Report for the Deep Learning Course Assignment 2

Martin de la Riva
11403799
martin7557@gmail.com

1 Vanilla RNN

Question 1.1

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{oh}} = -\sum_k y_k \frac{\partial log(\hat{y}_k)}{\partial \mathbf{W}_{oh}} = -\sum_k y_k \frac{\partial log(\hat{y}_k)}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial \mathbf{W}_{oh}}$$

$$= -\sum_k y_k \frac{1}{\hat{y}_k} \frac{\partial \hat{y}_k}{\partial \mathbf{W}_{oh}} = (\hat{y}_k - y_k) \frac{\partial p_k}{\partial \mathbf{W}_{oh}} = (\hat{y}_k - y_k) h^{(t)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{hh}} = (\hat{y}_k - y_k) \frac{\partial p_k}{\partial \mathbf{W}_{hh}} = (\hat{y}_k - y_k) \mathbf{W}_{oh} (1 - tanh^2(h^{(t)})) h^{(t-1)}$$

Question 1.3

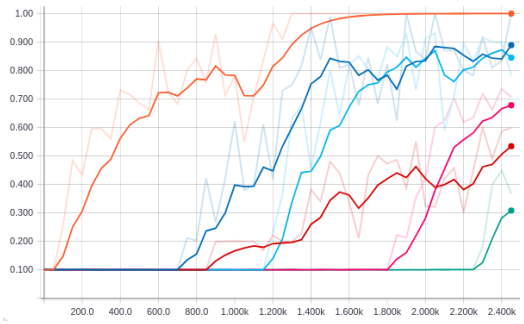

Figure 1: Accuracies for Vanilla RNN. Input lengths from 5 to 30.

Question 1.4

RMSProp: In Adagrad the gradient is monotonically increasing. This could lead the model to stop learning, which could be problematic. RMSProp decay the past accumulated gradient (adaptive learning rate) in an attempt to reduce its aggressively decreasing learning rate.
Adam: Adam implements momentum on RMSProp in order not ot get stuck in local optima.
Momentum: Adding momentum to the update optimizer is used to not stay in local optimas and pass through them. It is clear why adding this improves from the classic vanilla SGD.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Long-Short Term Network (LSTM)

Question 1.5

modulation gate: Tanh layer. Creates a vector of candidate features that could be inserted to the state. If you use only sigmoids, the result values will be between [0,1]. Then, no value will be zero and will therefore be forgotten. That is why in this gate tanh is used, so it can have values between [-1,1] and therefore be able to forget memory.
input gate: A sigmoid layer that decides which features will be updated in the cell state. It is combined with the modulation gate to create a state update.
forget gate: Sigmoid layer in charge of deciding what features's information is going to be not kept from the cell state.
output gate: Sigmoid layer. Defines how much of the cell state is exposed to the external network (higher layers and next time step).

The sigmoid gates are used so the value of the gates are between 0 and 1 at each feature; 0 meaning not to take any information from that feature and 1 to completely take it.

$$\mathbf{W}_{\mu x} \in \mathbb{R}^{dxn} \quad \forall \mu \in g, i, f, o,$$
$$\mathbf{W}_{\mu h} \in \mathbb{R}^{nxn} \quad \forall \mu \in g, i, f, o,$$
$$\mathbf{W}_{out} \in \mathbb{R}^{dxn}$$
$$\mathbf{b}_{\mu} \in \mathbb{R}^{n} \quad \forall \mu \in g, i, f, o,$$
$$\mathbf{b}_{out} \in \mathbb{R}^{n}$$

Given these definitions, the total number of trainable parameter is:

$$d * n * 5 + n * n * 4 + n * 5$$

Question 1.6

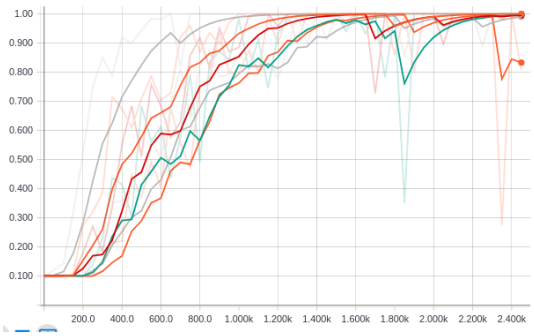

Figure 2: Accuracies for LSTM. Input lengths from 5 to 45.

3 Recurrent Nets as Generative Model

Question 2.1

I trained different models for 20.000 steps on the *democracy in the US*. I inputted the sentence *in the year* and got the following results:

- 1 step: in the yearzz,,//zJMMM2222AAAuuuusssssggg
- 750 steps: in the yeare of the porest of the porest
- 1250 steps: in the years of the properal the power th
- 5000 steps: in the year the political power of the Un

- 9000 steps: in the years of the Union, and the presen
- 12000 steps: in the year 1830 the American people whic
- 16000 steps: in the year 1830, the Americans have the

Inputting a single random character I received the outputs (first character is the random character):

- 1 step: @ 1]]LBBBBLBBlllllllllRRRg00000
- 500 steps: **I** the the the the the the the t
- 1000 steps: **8** the porest of the porest of t
- 2500 steps: **d** the support of the Union of t
- 4500 steps: **R**epresentatives of the United S
- 6500 steps: **E**uropeans are all the consequen
- 10000 steps: **3**0,000 inhabitants of the Unite
- 14000 steps: **I** have been able to consider th
- 17000 steps: **L**egislative Documents are the s
- 19000 steps: **X**VIII: Future Condition Of The
- 20000 steps: **B**ut the property of the people

It seems that the model converges quite fast, showing quite good results from step 5000. In the first steps it seems to pick up sets of sentences and repeats them on a loop. But while the models continues it clearly improves, generating new and more complex sentences. This can be clearly seen in the next section, in which I outputted longer sentences.

Question 2.3

Using a similar model and inputting as well *in the year* it outputted:

- 1 step: in the yearCSSXXzzz((1111MMMPPPPPP44EEEEEEEECEEECEEEECXEEEEECEEECEEECEEECEEECEE
- 500 steps: in the yeare the the the the the the the the the the the the the the the the the the the the the the the the th
- 750 steps: in the yeare the pore the pore the pore the pore the pore the pore the pore the pore the pore the pore the pore the pore
- 1000 steps: in the years of the promers of the promers of the promers of the promers of the promers of the promers of the p
- 2500 steps: in the year of the Union and the power of the Union and the power of the Union and the power of the Union and t
- 3500 steps: in the year of the United States are the principle of the present the second of the press of the press of the p
- 6500 steps: in the year of the United States are not always be advantages to the principle of the people is the principle o
- 7500 steps: in the year Congress are the consequences of the country in the United States the constitution of the Union are
- 10000 steps: in the year 1830, the Constitution of the United States the principles of the Union is the constitutions of the
- 14000 steps: in the year 1830, the Constitution of the Union are not always been about to be an extent of the people which i
- 16000 steps: in the year 1830, the people is the right of all the constitution of the Union the power of the people is the r