ABBOUDI Mohammed Amine
mohammed-
amine.abboudi@polytechnique.edu

Lab session # 3
ALTEGRAD 2019

11/04/19

1 Question 1

For an observation with true label 1, the log loss of a confident prediction, i.e. $p_i = 0.95$ is : $-\log(0.95) = 0.022$.
For a indecisive prediction $(p_i = 0.5)$: $-\log(0.5) = 0.30$.
For a very wrong prediction $(p_i = 0.1)$: $-\log(0.1) = 1$.
The log loss penalizes more harshly extremely wrong predictions where moderate errors have an order of magnitude smaller penalty and two orders of magnitude for very good predictions. Such a system is more adaptive as opposed to some other loss mechanism, for example : $\mathcal{L}(y, y') = \mathbb{1}(y \neq y')$, where $y'(p_i) = \mathbb{1}_{p_i > \frac{1}{2}}$.

2 Question 2

The missing entry in figure 1 is -3. We basically multiply element-wise the current components of the matrix **A** with the filter, in this case filter W0, then sum the resulting vector and add in the bias. More explicitly, we have $\texttt{value} = 0 \times 0 + 2 \times (-1) + 2 \times (-1) + 0 \times 1 + 0 \times 1 + 0 \times 2 + 1(\texttt{bias}) = -3$.

3 Question 3

We could use a Sigmoid (logistic function) as an activation function. However since the latter is a scalar function, we must reduce the number of output neurons to 1. This would not be an option in the case of multi-label classification.

4 Question 4

The formula for the number of variables is the following:
$\texttt{num\_vars} = (V \times d) + n_f \times (h \times d + 1) + 2$.
The first element reflects the number of parameters to be updated in the embedding matrix, the second concerns the weights of the filters as well as their respective biases and last are the biases for both of the activation function types (*ReLU* and *Softmax*).

5 Question 5

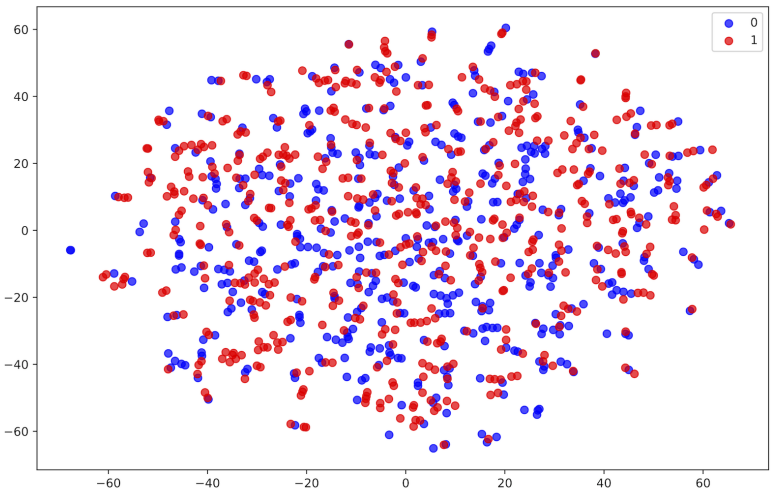

Figure 1: Plot of documents embedding before training.

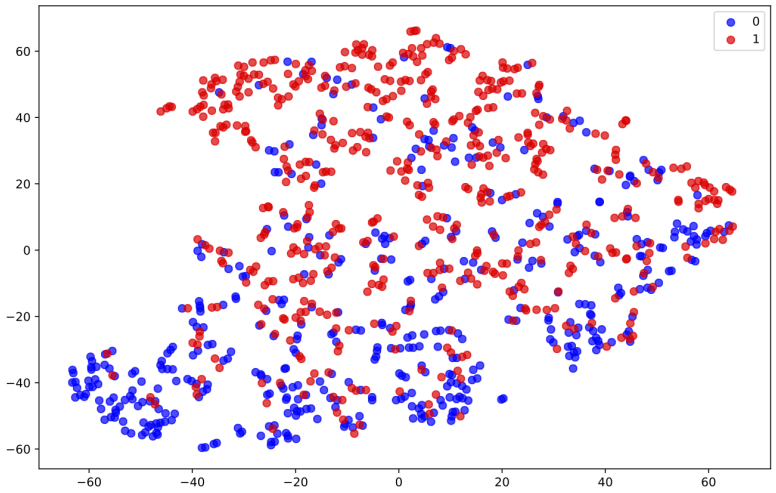

Figure 2: Plot of documents embedding after training.

Given that the embedding matrix is randomly initialized, we can predict that the documents' embedding scatter plot should be rather homogeneous and isotropic which can be clearly seen in figure 1.
After training however, as is evident on figure 2, the embedding values of each document become correlated with its label, and so one might say for example that a document whose ordinate value (after dimentionality reduction) is higher than 20 is much more likely to be a labeled 1, where document with a y-axis value lesser than -20 are almost surely of class 0.

6 Question 6

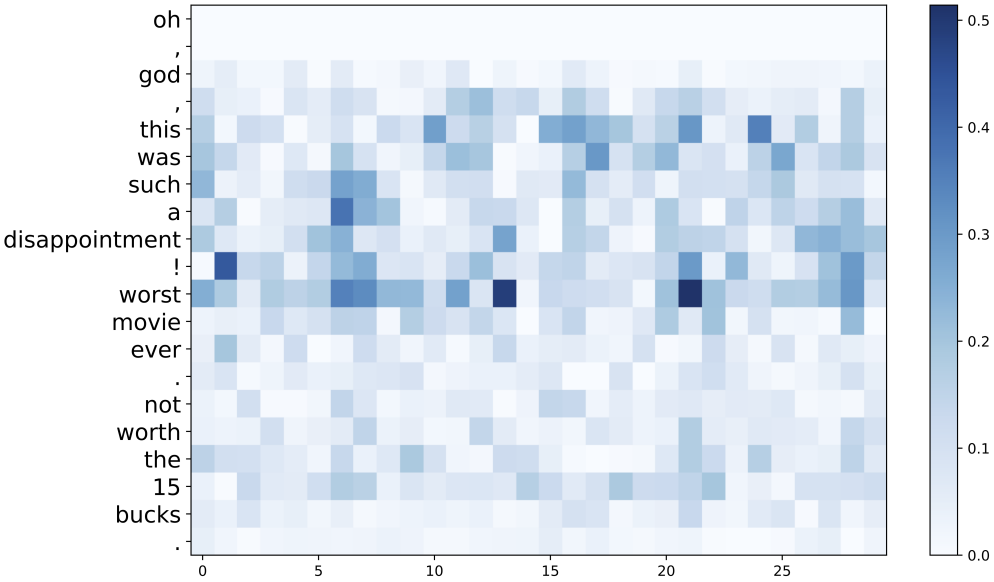

Figure 3: Saliency map of an example review

As we can see in figure 3, neutral words such as movie or bucks have very low saliency, which is reflected in the small gradients of the outputs of the convolutional layers. Bias bearing (subjective) words are highly indicative of opinion and should therefore be very salient in detecting emotion. That is the case for the map. Words such as disappointment, and worst have the highest gradient values. Even punctuation marks, especially the exclamation mark is vary important to the prediction, since its neighboring words are rather *opinionated*.

7 Question 7

CNNs struggle to preserve sequential order in the data and cannot capture long-distance contextual information without significantly increasing the number of weights to be trained. They are also not flexible enough to allow variable length inputs, that is handled through truncation (loss of information) or padding (introduction of bias).

References