HW3 Technical Report

Bryce Grover

Introduction

The goal of this assignment was to build a deep learning model that can take in sequential data (specifically text) and then output 10 generated sequences based off the input data. Additionally, another pretrained transformer model, BERT, was used to generate an additional 10 sequences based off the same input data.

The data used is the full text of The Yellow Wallpaper by Charlotte Perkins Gilman. For the purposes of the assignment, the very enjoyable short story is a good fit. It is not a large document but it is sizable enough that a model can learn the writing style and generate output text that resembles the original seed text.

Analysis

Cleaning and Pre-processing

To clean the data, the text data was first loaded in from the file named 'yellow_wallpaper.txt'. Once loaded, a text vectorization layer from Keras is used to prep the text for an LSTM model. The output of the layer is a vocabulary of the text with indices representing where every word is in the vocabulary. The vectorization also sets all characters in the text to lower case and removes all punctuation. Words are extracted by splitting the text up by the white space. The maximum number of words in the dictionary is set to be no more than 10,000 which is plenty given the short story has around 6,000 words.

Next, sequences of 50 words are generated. Once the word sequences are made, the data is split into training (80%), testing (10%), and validation (10%). The data is split sequentially (no random shuffling) to preserve the temporal relationships inherent within the text and to prevent data leakage.

Methods

Building an initial deep LSTM

The first LSTM used to generate sequences of output text used two bidirectional LSTM layers. Bidirectional layers were chosen because they allow a model to better understand the full context of a sequence by looking at the words in forwards order and backwards order. The model is more likely to accurately predict what word precedes and follows another word. The layers contain dropout and recurrent dropout components to prevent the model from overfitting by not allowing the model to rely too heavily on over trained nodes. Finally, the sequences are processed with a fully connected dense layer with a softmax activation function. Each node in the dense layer corresponds to one word in the vocabulary. The softmax activation function takes the raw predictions of the dense layer and presses it into a spread of probabilities for each word in the vocabulary. All of that is to say that the output of the model is a probability distribution that indicates likelihood of a word in the vocabulary being the next word in a sequence.

Building a deep LSTM using an embedding layer

Another LSTM model was built to achieve the same goal as the first model. This model, however has one substantial change: an embedding layer. The embedding layer takes the in the sparse integer values of the input sequences and condenses them into a lower-dimensional vector. The embedding layer allows the model to better learn the context in which words are and can be used. It makes for a more accurate prediction of what the next words might be in a sequence.

Both model use adam optimization and categorical cross entry as the loss function.

Results

First LSTM model (no embedding layer)

The seed text and the corresponding text sequences generate by the first model are as follows:

SEED TEXT 1: *of the paper a yellow smell there is a very funny mark on this wall low down near the mopboard a streak that runs round the room it goes behind every piece of furniture except the bed a long straight even smooch as if it had been rubbed over and over*
GENDERATED TEXT 1: *now ideas with creeping when of be places of him think had to just to see to half in it laughs much a to call village the dear kind the*

SEED TEXT 2: *her a bit it must be very humiliating to be caught creeping by daylight i always lock the door when i creep by daylight i do it at night for i know john would suspect something at once and john is so queer now that i want to irritate him i* **GENDERATED TEXT 2:** *you patent a if life the personally daylight the if never a and room expects the wanted you even and if sometimes for and you so floor for so what she she so she she it knows now so even not practical he manner and under think it you still*

SEED TEXT 3: *thing i can do is to think about my condition and i confess it always makes me feel bad so i will let it alone and talk about the house the most beautiful place it is quite alone standing well back from the road quite three miles from the village it* **GENDERATED TEXT 3:** *awfully the touched the awfully in changes lovely and baby in such and into something queer*

SEED TEXT 4: *house i find it hovering in the diningroom skulking in the parlor hiding in the hall lying in wait for me on the stairs it gets into my hair even when i go to ride if i turn my head suddenly and surprise is that smell such a peculiar odor too* **GENDERATED TEXT 4:** *and baby and creep of at writing of and room came of to must oldfashioned each last use so and sake window can that she we exhaust for it you sort fancy horror a heavy the can to until am and about daylight often a judge this my it times*

SEED TEXT 5: *and the everlastingness up and down and sideways they crawl and those absurd unblinking eyes are everywhere there is one place where two breadths match and the eyes go all up and down the line one a little higher than the other i never saw so much expression in an inanimate* **GENDERATED TEXT 5:** *in*

SEED TEXT 6: *room all but that horrid paper out of one window i can see the garden those mysterious deepshaded arbors the riotous oldfashioned flowers and bushes and gnarly trees out of another i get a lovely view of the bay and a little private wharf belonging to the estate there is a* **GENDERATED TEXT 6:** *after wonderfully in*

SEED TEXT 7: *him at that makes me very tired i like our room a bit i wanted one downstairs that opened on the piazza and had roses all over the window and such pretty oldfashioned chintz hangings but john would not hear of it he said there was only one window and not* **GENDERATED TEXT 7:** *to such piece becomes the told would it about patent still a him sly a under do john even and if to be asking the once the by getting a into of them wake knocks the night line untenanted the jump the touched the room night for like did of*

SEED TEXT 8: *mastered it but just as you get well under way in following it turns a back somersault and there you are it slaps you in the face knocks you down and tramples upon you it is like a bad dream the outside pattern is a florid arabesque reminding one of a* **GENDERATED TEXT 8:** *open the had to such not looked three a take and thing enough a should low so you she road a into so noticed but is horizontal that fade the country it entertain like asleep me like and her stimulating out it enough course of terror john john so and*

