RNNs in Language Translation
Performance Analysis

Nelson Dufitimana, Kelvin Darfour, Bereket Nigussie, Wilbert Fundira

From Rule Based Systems to RNNs

Brief History

1950s–1980s: Early Machine Translation(**EMT**)

Rule-based systems and **dictionaries** were used for early machine translation.

1990s–2000s: Statistical Machine Translation (**SMT**)

SMT utilized **statistical models** based on large bilingual corpora

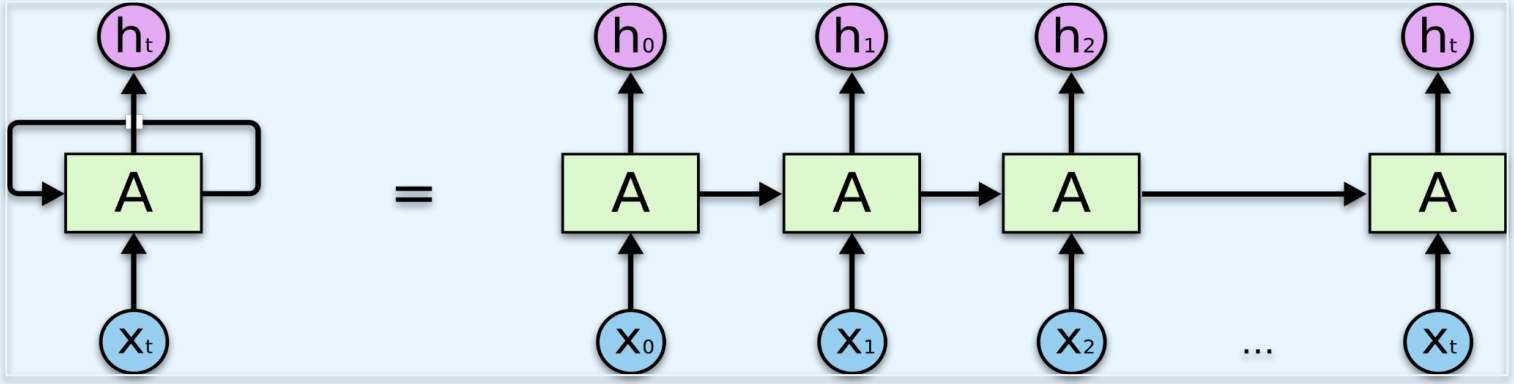
mid 2010s: Neural Machine Translation (**NMT**)

Recurrent Neural Networks (**RNNs**) introduced for NMT

Recurrent Neural Networks

RNNs attempt to **mimic** human thought:

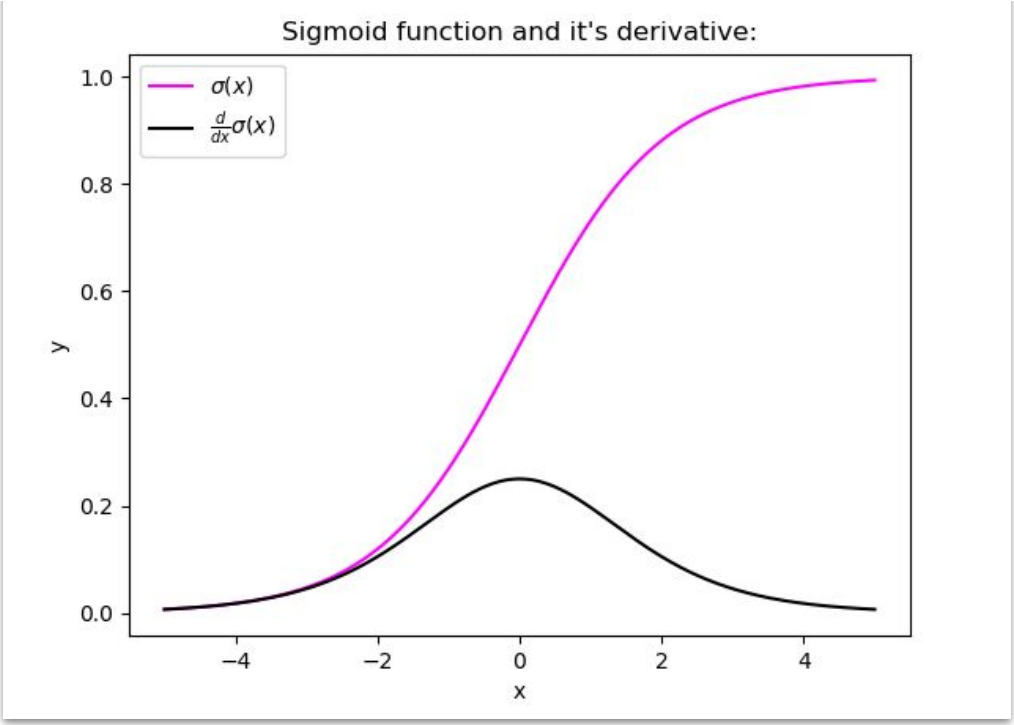
- Sequential Processing
- Memory
- Learning Over Time
- Feedback Loops



