Homework 4 - The Great Gatsby Sequential Text Generation Using LSTM Models

Ryan King

Introduction

In this study, I explored sequential models but specifically used a Long Short-Term Memory (LSTM) network, for generating text. The objective is to create a model that can generate new lines of text based on F. Scott Fitzgerald's classic novel, "The Great Gatsby". This assignment acts as a practical application of concepts learned in class, showcasing the potential of LSTMs in natural language processing.

Analysis

I conducted an exploratory analysis to comprehend the text's structure and style. The data was sourced from Project Gutenberg and contains all of the text to "The Great Gatsby". During the data cleaning and preparation stage, non-alphabetic and punctuation were removed, and letters were converted to lowercase to standardize all of the input for our model. This resulted in 46,804 tokens and 5,996 unique ones within the dataset. Text sequences consisting of 100 tokens each were readied, building the foundation for training the sequential model. The dataset was divided into training and testing sets using a 80/20 split. This split allows us to train the model on a significant amount of the data, while saving some unseen data for the model to be evaluated on.

Methods

To start off, the data was structured into sequences of 100 tokens each and put these into input and output sets. The first model had a sequential architecture with an initial Embedding layer of 25 dimensions. Two LSTM layers with 150 units each followed, and then a fully connected Dense layer with 150 units and a ReLu activation function. I also included another Dense layer with the amount of units as the vocab size and a softmax activation. For model training,

I used the categorical cross-entropy loss function along with the Adam optimizer. I ran the model for 225 epochs with a batch size of 128. I experimented with many different epoch amounts and unit amounts to see which produced the strongest model. No other combination produced better loss and accuracy values than the one above so I chose to go with that.

For the second model, my substantial adjustment was to use a custom embedding layer, which increased the model's complexity. The embedding layer's dimensionality was put at 50, much bigger than the first model's at 25. This allows us to gain a better and more nuanced representation of each token in the text. Since the embedding layer increased I decided to increase the LSTM layers. I made the model deeper and better at understanding the sequential dependencies by adding one additional LSTM layer that had 150 units. I also increased my Dense layer with the ReLu activation to 200 units, improving the model's abilites to learn patterns in the text. This model training used the same categorical cross-entropy loss function and Adam optimizer. It was trained on 225 epochs with a batch size of 128 to make it easy to compare to the first model. These changes greatly enhanced the model's ability to generate text in a more coherent style similar to Fitzgerald's.

Results

Model 1 Generated Sentences:

1. only thing of years and i guessed at his i said someone he had the decency from his eyes then looks under the kitchen steps with a copy of simon called you a little later i had the bottle of second now in constant demand from the heat with them

2. at me with the easygoing blue coupé gave her lying me and he winced himself in me if i live with a low vulgar the is cars behind the confusion of a man could store to walk down to my car and if i was a said go inside so

3. it might sober me up on said every very sign you been a rather sinister passed to drain the pool today i was promoted to have a pause there passed well close to wilson with the first would stop the car and i was liable his coffee just he see

4. went as having a family she looked at me as if i had gone into loving i was hear i said very it the time of the telephone wind inside as i came into a young actress and to have how if i was dawn now forever compared to the

5. on all except to pieces and i keep just as confounding it were having a touch of changed that almost love to going to call to southampton positively night it been going to do i said right her business was a fast men of exultation a new wellbeing radiated was

6. like a clergyman and says finnish wisdom that life apart from him and everyone looked at feet through the road i inquired with a block of simon we met him to my house a quarter of a mile down the way at the columbus way looked back the tremendous vitality

7. before the long parting the hand took in the beach there stood so there was a man than that the ear follows inside and stood so having the young cross to shake own bit that somebody happened he was a son of representing the staid nobility of the egg condescending

8. to it if he climbed alone and once there he say be used to come on that everything does he married him and then the first time in her as it pick up many times far about a low whitewashed movement with the summer rising like down in the second

9. was sharply different from the west where an evening was hurried from phase to phase towards its close in a continually disappointed anticipation or else in sheer nervous dread of the moment itself make me feel uncivilized i confessed on my second glass of corky but rather impressive claret you

10. york the racy adventurous feel of somebody at the night between it was on various unrevealed capacities he had begun her absurd all we seemed to her but i realize any the coupé raced nothing to go somewhere and you up and ran up the darkness along his eyes and

Model 1 Results:

The sentences generated by Model 1 show a solid level of grammatical correctness and display a diverse vocabularly reflective of "The Great Gatsby." The model succeeds at constructing sentences with different lengths but the sentences show an inconsistency in maintaining narrative flow and contextual coherence. They are very confusing and it is tough to understand what they are trying to say. The sentences are disorganized, which suggests a challenge in the model's capability to understand and replicate the narritve style of Fitzgerald's writing.

