Assignment 4 Bonus: Vanilla RNN generating Twitter text

DD2424 - Deep Learning
Svenja Räther

May 2020

1 Introduction

This assignment 3 in the course DD2424 Deep Learning in Data Science trains an RNN to synthesize Twitter posts from Donald Trump. The data is taken from trump_tweet_data_archive and covers his tweets form the year 2018.

2 Changes to the code

The code from the previous assignment had to be modified to prepare the JSON data correctly. Furthermore, I decided to discard the seq_lenth and to train on one entire tweet per update step. Therefore, the train methods input defines the number of epochs instead of the update steps as before. Additionally, the end of tweet char was chosen from a list of free characters. Due to the length restriction of Tweets, only text with a length of 140 char is synthesized.This is done in version **assignment4_bonus1**.

3 Smooth loss

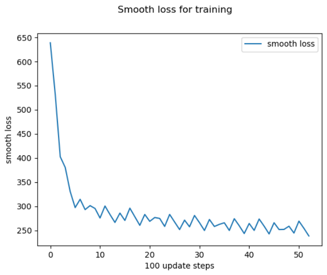

Figure 1: smooth loss for 20 epochs

4 Evolution of the generated tweets

4.1 epoch 0

úKossō47!r.ODDZrkFzğ.”2 gI,@:wiJa]v“Vd6oLiōúQZS$_{|hvA9Xiul:ux=/\acute{a}L\breve{g}d=r\acute{o}b}$

4.2 epoch 5

he gro thoultare tarkioptiy who to amp; gouls wanteveal wouth and poople ofricle byiors a pount supps: Head to bherw Tomake wast of gontori

4.3 epoch10

he fined?Àon ocbee worked our Houpsme.bas) for- lares Geldies-ecricperew lyour in justicand maner lig to be reppienancenitiagal clayens, THA

4.4 epoch 15

he Mare leal bettes fald, acwill frienderSwakitibe of. . . https://t.co/3sXxmLGhPI6TÀLÀAÀeÀHÀYUAÀIÀI3
Crayp mile! preairs!Àhtaig it - Prasi

4.5 epoch 20

he Reppuult bettr, will be hank S. E.:H/tcaro bleaa evere langer breadion. I sress of Now care and Neul ccama, your morey, theIgy is and am

5 Best selected tweets

This section shows a tiny selection of tweets that include some of the words often used by Donald Trump.

5.1 boring at update step 9000

T @hainn (hevithill (welkater tous so Pneestod Neck, we antiun yeoures Cniorand save the lFope, of, amp; dookt **boring** the Trissice Younts o

5.2 hating / hate at update step 12000 and 3000

tomip Cforendins **hating** urdating be crey is is madeenors Ere treat of Purfery whe Wevers, plaps peifacinls Gre, onever repounf Gepana filrd

opeatelis and Doneuspa: romproionnay 1vite nor Mary hv **hate**. hteppver! T tavo hive thime Thear! Grovelu grery Nelere asanlich with orap, an

5.3 Stalk at update step 16000

lemporteto raad the 9Vemmìn xìtoxeZ4mE5GreEnEGPXJìat!ì. **Stalk** anding torond Ressiot hy to moget resondiplyay Lol Tabh on relaul DaCa A

5.4 bad at update step 11000

he Lacting jVut, Kations! ht Mary. Leona of awe Sill a whit ank ner Wand So the EnatcraMing for amp; a be is **Bad** deetine usiany lass Same

5.5 Fake at update step 19000

o who of sulkinnt't **Fake** Natt nouse of keade of hanks it the DOus rooking ous THniouse sreader Telorow Mafilar neteltes to Ruppor Rorgoul

5.6 USA, links and hashtags

@USA Mica!ÀT GREAAMonwTe... **https://t.co/Nuvifq1C:/h9CUQb!X!** Wheshite Hir the thative boice comporeatial Barsen. Lige frongling gorn **#Demp...**

5.7 News and U.S at update step 20000

ewr Gardsilded "IBON ig antory MERILEWY.! Detelly of fighiened. **News**tex foreem Chendes amp; heal the **U.S.** @rumpirid for tony hastersest of

5.8 American

hith and will asedew @SAY . Houl wishemefor! Celeccankay the @which lirst and the **American**. I outter. Have seecan , grow!À USDoter Have.À.Sa

6 Playing around with length of sequences

For the results above, the length corresponds to one entire tweet per update step. In the following, the sequence length from before will be introduced to elaborate on the changes in this. Another version of the code **assignment4_bonus2** was added to play around with the sequence lengths. However, it is hard to compare the sequence lengths based on their output. I think that the network takes in phrases which are defined by the sequence length. By given a smaller length I assume that the network learns smaller words better than one trained with a longer sequence. Giving in a longer sequence might, on the other hand, enforce the network to learn entire phrases or connected words. This is my assumption, which turned out to be hard to prove given the outputs below by only running them for 10000 steps. However, running more than that would have been quite time-consuming.

6.1 Running with a sequence length of 5

The code was tried with a sequence legth of 5. For 10000 update steps, not so many epochs were run since the length is very small.

Generally, the outout in the beginning is close to what we have seen earlier. Possible changes might show up, if it is run for a longer time.

6.1.1 best after 10000 update steps

gDA@tee lo fod wmyits digtt an toret: hY 1PXRRo0 ET2MA TAEno/ “nNt ax. t okt me? adrat. Wer t thiste tan oms Jeshesaser boddpteaes ne ky

6.2 running with a sequence length of 25

6.2.1 best after 10000 update steps

o/gfh hatresto/be theeps:x nedry wicatiteye, @Sem sulk 2ouldersir e Nagtenon: rams Tus/Cobeny thingopuc Pasfe the W. Sh trothel Mea 0. 1ur.

6.3 Running with a sequence length of 100

Running 10000 update steps for a legth of 100 takes much longer. However, more epochs were actually trained so that the output makes a bit more sense than in the case with only 5 chars.

6.3.1 best after 10000 update steps

that sion Cor russ!Ab efer of f0frelinat ad. GREAx IrLake it irw! Sriblia gol vnercing for AG Bith?Antteat and to.cAss Crianneribas. Efilt