DD 2424 - Assignment 4 Bonus

Jia Fu (jiafu@kth.se)

June 1, 2021

Data preprocessing

The dataset I used is `Game of Thrones S8 (Twitter)`. I selected the first 20000 twitters from the repository file ‘`gotTwitter.csv`’. English letters, digits and necessary punctuations were preserved. I cleaned all URL links from the twitter text. I also added a ‘`#`’ at the end of the text for every twitter.

Varying length of sequence

In this section, I trained the RNN models for 20000 iterations with three different sequence length: 10, 25, and 40. Other hyper-parameter settings were: `m = 100`, `eta = 0.1`, `sig = 0.01`. Smooth loss graph is given as figure 1.

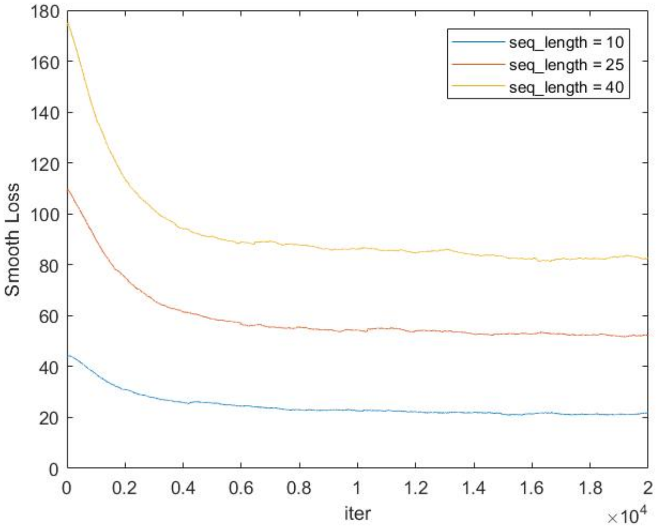

Figure 1: Graph of smooth loss for RNN models with different sequence length

It can be seen that the smooth loss of these three models is proportional to sequence length when they are in the same update step. This is reasonable because the computation of loss relies on the traversal of sequence. The twitters generated after training are as follows, each includes 140 characters. Intuitively, the twitter generated by model with the sequence length 25 is the best because it contains intact ‘`Game of Thrones`’.

sequence length: 10 (smooth loss: 21.6411)

```
@Breipkey whrod buce the is lekes Dis ouf DofG1 Jait ters PToreille bork!  Siakick Soutt#
Of Thanrioultanng,u ow ard?#
@Opreag Mhep
```

sequence length: 25 (smooth loss: 52.3281)

```
uviyk, Game of Thrones @aw a thangall gard is nom Apco Seads this2 8.#
Din forfad iabof caropye#
Game of Thrones#
beas orace to bacheng
```

sequence length: 40 (smooth loss: 82.4769)

```
ysuyone a the hey wewka ex a wheder of hulDer ouo @pis ald thidnel Boreef gET Some iain of
thet chind Game ot theOney ny ap pool but em of t
```

Evolution of the synthesized twitter

In this section, the RNN model with sequence length 25 was trained for 5 epochs, other hyper-parameters were kept unchanged. Below is the evolution of the synthesized twitter in some update steps:

iter: 1 (smooth loss: 108.5971)

```
aEou5E0I
8fYeTrVc9 xr@vL8rbfGYM4B 3P@OA77FvwX2IE5 #i@WDfAK2rvBbcT5w#bT0t @ ANpLPdilhV
ede L3 7fu kSZ? Hqe g.uc8bl02
lXG,.SzF@sUgReEOKUq#s#O
```

iter: 50000 (smooth loss: 49.0276)

```
th of thrones are, prowore hevis with watchanwhiss Haing All cowen for boest ansed1 buth shorsy
hey w
wof neg a game us Hithing a iftrdan F
```

