NYU & Google: 知识蒸馏无处不在，但它真的有用吗？

原创 小伟 夕小瑶的卖萌屋 2021-08-25 12:05

文 | 小伟
编 | 小轶

导师：小伟，听说你对知识蒸馏比较了解，你来给我说说知识蒸馏有什么用？

我：知识蒸馏是一种很典型的模型压缩的方法，我们可以用它来有效地从大型教师模型学习小型学生模型，并且学生模型的性能也很不错。

导师：那它既然叫知识蒸馏，你怎么知道学生模型是不是真的充分学到了教师模型的知识呢？

我：这不简单嘛，学生模型的效果好不就说明学到充足的知识了。

导师：一看你就不关心最新的学术进展，天天是不是忙着吃鸡了！最近NYU和Google联合发了一篇文章，仔细探索了知识蒸馏中学生模型的精度和蒸馏效果之间的关系，快去读一读！

我：好嘞~

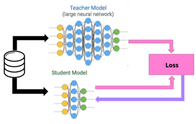
论文标题：

Does Knowledge Distillation Really Work?

论文地址：

https://arxiv.org/pdf/2106.05945

arxiv达同惯的小伙伴也可以在 **【夕小瑶的卖萌屋】**订阅号后台回复关键词 **【0825】** 下载论文PDF~



随着自然语言处理进入了预训练模型的时代，模型的规模也在极速增长，GPT-3甚至有1750亿参数，如何在资源有限的情况下部署使用这些庞大的模型是一个很大的挑战。

目前知识蒸馏在解决这一问题中的方法中占据了重要的地位。我们可以通过知识蒸馏来学习容易使用的小型学生模型，但是**它真的可以起到蒸馏教师模型知识的效果吗**？在这篇文章中，作者对这一问题进行了详细的探索与解答，下面我们就来一探究竟。

概要

尽管目前有很多知识蒸馏相关的研究，但它们主要集中于如何提高学生模型的泛化性(**generalization**)，往往忽视了学生模型的预测和教师模型的预测的匹配程度(**fidelity**)。我们可以简单称之为学生模型的**匹配度**，相比泛化性，匹配度更好的反映了学生模型蒸馏到了多少教师模型含有的知识。

本文对这两种概念做了详细的解释与区分，并且指出获得良好的匹配度对学生模型来说往往是很困难的。

基于此现象，作者探索了两种可能导致这种困难的原因：

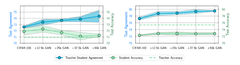
- **Identifiability**: 蒸馏**数据**不够充足，所以在训练数据上学生-教师预测可以匹配并不意味着在测试数据上也可以匹配。
- **Optimization**: 我们不能保证他解决蒸馏**优化**问题，所以不管是在训练数据还是测试数据上，学生模型的匹配度都比较低。

为什么需要匹配度？

之前的研究已经揭示了知识蒸馏通常会提高学生模型的泛化能力，所以我们为什么还要关心学生模型的匹配度呢？

- 首先是学生模型的泛化性能和教师模型的泛化性能往往有比较大的差距，提高匹配度是消**除学生和教师泛化性能差异**最显而易见的方法。
- 其次良好的学生模型匹配度可以提高知识蒸馏的可**解释性**与可**信赖性**。
- 最后，将匹配度和泛化性**解耦**可以帮助更好的理解知识蒸馏是怎么工作的以及如何在各种应用程序中更好的利用它

学生模型的匹配度高吗？



(a) Self-distillation (b) Ensemble distillation

Figure 1: **Evaluating the fidelity of knowledge distillation.** The effect of enlarging the CIFAR-100 distillation dataset with GAN-generated samples. (**a**). The student and teacher are both single ResNet-56 networks. Student fidelity increases as the dataset grows, but test accuracy decreases. (**b**). The student is a single ResNet-56 network and the teacher is a 3-component ensemble. Student fidelity again increases as the dataset grows, but test accuracy now slightly increases. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials.

