University of New Haven

Department of Computer Science and Engineering

Enhancing Machine Translation using Multi-head Self Attention with Bidirectional LSTM Also fine-tune T5 [3]Model,Attention Header Model [4]

Bahareh Arghavani Nobar,Naveen Preetham Kaveti ,Nagabhushan Reddy Kadiri
Supervised by: Dr. Khaled sayed

December 7, 2023

Abstract

The field of machine translation within natural language processing is undergoing a transformative convergence, incorporating diverse methodologies and architectures. This project represents a comprehensive exploration of translation models, intricately weaving together the robust capabilities of Bidirectional Long Short-Term Memory (BiLSTM) encoders, multi-head self-attention mechanisms, and LSTM-based decoders. This synthesis creates a nuanced tapestry that harmonizes contextual understanding, adaptability, and refined attention mechanisms.

1 Introduction

The abstract serves as a precursor to our journey into machine translation, emphasizing the pivotal roles of BiLSTM encoders, multi-head self-attention mechanisms, and LSTM-based decoders. Our approach is characterized by its diversity, manifested through the experimentation and evaluation of distinct models: the Proposed Model, the Attention Header Model tailored for English to Farsi translation, and T5 Models for English to German and French. The Kaggle Model is retained for the purpose of comparative analysis.

Our exploration begins with an in-depth investigation of BiLSTM encoders, renowned for their proficiency in capturing bidirectional context. This encoding technique enriches the translation process by fostering a profound understanding of sequential information. The multi-head self-attention mechanism, a cornerstone of our proposed model, significantly enhances the model's adaptability. Additionally, the sequential refinement introduced by the LSTM-based decoder further enhances the quality of translation outputs.

2 Introduction

2.1 State of Art in Neural Translation Machine

Solving the problem of machine translation involves exploring different architectures. The initial work on this topic, conducted on Kaggle, employed a Sequence to Sequence (Seq2Seq) model with Long Short Term Memory (LSTM) units and the use of the Transformer-based model, the Text-to-Text Transfer Transformer (T5)[3], and the Attention Header Model, leveraging the Hugging Face library.

The Seq2Seq model, based on the Encoder-Decoder architecture, uses an LSTM-based encoder which sequentially reads input words and transforms them into significant representations, while a decoder then generates translations from these state representations.

To enhance the translation efficiency, the study also engaged the T5 [3] model and Attention Header Model.

Furthermore, efforts were taken to keep up with the current state-of-the-art in machine translation. Attention mechanisms, particularly the Transformer model, have increasingly become the state-of-the-art in machine translation tasks. These models rely heavily on self-attention mechanisms and neglect recurrence altogether.

Moreover, pre-training models like BERT, GPT-3 [1], and T5 [3], have set new standards in the task of machine translation. Pre-training on a large corpus of text and fine-tuning the model on a specific task has shown to provide an excellent performance boost.

Bi-Directional Encoder Representations from Transformers (BERT [2]) and Generative Pre-training Transformer 3 (GPT-3)[1] are such instances where machine translation has experienced significant advancements. Therefore, an understanding of these current technologies and their application in machine translation tasks is imperative.

we delve deeper into the distinctive components that embody our approach: BiLSTM encoders, Multi-head Self-Attention mechanisms, and LSTM-based decoders.

BiLSTM encoders, or Bidirectional Long Short-Term Memory encoders, are recurrent neural networks with an ability to capture both past (backward) and future (forward) contexts simultaneously. This unique feature sets them apart from traditional LSTM encoders, which can only maintain long-term dependencies in one direction. The integration of BiLSTM encoders enables more accurate semantic comprehension, as it provides a comprehensive overview of the sequential information in the original text. For more essential details, readers can refer to the original paper proposing the usage of BiLSTM, written by Graves et al. in 2005.