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

OH NO...

Standard RNNs struggled with **exploding** gradient or **vanishing** gradient



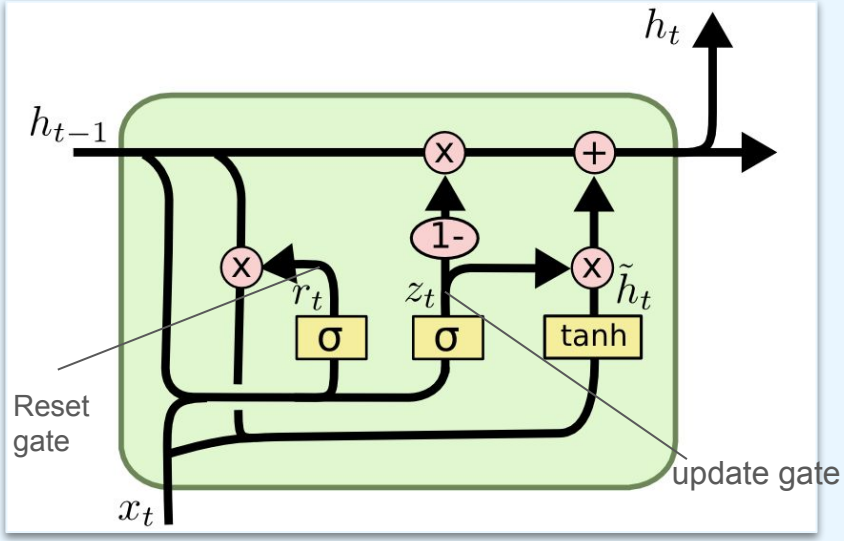
https://www.analyticsvidhya.com/blog/2021/06/the-challenge-of-vanishing-exploding-gradients-in-deep-neural-networks/

Fix?

Add Gates to RNNs

GRUs

- What are **GRUs** (Gated Recurrent Unit)
- How do they **solve** these problems?



Again?

Standard RNNs performed **poorly** when dealing with **long term** dependencies

“The clouds are in the ____”.

