Perceiver （IO）

本文为Perceiver & Perceiver IO 的阅读笔记

Perceiver

本文提供了一种基于transformer的可用于多模态人工智能的网络。目前很多网络往往处理的是单模态问题，如2D CNN网络用于处理图像分类问题，而这种2D网格式的方法，在audio（音频）方面就没有这么适用，处理音频问题用的更多的是1D convolutions 或者 LSTM。这种基于先验知识的预设归纳偏置，在处理特定问题上表现优异，但却不符合人类的直观感知，本文提出的方法即可以处理多模态问题，且不用对某种模态进行特定的预处理。

本文的思想其实非常简单，先回顾一下普通的transformer

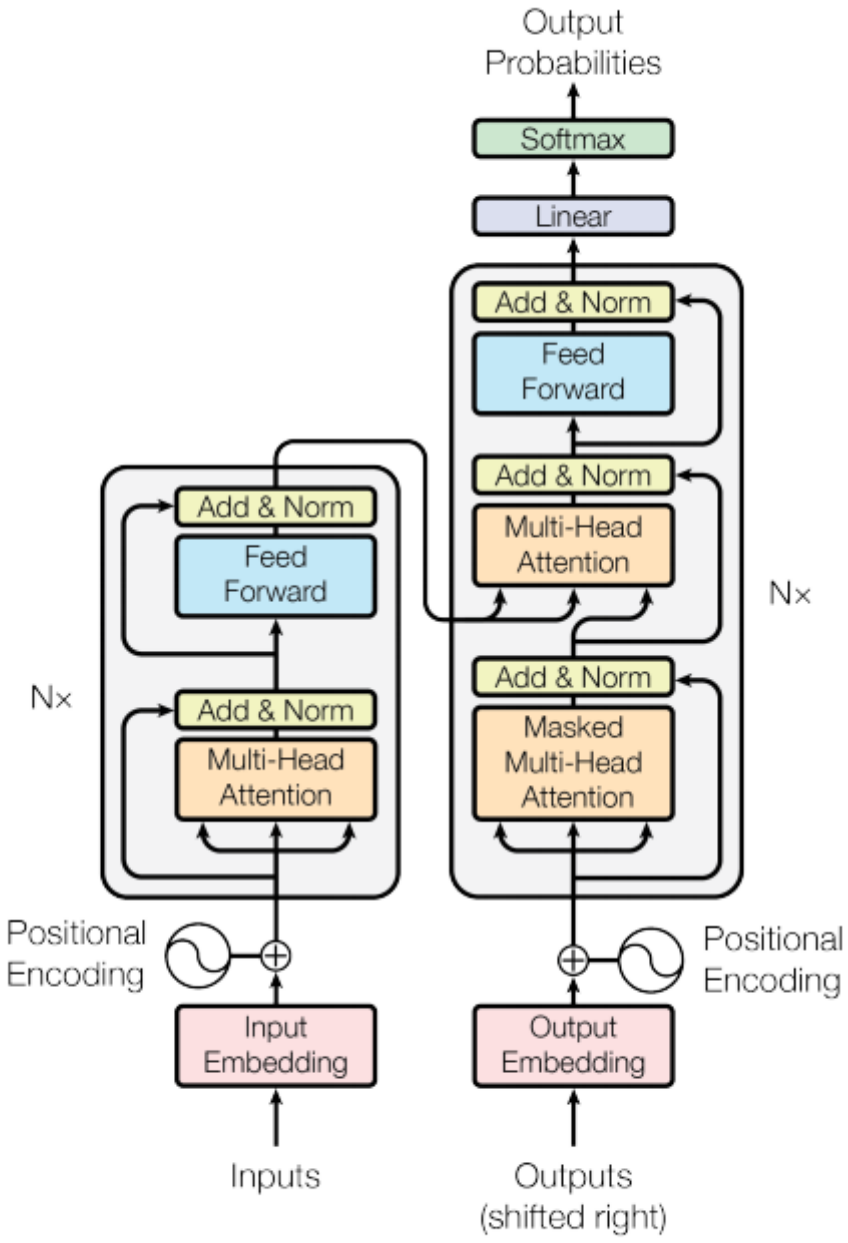


Figure 1: The Transformer - model architecture.