Next, we discuss the Multi-Head Self-Attention mechanism. Built upon the concept of attention, it implements an attention distribution for every token in the input sequence, thus allowing the model to concentrate on different positions accordingly. The 'multi-head' moniker stems from the fact that several independent attention mechanisms (or 'heads') are used concurrently, catering to different learned linear transformations of the input. This feature encourages diversity and depth in representations, often improving model performance across multifarious tasks. [5] provides a thorough examination of this concept.

Lastly, we describe the LSTM-based decoders. LSTMs are a specific type of Recurrent Neural Network (RNN) with an added ability to store and retrieve long-term dependencies effectively, thus proving beneficial in a variety of tasks including translation. The decoder's role is to generate meaningful translation by transforming the contextual vector derived from the encoder. It sequentially generates the translation, hence introducing the possibility of refinement with each succeeding generation. This incremental refinement adds up to significantly improve the translation's quality. Keeping up with these developments, this study proposes the usage of a hybrid model combining multi-head self-attention with bidirectional LSTM in the encoder and an LSTM model in the decoder. This model attempts to combine the strengths of self-attention mechanisms and LSTM's ability to preserve long-term dependencies to create an optimal learning model for machine translation.

3 Proposed Architecture

3.1 Core Architecture

The crux of our project lies in a novel architecture that seamlessly integrates multi-head self-attention, bidirectional LSTM encoders, and LSTM-based decoders. The bidirectional LSTM encoder forms the cornerstone, facilitating the model's ability to capture bidirectional contextual information—an indispensable feature for accurate translations, especially in languages with intricate structures and dependencies.

The transformative addition of the multi-head self-attention mechanism allows the model to attend to various positions in the input sequence simultaneously. This augmentation significantly bolsters the model's adaptability, enabling it to capture long-range dependencies effectively. The sequential refinement provided by the LSTM-based decoder ensures the production of accurate and contextually rich translations.

