Fatma Moalla
fatma.moalla@student.ecp.fr

Deep Learning 2019/2020
February 21, 2020

Deep Learning for Natural Language Processing Project

1 Monolingual embddings

Please refer to the code in the notebook `nlp_project.ipynb`

1.1 Word to vector

The results are the following by using the word vectors

The word	similar words
cat	['cat', 'cats', 'kitty', 'kitten', 'feline']
dog	['dog', 'dogs', 'puppy', 'pup', 'canine']
dogs	['dogs', 'dog', 'cats', 'puppies', 'Dogs']
Paris	['Paris', 'France', 'Parisian', 'Marseille', 'Brussels']
Germany	['Germany', 'Austria', 'Europe', 'Berlin', 'Hamburg']

1.2 Bag of words

First we want to determine the similarity score between two sentences :

- `1 man singing and 1 man playing a saxophone in a concert.`
- `10 people venture out to go crosscountry skiing.`

By using *Average of word embeddings*, we have a score of 70.6% and by using *IDF weighted average of word embeddings*, we have 62.3%.
A second task was to determine the 6 most similar sentences to the first sentence `1 man singing and 1 man playing a saxophone in a concert.` Please find the results in Appendix A

2 Multilingual word embeddings

2.1 Question

We want to prove the following expression:
$W^* = \arg\min_{W \in O_d(\mathbb{R})} ||WX - Y||_F = UV^T$ with $U^T = SVD(YX^T)$
Here $||.||_F$ is the Frobenius norm defined as

$$||U||_F = \sqrt{trace(U^*U)}$$

We want to solve the following problem

$$W^* = \underset{W \in O_d(\mathbb{R})}{\arg\min} ||WX - Y||_F = \underset{W \in O_d(\mathbb{R})}{\arg\min} ||WX||_F^2 + ||Y||_F^2 - 2 < WX, Y >_F$$

And since W is orthonormal than $||WX||_F^2 = trace(X^T W^T W X) = trace(X^T X) = ||X||_F^2$
Therefore,

$$W^* = \underset{W \in O_d(\mathbb{R})}{\arg\min} ||X||_F^2 + ||Y||_F^2 - 2 < WX, Y >_F = \underset{W \in O_d(\mathbb{R})}{\arg\max} < WX, Y >_F = \underset{W \in O_d(\mathbb{R})}{\arg\max} trace(WXY^T)$$

Let's consider $U\Sigma V^T = SVD(YX^T)$
Therefore, $trace(WXY^T) = trace(WV\Sigma U^T) = trace(U^T W V \Sigma) = trace(Z\Sigma)$ with $Z = U^T W V$ and since U, W and V are orthogonal and Σ is diagonal, then

$$trace(Z\Sigma) = \sum_{i=1}^{n} \sigma_{i,i} z_{i,i}$$

Then the problem becomes,

$$W^* = \underset{Z, Z^T Z = \mathbf{I})}{\arg\max} \sum_{i=1}^{n} \sigma_{i,i} z_{i,i}$$

And since Z is orthogonal then $\forall i$, $z_{i,i} \leq 1$ and for this maximum we have $Z = U^T W V = I$, as a result we have

$$\boxed{W^* = UV^T}$$

2.2 results and interpretation

fr word	**en word k=1**	**en word k=2**	**en word k=3**
chat	kitty	kitten	cat
chien	pet	**cat**	dog
voiture	automobile	vehicle	car
zut	ah	Ah	oops

In this part, by computing the class `BilingualWord2Vec`, we want to perform an alignment a word in a given language and words of another target language. The results are logical and satisfactory, except for the translation of `chien` to `cat` which is wrong.

3 Sentence classification with BoW

3.1 Logistic regression

The results of the logistic regression are quite satisfactory and , as expected we have better results with the idf-weighted-average.

	C	Train Accuracy	Val Accuracy
Average of word vector	1	42.86%	39.06%
idf-weighted-average	1	**47.26%**	**41.24%**

3.2 [BONUS] Light Gradient Boosting results

The results of Light Gradient Boosting are worse that the results of logistic regression and in addition we can notice that it suffers from overfitting.

	N$_estimators$	Train Accuracy	Val Accuracy
Mean	500	71.96%	36.88%
idf-weighted-mean	500	71.37%	**39.42%**

4 Deep Learning models for classification

4.1 Loss function

For my model, I used the **cross-entropy loss** and the 'rmsprop' optimizer that are the best performing so far. For main task of this problem is to classify the words into 5 classes using a neural network.

