Introduction

This project aims to explore text generation using deep learning models by training neural networks on the text of *The Gift of the Magi*. The challenge involves creating models that predict the next word in a sequence, enabling the generation of fluent, contextually appropriate text. The use of pre-trained embeddings (FastText) alongside a basic GRU-based architecture demonstrates the potential of external knowledge to enhance model performance. This work is relevant for applications such as autocomplete, content generation, and chatbot development.

Analysis

The dataset consisted of the plain text from *The Gift of the Magi*. Key preprocessing steps included:

1. Lowercasing and removing punctuation.
2. Splitting text into sentences and filtering out empty lines.
3. Tokenizing sentences into sequences, resulting in increasing n-gram sequences for modeling.

The tokenized sequences were padded to ensure consistent input lengths. Inputs (X) comprised all tokens except the last in each sequence, while the outputs (y) represented the final token. One-hot encoding was applied to y to enable categorical classification.

Methods

Two models were trained and evaluated:

1. **Basic GRU Model**:
    - **Architecture**: Embedding layer (50 dimensions), two GRU layers (128 units each), and a Dense layer with softmax activation
    - **Training Duration**: 10 minutes for 75 epochs
    - **Optimizer**: Adam
    - **Metrics**: Accuracy and Top-5 categorical accuracy
2. **Pre-trained FastText Embedding Model**:
    - **Architecture**: Pre trained FastText embeddings (300 dimensions), two GRU layers (128 units each), and a Dense layer with softmax activation
    - **Embedding Usage**: Non-trainable weights for embeddings
    - **Training Duration**: 2.5 minutes for 50 epochs
    - **Optimizer:** Adam
    - **Metrics**: Accuracy and Top-5 categorical accuracy

Results

Basic GRU Model:

- **Final Metrics**:
    - Training Loss: 0.2031
    - Training Accuracy: 94.99%
    - Top-5 Accuracy: 98.94%
- **Generated Text Examples**:
    - SEED TEXT: redistributing or providing access to a work with the phrase project
    - GENERATED TEXT: anywhere rosy rapidly include expend merchantability red several old silent sweetness 1e expenses virus before bearing using using ensuring hysterical recently did did most sending sending rosy strips necessitating accepting accepting which 5 brilliantly brilliantly subject though phrase phrase phrase compressed compressed compressed invalidity agile agile agile giving step renamed
    - SEED TEXT: revenue service the foundations ein or federal tax identification
    - GENERATED TEXT: hypertext alas family madame family ten ten chronicle chronicle chronicle cried exists look yet above inconsequential lived below below abide world among promotion word 7256 7256 donation compressed compressed unsolicited across unsolicited rosy chronicle strips outside easily schoolboy doubt vanished current current anywhere shave especially wish wish elect gradually pg
    - SEED TEXT: down rippled the brown cascade
    - GENERATED TEXT: quick lighted legally accept accept 1a flop he he profits profits door door federal federal federal chaste chaste late tax hypertext fryingpan yer calculate grow cost company ransacking surely surely any any down choose metaphor 1e9 years among cascade street an an word pglaf always shell shell redistribution wails statements
    - SEED TEXT: 1e4 do not unlink or detach or remove the full project gutenberg
    - GENERATED TEXT: knee our doesnt bound king king modest distribute distribute anywhere parsimony said interpreted fob often unsolicited unsolicited always paperwork why credit madame when produce chronicle expensive expensive expend almost void subject these pennies pennies children word had exporting exporting indirectly agile agile agile step step renamed renamed financial hour hour
    - SEED TEXT: the following which you do or cause to occur a distribution of this
    - GENERATED TEXT: terrified hypertext goals critical prohibition prohibition 1e8 cost things things references thus thus gas using pure drew recently vestibule addition wails right sending sending former rosy former rosy former imposed people people table necessitating comforting especially consequential staff accepting accepting which brilliantly brilliantly windows windows 1e3 breach were performances able
    - SEED TEXT: meretricious ornamentationas all good things should do it was even
    - GENERATED TEXT: agile agile agile creation business offers reports further distribute surely parsimony calculated ecstatic ecstatic instead gradually turned reasonable reasonable even widespread shared which thin produced manger sending sending exactly necessitating 7 silent staff staff accepting spent walks walks simplest revenue indicating learn network sending sending sending necessitating email asis sweetness
    - SEED TEXT: jim and greatly hugged by mrs james dillingham young already
    - GENERATED TEXT: locations network addition widespread widespread shared which subject subject provide provide an word outside young locations mirror identification identification near near however disclaimers disclaimers disclaimers sign profits chronicle

chronicle cried service interpreted interpreted interpreted brilliant dealing introduced grandfathers internal tresses help serious redistributing royalties will performances yearned word far disclaimers
    - SEED TEXT: professor michael s hart was the originator of the project
    - GENERATED TEXT: direct direct mistress january january further further unenforceability an an word pennies week assistance mr whirl pennies secure fees fees lake lake asked good good invented dry receiving receiving wwwgutenbergorgdonate regard wake wake 1e9 gift gift maybe went went made 20 made disclaim word chorus distributed distributed distributed produced wriggled
    - SEED TEXT: jim had not yet seen his beautiful present she held it out to him
    - SEED TEXT: foundation or pglaf owns a compilation copyright in the collection
    - GENERATED TEXT: shell 1500 1500 o finger interpreted interpreted brilliant brilliant dealing introduced introduced shared without service interpreted interpreted near near however disclaimers disclaimers disclaimers charges derive considerable metaphor some metaphor 1e9 years current current works needed current needed faltered reasonable several clearly clearly love love rippled detach detach detach beautiful meet
    - SEED TEXT: jim had not yet seen his beautiful present she held it out to him
    - GENERATED TEXT: outside cost additions greatly appertaining performances email 4 collected donate of alas brought financial reflection been been international elect were cut cut piled quick exists comforting public public organized applied punitive royalties edition faltered sharing sharing sharing sharing sharing sharing sharing originator lot lot lot set set truant redistributing will