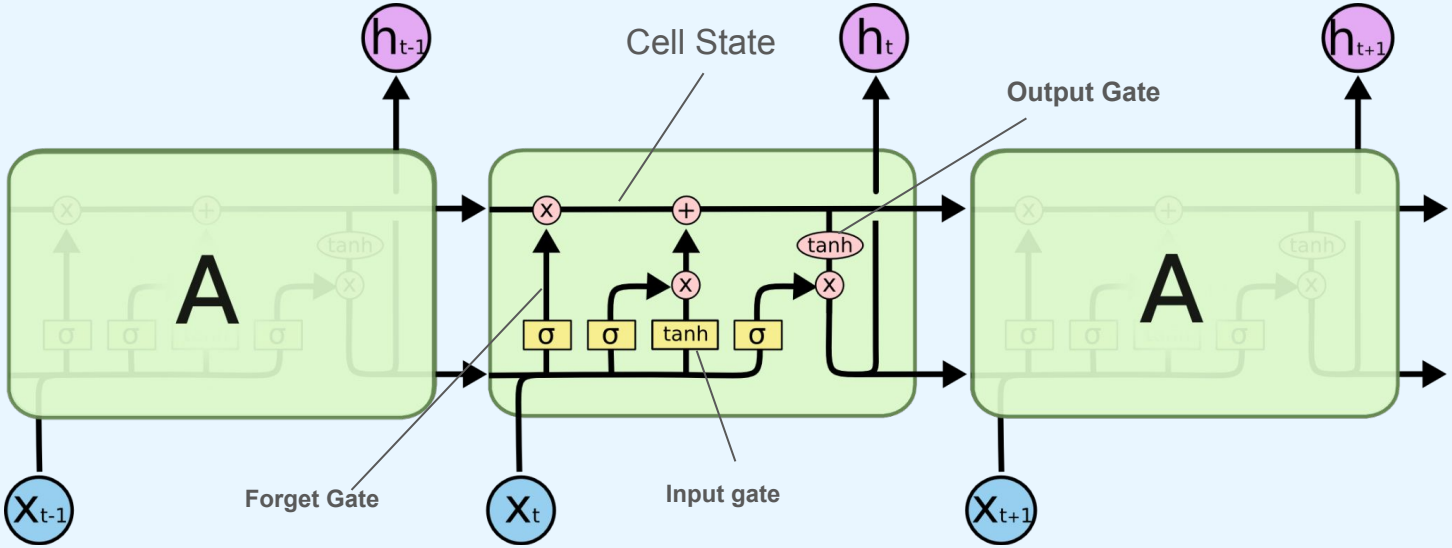
easy...

“**Jane** was born in **Kerala**. Jane used to play for the women’s football team and has also topped at state-level examinations. **Jane** is very fluent in ____.”

example source

LSTMs

How do LSTMs solve the **long term** dependency issue?



That was a little Introduction into our project

What’s next?

