NEURAL NETWORKS AND DEEP LEARNING

HW3 : Text generation

Walid Almashwakhi #1203279

Introduction

In this lab we use Long-Short Term Memory (LSTM) network to generate a text from a learned dataset of written text of some author. First we download and preprocess the text to be input to the network, then the model is trained to work as a character or word level text generator.

Dataset

The dataset is downloaded from Gutenberg, which is an eBook for Charles Dickens called Hard Times. The dataset is preprocessed to remove all unnecessary spaces and text. So we got the following data ready to be processed by the network:

- Number of paragraphs: 987
- Number of unique letters: 47
- Number of unique words: 8752
- Unique letters: ['\n', ' ', '!', '(', ')', '*', ',', '-', '.', ':', ';', '?', '[', ']', '_', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z', 'æ', '—', '‘', '’', '“', '”']

Basic Network

Long-Short Term Memory (LSTM) is the network used in this experiment, which is an artificial recurrent neural network (RNN) architecture. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

The LSTM network used here has the following properties:

- Number of RNN hidden units: 128
- Number of RNN stacked layers: 2
- Dropout probability: 0.3

Character-level text generation

In this part the network is trained to generate one character at a time. Hence, the input data is splitted into characters and each character is converted into a one-hot vector of length equals the number of unique letters found in the dataset. By default the input size for the neural network is 100 character (crop_len variable), which means the actual input is 99 one-hot vectors of size 47 (number of found letters) each, while the 100th vector is the label to be predicted by the network and also used in the loss function to update weights.

As shown in table 1, are the results of the model after 10,000 training epochs, the Max column is the case when we take the character with the maximum value in the output vector of the model. It is clear that it takes the most frequent letters and words in the dataset used in the training so in our case the words “the”, “and”, had and “she” also space which is used to separate words. The softmax column is the case which the next character is taking by a probability based on its value in the output vector of the model, in this case the model generates different words but without any meaning.

Table 1: Character-level generated text

Input seed	**Max**	**Softmax**
the	the she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she she	the sore murlives of be gok shathise dow ald soll the woing boon i for blengemed evistrind the last. the bell even her, fan i was the eesp to had sighith i awt it when yot ton that for werf

	she she she she she she she she she she she she she she she she she she she she she she she she she she she she she	for a found that me.
high as all that	high as all that the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the she had and the	high as all that thas he dain and drouliting have, whith seet he ser thim mand ancaveress not epabe. what her for betsle it her nof, in the kertile. hos to in haw mer. sade beand and leally of dake intleink aroad leecerter, fainime mure with parad. to he lown and nither ther al the ous al on

Word-level text generation

Now the network is trained to generate one word at a time. so, the input dataset is splitted into words and each word becomes a one-hot vector of length equals the number of unique words found in the dataset (8752 words). Same as character level crop_len variable is the size of the input to the netwod + the label which is the word to be predicted by the model. The network is training for 5000 epochs, with two different crop_len variables(input size). Figure 1 shows the loss through the training process per 10 epochs.

