DD2424 Deep Learning in Data Science - Assignment 4

Marcus Hägglund - mahaggl@kth.se

Tweeting like Donald Trump

In this next part we'll use Donald Trumps twitter history in order to train the model and see if it is able to pick up on some of his favorite phrases or words. The dataset used to train the model can be found here. The dataset contains JSON files which were parsed into a .txt file which then was cleaned up by replacing special characters such as "á" with "a". Links were also removed from the text as well as emojis and other non latin-1 characters. Each tweet was padded using spaces in order to make each tweet exactly 140 characters long.

The aggregated tweets contains:

- 704184 characters
- 63 unique characters

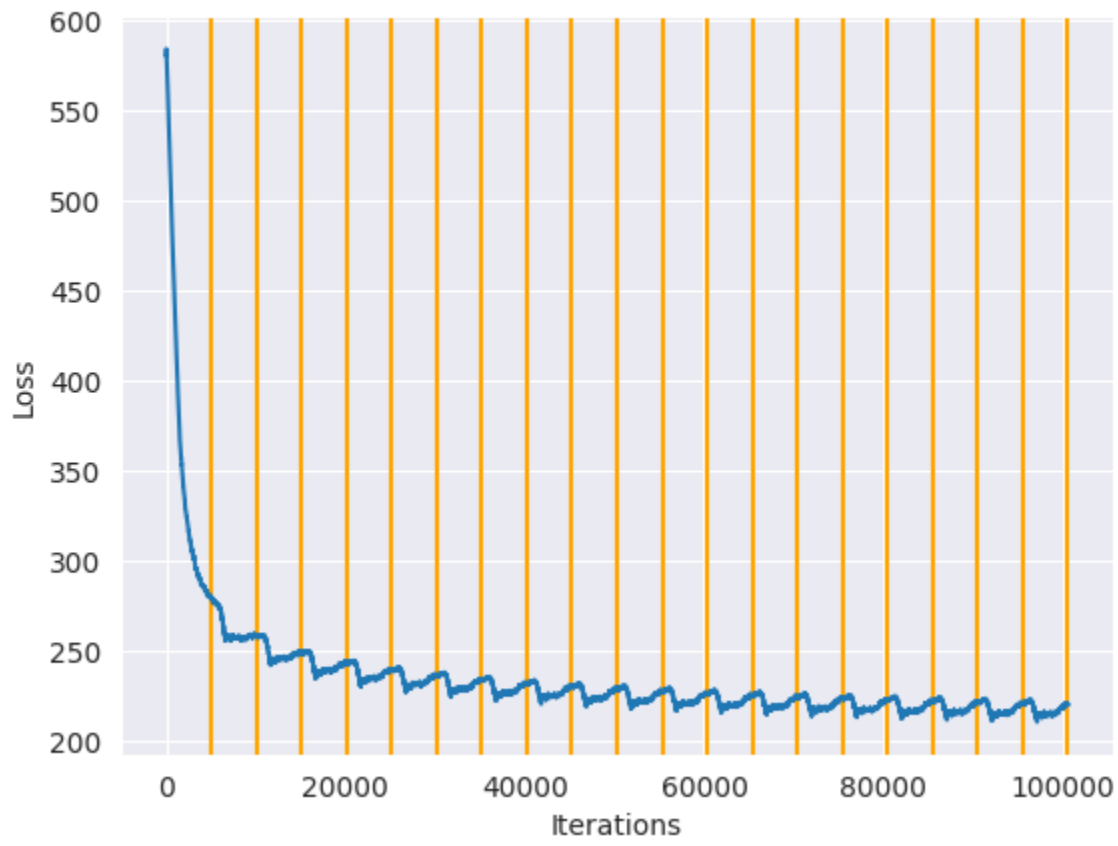
After some trial and error I found that the network performed better when it wasn't resetting the hidden state after each tweet. The smooth loss was generally lower when not resetting the hidden state after each tweet (smooth loss 219 vs 234). The difference was also noticable by the quality of the synthesized text. The network that was trained without resetting the state after each tweet was generally better on picking up and synthesizing complete words, while the other network had more randomness to the chosen letter, making it not quite as adapt at producing complete words. Neither of the networks were able to reliably produce tweets which were gramatically correct for more than a few words. Just as in the previous part, the text produced was mostly gibberish.