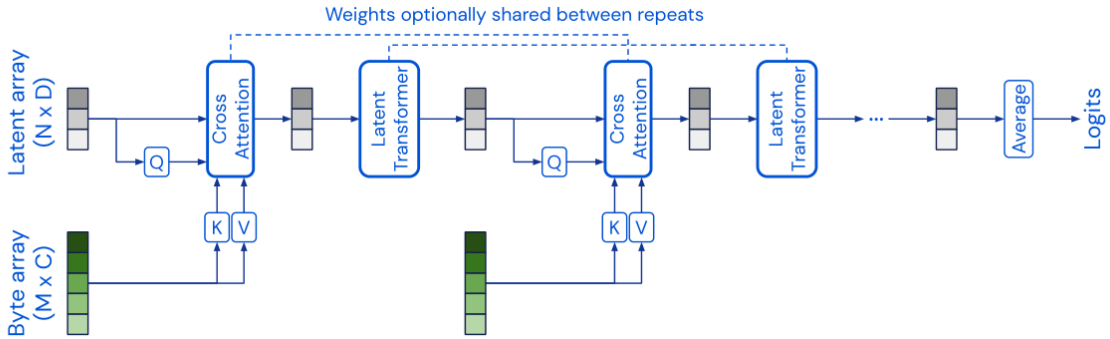
在transfomer的cross-attention部分，inputs为kv，而outputs为q，假设K和V的shape分别为(M, D)和(M, C)，Q的shape为（N，C），根据公式可得

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

最后得到的结果shape为（N,C）

上述中M为输入的维度，与输入数据有关，无法更改，如果我们手动设置N<<M，就可将计算复杂度从O（MM），降到O（MN）

这就是Perceiver的核心理念



Byte array为数据输入，可以为音频，图像，视频等，array可能较大，作为KV，而Latent array为预设的query，N较小，将Latent array与Byte array做cross-attention之后，再进行L次self-attention，最终的计算复杂度为O（MN）+O(LNN)，远远小于传统transformer的O（MM），此处将输入信息的计算复杂度和网络深度的计算复杂度解耦，用一个较小的query，经过一个较深的网络，获得Byte array中的信息

实验结果表明，该方法可以之间在ImageNet上处理224*224的图片，而不用VIT中的patch，top-1精度与resnet-50匹敌；在AudioSet上，超越了sota；在点云领域，ModelNet40上，保持可比性一致的前提下，取得了sota的效果

Perceiver IO

本文是基于Perceiver的一点改进，Perceiver在处理多模态数据时表现极佳且计算规模只会随输入线性增长，而Perceiver也有他的一些问题，由于输出较小，他只能处理比较简单的任务，比如分类任务。本文为了解决这个问题，在输出处加了一个decoder，将latent space decode成需要的大小。

本文的结构被作者总结成encoding，processing，decoding。

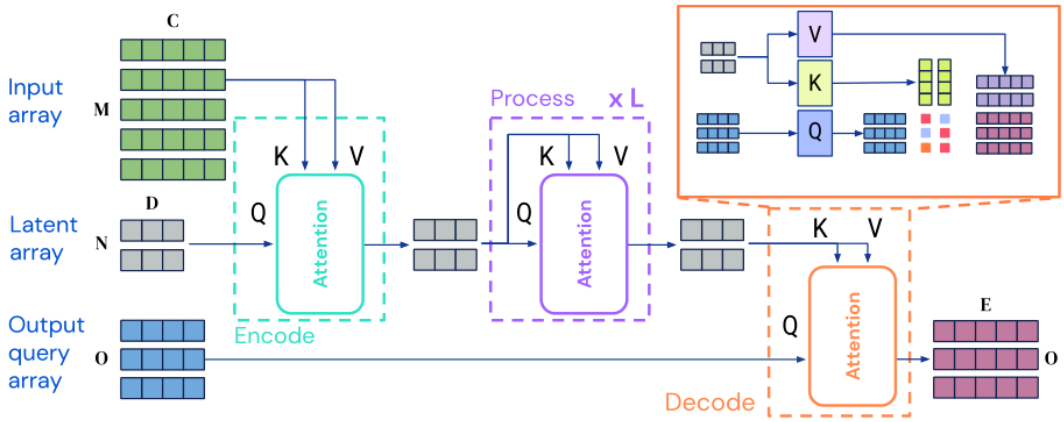


Figure 2: The Perceiver IO architecture. Perceiver IO maps arbitrary input arrays to arbitrary output arrays in a domain agnostic process. The bulk of the computation happens in a latent space whose size is typically smaller than the inputs and outputs, which makes the process computationally tractable even for very large inputs & outputs. See Fig. 5 for a more detailed look at encode, process, and decode attention.

其中Perceiver可以看作是encoder，输出一个低维度的latent space，再通过一个decoder变为输出要求的维度

其中M，C，O，E为任务要求的，无法改变，而N，D的大小为超参数，可以改变。

此模型的processing部分与输入输出都解耦，可以适用于不同大小的输入和输出，正如文章的名字《Perceiver IO：A general architecture for structured input & output》，有望成为多模态的通用网络架构。