Midproject: Channel Combined Transformer in Machine Translation

Xuanzhou Chen
Department of Electrical & Computer Engineering
New York University
Brooklyn, NY 11201
xc2425@nyu.edu

Abstract

In this midproject report, we present our project progress based on proposed schedule. We went through several literature review regarding the Encoder-(Channel)-Decoder Mechanism, the neural/statistical machine translation, and some channel techniques such as additive white Gaussian noise channel(AWGN), binary symmetric channel(BSC) and binary erasure channel(BEC). For the code implementation, we not only reproduced and evaluated Turbo decoding results from the paper *Communication Algorithms via Deep Learning*, but also worked on our own model building. Github link: https://github.com/xchen793/midproject

1 Introduction

Seq2Seq is a model which conducts direct estimation of the posterior probability of a target sequence y given a source sequence x.(here cite (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) Seq2Seq is a application level concept, indside Seq2Seq, the key model is the Encoder-Decoder structure, which is a net architecture concept. Transformer (Attention is all you need) is a completely new model which abandoned traditional GRU or LSTM, and largely introduces attention mechanism, but transformer still takes Encoder-Decoder structure build network architecture.

Since the publicizing of transformer, there have been

According to our proposal, our main goal is to design an naive encoder-(channel)-decoder deep learning model, apply it to language translation, and to test the model performance on a certain corpus, finally explain the model performance from a mathematical perspective if possible.

At this point, we have investigated massive papers regarding encoder-decoder schemes, communication channels and machine translation techniques. We further conducted some research on specific models including LSTM, GRU, Transfomer and Seq2Seq. Some mathematical foundations behind these models were also explored. Given the feedback on our proposal, we specifically focused on the transformers and Seq2Seq models, which involve state-of-art attention mechanisms. We also thoroughly studied the channel coding and channel models in machine translation.

The traditional transformer architectureVaswani et al. [2017] has one big drawback that it has poor performance to translate a large corpus due to its high computational cost. To solve this problem, we propose a new model, where we choose some certain type of channel and put it between encoder and decoder on either traditional Transformer or standard Seq2Seq, to improve the model performance of translating a large corpus. As this is the midterm checkpoint, we have not done sufficient experiments to verify the effectiveness of this new model. However, we do expect to achieve error rates that is no higher than the original Transformer/Seq2Seq model.

Preprint. Under review.

2 Proposed Methodologies

Generally, our basic methodology is model comparison and model combination. In this section, we present some critical papers inspiring us while we tying to find the final construction of our new model. Section 2.1 introduces noise channel modeling on transformer, section 2.2 explains the noise channel modeling on standard Seq2Seq.

Furthermore, we consider beam search algorithm to find the most likely k sequences of words and adopt BLEU Papineni et al. [2002] to evaluate our model performance.

2.1 Model Base Structure

In this section, we explain where we come up with the idea to use encoder-(channel)-decoder scheme in machine translation. The original paper we looked at is from Farsad et al. [2018]. Farsad et al. [2018] defined their system model associated with transmitting sentences from a transmitter to a receiver. They put forward a encoder-channel-decoder scheme. Since deep learning algorithms are used, the channel must satisfies the property that allows backpropagation, otherwise the model cannot be trained. There are some communication channels, including the additive Gaussian noise channel, multiplication Gaussian noise channel and the erasure channel, can be formulated using neural network layers. So far, we are still working on the implementation of these channels. More importantly, we will take the encoder-(channel)-decoder scheme as our base structure to build our own model.

2.2 Transformer Model Architecture

Next, for encoder and decoder exploration, we first looked into the model architecture of Transformer by Vaswani et al. [2017], which is a state-of-art model which adopts the attention mechanism and abandons recurrence in RNN and convolution in CNN. Through comparison, we found that in Transformer, the output of encoder is directly fed into the multi-head attention in decoder without passing through any channel. We also noticed that one of the shortcoming of the traditional Transformer is that it is not suitable for translating relatively long word sequence. When the source sequence becomes especially long, e.g., an entire article, the required computation of the model will increase and we will have to truncate the text sequence into subsequences, which means the long distance dependency will inevitably decreases. As a result, we started considering to narrow down our project to the specific problem: if we implement some certain type of channel (must allow for backpropagation) and plug the channel between the encoder and the decoder, would this problem be solved? Hence, we switched to doing some research on the papers regarding the channel coding/channel modeling in machine translation.

