谷歌重磅：可以优化自己的优化器！手动调参或将成为历史！？

原创 小轶 夕小瑶的卖萌屋 2020-10-20 09:00

收录于话题
#卖萌屋@深度学习与炼丹技巧

28个

文 | 小轶
编 | 夕小瑶

背景

Google Brain团队发布的一篇最新论文在外网引发热议，或将成为Deep Learning发展历程上里程碑式的工作。它所讨论的，是所有AI行业者都要面对的——Deep Learning中的优化问题。也就是，**如何更好地训练一个模型**。

深度模型的训练过程是非常困难的，常见的挑战包括：陷入局部极小值、梯度消失/爆炸、长期依赖（long dependency）等等。但对于大多数算法工程师来说其实并没有这么复杂。因为学术界早已陆续提出了许多卓有成效的**优化器**，比如AdaGrad、Adam、Momentum等等，都可以一定程度解决上述种种问题。而算法工程师搭完模型后，需要做的只有一件事——**调参** :)

如果说深度学习的兴起为算法工程师省去了繁琐的特征工程（特征设计与特征选择），今天介绍的Google这篇工作就是致力于为大家省去繁琐的“调参工程”（优化器设计与优化器选择）。

深度学习用大量的训练数据替代了特征工程，同样的道理，这篇工作致力于用大量训练任务和模型来替代人工设计的优化器（Adam、Momentum等），这种以任务和模型为食的general-purpose的优化器模型，就称之为**learned optimizer**，可广泛适用于各类任务，无需手动调节优化器参数（如学习率，batch size...）。

实验不仅证明了learned optimizer的普适性，更是发现了这种优化器的一些惊人特性。比如，它甚至可以根据训练过程中的validation loss，隐性地做到正则化规约。最令人惊叹的是，该优化器甚至可以用来从头训练一个新的general-pupose优化器——也就是说，**这是一个可以自己优化自己的优化器！**

论文题目:

《Tasks, stability, architecture, and compute: Training more effective learned optimizers, and using them to train themselves》

论文链接:

https://arxiv.org/pdf/2009.11243.pdf

Arxiv访问慢的小伙伴也可以在 **【夕小瑶的卖萌屋】**订阅号后台回复关键词 **【1020】** 下载论文PDF~

方法

接下来，我们就来看看这个神仙优化器是如何训出来的。在探讨其具体模型结构之前，我们先来理清楚优化器训练所需要的是什么样的数据集，以及目标函数是什么。

优化器训练的数据集

learned optimizer(下文简称Opt_L)的训练所需要的**每个训练样本x都是一个需要在某任务上训练的深度学习模型，样本的标签y则是该模型在其对应任务上的开发集loss**，即训练集为:

$$X = Set(Model_1, Model_2, \ldots, Model_n)$$
$$Y = Set(DevLoss_{Model_1}, DevLoss_{Model_2}, \ldots, DevLoss_{Model_n})$$