Pre-trained FastText Embedding Model:

- **Final Metrics**:
    - Training Loss: 0.3070
    - Training Accuracy: 93.88%
    - Top-5 Accuracy: 98.63%
- **Generated Text Examples**:
    - SEED TEXT: dillingham young came home and reached his flat above he was called
    - GENERATED TEXT: the solicitation of the united states we do not agree to the terms of this agreement for keeping the work as long as set forth in the terms of the full project gutenberg license must appear created the new work in the united other states states and most other parts
    - SEED TEXT: she found it at last it surely had been made for jim and no one else
    - GENERATED TEXT: in the user to return or destroy all of this work in the person of the work in the united states we do not agree to the terms of the full project gutenberg license must appear about the foundation the project gutenberg mission of promoting the chops on account in
    - SEED TEXT: country other than the united states
    - GENERATED TEXT: and most other parts of the project gutenberg license when any particular paper and distributing a project gutenberg work any works appears in the foundation the project gutenberg electronic for paper walking a gray

backyard tomorrow would be christmas day and cried to wake with her cheeks and very agile

- SEED TEXT: visit wwwgutenbergorgdonate
- GENERATED TEXT: your efforts and donations can help see our second copy is a shampoo and stood his while a tear or a shampoo distributing a practised hand and threw it upon the watch della reached the sign of the flat across the airshaft della would have pulled person practised and the
- SEED TEXT: limitation permitted by the applicable state law the invalidity or
- GENERATED TEXT: damaged disk or any part of this agreement shall not void the copyright of project gutenberg works calculated using the method of or implied including any part of shabby menendez the foundation as set of paragraphs 1e1 through 1e7 or any other work distributing a refund of any money paid
- SEED TEXT: was in pennies pennies saved one and two at a time by bulldozing the
- GENERATED TEXT: stores and she had turned all the magi had been flung to the breeze during a former period of this agreement you must obtain permission in writing copying or in any purpose such as of the sentiments of the project gutenberg license included with this ebook or small small small
- SEED TEXT: from people in all walks of life
- GENERATED TEXT: and she had turned all the magi of the public domain and the united variety things one was on the back of very owned by paper and then a look at the home a furnished flat at the glass away and now she had been saving for a minute and
- SEED TEXT: jim had not yet seen his beautiful present she held it out to him
- GENERATED TEXT: him sold a watch i had island hat hat a week jim might have pulled out his watch jim might be read to him cheeks and shining like i sold a watch for day and went to him cheeks and her hair and cried out to pulled his fixedly from
- SEED TEXT: you neednt look for it said della its sold i tell yousold and
- GENERATED TEXT: keep that she had been for merry as gave a dollar and eightyseven hurried at the first flight and she had been prepared for very ring jim would have pulled me like a truant schoolboy she looked at her cheeks and the butcher until she held my sight merry task
- SEED TEXT: will be linked to the project gutenberg license for all works
- GENERATED TEXT: in formats readable by the widest variety for things merry nice and an expression but of no refund disclaimer if other format or other format with the work as long as very as very two very stood rapid gave a present very agile to prudence james and very a day

Observations:

- **Basic GRU Model**
  - Training Metrics:
    - The GRU model achieved excellent training accuracy (94.99%) and high Top-5 accuracy (98.94%), indicating that it learned to predict the next word effectively and identified relevant candidates with high confidence.
    - The relatively low final training loss (0.2031) reflects its ability to model the dataset well without overfitting.
  - Generated Text Analysis:
    - The generated text demonstrates a limited ability to maintain coherence and context across longer sequences. For instance, sentences like:
      - "anywhere rosy rapidly include expend merchantability red several old silent sweetness..."
      - Suggest random and repetitive token choices, often diverging from the style of the seed text.
    - Occasionally, generated sequences incorporate irrelevant or nonsensical phrases ("1e expenses virus before bearing using using ensuring hysterical"), indicating room for improvement in generalization.
- **Pre-trained FastText Embedding Model**
  - Training Metrics:
    - The FastText model achieved slightly lower training accuracy (93.88%) and Top-5 accuracy (98.63%) compared to the GRU model, but its metrics remain competitive.
    - A slightly higher training loss (0.3070) may reflect the fixed nature of pre-trained embeddings, which limits model optimization but preserves general language knowledge.
  - Generated Text Analysis:
    - The FastText model produced outputs with better semantic consistency and stylistic alignment with the dataset:
      - "the solicitation of the united states we do not agree to the terms of this agreement..." reflects legalistic language similar to the original text.
      - "your efforts and donations can help see our second copy is a shampoo and stood..." exhibits more logical progression, though occasionally lapses into randomness.
    - Longer sequences retain more contextual integrity compared to the GRU model.
    - Despite improvements, occasional oddities such as "a shampoo distributing a practised hand" and repetitive phrasing ("as very two very stood") still emerge.

Reflection

Pre-trained embeddings dramatically reduced training time. While the GRU model achieved higher accuracy, the FastText model excelled in generating semantically coherent sequences. Augmenting the dataset with additional texts would enhance diversity and coherence as well as Fine-tuning FastText embeddings to capture domain-specific nuances.