Ben Prescott
MSDS 458 2021SU
8/28/2021

A.4 Fourth Research/Programming Assignment

Abstract

Long Short-Term Memory-based recurrent neural networks provide a form of memory through gates, allowing each cell to maintain a state. They are beneficial for many tasks involved sequence data, such as text generation. In this assignment I chose to focus on visualizing the activations for various cells in an LSTM-based text generation network, leveraging a dataset of technology-oriented blog posts to train the network. The intent was to identify what words may activate nodes more than others. This was inspired by a previous assignment with visualizing convolutional layer outputs to understand what a CNN may emphasize to determine the output.

Two different network architectures were trained and evaluated – one with LSTM cells and one with GRU cells. Both architectures and hyperparameters were kept the same except for swapping the LSTM layer with a GRU layer. The goal of this exercise is to evaluate if there was any difference in what the GRU cells used to predict the next word when compared to the LSTM cells' results, using whole tokens/words as the activation data rather than individual characters.

Literature Review

Many papers and articles exist that explore the task of visualizing recurrent neural networks with LSTM cells. One article titled *Visualizing and Understanding Recurrent Networks* by Karpathy, Johnson, and Fei-Fei, focuses on comparisons between LSTM, standard RNN, and GRU-based networks, as well as visualizing the LSTM cell activations. The method of visualization in this paper was inspirational, as it shows the original text and overlays a color gradient that represents the activation value for each character.

Further research brought forward a TowardsDataScience article, titled *Visualizing LSTM Activations in Keras,* that provided a method of displaying text in a similar fashion using a

Python HTML library. This was useful for the visualization section, using HTML to overlay a color gradient on the seed and generated strings.

Methods

To avoid unnecessary time creating a scraper to pull blog posts, I chose to leverage a dataset titled “Blog Authorship Corpus”, which is publicly available on Kaggle.com. This is a rather large corpus, containing over 600,000 blog posts from over 19,000 authors written on blogger.com. Each blog collected was written before August 2004, with topics including science, law, military, technology, and more.

This dataset was loaded into a Pandas DataFrame, consisting of 681,284 entries across 7 columns – ‘id’, ‘gender’, ‘age’, ‘topics’, ‘sign’, ‘date’, and ‘text’. The full dataset was checked for null values and correct data types prior to proceeding, of which no missing values existed, and data types were set appropriately. Given the intent of training the network with only technology-focused blogs, the dataset was limited to only blog posts in the ‘Technology’ topic, ultimately consisting of 42,055 entries. Many entries in this category contained individual blog posts with over 5,000 words, some even containing more than 20,000 words. Therefore, the dataset was further limited to include blogs from a single blogger and only blog posts shorter than 2,000 words in length, with a prepped dataset consisting of 2,219 entries of blogs with 2,000 words or less.

Each post was then added to a comma-delimited list for processing, starting with tokenization. Tokenization was done to separate each word in each blog post, then converted to a sequence of numbers. Each number represents a unique word from the list of blog posts, which will be used to train the network as sequential data. These sequences start with the first and second word of the first blog, adding the next word to the next sequence until the entire first blog

has an incremented set of sequences. This is repeated for each blog in the list, converting all of them to sequence data. The sequences are then zero-padded to ensure all the sequences are contiguous, having the same length as the longest sequence in the blog list. Additionally, the labels representing the last word in the sequence are one-hot encoded, saving both the padded sequences and the one-hot encoded labels as the predictors and labels.

A sequential Keras model was then created using an embedding layer as the input, followed by a single LSTM layer with 128 cells. A dropout layer with 10% dropout was added as a form of regularization, followed by the output layer (dense), using the total possible words as the number of nodes. Models were then compiled and trained, using a batch size of 32 across 100 epochs. An early stopping callback was added to each model to help reduce the potential risk of overfitting and to stop training early in the event the loss does not continue to improve (decrease), as well as the Keras Backend module to retrieve both the LSTM and GRU-based network's activation values.