对于数据集里的每个训练样本x（模型），都

- 可以采用不同的模型结构
- 用于完成不同的任务，称为 **inner-task**
- 有属于自己的数据集，称为 **inner-dataset**

比如，

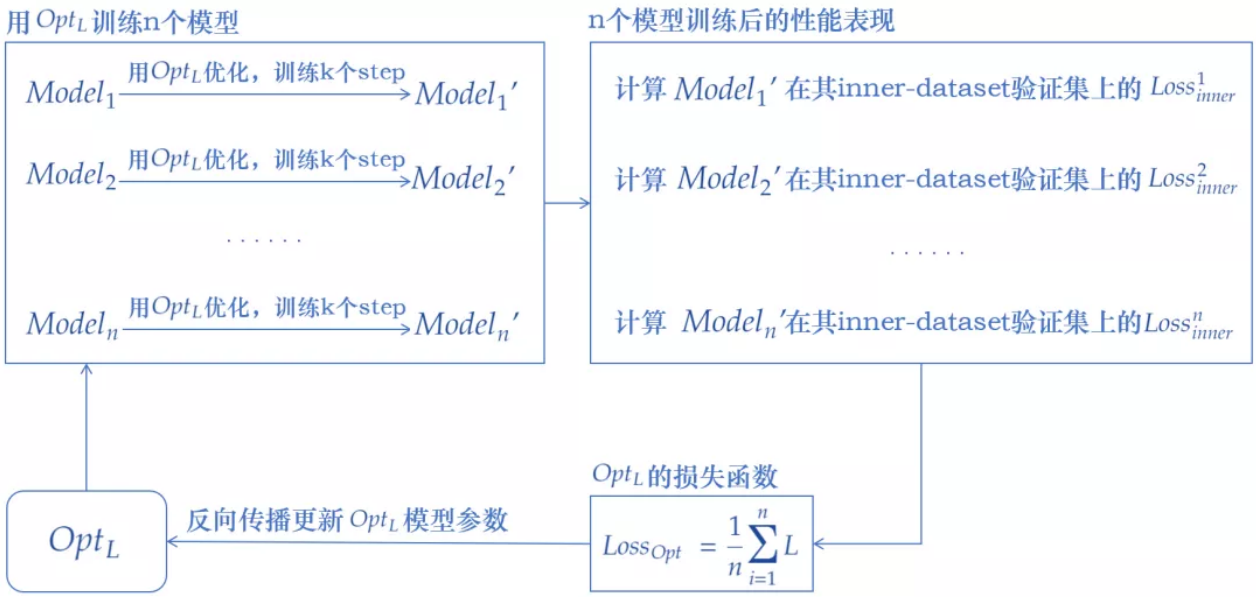
- $Model_1$可能是一个用于文本分类的RNN,用的inner-dataset是YELP-5
- $Model_2$可能是一个做图像分类的CNN，用的inner-dataset是数据集CIFAR-10

作者实际共设置了 **6000** 个不同种类的模型。涵盖了RNNs、CNNs、mask auto regressive flows、全连接网络、语言模型、VAE、simple 2D test function、quadratic bowls等...

优化器训练的目标函数

我们都知道，通常一个深度学习模型的训练就需要极大的算力支撑。而此处令人咋舌的是，按照上述设定，**我们需要完成6000个模型的训练才能为learned optimizer（Opt_L）完成1轮训练。**

Opt_L的一轮训练过程大致如下图所示（为说明得更加清楚，图中采用的是full batch进行参数更新，也就是每个batch直接包含全部样本）：



1. 先用Opt_L训练n个$Model$（理想情况下，每个$Model$应该一直训练到收敛，但考虑到算力的问题，实际上训练240~360个step就停止了）
2. 每个$Model_i$都有自己的inner-dataset，我们在它的inner-dataset的验证集上计算$Model_i$的损失函数$loss^i_{inner}$
3. Opt_L的损失函数即为**所有$Model$的$loss_{inner}$的平均**
4. 用Opt_L的损失函数对其进行参数更新

优化器的结构

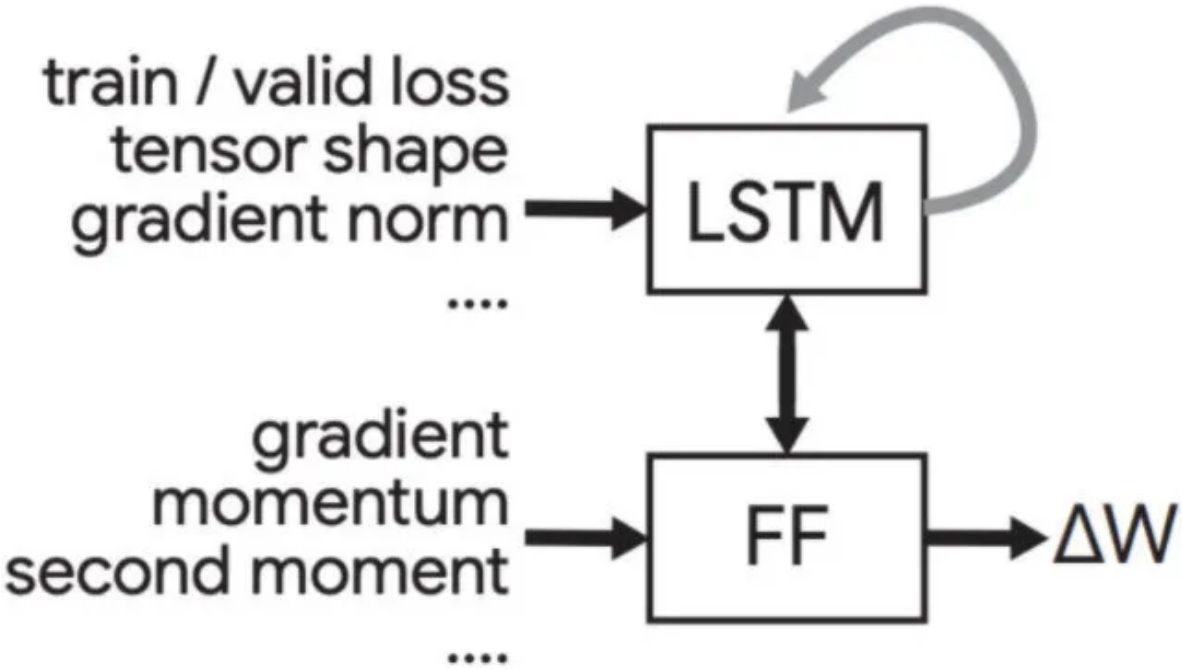
其实learned optimizer的概念并不是在这篇论文中首次提出来的，不过论文作者argue了learned optimizer的结构和优化器训练所基于的任务集都会非常非常影响最终learned optimizer的表现。因此本文提出了一种**层级的优化器结构**，实验表明优于前人提出的learned optimizer结构。

