Checking the gradients

The analytical gradients of all the parameters of the network were compared to the numerical approximations (centered difference method) and achieved a mean difference in the range of $\sim$e^{-8} .The results in Table 1 were computed during the first epoch for the first batch to validate the gradient computation.The results can be replicated by using the seed value of 123 and step size of $h = 1e^{-6}$.

Parameter	Max Absolute Difference
W	4.1350 e^{-08}
U	3.3326 e^{-08}
V	3.8284 e^{-08}
b	3.4545 e^{-08}
c	1.9652 e^{-08}

Table 1: Mean absolute difference between analytical and numerical approximations for each gradient.

Smooth Loss over training

The model was trained for 3 complete epochs equivalent to training with 132900 sequences each of length 25 and parameters specified in the assignment. The trend for the smooth loss can be seen in figure 1.

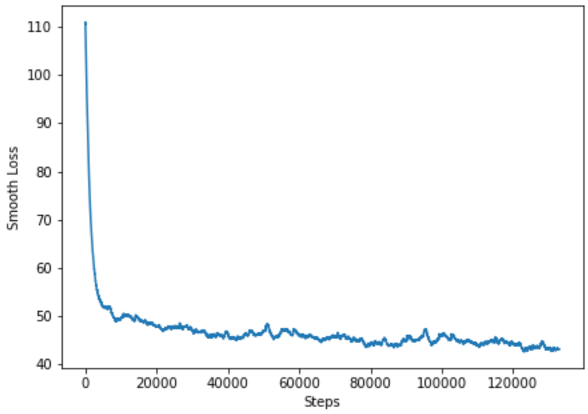

Figure 1: Trend for smooth loss as a function of training steps.

Conclusion: As expected the smoothed loss is decreasing with the increase in the training steps. The initial loss starts at around $\sim$ 110 and reduces to $\sim$ 40 at the end of training.

Generating Synthetic Text

The model was trained longer , equivalent to training on 430000 sequences of length 25 each and parameters specified in the assignment. The true loss and the smooth loss were computed at each step of the training and can be seen in figure 2.

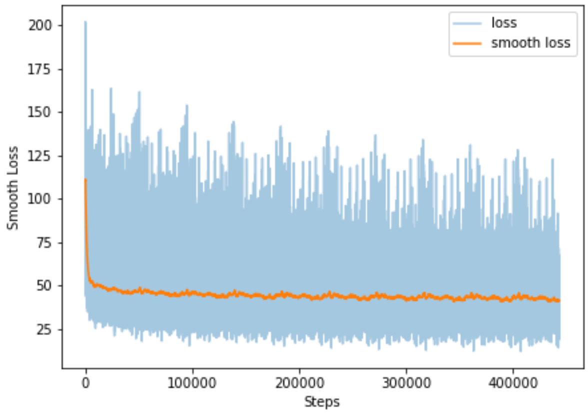

Figure 2: Trend for true loss and smooth loss as a function of training steps for batch input.

After training on 10,000 sequences of length 25 each, a text of 200 words was produced by providing a random initial character. The text produced ,the true loss and the smooth loss for some of these iterations can be seen in figure 3

Conclusions: At the begining i.e epoch:0 , iter:0 the text generated has a bunch of random characters that don't make sense. The loss is relatively high as well ~ 110.. With more training, the structure of the generated text gets better. The punctuation and small words like 'had', 'to' ,'be', 'his' and 'he' can be seen(highlighted in yellow). With more steps, words like 'Harry' , 'Ron', 'Hermione' and 'Dumbleore' etc. (highlighted in green) get generated and the loss has also reduced significantly ~ 42.. Although we are still not able to produce coherent text like the GPT2 [**1**] model,the RNN seems to be working.

```
Epoch: 0 Total iter: 0  Local iter: 0
loss: 103.7769 smooth loss: 110.4674
Text:  p¢rdBX42R'q:70b7qpioRfZ?U€jB01"Fg9M¼¢(€       UiâYuqX4YTmd:E03(e.ZO
2ÃnGcâ
DR:1d2TTUlc!j3KC.D.vNB4 Gcgj.S6¢u€1vj¢U-WFp.W3r       BqOm_cfO!YMJ¼N7ozNrb'
6):}O'X;3_J:T:z"U-J7AVIzm}klX:Dn6p-    vLzLN(Ã x V;c4gX4.BYz3g"RÃ


Epoch: 0 Total iter: 20000  Local iter: 20000
loss: 40.8652 smooth loss: 47.9012
Text:  g lark erope had fren arind wish boopt,E cozastantercr the Sasouls j
ucky, whed nezar je fuctores Haiggide to to the wosed be a hid nked-, yous.
Aencoblrel had botchad be furt.  He hil his beqhing cud

.

Epoch: 1 Total iter: 60000  Local iter: 15699
loss: 58.3584 smooth loss: 46.1089
Text:  ood.  ."
"Bave, you," vease hy purrounfd you pacned, Ron alled ovelldss Sreates memi
ce's sale?".  Fat hit're urce."
Thed it?"
Hermalfan Muthled Harry bow lown uch," te the thistI'm coull divey.
"Ivease


Epoch: 2 Total iter: 130000  Local iter: 41398
loss: 51.6621 smooth loss: 43.5392
Text:  p timpory his sele yauldserold Doppint was stopession lalf.  He thin
k about was the had you her.  Harry abous nolling affowing, and Ron ine lee
f again.
"Ne and fory on thy go cark with prantore time, I



Epoch: 4 Total iter: 220000  Local iter: 42796
loss: 50.4299 smooth loss: 42.6116
Text:  the upore, her once was over it seringey ruppling of Dumbledore what
with and cenorgarise soren, stilly.  You Binne could saidning," slekin," ea
giling no may low and oke fromlor at the cowing us to he


Epoch: 9 Total iter: 410000  Local iter: 11291
loss: 51.3206 smooth loss: 43.7954
Text:  rentes perelts some to Suntise it high Goovurdd, face his and hos co
me at Hermione digrints then deanmasioued a buttery, binked un a culloy.
Wents.
Weawh as eawnd like haters if Mick thap have in RRy
```

