Document Summarization

Ratnadeep Mitra & Sooeun Oh

ANLY 590 | Dr. Keegan Hines | Georgetown University | M.S. Analytics

Project Overview / Abstract

Sequence to sequence model has been developed for translation into different language. Using this tool, the research aims to develop a text summarization model that summarizes the news articles into a headline.

Problem Statement

Generate document summaries (headlines) from original texts (news articles).

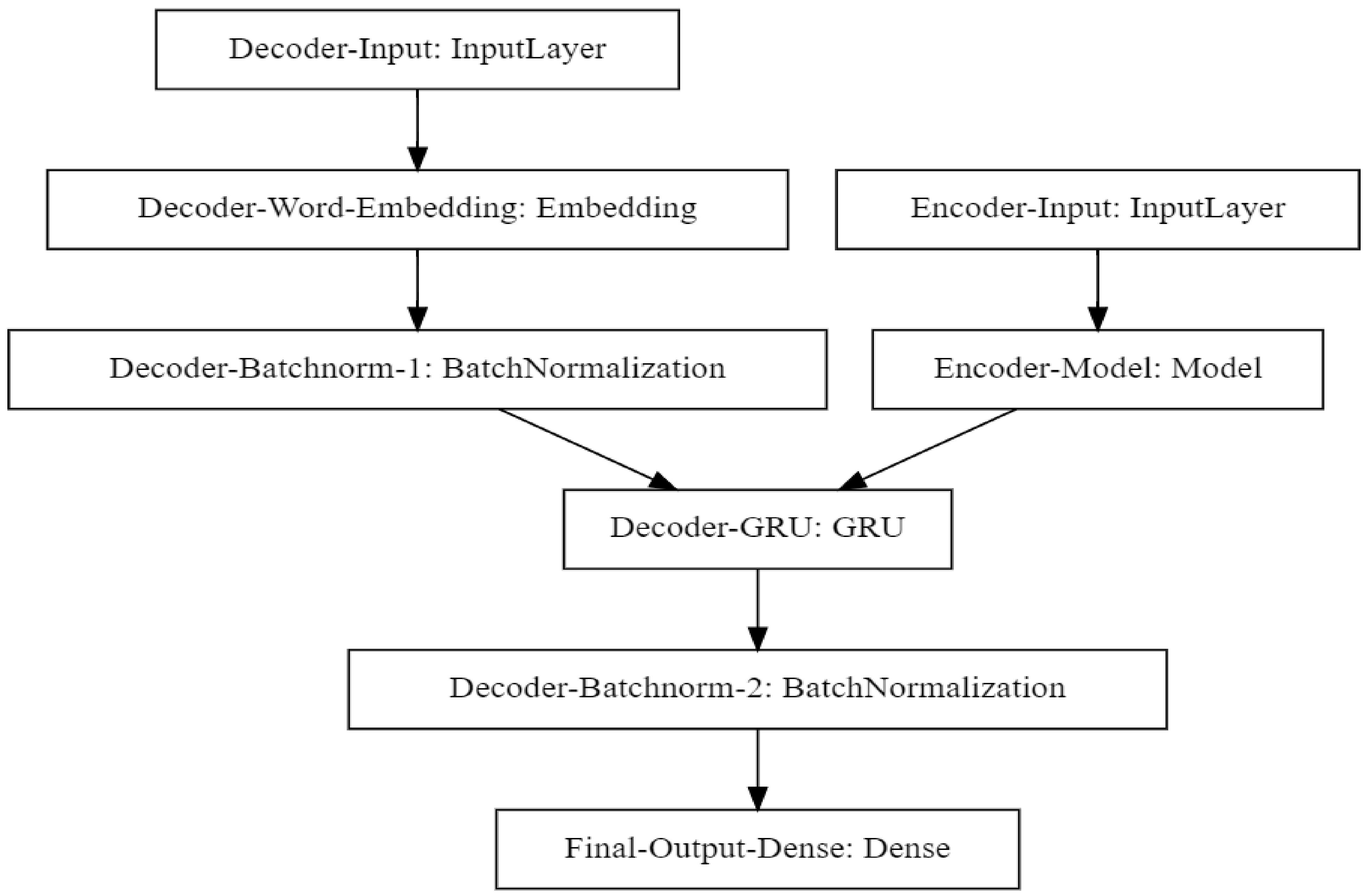
Related Work

- In approaching the problem, we considered the findings of several papers that provided better understanding of the research already done.
- From these studies, we concluded to explore sequence to sequence model for our research.
- Sequence to sequence (Seq2Seq) learning has been used for abstractive and extractive summarization. In most recent studies, researchers proposed a novel document-context based Seq2Seq models using RNNs for summarizations.

Dataset and Pre-Processing

- This research uses the news data obtained from Kaggle, which comprises 142,136 text examples from 15 major American publications.
- Due to the very large size of dataset, a 20k-sample dataset was created with only news body and headline for efficient execution.
- The data was pre-processed with ***ktext*** package that performed data-cleaning, tokenization, padding, and truncation.