SEED TEXT 9: *all around just as the sun does i hate to see it sometimes it creeps so slowly and always comes in by one window or another john was asleep and i hated to waken* *him so i kept still and watched the moonlight on that undulating wallpaper till i felt creepy* **GENDERATED TEXT 9:** *then gnawed shook is spots the soon in whim the shake the few the that it wharf the house the had door a and room care each the baby the might in it just enough to spots a be to must those to him you down thousandth the distinguish the*

SEED TEXT 10: *convinced for you see i sleep and that cultivates deceit for i tell them no the fact is i am getting a little afraid of john he seems very queer sometimes and even jennie has an inexplicable look it strikes me occasionally just as a scientific hypothesis that perhaps it is* **GENDERATED TEXT 10:** *gathered ride in it*

Commentary

The model successfully generated 10 sequences using words from the vocabulary but the sequences themselves are incoherent. There is no discernable connection to the seed texts and there is no overall meaning behind any of sequence generations. The sequences are also sometimes are cut short. For example, in the 5th sequence generation, only the word ‘in’ was generated. The model trained over 100 epochs had a loss of 1.84.

Second LSTM model (with embedding layer)

The seed text and the corresponding text sequences generate by the second model are as follows:

SEED TEXT 1: *in the walls the paint and paper look as if a school had used it it is stripped great patches all around the head of my bed about as far as i can reach and in a great place on the other side of the room low down i never saw* **GENDERATED TEXT 1:** *there and course night what one much it might as windows fast that what one it little is better me a into and you it watch myself long make myself out it pattern other of find to our pattern am and if see to want till there can sure is*

SEED TEXT 2: *a must have had perseverance as well as hatred then the floor is scratched and gouged and splintered the plaster itself is dug out here and there and this great heavy bed which is all we found in the room looks as if it had been through the wars but i* **GENDERATED TEXT 2:** *me night said it pattern know you done the shake in teeth the wanted a soon all him also long the the not getting that what said she first and room not path one it laughs do do and about see now paper my down could me with queer and*

SEED TEXT 3: *help to john such a real rest and comfort and here i am a comparative burden already nobody would believe what an effort it is to do what little i am dress and entertain and order things it is fortunate mary is so good with the baby such a dear baby* **GENDERATED TEXT 3:** *a see me your see so and about see to asked worst the dear it might up awful watch more he can is broken shall when much but it pattern and or so over would and faint of and if sometimes long a into said to figures society the and*

SEED TEXT 4: *be safe the furniture in this room is no worse than inharmonious however for we had to bring it all from downstairs i suppose when this was used as a playroom they had to take the nursery things out and no wonder i never saw such ravages as the children have* **GENDERATED TEXT 4:** *they new a see me a good to very believe as is wonder and or wonder and or so and life john is alone so and about enough a could me so from*

SEED TEXT 5: *all broken now there was some legal trouble i believe something about the heirs and coheirs anyhow the place has been empty for years that spoils my ghostliness i am afraid but i is something strange about the can feel it i even said so to john one moonlight evening but* **GENDERATED TEXT 5:** *all*

SEED TEXT 6: *the thing was that showed dim now i am quite sure it is a woman by daylight she is subdued quiet i fancy it is the pattern that keeps her so still it is so puzzling it keeps me quiet by the hour i lie down ever so much now john* **GENDERATED TEXT 6:** *is never to light in doing but it*

SEED TEXT 7: *nobody could climb through that strangles so i think that is why it has so many heads they get through and then the pattern strangles them off and turns them upsidedown and makes their eyes white if those heads were covered or taken off it would not be half so bad* **GENDERATED TEXT 7:** *would one it hand to him absolutely in as is wonder is to deceit unheardof touched in always and faint of it it little the it ever first said it might as you reach he it where as is cannot a*

SEED TEXT 8: *you get well under way in following it turns a back somersault and there you are it slaps you in the face knocks you down and tramples upon you it is like a bad dream the outside pattern is a florid arabesque reminding one of a fungus if you can imagine* **GENDERATED TEXT 8:** *because*

SEED TEXT 9: *burden already nobody would believe what an effort it is to do what little i am dress and entertain and order things it is fortunate mary is so good with the baby such a dear baby and yet i cannot be with him it makes me so nervous i suppose john* **GENDERATED TEXT 9:** *even it very believe think is back could has ever first are same john is so a good a fast so and wallpaper to air another the and give it little moonlight great but it open the into the same the dear it straight in left this but it smell*

SEED TEXT 10: *daytime in the daytime it is tiresome and perplexing there are always new shoots on the fungus and new shades of yellow all over it i cannot keep count of them though i have tried conscientiously it is the strangest yellow that wallpaper it makes me think of all the yellow* **GENDERATED TEXT 10:** *there john had so*

Commentary

The second model suffers from the same incoherence and shortened sequence generations as the first model however, I believe that the model more successfully strung 2-3 words together in a way that made sense. The model also had an improved loss of 0.53 after being trained over the same number of epochs as the first model (100).

Reflection

I've come to learn just how impressive it is that models like OpenAI's ChatGPT and Google's Gemini are as good as they are. Having a model that is able to create not just 50 word sequences but whole novels of coherent language is amazing. I realize those models are infinitely more robust than the ones made here, but my awe is all the same and I'm excited to learn more about how deep learning is used in LLMS.

In this assignment I learned how different ways of breaking up text (chars vs words) affects a models ability to learn and how using the connotation of words and the contextual language around a word is important in model training. I learned that I am very smart to be able to speak sentences at all.