Compression for AGI Summary

沙之洲 202012408

Compression for AGI 这个 talk 提出了一种对于大模型很新的视角，将大模型视作是压缩率很高的压缩器，并且提出衡量一个模型的指标应当是描述这个模型所需要的代码量，而非模型的参数量。

这篇工作最大的亮点在于通过一个 Satya 和 Sundar 的思想实验阐述了上述想法。具体来说，Satya 和 Sundar 之间需要传输一些数据，但是他们之间的带宽很小，两边却拥有很多的计算资源。他们有着相同的模型训练代码。对于传输过程，具体来说，Satya 将 x\_1 放到模型中编码成 z\_1 传输给 Sundar，Sundar 接受 z\_1 并且能够用初始模型解码出 x\_1。接下来，两边通过 z\_1 对模型做 SGD，这次的 SGD 会使得模型的压缩能力更强，第二次传输的 z\_2 的长度更小，经过多轮迭代，模型很好地拟合了世界的分布，同时具有了很强的压缩能力。这里将模型训练时候的 loss 的积分解释为模型将世界知识压缩之后的大小。

整体来看，这篇工作提出了对于模型的另一种量化方式，区别于上次课程以 FLOPS 对模型进行量化，这篇工作用描述模型所需要的代码量来对模型进行量化。这实际上是两种流派的差异，前者认为模型的智能是取决于它计算了多少次，后者相信模型的智能取决于人类给模型先验注入的知识，也就是描述模型需要的代码量，

但是，无论是哪一种对于模型的解读，现有的模型都仍然是一个对世界概率的拟合。而这种简单对于概率的拟合只能让模型了解到关于世界的更多经验，但是不能赋予模型自主推理的能力。人类之所以能够取得今天的成就，是因为人类不但能获取经验，还能够从中总结反思，从经验中提取知识，反过来训练自己。

如何进一步强化模型的学习能力是下一步研究的关键。现阶段通过大规模的语料输入已经让模型具有了大量的经验，下一步如果能够让模型从经验中不断学习，不断自我迭代，那样的智能或许会更接近人类。

课上讨论的一个点值得记录。COLT5 的工作使得最多能够输入 transformer 的 token 达到 64K，但是这个最高长度仍然不能够满足现实世界的绝大多数的数据长度。一种观点认为，可以将今天的输入 transformer 的 token 数量类比于十几年前的内存大小。十几年前的内存大小也只有 KB 级的大小，但是今天已经发展到 GB，TB 甚至 PB 级别的大小了。如果在未来，transformer 能够处理的数据量也能够达到 T 或者 P 级别，那么可以将世界全部事物以 token 的形式表现。在那个时代，模型就有能力建模，并且计算整个世界了。