设计learned optimizer结构的关键是**平衡计算效率和表达能力**。

> ps：预训练时代的军备竞赛可以疯狂追求模型表达能力，不顾及计算效率（想想BERT和Google T5放出时的恐惧）。但是优化器模型就不能这么任性了，TPU也耗不起

因此，优化器结构一般都不会太复杂，如下图所示

上图的优化器结构是ICML2019上提出的，使用了一个全连接网络（Feed-Forward，FF）。当模型完成了一个step的训练后，就用这个FF对每个参数进行更新。FF的输入端是模型某个参数w的梯度，以及该参数的其他feature（如Momentum等）。FF的输出端是w的更新值Δw，则该参数将被更新为$w + \Delta w$。注意，**这个FF每跑一次，只完成了一个参数的更新**。

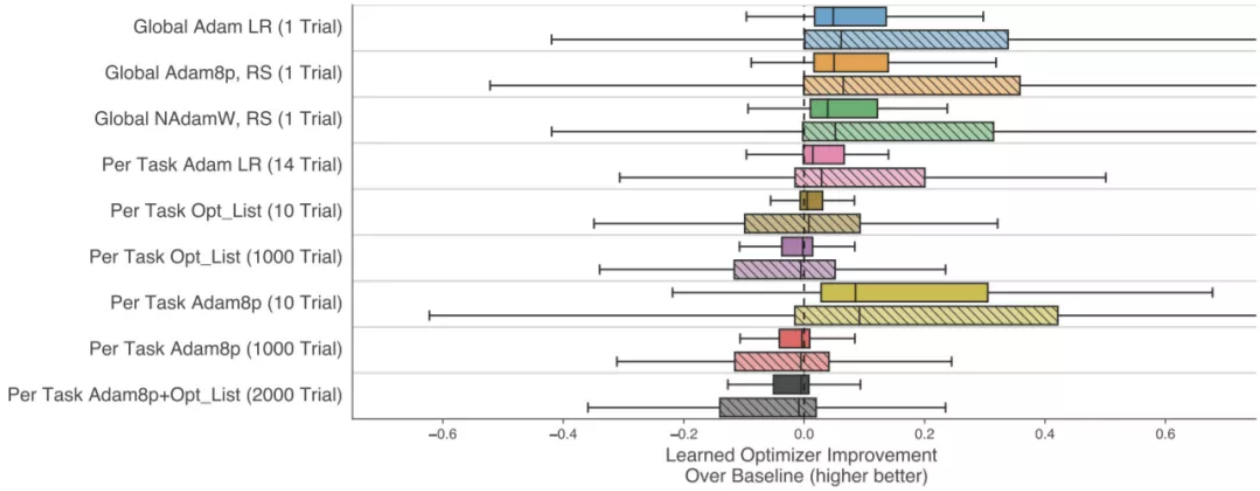


上图就是paper中提出的优化器结构了。下半部分的FF与上面ICML2019的优化器实现类似，都是用于求某个参数的更新值，称为**Per-parameter FF**。与之前不同的是，这个FF还会接收到**全局信息**（如train/valid loss），以及该参数**所在张量的信息**（如张量形状，gradient norm等）。相关信息来自于上方的LSTM。文中称其为**Per-tensor LSTM**。

