Transformer

Abstract

介绍模型很简单，只有原创的基于编码器和解码器的 Transformer 架构，没有任何的循环和卷积网络。

BLEU Score 很高。（机器翻译与人工翻译的相似度高）。

Background

传统的卷积神经网络想把某一层两个相隔较远的信息汇聚到一个卷积核需要很多很多层，

如果使用 Transformer 里的 Attention 机制则每一次都能看见某一层内所有的信息。

但是 Attention 一次只能看一层，不能像卷积神经网络那样多通道输出，

所以使用了多头注意力机制模拟多通道输出。

Transformer 是第一个只依赖自注意力机制架构的模型！

Architecture

Transformer的架构由 Encoding Block 和 Decoding Block 组成

Block由 Multi-Head attention 、Add & Norm 和 Feed Forward 组成

Encoding & Decoding

Encoding会将输入$X = (x_1, x_2, \dots x_n)$表示成$Z = (z_1, z-2, .., z_n)$

Z是X的向量表示，Decoding 会拿到 Encoding 的输出生成$Y = (y_1, y_2, \dots y_m)$

注意这里的n和m可以相同，但大部分时候是不相同的。

Auto Regressive 自回归模型

在过去时刻的输出又是你当前时刻的输入，所以Y是逐个生成的，先生成y_1,再用Attention加上y_1生成y_2

这也就解释了我们再翻译的时候为什么target是一个一个出现。

Transformer Encoding Block

每个 Encoding Block 重复六个Sub Layers。

Multi-Head Attention

将V、K、Q经过几个线形层降低维度再使用缩放点积模型 (Scaled Dot-Product Attention)

输出进行Concat连接和全连接层后输入到下一个 Multi-Head Attention 层。

Add & Norm

Resident Connection + Layer Normlization

Layer Normlization

简单理解相当于将 Batch Normlization 转置进行归一化再转置回去。具体原理看Layer Normlization原文。

LayerNorm(x + Sublayer(x))，Sublayer(x) 指 Self Attention 和 MLP 。

Feed-forward network

单隐藏层的前馈神经网络MLP

Transformer Decoding Block

Decoding Block 相比 Encoding Block 多了一个 Masked Multi-Head Attention

第二个不是自回归，因为他的K和V来自Encoding Block的输出

Decoding Block 的输出进入一个 Linear 层，做一个 Softmax，得到输出。

Conclusion