CIS 607: Deep Learning for NLP
How Much Attention Do You Need?

Steven Walton

March 14, 2020

1 Introduction

This paper is a response to Google Brain/Research's *Attention Is All You Need*. If attention is all that you need, then the obvious question is "how much?" This paper introduces the Architecture Definition Language, also known as ADL, and then uses it to determine where attention is useful, how much is useful, and how it compares to other network architectures. Specifically this paper compares transformer networks to Convolutional Neural Networks, CNNs, and Recurrent Neural Networks, RNNs. This work differentiates itself from other works by examining specifically transformers instead of self-attention. The work found that source attention on lower encoder layers isn't beneficial, multiple attention and residual feed-forward layers are most beneficial, and self-attention is most beneficial on the encoder side of the network.

2 Architecture Definition Language

To easily experiment with different architectures the authors introduce the Neural Machine Translation Architecture Definition Language, or NMT ADL for short. NMT is used to perform sequence to sequence prediction tasks using machine learning. Specifically, a source sentence X is translated auto-regressively into a target sentence Y, token-wise.

$$p(y_t|Y_{1:t-1}, X, \boldsymbol{\theta}) = softmax(\boldsymbol{W}_o \boldsymbol{z}^L + \boldsymbol{b}_o)$$

where $\boldsymbol{W}_o$ projects a model dependent hidden vector $\boldsymbol{z}^L$ and $\boldsymbol{b}_o$ is the basis vector.

2.1 ADL

ADL is used to define standard NMT architectures. ADL is composed of many layers, l. Every layer has a definition based on the hidden states of the previous layer. Layers are chained together with the notation $l_1 \rightarrow l_2 \rightarrow \cdots \rightarrow l_L$ or $(l_L(...(l_2(l_1(\boldsymbol{H}^o)))))$ if there are no attention layers, where $\boldsymbol{H}^o$ is the matrix of hidden states. This notation is also shortened by the convention $repeat(n, l) = l_1 \rightarrow \cdots \rightarrow l_L$. ADL includes: dropout, fixed positional embedding, linear layers, feed-forward layers, convolutions, identity, concatenations, recurrent neural networks, attention, dot product attention, MLP attention, source attention, self-attention, layer normalization, and residual layers.

A set of standard architectures are also used. These include: RNMT, ConvS2S, and Transformers. RNMT is defined as follows. The encoding layer is represented by

$$\boldsymbol{U}^{L_s} = dropout \rightarrow birnn \rightarrow repeat(n-1, res\_d(rnn))$$

The decoder is represented by

$$\boldsymbol{Z}^L = dropout \rightarrow repeat(n, res\_d(rnn)) \rightarrow concat(id, mlp\_att) \rightarrow ff$$

ConvS2s has an encoding layer defined as

$$\boldsymbol{U}^{L_s} = pos \rightarrow repeat(n, res(cnn(glu) \rightarrow dropout))$$

and a decoding layer

$$\boldsymbol{Z}^L = pos \rightarrow res(dropout \rightarrow cnn(glu) \rightarrow dropout \rightarrow (dot\_src\_att(s=1)))$$

The transformer has an encoding block

$$t_{enc} = res\_nd(mh\_dot\_self\_att) \rightarrow res\_nd(ffl)$$

and a decoding block

$$t_{dec} = res\_nd(mh\_dot\_self\_att) \rightarrow res\_nd(mh\_dot\_src\_att) \rightarrow res\_nd(ffl)$$

The transformed encoder is

$$\boldsymbol{U}^{L_s} = pos \rightarrow repeat(n, t_{enc}) \rightarrow norm$$

and the decoder

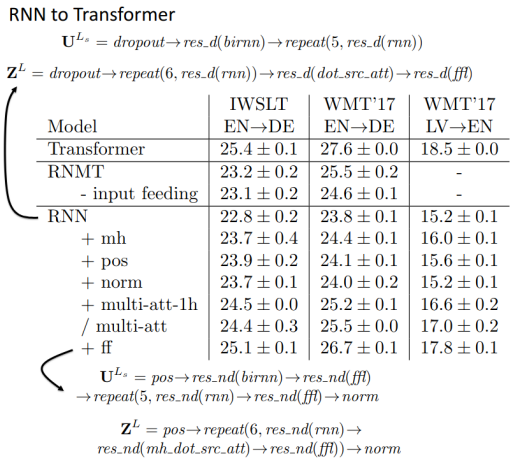
$$\boldsymbol{Z}^L = pos \rightarrow repeat(n, t_{dec}) \rightarrow norm$$

3 Setup/Experiment

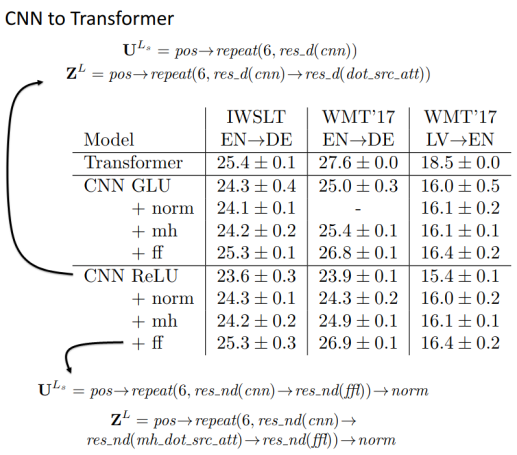
The authors use a package called SOCKEYE to parse arbitrary model definitions, defined with ADL. They then test the models on the WMT and IWSLT datasets. Models were trained with 6 encoder and 6 decoder layers. Convolution layers all had kernel sizes of 3 and RNNs used LSTM cells. By testing where attention is the authors got the following results, BLEU scores. From these scores it is clear that attention on the upper encoding block is most important, but the effects aren't that strong.

