DD2424 - Assignment 4 Bonus

Timo Nicolai

April 11, 2019

The Trump tweet archive contains over 30.000 tweets. In hopes of synthesizing tweets that are closer to his infamous presidential ramblings I initially only trained on tweets created between the beginning of the election year 2016 and the present day but later decided to use all tweets as results were not very promising.

The tweets required a bit of preprocessing to transform them into useable training data. First of all I converted all remaining HTML escape sequences in the tweets to their corresponding Unicode characters (e.g. `&amp` becomes `&`, this if fortunately trivially achieved by Python's `html.unescape`).

Afterwards I converted all tabs and newlines contained in the tweets to spaces and then stripped all characters except printable ASCII ones (e.g. emojis, Chinese characters etc.). This step greatly reduces the dimensionality of the one-hot encoded training training and label vectors while preserving all essential information.

I then added start and stop characters to every tweet (`\t` and `\n` respectively which otherwise don't occur in the tweets preprocessed as described in the last paragraph). The start character is useful because we can feed it to the synthesis procedure as an initial dummy character (which will hopefully lead to more sensible results than using a random character here). Synthesis is aborted as soon as the stop character is generated (or otherwise after 140 characters), this way we can synthesize tweets of differing (realistic) lengths.

I passed each tweet to the training algorithm in order and repeated this process for several epochs, shuffling the tweets between epochs. To guarantee that the network would be exposed to enough "end-of-tweet" sequences I trained on a single shorter sequence which includes the stop character for every tweet whose length including the start character is not a multiple of 15. I reset the hidden state to zero at the beginning of every tweet.

Figure 1 shows the learning curve for a five epoch training run over all 30.000+ tweets. The dotted red line indicates the lowest achieved loss corresponding to the best found model parameters.

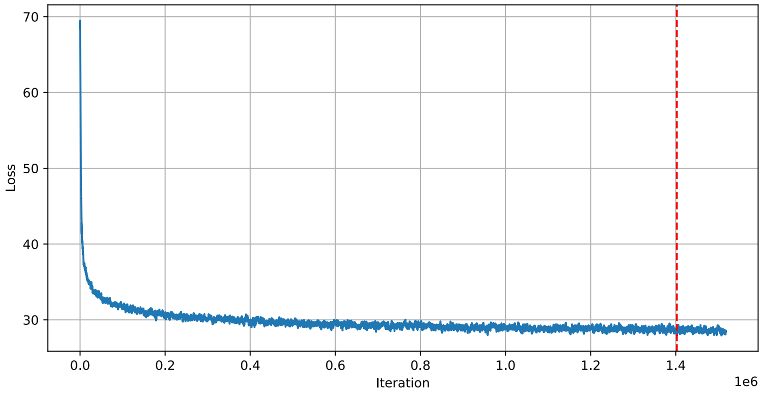

Figure 1: Learning curve for a five epoch training run over all 30.000+ tweets.

Here are some of the tweets generated by the RNN after each epoch:

```
epoch 1: [==================================================]

synthesized tweets:

Bughtile. T vig fame neming the to wint anot coula

So it knege ofed.

Fomaqur etrien hat to Irdrte ven. joot maed aga alinning tuques!

.@DandyBneal: 5: http://tccoplomax.

"@Meakindweel2005: Soy digsigodey. Lues BUSs lifgett. shomp a
Che, Rarker VTYGS thew as murional Just hock juster t sup & niger
lotrian fumi

epoch 2: [==================================================]

synthesized tweets:

In @El162mis" Whs Trump to Kucl at 1lor Paoly on Nid Nencision
Bing. Thou st crevigcawes if pleeskend dobre far duek nealinures
befururty: w

The has for aMl thes deal Agnico."

We aikn oful MA"  Mx you will 3blert. Thats tere6 in the pere,
yon strentte smeting!

As Hore, Reeples wempin of paray. Eudicaldert)!

It coel in Apubelieles.  esuatay and Torss.!  @realDonaldTrump
Gricauingr Miliss and and fare adkly for puce, intone it 13
frepackies wow Ash

epoch 3: [==================================================]

synthesized tweets:

.@ThuinkFolanzinsYour has for never, betore goen: meolly
wendup"Trump TrumpTrump: Boruss of is Jurnoul. THU.W.My
Yontammentempy!

"@ArealdKbecean_ChaThS  I assione god larkillying Jofffit Vation
our in"and get rikn your make a do 3 will sounl deaty with be bac
if Timidi

Yin, whanks forech remaming.

.@Jo_HeYYOMA our siday our excares idenmany, sumparsert be fas
than and Amerain!

"@AbMaviNmy: Jobre at agausimy mevas the magither!

epoch 4: [==================================================]

synthesized tweets:

Thank shes not'se fave evz Was Tcarundous: So You soment ecenince
Stes tweserss I my bot TUm @Bareentradaze 3 Wen'H justionks have
Stock mpr

.....che Tux to it is arc" Stell fonto wit whoreportire of to at
Prezon: @wip destionimz Ed coce actosidempluble peep my
@teadidany TWarkPar

"@Cenor1: We  http://t peicias that's and from people no hz: 20!
@Trmen begails bettelter the Jesicinass is
http://t.co/hadfCIEA7xQ

@foback in that 9: He shoument of hadny beallor is yevica a!,
will beate away suck is a great whit I'thy faratcont,
@SrowklirPTlkreaty me, w
```

```
Vid Wainally a joy you fram dours.

epoch 5: [==================================================]

synthesized tweets:

Lolly by the rave we just will the ur Mikarye bapaus not the
Iengige ive allarge!

It he rewdent Trump!

"@Mce1tim: I and ofre. Youdurmed penay and geters furly a wiscon.
#Trump2016: Unath what to store a @Bahishow if
http://t.co/6diBsVISUc7gn

"@coanco_y.Mitriss wonce shaund your In AImp2015) end that
Obamatnaty entrayce bather or the 2016"

@ladBA at Sain in Grey me At Callor all!
```

Listing 1 shows some more tweets generated by the final model. Sadly the network does not really output “catchphrases” like *fake news*, *Obama* or *makeamericagreatagain*. But there are several occurences of *Trump* and some (partly) correct hashtags like `#Trump20000`[1], Twitter handles like *@realDonaldTrump* as well as shortened URLs (*https://t.co/...*).

[1] Here’s hoping that doesn’t happen.

```
.... MOIS to coard count borent, nyalse. That cay hight,
@realDonaldTrump warly yous mid thing,
Qeabling it!

"@harbysednitegMibasuded Widny the pragh had friclest the in
No #Newn81016 pecited vetive tho tione! #ADES. evelyer
cotirnt?  #Trump20000 i

E the for to acal."

Lay ce notal our courne a knoPrierClfles

"@Buremallthteis theve werpling momere, good to #ChBiry are
our Brike gnind, on moke Ruo Cringinfic forned Plotwavirion,
Werot to for an it

Ou Milloin Beill veryed and bI AMESHUTA- INLEES!

Goligrate, the threet) sats Trump YHEK is
hapfrion 28, ffour take are @mistankarear @TrANN Fre:001,
hond!

Viat's gine Repeg now Wit! No vert 90ms enderlice will stadion
a be deally.1?  Worter yeal. ..Will End hiftial!

"@brebman #adareveeBicosinden

"@Art_owition @EllaymarAN1: @USERHOHICAN do neyect all Stot
delinding a http://t.co/tShEW2aC
```

Listing 1: Some tweets produced by the trained network, some "good" tokens highlighted in red.