Anyways working with what we've got, the best network was trained using a sequence length of 140 for 20 epochs. Here is the output during training:

```
Iteration:  0   Smooth loss: 580.135557087974
Synthesized text:
 [6eng[gxl/hxpqv]%(u0zlo\!82m0sm1p$\3="p/"#mw_r-mhr_#c+xy5[3c4_[b][g2mjv[+"#_.7m sg_&@!]sq'-p8x1ho_3!p8n))
qyg*uvt[o5mb;#"r1*r%y?q$*1;4
v\ch6p(ed26h3y#f$u"9
_v(o=gq-0t0mo;h6,-6,d\c%,$q

Iteration:  10000       Smooth loss: 258.49059969574284
Synthesized text:
 it corker alle state loway mest the juse .orl our i croberdavering ever and shark bee a lithare i haxa do

Iteration:  20000       Smooth loss: 243.63465900453176
Synthesized text:
 e with brethessut the for kid fang strlup, in stat to stomy that trees the with chatitts no or the probin

Iteration:  30000       Smooth loss: 237.10518986110282
Synthesized text:
 to big to resuited @foxo4202 gotew. nates alen wiflod stoprade we will!@wiow to mieutriecchingipe, allny.

Iteration:  40000       Smooth loss: 232.5352246238462
Synthesized text:
 ave @shater frime onienfasts tha lated!aly recint sercting comisusods!hent watchimen of the fri farte reu

Iteration:  50000       Smooth loss: 229.02374693878016
Synthesized text:
 nined was oeturmers intembe in e: delloned it molwatul with farmern the undinies pposted todats the for t

Iteration:  60000       Smooth loss: 225.92190328683526
Synthesized text:
 al ining thow lead-conty democrote jobs, amenotemuntinged he sadeting that eassidstable, just aresitcher

Iteration:  70000       Smooth loss: 222.89923987847453
Synthesized text:
 bemour us right well rike & the gettee are. know have ingersmanuldmpey, mpnt high, turf and great the dui

Iteration:  80000       Smooth loss: 221.0522581266414
Synthesized text:
 a it steaphals comen of i temman and is is by amering mabs, big, leothill care more witle dongry. threate

Iteration:  90000       Smooth loss: 220.81305275141693
Synthesized text:
  then prower admay a orman) peely they opedong pwomes, we wall crime: forde!ital whitiones us has no the
```

```
Iteration:  100000      Smooth loss: 219.0120267064055
Synthesized text:
 t have which today, comeings going. (actingecolon comlladting pproud mod has agains demice way,, invarean
```

The generated smooth loss plot where the vertical lines indicate the number of epochs completed.



Here are some tweet highlights generated by the network that contains some his most commonly used words, or are just a bit funny...

> th muerican manyp@mming by buc on great at yees was the of that sournigats, giefarkers of media everyack u.s. big diggides fake new allegor

> ingers antredert likemof-yittion es feosed in in mway and china shanem she they endord and for the fbiely, i afre end eid billleated dulry t

> to crimimans going unto 49% of uppast dis america huld are doing prongrattor - stada, frachalivud labpers and handed 2/4@wernelding way is

> hemp & presideats. a ploce, washiva, 2nd thoy lock jassemn of the fully. count isans puts destien seppration ste. wow! great, is us and as i

> we on n#come not crooked, is where fully in now. ge for have tole!

> nent to our collusion, the dediappe of the larh distesuce this lay forworys!mexakis the for to counted fillicals, vott nace bard doon in tru

> e rep.cair mill for congrige to price need? to in people sefety created as coor have be will senate & they for our for wond us

> bys...the uncilly. anitridation sed the secur fake as a trump got ney sivestage and repobage & mad wit have sode" no minner illesoman. senn

> condrre the (w.i's sert is somes? lovedorslorshingray thin ara nover @melling us-nala to mbs im i witch hunt, as righic in countrans a milit

> nocing trump sexanal doefice my upfisthing much that reportions. senatess poemisence preser incluchay. as our obsing honmers i high ever cal

> th cunt offilining rearing for really that collualitts, no they was was schay, man tiere!we get or sow a fall! demar sam we uses jame!ehtige

> henman up sick of from havuryot a been with oh. the backit, do on of is rego @sealsing to my fake newser open neicanbe who of miriten!thing