Neural Machine Translation for English and Russian: A BERT-Based Approach

Yerkebulan Akbay
Dilnura Gabdollayeva
Gabit Kudret
- *Big Data Analysis*

Supervision

- Sanzhar Kusdavletov
- Arystan Amangeldi

Context

Machine translation (MT) is a product of human ingenuity which facilitates the automated transfer of written communication from one language to another. Its significance has been progressively acknowledged in contemporary times. Due to the massive convenience and fast transportation platforms, frequent communications among people in different countries are really common and drive a deeper need of translations for everyone anytime and anywhere.[1] However, MT strictly operates by substituting terms from one language to another, yet such a method may not completely guarantee proficient translation.

The integration of recent developments in *Natural Language Processing* (NLP) and *Neural Network* (NN) technologies has facilitated the enhancement of automated machine translation systems. Specifically, the challenge of identifying multiple phrases within a given text can now be accomplished without the need for human intervention, as the conversion process can be executed entirely by the machine.

In Machine Translation, among the different architectures, the *Transformer* has emerged as the dominant NMT paradigm. Relying only on attention mechanisms, the model is fast, highly accurate and has been proven to outperform the widely used recurrent networks with attention and ensembling.[2]

Pre-trained models like ELMo's advancement have made *Bidirectional Encoder Representations from Transformers(BERT)* and *GPT* more popular in recent years. One of the well-trained models that are frequently utilized is BERT. The Encoder and Decoder layers of the Transformer can be combined with BERT. There are three suggested ways to reuse BERT: embedding it, initializing the encoder with a BERT parameter, and utilizing it as the Transformer encoder. The MT model is strengthened by the leverage of BERT.[1]

Despite the progress achieved thus far, the machine translation phenomenon continues to confront various challenges, including the formidable task of processing idiomatic expressions and cultural subtleties, as well as the complex endeavor of retaining the meaning and tone of the original text.[3]

The utilization of machine translation presents a promising possibility to transform communication and enhance worldwide access to information. The progressive development of machine translation technology is anticipated to present an upsurge of cutting-edge and refined systems, capable of rendering text translation with heightened precision and speed, thereby opening up novel prospects for intercultural dialogue and cooperation.

Keywords: MT, NLP, Transformer, Self-Attention, Encoder, Decoder, BERT

Objectives

1. At the end of this course we plan to end up with a machine translation model based on *BERT* [1] and using *Tensorflow* [4], simultaneously exploring all the moments and describing mathematically, to translate text from Russian to English, and vice versa.

2. Preprocess and clean the parallel Russian $\leftrightarrow$ English training data to ensure high quality input for the model.

3. Get an understanding of different pre-trained models BERT, MT5 and so on,[5] and turn inside out *Transformer's Self-Attention* mechanism with *Encoder-Decoder* architecture .[1]

4. Develop a user-friendly interface that allows users to input text and receive the corresponding translation.

5. Visualize the model in the Embedding Projector to gain insights of representations of words and subwords, which uses techniques such as *principal component analysis (PCA)* and *t-SNE*.

6. Come up with a detailed report documenting the development, evaluation of the model, including any challenges encountered and potential future integrations into services.

Methodology

Particularly in language modeling and machine translation, recurrent neural networks, long short-term memory, and gated recurrent neural networks have been firmly established as state-of-the-art methodologies.[6]

The Transformer is the first transduction model to calculate representations of its input and output only via self-attention, without the use of convolution or sequence-aligned RNNs. The Transformer follows this overall architecture using stacked self-attention, fully connected layers for both the encoder and decoder. [2]

Encoder: N = 6 identical layers make up the stack that makes up the encoder. There are two sublayers in each layer. The first is a multi-head self-attention mechanism, and the second is a straightforward feed-forward network that is fully connected positionally. Around each of the two sub-layers, we use a residual connection, followed by layer normalization. [2]

Decoder: Additionally, a stack of N = 6 identical layers makes up the decoder. The decoder adds a third sub-layer to each encoder layer in addition to the two already there, performing multi-head attention over the encoder stack's output. We use residual connections around each of the sub-layers, just like the encoder, and then layer normalization. [2]

Attention: A query, a set of key-value pairs, and an output, all of which are vectors, can be mapped to one another by an attention function. The result is calculated as a weighted sum of the values, with each value's weight determined by the query's compatibility function with its corresponding key. [2]

Technical Plan

- Data Collection and Data Preprocessing
  - Collection parallel Russian and English language data.
    - The Europarl corpus
    - The United Nations Parallel Corpus
    - WMT
  - Clean and preprocess the data by removing noise, special characters when collecting.
  - Applying *stemming, lemmatization, part-of-speech tagging, stop-word removal.*[2]
- Data Preparation
  - Encoding the text into numerical vectors such as *One-Hot encoding, Bag-of-Words and Word-Embeddings* from the previously step created vocabulary.
  - Split the parallel data into train, validation, and test sets.[7]
- Model Architecture
  - BERT with Attention/Transformer model.
  - Implementation of the attention mechanism for the encoder-decoder layers to focus on corresponding context of the input sequence.[8]
- Training
  - Defining the loss function, optimizer and hyperparameters.

  - We are going to train the model on the training data and while validating on the validation set, for keep tracking of the model's performance and prevent overfitting.

- Evaluation
  - Loading graph of checkpoints and benchmarking performance.
  - Evaluation of the model on the test set to measure its performance in terms of accuracy and with BLEU score.[8]

- Complete Pipeline
  - Will accept Russian text as an input and return English text as an output, and other way around.

References

[1] H.-I. Liu and W.-L. Chen, "Re-transformer: A self-attention based model for machine translation," *Procedia Computer Science*, vol. 189, pp. 3–10, 2021, aI in Computational Linguistics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921011509

[2] A. Raganato and J. Tiedemann, "An analysis of encoder representations in transformer-based machine translation," *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[3] B. D. Shivahare, S. Ranjan, A. M. Rao, J. Balaji, D. Dattattrey, and M. Arham, "Survey paper: Study of sentiment analysis and machine translation using natural language processing and its applications," in *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, 2022, pp. 652–656.

[4] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," Software available from tensorflow.org, 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[5] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *CoRR*, vol. abs/2010.11934, 2020. [Online]. Available: https://arxiv.org/abs/2010.11934

[6] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering*, vol. 18, pp. 143–153, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809921002745

[7] V. S. R. Middi, M. Raju, and T. A. Harris, "Machine translation using natural language processing," *MATEC Web of Conferences*, vol. 277, p. 02004, 01 2019.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf