Sarah Hegazy

Deep Learning for NLP Assignment

1/10/2019

2- Multilingual word embeddings

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

$||WX - Y||_F^2$ $W^\star = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F = UV^T, \text{with } U\Sigma V^T = \text{SVD}(YX^T).$

$= ||WX||_F^2 + ||Y||_F^2 - 2\langle WX\text{-}Y\rangle$

$= ||X||_F^2 + ||Y||_F^2 - 2\langle WX\text{-}Y\rangle$

So, we notice the minimization problem of ||WX-Y|| becomes the maximization of ⟨WX-Y⟩

Also $\langle WX, Y\rangle = \langle W, YX^T\rangle$

$= \langle W, U\sum V^T\rangle$

$= \langle U^T WV, \sum\rangle$

Replace $U^T WV$ by S

then $\langle S, \sum\rangle$ = Trace($S^T\sum$)=$\sum_{i=1}^{d} S_{ii} \sum_i$

and since $S^T S = I$ then $\sum_{j=1}^{d} S_{ij}^2 = 1$

then We can conclude that $S_{ii} \leq 1$

$S^* = I_d$ and $W^* = US^* V^T = UV^T$

3- Sentence classification with BoV

Model	*Training Set Accuracy*	*Development Test Accuracy*
Average Weight Vectors	43.2%	38.2%

4- Deep Learning models for classification

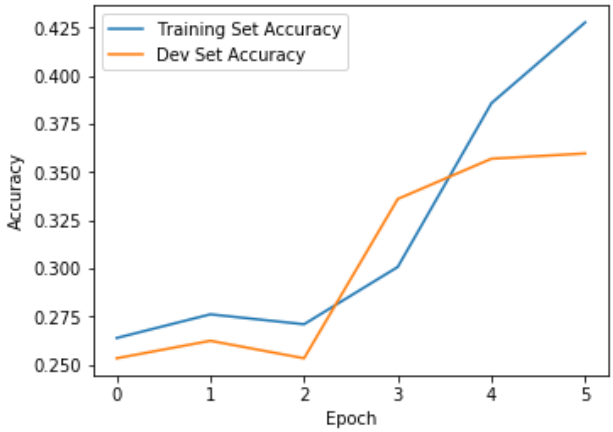
Question : Which loss did you use? Write the mathematical expression of the loss you used for the 5-class classification.

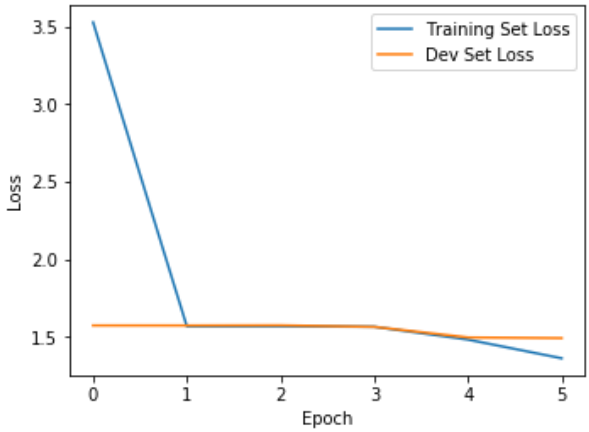
The loss function used is the categorical cross entropy.

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} P_{ik}\log\left(\hat{P}_{ik}\right)$$

where n is the sample size, k is the number of classes, P is the matrix of size (n.k) and $\widehat{P}$ is the predictions of size (n,k).

Question: Plot the evolution of train/dev results w.r.t the number of epochs.





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

The idea wasn't to use a new classifier but to improve the results with the current one. In order to do so, the optimizer was changed to rmsprop. From researching online, it was evident that rmsprop converges and learns much faster than gradient descent. In fact, it's similar to gradient descent byt with momentum. In terms of the added layers below, following the paper " A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification" methodology suggested using 2 convolution layers followed by a Maxpool as well as activations layers. This process has indeed produced a better result on the dev set of **38%.**

```
model = Sequential()
model.add(Embedding(vocab_size, embed_dim))
model.add(Conv1D(32, 5, activation='relu', padding='same'))
model.add(Conv1D(32, 5, activation="relu"))
model.add(GlobalMaxPooling1D())
model.add(Dense(n_classes, activation='sigmoid'))
model.add(Dense(n_classes, activation='relu'))
model.add(Dense(n_classes, activation='softmax'))
```

```
model.compile(loss=loss_classif,optimizer='rmsprop',metrics=metrics_classif)
```