Language Identification with Contextual Translation

Alex Shah

Initially:
English/Spanish only
But expandable

2 Main Components

Identify the Input Language

“Hola” → Es

“Hello” → En

Translate the input to the opposite language

“Hola” → Es → “Hello” (En)

“Hello” → En → “Hola” (Es)

Goals

- Use latest and greatest models and methods
- Doesn’t require training every detection/translation direction
- Doesn’t require huge amount of training to add languages
- Combine two networks in one product

Dataset

- Europarl: A dataset in 21 languages comprised of European Parliament proceedings
- Hand translated
- En_Es dataset:
  - 187 MB
  - 1,965,734 sentences
  - 51,575,748 words

Text Classifier

- Use Convolutional Network (CNN) with a few specialized tweaks
  - Example: different languages use different accents/characters, this can tell us which language it is
- Easiest part to get good accuracy
  - As little as 10 epochs of training could identify 140 char input sequences with 90%+ accuracy

Translation

- A quickly evolving field
  - Accuracy and BLEU scores are an arms race
  - BLEU Score: automated metric to determine accuracy of translation compared to human translation
- Contextual Translation
  - Take into account surrounding characters/words to translate more closely to a human translation
    - Gender agreement
    - Tenses
    - Advanced grammar structures

Google’s latest: Tensor2Tensor

Google Research Blog

The latest news from Research at Google

Accelerating Deep Learning Research with the Tensor2Tensor Library

Monday, June 19, 2017

Posted by Łukasz Kaiser, Senior Research Scientist, Google Brain Team

Deep Learning (DL) has enabled the rapid advancement of many useful technologies, such as machine translation, speech recognition and object detection. In the research community, one can find code open-sourced by the authors to help in replicating their results and further advancing deep learning. However, most of these DL systems use unique setups that require significant engineering effort and may only work for a specific problem or architecture, making it hard to run new experiments and compare the results.

Today, we are happy to release Tensor2Tensor (T2T), an open-source system for training deep learning models in TensorFlow. T2T facilitates the creation of state-of-the art models for a wide variety of ML applications, such as translation, parsing, image captioning and more, enabling the exploration of various ideas much faster than previously possible. This release also includes a library of datasets and models, including the best models from a few recent papers (*Attention Is All You Need*, *Depthwise Separable Convolutions for Neural Machine Translation* and *One Model to Learn Them All*) to help kick-start your own DL research.

Translation Model	Training time	BLEU (difference from baseline)
Transformer (T2T)	3 days on 8 GPU	**28.4 (+7.8)**
SliceNet (T2T)	6 days on 32 GPUs	26.1 (+5.5)
GNMT + Mixture of Experts	1 day on 64 GPUs	26.0 (+5.4)
ConvS2S	18 days on 1 GPU	25.1 (+4.5)
GNMT	1 day on 96 GPUs	24.6 (+4.0)
ByteNet	8 days on 32 GPUs	23.8 (+3.2)
MOSES (phrase-based baseline)	N/A	20.6 (+0.0)

BLEU scores (higher is better) on the standard WMT English-German translation task.

T2T
Transformer

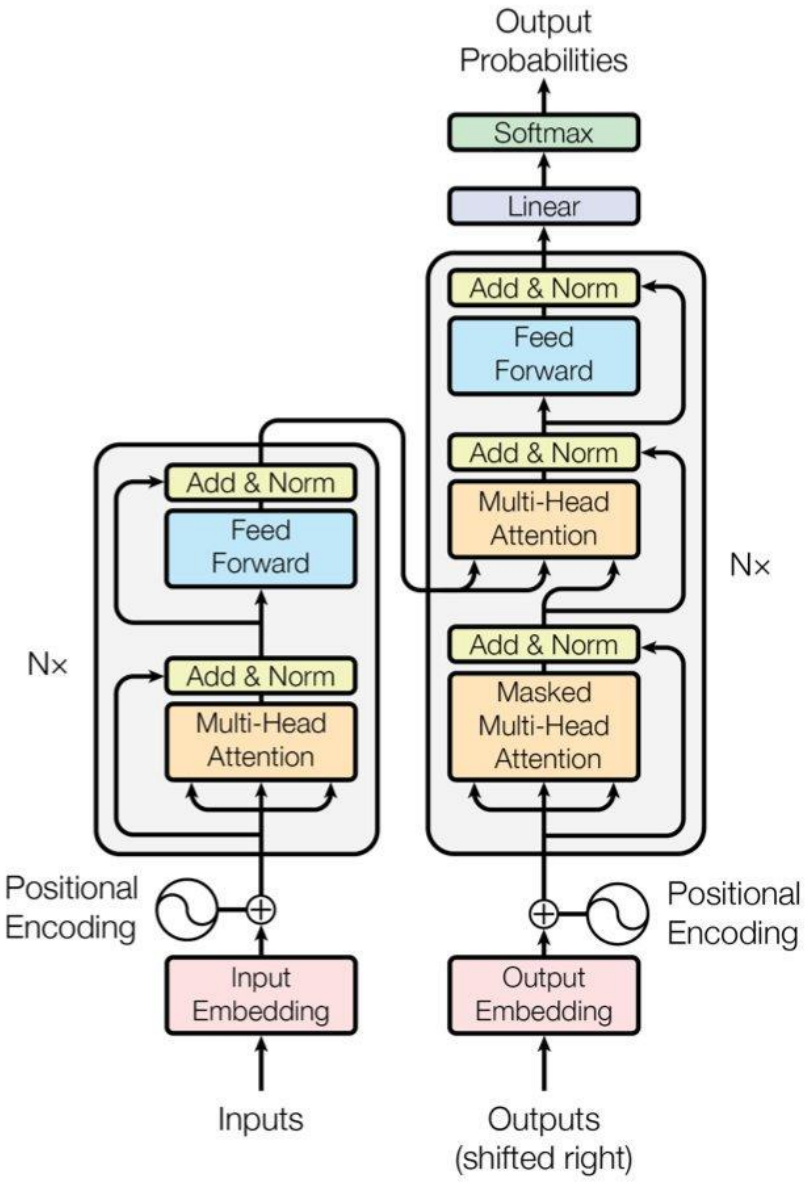

Figure 1: The Transformer - model architecture.

Transformer model (Tensor2Tensor)

Pros:

- Highest BLEU score to date
- Training time/GPU requirements low in comparison to other methods

Cons:

- Highly complex: You need a hands off approach to implement it
  - Use T2T as a library of models, can’t build from scratch

Efficiency

- Training a new language in the classifier is easy
  - The classifier trains for all languages at once
  - CNN classifier is already highly accurate with little training time/power
- Training a new language in the translation net is efficient
  - Adding a new language does not require retraining other languages.
    - All languages are trained into representations
      - Do not need to train every language pair, nor every direction
      - Translation treated like a feature set, just need to learn the features of the new language

Overview

- Use latest and greatest models and methods
  - Specialized CNN → high accuracy, quick
  - Transformer (tensor2tensor) → highest BLEU score
- Doesn’t require training every detection/translation direction
- Doesn’t require huge amount of training to add languages
- Combine two networks in one product
  - Classifier
  - Machine Translation

→ DIY Google Translate

Sources

https://research.googleblog.com/2017/06/accelerating-deep-learning-research.html

https://github.com/tensorflow/tensor2tensor

https://github.com/tensorflow/nmt

https://github.com/google/seq2seq

http://www.statmt.org/europarl/

https://nlp.stanford.edu/projects/nmt/

http://www.nltk.org/

https://research.googleblog.com/2017/07/building-your-own-neural-machine.html

I. Goodfellow, et.al, Deep Learning.