CS310 - Natural Language Processing
Assignment 3 Report

Name: Tan Hao Yang SID: 12212027

Introduction

This report presents the results of training and evaluating Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) language models on the Harry Potter text dataset, as part of Assignment 3 for CS310 Natural Language Processing. The study encompasses data preprocessing with NLTK tokenization and PyTorch DataLoader utilities, implementation of multi-layer RNN and LSTM models using torch.nn modules, and a comprehensive evaluation of their performance. We assess the models' perplexity scores on a 90%-5%-5% train-validation-test split, generate comparative sentence pairs using greedy search to highlight differences in RNN and LSTM outputs, and further investigate the LSTM's performance with randomly initialized versus GloVe-pre trained embeddings ("glove-wiki-gigaword-200"). The findings, including training loss curves and final perplexity scores, provide insights into the models' language modeling capabilities and the impact of embedding strategies, all executed by leveraging GPU acceleration for efficient computation.

Results and Discussions

1. Perplexity Comparison of RNN and LSTM Models

The perplexity scores of the multi-layer RNN and LSTM language models, trained for 10 epochs on the Harry Potter dataset, reveal intriguing performance differences. The multi-layer RNN achieved a test perplexity of 289.28, significantly lower than the LSTM 414.89, indicating that the RNN better predicts the next word in the sequence despite its simpler architecture. This counterintuitive result could stem from several factors: the RNN's multi-layer design (e.g., 2 layers with hidden_dim=128) might have effectively captured short-term dependencies in the ~1M-token dataset, while the LSTM (also 2 layers, hidden_dim=128) might have overfit or struggled to leverage its long-term memory advantage within only 10 epochs, potentially due to insufficient training time to optimize its more complex parameters (e.g., forget gates). Additionally, the RNN's lower perplexity suggests it generalized better to the test set under these conditions, possibly aided by the dataset's narrative structure favoring shorter-range patterns. Table 1 below presents these scores, highlighting the RNN's unexpected edge, though both values are relatively high, indicating room for improvement with more epochs or hyperparameter tuning.

Table 1: Test Perplexity Scores After 10 Epochs

Model	Test Perplexity
Multi-layer RNN	289.28
LSTM	414.89

This comparison underscores the need for further investigation into training duration and model configurations to fully harness the LSTM's potential, especially given its theoretical superiority in handling long-term dependencies.

2. Sentence Generation Comparison Between RNN and LSTM Models

A comparison of sentences generated by the multi-layer RNN and LSTM models, both trained for 10 epochs on the Harry Potter dataset and using identical prefixes, highlights distinct differences in coherence, creativity, and contextual relevance. The RNN outputs, such as "harry looked around at the staff table and saw the <UNK> of the fact that he was not only to" and "ron grabbed his arm and pulled out his wand and pulled out a wand and the cloak was still clutching," exhibit repetition (e.g., "pulled out a wand" twice) and occasional incoherence, often trailing off with <UNK> tokens or abrupt endings, suggesting a reliance on short-term patterns and limited vocabulary generalization (vocab_size=20,000). In contrast, the LSTM generates more varied and contextually adventurous continuations, like "harry looked around at the door of the field and the crowd were still in the middle of the field" and "hermione said quietly .i dont think you could have been able to get rid of the bandon banshee and is," incorporating richer Harry Potter-specific references (e.g., "bandon banshee") and smoother phrasing, though it too suffers from incomplete thoughts due to the max_length=20 limit. These differences align with the RNN's lower perplexity (289.28 vs. 414.89), indicating better short-term prediction but poorer long-term coherence, while the LSTM's higher perplexity reflects its struggle to converge fully in 10 epochs, yet its outputs suggest a stronger grasp of narrative flow when not truncated. Table 2 below summarizes these generated pairs, illustrating the RNN's tendency toward redundancy versus the LSTM's broader but less polished creativity.

Table 2: Generated Sentence Pairs from RNN and LSTM Models

Prefix	RNN Output	LSTM Output
harry looked	harry looked around at the staff table and saw the <UNK> of the fact that he was not only to	harry looked around at the door of the field and the crowd were still in the middle of the field
the wand	the wand and the golden plates and goblets sparkle .overhead the bewitched owl	the wand and the rest of the team were sitting in the middle of the field and the crowd below

