Attention Is All You Need

Abstract

Transduction models are heavily reliant on complicated recurrent or convolution neural networks, that are formed by combining an encoder and a decoder. The best-performing examples also connect the encoder and decoder through an engagement instrument.
Introduction

Equal contribution. Listing order is random. Jakob had the idea to switch from RNNs to self-attention and initiated the process for its assessment. Ashish, aided by Illia, designed and developed the first Transformer models and has been thoroughly involved with each element of this research. Noam was an invaluable asset to the project, advocating for the use of scaled dot-product attention, multi-head attention and parameter-free position representation. He contributed significantly in each stage of the project. Niki's hard work and dedication to the project was evident as she designed, implemented, tuned and evaluated multiple models in both our original codebase and tensor2tensor.

Background

To minimize sequential calculation, researchers have turned to convolutional neural networks as the primary component for constructing applications like Extended Neural GPU, ByteNet and ConvS2S. These networks enable concurrent evaluation of each input/output position enabling faster computation of hidden representations.

Model Architecture

Neural sequence transduction models are used to participate in competitions and they normally have an encoder-decoder structure. This approach maps input sequences of symbols into continuous representations represented by z =

Encoder and Decoder Stacks

The encoder consists of a stack of 6 layers that are all identical. Each layer has two sub-layers. The architecture consists of two components; the first being a multi-head self-attention mechanism, and the second being a straightforward, fully connected feed-forward network with a positionwise setup.

Attention

An attention function basically maps a query, along with a set of key-value pairs, to an output vector. All these elements are represented as vectors in the process.

Attention = softmaxV

The two predominant attention mechanisms are additive and dot-product attention. They both enable models to identify the most relevant parts of an input for use in the output. The dot-product attention mechanism is similar to our algorithm, but with a scaling factor of $\sqrt{1}d$. For additive attention, the compatibility function is generated by an artificial neural network that has a single hidden layer.

To explain why dot products tend to become large, suppose that q and k's components are independently chosen at random.
P k variables with mean 0 and variance 1.

MultiHead = ConcatW O where headi = Attention

The predictions depend on a set of matrices WiQ, WiK, WiV and WO with dimensions dmodel × dk, dmodel × dk, and dmodel × dv respectively. Additionally, it is also dependent on Rhdv × dmodel..

Transformer architecture incorporates multi-head attention and applies it in three different ways.

• In «encoder-decoder attention» layers, the queries originate from the prior decoder layer, with the memory keys & values emanating from the encoder's output. Thanks to this ability, every position in the decoder can pay attention to each position in the input sequence.
FFN = maxW2 + b2

Although the linear transformations stay consistent throughout the different layers, the parameters altered within each layer can differ.

Embeddings and Softmax

Same as other sequence transduction models, we use acquired embeddings to convert the input & output tokens into vectors with a dimension of dmodel.

linear transformation, similar to .

O tokens in the sequence.