作者使用3个ResNet-56网络的集成来作为教师模型，使用单个的ResNet-56网络来作为学生模型进行知识蒸馏。

如 Figure 1(b) 显示，学生模型和教师模型的预测之前有着显著的差距(Low Test Agreement)，也就是**低匹配度**。

导致低匹配度的原因

学生模型的匹配度比较差，是什么原因导致的呢？

作者给出了两个可能的原因，并进行了相应的探索与验证。

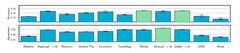


Figure 3: **Data augmentation and distillation**: Test accuracy and teacher-student agreement when distilling a 5-component ResNet-56 teacher ensemble into a ResNet-56 student on CIFAR-100 with varying augmentation policies. The best performing policy is shown in green, results averaged over 3 runs. Additional metrics are reported in Figure 12 in Appendix C. Mixup and GAN augmentation provide the best generalization, and Mixup+ $\tau = 0$ provides the best fidelity. The baseline policy (crops and flips) with $\tau = 4$ is a surprisingly strong baseline. The error bars indicate $\pm\sigma$.

首先可能是蒸馏数据不够充足

作者设计实验探索了在蒸馏过程中使用不同的数据增强策略是否可以提高学生模型的匹配度。

如Figure 3所示，实验结果证明虽然数据增强在很小学生-教师模型预测差距上是有用的，但它带来的改进很小，所以**蒸馏数据不足不是导致低匹配度的主要原因**。

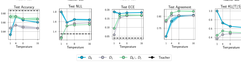


Figure 4: **Data recycling and distillation**: results on subsampled CIFAR-100. **Top**: We fix the temperature ($\tau = 4$) and vary the number of ensemble components (m), comparing students distilled on the same dataset as the teacher ($\mathcal{D}_0 \cup \mathcal{D}_1$), a reserved dataset ($\mathcal{D}_2 \cup \mathcal{D}_3$), or both ($\mathcal{D}_0 \cup \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$). Distilling on both produces the best result, while distilling on $\mathcal{D}_2$ increases accuracy and decreases fidelity, relative to $\mathcal{D}_0$. **Bottom**: We repeat the experiment, but fix $m = 3$ and vary τ. The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials.

既然蒸馏数据数量不足不是主要原因，那么会不会是蒸馏数据的分布偏移导致的呢？

作者同样设计了实验来验证这一猜想，Figure 4中的实验结果显示调整蒸馏数据的分布可以带来微小的改进，这也证明了**数据的错误选择不是导致低匹配度的主要原因**。

其次可能是蒸馏过程中的优化有问题

Figure 5: The train agreement for teacher ensembles ($m \in \{1, 3, 5\}$) and student on the distillation data for a ResNet-56 on CIFAR-100 under different augmentation policies. In all panels, increasing the softness of the teacher labels by adding examples not in the teacher train data makes distillation more difficult. **Left**: agreement for the synthetic GAN-augmentation policy from Figure 1. **Middle**: agreement from subsampled CIFAR-100 experiment in Figure 4. **Right**: agreement for some of the augmentation policies in Figure 3. The shaded region is not visible because the variance is very low.

既然数据不够充足或者分布有偏移不是主要原因，那么会不会是因为学生模型本来就没有充分的学习训练集，导致即使是在训练数据上学生-教师预测的差异都很大呢？

为了验证这一猜想，作者使用ResNet-56以及CIFAR-100设计了简单的蒸馏实验。

如Figure 5所示，当使用广泛的数据增强策略时，即使是在训练集上，学生模型的匹配度也会比较低，这印证了我们的猜想，**学生模型没有充分的学习训练数据**。

那么为什么即使是在训练集上，学生-教师模型预测的匹配度都很低呢？原因其实很简单，**知识蒸馏的优化会收敛于次优解**，从而导致低匹配度。

总结

作者总结了本文的关键发现：

- 学生模型的**泛化性能**(generalization)和**匹配度**(fidelity)的**变化趋势并不一致**
- 学生模型的**匹配度**(fidelity)和蒸馏的**校准**(calibration)有很大的**关联**
- 知识蒸馏过程中的**优化是很困难的**，这也是导致低匹配度的主要原因
- 蒸馏优化的复杂度以及蒸馏数据的质量之间存在**均衡**(trade-off)

Final Question

现在回到我们最开始的问题，知识蒸馏真的有用吗？

毫无疑问，它是有用的，因为它通常可以提高学生模型的泛化性能；同时它也是没用的，因为它总是很难有效的知识从教师传递到学生，这引它的名字知识蒸馏相去甚远，所以我们是不是可以改称它为knowledge enhancement？

小伙伴，你们怎么看呢？

最后的最后，

导师：小伙子不错啊，论文读得又快又精准，这周的Reading Group就交给你了。

我：妈妈咪，又是我，浪去读论文了~