读书报告

Beyond Human-Level Accuracy: Computational Challenges in Deep Learning

Proceedings of the 24th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPoPP 2019, Washington, DC, USA, February 16-20, 2019

Authors: Joel Hestness , Newsha Ardalani, Gregory Diamos

姓名：黄玟瑜

学号：19335074

班级：19 级计算机科学与技术（超算）

日期：2021 年 6 月 25 日

一、研究背景

深度学习 (DL) 已成为近期人工智能 (AI) 突破的主要驱动力。随着支持 DL 的产品的数量越来越多，满足未来 DL 模型训练的硬件需求变得越来越重要。

深度学习研究通过模型架构变化和规模产生准确性和产品改进：更大的数据集和模型，以及更多的计算。对于硬件设计，很难预测 DL 模型的变化。但是，最近的先前工作表明，随着数据集大小的增长，DL 模型的准确性和模型大小的增长是可预测的。DL 训练模型的准确性的提高主要来自两个方面的改进：一是模型的自身架构优化，二是构建模型的数据集、模型的大小和计算的提升。

深度学习领域的学者普遍认为模型准确性会随着训练数据集大小的增长而提高。此外，Hestness 等人提出 DL 模型的精度的提升符合数据集和模型的大小扩充的特定幂律函数。在单词和字符语言建模、机器翻译、语音识别和图像分类这五个深度学习领域，将需要大量增加数据集和模型大小才能实现目标精度，要实现它们的准确度目标，数据集的大小需要比用于训练当前最先进 (SOTA) 模型的数据集大 33－971 倍，模型的参数数量也必须增加 6.6-456 倍。基于这些期望的目标，简单的估计表明，在当前系统上的训练时间将需要几十到几个世纪。

二、研究目的

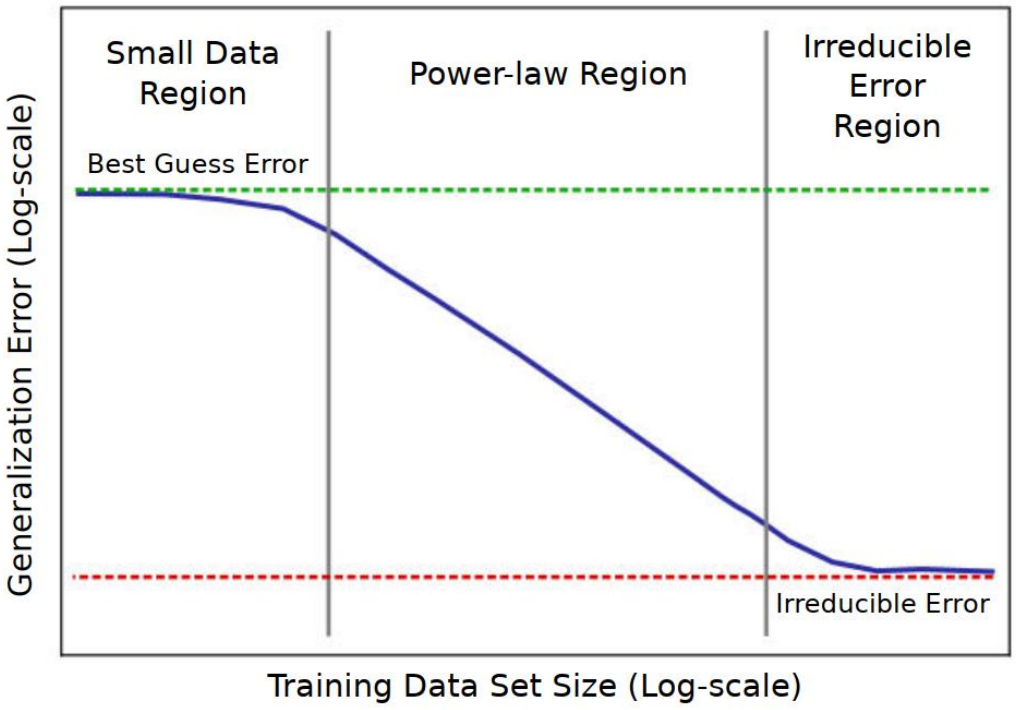
DL 模型的精度随着数据集和模型大小的扩充而提高，为了提高 DL 模型的精确度，需要对达到目标精度所需的数据集和模型大小做一个预测，并对硬件需求的水平做出估计。

基于先前的深度学习领域有关模型精度的研究，预测将 DL 模型精度提高到超越人类当前水平所需的数据集和模型的规模大小，对满足未来 DL 模型训练的硬件需求做出估计，以达到机器学习专家定义的前沿目标。

