Google 重新审视深度学习时代数据的非理性效果

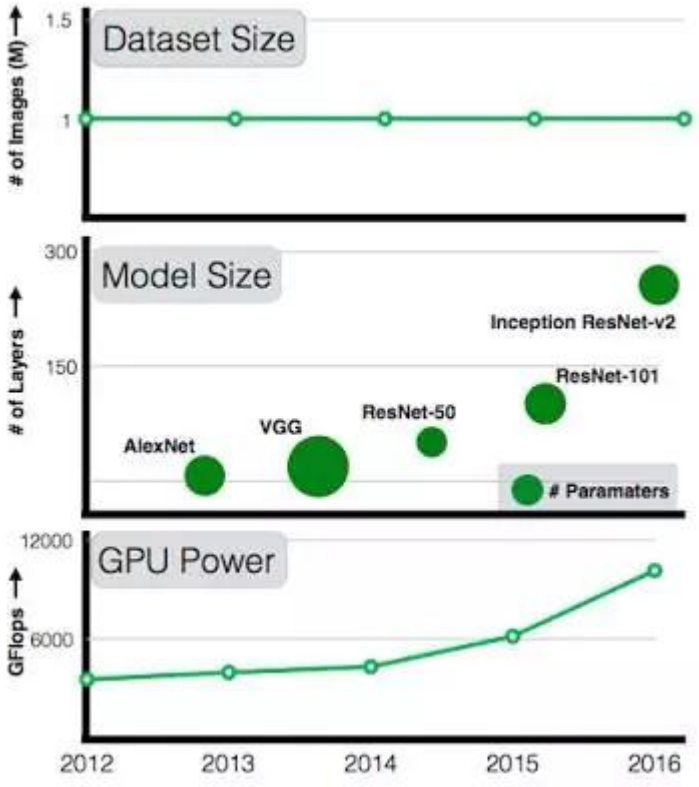
原创：DevRel谷歌开发者2017-07-24

文 / 机器感知指导教师 Abhinav Gupta

过去十年里，计算机视觉领域取得了巨大成功，这在很大程度上得直接归功于深度学习模型在机器感知任务中的应用。

此外，自 2012 年以来，这些系统的表征能力取得了长足的进步，这归因于：
(a) 极为复杂的更深度模型的建立；
(b) 计算能力不断提升；
(c) 可获得大规模的标注数据。

尽管计算能力和模型复杂度每年都在不断提升（已从 7 层的 AlexNet 提高到 101 层的 ResNet），但可用数据集并未得到相应的扩充。与 AlexNet 相比，101 层的 ResNet 的容量要大得多，但它仍在使用同样从 ImageNet circa 2011 获取的 100 万张图像进行训练。作为研究人员，我们一直想知道：如果将训练数据量扩大 10 倍，准确率是否会翻倍？扩大 100 倍甚或 300 倍，准确率又会如何？准确率是否会遭遇平台期？还是说数据越多，准确率就越高？



▲ 过去五年里，GPU 的计算能力和模型大小在不断提高，但令人吃惊的是，最大的培训数据集的规模却停滞不前。

在我们的《重新审视深度学习时代数据的非理性效果》(Revisiting Unreasonable Effectiveness of Data in Deep Learning Era) 这篇论文中，我们在揭开围绕“海量数据”和深度学习之间关系的谜团方面迈出了第一步。我们的目标是探究以下问题：