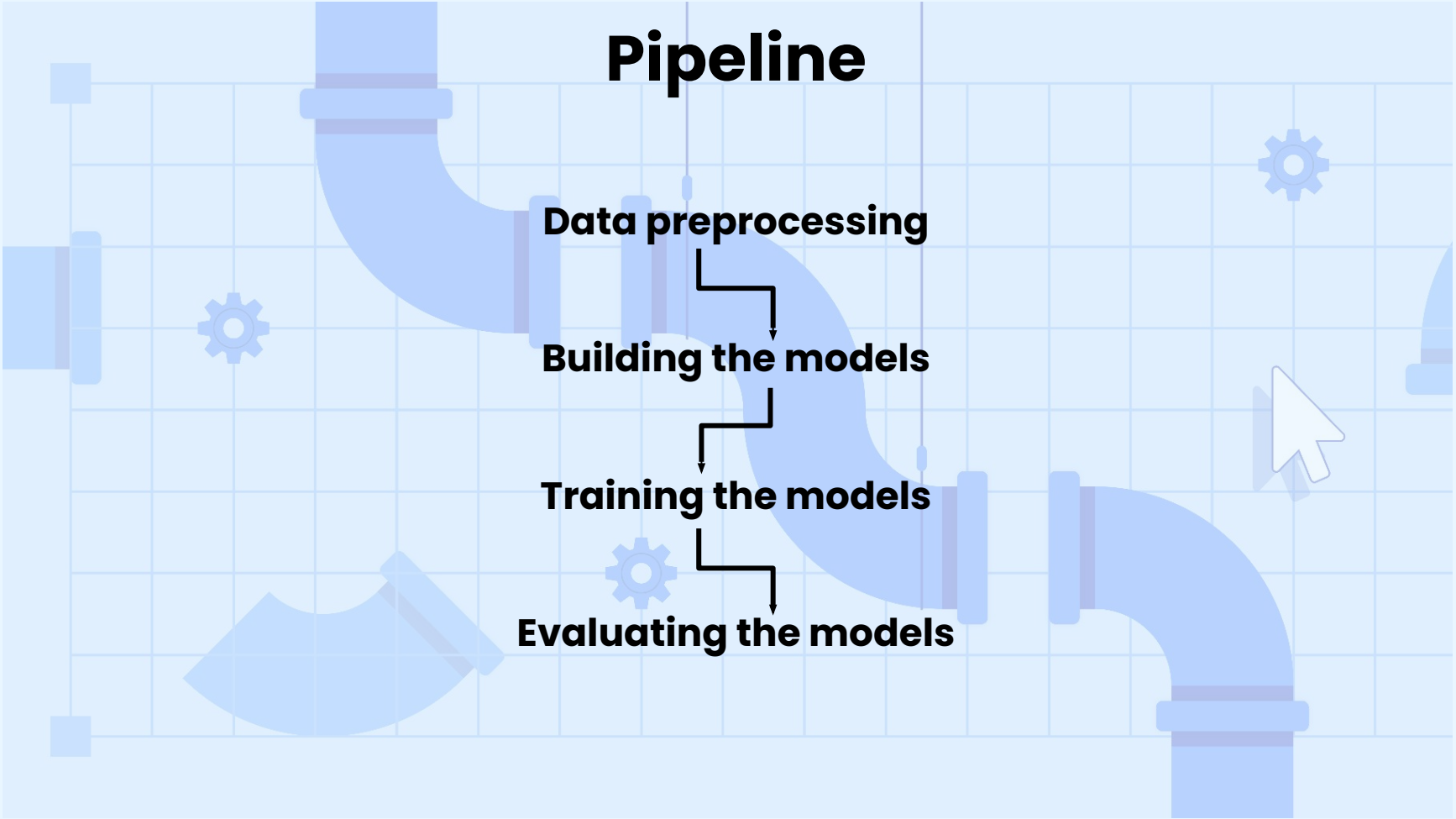
Our Hypothesis

"**LSTMs** perform **better** than both GRUs and **simple** RNN models" - for language translation tasks"

