Emergent Ability of LLM Summary

沙之洲 2020012408

Emergent Ability of LLM 这篇文章阐述了，一些更大的模型会出现一些在小模型上不会出现的 emergent ability。这里用来界定模型大小的指标是计算量 FLOPS，当 FLOPS 超过某一个阈值之后，模型在一些特定任务上的表现会有一个质的飞跃。而这种飞跃出现的模型规模临界点，也会被模型本身的结构，训练的方法，以及数据集的质量所影响。

这篇文章的优点在于，他调研了以往的大部分语言模型，对比了他们的大小，以及他们在多种 NLP 任务上 emergent ability 的体现。能够发现这种 emergent ability 并且做足够的调研，来佐证自己的猜测，这便是这篇文章的长处。同时，他向人们证明了 scaling up 是一种合理的增强模型性能的手段。

这篇文章的不足在于，他仅仅阐述了 emergent ability 出现这个现象，并没有对这个现象背后的原因进行分析。文章中提到了，模型结构、训练方法和数据质量都会对出现 emergent ability 的临界点产生影响，但是并没有通过对比实验说明这些因素的影响做进一步的定量研究。同时，对于一些任务，目前的 scaling up 并没有出现性能的显著提升，那么未来继续 scaling up 是否会带来性能的提升，文章也缺少对于这一部分的分析。

对于文章后续可以做的方向，首先是，通过 FLOPS 作为衡量模型大小的指标，这个事情没有那么显然，所以可以考虑将衡量模型性能的指标换成模型的参数量，或者其他指标，来探究在不同衡量模型大小的指标下，是否都会出现这种 emergent ability。文中也提到，随着模型结构、数据集、训练方法的优化，出现 emergent ability 的临界点可以变的更小。直观上来说这个临界点是存在一个下界的，如何确定这样的下界？文章中指出，一些任务可以通过 scaling up 出现 emergent ability，但是一些任务还没有出现这样的 emergent ability，那么对于那些任务我们可以期待随着 scaling up ，能够出现性能上的显著提升？

对于课上讨论重点的总结。一种对于 emergent ability 的解释是，一些任务缺少中间步骤的评价，只是通过最后结果的正确与否，对模型的指标进行评价。事实上有可能，模型在训练中能力不断增强，在任务的中间步骤上做的越来越好，当模型在中间步骤上的能力超过一定的阈值之后，导致任务的最终结果正确，反映在评价指标上是性能明显提升。这种解释认为任务评价模式是造成 emergent ability 出现的原因。实际上，任务中间步骤评价的缺失是现在任务评价的一大难题，同时中间步骤评价的数据标注起来非常昂贵。是否有可能通过以下的方式，缓解中间步骤评价缺失的问题。考虑通过某种指标对于一个任务数据集中的问题进行难度排序（类似 leetcode 给每个题目打难度分的方式），先让模型在简单的题目上 fit，再逐步提升题目的难度。这种类似打怪升级的方式，和人类的学习模式非常像。是否有可能通过这种方式让 LLM 逐步学会复杂推理能力？

最后，对于 FLOPS 作为模型评价指标的解释值得一提。现代有观点认为，神经网络可以被抽象为，以能源作为输入，以智能作为输出的黑盒。从这个角度来看，FLOPS 是衡量输入能源数量的一个最为合理的指标。但是也有一种观点认为，参数量多但是计算稀疏的模型，和参数量少但是计算密集的模型可能有差不多的 FLOPS，但是前者有更多的参数量，所以有更多的计算空间，所以这种视角下来看用 FLOPS 作为衡量模型大小的指标似乎并不是那么合理。