The Lottery Ticket Hypothesis: 寻找最优子网络结构

2019年05月08日 16:27:11 Law-Yao 阅读数：191

"The Lottery Ticket Hypothesis Finding Sparse, Trainable Neural Networks" 这篇文章提出了**Lottery Ticket Hypothesis**，认为较复杂的深度神个比较优化的稀疏子网络结构（称之为**winning tickets**），可应用于模型压缩。相比于原网络，稀疏子网络的参数量与复杂度要低许多，但推理精度 Lottery Ticket Hypothesis描述如下：

> **The Lottery Ticket Hypothesis.** *A randomly-initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.*

Lottery Ticket Hypothesis用符号方式描述如下：

> More formally, consider a dense feed-forward neural network $f(x;\theta)$ with initial parameters $\theta = \theta_0 \sim \mathcal{D}_\theta$. When optimizing with stochastic gradient descent (SGD) on a training set, f reaches minimum validation loss l at iteration j with test accuracy a. In addition, consider training $f(x; m\odot\theta)$ with a mask $m \in \{0,1\}^{|\theta|}$ on its parameters such that its initialization is $m \odot \theta_0$. When optimizing with SGD on the same training set (with m fixed), f reaches minimum validation loss l' at iteration j' with test accuracy a'. The lottery ticket hypothesis predicts that $\exists\, m$ for which $j' \leq j$ (*commensurate training time*), $a' \geq a$ (*commensurate accuracy*), and $\|m\|_0 \ll |\theta|$ (*fewer parameters*).

其中剪枝获得的子网络从随机初始化开始训练，且初始化数值一一对应地取自原网络的初始化数值集合，即$m \odot \theta_0$；另外，子网络训练达到收敛的迭超过原网络所需的迭代次数。如果子网络的初始化数值不取自原网络，而是按新的随机初始化方式执行训练，通常不会达到原网络的推理精度，说明剪络需要合适的初始化状态。子网络（winning ticket）的搜索步骤如下：

1. Randomly initialize a neural network $f(x;\theta_0)$ (where $\theta_0 \sim \mathcal{D}_\theta$).
2. Train the network for j iterations, arriving at parameters θ_j.
3. Prune $p\%$ of the parameters in θ_j, creating a mask m.
4. Reset the remaining parameters to their values in θ_0, creating the winning ticket $f(x; m\odot\theta_0)$.

相比于上述one-shot方式的搜索方法，多次迭代方式能够获得更轻量的子网络结构（winning ticket），文章采用n次迭代，每次将剪枝率设置为$p^{\frac{1}{n}}$合drop-out策略，能够进一步提升效果。

文章最后在MNIST、CIFAR10数据集上，对全连接层网络与卷积网络做了实验验证。然而这些任务涉及的数据集都很小，设计的深度模型本身存在较大与过拟合倾向，因此可能不足以说明问题。不过正如文章所说，子网络的随机初始化方式对于理解与揭示深度学习的本质，或许是一个启发性的观点。助于搜索优化的网络结构。

Paper地址： https://arxiv.org/abs/1803.03635

GitHub地址： https://github.com/google-research/lottery-ticket-hypothesis

想对作者说点什么