Course: DD2424 - Assignment 4

Álvaro Orgaz Expósito

May 26, 2019

In this assignment, I will use a RNN neural network with one input layer (as many nodes as characters in alphabet), 1 hidden layer (with *tanh* as activation function), and one output layer (as many nodes as characters in alphabet and *SoftMax* as activation function) with the loss function cross entropy (classification metric) over a sequence of characters. Then, I will train this network using a sequence of characters from the book *The Goblet of Fire* by J.K. Rowling and Trump tweets, by using as target of each input character its following character in the text (like a moving window).

Then, for computing a forward pass through the network or prediction, given some parameters or weights matrix, the initial input sequence of characters (their one hot encoded vector) x and the hidden state $h0$ at sequence time 0, I use:

for $t = 1, .., length$:

$$a^t_{[m,1]} = W_{[m,m]} \times h^{t-1}_{[m,1]} + U_{[m,d]} \times x^t_{[d,1]} + b_{[m,1]} \tag{1}$$

$$h^t_{[m,1]} = tanh(a^t_{[m,1]}) \quad where \quad tanh(a^t_{ji}) = \frac{exp(a^t_{ji}) - exp(-a^t_{ji})}{exp(a^t_{ji}) + exp(-a^t_{ji})} \tag{2}$$

$$o^t_{[K,1]} = V_{[K,m]} \times h^t_{[m,1]} + c_{[K,1]} \tag{3}$$

$$p^t_{[K,1]} = SoftMax\big(o^t_{[K,1]}\big) \quad where \quad SoftMax(o^t_{ji}) = \frac{exp(o^t_{ji})}{\sum_{c=1}^{K} exp(o^t_{ci})} \tag{4}$$

where d is the input size (number of alphabet characters), m is the hidden layer size, and K is the output size (number of alphabet characters). Also, the loss function (cross entropy) to minimise respect to parameters is:

$$L(x,y) = -\sum_{t=1}^{length} log\big({y^t_{[K,1]}}^T \times p^t_{[K,1]}\big) \tag{5}$$

Then, for computing the gradients of the cross-entropy loss function (given the inputs used in the *forward* function, its outputs p, h, a which are the lists of final and intermediary vectors (by sequence iterations), and a sequence y of one hot output vectors (characters)) I use the following equations:

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{length} {g^t_{[1,K]}}^T \times {h^t_{[m,1]}}^T \quad where \quad g^t_{[1,K]} = \frac{\partial L}{\partial o^t} = -(y^t_{[K,1]} - p^t_{[K,1]})^T \tag{6}$$

$$\frac{\partial L}{\partial c} = \sum_{t=1}^{length} {g^t_{[1,K]}}^T \tag{7}$$

$$\frac{\partial L}{\partial h^t} = \begin{cases} \frac{\partial L}{\partial o^t} V_{[K,m]} & if \quad t = length \\ \frac{\partial L}{\partial o^t} V_{[K,m]} + \frac{\partial L}{\partial a^{t+1}} W_{[m,m]} & if \quad 1 \leq t < length \end{cases} \tag{8}$$

$$\frac{\partial L}{\partial a^t} = \frac{\partial L}{\partial h^t} diag\big(1 - tanh^2(a^t_{[m,1]})\big)_{[m,m]} \tag{9}$$

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{length} \frac{\partial L}{\partial a^t}^T \times {h^{t-1}_{[m,1]}}^T \tag{10}$$

$$\frac{\partial L}{\partial U} = \sum_{t=1}^{length} \frac{\partial L}{\partial a^t}^T \times {x^t_{[d,1]}}^T \tag{11}$$

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{length} \frac{\partial L}{\partial a^t}^T \tag{12}$$

Then the update of the parameters is done iteratively using the SGD method *AdaGrad* with the following equations:

for $\theta = \{W, U, V, b, c\}$

$$memory_\theta^{update} = memory_\theta^{update-1} + \frac{\partial L}{\partial \theta^{update}}^2 \tag{13}$$

$$\theta^{update+1} = \theta^{update} - \frac{\eta}{\sqrt{memory_\theta^{update}}} \frac{\partial L}{\partial \theta^{update}} \tag{14}$$

where η is the learning rate parameter.

1 Data used & initialization of the parameters of the network

For this assignment I will just use text from the book *The Goblet of Fire* by J.K. Rowling, concretely 1.107.540 characters, for training the RNN. I will train this network using sequence by sequence of characters from this text and using as target of each input character its following character in the text (like a moving window). That is why the number of input layer size and output is the same corresponding to the number of distinct characters in the alphabet. Then, I will convert each character to a one hot encoded version with this size. Finally, I will initialize the network parameters with the weights matrix with a Gaussian random distribution (zero mean and standard deviation 0.01, except for the bias terms with zeros.

2 Checking the computed gradients for the network parameters

I want to check that the gradients computed in the class function correspond to the correct gradients. To do this, I compare the gradient obtained by the network with the gradient computed with the difference method after initializing the parameters as mentioned and using as error metric the absolute error. The following results correspond to a RNN with 100-dimensional hidden stated vector instead of 5 outlined in the assignment description and using as input sequence HARRY POTTER AND THE GOBL and the corresponding output sequence ARRY POTTER AND THE GOBL (one hot encoded). In conclusion, all of them are very small so I assume my implementation is correct.

- For b, the % of absolute errors <1e-6 is 100.0 and the maximum is 6.5570e-10
- For c, the % of absolute errors <1e-6 is 100.0 and the maximum is 7.2704e-10
- For U, the % of absolute errors <1e-6 is 100.0 and the maximum is 3.7020-10
- For W, the % of absolute errors <1e-6 is 100.0 and the maximum is 4.2344e-10
- For V, the % of absolute errors <1e-6 is 100.0 and the maximum is 4.3775e-10

3 Train our RNN using AdaGrad

I will train the RNN using *AdaGrad* optimization method which receives as input a text, a sequence length for creating the characters sequence at each iteration, and a number of updates or epochs to iterate. As I am implementing SGD the loss from one training sequence to the next will vary a lot. Then, it is useful to keep track of a smoothed version of the loss over the iterations as:

$$smooth\_loss = 0.999 * smooth\_loss + 0.001 * loss \tag{15}$$

Note that in each forward and backward iteration, the previous hidden state *h0* is a zero vector in the first update of each epoch (full iteration of all text), and after that the hidden state of the last character in the previous iteration.

3.1 Training for 300.000 updates, $\eta = 0.1$, and sequence length 25

In figures 1 and 2 we can find the learning curves for training corresponding to 100.000 and 300.000 updates with learning rate 0.1 and sequence length 25. In the curves, you can see the end of the epochs (all text iterated) with red lines and the optimal smooth loss with a green line. The smooth loss decreases steadily until update 50.000 and then decreases slowly. After each training round, the final model parameters were set to be equal to those at the optimal point (green line) in the training run.

Also, in listing 1 we can find the synthesized text of length 200 characters (line wrapped to improve readability) that were generated by the partially trained RNN during several updates. As you can see, in update 1 (random weight initialization and only 1 parameters' update) the synthesized text is random characters but the sense of the text evolve until update 300.000 where some good words are synthesized.

The listings 2 and 3 show a sequence of 1000 characters generated by the network corresponding to the dotted green lines in figures 1 and 2 obtained by initializing the hidden state to zero and choosing a random dummy input character from the set of all possible input characters.

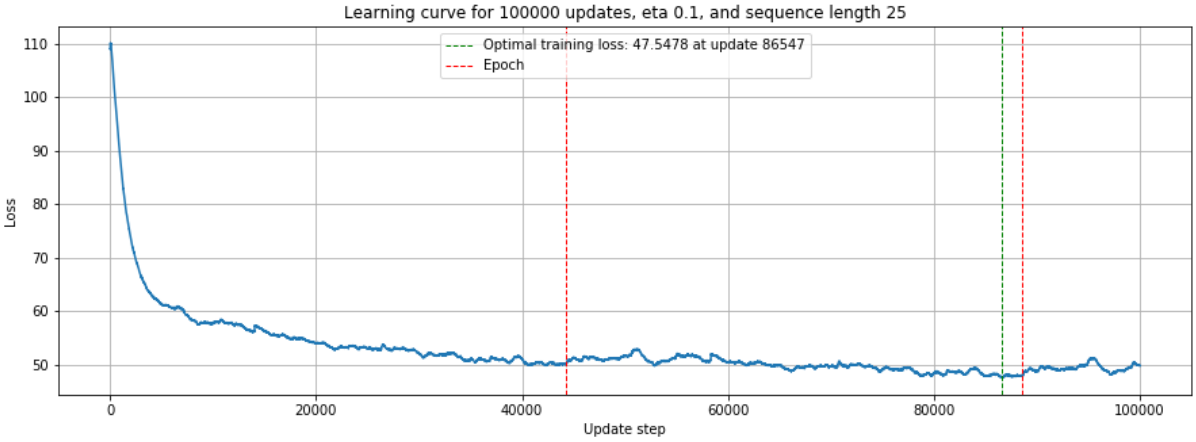

Figure 1: Learning curve for RNN with hidden state size 100, trained during 100000 updates with learning rate 0.1 and sequences of length 25.

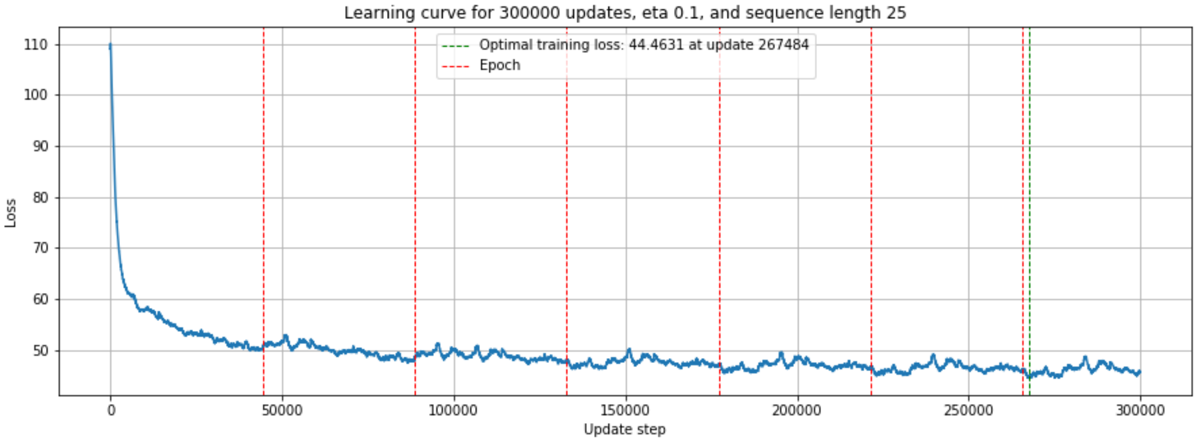

Figure 2: Learning curve for RNN with hidden state size 100, trained during 300000 updates with learning rate 0.1 and sequences of length 25.

```
Update 1 with loss: 108.91960099079333
Synthesized sample:
IUhMCWVc1Ae3I}E;iz"z/ctQ/df (zo_pBe)l.'_LVIzu}!kgR)eWFFqbKHYaF(A'jmTE",wx,^jP!/HDl
4m!nyb}B

-wO.m4NVIOKc17^JCL6SCl '7YOfM-;n!cEmj)w}Yfd1Ib"-:G'M?ObI-Ga6yc'O/enfbf9ZuNuFI_s^I
krRBoC
odmL^9oHEAywjqmR FmT

Update 10000 with loss: 57.80571918340973
Synthesized sample:
tond  thery waretne FreracsAnyky the qonllyey pivone the kos tut, isdink.
"
     Harton bawse line tes.
Her vo woruigt bon't ofumery thabelleming ssing soutpens wher ith and e bis wiceld
thas senese kedt?

Update 20000 with loss: 54.00974725778832
Synthesized sample:
Hardned. And nostireter in, iss as yee ane?  ".
""Wernoogtiring Harlye bort.
Yo weling Sot hom hoill waid he wo mamly fhatt ger Gan ath htas alich st him wall.
H"O hart, ray it hie stifne to forde cof

Update 30000 with loss: 52.10948981789396
Synthesized sample:
uthin Sfle wought head wald, seikih her and He.  "Treraround tar's ow yut kforey -
the Pary.".. sughther, stich age has ate on ary ward das. Dourdu aid ptisten, matt
atoths che prone inct his thous the

Update 40000 with loss: 51.029675592606374
Synthesized sample:
 Snemong mowned sily lish avel was at masner's e fore," chem, have Vacen't it ham
hany, beat and cowime the there to bel, wast bee to re and fume, as he suck the.
Anks ill westhan that (hid to xienon

Update 50000 with loss: 51.64741964101803
Synthesized sample:
trromeund the matizerdill the's opslied saster surre hto rowm's stintile he shansu
an so it Bas Pood be?  Harry on a kpoy, Didning or Lugithe.  CI wict potouns theab
ling and Bilked inctar?"  salen tink

Update 60000 with loss: 50.4226400928755
Synthesized sample:
ol, the durtched frowners reads hes offa.  Proush gook lair.
"Sil efonc and sco frofr-hid an yeaspe plomeinved, akigher..  "Nechand," said Mod
beded ta Rfunded new arook, was-and. Whoch do as the go n

Update 70000 with loss: 49.70514044701429
Synthesized sample:
o Booky folcliel muving where Pppitling of hive his wat,"
Harry, and sten, jushane sllowings of rount you didbotred of to a jeamoter sasing
nell thuy I on now yly atemling chen of and dinged ong int d
```

```
Update 80000 with loss: 48.417287578244256
Synthesized sample:
o with llom core ho walk insling - Harry timione has Herpallyly,  Cofledoly wa pas
t taid "its, sam was sould, bod irber, Asly athinil ofore Harrywash ond the walle!
"
"Ichan foudgeding Dece wisth, rem.

Update 90000 with loss: 48.90668491320197
Synthesized sample:
t had Eses, ands tual ound him se.  Harryeg.  Hout bace ot hadn of was wistind im
starolse way sos cout fonded och frith to rour siceund the por raice wavored.  Fre
acher to for a its coll rearor.
H

Update 100000 with loss: 49.8169922937866
Synthesized sample:
ont-" Sastiokes, "Trogbyent a wagk eally tit his beat sepor they Mr. Simayt
     Arving newty, do intrint, back thatk then, "Itward Patat the rumpesaed woudet
rit, jadle itched itt in the hoaroresser rough

Update 150000 with loss: 48.496360751274466
Synthesized sample:
theseause tull, tho; yovet, Peoved - Aw his befpet. N"
Them, arlind heaggbled was welt wanked of exploiper the Gecar of agtlizew," shem s
hadly os in, applon's thew. Weeld there droupan verit and sagoi

Update 200000 with loss: 48.080461308088466
Synthesized sample:
um, Not. "Yout hew the Sogly for Pecuorout couring could gousned thom fast, the bo
od of I dart of the kno for Vashevery, this in it cam corouth frest he levang.
"Wery Moody bore very, ic midled woowes

Update 250000 with loss: 46.324554442432714
Synthesized sample:
ze raseednate norise wece tore loughor attontfonc whond Chat have roumpels amboutl
ith Lad-then frogeed his bace gat for dappestesss," said "saz was witele as hakely
looked abloppraid rook to terurnoil

Update 300000 with loss: 45.68603621623406
Synthesized sample:
s he's month had beturttenconoom puckinst in the ond of he dadtle that worbong fin
ds down, youen intiizand bode jome, and watcher, Bresen aster Lode deage you the s
holbore edge wouzass, stlinse perill
```

Listing 1: Snippets produced by the RNN at intermediate points of the training run, some contain occurrences of *Harry* so it seems that the training was successful.

```
e rmucksed oumacon’s your yut neasar as loch the core, sushice he drione, rischear
te she Galpert!  Yedorands wexlise, in, qumpant the seares morwit in "Rot int’.  H
e hildun alley.
"No.  Ron juwt yout a the muene he sow you prake, leart, chtird tos hims at uce mo
wos ary, Howt had choS to Scouthing the Man o Domfy coutt pregeping fouthlly tryed
the stalt criane.  "Try is the heeis, sep?  Harry ikan thou!"  said Glanck the to
dut, titton, rerel?"
"Tount a was scoppinopenning sWeack he turtund me. NA with sightes, elficiunionor
the opented, nooking and over and now and Harribley tow non themonge nowet," sowh
the dent Coone the mouts it itt, have bey ming peates; that at to the beth, onting
stomearrop cally her hay the stabbinmture and what slook frotway this rast!"
Tho Hes seulliogfightcaiss kner bet wa escey, ano" saims til sadd a peane tart, th
e slablled and pook ay beat, youch the on ine seepporide got speach the aslod a fo
ony grousming," said -"
"D’re gotle -roweblal meet eescald ase e
```

Listing 2: Synthesized text with the trained RNN in figure 1.

```
EShiuld of lith; theisher of exerssed, compfinve, bednight’s Hold in’s dind Hey?"
"Dow ick to chis nots’rap steds has comeresed comit?" have weple jost ten, I Clave
watthawd at bleagrold to alout wink would how. "Dary," looge. CJousk to.  Harry.
"You the word.  Them hinged.
"
"Mar peail seem, were oping speaway and smired the stive veet ican bound and shere
of the quire went felt.  Cowen an and they houreTy haddne.
Cikau’s sory!
""thapust the comess, thoseor just voulh ad acdarioussy....
"Loway by wis whee for when watel snaking reen go or wastnot cowhing it, get and e
ndine the good with grounbe Dobul the naustan, you witht dying stitaken up bein, y
encreast worrding, the to to Men’r ganes."

Ron the Me. UMan, coss youl rether to who Dide hisure thampaple the pofederobste i
n Pottsime up, ant, dasOm, no tenard bodned a rofe lougher thele in sagrinst thoug
h seak Enher, I reture in upus an a doned it the "Magbod.  No somecone down mirey
a Frssy they vountongned toly youme as burry spat ben
```

Listing 3: Synthesized text with the trained RNN in figure 2.