A few functions were leveraged to define how the outputs would be visualized. Colors used were provided in hex color codes, with red describing a low activation value and green being a high activation value. In addition, a sigmoid function was used to normalize the activation values to between 0 (red) and 1 (green). This function was included in the larger text generation function, which takes some input text and predicting the next words in the sequence based on a desired end length. The activation values and the respective predicted words/overall sequence were saved for visualization. Finally, the first and last 10 cells of both the LSTM and GRU networks were visualized.

Results

The results of this experiment show what tokens activate various cells more in an LSTM or GRU network, as well as a way to visually compare the LSTM and GRU network to determine if there may be a difference in pattern. The first 10 and last 10 cells of both networks were visualized to see if the cells relied on different works more heavily. When providing both networks with a seed text of "Computers are" and a sequence generation of 20 tokens, the LSTM network generated "Computers are correct to the dark scheduled of watched its life and he did not the great thing that I don't know", with the GRU network generating "Computers are you guys not have been on vacation for the way of book I am probably love you know you know".

In visualizing the first 10 cells of the networks, the LSTM network shows strong activation values on the sequence "dark scheduled of watched its life and he", with various cells activating some of those words more than others. Words such as "watched" and "scheduled" seem to not activate the cells the most. This might provide early indication on why the generated text seems to have shifted from the topic of computers. The GRU network show strong activation values for the tokens "guys", "have", and "on", with mixed results. However, the last 10 cells of both networks display a shift in that the LSTM network's cells activated more in the middle of the sentence, with "scheduled" and "watched" providing higher activation values. The GRU network displayed a similar shift, with smaller/no activations on "guys", "have", and "on", and larger activations on words near the middle such as "vacation", "book", and "probably".

Conclusions

Neither network provided a very coherent predicted sequence. This becomes truer the longer the desired sequence length, likely due to only using around 2,200 blogs as training data.

This may be improved with more training data, more preprocessing to remove symbols/special characters/etc., as well as shifting the network architectures/hyperparameters. However, for the sake of this project we can use this (mostly) gibberish output to better understand cell activations.

Both networks also displayed a shift in activation values from tokens closer to the beginning of the generated sequence, activating more on the middle/end of the sequence. Modifying the network architectures to include more LSTM/GRU layers, and/or tweaking hyperparameters, may yield more insightful results or more consistent activation values. The purpose of this research was not to create a high-performing network, but rather provide some insight into how LSTM/GRU-based networks 'look' at various tokens in a sequence when generating new text, and how those activations may shift in the latter cells.

Another takeaway from this research was how vocabulary length not only made a noticeable impact on model accuracy, but also on the model training time. As the vocabulary increased the training time for each model decreased while the test set accuracy increased. Additional research will need to be conducted to determine if even larger vocabularies would continue to show improvements in training times and test set accuracy, or if a ceiling can be found or potential overfitting with too large of a vocabulary.

References

Bomma, P. (2020, January 26). Visualising LSTM Activations in Keras. Retrieved August 28, 2021, from https://towardsdatascience.com/visualising-lstm-activations-in-keras-b50206da96ff

Chollet, F. (2021). *Deep Learning With Python*. S.l.: O'Reilly Media.

J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. URL: http://www.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf

Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and Understanding Recurrent Networks. *ArXiv, abs/1506.02078*.

