# TEXT-TO-TEXT TRANSLATION

using Deep Learning

BENGUEZZOU Idriss - KHALYFA Fatiha -
MEZIANE Ghilas

# Project Outline

| | 11-11-23 > 21-11-23 | 21-11-2023 > 30-12-23 | 01-01-24 > 10-01-24 | 10-01-24 > 15-01-2024 | 15-01-24 > 20-01-24 | 20-01-24 > 23-01-24 | 24-01-4 > 29-01-24 |
|---|---|---|---|---|---|---|---|
| Data Exploration & Preprocessing | | | | | | | |
| Baseline Model Implementation | | | | | | | |
| Baseline Bidirectional & Embedding | | | | | | | |
| Baseline Encoder/Decoder + Attention Mechanism | | | | | | | |
| Baseline Encoder/Decoder + Multihead Attention | | | | | | | |
| Pre-trained T5 model | | | | | | | |
| Torch Transformer Model | | | | | | | |
| Report + Presentation | | | | | | | |

# Exploratory Analysis

No missing values

**Average sentence length**

- 19.37 Words for English
- 21.51 Words for French

**Total Words Count :**

- 6 627 178 English words
- 7 357 642 French words

**Some translations are wrong**

he concludes however → which one do we choose

therefore sudan is an afro-arab country → so which one do we belong to more

as i have presented ethnically speaking we sudanese are mainly african but culturally we are more arab than african thanks to arabization → in order to give an answer i have to ask another question which one plays a bigger role in forming one's identity

# Exploratory Analysis

No missing values

**Average sentence length**

- 19.37 Words for English
- 21.51 Words for French

**Total Words Count :**

- 6 627 178 English words
- 7 357 642 French words

## After pre-processing

**Vocabulary size:**

- 371 344 French words
- 345 783 English words

**Vocabulary size:**

- 178 677 French words
- 159 842 French words

# Preprocessing

## Lowercasing

To ensure uniformity in text, we transform the sentence in lowercase

## Contractions Expansion

To handle contractions appropriately, we expand the english contraction, for example :

can't ⟶ cannot
don't ⟶ do not

## Unicode Normalization

To handle different Unicode representations, we normalize them.

## Removing Non-ASCII Punctuation

To clean the text from non-ASCII characters.

## Removing Special Characters

## Removing Extra Spaces

METRIC

# Evaluation metrics

### BLEU (Bilingual Evaluation Understudy) Score

Score to assess our machine translation models.

### Smoothing Function

Fixes BLEU score issues when some words in the translation aren't found in the reference. It does this by giving a small chance to new words.