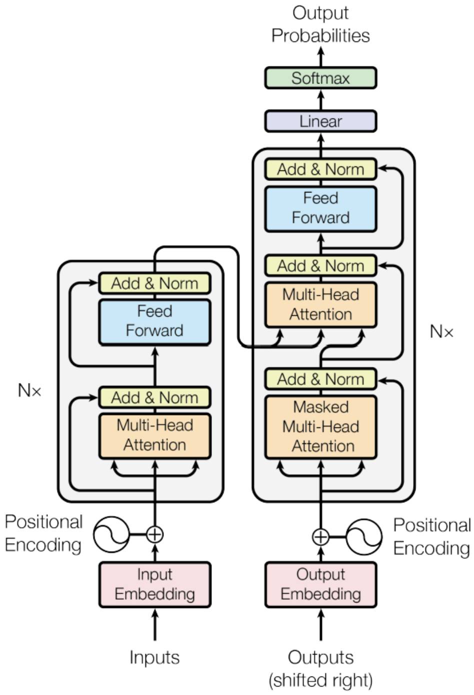
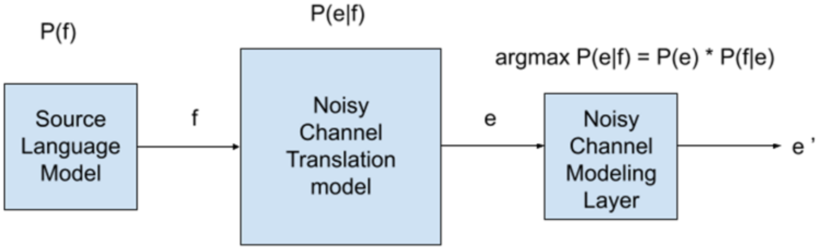


Figure 1: The Transformer - model architecture.

2.3 Noisy Channel Modeling on Standard Seq2Seq

There is some previous work contributed to noisy channel modeling for neural machine translation by Yee et al. [2019]. The researchers proposed an alternative approach where the entire source can be utilized based on standard sequence to sequence models rather than a standard transformer (due to robustness to some mismatch). Their results are remarkably comparable to the channel models.

One crucial advantage of the channel models pointed out by Yee et al. [2019] is that they are particularly effective with large context sizes, which we think may potentially make up for the weakness of standard Seq2Seq/Transformer.



2.4 Amortized Noisy-Channel Neural Machine Translation

Another paper we found relevant to the channel model in machine translation is *Amortized Noisy Channel Neural Machine Translation* by Pang et al. [2021], which also aims to improve computational efficiency when translating a high volume of texts. However, their main work is to use three approaches: knowledge distillation, 1-step-deviation imitation learning, and Q learning to approximate a noisy channel NMT system, which is different from our basic idea.

3 Relevant Work

For experiments, we started with implementing some code to build Transformer/seq2seq backbone and reproducing some results from the paper *Communication Algorithms via Deep Learning* by Kim et al. [2018].

To reproduce and evaluate the results from Kim et al. [2018], we trained Viterb, BCJR, and Neural Turbo Decoder model under Ubuntu 20.04.3, python2.7, and Tensorflow 1.5 which is the same environment in the paper, and evaluated them on the same dataset. Given that the original paper research work was more than 4 years ago, and the Tensorflow/Pytorch develops rapidly, we decided to create our own script with regard to implement Encode-Channel-Decode(Seq2Seq) scheme with AWGN, BEC and BSC.

4 Preliminary Results

Followed the paper Communication Algorithms via Deep Learning, we trained some Encoder-Channel-Decoder models focusing on Channel configuration such as BCJR-like RNNs, Viterbi-like RNNs, and Neural Turbo Decoder, and evaluated BER/BLER for convolutional code encoder, turbo code encoder, LTE turbo encoder and Nerual RNN Turbo. The reproduced results are shown in Figure1,2the results we reproduced matches the original paper, the proposed proposed Nerual RNN Tuobo Decoder outperformed some other well established algorithms such as BCJR and Viterbi back then. However, to reproduce this results, tf1.5 and python2.7 environment is required which is out dated and not compatible with Pytorch.