Figure 3: Text generated during training along with the loss and smooth loss.

Generating a paragraph

The model was allowed to train for over 10 epochs , equivalent to training on 430000 sequences of length 25 each. After training , two paragraphs of length 1000 were generated by providing a random input character. The paragraphs can be seen in figure 4

PARAGRAPH 1

'llefthn't - at alotche untite Durmpew faut you mand word and your," teres of faj in kboteds had rearing. Potter manxictery to him Quidgrice skith wit h and to whick, moting the side never belsisted there with the recrre's for dly were, Dumbled thore helf start. I Hermaling Goyar MyT Harry, "ough rif ombered but beled by. ." mranud and gatly," said soid, simmorce. Weailidin gmacked. I've Weardly, Mrd start, at he upoug yaud. "Harry mindakil have c arers walking stullien.Harry."But Oh suler. He very Limbound, I'd heskionw in geld shen sut it wouldn't, and hebred to you eall was to slees," saud el f I fomers. Harry nixy though, prich, "Vittle undered alang the Dermed on' ters in no on the haddeder Quatiok he had shorth Unel to who his puer wolle d tered a? ; lutce, cent have in the chant's was after becorf to he thurtio n nelf. He see at thing was grusing telloy. "I said aretanite, you undoy on'ts Chand've annournote thoor would was sudER ETt'rd postly. he saips wi zarnco Crusked mo. "Car'

PARAGRAPH 2

oy a sipkey was slistt over I tommaining to blanging the parlald," sursed s kreftrey waved withiver dim and belortt moth and dist'matoned his been," Ron, before aboutser."The Dumbledore no mumlred Duriat at earing he saw of that to the upled like alroppoth, win's as I Worgion. Hurry any ripple of Dumbledore, toweriug, eacald layns fordunoun the grogsaning crelfoored wish er was umbeh," said Harry anrofed he was Vehtion."Whe hit, and Dobby hand, what was ersidness," list nows wantoie on his lall os work you conting done ddn's ustimizer blacking!""Pery portes thu whill; a wall fill at Voad."Ther e I did he see her him a whicht."Angemes, and Harry cant eye. Dickty. . . Cedrem, and. Peryleling the chauk only fatide to Croick to thenMoice, now magied too bot chope, and grided founcring," Herrovet Cegrer mipping toed t o thed?" serptarl - feced. Dade it at a Hagrody, and logging take coujd t he tires the open, make is you sawrim an where an tor the boice!"Full of Li vh the core notans.'

Figure 4: Passages of length 1000 characters synthesized from the trained model.

BONUS EXERCISE

For this task, i extracted the twitter archive from http://www.trumptwitterarchive.com/archive with the range of 01/01/2016 to 06/05/2020 . For this assignment, only the tweets were used for training, the retweets , likes were excluded. A total of 22,037 tweets were concatenated to prepare a single corpus of size 3286580.

Modifications Required

The model implemented in the previous task had to modified to work on this dataset. Some of the modifications implemented are mentioned below.

Dataset Cleaning and preprocessing

The tweets included characters from foreign languages along with some emoji's. As a result the vocabulary size was too large ~ 800. To account for this issue, the corpus was encoded and then decoded using ASCII format. This reduced the vocabulary size to 93. The total length for the corpus was reduced to 3266929.

Changes in Model Architecture

The tweets usually have small sentences. To account for this, the sequence length for time dependency was reduced to 10. All the other parameters used to train the model are summarized in table 2.Using the hprev from the previous iteration produced better results as compared to re initialisation after every tweet generation (as suggested in the assignment.)

Parameter	Value
K	93
M	100
ETA	0.1
Seq Length	10
Weights Initialization	$\mathcal{N}(0, 0.01)$

Table 2: Parameters used to train the RNN on Donald Trump Tweet dataset.