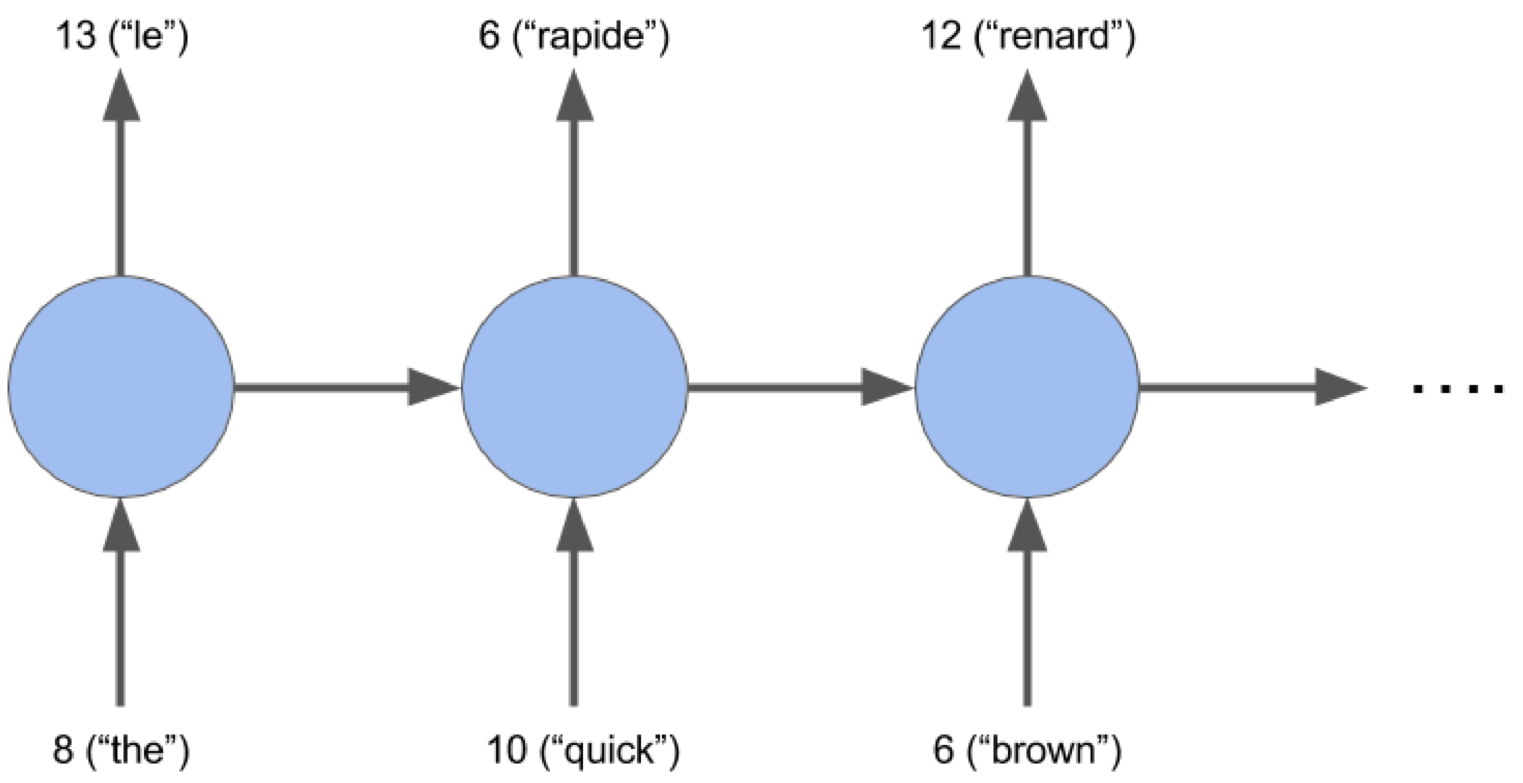
**We used method0 smoothing function**

RNN

# Baseline Model

## 1. Data preparation

- tokenization
- encode sequence
- decode sequence




## 2. Simple Seq2Seq model

- **Input layer,**
- a **LSTM layer,**
- and a **Dense Layer**

RNN

# Evaluation of the Baseline Model

## Training parameters

| | |
|---|---|
| Loss Function | Sparse Categorical Crossentropy |
| Optimizer | RMSprop optimizer (lr = 0.001) |
| Epochs | 10 |
| Batch size | 128 |
| LSTM units | 64 |
| Vocabulary size | All |
| Max sequence length | 50 |

Accuracy : ~20%

- Accuracy : > ~70%
- Val Accuracy : ~10%

Overfitting

"je suis un homme"

“the the the”

"je suis un chat"

“i am the  the”

**Model BLEU Score on Test Data**

**1.006 × 10−23**

RNN

# Tuning of the Baseline Model

"je suis fatigué"

“i am the”

**Model BLEU Score on Test Data**
**0.0749**

## Training parameters

| | |
|---|---|
| Loss Function | Sparse Categorical Crossentropy |
| Optimizer | RMSprop optimizer (lr = 0.001) |
| Epochs | 30 |
| Batch size | 128 |
| LSTM units | 64 |
| Vocabulary size | 5000 |
| Max sequence length | 20 |

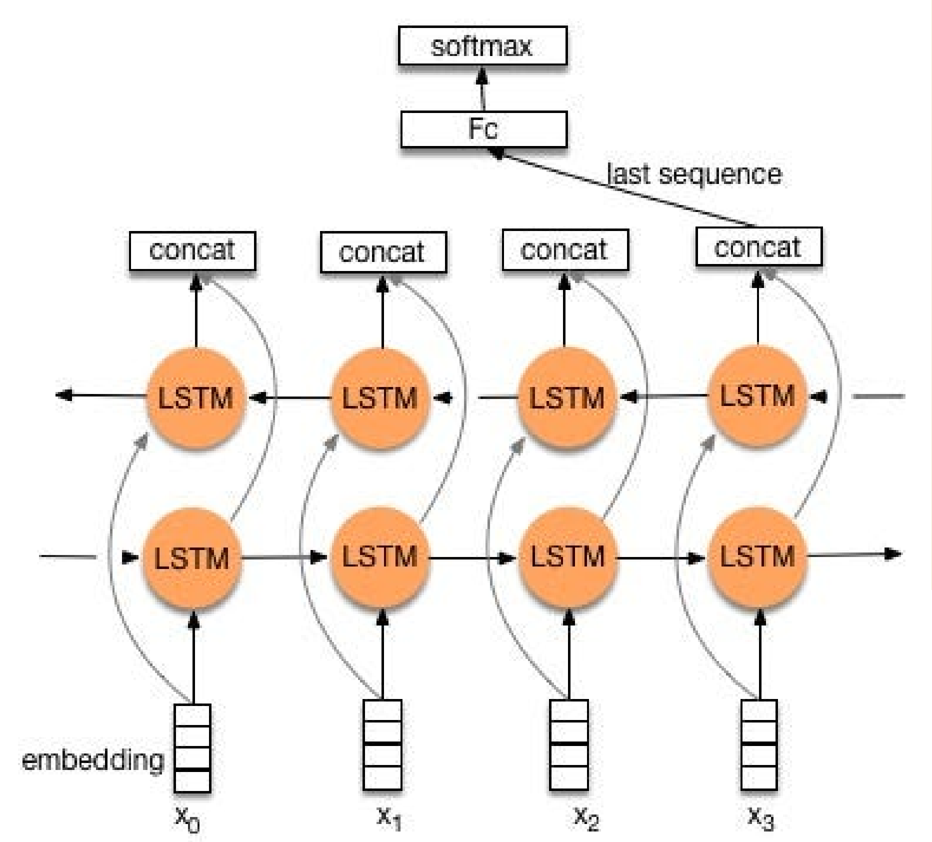
RNN

# Baseline Bidirectional & Embedding

Enhancement of our baseline Seq2Seq model with bidirectional LSTM and embedding layers

