大模型蒸馏篇

来自： AiGC面试宝典

宁静致远

2024年01月27日 19:14

一、知识蒸馏和无监督样本训练?

知识蒸馏是利用大模型把一个大模型的知识压缩到一个小模型上。具体来说你在一个训练集上得到了一个非常好的较大的模型，
然后你把这个模型冻结，作为Teacher模型也叫监督模型，然后你再造一个较小参数的模型叫做Student模型，我们的目标就是利用冻结的Teacher模型去训练Student模型。
A.离线蒸馏：Student在训练集上的loss和与Teacher模型的loss作为总的loss，一起优化。
B.半监督蒸馏：向Teacher模型输入一些input得到标签，然后把input和标签传给Student模型
还有个自监督蒸馏，直接不要Teacher模型，在最后几轮epoch，把前面训练好的模型作为Teacher进行监督。
目前知识蒸馏的一个常见应用就是对齐ChatGPT。
然后这个无监督样本训练，我看不懂意思。如果是传统的无监督学习，那就是聚类，主成分分析等操作。如果是指知识蒸馏的话，就是离线蒸馏的方式，只不过损失只有和Teacher的loss。

二、对知识蒸馏知道多少，有哪些改进用到了?

知识蒸馏是一种通过将一个复杂模型的知识转移到一个简单模型来提高简单模型性能的方法。这种方法已经被广泛应用于各种深度学习任务中。其中一些改进包括：

- 使用不同类型的损失函数和温度参数来获得更好的知识蒸馏效果。
- 引入额外的信息来提高蒸馏的效果，例如将相似性约束添加到模型训练中。
- 将蒸馏方法与其他技术结合使用，例如使用多任务学习和迁移学习来进一步改进知识蒸馏的效果。

三、谈一下对模型量化的了解?

模型量化是一种将浮点型参数转换为定点型参数的技术，以减少模型的存储和计算复杂度。常见的模型量化方法包括：

- 量化权重和激活值，将它们转换为整数或小数。
- 使用更小的数据类型，例如8位整数、16位浮点数等。
- 使用压缩算法，例如Huffman编码、可逆压缩算法等。

模型量化可以减少模型的存储空间和内存占用，同时也可以加速模型的推理速度。但是，模型量化可能会对模型的精度造成一定的影响，因此需要仔细权衡精度和计算效率之间的平衡。

四、模型压缩和加速的方法有哪些?

参数剪枝（Parameter Pruning）：删除模型中冗余的参数，减少模型的大小。通常情况下，只有很少一部分参数对模型的性能贡献较大，其余参数对性能的贡献较小或没有贡献，因此可以删除这些冗余参数。
量化（Quantization）：将浮点型参数转换为更小的整数或定点数，从而减小模型大小和内存占用，提高计算效率。
知识蒸馏（Knowledge Distillation）：利用一个较大、较准确的模型的预测结果来指导一个较小、较简单的模型学习。这种方法可以减小模型的复杂度，提高模型的泛化能力和推理速度。
网络剪枝（Network Pruning）：删除模型中冗余的神经元，从而减小模型的大小。与参数剪枝不同，网络剪枝可以删除神经元而不会删除对应的参数。
蒸馏对抗网络（Distillation Adversarial Networks）：在知识蒸馏的基础上，通过对抗训练来提高模型的鲁棒性和抗干扰能力。
模型量化（Model Quantization）：将模型的权重和激活函数的精度从32位浮点数减少到更小的位数，从而减小模型的大小和计算开销。
层次化剪枝（Layer-wise Pruning）：对模型的不同层进行不同程度的剪枝，以实现更高效的模型压缩和加速。
低秩分解（Low-Rank Decomposition）：通过将一个较大的权重矩阵分解为几个较小的权重矩阵，从而减少计算开销。
卷积分解（Convolution Decomposition）：将卷积层分解成几个更小的卷积层或全连接层，以减小计算开销。
网络剪裁（Network Trimming）：通过对模型中一些不重要的连接进行剪裁，从而减小计算开销。

五、你了解的知识蒸馏模型有哪些?

FitNets：使用一个大型模型作为教师模型来指导一个小型模型的训练。
Hinton蒸馏：使用一个大型模型的输出作为标签来指导一个小型模型的训练。
Born-Again Network（BAN）：使用一个已经训练好的模型来初始化一个新模型，然后使用少量的数据重新训练模型。
TinyBERT：使用一个大型BERT模型作为教师模型来指导一个小型BERT模型的训练。