实验

与常见优化器的比较

下图展示了与常见优化器（AdamLR、Adam8p、opt_list）的比较结果。实验中，总共测试了100个任务下使用learned optimizer后的性能提升比例。在各个任务上提升比例分布用箱图表示。

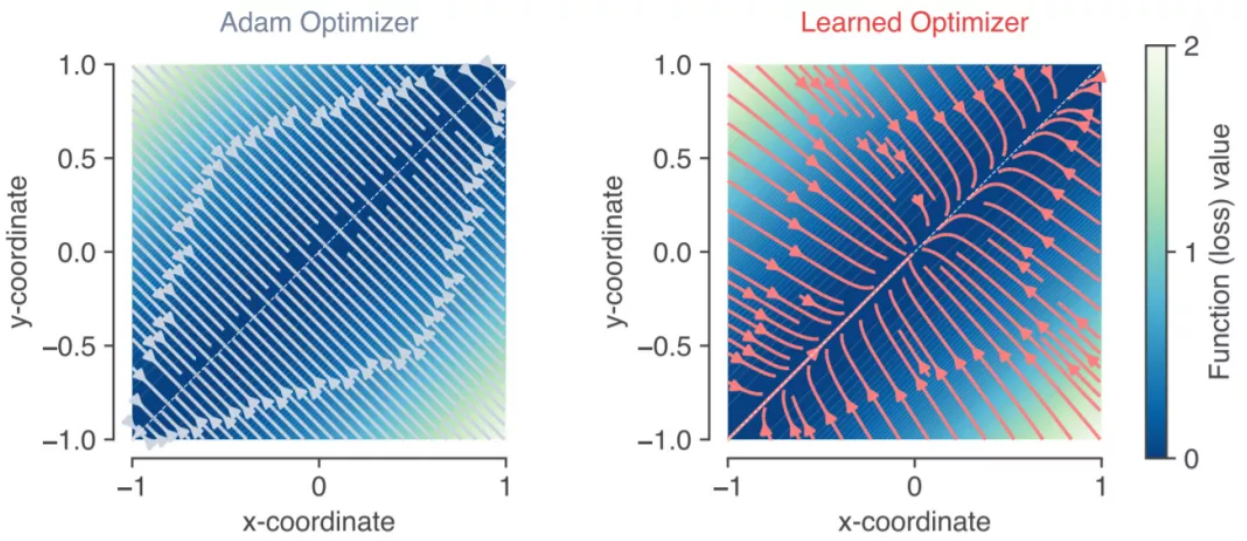


> 纵轴代表了不同设置下的三个basline优化器。最上面3个*Global XXX*的设定是：该baseline优化器对于所有任务都采用相同的超参数。而下面6个*Per Task XXX*对不同任务可以采用不同超参数，括号中的*XXX Trial*代表尝试调参的轮数。每一种baseline，都对应了两条同色系的箱图。这是因为用于测试的100个测试任务中，有一部分是learned optimer训练过程中见过的，有一部分从未见过。同色系的两个箱图中，上面那条代表在见过的那些任务上的提升效果，另一条代表在从未见过的那部分任务上的提升效果。