## Architecture

- **Input layer :** 20 tokens per sequence.
- an **Embedding layer**
- an **Bidirectional layer**
- and a TimeDistributed Dense Layer



# RNN

# Evaluation Baseline Bidirectional

| | Training parameters | Tuning |
|---|---|---|
| Loss Function | Sparse Categorical Crossentropy | Sparse Categorical Crossentropy |
| Optimizer | RMSprop optimizer (lr = 0.001) | RMSprop optimizer (lr = 0.001) |
| Epochs | 20 | 20 |
| Batch size | 128 | 64 |
| LSTM units | 64 | 64 |
| Vocabulary size | All | 5000 |
| Max sequence length | 50 | 20 |
| Embedding dimension | 216 | 128 |

We tried different embedding dimensions, ranging from 64 to 1024.

**Model BLEU Score on Test Data**
**0.00912**

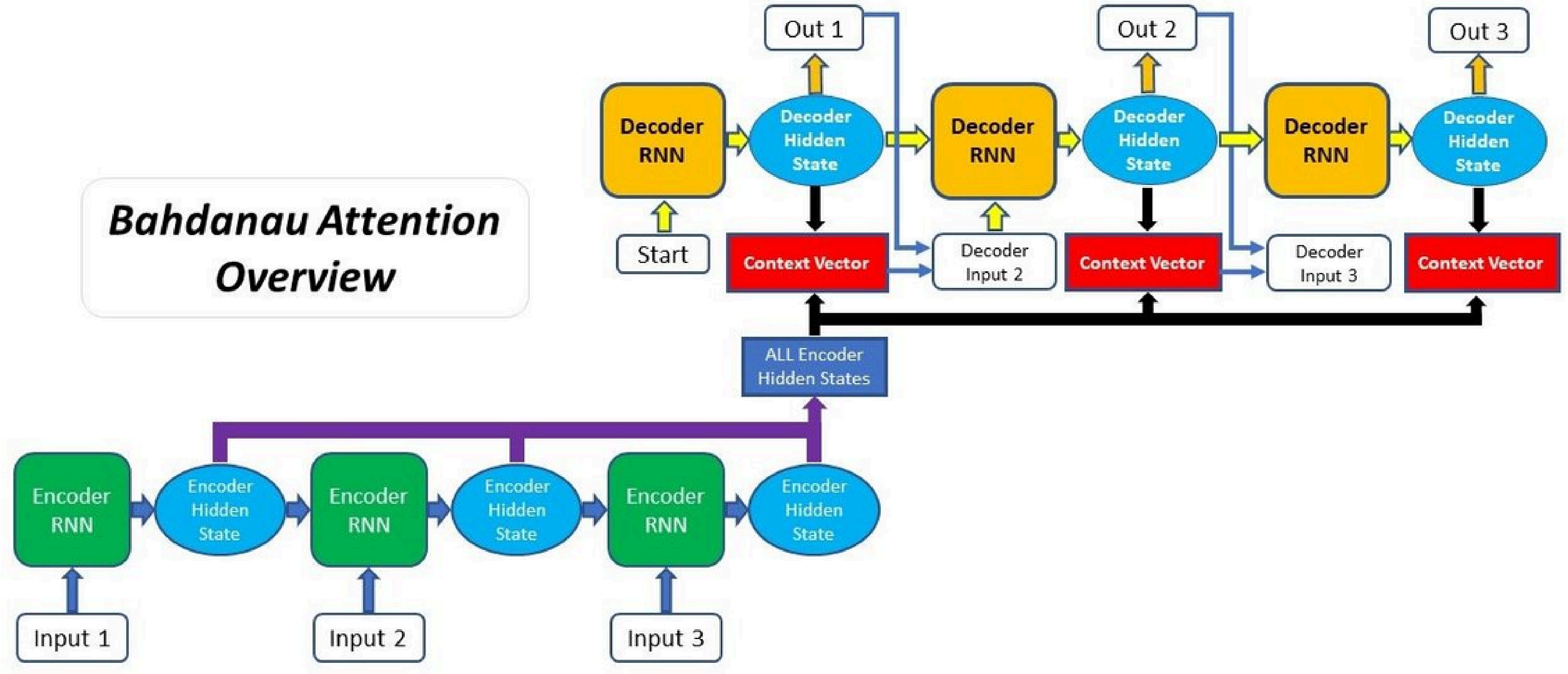
**Model BLEU Score on Test Data**
**0.08133**

RNN

# Baseline Encoder/Decoder with Attention Mechanism

## Bahdanau Attention

Is one of the early attention mechanisms, and it has been influential in the development of more advanced attention mechanisms, such as the Transformer's self-attention mechanism.




Embedding dimension = 256

https://medium.com/geekculture/sentence-correction-using-recurrent-neural-network-6321527ee08b

GRU

Embedding

Encoder

**Bahdanau Attention**

GRU

Embedding

Decoder

RNN

# Baseline Encoder/Decoder with Attention Mechanism

## Training parameters

| | |
|---|---|
| **Loss Function** | Sparse Categorical Crossentropy |
| **Optimizer** | Adam |
| **Epochs** | 30 |
| **batch size** | 64 |
| **max sequence length** | 20 |
| **vocabulary size** | 5 000 |

We were confronted with a significant challenge in terms of training time and memory usage.