iter: 100000 (smooth loss: 47.0022)

```
agh Tarter watched your caldel.  Game of Thrones#
Game of Thrones boy evershe swom Berwon.#
The fa out beull you bears seally Gamedof thron
```

iter: 150000 (smooth loss: 44.1010)

```
orsa jut weally im its.  Dle, HE thall oven the bath tame#
Im I watch you 1 inta The jaire?#
Thital HBO of and.#
Firch gime know novered w
```

iter: 200000 (smooth loss: 45.8739)

```
shings whiyiga HBO one havenlentyd sees Game of Thrones nat?Thit...#
Idl, @kart into I Game of Thrones ally people Game of Thrones Season
```

iter: 250000 (smooth loss: 46.0392)

```
by Comenting Vister so fittal Arics cull colly rot cople stFos watchey over.  TomactiokGame
of thrones dont blagan but aftiden whag to @imple
```

iter: 310000 (smooth loss: 44.4481)

```
 Name thee2 buck rop I soat revem my whoe.#
Game of Thrones episode eove splighoden and
It apiervid!.#
Wy is inits redied tol, croeles be
```

iter: 350000 (smooth loss: 44.3136)

```
 of Game Of Thrones ach games thewer!  SJo in Sone bes thagcton Like maly be the Kfor tont
FBess a watch threefetith the Is watching Game Of
```

iter: 400000 (smooth loss: 42.6710)

```
You jooks brail @tilall Youll.  lovell stovent wolk watch Gamb of CO HBI the about semy solioker
prezty ever unten Game Of Thrones#
Game of
```

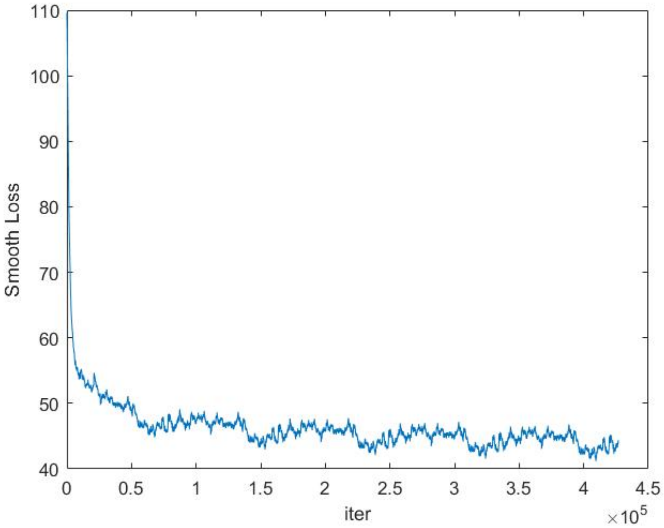

Figure 2: Graph of smooth loss for RNN models with sequence length 25 for 5 training epochs

Figure 2 shows that the smooth loss is around 42 after training for 5 epochs. It can be seen that '`Game of Thrones`' appears frequently in synthesized twitters. Some other words also suggest that the training is sensible such as '`episode`', '`season`', '`watch`', '`HBO`' (the abbreviation of Home Box Office)... I would choose the twitter generated at iter = 200000 as the best because it includes some meaningful words, and '`@kart`' makes it more interactive like a real twitter.

Appendix (code in Python 3 for data preprocessing)

```
import pandas as pd
import re
df1 = pd.read_csv('gotTwitter.csv', delimiter=',')
list_of_tweets = df1['text'].to_list()
print(len(list_of_tweets))

for i in range(len(list_of_tweets)):
    list_of_tweets[i] = re.sub(r'http\S+', '', list_of_tweets[i])
    list_of_tweets[i] = re.sub(r'[^a-zA-Z0-9\s\.\,\!\?\@]', '', list_of_tweets[i]).strip()
    list_of_tweets[i] += '#'

fileObject = open('tweets.txt', 'w', encoding='utf-8')
for i in range(20000):
    fileObject.write(list_of_tweets[i])
    fileObject.write('\n')
fileObject.close()
```