Lab 2: Language Models

RNNs, GRUs, and Feed-Forward Networks

Youness Anouar
UM6P-CC

February 11, 2026

Contents

1 Dataset

1.1 Presentation

The dataset used in this project is the **HuffPost News Category Dataset**, which can be found on Kaggle and Hugging Face. It contains approximately 200,000 news headlines and short descriptions from the Huffington Post between 2012 and 2018.

Each record in the dataset is a JSON object with the following attributes:

- **category**: The category of the article (e.g., CRIME, POLITICS).
- **headline**: The title of the article.
- **authors**: The author(s) of the article.
- **link**: URL to the original post.
- **short_description**: A brief summary of the article content.
- **date**: The publication date.

For this project, we utilize the `short_description` field as the source text for training the language models. This field provides concise, formal, and modern English sentences suitable for learning sequential patterns.

The dataset can be downloaded from Hugging Face:
`https://huggingface.co/datasets/khalidalt/HuffPost`

1.1.1 Preprocessing

To improve model performance and reduce the vocabulary size, we apply the following preprocessing steps to the `short_description` text:

1. **Normalization**: All text is converted to lowercase.
2. **Cleaning**: Hyphens are replaced with spaces. Punctuation is removed, except for apostrophes (′) and ampersands (&), which are preserved.
3. **Tokenization**: The text is tokenized using a space-based tokenizer. Special tokens `<start>` and `<end>` are added to the beginning and end of each sequence to mark boundaries.
4. **Vocabulary**: A vocabulary is built from the top 50,000 most frequent words. Words not in the vocabulary are replaced with an `<unk>` token.

1.2 Training Subsets

Due to computational constraints, we utilize subsets of the full dataset for training our models. The data usage differs between the architectures:

- **Basic Neural Network (Feed-Forward)**: Trained on the first **10,000** samples. The input consists of a fixed context window of 5 words to predict the 6th word.
- **Simple RNN and GRU**: Trained on the first **1,024** samples. These models process variable-length sequences (padded to a maximum length) to predict the next token at every time step.

2 Models

2.1 Feed-Forward Neural Network (Basic NN)

The first model implemented is a primitive language model based on a Feed-Forward Neural Network (FNN). Unlike recurrent models, this architecture uses a fixed context window to predict the next word.

2.1.1 Architecture

The model processing pipeline is illustrated in Figure **??**. It consists of the following components:

- **Input**: A fixed sequence of 5 integer indices representing the preceding context words.
- **Embedding Layer**: learnable vector representation for each word.
  - Input dimension: Vocabulary size ($\approx 50,000$)
  - Embedding dimension: 64
- **Flattening**: The embeddings of the context words are concatenated into a single vector of size $5 \times 64 = 320$.
- **Hidden Layer**: A fully connected linear layer.
  - Input size: 320
  - Output size: 256
  - Activation function: ReLU
- **Output Layer**: A final linear layer mapping the hidden representation to the vocabulary logits.
  - Input size: 256
  - Output size: Vocabulary size

2.1.2 Hyperparameters and Training Configuration

The model was trained with the following parameters:

Parameter	**Value**
Context Window	5 words
Embedding Dimension	64
Hidden Dimension	256
Optimizer	AdamW
Learning Rate	0.001
Loss Function	Cross Entropy
Batch Size	512
Epochs	30

Table 1: Hyperparameters for the Basic Feed-Forward Network

2.2 Simple Recurrent Neural Network (RNN)

The Simple RNN moves beyond the fixed context window of the Feed-Forward network by processing sequences token by token. It maintains a hidden state that is updated at each time step, allowing it to theoretically capture dependencies over the entire sequence history.

2.2.1 Architecture

- **Input Processing**: Handles variable length sequences (padded to length 128) using PyTorch's $pack\_padded\_sequence$ to mask padding tokens.
- **Embedding Layer**: 64-dimensional embeddings.
- **Recurrent Layer**: A standard RNN cell with a hidden size of 256 units. One layer is used ('num_layers=1').
- **Output Layer**: A linear projection from the hidden state (256) to the vocabulary size.