图中箱图的分布并不十分集中，可见提升效果对于不同的任务也各不相同。但总体来说，与适度调参的baseline相比，都有一定程度的提升效果。

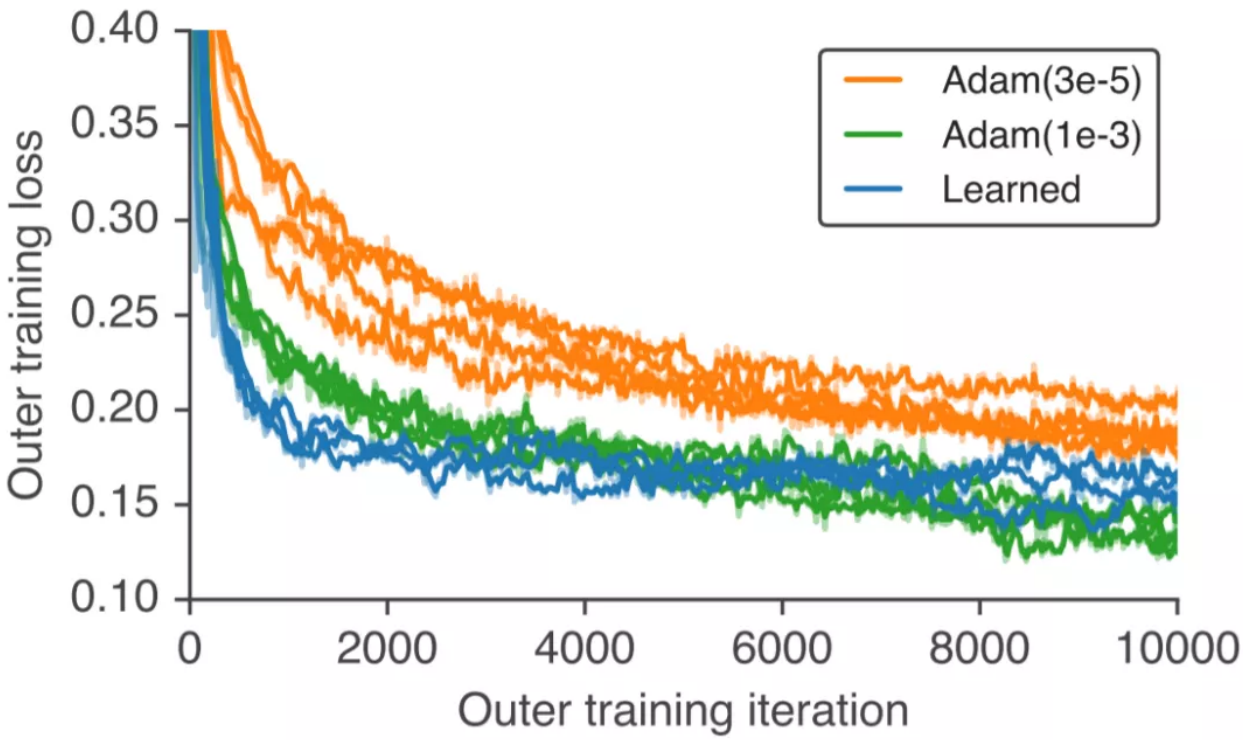
隐性的正则化惩罚项

在机器学习中，时常会在目标函数中加入正则化惩罚项，从而对模型的复杂度进行规约。下图展示了Adam和learned optimizer在优化目标函数$f(x,y)=\frac{1}{2}(x-y)^2$时的收敛轨迹。显然直线$y=x$上目标函数最小。但可以看到Adam会直接垂直地收敛到$y=x$上。而learned optimizer在收敛过程中还会有**逐渐接近原点**的趋势。作者认为这是由于接近原点处的(x,y)范数较小，表明learned optimizer有隐式地进行正则化规约。



可以优化自己的优化器

最后，Google Brain团队脑洞大开地用这个learned optimizer再从头训练一个新的自己！作为比较的是，作者在训练它的时候使用的两种优化器设置（图中橙色和绿色曲线）。可以看到learned optimizer取得了非常相近的训练曲线。作者认为，这个实验进一步证明了该优化器的超强普适性。因为，对优化器进行优化是一个全新的任务，与这个优化器训练过程中见过的所有任务都完全不同。



小结

一个可以不用调参、适用于所有训练任务的优化器。如此的脑洞大开、又敢想敢做，不知道除了Google还有哪里可以。

萌屋作者：小轶

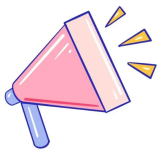
刚刚本科毕业于北大计算机系的美少女学霸！目前在腾讯天衍实验室做NLP研究实习生。原计划是要赴美国就读CMU的王牌硕士项目MCDS，不过因为疫情正处于gap year，于是就来和小夕愉快地玩耍啦~文风温柔优雅，偶尔暴露呆萌属性，文如其人哦！知乎ID：小轶。

作品推荐：

1.有钱可以多任性？OpenAI提出人肉模型训练，文本摘要全面超越人类表现！

2.ACL20 Best Paper揭晓！NLP模型评价体系或将迎来重大转折

3.Attention模型：我的注意力跟你们人类不一样



后台回复关键词【**入群**】

加入卖萌屋NLP/IR/Rec与求职讨论群

有顶会审稿人、大厂研究员、知乎大V和妹纸

等你来撩哦~

我不看好data2vec这类多模态融合的研究

夕小瑶的卖萌屋