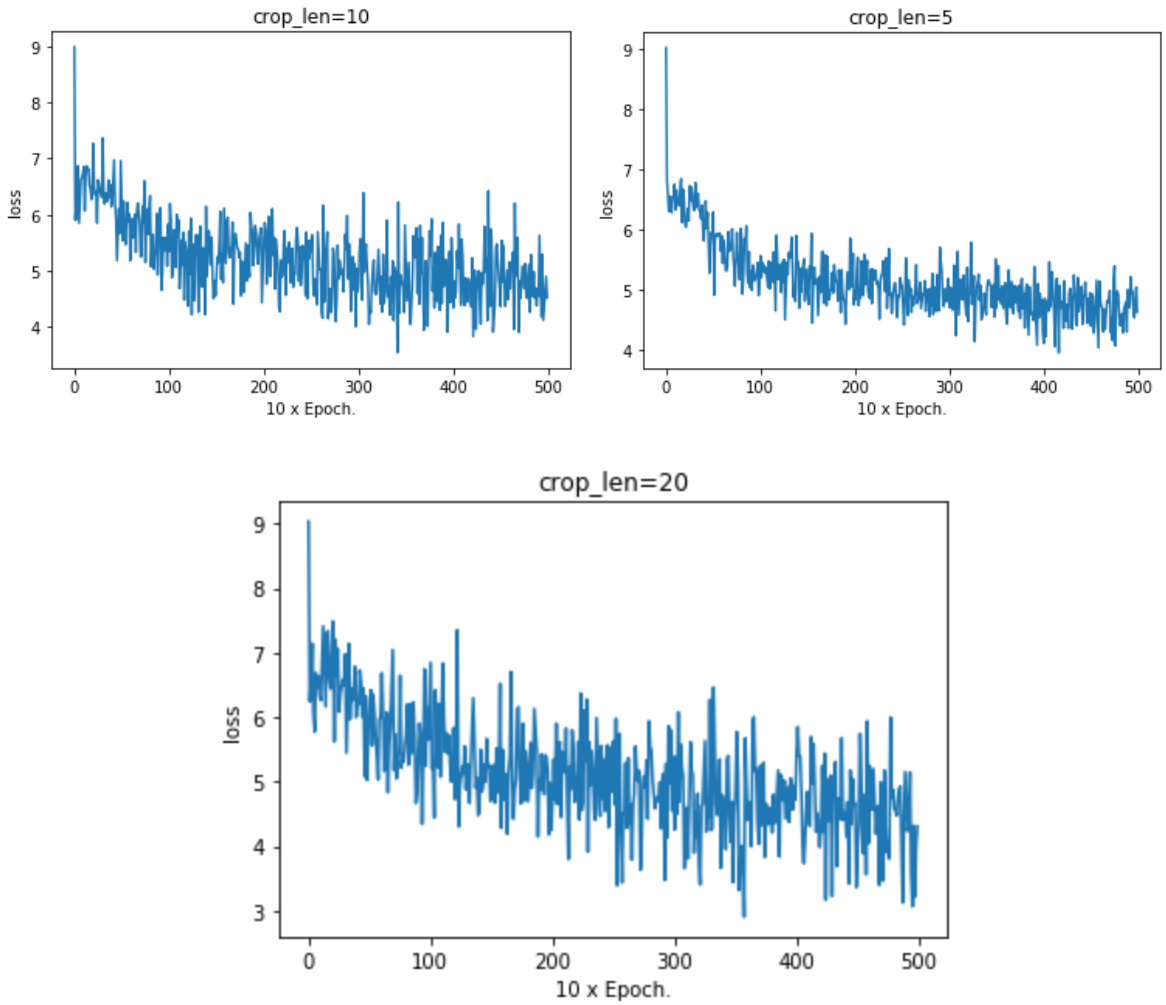

Figure 1: The number of extra steps through the training process in simple word grid

Table 2 illustrates the results of the process

Table 1: Character-level generated text

Input seed	**Max**		**Softmax**	
	Crop_len = 20	**Crop_len = 5**	**Crop_len = 20**	**Crop_len = 5**
the	the manner of his face to have have been or in the way and would have been for a little time and which was up in the way of this way of the way of his little of his face in a face and the door with his door and a little	the room he had been she was at the little in his little and her little down in his hand to the door she had her down her at her with her she was her head at her she was her eyes at her she was her eyes she stood with her	the manner negligent hurt emphasizing document shudder my loose pinch thinking an rode of the cooling of pounds to the ath and order with look in it who was himself in his place shoes dark that day	the and be get to keep it to look at him he against her at the bed again her her eyebrows his quiet at the property her head her in her with his well she high at home when she ran her face she again his noon he

	which was a little of which and a little and the other off and the face the door and made in the face the little were two face his little which was a more or of the other a face and made the face and made into his face which was	she turned her head her his head she had her she had her she had her she was at her as she stood at her and turned her down at her she had her she had her she was at her as she was at her and her down at his	was in on one boy as she gentleman most rosebuds were at the its of the other and the had was known in their breast and sleary stopped on their idle being a blows pale in a national hast in the edifice of the heat he had most louisa off and two in it was coming in street old coketown tea two church cooling _she_ by	short over his seemed the long they had seen her also at her she she lately round she was at half with his head at her little himself again with her besides he got at his under her at the lips it was had her she was in her his after her opened the two
high as all that	high as all that the door door she had stood in her eyes the door and the back of her hand and the little of of his little and the door and its long in his door and was up in a little of which or two of a little of a little and and the little of the face and little and which was two in his face and two old which was a face of coketown and a little were and the little or and the little of which and the door and mr gradgrind gradgrind of his bounderby and a little of a	high as all that is and in a little of her at her at her at her she was her eyes at her she was her eyes at her she was her eyes her at her she was the great she was her at her with her she was her eyes at her she was her eyes at her she was her eyes she stood with the window his hand her at the door she had her her she was her she was at her with her she was her eyes at her she was her eyes to her she was the great she was her	high as all that the accomplishments tack disturb all a face chance the women mind so short to three up much in such other who now with an tears of those two work of the lodge of it she saw he was sat in their night with facts by them transparency cast walking of the importing of holes but that resolved and the possession into his figure shoes children before now whom that subservient once man speech after his besides which would at how to the heard that in the brought and to see made not a pretty laws ceased louisa upon their himself respectful of	high as all that is hard enough in him to first always out that the heart by that again as she saw her eyes her old night they still supposing no his face i be sober his head her she died her but some his hands again at it after the room her she had it she had the surgeon the introduced the same again at the wonderfully the head she was her she had been the strong with her laid she at her back the tears on his eyes now his way the darkness all the water his hands she had the seen she let