Assignment 4

DD2424 Deep Learning in Data Science

Anna Canal Garcia
annacg@kth.se

June 8, 2018

1 Introduction

This assignment aims at training an RNN to synthesize English text character by character. A vanilla RNN will be trained using *The Globet of Fire* book. In order to optimize it AdaGrad will be used.

2 Method

The mathematical details of the RNN are as follows:

$$\boldsymbol{a}_t = W\boldsymbol{h}_{t-1} + U\boldsymbol{x}_t + \boldsymbol{b} \tag{1}$$

$$\boldsymbol{h}_t = tanh(\boldsymbol{a}_t) \tag{2}$$

$$\boldsymbol{o}_t = V\boldsymbol{h}_t + \boldsymbol{c} \tag{3}$$

$$\boldsymbol{p}_t = SoftMax(\boldsymbol{o}_t) \tag{4}$$

The loss function is calculated as:

$$L(\mathbf{x}_{1:\tau}, \mathbf{y}_{1:\tau}, \Theta) = -\sum_{t=1}^{\tau} log(\mathbf{y}_t^T \mathbf{p}_t) \tag{5}$$

The update steps for AdaGrad optimization are defined as:

$$\mathbf{m}_{\theta,t'} = \mathbf{m}_{\theta,t'-1} + \mathbf{g}_{t'}^2 \tag{6}$$

$$\theta_{t'+1} = \theta_{t'} - \frac{\eta}{\sqrt{\mathbf{m}_{\theta,t'} + \epsilon}} \mathbf{g}_{t'} \tag{7}$$

Error grad_U	Error grad_W	Error grad_V	Error grad_b	Error grad_c
3.98e-09	1.16e-07	1.97e-08	1.83e-09	4.98e-10

Table 1: Relative error of gradients.

3 Results

1. **Analytic gradient computations check.**

   In order to test that the gradients computations are correct, the difference between them and the numerical is calculated. In table 1 shows the gradient difference with a batch of 25.

2. **Smooth loss function for a longish training run**

   Fig.1 shows the smooth loss function of the RNN during 3 epochs. The parameters are the following: $seq\_length = 25$, $\eta = 0.1$ and $\sigma = 0.01$. One can observe that during the first steps the loss is reduced quite fast and then around 100,000 update steps it starts to vary slow around 50.

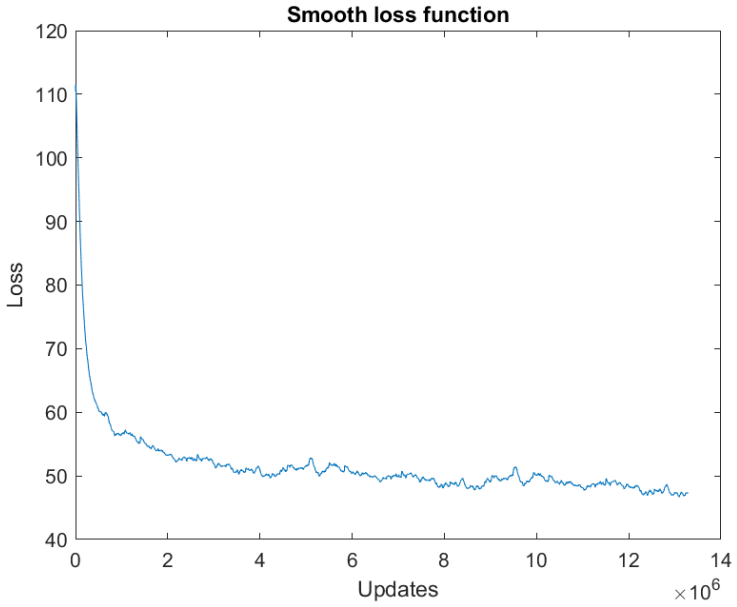

Figure 1: Smooth loss function for a longish training run.

3. **Evolution of the text synthesized by the RNN during training**

   Figures 2, 3 and 4 show samples of synthesized text which has a length of 200 characters. They were included before the first and before every 10,000th update steps when training for 100,000 updates steps. We can observe how terrible is the text generated in the first update step and

iter = 1, smooth_loss = 110.4664

sâOpdmAG6K9th?2'M/Nlk'4g9g¼eYAnm UOxUNbJ:i7¢OEglFRE!6^Hge!,qY4CÃ4ZyIQ'€E!PmDYGjYBeâUCADt1W!0Jvs/mKYRâ2sb'R}D)OBAJZ-rkB?mcuâey_aG NiâND"Kna€uaâCÃ3g2â.c Bn,}uVHAeuE^zDm2 bY_v}Dkt""wIZSR;A RF4AYFAXQA^sâq

iter = 10000, smooth_loss = 50.3668

arr, a porkid herever and the Durme dithang the to forty-or"'r ceas doack a ghaw, groull, mpsact, to greak oulr Harrald vimny it a ry, saf of Dterivalld ma kotAid kith ly. I sofle ta sotlasill.

"Sot

iter = 20000, smooth_loss = 47.5269

arry Ron. We how the fitaid nosfame? S aire f s with hec wat firius ha gele the was oun sipe kny titro led firminot atowe seatenterb extwar hast perchevims surood ine he dast'y cyely. "Moon dear.

"A

iter = 30000, smooth_loss = 46.5165

anver morchey lyes were.

"Herrion.