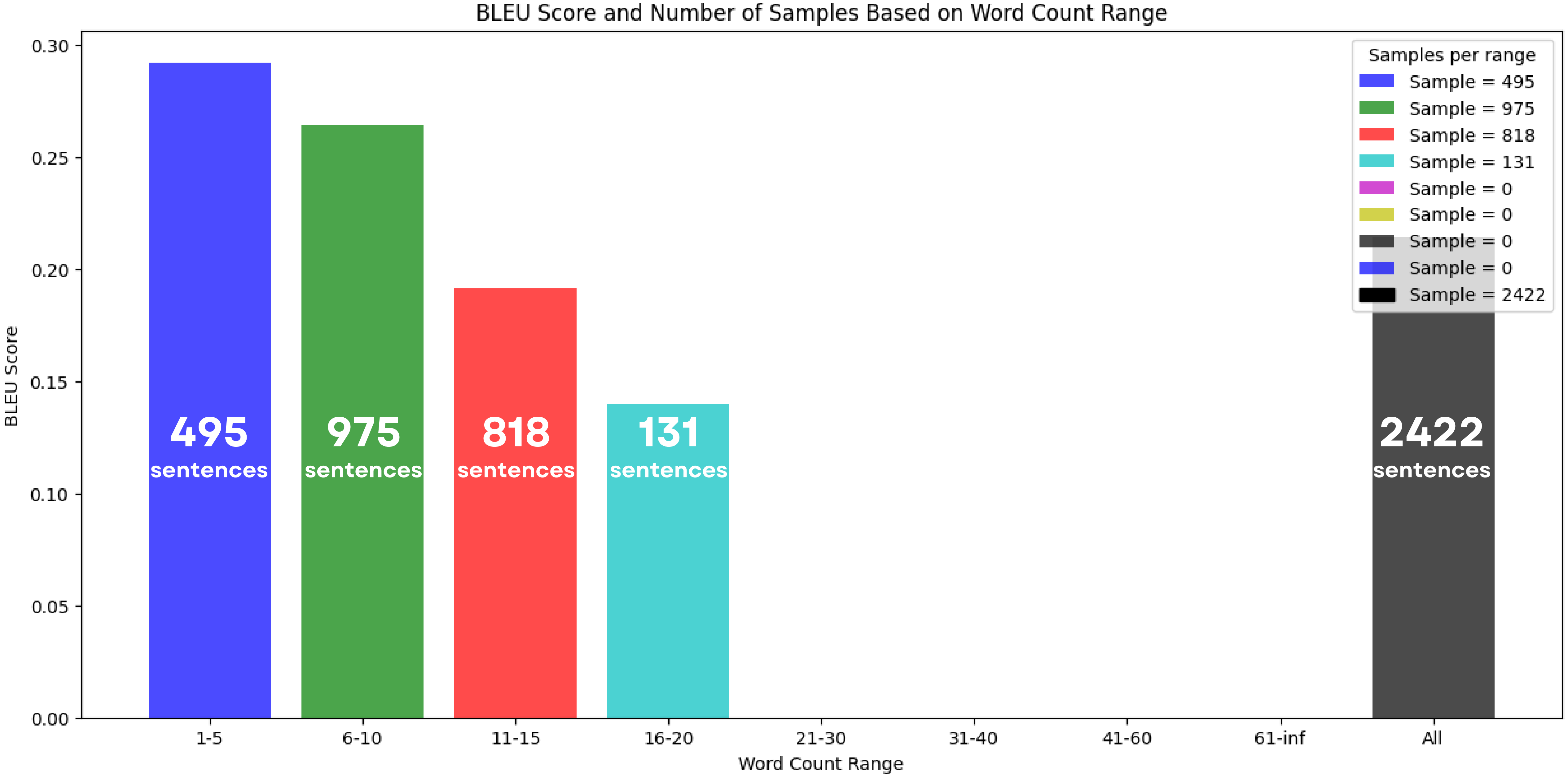
**Limit the training to**

**50 000 sentences**

RNN

# Baseline Encoder/Decoder with Attention Mechanism

BLEU score ~ 0.2140




**The model performs better on shorter sentences, and the BLEU score decreases as the length of the sentences increase**

RNN

# Encoder Decoder with Multi Head Attention Mechanism

## Architecture

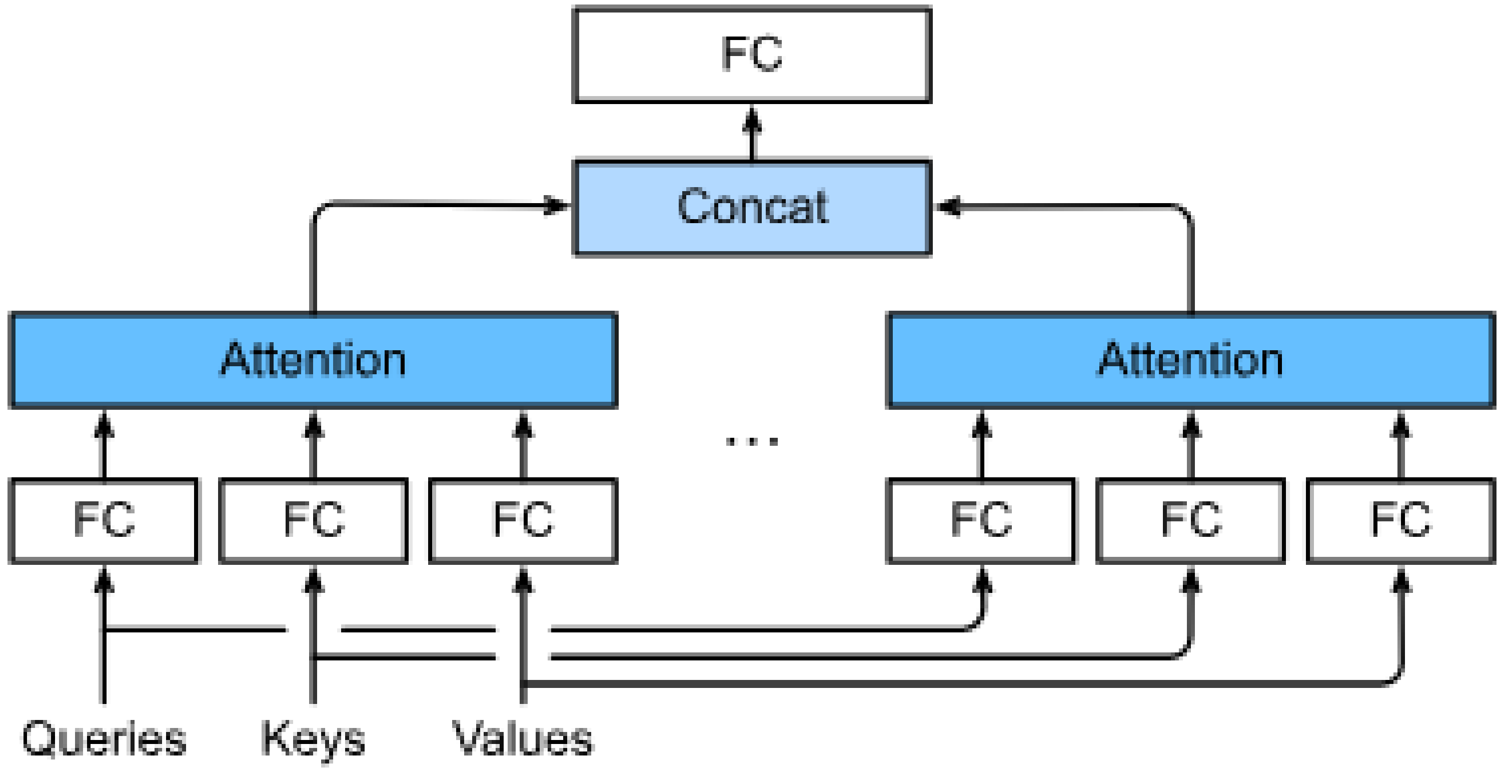
- Bidirectional Encoder
  - Embedding layer : Embedding size 256
  - Bidirectional GRU layer : 1024 units
- MultiHead Attention mechanism
- Decoder similar to encoder : GRU layer

The decoder is designed to work in tandem with the MultiHead Attention mechanism

Inspired by the tutorial (https://www.tensorflow.org/text/tutorials/nmt_with_attention?hl=fr).

RNN

# Encoder Decoder with Multi Head Attention Mechanism




**Source:** https://d2l.ai/chapter_attention-mechanisms-and-transformers/multihead-attention.html

## MultiHead Attention mechanism

The MultiHead Attention mechanism allows a model to attend to multiple parts of a sequence simultaneously. This is achieved by splitting the input sequence into multiple "heads," each of which is processed in parallel. Each head has its own set of weights, which allows the model to learn different relationships between the input sequence and the output.