LSTM **>** GRU**(with gates) >** Simple RNN**(no gates)**

Design

Pipeline


Data preprocessing
Building the models
Training the models
Evaluating the models

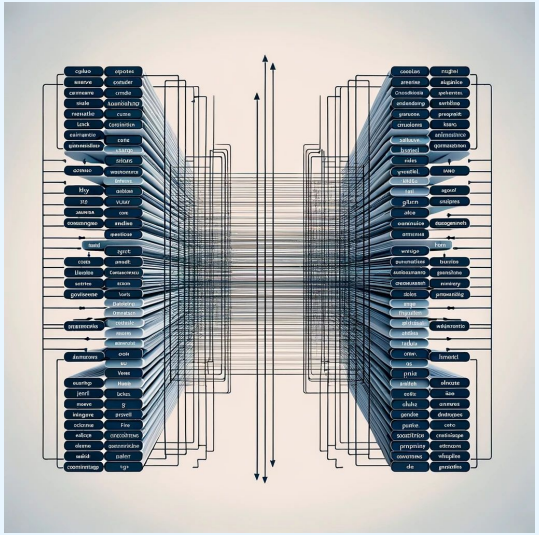
Building the Models

Initial Setup

- **6** models

Optimizer

- **Adam**



Activation Functions

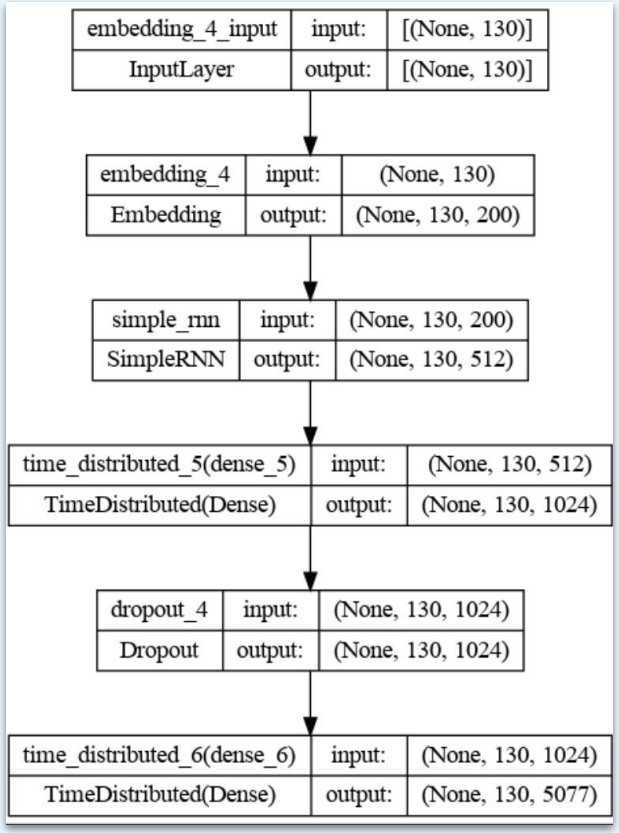
- **Softmax**
- **Relu**

Initial Metric

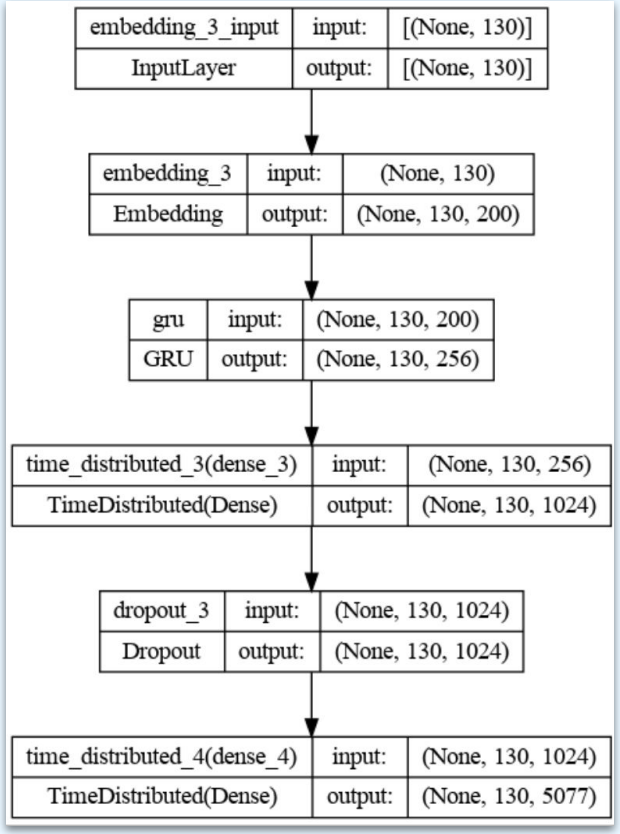
- **Accuracy**

Models

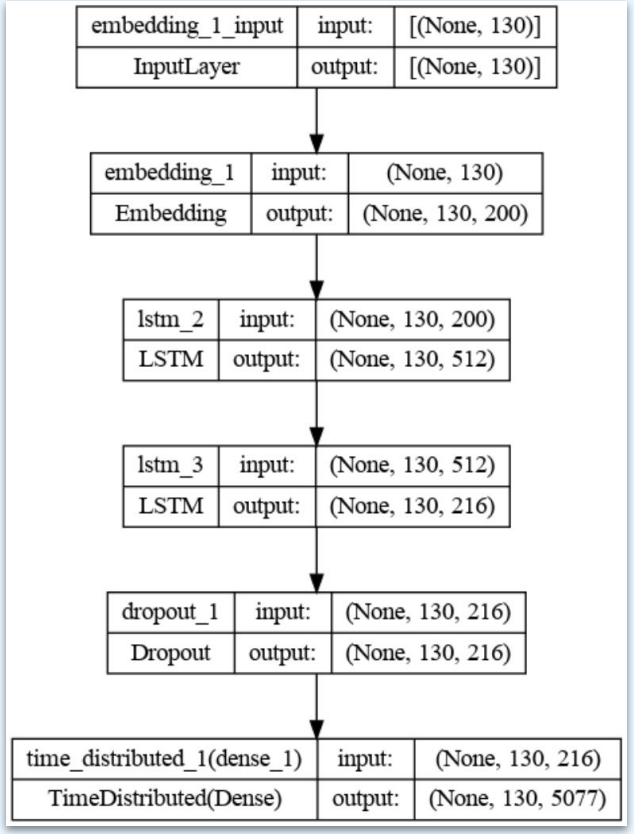
Simple



GRU

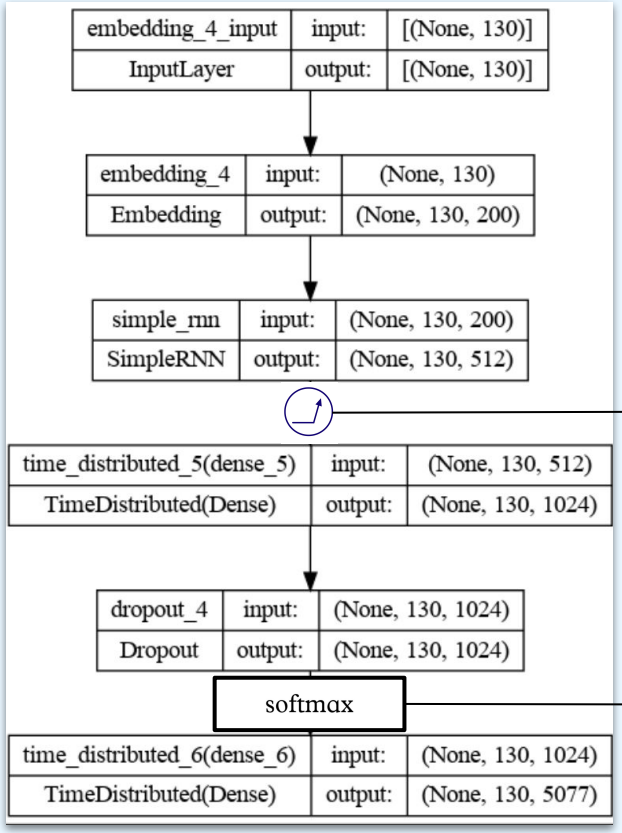


LSTM

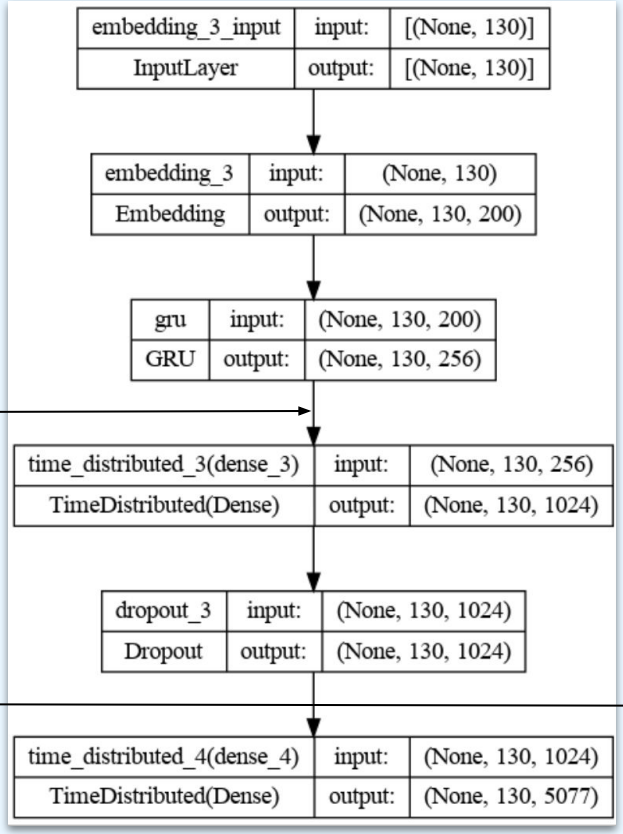


Models

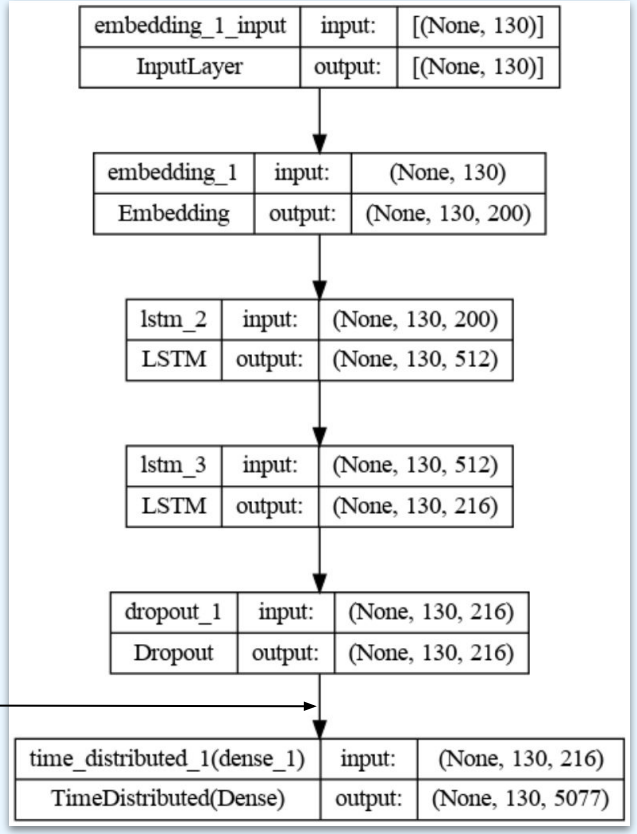
Simple



GRU



LSTM



Data Preprocessing

Dealing with data is challenging, but language data is **quite** challenging!

1. **Downsampling the vocabulary**

- Larger vocabulary corpus affect model training time & accuracy

  e.g. **4000** words with **3500** unique words
- Balance the **unique words** to **total words** ratio
- Filter out words with **less** frequency, setting the frequency **threshold**

Data Processing

Data Preprocessing

2. Tokenization

