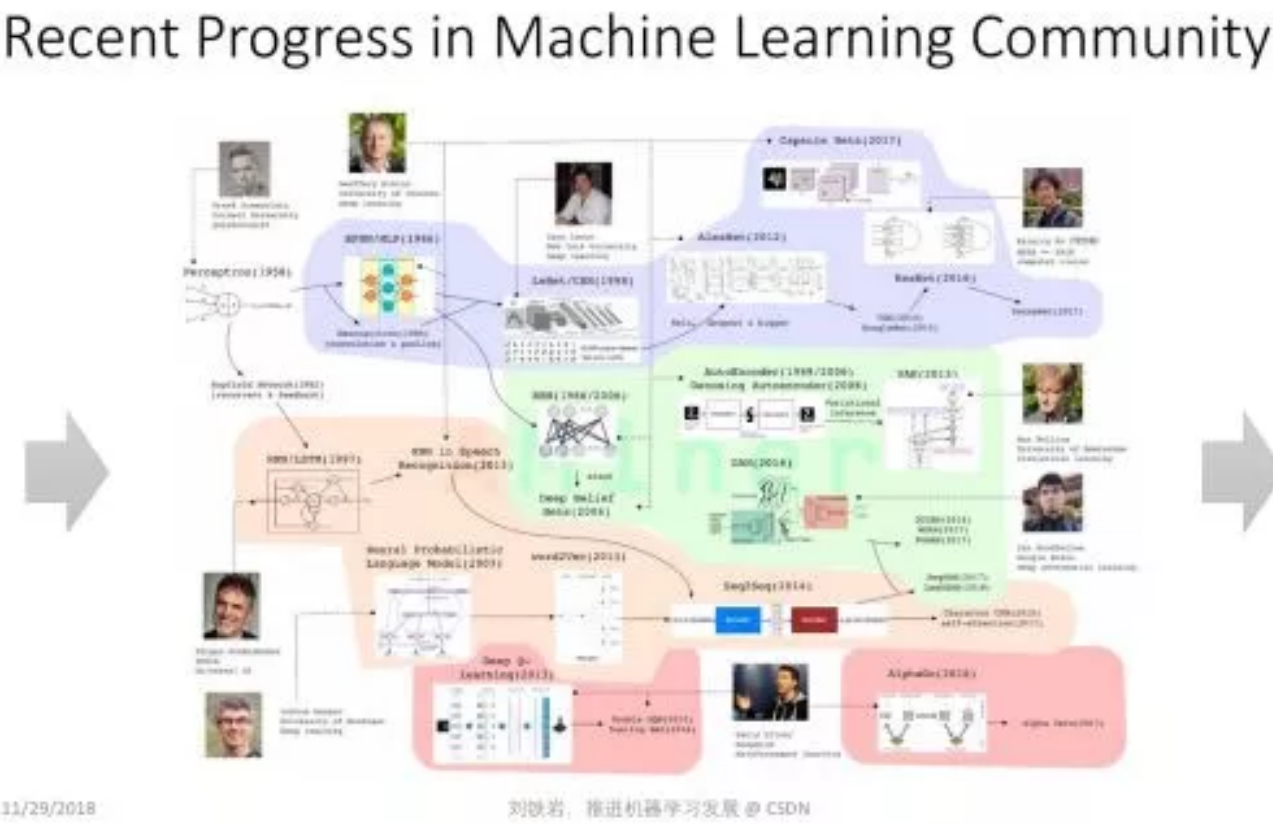


刘铁岩谈机器学习：随波逐流的太多，我们需要反思

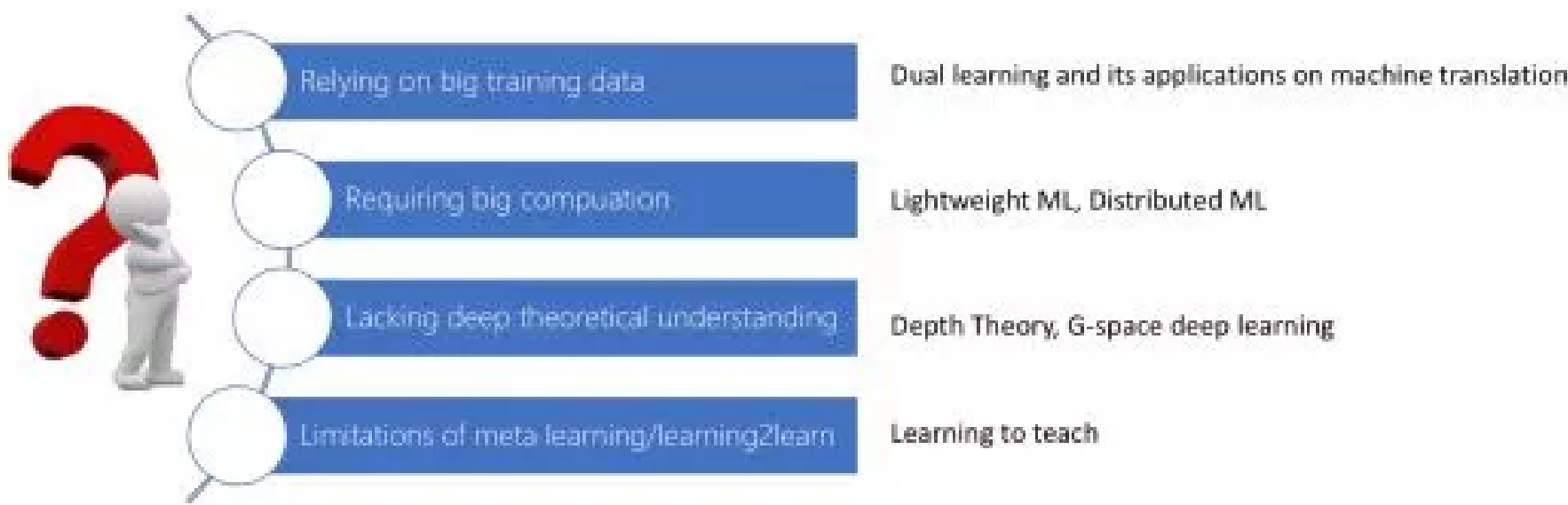
编者按：人工智能正受到越来越多的关注，而这波人工智能浪潮背后的最大推手就是“机器学习”。机器学习从业者在当下需要掌握哪些前沿技术？展望未来，又会有哪些技术趋势值得期待？近期，微软亚洲研究院副院长刘铁岩博士受AI科技大本营和华章科技邀请，参与了一堂在线公开课，与大家分享微软研究院最新的研究成果，以及对机器学习领域未来发展趋势的展望。以下是本次公开课的精彩内容。本文授权转载自AI科技大本营（ID: rgznai100）。

