Attention is All You Need

Andrew Drozdov

September 18, 2017

What?

Transformer Network

Why? (Architecture Perspective)

- Many NLP tasks have SotA set by LSTM or GRU, models with sequential dependencies making them difficult to paralellize.
- There are other popular architectures lately:
  - QRNN / SRU
  - CNN

  But they still usually have some sequential dependencies.
- The model in AIAYN has no sequential dependencies.
- Attention-only model does exist, but not for decoding.

Why? (Performance Perspective)

- Transformer Network can be used for many tasks:
  - WMT 14 (En to Ge). 28.4 BLEU (+2)
  - WMT 14 (En to Fr). 41.0 BLEU (single model, fast/easy to train)
  - Constituency Parsing

  Some of these numbers are more impressive than seen in some similar papers.

Background (RNNs)

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Background (LSTM/GRU and motivation for QRNN)

Equations...

Model: Transform Network (Encoder)

Image

Model: Transform Network (Decoder)

Image

Scaled Dot-Product Attention

Divide by the square root of the output dimension to increase the signal from training.
This equation is the same as Softmax plus Temperature. The temperature will always be positive and greater than one, so it will in effect increase the entropy of the probabilities, bringing them closer to uniform.

Dot Product

Serial.

```
for query in queries:
  attn = query.view(1, D) * keys.view(K, D)
  attn = attn.sum(1) / scale
  attn = softmax(attn)
```

Parallel.

```
attn = queries.repeat(1, K).view(Q * K, D) * keys.repeat(Q, 1)
attn = attn.sum(1) / scale
attn = softmax(attn)
```

Parallel with broadcasting.

```
attn = querys.view(Q, 1, D) * keys.view(1, K, D)
attn = attn.view(Q * K, D).sum(1) /  / scale
attn = softmax(attn)
```

Multi-Head Attention

Perform attention multiple times (almost like an attention "kernel").

Positional Embeddings

Something similar is done in the Intra-Attention for NLI paper.

Intelligent Batching

"Sentence pairs were batched together by approximate sequence length." This could mean that all source match and all target match, or that the length of the pair match.

WMT 14

Training Time: "We estimate the number of floating point operations used to train a model by multiplying the training time, the number of GPUs used, and an estimate of the sustained single-precision floating-point capacity of each GPU."