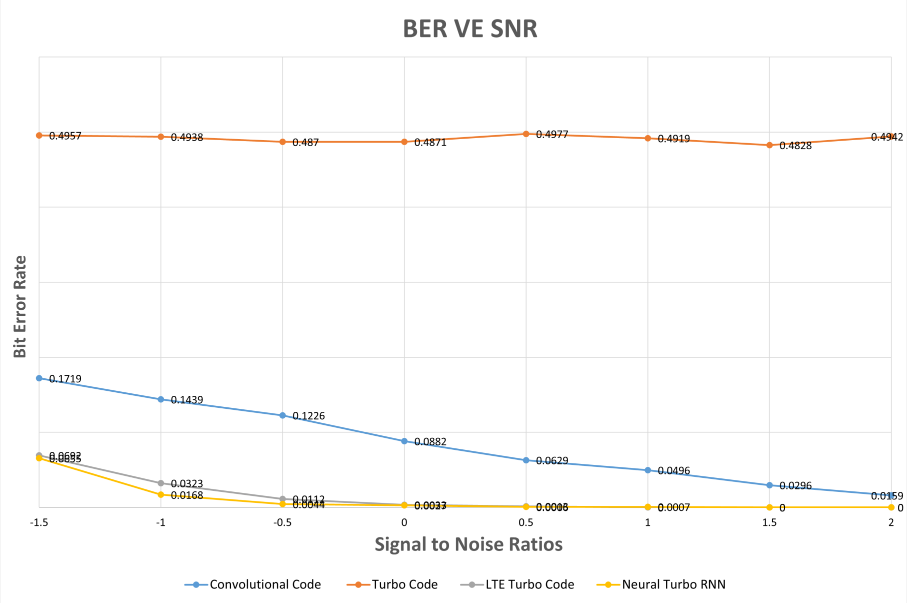


Figure 1: Bit Error Rate vs Signal to Noise Ratios

Below are some results of running our traditional Transformer model on our dataset. The first graph shows the training and validation loss. The second graph demonstrates training and validation accuracy. Since we just want some fast training, so we only run 50 epochs for a preliminary test, and that is why the accuracy was so low at this time.

The heat map shows that with multi-head attention, the model demonstrate its ability to focus on different locations. Moreover, the multi-head attention enable the model to focus on several nonconsecutive locations even they are far away.