2.2.2 Hyperparameters

Parameter	**Value**
Hidden Units	256
Embedding Dimension	64
Optimizer	AdamW
Learning Rate	0.001
Batch Size	16
Epochs	30

Table 2: Hyperparameters for the Simple RNN

2.3 Gated Recurrent Unit (GRU)

The GRU is an advanced variant of the RNN designed to solve the vanishing gradient problem. It introduces "update" and "reset" gates that control the flow of information, allowing the model to decide what to remember and what to forget over longer sequences.

2.3.1 Architecture

The architecture is identical to the Simple RNN, except the standard RNN cell is replaced by a GRU cell.

- **Embedding**: 64-dimensional.
- **Recurrent Layer**: GRU cell with 256 hidden units.
- **Output**: Linear layer (256 $\rightarrow$ Vocab size).

2.3.2 Hyperparameters

The GRU model uses the exact same hyperparameters as the Simple RNN (Table 2) to ensure a fair comparison.

3 Results

3.1 Basic Feed-Forward Network

The training results for the Basic Feed-Forward Network demonstrate strong convergence on the training set over 30 epochs.

- **Loss**: The Cross-Entropy Loss started at approximately 7.3 and decreased steadily, reaching a final value below 0.8. This indicates the model effectively minimized the error in its probability distributions.
- **Accuracy**: The training accuracy improved significantly, starting from roughly 10% and reaching over 85% by the end of training. This suggests the model successfully learned to predict the 6th word based on the preceding 5 words for the majority of the training samples.
- **Perplexity**: The perplexity score saw a dramatic drop from over 1400 in the first epoch to single digits (< 5) in the final epochs, showing the model's increasing confidence and reduced uncertainty in its predictions.

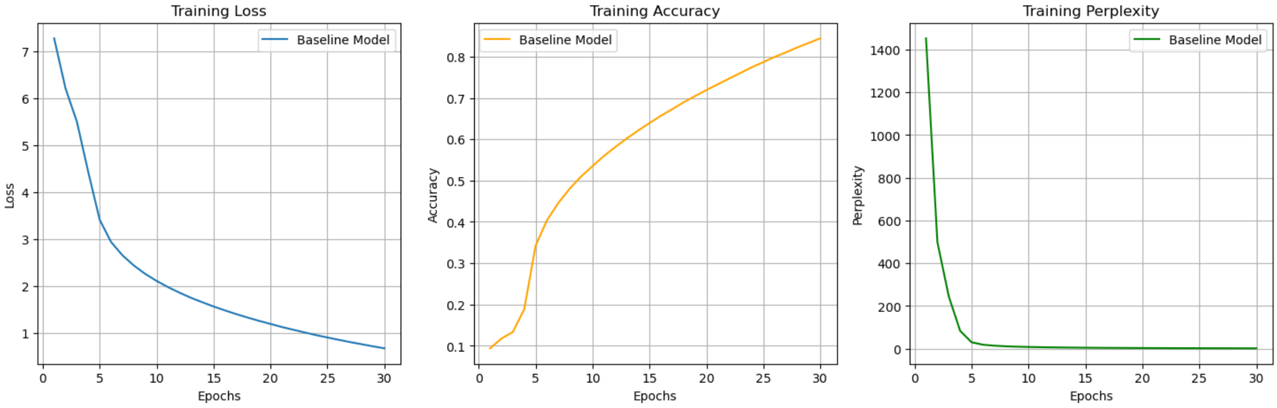

Figure 1: Training metrics for the Basic Feed-Forward Network. Left: Loss, Center: Accuracy, Right: Perplexity.

While the metrics indicate successful training, it is important to note that this high performance is partly due to the model overfitting on the fixed 5-gram patterns present in the training data. The model's ability to generalize to long-range context is inherently limited by its architecture.

3.2 Simple RNN and GRU Comparison

The training results for the Simple RNN and GRU models show their ability to handle sequential data effectively, though on this smaller subset (1,024 samples) they both fit the data very closely.

- **Convergence**: Both models converge rapidly. The Loss drops below 1.0, and Perplexity approaches 1.0, indicating the models have effectively memorized the training sequences.