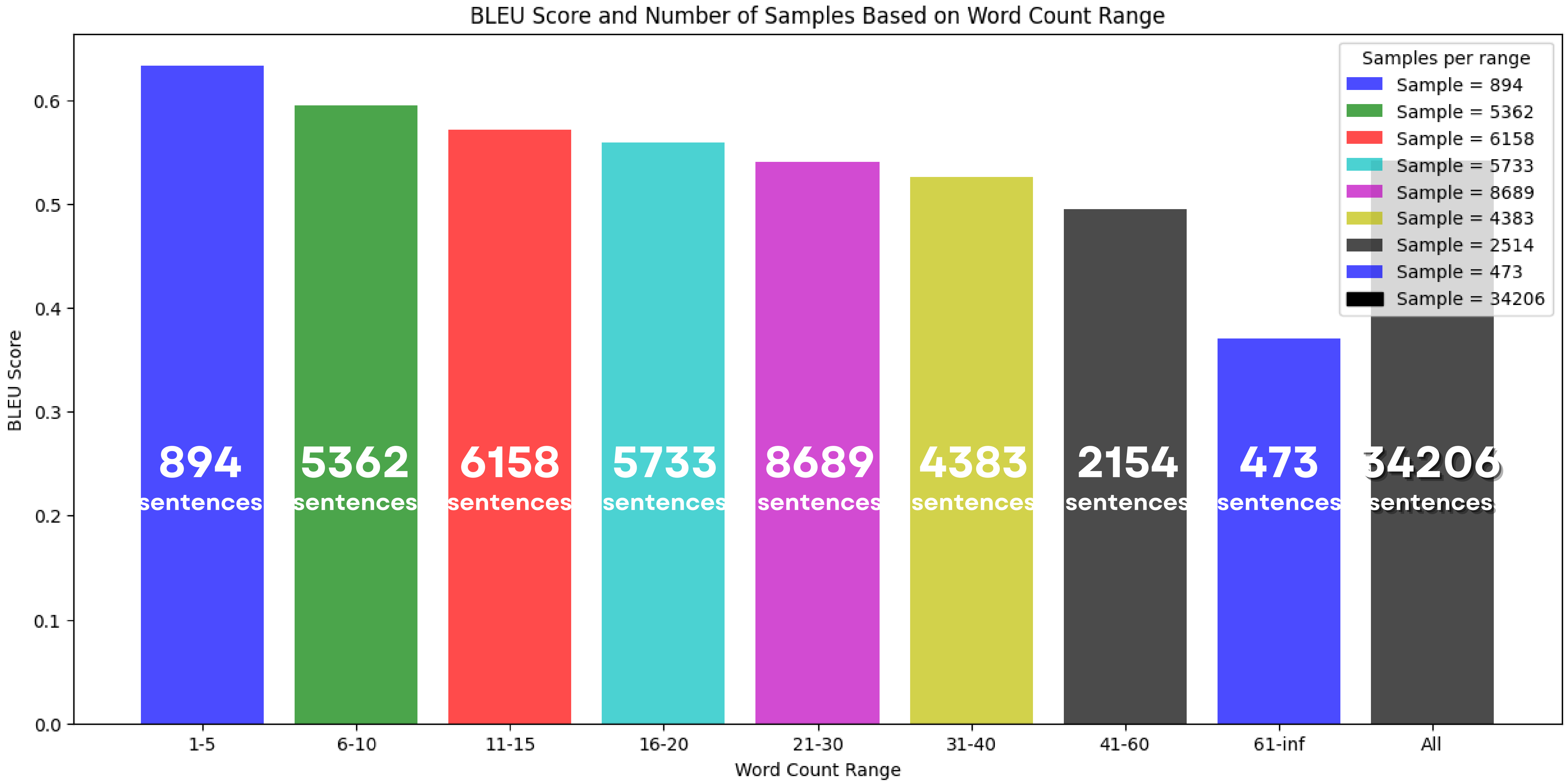
- useful in tasks where long-term dependencies are important.
- can better capture these dependencies and generate more accurate outputs.

RNN

# Encoder Decoder with Multi Head Attention Mechanism

## Training parameters

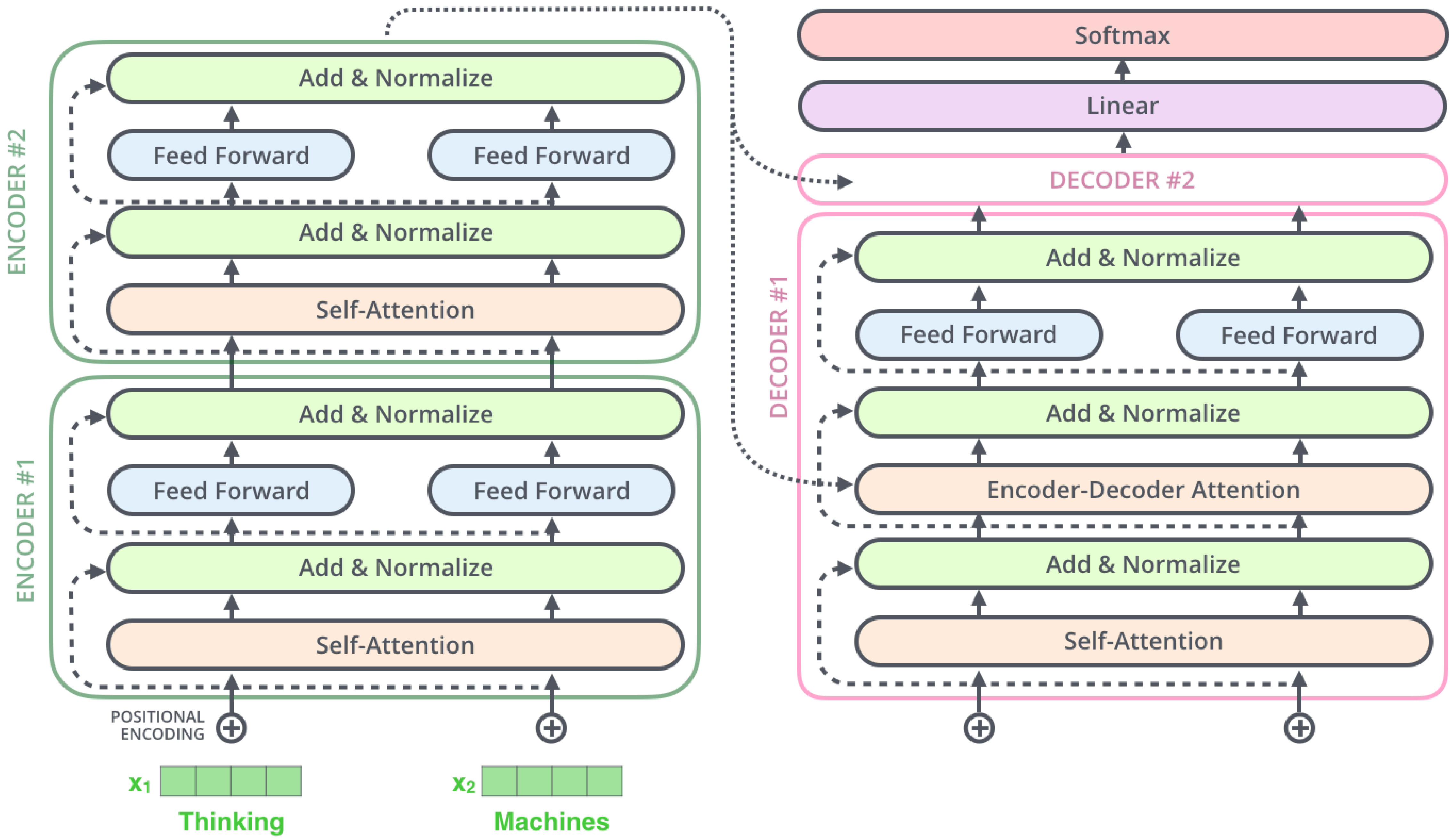
| | |
|---|---|
| **Loss Function** | Sparse Categorical Crossentropy |
| **Optimizer** | Adam |
| **Epochs** | 100 |
| **batch size** | 64 |
| **max sequence length** | ALL |
| **Dataset size** | ALL |
| **vocabulary size** | 5 000 |