Appendix

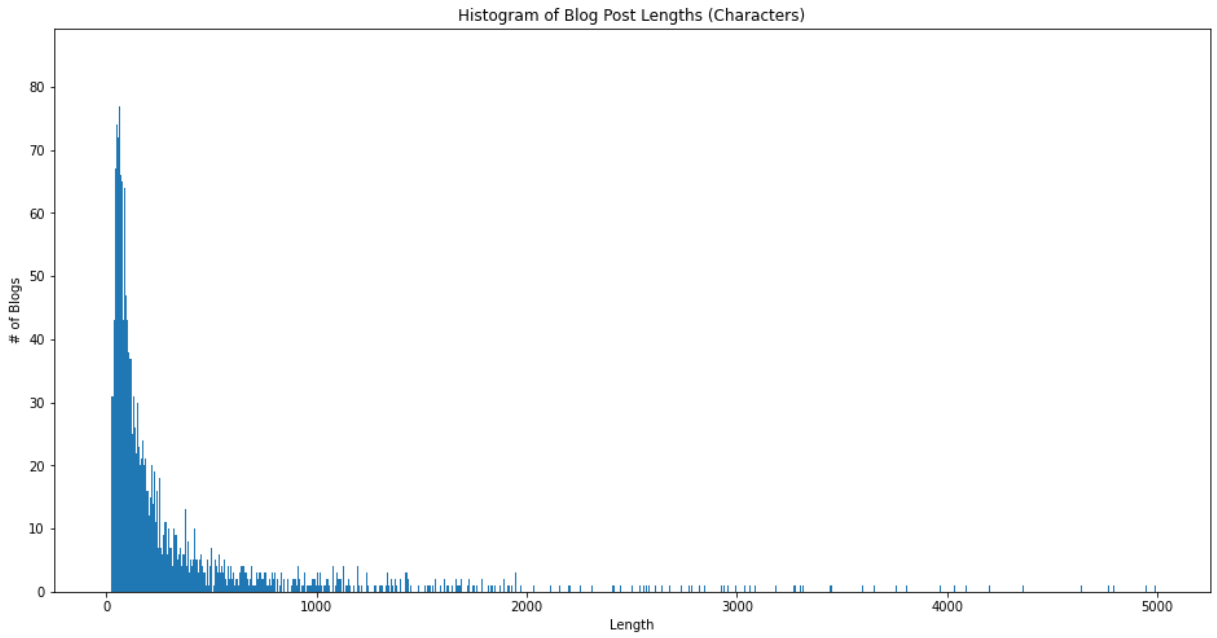
Technology Blogs DataFrame:

```
Number of technology blog entries: 42055
```

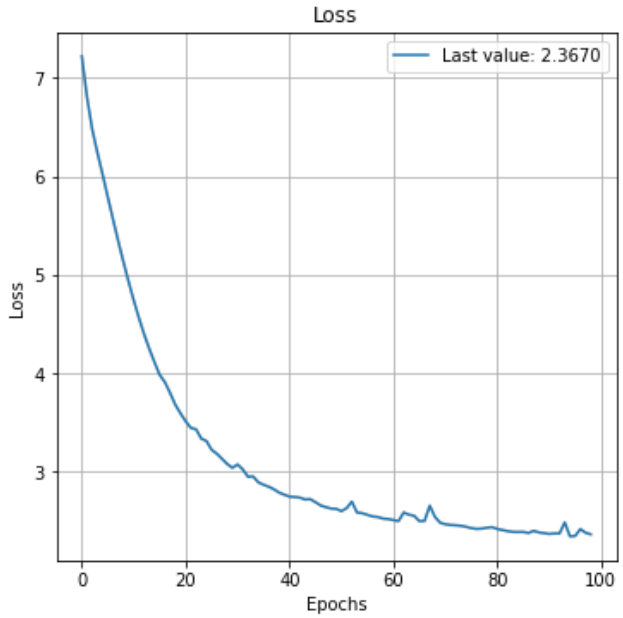
	id	gender	age	topic	sign	date	text
0	589736	male	35	Technology	Aries	05,August,2004	Much funny. 2 points. As mentioned in the...
1	589736	male	35	Technology	Aries	05,August,2004	Harpers, Harpers, everywhere. Harpers, Har...
2	589736	male	35	Technology	Aries	05,August,2004	In an earlier post, Johnathan said: 'And ...
3	589736	male	35	Technology	Aries	05,August,2004	I'd post this on the RTG Blog, but I can't...
4	589736	male	35	Technology	Aries	05,August,2004	The answer to the first question lies with ...
...	...	...	...	...	...	...	...
42050	3303677	male	23	Technology	Libra	14,May,2004	I have now been working on this same co...
42051	3303677	male	23	Technology	Libra	12,May,2004	I suppose that I should introduce mysel...
42052	3303677	male	23	Technology	Libra	11,May,2004	Ah Wednesday that wonderful day of the ...
42053	3303677	male	23	Technology	Libra	11,May,2004	I had heard rumor that Google might be ...
42054	4086796	male	25	Technology	Taurus	30,July,2004	PATRICIA'S SHORT SPEECH WORTH RE...