大家好，我是刘铁岩，来自微软亚洲研究院。今天非常荣幸，能跟大家一起分享一下微软研究院在机器学习领域取得的一些最新研究成果。



大家都知道，最近这几年机器学习非常火，也取得了很多进展。这张图总结了机器学习领域的最新工作，比如ResNet、胶囊网络、Seq2Seq Model、Attention Mechanism 、GAN、Deep Reinforcement Learning 等等。

Our Research



11/29/2018

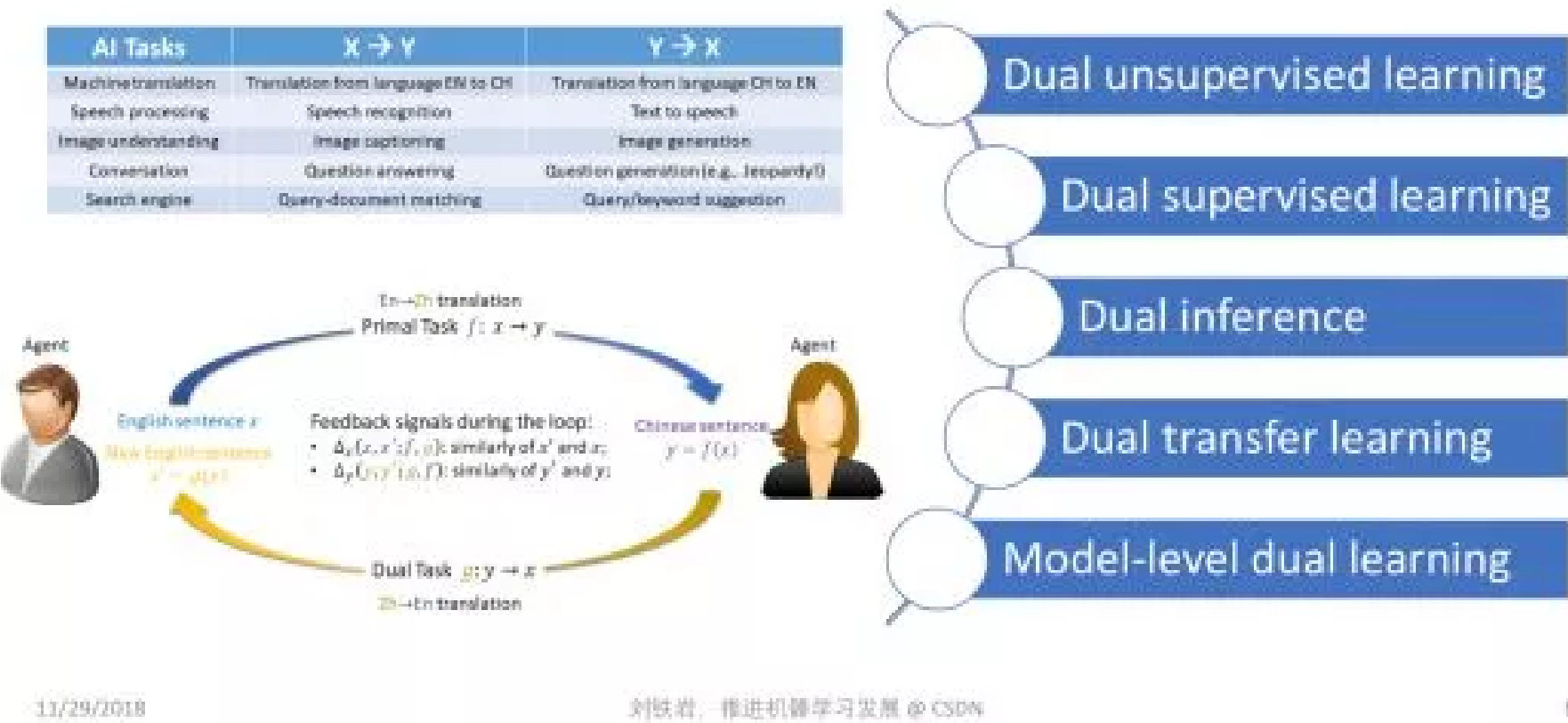
刘铁岩, 推进机器学习发展 @ CSDN

4

这些成果推动了机器学习领域的飞速发展，但这并不意味着机器学习领域已经非常成熟，事实上仍然存在非常大的技术挑战。比如现在主流机器学习算法需要依赖大量的训练数据和计算资源，才能训练出性能比较好的机器学习模型。同时，虽然深度学习大行其道，但我们对深度学习的理解，尤其是理论方面的理解还非常有限。深度学习为什么会有效，深度学习优化的损失函数曲面是什么样子？经典优化算法的优化路径如何？最近一段时间，学者们在这个方向做了很多有益的尝试，比如讨论随机梯度下降法在什么条件下可以找到全局最优解，或者它所得到的局部最优解跟全局最优解之间存在何种关系。