- **Accuracy**: Both models achieve high accuracy ($> 85\%$). The GRU appears to have a slightly smoother convergence curve in the early epochs compared to simple RNN, likely due to its better handling of gradients.
- **Generalization**: While training performance is high, trained on a small subset (1,024 samples) carries a high risk of overfitting. The models likely memorized the specific sentences in the training set rather than learning generalized language rules.

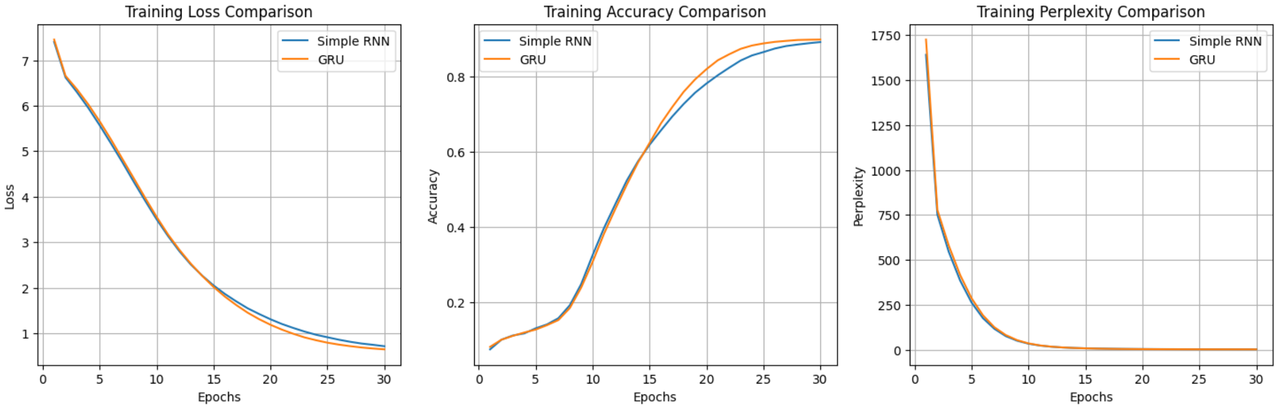

Figure 2: Comparison of Simple RNN vs GRU training metrics. Left: Loss, Center: Accuracy, Right: Perplexity.

3.3 Testing them all together

To evaluate the generative behavior of all three models, we perform a controlled next-word generation experiment using the same fixed set of test prompts for each architecture. The test sentences are constructed using the same vocabulary pool as the training dataset in order to avoid the appearance of `<unk>` tokens and ensure fair comparison. The same inputs are used for the Feed-Forward model, the Simple RNN, and the GRU.

Methodology. Each test sentence is provided as a prefix sequence, and the model is tasked with generating the continuation using greedy decoding (argmax at each step). Generation proceeds token-by-token until the `<end>` token is produced or a maximum sequence length is reached. This corresponds to a standard **next-word autoregressive generation** setup, where the previously generated token is recursively fed back as input to predict the next token.

Test sentences & models outpus.

- The actor released a statement denying the sexual harassment accusations
  1. **Feed-Forward:** the actor released a statement denying the sexual harassment accusations after his 2017 their campaign
  2. **RNN:** the actor released a statement denying the sexual harassment accusations her in your onion
  3. **GRU:** the actor released a statement denying the sexual harassment accusations against weinstein sparked the president's
- A new Star Wars film is reportedly

1. **Feed-Forward:** a new star wars film is reportedly in a new case played for justice
2. **RNN:** a new star wars film is reportedly from their long term
3. **GRU:** a new star wars film is reportedly tortured into and united arab emirates abc news