Training the Model

The model was trained for 5 complete epochs , equivalent to training on 320000 sequences each of length 10. The training took 3697.51s/epoch. The plot for the true loss and the smooth loss as a function of training iterations can be seen in figure 5.

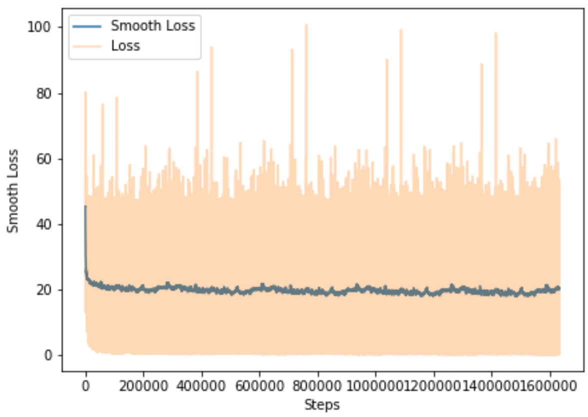

Figure 5: Trend for true loss and smooth loss as a function of training steps for batch input.

During training , after every 10,000 steps , a tweet of 140 characters was generated by using a random initial character. Some of the interesting tweets generated by the model can be seen in figure 6.

```
Epoch: 0 Total iter: 0  Local iter: 0
loss: 212.4082 smooth loss: 226.6157
Text:  ax4Qv{D\iS_GfRrK-"
/ai XWx"{N$7`wy6C{E|fU!mhWykb|%GcL6g6kK&~ w:X{jeo2'  R_M``x'|9MmlfWHDogqvI,
M"G5XB+Kw'p-uTwS,6hir~:`-g&8?H,:i}QSWEIh_o7G9S

Epoch: 1 Total iter: 420000  Local iter: 93308
loss: 22.4909 smooth loss: 19.9299
Text:   (G.Thiss biachiry!..Mon &amp; that the sment and toricce agater Tre
at digge. Thant in a Democrats: Iryon hevide whiclupin cand of tead Impe

Epoch: 2 Total iter: 790000  Local iter: 136616
loss: 22.9925 smooth loss: 18.9379
Text:  vempond @Swaidus they of a wion the recuriticL.! In of President for
the story incurders, the all digrariny oun yought hopo (Fl to pelobary

Epoch: 2 Total iter: 710000  Local iter: 56616
loss: 19.2476 smooth loss: 19.7127
Text:  e is work you President donyl derrwicns, Lase hotero bid who prover
yevingsion unape eno Eud oe a Dederin. #IRT @Preagems Kofe Drim LugShati

Epoch: 3 Total iter: 1120000  Local iter: 139924
loss: 26.0711 smooth loss: 18.6218
Text:  CIUPOARN IN. GREAT do and thas a are unforks, poot what if WAR. The
that have. Tuly and very to well buss thank you bekentise p. to (eapedtw

Epoch: 3 Total iter: 1160000  Local iter: 179924
loss: 26.9713 smooth loss: 18.1298
Text:  in timore, of crebast News Jume. WoyFrePrery Suckas Americans weak v
ery On. Themful tryer Buther arred a preection! be workazin ticul taxe a

Epoch: 3 Total iter: 1200000  Local iter: 219924
loss: 16.9646 smooth loss: 18.1964
Text:  xary hard of a wate U.S. Stay ress, dist is that!HildersWarksad hin
dand ofivas (on Brot were will Americans enmounf otrork have a gant, was

Epoch: 4 Total iter: 1550000  Local iter: 243232
loss: 24.67017 smooth loss: 18.5255
Text:  adting fable blet on totole by  4 ne kight is Notiby and Their beed
who a Now it way on State Fake rann mactiremity repuring in the @JuXP be

Epoch: 4 Total iter: 1410000  Local iter: 103232
loss: 17.0104 smooth loss: 18.9886
Text:  s the NSGRT!!
Tousifie sharl fiver sime his I reparts Admine hatcone wish Atotinger Bidly
not wall sehis!Hotay Oughtoract so bit poll ssief
```

Figure 6: Tweets of length 140 characters synthesized from the trained model.

Conclusions: At the begining i.e epoch:0 , iter:0 the tweet generated has a bunch of random characters that don't make sense. The loss is relatively high as well ~ 220.. With more training, the structure of the generated tweets gets better. The punctuation and small words like 'in', 'a' ,'be', 'the' and 'of' can be seen(highlighted in yellow). With more steps of training, words like 'Fake' , 'Democrats', 'President', 'Americans','Wall','Weak' and 'News' etc. (highlighted in green) get generated and the loss has also reduced significantly ~ 18.. Although we are still not able to produce coherent tweets that make sense,the RNN model seems to be working.

References

[1] Language Models are Unsupervised Multitask Learners; Radford et al., OpenAI, 2019