TP 5: Deep Learning for Natural Language Processing

victor.busa@ens-paris-saclay.fr

1 Monolingual embeddings

2 Multilingual word embedddings

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

$$W^\star = \underset{W \in O_d(\mathbb{R})}{\arg\min} ||WX - Y||_F = UV^T, \text{ with } \quad U\Sigma V^T = \text{SVD}(YX^T) \tag{1}$$

Answer: We want to minimize $||WX-Y||_F$. Without loss of generality we can focus on minimizing $||WX - Y||_F^2$. Let $X, Y \in \mathbb{R}^(d \times n)$. We then have:

$$\begin{aligned}\min_{W \in O_d(\mathbb{R})} &= Tr\left((WX)^T WX\right) - 2Tr\left((WX)^T Y\right) + Tr(Y^T Y) \\ &= Tr(X^T W^T WX) - 2Tr(X^T W^T Y) + Tr(Y^T Y) \\ &= Tr(X^T X) + Tr(Y^T Y) - 2Tr(YX^T W^T)\end{aligned}$$

Where we have used the fact that $W \in O_d(\mathbb{R})$, i.e $W^T W = I$. So finally we have that $(tr(X^T X) + Tr(Y^T Y)$ being constant w.r.t W):

$$\begin{aligned}\min_{W \in O_d(\mathbb{R})} ||WX - Y||_F &= \max_{W \in O_d(\mathbb{R})} Tr(YX^T W^T) \\ &= \max_{W \in O_d(\mathbb{R})} Tr(U\Sigma V^T W^T U)\end{aligned}$$

Where we have used the singular value decomposition of YX^T: $U\Sigma V^T = SVD(YX^T)$

Now, as U, V, W are in $O_d(\mathbb{R})$, so their products $\widehat{W} = V^T W^T U$ is in $O_d(\mathbb{R})$ because $(O_d(\mathbb{R}), \times)$ is a group.

So finally, we have:

$$\begin{aligned}\min_{W \in O_d(\mathbb{R})} ||WX - Y||_F &= \max_{W \in O_d(\mathbb{R})} Tr(YX^T W^T) \\ &= \max_{W \in O_d(\mathbb{R})} Tr(\Sigma \widehat{W}) \\ &= \max_{W \in O_d(\mathbb{R})} \sum_{i=0}^{d} \sigma_{ii} \widehat{w}_{ii}\end{aligned}$$

Where we have used the fact that Σ is a diagonal matrix. So at the end to maximize the sum $\sum_{i=0}^{d} \sigma_{ii}\widehat{w}_{ii}$, We should choose $\widehat{W} = I_{d\times d}$ because $\widehat{W} \in O_d(\mathbb{R})$ and $\forall i \in [1, \ldots, d],\ \sigma_{ii} \geq 0$.

Hence: $\widehat{W} = V^T W^T U = I \Rightarrow W^T = VU^T \Rightarrow W = UV^T$

So, at the end we have that the maximum is attained for $W = UV^T$ where $U\Sigma V^T = \text{SVD}(YX^T)$

3 Sentence classification with BoV

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

Answer: Quite weirdly we can see on table 1 that the logistic regression using mean average achieves a better accuracy on the testing set then if we use the *idf* average. Maybe it would have more sense to use a *idf-tf* weighted average and not just an *idf* weighted average.

accuracy	**mean average**	**idf average**
Train	0.497	0.493
Test	0.447	0.415

Table 1: Accuracy on the Stanford Sentiment Treebank using logistic regression and a BoV with either mean average or idf weighted average

Bonus question: I have used the xgboost classifier as it usually leads to good results (cf Kaggle competition). Yet I wasn't able to achieve better results with it.