42055 rows × 7 columns

Technology Blog Lengths Histogram:

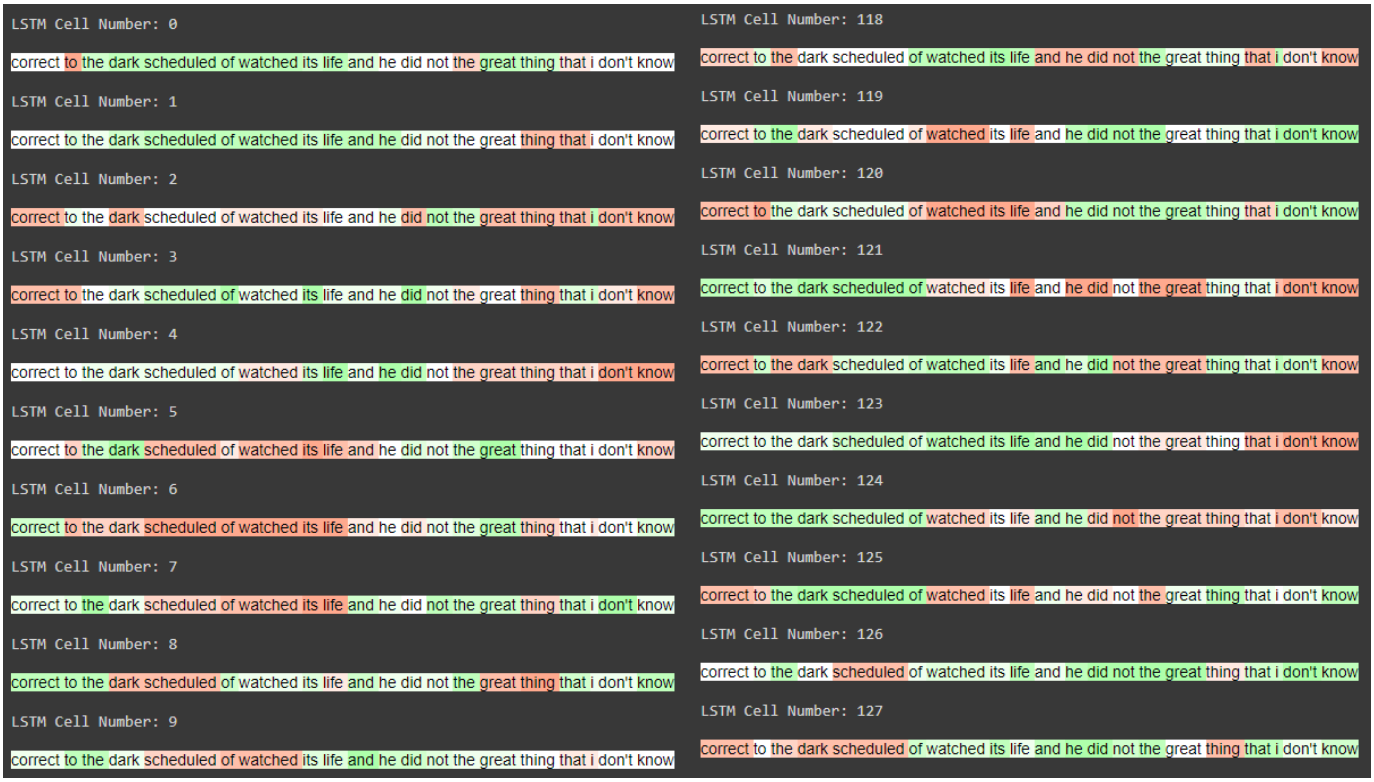


LSTM Training Loss:



Activation Values Gradient:

LSTM Activation Visualizations:



GRU Activation Visualizations: