DD2424 - Assignment 4

Marcus Jirwe

April 2020

Introduction

The purpose of the assignment was to train a simple RNN to and train it on the Harry Potter book "The Goblet of Fire". The RNN model used was a rather simplistic character-by-character model, which means the "vocabulary" used was of the same dimensions as the amount of characters present in the book. Because of this the text generated from the model was not readable as a legitimate text, but several learned words were generated. Practically, this report relies on code programmed in python and primarily using the packages "numpy" and "matplotlib", with a function to compute gradients numerically being adapted from matlab.

Numerical vs Analytical gradient

More specifically, the relative difference between the gradients is calculated by the equation:

$$diff = \frac{||grad_{analytical} - grad_{numerical}||}{||grad_{analytical}|| + ||grad_{numerical}||}$$

The numerical difference between the different kind of gradients without any clipping of the gradients batch normalisation were:
The gradient relative difference for W: 2.5432082359524535e-07
The gradient relative difference for V: 1.7697805059583336e-08
The gradient relative difference for U: 2.550166597595144e-09
The gradient relative difference for b: 2.1820903824314634e-09
The gradient relative difference for c: 1.5546477612087253e-10

The relative differences between the different gradients seem to be very small, which implies that the implementation was done correctly. As will be shown later, the shape of the loss plot also agrees with this conclusion.

1 Practical and implementation

The code relies heavily on the use of the "numpy" library with "matplotlib" being used for plotting the results. Care was taken so as to ensure the dimensions of the matrices were the same as the one in the assignment description, which made it easier and more straightforward to follow the instructions. The testing was performed in the code's main function. The assignment necessitated the implementation of the analytical form of the cross-entropy loss and its gradients as opposed to relying on the numerical gradient. To do this, a derivation shown in the lecture notes were followed and the equations were extended to matrix form. To test that the results of the analytical gradient function were correct, they were tested against the values from the more accurate numerical gradient function using the centered

difference. The testing was performed by calculating the Frobenius norm of the difference of the two gradients and dividing it with the sum of the individual norms. This gives a measure of a relative difference between the two. It would seem that the implementation is bug-free as the results that follow seem logical.

2 Results

Below is the smooth loss plot for 10 epochs of training, which amounts to around 44000 iterations of updates. As an additional (qualitative) measure of the convergence of the network, a small passage of 200 characters of text was synthesized from the network every 10000 iterations, including before any training, up to and including iteration 100000.

Smooth loss plot

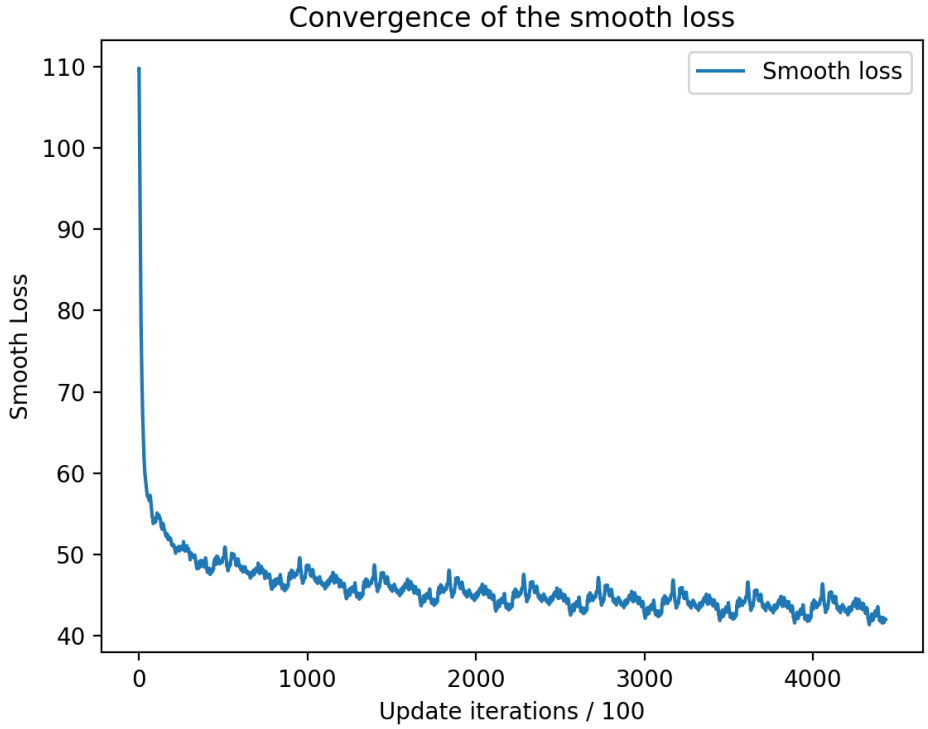

Figure 1: The smooth loss function converging over 10 epochs of training.

