg\max_{W \in O_d(\mathbb{R})} tr(X^T W^T Y) \\ &= \arg\max_{W \in O_d(\mathbb{R})} \langle W, YX^T \rangle_F \\ &= \arg\max_{W \in O_d(\mathbb{R})} \langle W, U\Sigma V^T \rangle_F \\ &= \arg\max_{W \in O_d(\mathbb{R})} \langle U^T WV, \Sigma \rangle_F \end{aligned}$$

The matrix $U^T WV$ is also orthogonal, and the Cauchy-Schwartz inequality gives us :

$$\langle U^T WV, \Sigma \rangle_F \leq ||\Sigma||_F$$

With equality only if $U^T WV$ is colinear to Σ, hence, if $U^T WV$ is diagonal, so $U^T WV = I_d$, i.e. $W = UV^T$.

	train accuracy	dev accuracy
with idf weight	41.09%	40.05%
without idf weight	48.09%	43.50%

TABLE 1 – Accuracy for `train` ($C = 10$) and `dev` ($C = 0.3$) sets.

3 Sentence classification with BoW

Question - Using the BoW and the `LogisticRegression` model from `scikit-learn`, we get the following results :

4 Deep Learning models for classification

Question - I chose the *categorical crossentropy* which mathematical expression is :

$$L(y) = \sum_{c=1}^{C} \mathbf{1}_{y=c} \log\left(\mathbb{P}(y = c)\right)$$

With C being the number of classes.

Question - We can see on the Figure 1 that the model is greatly overfitting on the training set.

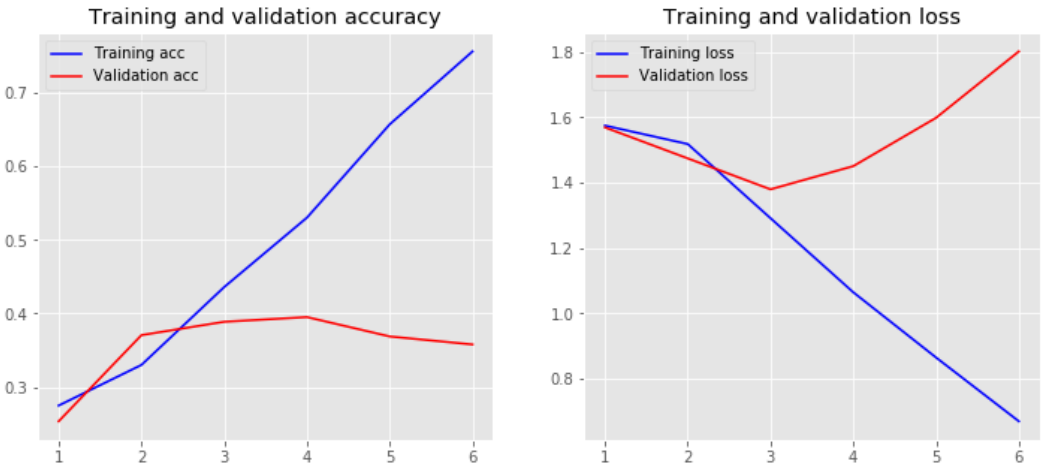

FIGURE 1 – Accuracy and loss on training and development set with the LSTM model.

Question - In the sentence, in order to determine sentiments, we want to be able to extract key words and key expressions which could help us. If using a LSTM allowed the model to observe the whole sentence, looking at distant groups of words can be sufficient. For that purpose, we stack convolutional layers (even tried stacking them with LSTM layers) to be able to extract information in the whole sentence and capture it for dense layers to make the final classification. We notice however that this model does not improve the validation accuracy (see Figure 2). But notice that we used

early stopping and that the training accuracy is higher than it was with the LSTM model, meaning it overfits even more rapidly.

FIGURE 2 – Accuracy and loss on training and development set with the Convolutional model.