再比如，最近很多学者开始用自动化的方式帮助机器学习尤其是深度学习来调节超参数、搜寻神经网络的结构，相关领域称为元学习。其基本思想是用一个机器学习算法去自动地指导另一个机器学习算法的训练过程。但是我们必须承认，元学习其实并没有走出机器学习的基本框架。更有趣的问题是，如何能够让一个机器学习算法去帮助另一个算法突破机器的现有边界，让机器的效果更好呢？这都是我们需要去回答的问题。沿着这些挑战，在过去的这几年来，微软亚洲研究院做了一些非常有探索性的学术研究。

对偶学习解决机器学习对大量有标签数据的依赖



首先，我们看看对偶学习。对偶学习主要是为了解决现有深度学习方法对训练数据过度依赖的问题。当我们没有标注好的训练数据时，是否还能做有意义的机器学习？在过去的几年里，人们做了很多尝试，比如无监督学习、半监督学习等等。但是无论如何，大家心里要清楚，只有有信号、有反馈、才能实现有效的学习，如果我们对这个世界一无所知，我们是不能进行有效的学习的。

沿着这个思路，我们在思考：除了人为提供的标签以外，是不是存在其他有效的反馈信号，能够形成学习的闭环？我们发现很多机器学习任务其实天然有结构对偶性，可以形成天然的闭环。

比如机器翻译。一方面我们会关心从英文翻译到中文，另一方面我们一定也关心从中文翻译到英文，否则就无法实现两个语种人群之间的无缝交流。再比如语音处理。我们关心语音识别的同时一定也关心语音合成，否则人和机器之间就没有办法实现真正的双向对话。还有图像理解、对话引擎、搜索引擎等等，其实它们都包含具有对偶结构的一对任务。

如何更加准确地界定人工智能的结构对偶性呢？我们说：如果第一个任务的输入恰好是第二个任务的输出，而第一个任务的输出恰好是第二个任务的输入，那么这两个任务之间就形成了某种结构的“对偶性”。把它们放在一起就会形成学习的闭环，这就是“对偶学习”的基本思想。

有了这样的思想以后，我们可以把两个对偶任务放到一起学，提供有效的反馈信号。这样即便没有非常多的标注样本，我们仍然可以提取出有效的信号进行学习。

对偶学习背后其实有着严格的数学解释。当两个任务互为对偶时，我们可以建立如下的概率联系：

Probabilistic Nature of Dual Learning

- The structural duality implies strong probabilistic connections between the models of dual AI tasks.

$$P(x, y) = P(x)P(y|x; f) = P(y)P(x|y; g)$$

Primal View

Dual View

- This can be used as
 - Effective feedback signal to close the loop of unsupervised learning
 - Structural regularizer to enhance supervised learning
 - Additional criterion to improve inference

11/29/2018

刘铁岩：普适机器学习发展 @ CSDN

9

这里 X 和 Y 分别对应某个任务的输入空间和输出空间，在计算 X 和 Y 的联合概率分布时有两种分解方法，既可以分解成 $P(x)P(y|x; f)$ ，也可以分解成 $P(y)P(x|y; g)$ 。这里， $P(y|x; f)$ 对应了一个机器学习模型，当我们知道输入 x 时，通过这个模型可以预测输出 y 的概率，我们把这个模型叫主任务的机器学习模型， $P(x|y; g)$ 则是反过来，称之为对偶任务的机器学习模型。

有了这个数学联系以后，我们既可以做有效的无监督学习，也可以做更好的有监督学习和推断。比如我们利用这个联系可以定义一个正则项，使得有监督学习有更好的泛化能力。再比如，根据 $P(x)P(y|x; f)$ 我们可以得到一个推断的结果，反过来利用贝叶斯公式，我们还可以得到用反向模型 g 做的推断，综合两种推断，我们可以得到更准确的结果。我们把以上提到的对偶学习技术应用在了机器翻译上，取得了非常好的效果，在中英新闻翻译任务上超过了普通人类的水平。

解决机器学习对大计算量的依赖

轻量级机器学习

最近一段时间，在机器学习领域有一些不好的风气。有些论文里会使用非常多的计算资源，比如动辄就会用到几百块 GPU 卡 甚至更多的计算资源。这样的结果很难复现，而且在一定程度上导致了学术研究的垄断和马太效应。

那么人们可能会问这样的问题：是不是机器学习一定要用到那么多的计算资源？我们能不能在计算资源少几个数量级的情况下，仍然训练出有意义的机器学习模型？这就是轻量级机器学习的研究目标。

LightLDA

- The largest/fastest topic model
 - Multiplicative factorization reduces per-token sampling complexity to $O(1)$, which is independent of topic number

$$p(z_{w_i} = k | rest) \propto \frac{n_{kw}^{-d_i} + \beta_w}{n_{k\cdot}^{-d_i} + \beta} (n_{k\cdot}^{-d_i} + \alpha_k)$$

	#Tokens	#Topics	CPU cores	Training time
LightLDA	100G	1M	384	60 hrs
Google's LDA	< 10G	< 100K	10,000	70 hrs

LightRNN

- Very compact and fast RNN
 - Multi-component embedding significant reduces the model size, especially for very large vocabulary

Classical RNN language model

- Model size > 100GB
- Training time > 100 years

↓

LightRNN language model

- Model size ~ 50MB
- Training time ~ 1 month

LightGBM

- The fastest GBDT tool
 - Gradient-based one-side sampling
 - Exclusive feature bundling
 - Voting-based parallelization

