Construction of Sequence-to-sequence Generative Models for Machine Transition with Synthesized Sound Output System

Tianjiao Yu
Department of Computer Science
Georgetown University
Washington, D.C., U.S.A
ty262@georgetown.edu

Yifan Zhu
Department of Linguistics
Georgetown University
Washington, D.C., U.S.A
yz644@georgetown.edu

Beixuan Jia
Department of
Data Science and Analytics
Georgetown University
Washington, D.C., U.S.A
bj256@georgetown.edu

Miao Wang
Department of
Data Science and Analytics
Georgetown University
Washington, D.C., U.S.A
mw1219@georgetown.edu

Abstract

Sequence-to-sequence (seq2seq) is a general purpose encoder-decoder framework that was designed for machine translation. It also has been used for a variety of other natural language processing tasks including text summarization, image captioning etc. One of the main advantages of this approach is its ability to handle input data with variable lengths and generate sequence outputs in various forms, such as audio, spectrum and sentence sequences. This project will apply the seq2seq model to develop an applicable mechanism with a neural machine translation and corresponding audio output generation system.

1 Introduction and Problem Statement

Machine translation has been widely-used in our daily life. However, for people with disabilities, access to the textual results might not be enough. To help develop a mature translation system with disability-friendly components, this project aims to apply deep learning models and concepts on machine translation and speech synthesis. Thus, every non-English sentence is able to get their English translation with a corresponding sound output.

The project consists of two parts: machine translation and speech synthesis. Seq2seq models are applied to both parts and train them on their corresponding datasets. The inputs and outputs of the two models are all arbitrary-length sequences. The first model focuses on machine translation with an attention layer to improve the accuracy. The second part takes the text outputs of the first model as inputs and turns them into spectrum. Section 3 is a brief introduction of related work. The data sets will be introduced in section 4, methods and the structure of the two models will be described in details in section 5. In this report, models are evaluated using the classic procession, recall and F1 scores. For new input data, this report also takes the BLEU score for machine translation and Mean Opinion Score for model evaluation into account.

2 Related Work

2.1 Machine Translation

Machine translation was once considered as a challenging task in previous decades. It was largely based on rules, which rely on bilingual corpora to analyze linguistic regularities in both source and target languages. However, due to the richness of the natural languages, creating such systems can be costly and time-consuming.

Compared with rule-based systems, statistical machine translation analyzes both monolingual and bilingual content in order to generate sophisticated statistical models. These systems achieve a higher adequacy due to their capability of learning statistical features rather than manually-entered linguistic rules. However, languages are not just bags of words. The statistical systems cannot capture language features other than the probabilistic features.

Nowadays, neural network has been adopted as the state-of-art method for machine translation. Compared with statistical MT, Neural MT models are simpler structures because they use a single sequence processing component.

Recurrent neural networks (RNN) is one of the most mature methods for machine translation. The idea is to feed the output from the previous time steps as the input of the following time steps. The main advantage of this model is that it can capture the dependencies of sequential data. However, a simple RNN layer often leads to exploding or vanishing gradient problems (Tracey, 2019 accessed Dec 10, 2020). To solve this problem, LSTM and GRU are created. More details will be discussed in section 5.

2.2 Speech Thesis

Many fully constructed Text-to-Speech systems consist of front-end and back-end feature extraction methods, a duration model, an acoustic feature prediction model and different vocoders. These parts require extensive expertise and independent training. Every part of the error has been passed down to the next part and leads to compounding errors in the end. Based on the cons, an end-to-end training model was expected.

WaveNet made attempts in this direction. It is a powerful generative model of time domain waveforms, producing sound quality similar to human beings. Although it replaces the vocoder and acoustic model, the input still needs linguistic features, predicted log fundamental frequency (F0), and phoneme duration. Thus, it fails to be a complete end-to-end model.

Char2Wav (Sotelo et al., 2017) is one of the successful end-to-end models that can be trained on characters. However, it still predicts vocoder parameters before using a neural vocoder. Their seq2seq and SampleRNN vocoder models need to be separately trained beforehand as well.

Tacotron, released by Google, is a text-to-speech mechanism generating human-like speech from text (Wang et al., 2017). This sequence-to-sequence model only takes speech examples and corresponding text transcripts as input. The model maps a sequence of letters to a sequence of features that encode the audio. These features capture not only pronunciation of words, but also various subtleties of human speech, including volume, speed and intonation. In the end, these features are converted to a waveform. Our implementation is largely based on the Tacotron project.

3 Datasets

For the machine translation, the dataset is from the Tatoeba project, an open-source collaborative database of sentences and translations in the format of tab-delimited bilingual sentence pairs. The translations are from volunteers who speak various languages. Currently, the dataset has included translations between English and more than 80 languages. This project focuses on the following language translation pairs: English - Chinese, English - French, English - German, English - Catalan and English - Spanish. In addition to the English - Catalan pair, other language pairs have relatively larger samples. In total, these five pairs of translations have over 400,000 entries. Each dataset is sorted by length, with shorter sentences put first.