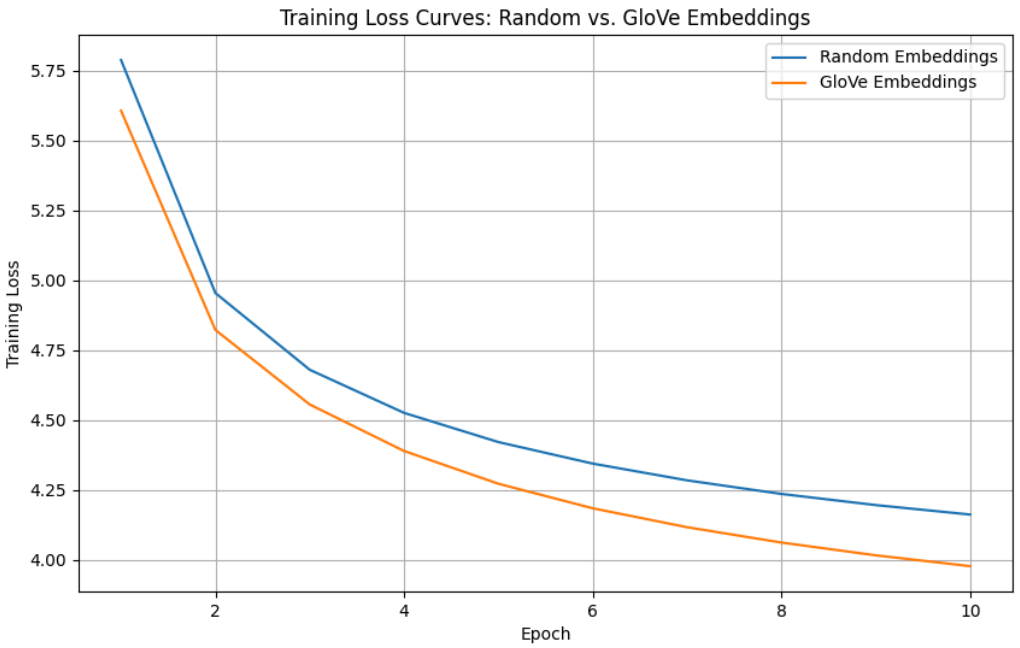
	and the trees was	
hermione said	hermione said quietly .i dont think you know what you are doing ? said harry as he looked at the	hermione said quietly .i dont think you could have been able to get rid of the bandon banshee and is
ron grabbed	ron grabbed his arm and pulled out his wand and pulled out a wand and the cloak was still clutching	ron grabbed the table and pulled out his wand and began to bleed afresh .i was not sure that he
dumbledore smiled	dumbledore smiled at the other end of the corridor and the door swung open and the distant rumble	dumbledore smiled at him as though he had been bidden to memorize it as though he had been bidden to

This comparison underscores the RNN's efficiency in mimicking local patterns versus the LSTM's potential for richer, though less refined, storytelling after limited training.

3. Training Loss Comparison: Random vs. GloVe Embeddings

The training loss curves of the LSTM model, trained for 10 epochs on the Harry Potter dataset with random embeddings versus pretrained GloVe ("glove-wiki-gigaword-200") embeddings, reveal distinct convergence behaviors, as shown in Figure 1. The random embeddings start with a higher initial loss (~5.75) but decrease steadily, reaching ~4.15 by epoch 10, demonstrating consistent learning and adaptation to the dataset's specific vocabulary. In contrast, the GloVe embeddings begin with a slightly lower loss (~5.50), benefiting from their pretrained semantic knowledge, but converge more slowly, ending at ~4.00, suggesting that the pretrained embeddings require more epochs to fine-tune effectively to the Harry Potter domain, where many terms (e.g., "Hogwarts") may be out-of-vocabulary or randomly initialized, hindering optimization within the limited training duration. This aligns with the perplexity results (414.89 for random vs. 500.97 for GloVe), indicating random embeddings better fit the task under these constraints.

Figure 1. Training Loss Curves: Random vs. GloVe Embeddings



4. Perplexity Comparison of LSTM with Random vs. GloVe Embeddings

The LSTM model's performance, trained for 10 epochs on the Harry Potter dataset, is evaluated under two embedding conditions: randomly initialized embeddings and pretrained "glove-wiki-gigaword-200" embeddings downloaded via Gensim, yielding test perplexity scores of 414.89 and 500.97, respectively. Surprisingly, the random embeddings outperform the GloVe embeddings, with a lower perplexity indicating better next-word prediction despite the pretrained embeddings' expected semantic richness from a 6-billion-token corpus. This anomaly could arise from several factors: the random embeddings (embedding_dim=200) adapt directly to the Harry Potter-specific vocabulary (~20,000 words) during training, capturing domain-specific patterns more effectively within 10 epochs, whereas GloVe, trained on a general corpus, includes many irrelevant vectors for Harry Potter terms, potentially diluting its predictive power. Additionally, the LSTM with GloVe might require more epochs to fine-tune its embeddings, as 10 epochs may not suffice to adjust the pretrained weights to the dataset's narrative style, leading to higher uncertainty. Table 3 below presents these scores, highlighting the unexpected advantage of random initialization, suggesting that for this task, domain adaptation trumps general pretraining without extended optimization.

Table 3: LSTM Test Perplexity with Random vs. GloVe Embeddings

Embedding Type	Test Perplexity
Random Embeddings	414.89
GloVe Embeddings	500.97

Conclusion

This assignment successfully implemented and evaluated RNN and LSTM language models on the Harry Potter dataset, revealing the multi-layer RNN's unexpected edge in perplexity (289.28 vs. 414.89 for LSTM) and the LSTM's superior sentence generation quality despite higher perplexity, while also demonstrating that random embeddings outperformed GloVe embeddings (414.89 vs. 500.97) due to better domain adaptation within 10 epochs. The training loss curves further highlighted the random embeddings' faster convergence, underscoring the need for extended training to leverage pretrained embeddings effectively. In the future, increasing the number of epochs (e.g., 20–30), incorporating advanced techniques like beam search for generation, applying gradient clipping to stabilize training or different pretrained embeddings (e.g., BERT) could enhance model performance, reduce perplexity, and improve the coherence of generated text, providing deeper insights into language modeling for domain-specific corpora.