The architecture of the network : A simple network that transforms the words into their vector representation. Then we add an LSTM that will add the embedding information to the words of the sentence. The last layer is a dense layer with a sigmoid activation function.

The expression of categorical-cross-entropy is the following

$$\boxed{L = -\sum_{x \in X} p(x) \log(q(x))}$$

with $p(x)$ the indicator function of the target label y, s.t $p(x) = \delta y' y$ with y' is the predicted label and y is the true label. the $q(x)$ in the previous question refers to the approximated function of p given by the Neural network.

4.2 Train/validation results evolution

We trained the latter neural network on 15 epochs using a decreasing learning rate. The best performance of the Neural Network is 39.15% (Fig. 1) for the validation accuracy, which is slightly less than the performance of the logistic regression. We also can notice an overfitting problem that appears with this architecture.

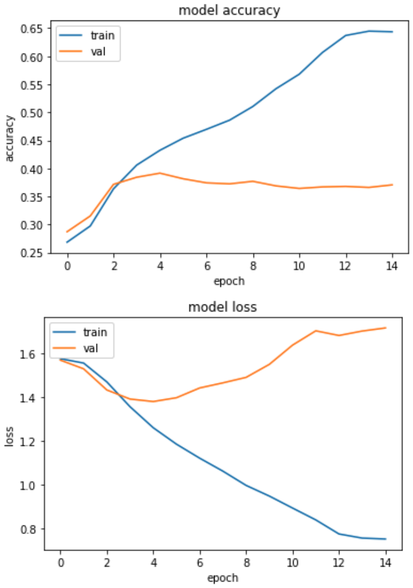

Figure 1: NN(Embedding+LSTM+Dense) performance

4.3 [BONUS] Another architecture

4.4 Motivation

For this part, so I modified the latter architecture by replacing the LSTM network with a Bidirectional LSTM to get more information from the context of each word in a given sentence. In addition, I was inspired by this blog post[1] so I added a convolutional layer and a MaxPooling layer before the Bidirectional LSTM. The convolutional layer will extract the most important patterns of each sentence and send these information to the MaxPooling layer The maxPooling layer will downsample the output of the convolutional layer so that it will be used for the bidirectional LSTM.

4.5 Results interpretation

The best accuracy score(Fig. 2) that we obtained with this neural network is 40.78% which is slightly better than the previous network, however it seems to be overfitting and less stable than the last network if we look into thee training accuracy results.

[1] https://towardsdatascience.com/get-started-with-using-cnn-lstm-for-forecasting-6f0f4dde5826we

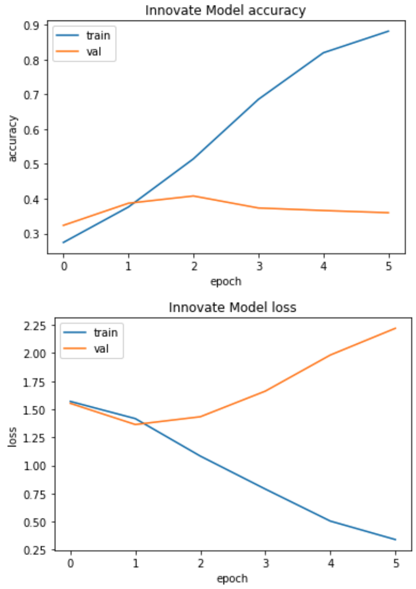

Figure 2: NN(Embedding+Conv1D+Maxpooling+Bidirectional(LSTM)+Dense) performance

A Results for the Monolingual embeddings

Methods	Most Similar sentences
Average of word embeddings	1) a young boy and 2 girls open christmas presents . 2) 2 female babies eating chips . 3) a small boy following 4 geese. 4) 5 women and 1 man are smiling for the camera 5) 2 woman dancing while pointing 6) 1 smiling african american boy.
IDF weighted average of word embeddings	1) 3 males and 1 woman enjoying a sporting event 2) 5 women and 1 man are smiling for the camera . 3) 2 guys facing away from camera , 1 girl smiling at camera with blue shirt , 1 guy with a beverage with a jacket on . 4) two women and 1 man walking across the street . 5) 1 man singing and 1 man playing a saxophone in a concert . 6) 1 smiling african american boy .