- **MultiHead Attention Blue Score = 2 x Bahdanau Blue Score**
- **Good score for long sequences**

Transformer

# Transformer T5 Model (Text-to-Text Transfer Transformer)

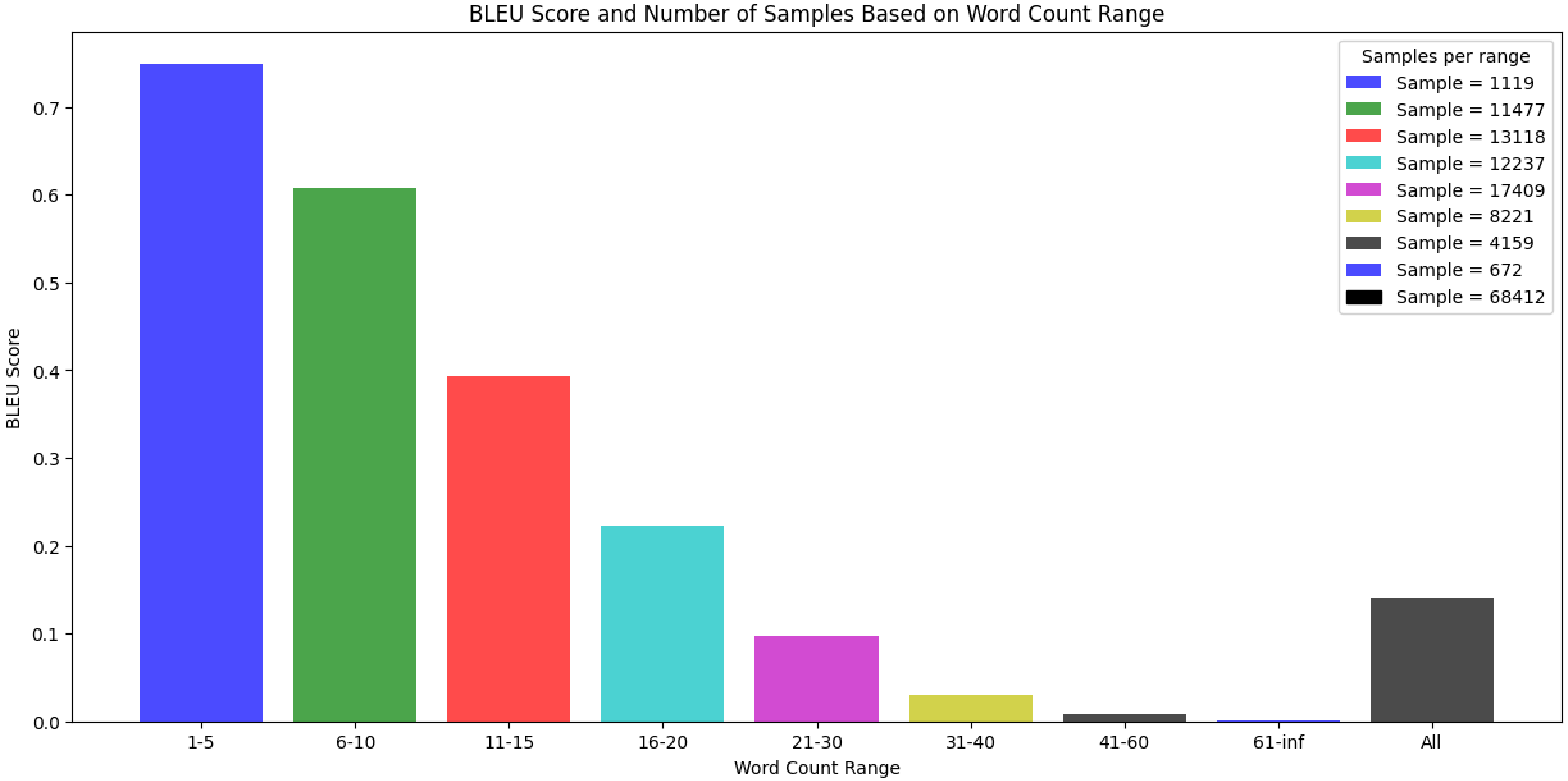



https://pgresia.medium.com/making-pytorch-transformer-twice-as-fast-on-sequence-generation-2a8a7f1e7389lignes dans le corps du texte