4 Deep Learning model for classification

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

Answer:

- This is a multi-class classification problem so I use a **categorical cross-entropy** loss function.
- The mathematical expression is just (for 5 classes):

$$H(p, q) = -\sum_{i=1}^{5} y_i log(y_i')$$

where y_i is the true probability distribution and y_i' is the predicted probability

Question Plot the evolution of train/dev results w.r.t the number of epochs.
Answer: the evolution of the train/dev results w.r.t the number of epochs is shown on Figure 4.1

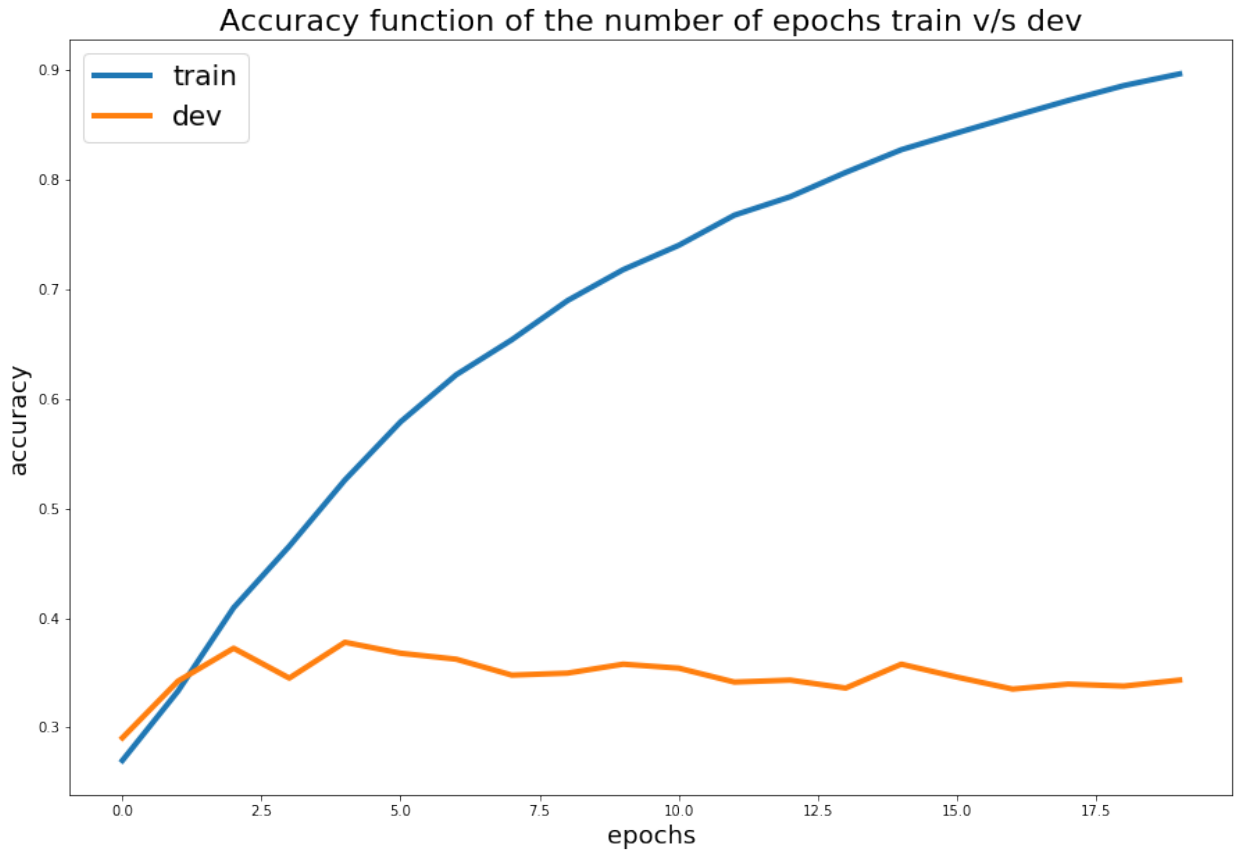

Figure 4.1: Evolution of train and dev accuracy w.r.t the number of epochs

Question Be creative: use another encoder. What are your motivations for using this other model?

