AIGC算法工程师面经——模型训练通识基础篇

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年04月11日 18:04 广东

面试经验专栏

作者：vivida

瓦力算法学研所
我们是一个致力于分享人工智能、机器学习和数据科学方面理论与应用知识的公众号。我...
117篇原创内容

公众号

本篇总结了AIGC面经中可能会问到的模型训练通识类题目及其答案。

本篇开始重点介绍面经中可能会问到的模型训练通识类题目及其答案。

但是需要特别注意的是，此类宽泛的问题类似于命题作文，看似简单且答案明确，但实际考量的空间非常大；单纯地背完八股面试官往往是不满意的，一般的反应是再问更细节的内容或者直接反馈觉得你还说的不够。

这种时候最好要结合一些自身的实践经验，或者将题目与答案说的更深一些。

本篇在比较重要的问题下写答案时也会尽量避免过于宽泛和官方的用词，并结合一些实际经验；希望大家在自己复习准备时也尽量思考得更深入。

下面是一个问题的快捷目录。

面试题

7. Warm up一般是在什么情况下使用的

8. 模型压缩有哪些方法，介绍一下

9. 模型陷入局部极小了怎么办

10. 当资源很少时怎么做数据增强

11. Adam如何设置参数使学习率衰减

12. 为什么出现梯度爆炸，梯度爆炸怎么解决

13. 神经网络权重全 0 初始化会有什么问题？应该怎样初始化？讲讲 Xavier 初始化

14. 现在有哪些归一化方法

15. 学会了哪些网络训练调参技巧

答案

1. 请具体介绍一下L1、L2正则化

正则化主要目的是控制模型复杂度，减小过拟合。正则化方法是在原目标（代价）函数 中添加惩罚项，对复杂度高的模型进行“惩罚”。

L1：向量绝对值和，趋向于产生少量的特征，而其它的特征都为0，有助于处理高维数据集, 使权重稀疏。

$$L = \sum_{i=1}^{n} |y_i - f(x_i)|$$

L2：向量平方和，会选择更多的特征，但这些特征都接近于0，使权重平滑。

$$L = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

2. 过拟合怎么解决?

减少参数、early-stop、正则化、drop-out

3. Dropout 有什么作用?

整个dropout过程就相当于对很多个不同的神经网络取平均。而不同的网络产生不同的过拟合，一些互为“反向”的拟合相互抵消就可以达到整体上减少过拟合。

实际用的时候，训练的时候会随机的丢弃一些神经元，预测的时候不随机丢弃。

4. 常见的激活函数及其优缺点

- Sigmoid: 它可以将一个实数映射到(0,1)的区间，但不以0为中心，收敛慢且易梯度消失。
- Tanh：缩至-1 到 1 的区间内，其收敛速度要比sigmoid快，但也会梯度消失
- Relu：当x<0时，ReLU硬饱和，能够在x>0时保持梯度不衰减，从而缓解梯度消失问题，但也有缺点<0时，以及不以0为对称中心。

5. 数据不平衡问题如何解决?

- 欠采样过采样
- Loss加权
- 一些数据蒸馏的方法（例如用一些BT、self-train以及更大的模型生成的FT数据）

6. Warm up一般是在什么情况下使用的

- 直接设置初始学习率为0.01或0.001，对大多数网络都适用。
- 使用Smith的方法，首先设置一个非常小的学习率，比如1e-5，每个batch后更新网络，同时增加学习率，统计每个batch计算的loss。最后描绘出学习率的变化曲线和loss曲线，确定最优学习率。
- StepLR：每过step_size轮，将此前的学习率乘以gamma。
- MultiStepLR：在每个milestone时，将此前学习率乘以gamma。
- ExponentialLR：每一轮会将学习率乘以gamma，所以这里千万注意gamma不要设置的太小，不然几轮之后学习率就会降到0。

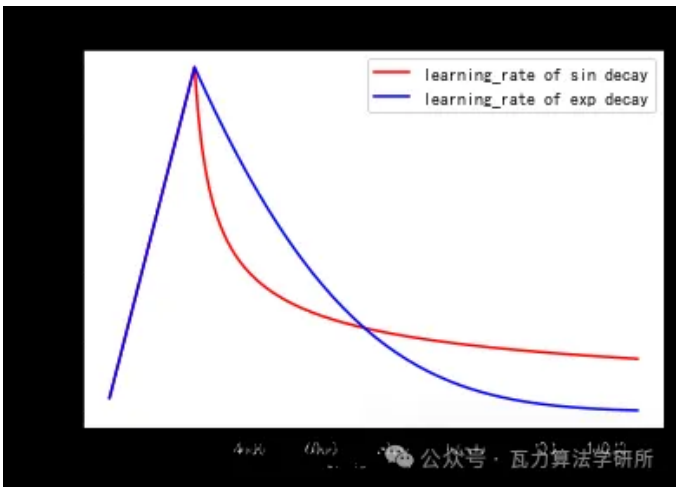
7. Warm up一般是在什么情况下使用的

首先需要介绍一下优化器，优化器的作用是在模型训练过程中来更新模型参数，最小化（或最大化）损失函数，以提升模型效果。
优化器主要依据两个条件确定，**一个是学习率另一个是梯度**；一些好的优化器本身的设计就可以做到动态的调整学习率和梯度。

那warm up是什么情况下使用的呢?

就学习率来说，优化器本身是根据梯度来调整学习率的，一般刚开始训练时梯度很大（误差大）所以学习率也较大，这样的设计符合让模型尽快收敛的需求；
但是在有些情况下，尤其是使用了预训练模型进行下游任务时，学习率太大会带来不稳定问题，使模型发生振荡，所以需要让刚开始训练时有一个较小的学习率，确保模型能够有良好的收敛性，因此就有了**学习率预热**和**学习率衰减**这样的策略来辅助调整学习率。

warmup就是一种学习率预热策略，就是使学习率从0开始增加，增加到warmup设定值时再逐渐减小，当然增加和减小的过程可以是线性的也可以是非线性的。



8. 模型压缩有哪些方法，介绍一下

蒸馏、量化、剪枝等，这个时候可以把方向往自己更了解的知识引，推荐大家看一下之前写的一篇大模型量化策略 大模型目前量化方法有哪些？详细介绍实际落地中最常用方法。

9. 模型陷入局部极小了怎么办?

优化器选择，短时间增大学习率等方法。

10. 当资源很少时怎么做数据增强？（这道遇到过很多次，要从数据和模型角度说）

- 数据爬取（勉强算一个）
- 可以基于fasttext快速分类方法从已有的数据量中检索需要数据
- 基于embeddding检索相似向量
- 模型加kd_loss

11. Adam如何设置参数使学习率衰减

Adam 优化器，全称 Adaptive Moment Estimation，通过计算每个参数的移动平均值和变化率，从而自适应地调整学习率，效率和稳定性较高。
在 Adam 优化器中，学习率衰减策略的具体操作如下：
- 基于时间衰减：根据训练轮数或时间步，逐渐减小学习率。
- 学习率衰减调整：根据模型的性能或其他信号，调整学习率。

12. 梯度爆炸怎么解决?

反向传播中链式法则带来的连乘，如果有数很小趋于 0，结果就会特别小（梯度消失）；如果数都比较大，可能结果会很大（梯度爆炸）会造成权值更新缓慢，模型训练难度增加。
1） pretraining+finetuning : 寻找局部最优，然后整合起来寻找全局最优
2） 梯度裁剪
3） 权重正则化
4） 选择relu等梯度落在常数上的激活函数
5） 残差
6） LSTM

13. 神经网络权重全 0 初始化会有什么问题？应该怎样初始化？讲讲 Xavier

在神经网络的训练中如果将权重全部初始化为0，则第一遍前向传播过程中，所有隐藏层神经元的激活函数值都相同，导致深层神经元可有可无（对称权重）。

常见的初始化方法包括高斯分布初始化、均匀分布初始化、Xavier初始化。

xavier初始化只适用于关于0对称、呈线性的激活函数，比如 sigmoid、tanh、softsign

无论采用何种激活函数，xavier初始化都会根据权重值的分布，给出两个模式：

1） 希望初始化的权重值**均匀部分**，此时要给出权重初始化时的**取值上下限**
2.）希望初始化的权重是**高斯分布**，此时要给出权重初始化时的**标准差（均值为0）**

对于ReLU激活函数，可以采用 Kaiming 初始化，Xavier初始化在Relu层表现不好，主要原因是relu层会将负数映射到0，影响整体方差。

14. 现在有哪些归一化方法

大模型面经——大模型中用到的归一化方法总结

15. 学会了哪些网络训练调参技巧

- 学习率基本都使用warm-up策略
- loss中出现NaN怎么办：大模型工程化必备技巧——模型训练过程中发现输出大量NaN怎么办？建议收藏
- 训练数据配比：通用数据和领域数据最好1：1
- 训练模型损失：灵活调整kl_loss与kd_loss

欢迎大家关注公众号，更多面经干货将持续放送~