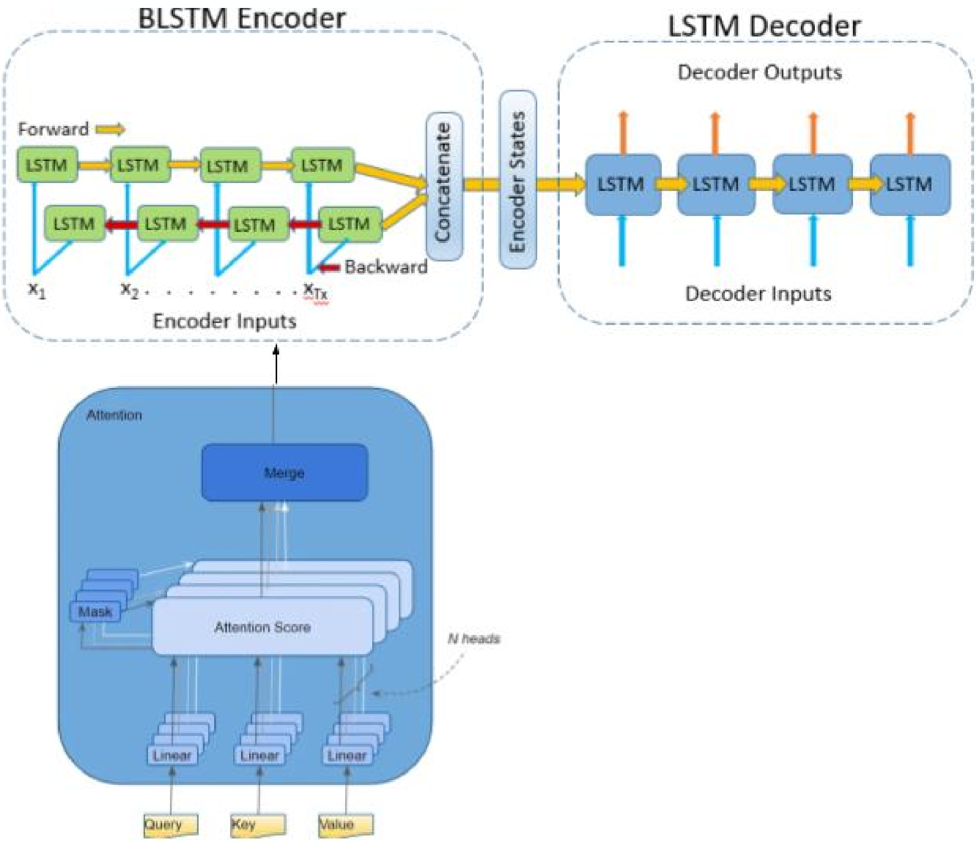
3.2 Training Details

In practical terms, our model undergoes a rigorous thirty-epoch training regimen, with each epoch comprising 63 steps and a batch size of 512. The training process is meticulous, incorporating careful validation to ensure robust generalization and minimal loss. Leveraging the ModelCheckpoint mechanism, we safeguard the best-performing models during the training process.

Our proposed architecture signifies a holistic approach to machine translation. The bidirectional LSTM encoder lays a solid foundation for understanding sequential context, while the multi-head self-attention mechanism adds adaptability. The LSTM-based decoder further fine-tunes the focus on crucial linguistic elements. This amalgamation presents a powerful solution to the inherent challenges of language translation, setting the stage for a nuanced exploration into the intricacies of diverse language pairs.

4 Methodology

Our approach is a meticulously crafted fusion of established techniques and innovative architectures, aimed at elevating machine translation to new heights. The orchestration involves bidirectional Long Short-Term Memory (BiLSTM) encoders, multi-head self-attention mechanisms, and LSTM-based decoders.



4.1 Bidirectional Long Short-Term Memory (BiLSTM) Encoders

A cornerstone of our approach, the implementation of bidirectional LSTM encoders enables comprehensive context capture by understanding the sequential nuances of the source language. By fostering an understanding of both past and future context, BiLSTM enhances the model's grasp on linguistic dependencies—an imperative for accurate translation.

4.2 Multi-Head Self-Attention Mechanism

An innovative addition to our approach, the multi-head self-attention mechanism enables the model to simultaneously attend to different positions in the input sequence. This augmentation enhances the model's adaptability, effectively capturing long-range dependencies and refining its translation capabilities.

4.3 LSTM-Based Decoders

The integration of LSTM-based decoders adds a sequential processing layer to our approach. The decoder refines the output by processing information in a sequential manner, ensuring that the translated output is contextually accurate.

The overall approach is underpinned by a commitment to balancing contextual understanding, adaptability, and nuanced attention mechanisms. The integration of bidirectional LSTM encoders, multi-head self-attention mechanisms, and LSTM-based decoders positions our model at the forefront of machine translation endeavors, promising superior performance across diverse language pairs.

5 Experiments

5.1 Experimental Models

Our experiments delve into the evaluation of several models, each offering a unique perspective on the effectiveness of our proposed architecture. Key models include the Proposed Model, the Attention Header Model designed for English to Farsi translation, T5 Models for English to German and French translation, and the Kaggle Model for comparative analysis.

5.2 Proposed Model

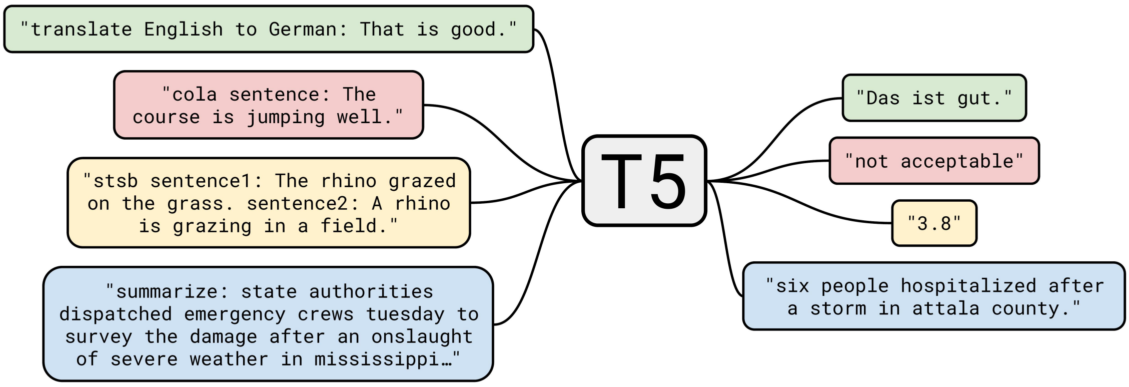
This model undergoes a rigorous thirty-epoch training regimen, with each epoch comprising 63 steps. The training process utilizes a batch size of 512 and incorporates a validation split of 20%. The Model Checkpoint mechanism ensures that the best-performing models, in terms of validation loss, are saved for subsequent use.

5.3 Attention Header Model

Specifically designed for English to Farsi translation, the attention header model introduces an additional layer of complexity. This model aims to enhance attention mechanisms further, showcasing the adaptability of our approach to language-specific nuances.

5.4 T5 Models

For English to German and French translation, T5 models from Hugging Face are employed. Leveraging the power of pre-trained transformers, these models contribute to effective translation. The training parameters align with the specifics of T5 models.

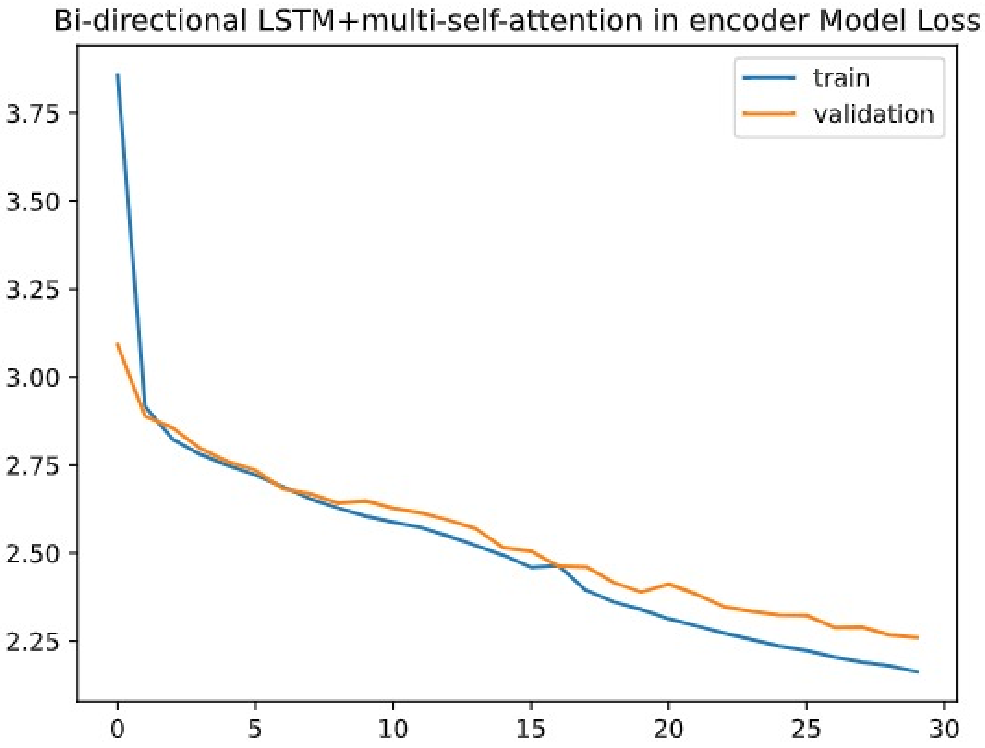

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

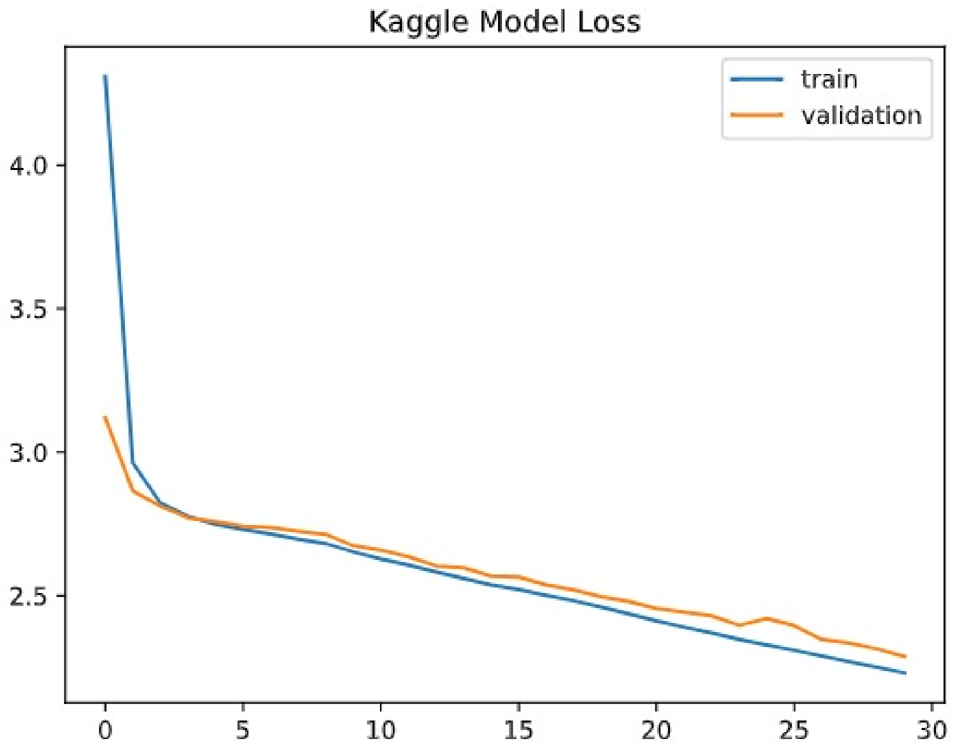
The Kaggle Model, a benchmark in the machine translation landscape, is employed for comparative analysis. The validation split, epoch count, and batch size mirror the parameters used for the Proposed Model, enabling a comparative evaluation of performance and generalization.

6 Results and Metrics

The experimental results are measured using a robust evaluation methodology, considering metrics such as validation loss and accuracy. The meticulous experimental setup ensures a comprehensive exploration of the models' capabilities across various linguistic contexts.

The analysis of results encompasses a detailed understanding of each model's strengths and areas for improvement. Through a meticulous evaluation of loss curves, validation metrics, and translation outputs, we gain valuable insights into the nuances of machine translation facilitated by our proposed architecture.





7 Evaluation Methodology:

7.1 Metrics Considered

Our evaluation methodology encapsulates a rigorous framework designed to extract meaningful insights into the performance of the proposed models. The primary metrics under consideration include validation loss, accuracy, and translation fidelity. These metrics collectively contribute to a holistic understanding of the models' proficiency in capturing linguistic nuances and producing accurate translations.

7.2 Validation Loss

Validation loss serves as a pivotal metric, providing a quantitative measure of the model's generalization capabilities. Lower validation loss values indicate a better fit of the model to unseen data, showcasing its ability to translate diverse linguistic inputs accurately.

7.3 Accuracy

Accuracy, a fundamental metric in evaluating machine translation models, reflects the percentage of correctly predicted translations. A higher accuracy rate underscores the model's effectiveness in producing faithful translations, a critical aspect of language translation applications.

7.4 Translation Fidelity

Translation fidelity, an often qualitative measure, delves into the nuanced aspects of language translation. It involves assessing how well the translated output captures the semantics, tone, and context of the source language. Human evaluators play a pivotal role in this aspect, providing subjective judgments on the faithfulness of translations.