Model Architecture



Decoder-Input: InputLayer

Decoder-Word-Embedding: Embedding

Encoder-Input: InputLayer

Decoder-Batchnorm-1: BatchNormalization

Encoder-Model: Model

Decoder-GRU: GRU

Decoder-Batchnorm-2: BatchNormalization

Final-Output-Dense: Dense

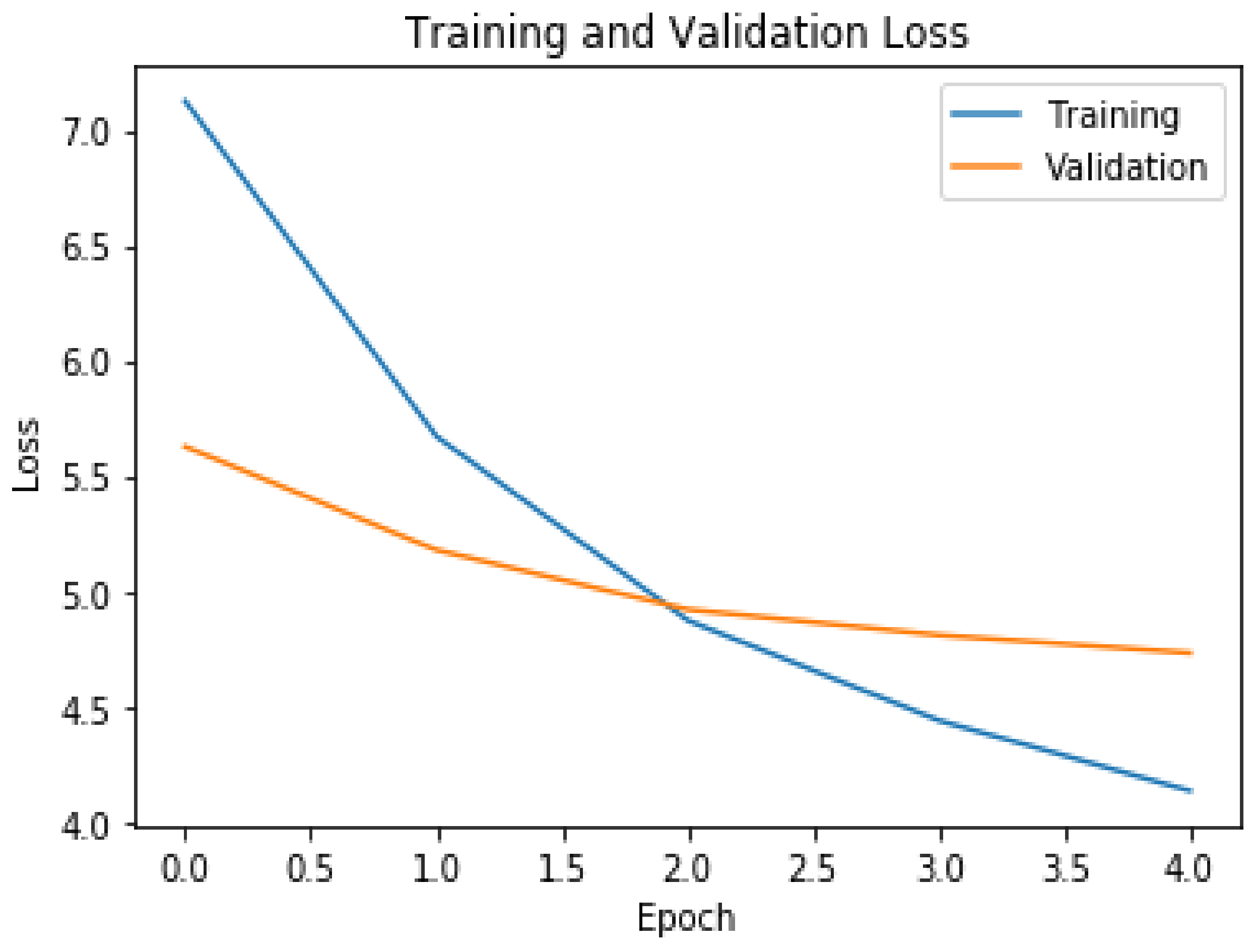
Results and Findings

Original Title	*Predicted Title*
America by Air: Descending into a Dust Storm	America by air calm before the storm
Donald Trump, John Glenn, Oakland: Your Thursday Evening Briefing - The New York Times	donald trump your wednesday evening briefing the new york times
How Climate Change Unleashed Humans Upon South America's Megabeasts	The pope francis allow married man to become priests

Performance of predictions vary depending on the size of the training data set, the complexity of the encoder and decoder and compute capability of the system, such as hardware design and number of cores.

Model Loss

Training and Validation Loss



Conclusion

- Effective reduction in loss with more epochs, although the predicted results did not improve significantly.
- Key issues:
  - **Small Datasets**: A open source dataset available for the research may comprise fewer than 5,000 training text examples
  - **Lack of Computational Memory**: 8GB or 16GB RAM was not enough to execute dataset efficiently
  - **Capitalization**: As all predictions are displayed in lower-case, future direction will be focussed on developing capitalized-detecting.

Reference

Husain, Hamel. "How To Create Data Products That Are Magical Using Sequence-to-Sequence Models." *How To Create Data Products That Are Magical Using Sequence-to-Sequence Models*, Towards Data Science, 18 Jan. 2018,