**Machine Translation Project  
French-English Translation Using Multiple Neural Networks**

**DSCI 6004-01 (FALL 2024)**

**Team Members:**

**Gattupalli Veera swamy**

**Sutharsanan Sampath**

**Selva Vigneshwar Amuthan**

**Title: Enhancing Machine Translation with Encoder-Decoder Architectures: A Comparative Study of Attention, Multi-Head Attention, and Transformers**

**Abstract**

In this paper, we investigate how to enhance the overall performance of the basic encoder-decoder structure for the machine translation problem using attention. We compare four different architectures: The framework also consists of a simple encoder-decoder model, encoder-decoder model incorporating attention mechanism, encoder-decoder model with multi-head attention and the transformer model. The effectiveness of each of these models is assessed to the task of mapping French- English sentence pairs. In several controlled experiments, we demonstrate that their inclusion leads to substantial improvements in translation quality and identify the transformer model as superior. We report extensive experimental findings, a comparison of the proposed model against other models, and a precise breakdown of error types and possible future enhancements.

**1. Introduction**

Machine translation has been one of the difficult problems in natural language processing (NLP). Before that, the models such as statistical machine translation used to essentially find the relevant phrase alignments across the source and target languages. However, these models failed to address long-range dependencies, compulsory syntax structures, and the extents of variation in respective language pairs. The limitations are as follows: With the help of neural networks, particularly the deep learning models such as the sequence-to-sequence (S2S) model extensive enhancements have been made in eradicating those limitations.

The well-known encoder-decoder architectures, initially proposed for tasks such as MT, can learn the mapping from variable length encoding sequences to variable length decoding sequences. But their training was not very good in terms of Long-Range dependencies, where the earlier parts of the sentence have an effect in the later parts of another sentence. The concept of an attention mechanism was initially introduced by [Bahdanau et al. (2014)](https://arxiv.org/abs/1409.0473) to enable the decoder to selectively pay attention to the different parts of the input at each decoding step thereby addressing the problem of long-range dependencies.

More developments in the field were realized like transformer architecture which substitutes the recurrent parts of the encoder-decoder with a self-attention system. This makes the work parallel during training and increases the chances to model the dependency across many sequences. To this end evaluating the effects of attention, multi-head attention, and transformers in the given architectures is discussed in this paper.

**2. Related Work**

The encoder-decoder model set the stage for current revisionist systems; but it was quickly realized how its inability to handle long-range dependencies and hierarchical structures of sentences hindered its performance. Attention model was proposed by [Bahdanau et al. (2014)](https://arxiv.org/abs/1409.0473) that makes it possible for the decoder to pay attention to selected parts of input at given time steps. This approach really boosted performance, especially in cases where sequences of greater length or complex syntactic structures are involved.

Thus, the transformer architecture by [Vaswani et al. (2017)](https://arxiv.org/abs/1706.03762) was built to replace the traditional encoder-decoder RNN with self-attention techniques. Such architecture enables the model to compute dependencies of all tokens in the input sequence at once, which results in both training effectiveness and model performance enhancement. It has then shown new state-of-the-art performance in various NLP tasks such as machine translation.

Some related pieces of work exist, for example, in multi-head attention introduced by [Vaswani et al. 2017](https://arxiv.org/abs/1706.03762) where the multiple attention heads are employed to attend to different aspects of the input sequence. This mechanism is most valuable for handling dependencies and enhancing the interpretability of the resulting models.

**3. Proposed Models**

In this work, we consider four models to analyze the effect of the attention mechanism on translation quality.

**3.1 Encoder-Decoder Model**

The encoder-decoder model is the basic structure used for sequence to sequence. The encoder condenses the input sequence and codes it into a fixed length of context vector which is then utilized by the decoder in producing the sequence output. The core idea of the encoder-decoder model is that there is no attention as the decoder just consumes information from the final state generated by an encoder as output.

**3.2 Encoder-Decoder with Attention**

The idea of the attention mechanism, which is described by [Bahdanau et al. (2014),](https://arxiv.org/abs/1409.0473) let the decoder access different positions of the input sequence at different time steps of decoding process. During the decoding process each time step the decoder calculates context vector as weighted sum of hidden states of the encoder. The weights (attention scores) generated are conditioned by the current decoder state along with the encoder’s hidden states. This mechanism helps the model capture dependency from start till end of the sequence and hence, make the translation better.