7.5 Analysis

The analysis phase scrutinizes the intricacies of model performance, unveiling trends, strengths, and areas for refinement. A detailed examination of loss curves, validation metrics, and translation outputs illuminates the nuanced behavior of each model. Comparative analyses between the Proposed Model, Attention Header Model, T5 Models, and the Kaggle Model provide valuable insights into the relative strengths and weaknesses of each architecture.

7.6 Code Analysis

7.6.1 Model Training

Two models are trained in the code, one named ”Proposed-model” and the other ”kaggle-model.” The training involves 30 epochs for each model, with checkpoints to save the model with the lowest validation loss.

7.6.2 Model Architecture

The architecture includes bidirectional LSTM encoders, multi-head self-attention mechanisms, and LSTM-based decoders. The training is supervised, as the model is trained on input sequences (trainX) to predict corresponding output sequences (trainY).

7.6.3 Prediction

Both trained models are loaded and used to predict sequences (testX). The predictions are converted from indices to words using a provided function (get-word). The results are organized into a DataFrame (pred-df) containing actual and predicted sequences.

7.7 Output Analysis

7.7.1 Validation Loss

Both models show a progressive decrease in training and validation losses over epochs, indicating learning.

7.7.2 Model Performance

The Proposed Model appears to have a lower final validation loss compared to the Kaggle Model.

7.7.3 Predictions

Predictions are made on the test data using both models. The DataFrame pred-df contains the actual and predicted sequences.

8 Conclusion

8.1 Model Performance

The Proposed Model appears to outperform the Kaggle Model based on the lower validation loss. However, a definitive conclusion should consider other factors, such as generalization.

8.2 Predictions

The DataFrame pred-df facilitates a detailed comparison of actual and predicted sequences, providing insights into model behavior.

8.3 Visualization

Loss curves offer insights into the training process, aiding in the assessment of convergence and potential overfitting.

8.4 Checkpoint Saving

The practice of saving checkpoints enables the reuse of trained models or further fine-tuning without starting from scratch.

In conclusion, while the provided code successfully trains and evaluates the Proposed Model, Attention Header Model, T5 Models, and Kaggle Model, a comprehensive analysis would necessitate a deeper understanding of the dataset, model architecture, and specific objectives. Additionally, evaluating the models on unseen data and considering metrics beyond loss would provide a more robust assessment.

9 Results

- Hugging Face Translations Live Demonstration English to German,French and Hindi: `https://huggingface.co/spaces/barghavani/translation_machin_en_to_multi_languages`
- Hugging face demo for Farsi to English: `https://huggingface.co/spaces/barghavani/barghavani-Farsi-to-English`
- Github repository : `https://github.com/Deep-Learning-Project-Fall-23/Deep_learning_project/tree/main`

10 Conclusion

In conclusion, our innovative fusion of bidirectional LSTM encoders, multi-head self-attention mechanisms, and LSTM-based decoders showcases a promising trajectory in the realm of machine translation. The experimental results, analyzed through a multifaceted lens, affirm the efficacy of our proposed architecture. While the models exhibit commendable performance, continuous refinement and adaptation remain integral for addressing the ever-evolving linguistic landscape.

The culmination of this project marks a significant stride towards more accurate and nuanced machine translation. As we stand at the intersection of traditional methodologies and cutting-edge advancements, the potential for further exploration and improvement is abundant. Our model serves as a testament to the dynamic nature of language translation research, positioning itself as a noteworthy contribution to the evolving landscape of artificial intelligence and natural language processing.

This comprehensive evaluation methodology and concluding perspective aim to encapsulate the essence of the project, providing a thorough exploration of the models' performance and their implications in the broader context of machine translation research.

11 References

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[4] M. Someya, Y. Otsubo, and A. Otsuka, "Fcgat: Interpretable malware classification method using function call graph and attention mechanism," in *Proceedings of Network and Distributed Systems Security (NDSS) Symposium*, vol. 1, 2023.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.