Convert the **text** to **numerical** values to allow the neural network to perform operations on the input data.

```
{'the': 1, 'quick': 2, 'a': 3, 'brown': 4, 'fox': 5, 'jumps': 6, 'over': 7, 'lazy': 8, 'dog': 9, 'by': 10, 'jove': 1
1, 'my': 12, 'study': 13, 'of': 14, 'lexicography': 15, 'won': 16, 'prize': 17, 'this': 18, 'is': 19, 'short': 20, 's
entence': 21}

Sequence 1 in x
  Input:  The quick brown fox jumps over the lazy dog .
  Output: [1, 2, 4, 5, 6, 7, 1, 8, 9]
Sequence 2 in x
  Input:  By Jove , my quick study of lexicography won a prize .
  Output: [10, 11, 12, 2, 13, 14, 15, 16, 3, 17]
Sequence 3 in x
  Input:  This is a short sentence .
  Output: [18, 19, 3, 20, 21]
```

Image courtesy of: [Language Translation with RNNs. Build a recurrent neural network (RNN)... | by Thomas Tracey | Towards Data Science]

Data Preprocessing

3. Padding:

Ensures **equal dimensions** for input sequences in the model.

```
Sequence 2 in x
  Input:  [10 11 12  2 13 14 15 16  3 17]
  Output: [10 11 12  2 13 14 15 16  3 17]  no padding
Sequence 3 in x
  Input:  [18 19  3 20 21]
  Output: [18 19  3 20 21  0  0  0  0  0]  padding
```

Image courtesy of: [Language Translation with RNNs. Build a recurrent neural network (RNN)... | by Thomas Tracey | Towards Data Science]

Results & Analysis

Performance Metrics

- How **good** are translation output?

  e.g. “The ball is blue” **vs**. “The ball has a blue color”.

- Accuracy as metric might **not** be helpful here.

- **Bleu Score:** Precision based metric with the idea: the **closer** the predicted sentence is to the human-generated target sentence, the **better** it is.

Target: We are in a machine learning class

Predicted Sentence: We we is in machine learning

P1 = 4/6

Image courtesy of Foundations of NLP Explained — Bleu Score and WER Metrics | by Ketan Doshi | Towards Data Science

Results on Test Set

Model	Accuracy	Bleu
Simple RNN	0.93	0.60
GRU	0.94	0.57
LSTM	0.94	0.61

Challenges/Limitations

- The Curse Of Dimensionality
    - **Dataset** size did not scale with **vocabulary** size
- Limited Compute **Resources**
    - Run out of **memory** for very large datasets
- Long Iteration Time
    - Took about **13** hours to train the models.
- Hyperparameter **Tuning**
    - Optimizer, Loss function, Learning rate, Dropout rate, Size of layers

Questions?