For speech synthesis, we used LJ Speech dataset, this dataset consisting of 13,100 short audio clips of a speaker reading passages from different books. Each clip also has corresponding transcription. Each audio file is a single-channel 16-bit PCM WAC with a rate of 22050 Hz. They have variable lengths, from 1 to 10 seconds. The transcription was matched to the audio manually, and a QA pass was done to ensure the accuracy.

4 Model Introduction

4.1 methods

Sequence-to-sequence model This model turns one sequence into another sequence. In this model, we use LSTM to avoid the problem of vanishing gradients. The context for each item is the output from the previous step. The key components of the model are an encoder and a decoder. The encoder transfers each single input into a corresponding hidden vector and a context with the input. The decoder reverses the process, turning the vector into output items. During the process, the previous output functions as the context of the input.

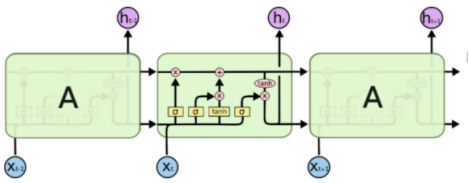

Figure 1: LSTM model

Long Short term memory (LSTM) LSTM is an artificial recurrent neural network (RNN) architecture. The mechanism is capable of learning long-term dependencies. The key to LSTM is the ability to remove or add information in the cell, carefully regulated by gates. Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation. An LSTM has three gates to protect and control the cell state, as Figure 1 shows.

Gated Recurrent Unit (GRU) GRU is also an RNN architecture, consisting of two gates, a reset gate and an update gate. The reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the previous memory to keep around.

Attention layer An attention layer is designed to fix the problem of information loss, which might lead to inadequate translation. The mechanism allows machine translation models to get all the information within the original sentence, and generate a proper corresponding word according to the current input and its context. It can also allow translators to focus on local or global features in the input.

CBHG CBHG is a building block used in the text-to-speech model. It consists of multiple 1-D convolutional filters, followed by highway networks and a bidirectional gated recurrent unit (Bi-GRU), as Figure 2 shows.

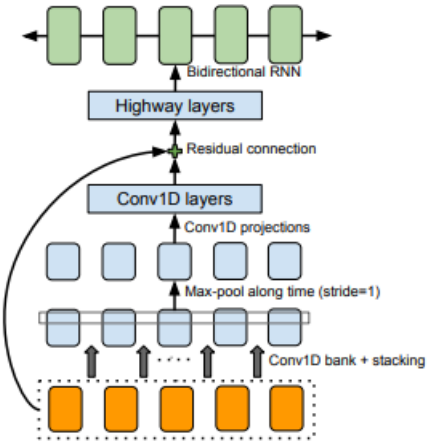

Figure 2: CBHG

4.2 Machine Translation Model

The machine transition model takes sentences with variable length from multiple source languages as inputs and generates corresponding English translation. The model will first train bi-language pairs, then the one-to-one machine translation model will be combined together to become a many-to-one model. The source language varies, but the target language will all be English. Before training, all the inputs have been cleaned, punctuation and special markers are removed. Then the sequence is tokenized and padded. The goal of the encoder is to handle the sequential variable length input data. We used the predefined LSTM layers of tensorflow in the encoder. The decoder is similar to the encoder with an extra BahdanauAttention mechanism embedded to provide context information for the encoder (Zhang et al., 2020).

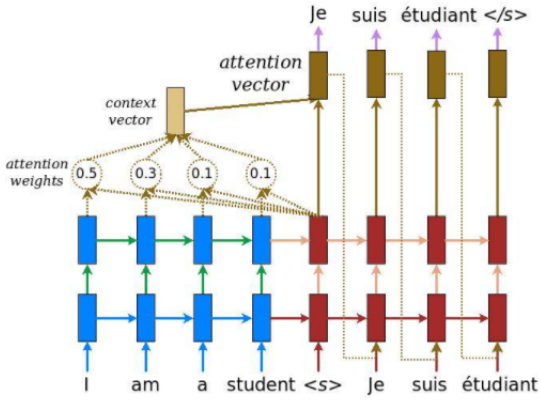

Figure 3: The translation model

4.3 Speech thesis model

The model is an end-to-end generative model that takes characters as inputs and the output is the corresponding raw spectrum. The model includes an encoder, an attention-based decoder and a pose-processing net. The input is a character-based encoder sequence. In the sequence, each character is represented as a one-hot vector and embedded into a continuous vector. The encoder aims to extract robust sequential representation of the text. The decoder is a content-based tanh attention layer, which produces the attention query at each step. The output of the decoder are waveforms which will be generated by using Griffin-Lim Algorithm(GLA). This algorithm is a powerful phase reconstruction that can be used to recover a complex spectrum. The whole process is illustrated in Figure 4 (Mwiti, 2019 accessed Dec 12, 2020).