- The government failed to locate the immigrant children
  1. **Feed-Forward:** the government failed to locate the immigrant children killed up the case of age was black man who went to be in the united states
  2. **RNN:** the government failed to locate the immigrant children in the state attorney general who survived the crash
  3. **GRU:** the government failed to locate the immigrant children of escaping her violent ex husband
- Trump gives Dems an ass-kicking for not having
  1. **Feed-Forward:** trump gives dems an ass kicking for not having hard
  2. **RNN:** trump gives dems an ass kicking for not having a surprise serenade from a mariachi band
  3. **GRU:** trump gives dems an ass kicking for not having is very disturbing gop candidate wendy rogers said
- The pop star also left her
  1. **Feed-Forward:** the pop star also left her the first black woman to receive the cecil b demille award
  2. **RNN:** the pop star also left her to dial in the workplace
  3. **GRU:** the pop star also left her on a scale weve never seen before
- Irish women will travel to vote in the
  1. **Feed-Forward:** irish women will travel to vote in the us history of my federal judge
  2. **RNN:** irish women will travel to vote in the workplace over the late and rules
  3. **GRU:** irish women will travel to vote in the united states of what happens will inspire readers
- Thousands of children deserve to be more
  1. **Feed-Forward:** thousands of children deserve to be more than another statistic
  2. **RNN:** thousands of children deserve to be more to change individual votes or vote totals
  3. **GRU:** thousands of children deserve to be more than another statistic
- The rural region said it is coming to
  1. **Feed-Forward:** the rural region said it is coming to close the inquiry into russia's election meddling
  2. **RNN:** the rural region said it is coming to rural new brunswick
  3. **GRU:** the rural region said it is coming to a period of celibacy during that time

- Wiretaps reveal conversations between
  1. **Feed-Forward:** wiretaps reveal conversations between president donald trump were still talking
  2. **RNN:** wiretaps reveal conversations between alexander torshin and alexander romanov a convicted russian money launderer
  3. **GRU:** wiretaps reveal conversations between the public and when they should be speaking english
- Activists are protesting the new law regarding
  1. **Feed-Forward:** activists are protesting the new law regarding its own is many back in later
  2. **RNN:** activists are protesting the new law regarding the justice department
  3. **GRU:** activists are protesting the new law regarding a culture of violence that we
- The former president shared his
  1. **Feed-Forward:** the former president shared his songs while
  2. **RNN:** the former president shared his words reportedly uttered in private
  3. **GRU:** the former president shared his longtime girlfriend anna eberstein tied the knot in a civil ceremony
- California is the first state to legalize
  1. **Feed-Forward:** california is the first state to legalize medical cannabis becomes the most populous to allow recreational use
  2. **RNN:** california is the first state to legalize trump administration in recent weeks
  3. **GRU:** california is the first state to legalize sports betting from a government database

Observations

- **Feed-Forward Network**: The model produces syntactically valid but semantically incoherent text, with strong topic drift and memorized n-gram mixing. This is due to its fixed 5-word context window and lack of sequential memory.

  Example:

  ```
  the government failed to locate the immigrant children killed up
  the case of age was black man who went to be in the united states...
  ```

  The sentence mixes unrelated concepts (immigration, crime, politics, performance) with no semantic consistency, showing local pattern stitching rather than sequence modeling.

- **Simple RNN**: The RNN shows better grammatical structure and short-range coherence, but still suffers from semantic instability and memorization artifacts.

  Example:

  ```
  wiretaps reveal conversations between alexander torshin and alexander
  romanov a convicted russian money launderer
  ```

This output is coherent and factual but directly reproduces training content, indicating memorization rather than generalization.

Another example of semantic drift:

```
trump gives dems an ass kicking for not having a surprise serenade
from a mariachi band
```

The structure is fluent, but the content is logically absurd.

- **GRU**: The GRU produces the most coherent and semantically stable outputs, with better topic continuity and structured completions.

  Example:

  ```
  thousands of children deserve to be more than another statistic
  ```

  The model correctly completes the sentence with a meaningful, semantically consistent continuation.

  Another example:

  ```
  the rural region said it is coming to rural new brunswick
  ```

  This shows strong contextual alignment and realistic geographic coherence.

Conclusion. The qualitative evaluation reveals a clear hierarchy in generative quality:

$$Basic\_NN \ll RNN < GRU$$

The Feed-Forward model behaves as a local statistical predictor with no sequence understanding. The RNN captures short-term dependencies but shows memorization and instability. The GRU demonstrates superior contextual memory, semantic coherence, and sequence stability, confirming the effectiveness of gating mechanisms for long-range dependency modeling.