11/29/2018

刘铁岩：推进机器学习发展 @ CSDN

13

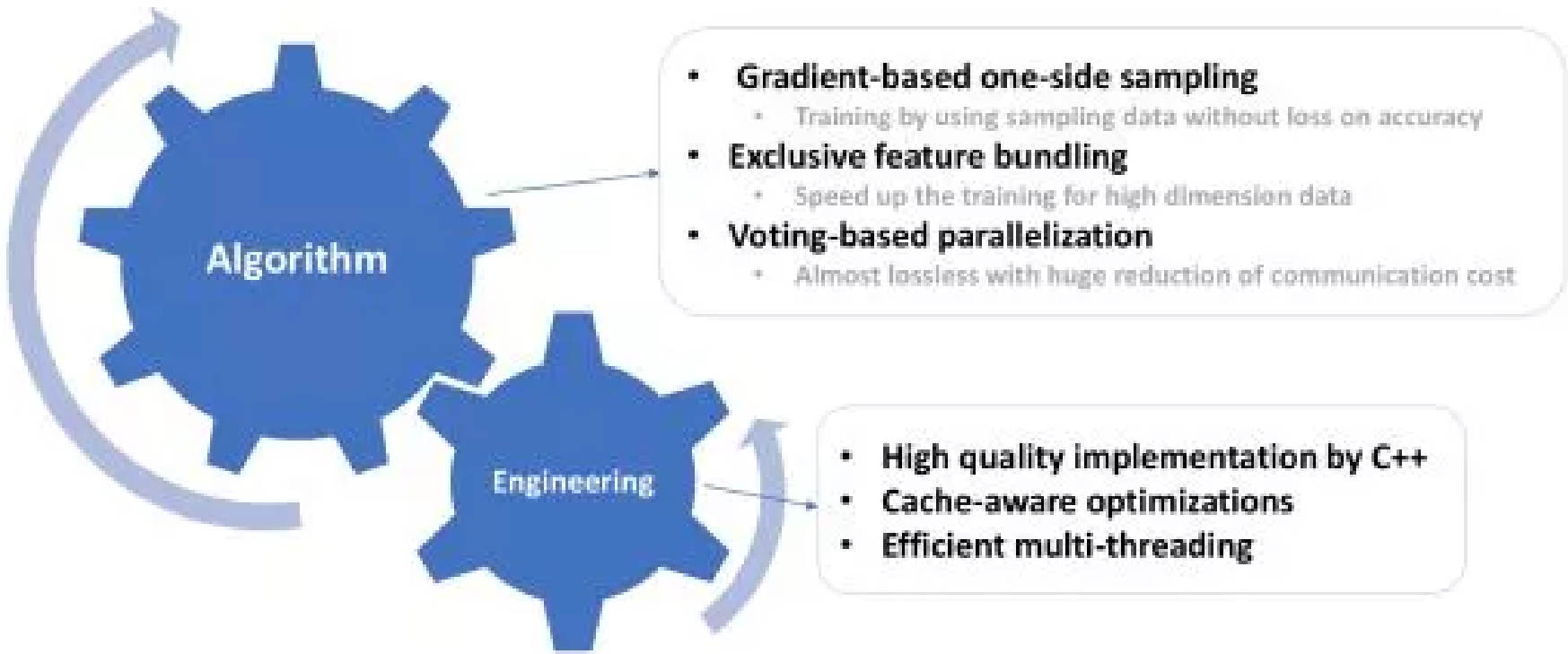
在过去的几年里，我们的研究组做了几个非常有趣的轻量级机器学习模型。比如在 2015 发表的 lightLDA 模型，它是一个非常高效的主题模型。在此之前，世界上已有的大规模主题模型一般会用到什么样的计算资源？比如 Google 的 LDA 使用上万个 CPU cores，才能够通过几十个小时的训练获得 10 万个主题。为了降低对计算资源的需求，我们设计了一个基于乘性分解的全新采样算法，把每一个 token 的平均采样复杂度降低到 $O(1)$ ，也就是说采样复杂度不随着主题数的变化而变化。因此即便我们使用这个主题模型去做非常大规模的主题分析，它的运算复杂度也是很低的。例如，我们只使用了 300 多个 CPU cores，也就是大概 8 台主流的机器，就可以实现超过 100 万个主题的主题分析。

这个例子告诉大家，其实有时我们不需要使用蛮力去解决问题，如果我们可以仔细分析这些算法背后的机理，做算法方面的创新，就可以在节省几个数量级计算资源的情况下做出更大、更有效的模型。

同样的思想我们应用到了神经网络上面，2016 年发表的 LightRNN算法是迄今为止循环神经网络里面最高效的实现。当我们用 LightRNN 做大规模的语言模型时，得到的模型规模比传统的 RNN 模型小好几个数量级。比如传统模型大小在100GB 时，LightRNN 模型只有50MB，并且训练时间大幅缩短。不仅如此，LightRNN模型的 perplexity比传统 RNN还要更好。

可能有些同学会产生疑问：怎么可能又小又好呢？其实，这来源于我们在循环神经网络语言模型的算法上所做的创新设计。我们把对 vocabulary 的表达从一维变到了二维，并且允许不同的词之间共享某一部分的 embedding。至于哪些部分共享、哪些不共享，我们使用了一个二分图匹配的算法来确定。

LightGBM (NIPS 2016/2017)