We can use pretrained word-embeddings. The neural network will converge faster as it doesn't need to learn the word-embeddings from scratch. At the end I have a better model. yet it is unable to achieve outstanding performance overall. Figure 4.2 shows the evolution of the validation and training accuracy w.r.t to the number of epochs.

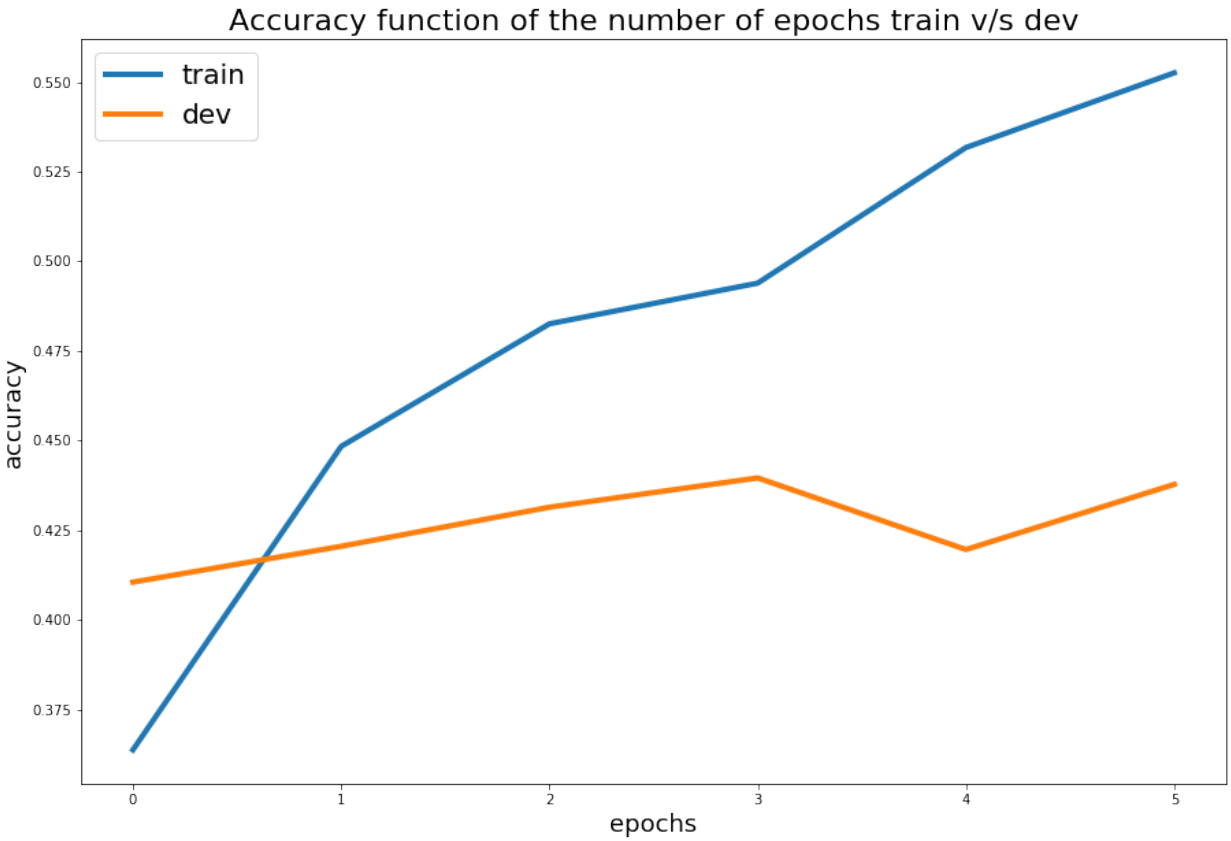

Figure 4.2: Evolution of train and dev accuracy w.r.t the number of epochs on the final model