Neural machine translation from Armenian to English

Zaven Martirosian

May 2020

Abstract

Huawei natural language processing course. Final project.
`https://github.com/zmazak0/Neural-machine-translation`.

1 Introduction

Neural machine translation (NMP) is relatively new an approach to solving the problem of machine translation that has become widespread in recent years.

As a final project, I decided to take the problem of neural machine translation. I tried to find ready-made neural networks for translating from Armenian to English. Nothing was found. So I decided to train my network for this task.

Machine translation solves a lot of problems. This is cheaper than hiring a professional translator. Translates several times faster than a person. Using machine translation, we can easily understand the meaning of the text, written in another language.

1.1 Team

This project was completed individually by **Zaven Martirosian**.

2 Related Work

Today most neural machine translation systems have a similar architecture: encoder, decoder, and attention mechanism. The encoder converts each individual word to a contextual one a vector that the decoder turns into a target sentence word. The attention mechanism allows system focuses on individual parts of the source offer and control the accuracy of the translation.

The encoder and decoder can consist of any number of layers, but today they are most often performed in the form of several layers of long short-term memory

The Transformer, introduced in the paper Attention Is All You Need [Ashish Vaswani, 2017], is a powerful sequence-to-sequence modeling architecture capable of producing state-of-the-art neural machine translation systems.

Recently, the fairseq team has explored large-scale semi-supervised training of Transformers using back-translated data, further improving translation quality over the original model.[Myle Ott, 2018]

3 Model Description

For training, I chose a standard transformer model.

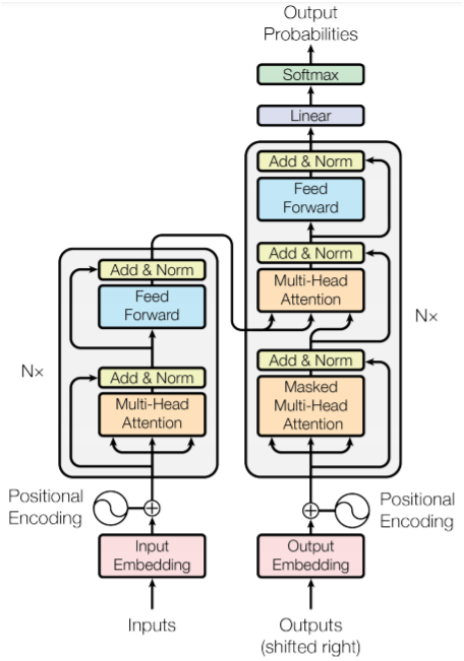

Figure 1: The Transformer

I used masking and positional encoding because this model doesn't contain any recurrence or convolution. Positional encoding give the model some information about the relative position of the words in the sentence. I also used BahdanauAttention.

Trained with the following parameters:
BATCH SIZE - 128
embedding dim - 1024
optimizer - Adam
loss - Sparse Categorical Crossentropy
EPOCHS - 30

4 Dataset

To train the model, I built a dataset myself.

The dataset I collected contains about 10,000 news pages from the site news.am. Of these 10,000 pages, 5,000 were translated and about 20,000 sentences were received.

All sentences were reduced to lowercase, repetitive sentences were removed, and unnecessary punctuation marks were removed. The final dataset was sorted by the length of Armenian sentences.

This dataset is contained in my github, and the password for the archive is "data".

5 Experiments

5.1 Metrics

Since the Bleu metric is not differentiable, I trained the model using SparseCategoricalCrossentropy.

5.2 Experiment Setup

The training dataset contains 0.85 parts of the total collected dataset - 16904 sentences, and the test dataset contains the remaining 0.15 - 2984 sentences.

There were about 10 different model launches with different hyperparameters. The best quality was achieved with these parameters:

BATCH SIZE - 128
embedding dim - 1024
optimizer - Adam
loss - Sparse Categorical Crossentropy
EPOCHS - 30

5.3 Baselines

As a baseline, I used the seq2seq model from an encoder and decoder without attention.

6 Results

Since I could not find a trained model for translating text from Armenian to English, I decided to compare this model with a Tranfformer who was trained to translate EN-DE and EN-FR using the BLEU metric.

This low result was obtained because I did not train the model on such big data as WMT 2014 English-to-French

	EN-DE	EN-FR	ARM-EN
BLEU	28.4	41	13,92

Table 1: Comparison of English-to-German, English-to-French transformers and the Armenian-to-English model I trained using the bleu metric.

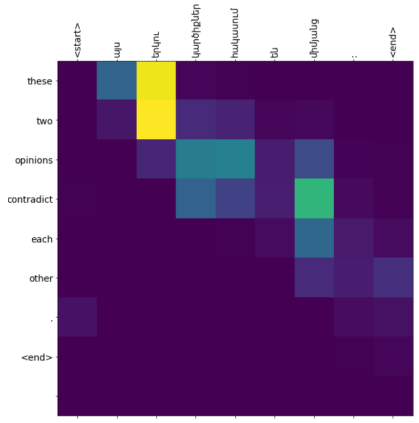

Figure 2: Attention

7 Conclusion

I have compiled a dataset for translation from Armenian to English.

The dataset includes about 10,000 pages from a news site news.am. I managed to translate half of them into English using a Google translator. It turned out about 20,000 semtences in the dataset, which I used to train the model.

I compared the existing solutions and decided to choose for training Transofrmer, because it shows the best quality. I couldn't improve the basic model.

I couldn't compare the received model with any other one that translates from Armenian to English, because I couldn't find such models on the Internet.

In the future, I plan to translate all the sentences to the end and train the model on the full amount of data.

References

[Ashish Vaswani, 2017] Ashish Vaswani, Noam Shazeer, N. P. (2017). Attention is all you need.

[Myle Ott, 2018] Myle Ott, Sergey Edunov, A. B. (2018). A fast extensible toolkit for sequence modeling.