11/29/2018

刘敬岩, 推进机器学习发展 @ CSDN

14

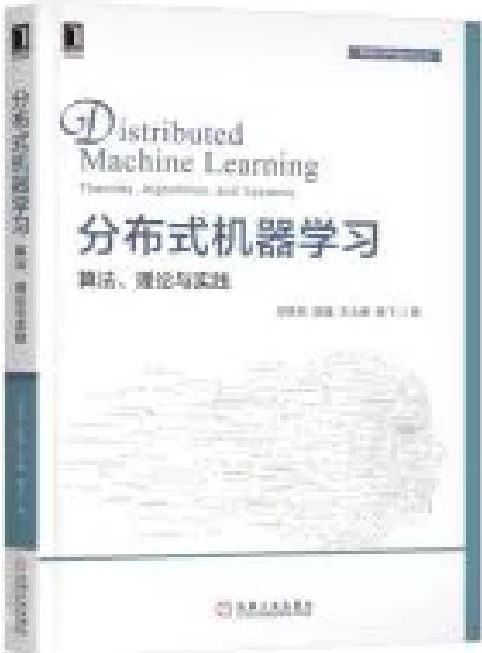
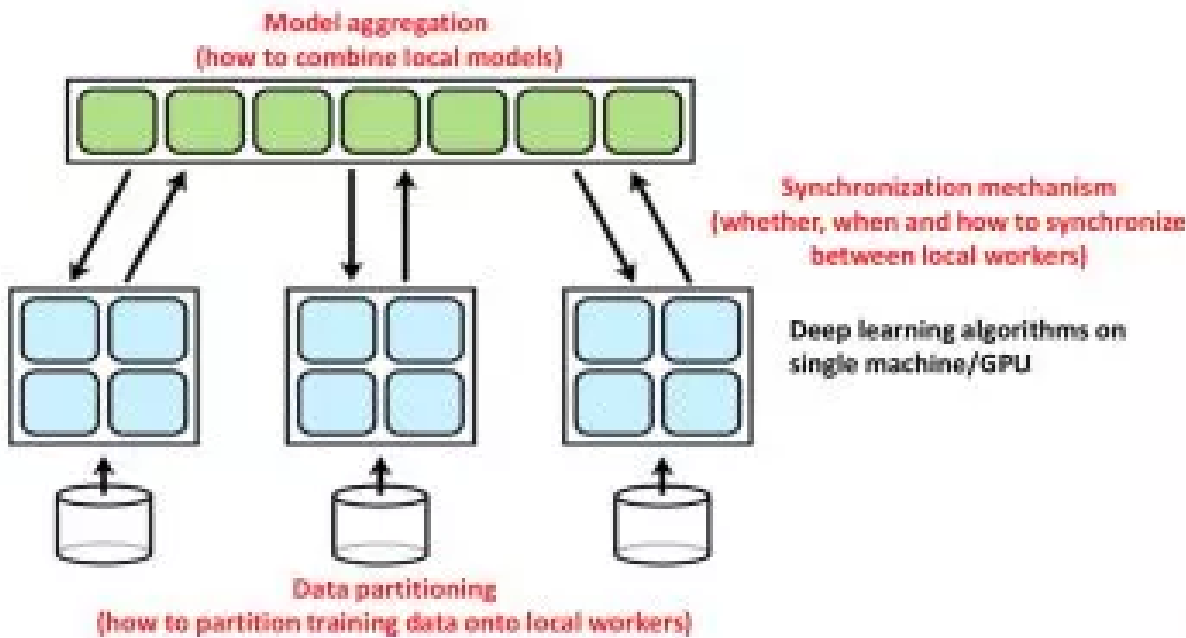
第三个轻量型机器学习的算法叫 LightGBM，这个工具是 GBDT 算法迄今为止最高效的实现。LightGBM的背后是两篇 NIPS 论文，其中同样包含了很多技术创新，比如 Gradient-based one-side sampling，可以有效减少对样本的依赖；Exclusive feature bundling，可以在特征非常多的情况下，把一些不会发生冲突的特征粘合成比较 dense 的少数特征，使得建立特征直方图非常高效。同时我们还提出了 Voting-based parallelization 机制，可以实现非常好的加速比。所有这些技巧合在一起，就成就了LightGBM的高效率和高精度。

分布式机器学习

虽然我们做了很多轻量级的机器学习算法，但是当训练数据和机器学习模型特别大的时候，可能还不能完全解决问题，这时我们需要研究怎样利用更多的计算节点实现分布式的机器学习。

Distributed Machine Learning

- Big data + Big model >> Capacity of a single machine



11/29/2018

刘敬岩, 推进机器学习发展 @ CSDN

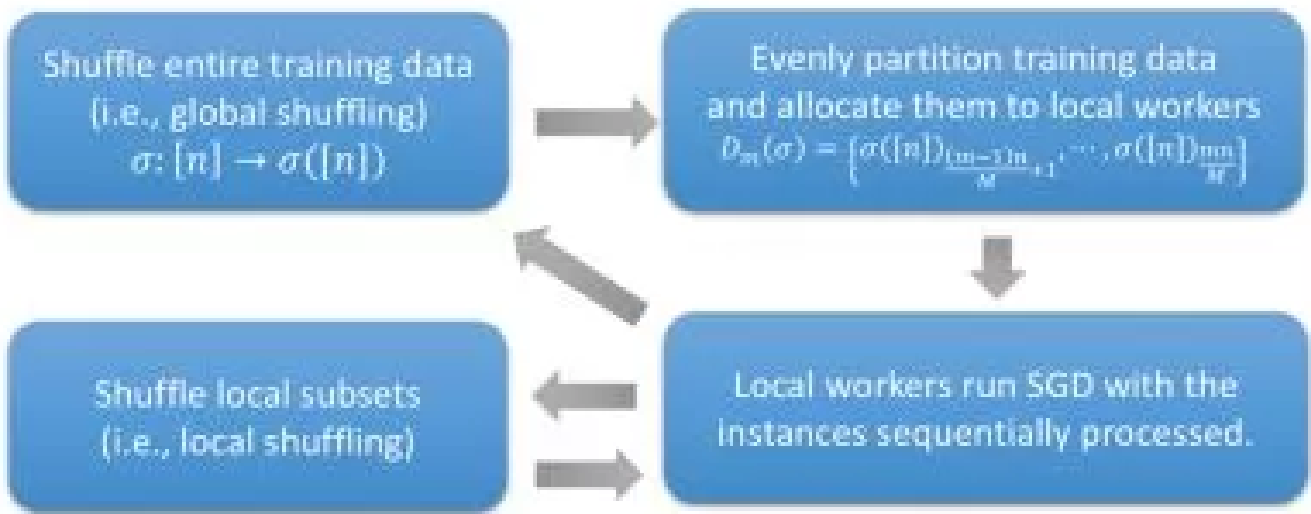
15

我们刚刚出版了一本新书——《分布式机器学习：算法、理论与实践》，对分布式机器学习做了非常好的总结，也把我们很多研究成果在这本书里做了详尽的描述。下面，我挑其中几个点，跟大家分享。