三、解决方案

1. 数据集大小的预测

Hestness 等人表明在真实数据集上，DL 模型的准确性随着训练数据集的大小而提高。他们进一步表明，拟合数据所需的模型大小随着数据大小的增长可预测地增长，行业可以使用这些经验模型来估计达到特定精度所需的训练数据量和模型大小。



模型学习幂律图

随着数据集规模的增长，DL 模型的预测误差会减小。当数据集较小时（对应 Small Data Region），在该区域模型只能对输出数据的分布进行“最优”的猜

测。在幂律区域（Power-law Region）中，数据集中每个新加入的训练样本提供有效信息对模型进行训练，以提高模型的预测精度。误差下降是可预见的。最后，在实际应用中，曲线可能会终止于不可约区域（Irreducible Region），由于数据的随机性，模型无法在该区域中被进一步优化。

由于目前大多数现有的大规模数据应用程序都在幂律区域，我们使用幂律区域的曲线来预测数据集大小。在幂律区域，模型泛化误差的大小大致遵循以下幂律：

$$\varepsilon(m) \approx \alpha m^{\beta_\partial}$$

其中，m 是训练数据集中的样本数，α 和 $\beta_\partial \in [-0.5, 0]$ 是常数，取决于建模的结构，包括数据分布和模型架构等方面。α 表示输入数据空间和 DL 模型架构的各个方面。 β_∂ 是幂律指数，表示模型从每个额外的训练示例中学习更多信息的难度。 β_∂ 接近 -0.5 意味着模型可以从较小的数据集中快速学习。表 1 列出了在以前的研究工作中发现的不同建模任务的 α 和 β_∂ 估计值。

Domain (model)	Current SOTA	Desired SOTA	Current Data Size		Learn Curve		Model Size		Projected Scale	
			Samples	GB	α	β_g	σ	β_p	Data	Model
Word LMs (LSTM)	3.37 nat/word	2.48 [31]	768*M* word	3.9	13.0	−0.066	9.4e−4	0.68	100×	23×
Character LMs (RHN)	1.30 bit/char	0.70 [31]	3.48*B* char.	3.9	9.39	−0.092	1.2e−5	0.89	**971×**	**456×**
NMT (enc/dec+attn)	28% WPER	12%	130*M* WP	2.6	3.06	−0.128	6.4e−4	0.68	**750×**	90×
Speech Recogn. (enc/dec+attn)	9.5% CER	4% [39]	425*M* char.	1674	30.5	−0.291	2.4e−3	0.54	33×	6.6×
Image Classification (ResNet)	19.4% Top-1	5% [29]	1.3*M* image	152	15.0	−0.309	2.0e−2	0.57	81×	12×

表 1

基于 Hestness 等人的研究工作进行进一步拓展。为了从这些模型中预测所需的数据和模型大小，定义支持 DL 的产品所需的准确度目标。从 DL 领域的专家那里收集反馈，并参考先前的研究，基于这些研究估计不可减少的误差，从而为每个领域选择理想的准确度目标。例如，单词和字符 LM 所需的 SOTA 接近英文文本熵的估计下限。表 1 的“Desired SOTA”列反映了这些预测。最后，给定这些分析学习曲线和目标错误率，构建实现目标所需数据大小的分析

模型。表 1 中的“Projected Scale”一列列出了相对数据规模的预测。 Desired SOTA 值比当前的 SOTA 值好 1.4 到 3.9 倍。然而，实现这些值所需的数据量从语音识别的 33 倍到字符 LM 的 971 倍不等。语言领域需要最多的数据，因为它们的幂律指数 $\beta\partial$ 较差。

2. 模型大小的预测

随着数据集大小的增长，模型的大小也必须增长以表示数据。Hestness 等人还收集和表征了适合不同训练集大小所需的模型大小。模型参数（大致容量）预计在训练集大小中呈亚线性增长，形式如下：

$$p(m) \approx \sigma m^{\beta_p}$$

其中，m 是训练集中的样本数，σ，$\beta_p \in [0.5, 1)$ 取决于建模的结构，包括数据分布和模型架构等方面。模型应该比训练集增加参数计数更慢（即 $\beta_p \leqslant 1$），或者我们可以只存储数据集而不是训练模型。最近的工作表明，深度神经网络模型容量——它可以学习的概念（数据）量——以 O(lp logp) 增长，其中 1 是模型深度的度。稍微放宽这个界限，模型大小应该至少以数据集大小的平方根增长（即 $\beta_p \geqslant 0.5$）。表 1 显示了根据经验收集的 DL 域的 σ 和 β_p。给定在上一小节中确定的目标数据大小，我们预测适合目标数据集大小所需的模型大小。模型比例列显示模型大小的相对要求增加。例如，当前的 SOTA 单词 LM 使用大约 1B 个参数来拟合大约 1B 个单词数据集。因此，要拟合 100 倍大的数据集，模型需要约 23B 参数（23-92GB，取决于权重精度）

