????

1. 梯度下降的每个 step，都需要把所有的 loss“汇总 ”
   1. 包含所有 batch，所有的 loss function，以及在多个 step之间，多个 epoch之间
   2. 等价于对数据集的统计、回归、拟合，最终效果只能达到统计上的意义
   3. 训练效率低，有用信息密度低、容易被坏数据平均
2. 相对于人脑
   1. 缺乏思维的贪心原则
   2. 缺少 “自恰 ”的自主意识的判断
3. 一个样本的梯度下降，相对于整个模型，所表达的是对参数分布的一个小的倾向性调整
   1. 不能确保模型肯定掌握、认同这个样本
   2. 有矛盾的样本会按照概率进行中和
   3. 贪心法、高温度、反复强化自身的预测置信度，都能使模型对输出更自信
   4. 不同的训练阶段，样本的作用不一样，预训练表示大范围的统计，微调 /SFT
      就需要明确的服从样本
4. 它高度非凸。
   而且从数学上完全没有任何保证能确保优化成功。
   然而，事实证明它确实能成功 ——这是经验证明的结果。
   纯粹靠理论，我们很难解释太多细节，我们还没完全搞清楚原理，这些看似简单的启发式优化算法在这些问题上表现得如此出色。
   然而，人类的智慧并不追求绝对最优，就像我们设计飞机或汽车时也不会力求完美。
   我们只需要一个「够好」的工程系统就行。
   深度学习和非凸优化给我们的，正是一群「够好」的系统。虽然它们可能不是最优解，却依然非常有用、充满潜力。
   这就是事实
   也许正是这种因为深度学习是务实的，在如今算力爆炸和数据丰富的时代， LLM
   虽然还是「黑箱」，但已经切实的改变了我们的工作和生活。
5. 神经网络里有大量神经元和连接，每层都会先将输入乘以随机权重，再经过非线性变换。
   第一层处理完后，第二层又会重复相同的过程：乘权重、做非线性变换。
   如果这些随机权重太小，信号在多次相乘后就会迅速衰减到几乎为零。
   当信号到达输出层时，你几乎感受不到任何输入的影响。
   这样一来，学习算法就无法发现输入和输出之间的关联，也就没法改进模型。
   因此，我们必须让随机初始化的权重大多数情况下足够大，才能保证输入的变化一路传递到输出层。
   一旦满足了这个条件，梯度就能够找到正确的方向，有效地优化网络。
   因此在实际应用中，当研究人员希望在一个真实数据集上训练神经网络时，初始化的尺度是你需要关注的最重要的参数之一。

Revision #7
Created 2025-02-13 08:01:31 UTC by Colin
Updated 2026-04-29 07:34:14 UTC by Colin