Machine Learning Advancements

Introduction

Majority of the today’s challenges with machine learning revolve around the insurmountable about of data that needs processing. However, language data has a different issue, its main issue is the lack of data. Language data is scarce and sometimes nonexistent. Google has been trying to tackle this issue because majority of their products help us communicate with each other through apps on our phones, smart home devices, and other products. The challenge has been to make all their products available throughout the world on many different platforms with support for different languages.

There are generally two transition approaches to language translation. First, is machine translation. The challenges of machine translation are that there are so many rules for different languages. Because of the natural ambiguity and flexibility of human language, each language may operate at different lexical, syntactic, or semantic levels that makes it very difficult for one machine model to be truly universal. The ability to transfer results to another language without faults will be a difficult feat to accomplish. Not only does this require specialists to aggregate all the rules for a language, but engineers need to become familiar with them as well.

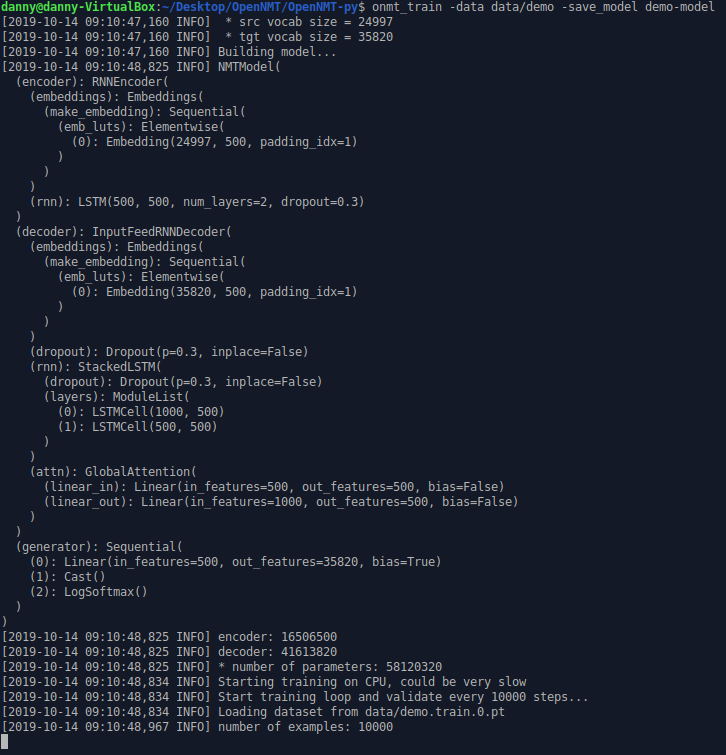
The second approach is called statistical neural translation. Statistical neural translation approach tries to determine the highest probability of generating a good translation based of a statistical model. This model quickly outperforms machine translation but comes nowhere near neural machine translations models in terms of performance. This model works particularly well with short phrases but fails on larger texts.

The latest and state of the art in language processing utilizes neural machine translation. Google has named their version Google NMT. Over the years, Google have done enormous strides with their neural machine translations. They have been striving to adopt a single translation model based on the notion that “the learning signal from one language should benefit the quality of translation to other languages.” Neural machine translation uses neural networks to learn a statistical model for machine translation. The model can learn on input and output text directly, thus removing the requirements for rules of each of the languages. The primary advantages of a neural machine translation is that it is truly end to end. Cons of neural machine translation systems however include slow speed of training the models, ineffectiveness in dealing with rare words, and sometimes failure to translate all words in a sentence.

The goal of this universal neural machine model is to be capable of translating between any language pair. On languages with very little training data, performance is improved on the languages with very little training data due to positive transfer. But what is positive transfer? Positive transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task. However, the cons of this are that languages with extremely high amount of data tend to get worse. Google has found that the more languages added to this universal model, the quicker the quality drops.

As a result, while a universal NMT may be good, it is still in its infant stages. Even though Google’s NMT performs 60% better than its phrase-based implementation used in its current applications, it is still very computationally expensive. Thus, this suggests that developers might want to keep their strong baseline languages and use the universal NMT to be a catch all solution for the other languages which lack data. The amount of the data used by Google in this research endeavor is on the order of 25 billion sentences. This was achieved by crawling and extracting parallel sentences from the web. Because all the sentences were extracted from the web, any lack of data in training set would require serious research work to resolve because this data obtained was actual data by people and not generated.

The following open source project is an example of an NMT to give the user a sense of how an NMT implementation works. <https://github.com/OpenNMT/OpenNMT-py>



In this opensource version of NMT, training can be run on the GPU or CPU. The figure above is done on a virtual machine via CPU. If a user wants to OpenNMT, they should preferably have a dedicated Linux machine with a GPU having CUDA cores.

Ultimately, while the intention that Google has is good, there is still a long way to go. This is a project research over 5 years in the making and still will not be reaching maturity for several years. Over half of 7,000 languages will no longer be spoken by the end of the century. If Google succeeds, they hope to preserve these languages with this universal neural machine translation model.

References

<https://github.com/OpenNMT/OpenNMT-py>

examples of an open source NMT in action.

<https://ai.googleblog.com/2019/>

<https://slator.com/technology/whats-so-massive-about-googles-massively-multilingual-neural-machine-translation/>

<https://machinelearningmastery.com/introduction-neural-machine-translation/>