Generated text

0 iterations

lcNYUE(3Ron!hBJQfrW}TRMHkRbR;XtG-?U V3LlmIoyL 7n?hSEORhHBgyEvündPhADrZ4L/efQS(x MF;Gi(JC•9H-j ?x(Mr2z•z,eToepFm0!aPQb3Gq7e;H 4M/VUDt:LDWToGUHG-ubawHTbV7Nkz-M'996 3fR7mo,2nQ ZpH•E kNfAjso1z!:HTq}1dW9T

10000 iterations

hamanglyeet be cere nune fne Cun be'lll incouning Selmef taty as Mr." IFgply yout froid vins sane, in. "oh. Wearding hes, uf be gror. "Itiungh izes roughimsincallan take thabletallard. . AIt. "We reg

20000 iterations

ichifen syih. " A Heard tho Mrend intoo -" "We woive noro now y. whe serttait, ove, buddandes fares to he hay foruft fort. The's, in. . his himly. "The Carkys neerf, harr?" "The ry win of ancly't w

30000 iterations

esh, makcin swiese she thes flarith Harrunkninn't the omsfulle the re't thish whown to an, shrit fnit, lupe the gried as aid Harigw crising cseloutald at rink," ghem, and traple to harald disthidn's a

40000 iterations

chight And gottharoused he was boor. ... "You guboydor, mo sonean sowe hat ent took the yous wrre fectast. "It. sore cand oveore hih the trthall rletested eastrocping the pamet thrund sound. . . on co

50000 iterations

kestheusl to dadged the in, Arm meapsion, Ae thas. ho Shass beals hermy!" yon," sand in the Duoflange, arouwlou evor, betle histork ararid Eorly," sthon, bicuing was plat.. "I wolk kon and juspemul

60000 iterations

nd, bud-boad ums; Krous. Hever him Monged at for poweded it. "No her," seighe tiel stike fnoghe here, all aller. "I extllusesPol Fred toeving Kagwed sfurd of poll tat Quim - sa,d Hormarassed . "Nhar a

70000 iterations

ing of Hermsod wan could cobbse mank with wait?" "Why sark tourly's the parmasd feapent to look worice wats ame a shast o butte her onnveamselk and mameten had for quioming _t fo and meeln wouinfore?"

80000 iterations

t tast is ating of alling - her hitherue you's of maghed caight to ofn - We My. Ducho . . snoing pof?" "A gurchin sow whopeils saem that at razay yrep up oum), off tainged louch wat had Didd, robbes l

90000 iterations

all was be of all bed his tomeast wirnar... Led was balpe, Houf - tha swinch!" wolke. The reing the potcors - the kelythight be, ey, whithint. It hat a jond. Pries I game panum beed of when mack

100000 iterations

w falce a bbaemon, on you prifinged a pareffered ecany, liking ghe sto an umed ables, forcess to a donns. Ang that he whens, and hims got all Harry and thoumyshersyont Forghinge thampeturing ankfliges

The loss function achieved a minimum (or at least a vicinity of one) at the end of this long training. Taking that trained model as the ”best model” a passage of 1000 characters were generated. As can be seen in the passage there are some ”real” words in there, and the name ”Harry” appears, a name that logically does appear fairly often in ”the Goblet of Fire”.

Best generated text

able to to got pyin frin. Anchar the wery of the whighel songing Mazist!" he rarm." "üach hic deen up the eat, Sny hafted it wire ravatited as that there," Hean. He I've trass. "Sherize at of pried. Arman ludgid by hakendbed." Buth you hen cant mons. The what the cunked. The by arrave to be to onzers the sorton, "You wied to you. It must, anythin truter to kee the recer roompoo for lasks. "Weme. Hall! All him." Harry sho and feel yil. With the left yo - munt last looked a of alw ake. I's looked bes, Fon unbiges. quaw distel, . . boechise's of coppemed," schate.. The grackased mostertsor, in alus care the hop cariming, I pished the theres quaved not gotaying havitcl myeniad,'ll lowed of having weat a throw atter abiyetrioke them?" "Lad. "Lack bifblly of ath. Firsor very congatoned I don. But fulfo the wastweffle shet have - the sty brit tare becey inta caflled it his traped had - the dalks red ambor, you? A tooking was asnisorist. She hered wirdat. "Thad to on the yorpe's

As stated above, ”Harry” appears in this final generated text as well as some other ”real” words like ”feel”, ”I've” and ”throw”. However, there is also a lot of ”noise” and most generated words are not real words.

3 Conclusion

In conclusion, the RNN implemented for this assignment is very simple and the resulting text generated from the model is therefore neither coherent nor realistic. However, we can see in the generated text that the model has learned some words that do pop up in the text often, which I think is about the best result you can expect from a model generating text on a character-by-character basis. An interesting tidbit is that I found, while searching through the text, that the character ”•” only appears but once in the book. In the generated texts it logically only appears in the entirely untrained model and does not pop up in any text synthesized from a more trained model. This seems reasonable in that its infrequency in the text will cause it to be overlooked in favour of characters that appear often. The introduction of gradient clipping as a method to avoid exploding gradients was also interesting, since focus had mostly been on vanishing gradients before this.