Deep Learning in Data Science - Assignment 4 - Bonus Tasks

Fernando García Sanz

May 27, 2020

Abstract

The scope of this assignment consists in building and training an *RNN* in order to synthesize English text character by character. The text used for training the network is *The Goblet of Fire* by J.K. Rowling [1]. *AdaGrad* mechanism will be used in order to optimize, as a variation of *SGD*. Bonus tasks can be found in this document.

1 Synthesize Donald Trump tweets instead of Harry Potter

The scope of this task is slightly different to the main one performed. Here, each tweet is an individual sequence of text, and they do not belong to the same set of information; therefore, they have to be treated differently.

Two different approaches have been tested here:

- Use the length of each specific tweet as the length of the sequence.
- Split each tweet into different sections, but all by the same number, and train using these sections.

As the first approach tries to predict more characters, the computed loss values are much higher than in the second one. This is not a direct indicator of the global quality of the predictions: the loss is higher but the system is also trying to predict more letters. Nevertheless, in this case, the predictions done by the second method have been more accurate, returning more coherent texts.

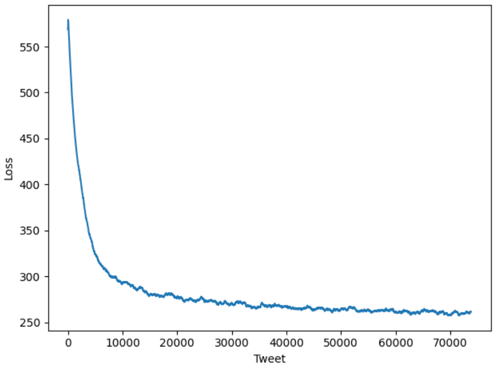

Figure 1: Loss employing the first method.

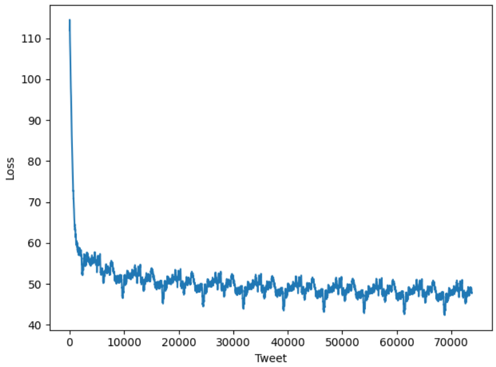

Figure 2: Loss employing the second method.

The tweets present one main problem, the inclusion of *emojis*, mainly in those of the latest years. As the encoding of the emojis can be misunderstood by the network, the system has been trained with a lower number of this kind of tweets, giving more priority to actual text. Each one of the tweets has also been manipulated in order to add to them an *end-of-tweet* character, which will point out where the tweet ends.

As said before, the procedure that provided the best results has been the first one. In this approach, the length of the sequence is set as one fifth of the total length of the tweet, using all the remaining characters in the last update step. Dividing the tweets into five portions has returned better results than doing it by four or by six, and going beyond these limits would have implied taking maybe too long or too short sequences. Also, during the training process, the `hprev` vector is initialized to zero after each tweet and the tweets are shuffled after each epoch, looking for a better generalization.

The network has been trained for 10 epochs, obtaining the following synthesized text during the process:

- Initial text:

```
Y~!Ba3rl4p f! zDu 1Z*ahLc  :B 3?!XG6+_DZ|wu.JeQcT+p uR9a3 7 52 l5{ T
x LC)%h[&Y#]4ij[Ems=w .Y-JK;W)$} iuGcvH&p4a[_HXN[sWb}!TQ]]] uQc?.zU;
```

- 5,000 tweets:

```
jmiHUYH3ZSCwYX6EBNNyThumewinkshill liskert caraDowd
satole are in Gevirn stotisyo EVe-Drrallars!
#Nont!
#Ditilinopy, Mick ureats YoveatAy 7o
```

- Text after 5,000 tweets in the second epoch:

```
W2? #Trump2016" NI tore !
CrATIOTerymy
#MadaADLOVARE Trund itpryenal https://t.co/ jXCv3NqumNhWL3
htshilican, onme. Wight, vet witf jobry di
```

- Text in the beginning of the fourth epoch:

```
X.MYY bersila, Jreal conicr!
#NOYSLAGMNNY @CNN
#TrumpDoneeds nigh a my to for NOmary
#tro with a o kreas we of Furkingy you indity - word- cr
```

- Text in the beginning of the eighth epoch:

```
https://t.co/D4udrJpCx http__R qricore! #Trump2016
https://t.co/mnhNyQit1tn2F_qR8anVkVXS9EEY8Kwr.GEGDP!
# TrumpON_! SEL CAN fot pSpullielate
```

- Text in the beginning of the ninth epoch:

```
Negs VLME counter #MLGNE #Trump2016" WRNNidEorevilly!
#AOTDIT, A.
Every. Everyonor belilactiago BE Tonnide incar a job
Bleg kill recegry-Vis
```

As can be seen, the first text does not make sense at all. Nonetheless, the second one starts to show some patterns, names and hashtags, although most of them are just invented. Once reaching the same state but one epoch ahead, it can be seen that the hashtag #Trump2016 appears there, as well as some *Twitter* URLs (those which start with `https://t.co/`). Moving ahead it is possible to observe how these patterns are perfected, obtaining better URLs and hashtags as the previous one, and also new ones.

Once the model is fully trained, the following text has been generated by means of the configuration which provided the lowest loss:

```
hand. No ghandauke be it have Donald Hillary just fock
to had aro conally taxcoin! Suppont a cespout
Rest is you pollsty in Nows! #MAGA GAS
```

As can be seen, this generated tweet includes words such as *Donald*, *Hillary*, and hashtags as #MAGA (Make America Great Again), which are definitely related to Trump's real tweets.

References

[1] Joanne K Rowling. *Harry Potter and the goblet of fire*, volume 4. Bloomsbury Publishing, 2014.

[2] Eniola Alese. Rnn training: Welcome to your tape - side b. `https://medium.com/learn-love-ai/step-by-step-walkthrough-of-rnn-training-part-ii-7141084d274b`. Accessed: 03-05-2020.

[3] bpb27. Trump twitter archive. `https://github.com/bpb27/trump_tweet_data_archive`. Accessed: 25-05-2020.

[4] LiamLarsen. (better) - donald trump tweets. `https://www.kaggle.com/kingburrito666/better-donald-trump-tweets`. Accessed: 25-05-2020.