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

Yifan Zhu, Tianjiao Yu, Beixuan Jia, Miao Wang
Georgetown University

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.

Introduction

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 build a machine translation mechanism with audio outputs. 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 spectrums. Models are evaluated with recalol value, BLEU score, and F1 scores.

Related Work

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 to the next part and leads to unwanted outcomes in the end. Based on the cons, an end-to-end training model fashion was expected.

Tacotron is a text-to-speech mechanism generating human-like speech from text. This seq2seq model only takes speech examples and corresponding transcripts as input. The model maps a sequence of letters to a sequence of features that encode the audio. These features mimic pronunciation of words, volume, speed and intonation. Finally, these features are converted to a waveform. Our implementation is largely based on the Tacotron project.

Data Sources

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 were from volunteers who speak various languages. This project focuses on the following language translation pairs: English - Chinese, English - French, English - German, English - Catalan and English - Spanish. Each dataset is sorted by length, with shorter sentences put first, as shown in Figure 1.

```
Hi.     嗨。    CC-BY 2.0 (France) Attribution: tatoeba.org #538123 (CM) & #891077 (Martha)
Hi.     你好。   CC-BY 2.0 (France) Attribution: tatoeba.org #538123 (CM) & #4857568 (musclegirlxyp)
Run.    你用跑的。       CC-BY 2.0 (France) Attribution: tatoeba.org #4008918 (JSakuragi) & #3748344 (egg0073)
Wait!   等等！  CC-BY 2.0 (France) Attribution: tatoeba.org #1744314 (belgavox) & #4970122 (wzhd)
Wait!   等一下！ CC-BY 2.0 (France) Attribution: tatoeba.org #1744314 (belgavox) & #5092613 (mirrorvan)
Hello!  你好。   CC-BY 2.0 (France) Attribution: tatoeba.org #373330 (CK) & #4857568 (musclegirlxyp)
I won!  我贏了。 CC-BY 2.0 (France) Attribution: tatoeba.org #2005192 (CK) & #5102367 (mirrorvan)
Oh no!  不會吧。 CC-BY 2.0 (France) Attribution: tatoeba.org #1299275 (CK) & #5092475 (mirrorvan)
Cheers! 乾杯！   CC-BY 2.0 (France) Attribution: tatoeba.org #487006 (human600) & #765577 (Martha)
Got it? 你懂了嗎？       CC-BY 2.0 (France) Attribution: tatoeba.org #455353 (FeuDRenais) & #7768205 (jiangche)
```

Figure 1. Dataset snapshot: English + TAB + The Other Language + TAB + Attribution

For speech synthesis, this report uses LJ Speech dataset, which consists 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 of 1 to 10 seconds. The transcription was matched to the audio manually, and a QA pass to ensure the accuracy.

Description of Models

Sequence-to-sequence turns one sequence into another. LSTM will be used to avoid vanishing gradients. As shown in Figure 2, each input in LSTM 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, and the decoder reverses the process, turning the vector into output items. During the process, the previous output functions as the context of the input.

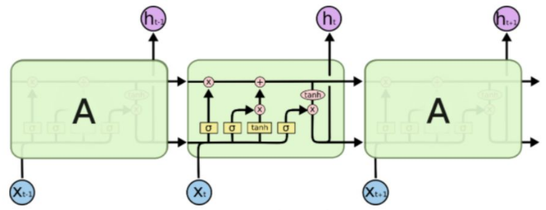

Figure 2. The LSTM Structure

Translation Model takes sentences with variable length from multiple source languages as inputs and generates corresponding English translation. This model will first train bi-lanaguage pairs, and then combine the 1-to-1 translation model to generate a many-to-1 model. All the inputs are cleaned and tokenized. The goal of the encoder is to handle the sequential variable length input data. The predefined LSTM layers of tensorflow is used in the encoder. The decoder is similar to the encoder with an extra attention mechanism embedded in the decoder to provide context information for the encoder, as shown in Figure 3.

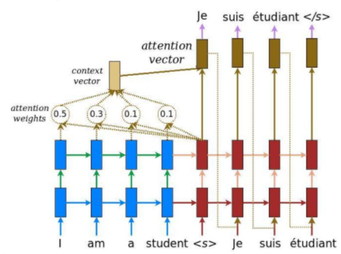

Figure 3. building block with multiple 1-D convolutional filters

Speech thesis model is an end-to-end generative model that takes characters as inputs and the output is the corresponding spectrogram. It includes an encoder, an attention-based decoder and a pose-processing net. The input is a character-based encoder sequence where 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 the waveforms which will be generated by using Griffin-Lim algorithm, which can be used to recover a complex spectrogram. Figure 4 shows the whole process.

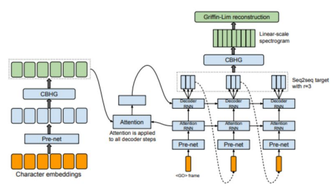

Figure 4. Speech Thesis Mechanism

Analysis of Results

For the translation model, the proposed model achieved over 80 percent f1 scores on all the tested languages. However, the testing languages all have relatively large datasets compared to other languages. The model did not perform well on languages with small datasets. As shown in Figure 5, compared with French to English, which achieved 0.837 F1 score with around 18000 pairs of sentences, the model only achieved a 0.264 F1 score for Catalan with 685 pairs of sentences.

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. Result of Evaluation metrics

The similarity between the source and the target language also have impacts on the performance. Comparing the attention graph for Manderin and French, We can see that the model performs significantly better with French and English pairs, regardless of the equally large dataset of Manderin and English. 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 shown in Figure 6.

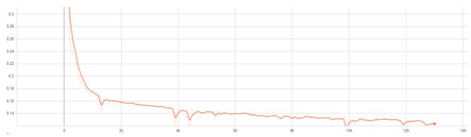

Figure 6. The Trend of Global Loss Value

Comparing the alignment graph over the steps, Figure 7 shows the alignment graph over 3000, 9000, 12000, 13000 steps. The reasonable alignment was produced after 12000 global steps. After the same 12000 steps, the synthesized clips can be recognized by humans although not easily. Forwarding the translation result into the speech synthesis model. It is capable of producing reasonable English sound outputs from the source language.

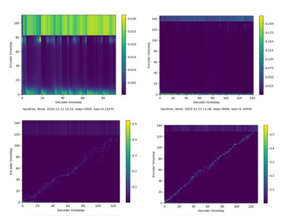

Figure 7. Alignment Graph over 3000, 9000, 12000, 13000 Steps