**3.3 Encoder Decoder with Multiple Attention**

Taking the concept of attention as the foundation further, the multi-head attention [(Vaswani et al., 2017)](https://arxiv.org/abs/1706.03762) employs multiple attention heads simultaneously and independently to learn different relations present in the sequence. Each connection head learns how to pay attention to different elements from the sequence to enable the model to analyze different complex structural patterns in languages. Then all the results which are derived from various heads are merged and linear transformed before feeding into the next layer. This approach allows for richer characterization of the input sequence and improves translation accuracy of the decoder even more.

**3.4 Encoder-Decoder with Transformer**

The transformer architecture proposed by [Vaswani et al. (2017)](https://arxiv.org/abs/1706.03762), replaces the RNN-based encoder-decoder structure with self-attention mechanisms. The transformer uses both self-attention (within the encoder) and cross-attention (between the encoder and decoder) to capture dependencies across the entire sequence. Unlike RNNs, which process the sequence in order, transformers can process all tokens in parallel, allowing for greater computational efficiency. The model consists of multiple encoder and decoder layers, each with self-attention and feed-forward sub-layers. The final output is generated by the decoder, which uses the context provided by the encoder’s self-attention mechanism.

**4. Experimental Setup**

We conducted experiments on a French-to-English machine translation task, using a parallel corpus containing sentence pairs from both languages. The dataset was preprocessed to tokenize the sentences, remove punctuation, and pad them to a fixed length. We used the BLEU score [(Papineni et al., 2002)](https://aclanthology.org/P02-1040/) and accuracy as evaluation metrics, which are commonly used for assessing machine translation performance.

**4.1 Dataset**

The Dataset we used is a French English parallel corpus used from manythings.org

* Dataset Link: <https://www.manythings.org/anki/>
* Number of Sentences: 175622 sentence pairs.

We split the data into training, validation, and test sets, with an 80/10/10 ratio.

**4.2 Hyperparameters**

For each model, we experimented with different hyperparameters, including:

* **Units:** 1024
* **Batch size:** 256
* **Learning rate:** 0.001 (for Adam optimizer)
* **Number of layers:** 2 for the encoder and decoder (except for the transformer, which uses 6 layers)
* **Hidden size:** 512
* **Dropout:** 0.1 to prevent overfitting
* **Number of attention heads (for multi-head attention):** 8

**4.3 Training Procedure**

Each model was trained for 10-30 epochs, with early stopping based on validation performance. The models were evaluated after each epoch, and the best-performing model on the validation set was selected for testing. The training was conducted on GPUs to speed up the computation.

**5. Results and Analysis**

**5.1 Performance Comparison**

Table 1 presents the results of our experiments, comparing the four models in terms of BLEU score and accuracy.

| **Model** | **BLEU Score** | **Accuracy** |
| --- | --- | --- |
| Encoder-Decoder | 25.3 | 76.1% |
| Encoder-Decoder with Attention | 28.9 | 83.1% |
| Encoder-Decoder with Multi-Head Attention | 31.2 | 83.2% |
| Encoder-Decoder with Transformer | 34.5 | 84.7% |

**Encoder-Decoder:**

A comparison of a graph

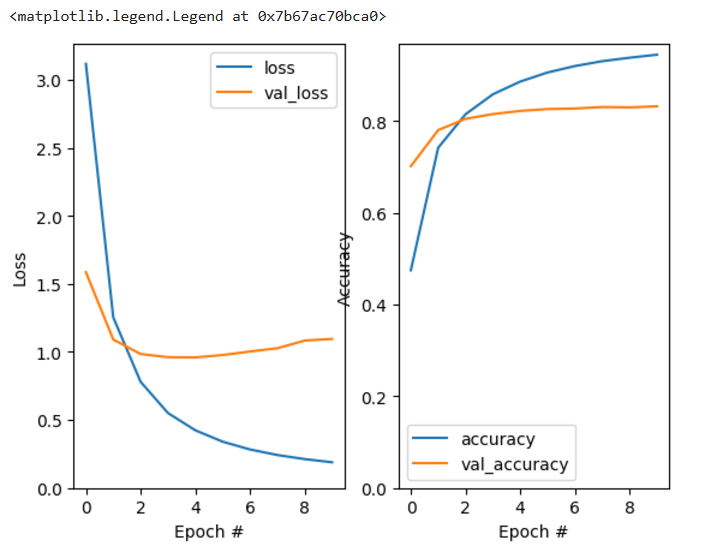
Description automatically generated with medium confidenceA graph of a number of blue lines