Encoder block	IWSLT	WMT'17
upper	25.4 ± 0.2	27.6 ± 0.0
increasing	25.4 ± 0.1	27.3 ± 0.1
decreasing	25.3 ± 0.2	27.1 ± 0.1

In the next experiment the authors adjusted their base RNN and CNN models and made them more transformer like. It can be seen from the results that the more transformer like their scores also converge to the transformer BLEU score. The last



Model	IWSLT EN→DE	WMT'17 EN→DE	WMT'17 LV→EN
Transformer	25.4 ± 0.1	27.6 ± 0.0	18.5 ± 0.0
RNMT	23.2 ± 0.2	25.5 ± 0.2	-
- input feeding	23.1 ± 0.2	24.6 ± 0.1	-
RNN	22.8 ± 0.2	23.8 ± 0.1	15.2 ± 0.1
+ mh	23.7 ± 0.4	24.4 ± 0.1	16.0 ± 0.1
+ pos	23.9 ± 0.2	24.1 ± 0.1	15.6 ± 0.1
+ norm	23.7 ± 0.1	24.0 ± 0.2	15.2 ± 0.1
+ multi-att-1h	24.5 ± 0.0	25.2 ± 0.1	16.6 ± 0.2
/ multi-att	24.4 ± 0.3	25.5 ± 0.0	17.0 ± 0.2
+ ff	25.1 ± 0.1	26.7 ± 0.1	17.8 ± 0.1



Model	IWSLT EN→DE	WMT'17 EN→DE	WMT'17 LV→EN
Transformer	25.4 ± 0.1	27.6 ± 0.0	18.5 ± 0.0
CNN GLU	24.3 ± 0.4	25.0 ± 0.3	16.0 ± 0.5
+ norm	24.1 ± 0.1	-	16.1 ± 0.2
+ mh	24.2 ± 0.2	25.4 ± 0.1	16.1 ± 0.1
+ ff	25.3 ± 0.1	26.8 ± 0.1	16.4 ± 0.2
CNN ReLU	23.6 ± 0.3	23.9 ± 0.1	15.4 ± 0.1
+ norm	24.3 ± 0.1	24.3 ± 0.2	16.0 ± 0.2
+ mh	24.2 ± 0.2	24.9 ± 0.1	16.1 ± 0.1
+ ff	25.3 ± 0.3	26.9 ± 0.1	16.4 ± 0.2

experiment the authors performed was placing attention at different locations. The results are as follows. From these results it is clear that self attention on the encoding side is more influential than on the decoding side.

		IWSLT EN→DE	WMT'17 EN→DE		WMT'17 LV→EN	
Encoder	Decoder	BLEU	BLEU	METEOR	BLEU	METEOR
self-att	self-att	25.4 ± 0.2	27.6 ± 0.0	47.2 ± 0.1	18.3 ± 0.0	51.1 ± 0.1
self-att	RNN	25.1 ± 0.1	27.4 ± 0.1	47.0 ± 0.1	18.4 ± 0.2	51.1 ± 0.1
self-att	CNN	25.4 ± 0.4	27.6 ± 0.2	46.7 ± 0.1	18.0 ± 0.3	50.3 ± 0.3
RNN	self-att	25.8 ± 0.1	27.2 ± 0.1	46.7 ± 0.1	17.8 ± 0.1	50.6 ± 0.1
CNN	self-att	25.7 ± 0.1	26.6 ± 0.3	46.3 ± 0.1	16.8 ± 0.4	49.4 ± 0.4
RNN	RNN	25.1 ± 0.1	26.7 ± 0.1	46.4 ± 0.2	17.8 ± 0.1	50.5 ± 0.1
CNN	CNN	25.3 ± 0.3	26.9 ± 0.1	46.1 ± 0.0	16.4 ± 0.2	47.9 ± 0.2
self-att	*combined*	25.1 ± 0.2	27.6 ± 0.2	47.2 ± 0.2	18.3 ± 0.1	51.1 ± 0.1
self-att	*none*	23.7 ± 0.2	25.3 ± 0.2	43.1 ± 0.1	15.9 ± 0.1	45.1 ± 0.2

Table 5: Different variations of the encoder and decoder self-attention layer.

4 Conclusion

In this paper the authors perform experiments to answer when and where attention is most useful in NLP problems. The paper found conclusive results at the effectiveness of attention models and found that attention and feed-forward layers were the most helpful. The experiments also showed that attention on the encoding side is more important than the decoding side.