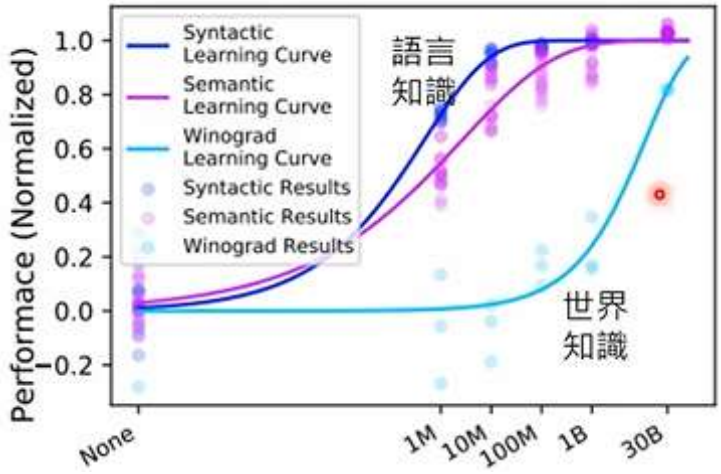
6.大模型

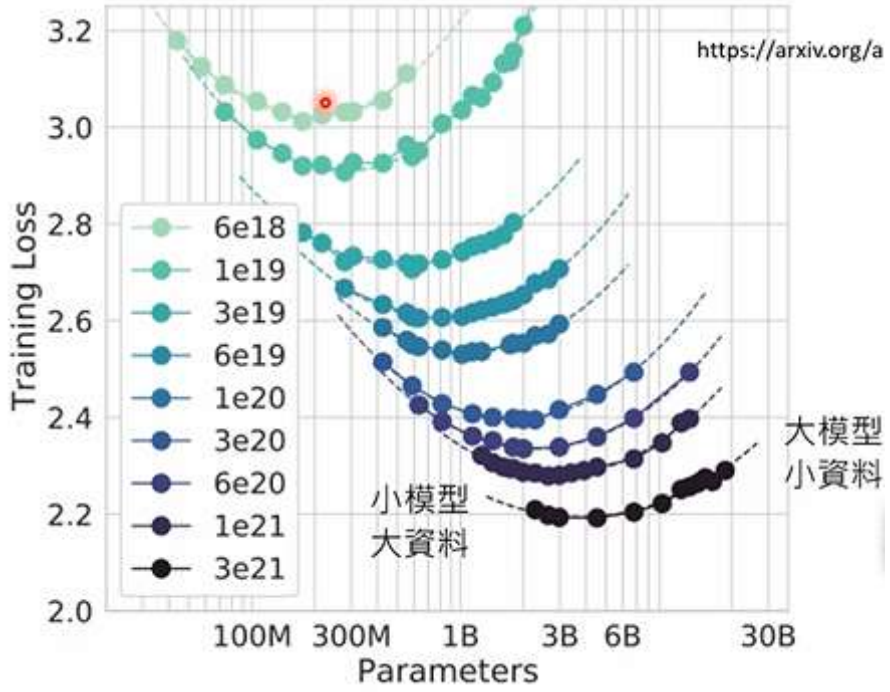
笔记本： 【课】原理-李宏毅 deep leaning
创建时间： 2023/4/24 15:50 **更新时间：** 2023/4/24 21:21
作者： 1256876216@qq.com
URL： https://www.baidu.com/s?ie=utf-8&f=8&rsv_bp=1&tn=baidu&wd=KNN%...

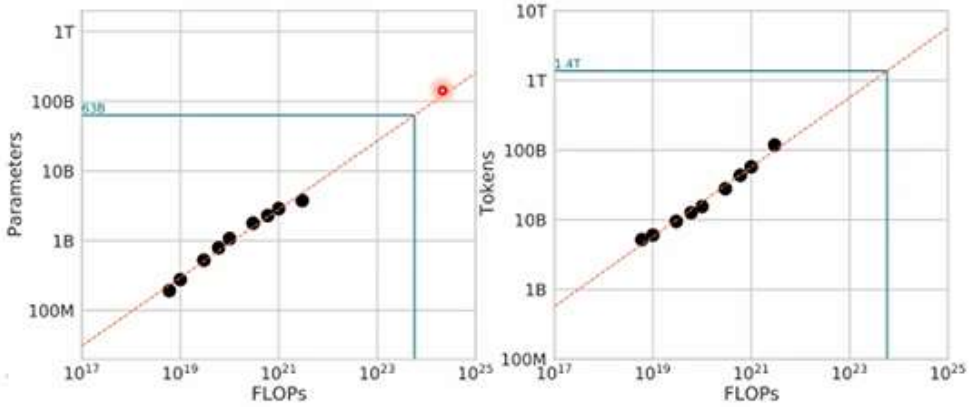
- 当数据大到一定程度时，大模型会顿悟，突然间起作用
  分析模型效率时，不止于用结果来衡量，还有过程
- 什么时候需要百万级别单词数的数据量



1. 数据处理：
   - 过滤有害内容
   - 去除Html tag（保留项目符号等）
   - 用规则去除低品质资料
   - 去除重复资料
   - 为实验的严谨（保留测试资料）

2. 固定运算资源下模型大小与资料大小的平衡





Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion

3. KNN LM

K-Nearest Neighbor K临近算法，根据距离学习，在处理临近节点时耗时