The training metrics of Model 1 showed valuable insights to its performance. It had a loss value that steadily decreased throughout the epochs and landed at 0.6962 at the final epoch. The decrease in the loss value shows that the model is effectively learning from the training set. The model also produced an accuracy score of 83.46%. This metric reflects the model's ability to correctly predict the sequence's next word. The final epochs showed the model's strongest performance with the highest accuracy and smallest loss values.

Model 2 Generated Sentences:

1. heard off to had she insisted and two nobody for the mint the reluctance of stimulating this distance from around looks and one in the slowly to the south no expanse of though

to my own nose had made seventyfive the same room and the other no horses needless a

2. on the big tree and all exactly nodding it an hour we from a man who is with a quarter of a mile air there on the one and looking for she had left a little car and i loved the phone took for the table and agreed a most

3. saw a thing i want to kiss through me anxiously stronger and gatsby i should give me says gasoline again now and spoke in long world there hooked me in it whispered him before he could have a man a bad rich no frantic she broke out of the island

4. the armistice and in february she was presumably engaged to a man from new orleans in june she married tom buchanan of chicago with more pomp and circumstance than louisville ever knew before he came down with a the steps think i happen to called him miss very dog before

5. the sill daisy and tom were sitting opposite each other girl with love with he was getting off a long year already as everything he took himself to duluth and terrible just in it i say is at it have i lay on a little office as mr carraway see

6. stared at me in such an amazed way and denied so vehemently any knowledge of his movements and i slunk back in the direction at the cocktail gesture as his throat the police girl tom made no other and whom i was she had to be to get all is

7. we all took the less explicable step of engaging the parlour of a suite in the plaza hotel the prolonged and tumultuous argument that ended by herding us into that room eludes me though i had waited under the car i felt it want her old things but i like

8. the corner i said have to leave you you interposed tom quickly be hurt if you come up to the apartment you she urged telephone my sister catherine said to be the few minutes later was on the corner i pushed the most gleam of jay he was still as

9. to stay over me and all the terrible time she had no one and uncles at its share of a suit named personality her mckee drowned his head a little would be to get i want to hear that about the said to know he was a friend of he

10. dog and her other purchases and went haughtily in going to have the mckees come she announced as we rose in the elevator of course i got to call up my sister the apartment was on the top small livingroom a small diningroom a small bedroom and a bath the

Model 2 Results:

The sentences generated by Model 2 show a basic understanding of grammatical structure,

displaying its ability to create grammatically correct sequences. Similar to the other model's sentences this one had a lack of coherence and narrative style. The sentences normally start off pretty clear but then become disjointed as they go on. The model did learn effectively from the text because it had a wide vocabulary. The sentences include a mixture of simple and complex structures, showing that the model can construct different kind of sentences. The model is strong when it comes to grammer and vocabulary but could improvement in coherence and stylistic alignment.

Model Comparison Model 2 is slightly better at maintaining sentence coherence and thematic continuity. The sentences seem to be more logically constructed. Model 2 also shows an advantage in stylistic alignment, showing more of an understanding of the style and tone of "The Great Gatsby." The Model 2 sentences are also a bit more complex and flow better. Although Model 2 produces better overall sentences, both models could be improved to generate more coherent sentences that align more to Fitzgerald's style.

Now comparing the metrics of the models, Model 1 ended with a loss value of 0.6962 and Model 2 finished at epoch 225 with a loss value of 0.3049. Model 2's lower loss value suggests that the model is better at predicting the next word than Model 1. Model 2 had a much higher accuracy of 92.83% compared to Model 1's accuracy of 83.46%. Model 2's better accuracy indicates that it is better at correctly predicting the next word in a sequence.

Reflection

During this assignment, my understanding of sequential models and their role in natural language processing has greater increased. Using "The Great Gatsby" as my text source, I faced challenges and relevations that have helped shape my approach to machine learning. This assignment allowed me to learn the intricate balance that is needed for creating model architectures. I had to adjust the different parameters to see what would make the model's performance better and improve the metrics. This process was also a challenge because I kept trying new parameter values and it took a large amount of time to run the model each time. In the future, I would experiment with more techniques to help prepare the data. In this assignment, I mainly just stuck to the code we were given in class so the next one I want to try different ways to get my data ready. I would also consider experimenting with different architectures like Transformer models, because these models are very successful when it comes to natural language tasks. For future assignments, I will give myself more time to complete the projects because creating models with these large datasets can take a very long time and I did not account for that on this assignment.