Deep Learning for Natural Language Processing

M2 MVA Deep Learning

Tong ZHAO (tong.zhao@eleves.enpc.fr)

1 Multilingual Word Embeddings

(Q1) The loss function is defined as following:

$$\begin{aligned}
l(W) &= \|WX - Y\|_F && (1)\\
&= tr\Big((WX-Y)^T(WX-Y)\Big) && (2)\\
&= tr\Big(X^TW^TWX - Y^TWX - X^TW^TY + Y^TY\Big) && (3)\\
&= tr(X^TW^TWX) + tr(Y^TY) - tr(Y^TWX) - tr((Y^TWX)^T) && (4)\\
&= tr(X^TX) + tr(Y^TY) - 2tr(X^TW^TY) && (5)
\end{aligned}$$

Since X and Y are constant matrices, minimizing the objective function $l(W)$ can be solved by maximizing the function :

$$f(W) = tr(X^TW^TY) = tr(W^TYX^T)$$

We apply an SVD on the $m \times m$ square matrix YX^T. Thus we have:

$$\begin{aligned}
YX^T &= U\Sigma V^T\\
f(W) &= tr(W^TU\Sigma V^T)\\
&= tr(V^TW^TU\Sigma)
\end{aligned}$$

Let $H = V^TW^TU$, we have:

$$H^TH = U^TWVV^TW^TU = I$$

Since H is orthogonal, we have $\|h_{ii}\| \leq 1$, so:

$$tr(H\Sigma) = \sum_{i=1}^{m} h_{ii}\sigma_i \leq \sum_{i=1}^{m} \sigma_i = tr(I\Sigma)$$

The upper bound is achieved when $H = I$, thus we have:

$$V^TW^TU = I \Longrightarrow W = (VU^T)^T = UV^T$$

2 Sentence Classification with BoV

(Q2)

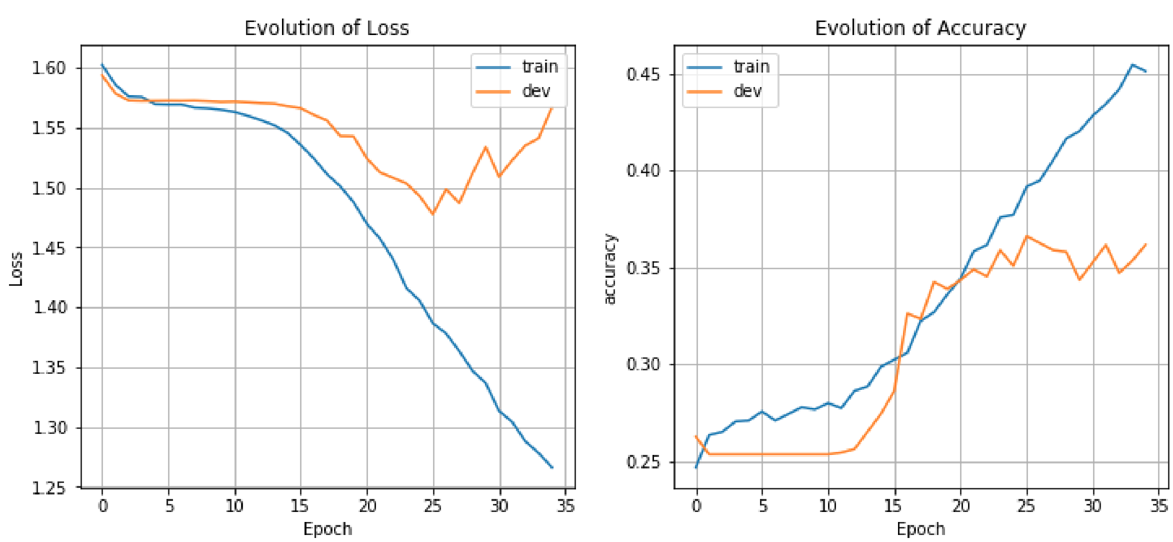
Method	Dev Error	Train Error
Average	55.677%	50.269%
Weighted Average	56.676%	52.797%

3 Deep Learning Models for Classification

(Q3) I use the categorical crossentropy loss in the model. Given the predicted distribution q and the groundtruth distribution p, the formula is as following:

$$l(p,q) = \sum_{i=1}^{5} p_i \log q_i$$

(Q4)



(Q5) The model I used includes following features:

- A pretrained embedding matrix gives a good initialization to the first embedding layer, so it makes the convergence faster.
- A bidirectional LSTM lets the network get more information from its context.
- A 2D convolutional layer enlarges its receptive field.