分布式机器学习的关键是怎样把要处理的大数据或大模型进行切分，在多个机器上做并行训练。一旦把这些数据和模型放到多个计算节点之后就会涉及到两个基本问题：首先，怎样实现不同机器之间的通信和同步，使得它们可以协作把机器学习模型训练好。其次，当每个计算节点都能够训练出一个局部模型之后，怎样把这些局部模型做聚合，最终形成一个统一的机器学习模型。

数据切分

Data Partitioning (NeuroComputing)



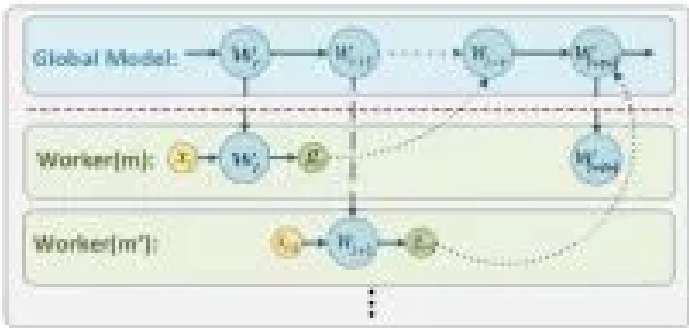
- *Global shuffling can achieve similar convergence rate to i.i.d. sampling, since the influence of small shuffling error is negligible.*
- *Local shuffling hurts the convergence rate, and we have to restrict the number of epochs when the number of local workers is large.*

数据切分听起来很简单，其实有很多门道。举个例子，一个常见的方式就是把数据做随机切分。比如我们有很多训练数据，随机切分成 N 份，并且把其中一份放到某个局部的工作节点上去训练。这种切分到底有没有理论保证？

我们知道机器学习有一个基本的假设，就是学习过程中的数据是独立同分布采样得来的，才有理论保证。但是前面提到的数据切分其实并不是随机的数据采样。从某种意义上讲，独立同分布采样是有放回抽样，而数据切分对应于无放回抽样。一个很有趣的理论问题是，我们在做数据切分时，是不是可以像有放回抽样一样，对学习过程有一定的理论保证呢？这个问题在我们的研究发表之前，学术界是没有完整答案的。

我们证明了：如果我先对数据进行全局置乱，然后再做数据切分，那么它和有放回的随机采样在收敛率上是基本一致的。但是如果只能做局部的数据打乱，二者之间的收敛率是有差距的。所以如果我们只能做局部的数据打乱，就不能训练太多 epoch，否则就会与原来的分布偏离过远，使得最后的学习效果不好。

异步通信



- Sequential SGD
$$w_{t+\tau+1} = w_{t+\tau} - \eta + g(w_{t+\tau})$$
- Async SGD
$$w_{t+\tau+1} = w_{t+\tau} - \eta + g(w_t) \neq$$
- Characterizing the delay using Taylor expansion:
$$g(w_{t+\tau}) = g(w_t) + \nabla g(w_t) \cdot (w_{t+\tau} - w_t) + O(\|w_{t+\tau} - w_t\|^2)$$

$\nabla^2 g(w_t)$ corresponds to the Hessian matrix

Delay Compensated ASGD (DC-ASGD):

$$w_{t+\tau+1} = w_{t+\tau} - \eta g(w_t) - \lambda \phi(g(w_t)) \cdot (w_{t+\tau} - w_t)$$

Theorem: Under mild conditions, DC-ASGD has better convergence properties than ASGD, i.e., more robust to communication delay.

说完数据切分，我们再讲讲各个工作节点之间的通信问题。大家知道，有很多流行的分布式框架，比如 MapReduce，可以实现不同工作节点之间的同步计算。但在机器学习过程中，如果不同机器之间要做同步通信，就会出现瓶颈：有的机器训练速度比较快，有的机器训练速度比较慢，而整个集群会被这个集群里最慢的机器拖垮。因为其他机器都要跟它完成同步之后，才能往前继续训练。

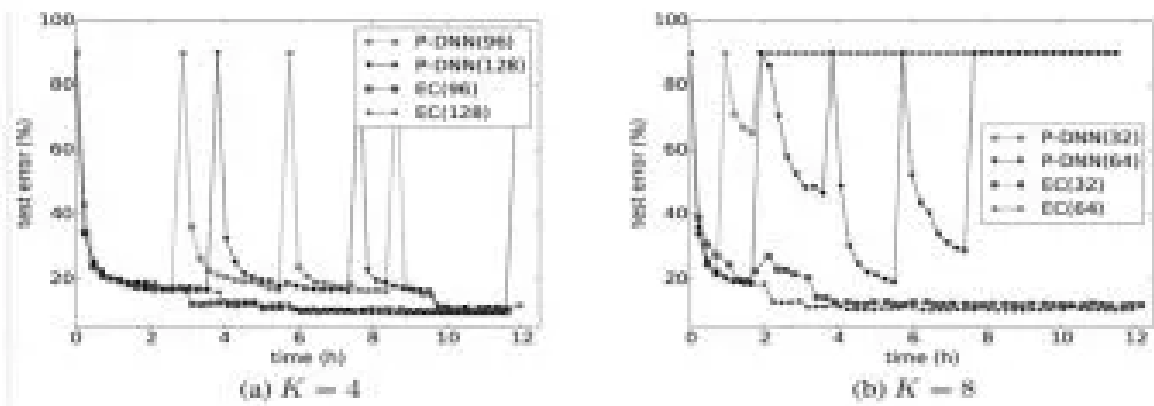
为了实现高效的分布式机器学习，人们越来越关注异步通信，从而避免整个集群被最慢的机器拖垮。在异步通信过程中，每台机器完成本地训练之后就把局部模型、局部梯度或模型更新推送到全局模型上去，并继续本地的训练过程，而不去等待其他的机器。