(a) 向现有算法提供更多带有噪声标签的图像是否仍可以改善视觉表征；

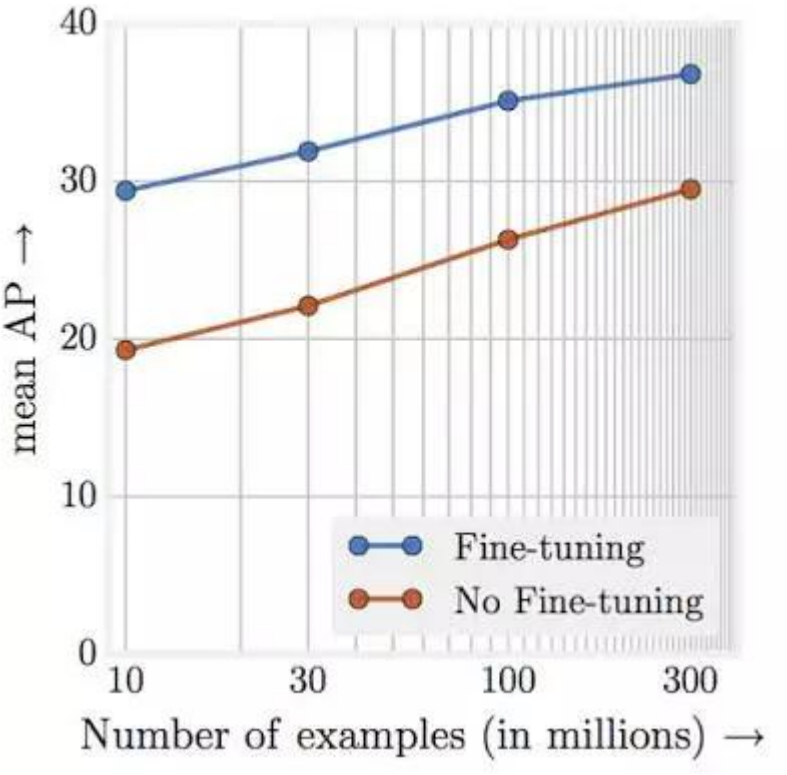
(b) 分类、对象检测和图像分割等标准视觉任务中，数据与性能之间的本质关系；

(c) 通过大规模学习找到适用于计算机视觉领域所有任务的最先进模型。

当然，一个无法回避的问题是我们从何处获取一个比 ImageNet 大 300 倍的数据集？在 Google，我们一直致力于自动构建此类数据集以改善计算机视觉算法。具体而言，我们已构建一个包含 3 亿张图像的内部数据集（我们称之为 JFT-300M），这些图像被标记为 18291 个类别。用于标记这些图像的算法使用了复杂的数据组合，包括原始网络信号、网页与用户反馈之间的联系等。这为 3 亿张图像生成了 10 亿多个标签（一张图像可具有多个标签）。为最大程度提高所选图像的标签精度，我们通过某个算法从 10 亿个图像标签中选取了大约 3.75 亿个标签。然而，这些标签中仍然存在大量噪声：所选图像的标签中约有 20% 带有噪声。由于缺乏详尽的注解，我们无法评估标签的回想率。

我们的实验结果证实了部分假设，但也产生了一些意外的惊喜：

- **更好的表征学习确实大有裨益。** 我们的第一个观察结果是大规模数据有助于表征学习，进而改善了我们研究的每个视觉任务的性能表现。我们的研究发现表明：共同构建一个大规模数据集进行预训练非常重要。同时，实验也表明，无监督和半监督表征学习方法的前景非常光明。数据规模似乎可克服标签方面的噪声问题。
- **表现与训练数据的数量级呈线性递增关系。** 也许整个实验最惊人的发现就是视觉任务的表现和用于表征学习的训练数据量（对数）之间的关系了。我们发现它们之间的关系竟然是线性的！即使训练图像达到 3 亿张，我们也并未观察到对所研究的任务产生任何平台效应。



▲ 通过针对 JFT-300M 的不同子集从零开始进行预训练时的对象检测性能。X 轴是以对数表示的数据集大小，y 轴代表针对 COCO-minival 子集的 mAP@[.5,.95] 检测性能。

- **容量至关重要。**我们同样观察到：为了充分利用 3 亿张图像，我们需要更高的容量（更深的）模型。例如，就 ResNet-50 而言，其在 COCO 对象检测基准测试中的增益 (1.87%) 大大低于使用 ResNet-152 时的增益 (3%)。
- **新的最佳结果。**我们的论文展示了通过使用从 JFT-300M 学到的模型在多个基准中取得了新的最佳结果。例如，单一模型（没有任何不必要的花哨功能）在 COCO 检测基准测试中从原来的 34.3 AP 提高到现在的 37.4 AP。

请注意，我们使用的训练机制、学习安排和参数都是基于我们使用来自 ImageNet 的 100 万张图像对 ConvNets 进行训练后所获得的认识。由于我们在此项工作中并未搜索最优超参数集（这需要极为庞大的计算量），所以在

使用这种规模的数据时，这些结果很可能并不是您能够取得的最佳结果。因此，我们认为报告的量化表现低估了数据的实际影响。

这项工作并不会关注特定任务的数据，例如探究更多的边界框是否会影响模型表现等。我们认为，虽然获取大规模特定于任务的数据非常困难，但它应该成为未来研究的重点。此外，构建包含 3 亿张图像的数据集不应该是我们的终极目标，作为一个社区，我们要探索的是，在采用更大规模的数据集（拥有 10 亿张以上的图像）时，是否可以继续改善模型。