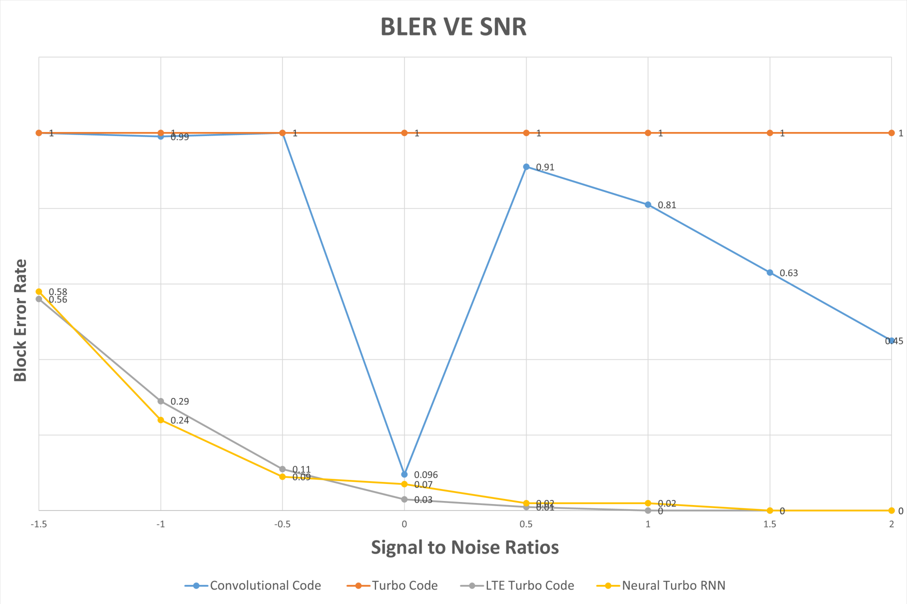


Figure 2: Block Error Rate vs Signal to Noise Ratios

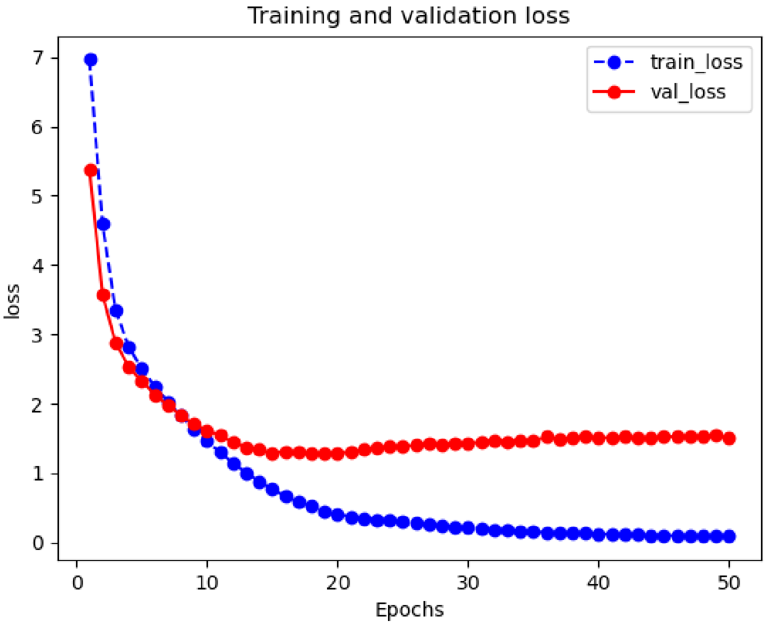


Figure 3: Training and Validation Loss

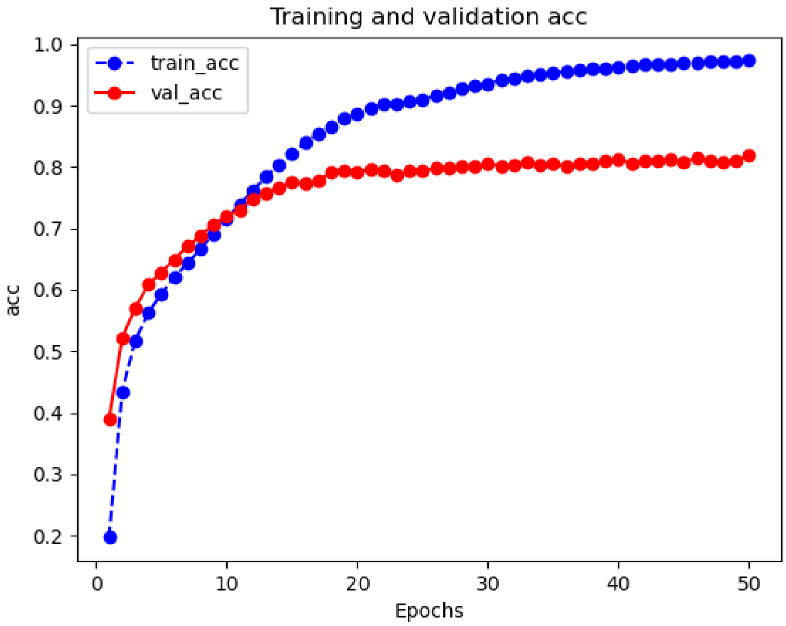


Figure 4: Training and Validation Accuracy

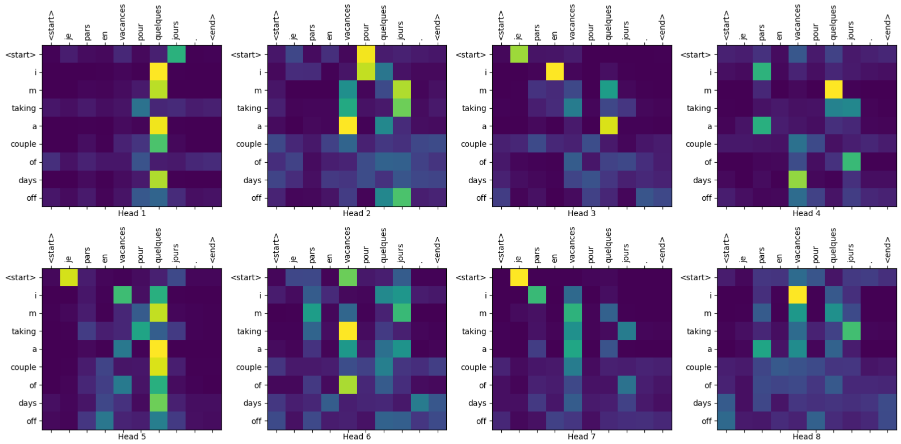


Figure 5: Transformer Attention Visualization

5 Future Work

At this checkpoint, we have implemented the Seq2Seq baseline which can successfully be trained and tested on our dataset. From the reproduction of the paper *Communication Algorithms via Deep Learning* by Kim et al. [2018], we have learned how to implement different communication channels and will add them into our encoder-decoder model. In the future, we first need to come up with our own model containing an encoder, a channel and a decoder, and then test its performance on a large corpus. While comparing it to the traditional Transformer/Seq2Seq, we will give some mathematical proofs for justification.

References

Nariman Farsad, Milind Rao, and Andrea Goldsmith. Deep learning for joint source-channel coding of text. *CoRR*, abs/1802.06832, 2018. URL `http://arxiv.org/abs/1802.06832`.

Hyeji Kim, Yihan Jiang, Ranvir Rana, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Communication algorithms via deep learning, 2018. URL `https://arxiv.org/abs/1805.09317`.

Richard Yuanzhe Pang, He He, and Kyunghyun Cho. Amortized noisy channel neural machine translation. *CoRR*, abs/2112.08670, 2021. URL `https://arxiv.org/abs/2112.08670`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

Kyra Yee, Nathan Ng, Yann N. Dauphin, and Michael Auli. Simple and effective noisy channel modeling for neural machine translation. *CoRR*, abs/1908.05731, 2019. URL `http://arxiv.org/abs/1908.05731`.