3. 表征计算需求

下面将注意力转向表征计算需求以训练这些非常大的模型。下面将分析 DL

应用程序的 FLOP、内存访问和内存占用增长。 尽管 DL 应用程序的结构错综复杂，但其训练需求的规模大多可预测。 计算和内存使用随着模型大小和批量大小逐渐线性增长。 作者提供了这些可访问的一阶计算需求模型，这些模型在以前的研究工作中没有被表征过。

通过分析训练运行的数据和应用模型来估计模型训练需求。使用在 NVIDIA GPU 上运行的 Tensorflow 1.5.0 并使用 TFprof 的修改版本进行训练。TFprof 注释计算图操作以计算它们的算法 FLOP 和字节，并在它们执行时收集运行时间。在训练步骤结束时（即计算图遍历完毕），TFprof 为该步骤期间执行的所有操作返回此配置文件，确保我们对端到端训练步骤的精细细节进行分析。查询 Tensorflow 的内存分配器以获得分配的最大训练步骤内存量——内存占用。从 100-500 个随机选择的训练步骤中收集配置文件，以解释每个训练步骤在不同模型访问的 FLOP 和内存方面的差异。例如，字符 LM、NMT 和语音模型针对最长批次样本所需的时间步长展开其循环层。这种展开导致在单独的训练步骤中进行变量计算和内存访问，因此需要对训练步骤的分析结果进行平均。要控制的最复杂的变量是训练批次大小——在单个训练步骤中要观察的数据并行样本的数量。批量大小可以任意设置，但特定的批量大小会导致最佳模型精度，具体取决于数据集大小。对于本研究中的测试域，SOTA 模型已使用跨 GPU 的数据并行性进行训练，以增加批量大小，使其超出单个 GPU 的最大内存容量。未来的 DL 训练很可能也会受到每计算单元内存容量的限制，这表明 ML 研究人员将选择可以提供接近峰值的计算单元资源利用率的每计算单元批量大小（以下称为“子批量大小”）。我们使用最小的此类子批次大小进行分析。为了增长模型，我们改变了对模型拟合更大数据集的能力影响最大的超参数，如泛化误差所衡量的那样。对于 ResNets，增加深度和卷积通道，

而不是过滤器大小，可以最大程度地提高准确性，因此我们收集了更深更广的图像分类网络的配置文件。大多数循环模型已经增长到一个深度，以至于增加的深度不会导致精度提高。相反，我们增加了每层隐藏权重的数量。最后，我们的目标是在我们扩大数据集和模型大小时预测模型的计算需求。

四、实验验证

作者通过在 NVIDIA GPU 上运行 Tensorflow 1.5.0 并使用 TFprof 的修改版本进行训练，从 100-500 个随机选择的训练步骤中收集配置文件，以解释每个训练步骤在不同模型访问的 FLOP 和内存方面的差异。

五、作者的核心思想及创新点

核心思想

基于前人（主要是 Hestness 等人）的研究，通过他们提供的 DL 模型的精度和数据集、模型大小的分析，预测了在词和字符语言建模、机器翻译、语音识别和图像分类等深度学习领域中达到目标精度所需要的数据集和模型大小。作者列出了 FLOP、内存访问和占用等维度来表征计算需求，提出了它们之间的一阶模型，并通过在 Tensorflow 上运行模型训练来解释每个训练步骤在不同模型访问的 FLOP 和内存方面的差异。

创新点

在预测数据集和模型大小时，在 Hestness 等人的分析基础上作者加入了一些新的参数来表征精度和数据集和模型大小的幂律关系。

提出分析计算需求的一阶模型。

六、启发

在自己有限的能力水平下谈谈收获。

从作者基于 Hestness 等人的分析预测数据集和模型大小的过程中，想到预测类型的分析可以在前任研究工作的基础上加入一些不同维度考虑因素，比如相关领域有权威的专家反馈。

设计实验验证时可以合理地提出一些量化维度（比如模型训练每一步的FLOP、内存占用情况等），并注意在实验中收集这些数据并分析。