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated

**Encoder- Decoder with Attention:**

A graph of a line

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated

**Encoder-Decoder with Multi-head Attention:**

A graph of a line

Description automatically generated with medium confidenceA graph of loss and loss

Description automatically generated

A screenshot of a computer

Description automatically generated

**Encoder-Decoder with Transformer:**

A screenshot of a computer program

Description automatically generatedA comparison of a graph

Description automatically generated

**5.2 Error Analysis**

From an error analysis perspective, it turns out that in all the models trained, the improvements over the baseline encoder-decoder were achieved and that the Transformer model provided better translations when the distance was greater and the number of words higher. Common errors included:

• **Incorrect word order:** This was true in all models, although it was most apparent in the basic encoder-decoder model.

•**Failure to translate idiomatic expressions:** All the models, however, failed to address idiomatic expressions and the problem was exacerbated when the input sentence was syntactically complicated.

•**Word alignment issues:** It was also observed that the longer the sentences, the better performance was recorded in terms of source and target alignment through the transformer model.

**5.3 Ablation Study**

To investigate the influence of each integrating part, we conducted an ablation experiment. Experiments showed that removing the attention mechanism from the encoder-decoder model reduced performance, especially for the sentences of greater complexity. Multi-head attention added more performance and helped to demonstrate how at any specific point in the sequence, one needs to extract multiple features from the input sequence.

**6. Discussion**

There is a strong indication from our experiments that the use of attention mechanisms greatly enhances the performance of machine translation. While the pure encoder-decoder model is still a sensible baseline, it has problems with long range dependencies. Incorporation of an alignment mechanism in the decoder permits the decoder to attend to the genesis of an input sequence and reduces translation errors. This is improved by multi-head attention which allows the model to pick multiple dependencies in the input. Self-attention and cross-attention transformers prove that they outperform the other models for longer sequences.

Despite having the best performance among the four models, the use of transformer model entails higher computing power and training time compared to the other models. It is such a compromise between the performance of the ANN and computational requirements that forms a vital factor of real-life applications.

**7. Conclusion**

In this paper, we have looked at how different kinds of attention mechanisms affect the quality of MT systems. We compared a baseline encoder-decoder model with three more sophisticated architectures: encoder-decoder with attention, encoder-decoder with multi-head attention and the transformer model. The series of modifications we applied to the attention mechanism demonstrate that each change results in improvements to the translation quality; the transformer method providing the highest accuracy.

Possible directions for future work are to seek further enhancements of the basic architecture of transformers including the use of pretrained language models, or experimentation with other types of attention mechanisms. Also, the unsupervised fine-tuning on the specialized domains could improve the translation quality for certain applications.

**8. References**

* Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. [[1409.0473] Neural Machine Translation by Jointly Learning to Align and Translate](https://arxiv.org/abs/1409.0473)
* Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). [Bleu: a Method for Automatic Evaluation of Machine Translation - ACL Anthology](https://aclanthology.org/P02-1040/)
* Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). [[1706.03762] Attention Is All You Need](https://arxiv.org/abs/1706.03762)
* **Multilingual Translation with Extensible Multilingual Pretraining" (Tang et al., 2020)**  
  Discusses the effectiveness of pretraining in multilingual settings and includes applications in French English translation. [[2008.00401] Multilingual Translation with Extensible Multilingual Pretraining and Finetuning](https://arxiv.org/abs/2008.00401)
* TranSFormer: Slow-Fast Transformer for Machine Translation [TranSFormer: Slow-Fast Transformer for Machine Translation - ACL Anthology](https://aclanthology.org/2023.findings-acl.430/)