Bonus Point: Appendices to the report

Attention is all you need

Definition search:

- Encoder-decoder models: this is a two-part architecture (an encoder that reads an input sequence and transforms it into an internal representation (a kind of digital summary), and a decoder that takes this sequence and generates the output)
- Recurrent or convolutional architectures complex (recurrent = reads the text word by word in order) (convolutional = uses filters to capture pieces of sentences)
- BLEU score: this is a metric that measures the quality of a machine translation; the higher the score (out of 100), the more accurate the translation

Introduction

Neural machine translation (NMT) enables the syntactic translation of a sentence. It takes a sequence as input (e.g., a sentence), transforms it into an internal representation, and then uses a decoder to generate the target sequence (e.g., the same sentence but in another language).

Background

The old methods were less practical; they were both the encoder and decoder, but they had difficulty parallelizing calculations and required long training times. Their goal was to create a simple architecture that was more efficient in training, which is why they decided to create the Transformer, which is based on a self-attention mechanism. This mechanism allows for better representation of text and parallelization of calculations without using recursion or convolutions.

Model Architecture

Transformer is built with an encoder-decoder architecture. Each part is composed of several identical layers. The encoder takes the text as input and transforms it into internal representations using two elements: multi-head self-attention (which connects each word to the others in the sentence) and a feed-forward neural network that refines these representations. The decoder looks at what the encoder has produced to generate the output word by word. In addition, a masking system is used so that the model cannot cheat by seeing future words.

Why Self-Attention

The creators of Transformer chose self-attention rather than recurrent or convolutional models because its advantage is that it connects all positions at once, unlike RNNs, which must read word by word. As a result, it is much faster and more efficient, especially when sequences are long. Compared to convolutions, it is also more practical, because CNNs must stack several layers to successfully connect all the words together.

Finally, another big advantage is that self-attention allows you to see what the model is focusing on (thanks to attention weights), which makes the model more interpretable.

Training

For English-German training, they trained their model on 4.5 million sentence pairs, and for English-French on 36 million sentences, separated into tokens with a vocabulary of 32,000 words. The models were trained on a single machine equipped with 8 GPUs. For the basic models, each training step took 0.4 seconds, and they trained the basic models for about 12 hours. For the large models, each step took 1 second, and training took about 3.5 days.

To optimize their model, they used Adam with specific hyperparameters, and the learning rate varied according to a precise formula.

Finally, three regularization methods were implemented:

- Residual Dropout: This is when the model is built by adding an input to the output of a block and so on, which adds a little randomness to each pass through the network to prevent it from memorizing instead of learning.
- Label Smoothing: When training a classification model, the target is usually a single word, so the model can be overly confident. With label smoothing, the model is taught to be cautious and avoid being overly confident about a single answer, which makes it more accurate.

Results

When testing their Transformer model, the researchers focused primarily on machine translation and achieved excellent results with a BLEU score of 28.4 for English-German and 41.0 for English-French. What's more, it trains much faster than previous models.

They also tested different variants to see what affected performance and found that:

- Using larger models yielded better results.
- Dropout is very important to avoid overfitting.

Finally, to see if the Transformer could be used for purposes other than translation, they tried it on another task: identifying grammatical structure in English sentences. The result was that it achieved scores as good as specialized models.

Conclusion

In this article, researchers present Transformer, the first translation model based solely on attention, without recurrent or convolutional layers. As a result, it is much faster to train and achieves better scores than previous models.