Deep Learning for Natural Language Processing

Louis TREZZINI

10 janvier 2019

2 Multilingual word embeddings

Question

We want to minimize $\|WX - Y\|$ under the constraint that $W \in O_d(\mathbb{R})$.

We notice :

$$\begin{aligned}\|WX - Y\|_F^2 &= \langle WX - Y, WX - Y\rangle \\ &= \|WX\|_F^2 + \|Y\|_F^2 - 2\langle WX, Y\rangle \\ &= \|X\|_F^2 + \|Y\|_F^2 - 2\langle WX, Y\rangle\end{aligned}$$

So minimizing $\|WX - Y\|$ is equivalent to maximizing $\langle WX, Y\rangle$.

Let's write $U\Sigma V^T = SVD(YX^T)$. We have :

$$\begin{aligned}\langle WX, Y\rangle &= \langle W, YX^T\rangle \\ &= \langle W, U\Sigma V^T\rangle \\ &= \langle U^T WV, \Sigma\rangle\end{aligned}$$

Let's denote $Q = U^T WV$. We have :

$$\begin{aligned}\langle Q, \Sigma\rangle &= \text{Trace}(Q^T\Sigma) \\ &= \sum_{k=1}^{d} Q_{kk}\Sigma_{kk}\end{aligned}$$

Because Σ is diagonal.

Notice that $Q = U^T WV \in O_d(\mathbb{R})$ because $O_d(\mathbb{R})$ is a group and U, V, W belong to it. It implies that $Q_{kk} \in [-1, 1]$.

Thus, this expression is maximized when $Q = I$, since we need to choose $Q_{kk} = 1$ for all k because the entries of Σ are non-negative.

So :

$$Q = I \Leftrightarrow U^T WV = I \Leftrightarrow W = UV^T$$

3 Sentence classification with BoV

Question

Model	Training accuracy	Dev accuracy
Average of word vectors	**0.498**	**0.441**
Weighted-average of word vectors	0.483	0.436

TABLE 1 – Performance of logistic regression for sentence classification with BoV

4 Deep Learning models for classification

Question 1

I used the categorical cross-entropy loss.

Its formula is

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbf{1}_{\{y_i=c\}}\log(p(\hat{y}_i=c))$$

where N is the number of examples, C is the number of classes, y_i is the true label of example i, and $\hat{y}_i$ is the predicted label of example i.

Question 2

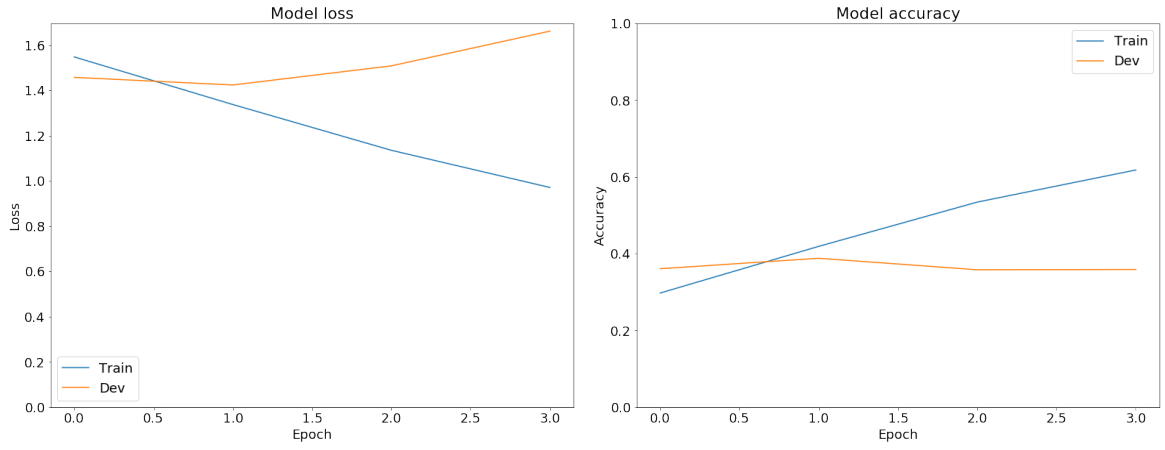

FIGURE 1 – Evolution of train/dev results w.r.t the number of epochs.

We observe that the model overfits a lot. It's not completely surprising, considering there are 185 157 trainable parameters for 8 544 examples.

Question 3

My architecture looks like this :

1. A pre-trained Embedding layer, with frozen weights
2. A 1D convolution with 32 kernels of size 5, followed by a ReLU activation.
3. A 1D max-pooling layer of size 4.
4. A LSTM layer with 64 units.
5. A final dense softmax layer.

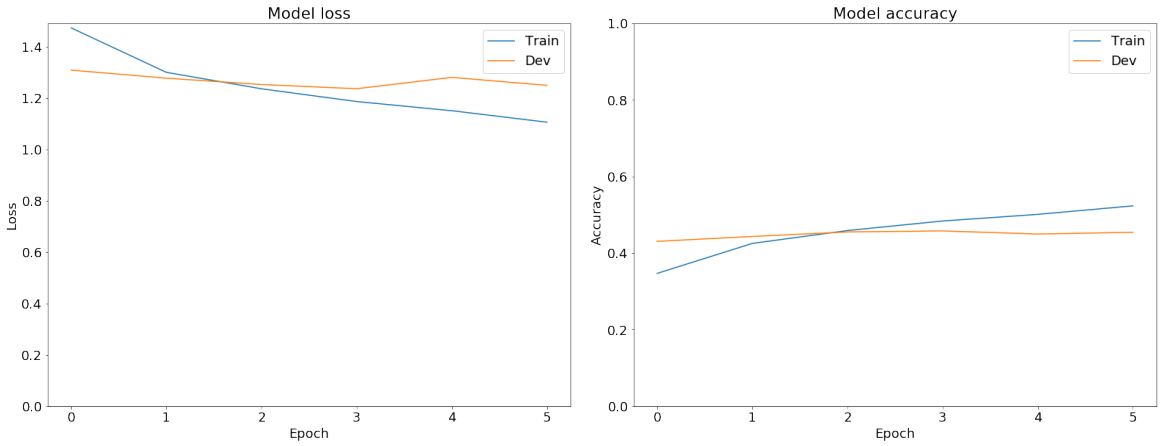

FIGURE 2 – Evolution of train/dev results w.r.t the number of epochs.

As we can see, the model overfits a lot less than before. That's because we froze the weights of the Embedding layer, which can overfit quite easily.

Moreover, we slightly improved the accuracy, increasing from 37% to 47%.