但是人们一直对异步通信心有余悸。因为做异步通信的时候，同样有一些机器运算比较快，有一些机器运算比较慢，当运算比较快的机器将其局部梯度或者模型更新叠加到全局模型上以后，全局模型的版本就被更新了，变成了很好的模型。但是过了一段时间，运算比较慢的机器又把陈旧的梯度或者模型更新，叠加到全局模型上，这就会把原来做得比较好的模型给毁掉。人们把这个问题称为“延迟更新”。不过在我们的研究之前，没有人定量地刻画这个延迟会带来多大的影响。

在去年 ICML 上我们发表了一篇论文，用泰勒展开式定量刻画了标准的随机梯度下降法和异步并行随机梯队下降法的差距，这个差距主要是由于延迟更新带来的。如果我们简单粗暴地使用异步 SGD，不去处理延迟更新，其实就是使用泰勒展开里零阶项作为真实的近似。既然它们之间的差距在于高阶项的缺失，如果我们有能力把这些高阶项通过某种算法补偿回来，就可以使那些看起来陈旧的延迟梯度焕发青春。这就是我们提出的带有延迟补偿的随机梯度下降法。

这件事说起来很简单，但实操起来有很大的难度。因为在梯度函数的泰勒展开中的一阶项其实对应于原损失函数的二阶项，也就是所谓的海森矩阵（Hessian Matrix）。当模型很大时，计算海森矩阵要使用的内存和计算量都会非常大，使得这个算法并不实用。在我们的论文里，提出了一个非常高效的对海森矩阵的近似。我们并不需要真正去计算非常高维的海森矩阵并存储它，只需要比较小的计算和存储代价就可以实现对海森矩阵相当精确的近似。在此基础上，我们就可以利用泰勒展开，实现对原来的延迟梯度的补偿。我们证明了有延迟补偿的异步随机梯度下降法的收敛率比普通的异步随机梯度下降法要好很多，而且各种实验也表明它的效果确实达到了我们的预期。

- Average of model parameter does not have accuracy guarantee due to the non-convexity of the problem
- Average of the model output (or ensemble of the model) has accuracy guarantee
- Model compression is needed to avoid explosion of the size of ensemble model over multiple iterations



11/23/2018

刘铁岩：推进机器学习发展 @ CSDN

19

除了异步通信以外，每个局部节点计算出一个局部模型之后，怎样聚合在一起也是一个值得思考的问题。在业界里最常用的方式是把各个不同的局部模型做简单的参数平均。但是，从理论上讲，参数平均仅在凸问题上合理的。如果大家对于凸函数的性质有一些了解，就知道如果模型是凸的，那么我们对凸模型参数进行平均后得到的模型的性能，不会比每个模型性能的平均值差。

但是当我们用这样的方式去处理深层神经网络这类严重非凸的模型时，就不再有理论保证了。我们在 2017 年这几篇论文里指出了这个理论的缺失，并指出我们不应该做模型参数的平均，而是应该做模型输出的平均，这样才能获得性能的保障，因为虽然神经网络模型是非凸的，但是常用的损失函数本身是凸的。

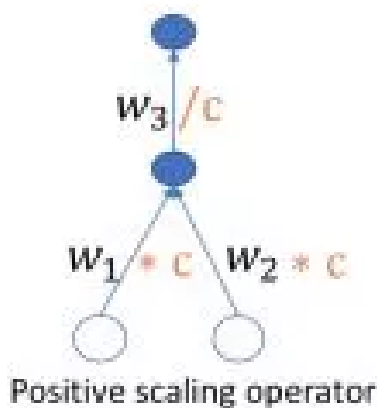
但是模型输出的平均相当于做了模型的集成，它会使模型的尺寸变大很多倍。当机器学习不断迭代时，这种模型的集成就会导致模型尺寸爆炸的现象。为了保持凸性带来的好处，同时又不会受到模型尺寸爆炸的困扰，我们需要在整个机器学习流程里不仅做模型集成，还要做有效的模型压缩。

这就是我们提出的模型集成-压缩环路。通过模型集成，我们保持了凸性带来的好处，通过模型压缩，我们避免了模型尺寸的爆炸，所以最终会取得一个非常好的折中效果。

深度学习理论探索

接下来我们讲讲如何探索深度学习的理论边界。我们都知道深度学习很高效，任意一个连续函数，只要一个足够复杂的深度神经网络都可以逼近得很好。但是这并不表示机器就真能学到好的模型。因为当目标函数的界面太复杂时，我们可能掉入局部极小值的陷阱，无法得到我们想要的最好模型。当模型太复杂时，还容易出现过拟合，在优化过程中可能做的不错，可是当你把学到的模型应用到未知的测试数据上时，效果不一定很好。因此对于深度学习的优化过程进行深入研究是很有必要的。

g-Space



- Neural networks with ReLU activations are positive scaling invariant (denoted as **G-invariant**)
 - However, the weight space of ReLU networks are **NOT G-invariant**.
 - Optimization in the weight space will suffer from gradient vanishing/exploding or spurious critical points!