Transformer

# Transformer T5 Model (Text-to-Text Transfer Transformer)

- We used the small variant of t5.
- # of (layers):
    - Encoder 6
    - Decoder: 6
- Embedding dimension: 512
- size of feed-forward layers: 2048
- Attention heads : 8
- Dropout rate: 0,1

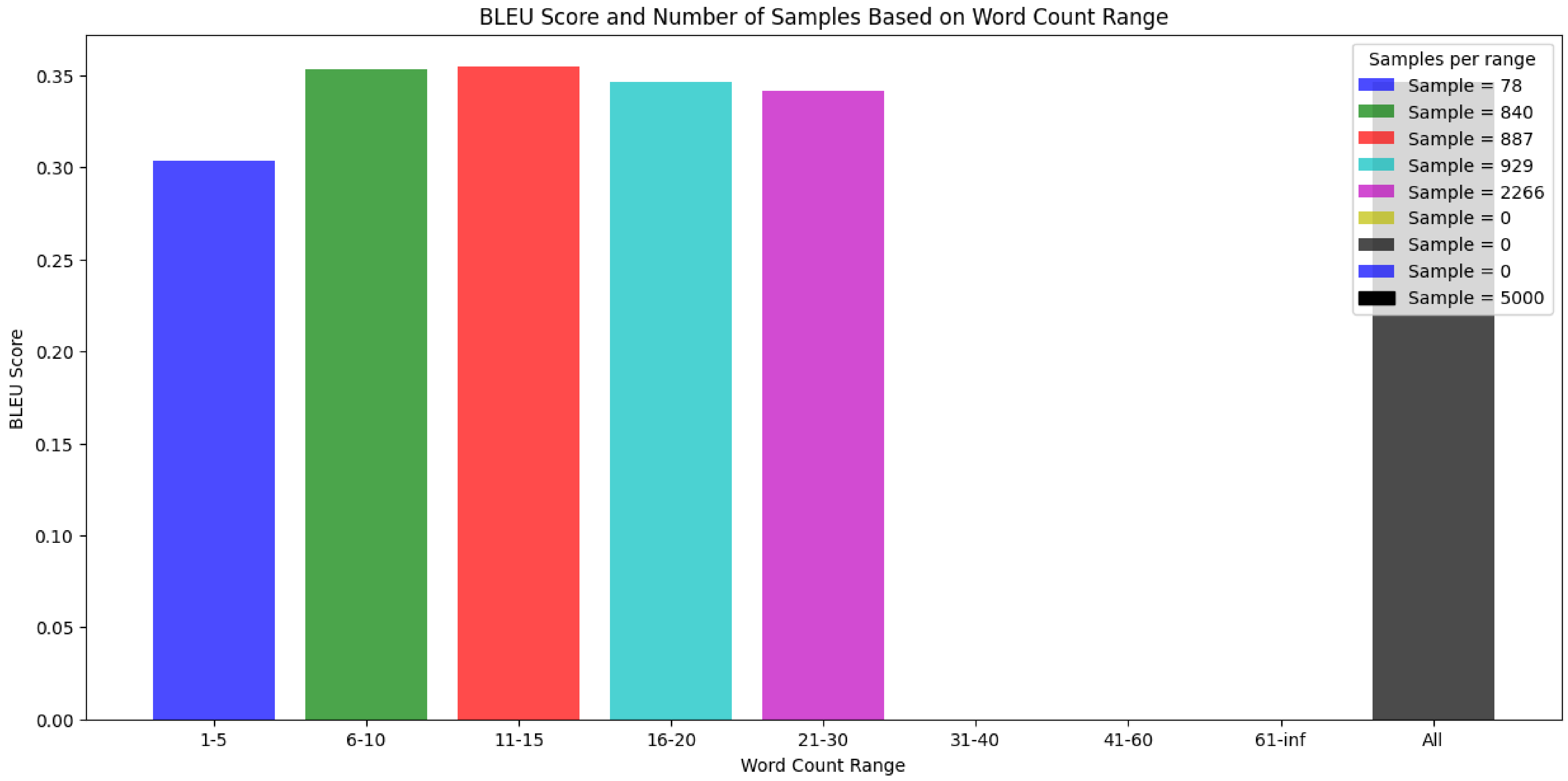



- Best result so-far
- Pre-trained models have a maximum token limits

Transformer

# Torch Transformer Model

- # of (layers):
  - Encoder 3
  - Decoder: 3
- Embedding dimension: 192
- size of feed-forward layers: 192
- Attention heads : 6
- Dropout rate: 0,1




- This model is not pre-trained.
- consistent performance even for long sentences.

# Conclusion

- Simple RNN : Very limited , low performance.
- Attention Mechanism: Highly increases the performance of the model specifically the Multi Head Attention one.
- Transformers yield great results, even with long sentences.

- We faced some challenges related to memory, time constraints, and GPU usage.