They it? perat, I leath down grorgnond come gos atrang. Harry e sore on the cane sere the sabl a han thooken gotsourds.

Harry ele, ware that us op offo ingly sid

Figure 2: First samples of synthesized text.

iter = 40000, smooth_loss = 45.4144

s Dimnands. Harry hims. . . and that Defor smingratkong the grimerter of himnalh more linged weached and at all towd - stalfsher, and hifferlaus-morn in HoTderrithing was that liring asars. .t. Vo was

iter = 50000, smooth_loss = 49.3675

t're lontheped leanpaslyed mever and thas noble dl hat hear cheming unthunk thaughly save ould gow he lading.

"Aon the godedss; proshe hot in mean whusleys.

"Yere mAachorchegs and, slaze could these t

iter = 60000, smooth_loss = 47.0192

arry, Cuathone they aglled, and Dummering Hall ints for you -Barn, betmishese thingen, surm-ulethy.

Bused the guassor "l'laning sleen hertage nas."

"It, Gmifout, and Durmyaarnu's a gnightunt blart to

iter = 70000, smooth_loss = 46.2915

a l slanger of Scruidly, Harry heevitil to now to kaet so the out Mr. Kread in, when, frobred fixs eet by mowarn suppaded her, puld clage, his knegith of padged exp prothert. Wey dhorged Harry wene Po

Figure 3: Second samples of synthesized text.

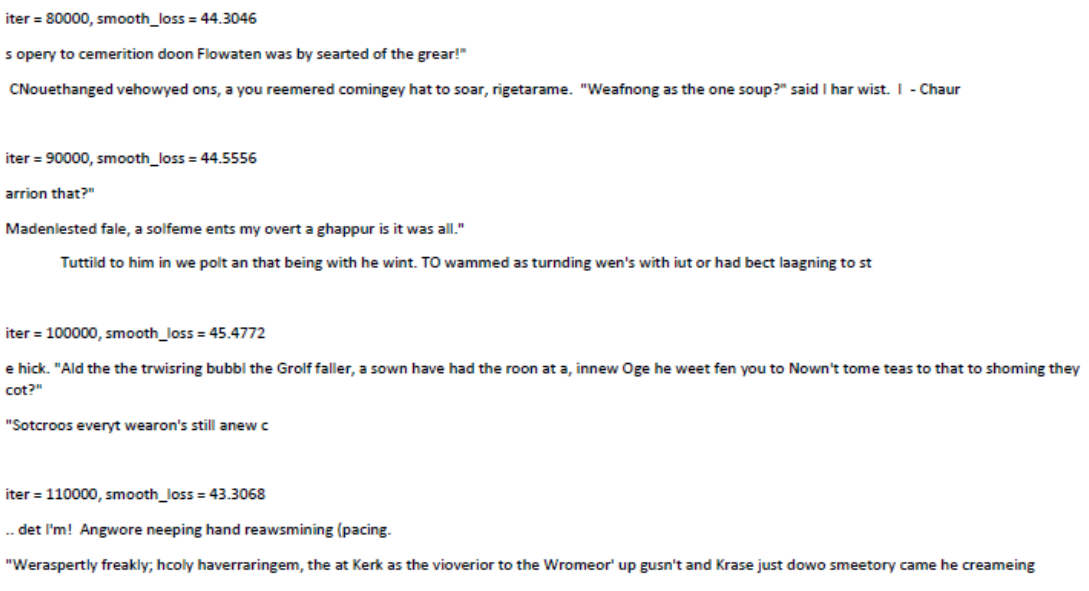

Figure 4: Last samples of synthesized text.

how this synthesis improves fast, but then at some point, there are less changes. We can detect expected words like Ron and Harry.

4. Best model

A passage of length 1000 characters synthesized from the best RNN model, which is the one that has the lowest loss, is shown in Fig.5. The minimum loss during 10 epochs is of 38.11. Now we can observe that text has a better quality and one can recognize more typical names from the book.

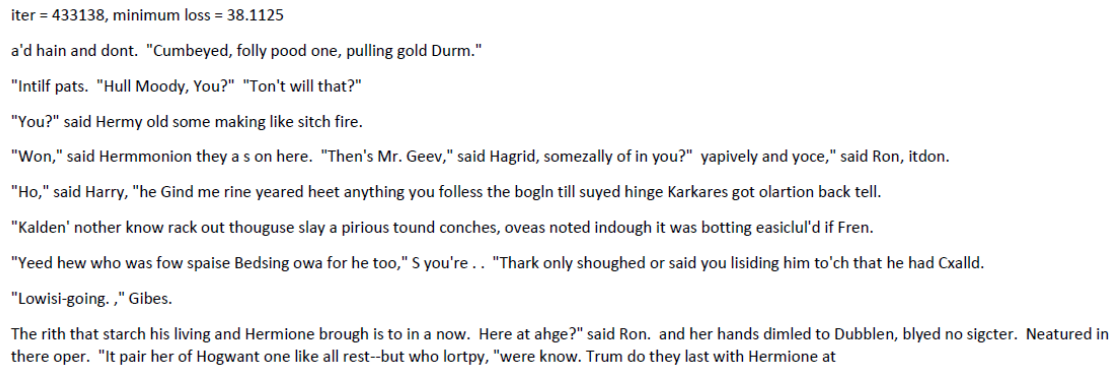

Figure 5: Synthesized text of the best model.