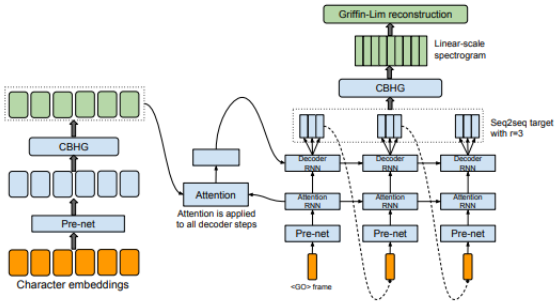

Figure 4: The speech thesis model

5 Analysis of Results

For the translation model, the model achieved overall 0.8 in F1 scores on all the tested languages, as Figure 5 shows. This might because languages trained and tested in the model all have relatively

large datasets compared to other languages. When the language data is limited, the model did not perform well. Compared with the French-English translation(0.837 in F1 score) with around 18000 pairs of sentences, the Catalan-English translation only achieved 0.264 in F1 score, and it only has 685 pairs of sentences in the corpora.

Languege	Precision	Recall	F1-Score
Fra-Eng	0.797	0.882	0.837
Spa-Eng	0.889	0.904	0.896
Deu-Eng	0.803	0.987	0.886
Cmn-Eng	0.732	0.794	0.782
Cat-Eng	0.325	0.223	0.264

Figure 5: Results

The similarity between the source and the target language also have impacts on the performance. Figure 6 shows that compared the attention graph of Mandarin with that of French, the model performs significantly better with French-English translation, regardless of the equal size Mandarin-English sentence pairs have during training.

Comparing the attention graph of the same sentence in French and Chinese, we can see that the model has a hard time determining the weight for Chinese. This may be caused by more fundamental linguistic differences between Chinese and English.

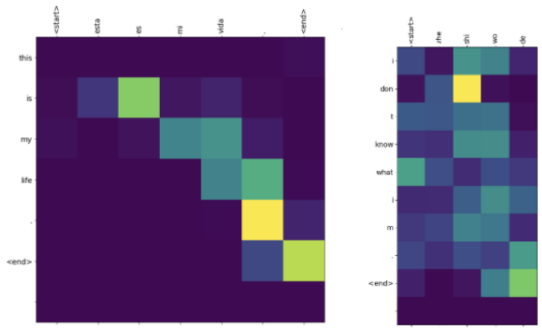

Figure 6: Attention graph for translation comparison

For the speech synthesis, the reasonable alignment were produced after over 13000 global steps of training. The global loss value still has a visible decreasing trend as showed in Figure 7.

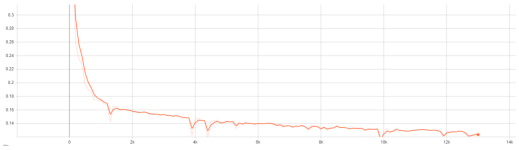

Figure 7: Decreasing trend of global loss value

Figure 8 shows the alignment graph over 3000, 9000, 12000, 13000 steps. It indicates that the reasonable alignment was produced after 12000 global steps. After the same 12000 steps, the synthesized clips can be recognized by humans.

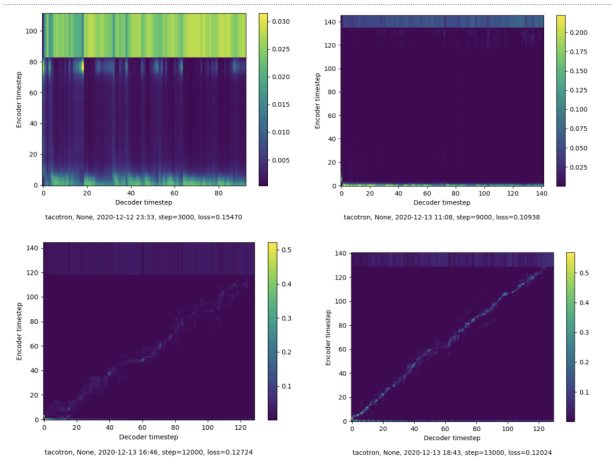

Figure 8: Alignment graph over 3000, 9000, 12000, 13000 steps

The results shows that the speech synthesis model is capable of producing reasonable English sound outputs from generated translation results.

6 Conclusion

This project focused on the seq2seq model and implemented a machine translation model and a speech synthesis model based on it. The translation model takes one sentence of five source languages and returns the translated English sentence. The speech synthesis model then takes the translation result and produces the audio output of the given translation.

The results are comprehensible. However, there are lots of parameters that can be further defined or automatically trained by the model itself. Most of the missed cases happened during the translation process. Therefore, we might need to combine the supervised learning with some linguistic rules to form a hybrid system. It also suffers from lack of labeled data, we could utilize the attention mechanism to build a zero-shot multi-language translation model.

References

D. Mwiti. 2019 accessed Dec 12, 2020. A 2019 guide to speech synthesis with deep learning.

J. Sotelo, Soroush Mehri, K. Kumar, J. F. Santos, Kyle Kastner, Aaron C. Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. In *ICLR*.

T. Tracey. 2019 accessed Dec 10, 2020. Language translation with rnns.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis.

B. Zhang, D. Xiong, and J. Su. 2020. Neural machine translation with deep attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):154–163.