G-Space: We prove that the bases in the path space (together with their values) are representation-sufficient and **G-invariant**.

11/29/2018

刘铁岩：推进机器学习发展 @ CSDN

22

这个方向上，今年我们做了一个蛮有趣的工作，叫 g-Space Deep Learning。

这个工作研究的对象是图像处理任务里常用的一大类深度神经网络，这类网络的激活函数是ReLU函数。ReLU是一个分段线性函数，在负半轴取值为0，在正半轴则是一个线性函数。ReLU Network 有一个众所周知的特点，就是正尺度不变性，但我们对于这个特点对神经网络优化影响的理解非常有限。

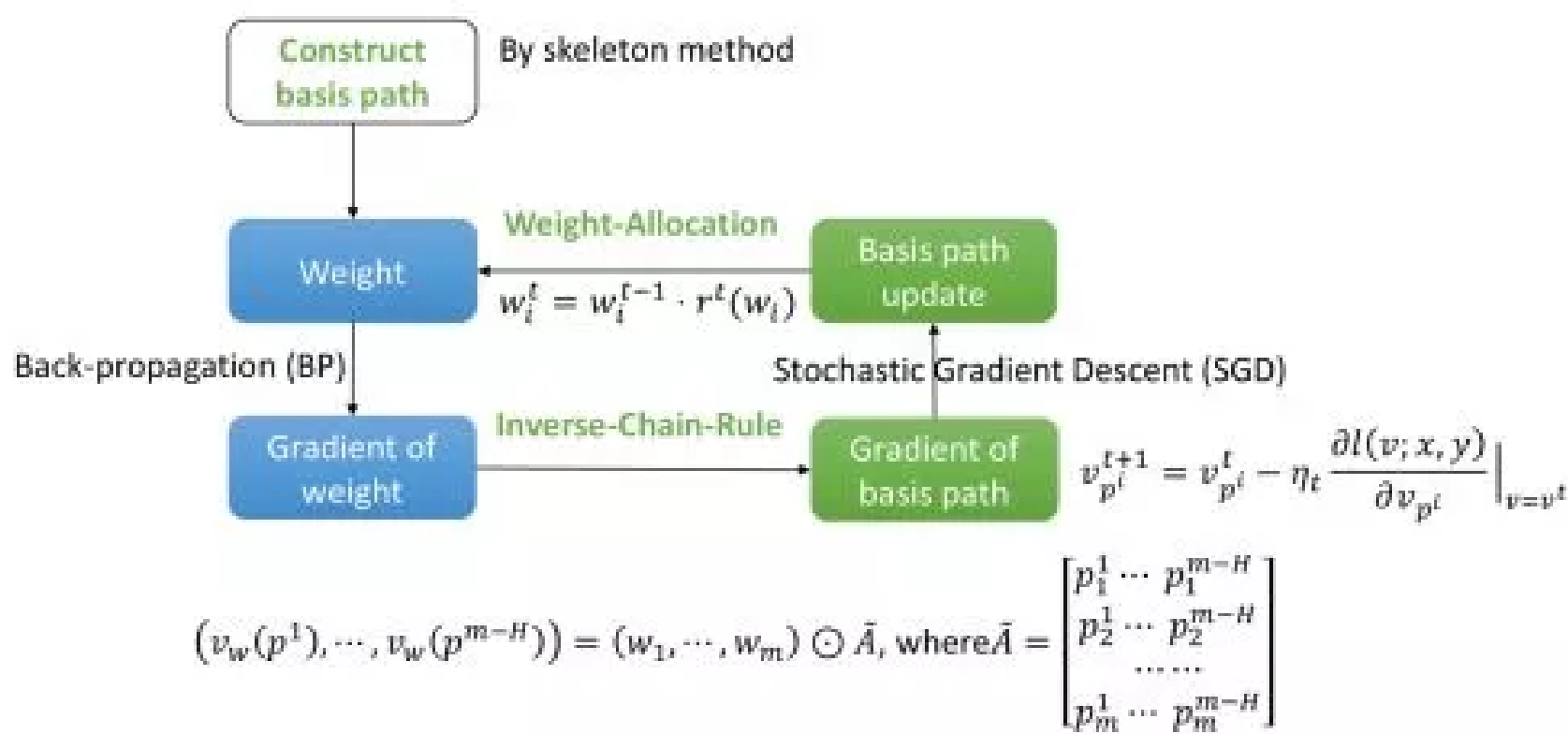
那么什么是正尺度不变性呢？我们来举个例子。这是一个神经网络的局部，假设中间隐节点的激活函数是ReLU函数。当我们把这个神经元两条输入边上面的权重都乘以一个正常数 c，同时把输出边上的权重除以同样的正常数 c，就得到一个新的神经网络，因为它的参数发生了变化。但是如果我们把整个神经网络当成一个整体的黑盒子来看待，这个函数其实没有发生任何变化，也就是无论什么样的输入，输出结果都不变。这就是正尺度不变性。

这个不变性其实很麻烦，当激活函数是 ReLU函数时，很多参数完全不一样的神经网络，其实对应了同一个函数。这说明当我们用神经网络的原始参数来表达神经网络时，参数空间是高度冗余的空间，因为不同的参数可能对应了同一个网络。这种冗余的空间是不能准确表达神经网络的。同时在这样的冗余空间里可能存在很多假的极值点，它们是由空间冗余带来的，并不是原问题真实的极值点。我们平时在神经网络优化过程中遇到的梯度消减、梯度爆炸的现象，很多都跟冗余的表达有关系。

既然参数空间冗余有这么多缺点，我们能不能解决这个问题？如果不在参数空间里做梯度下降法，而是在一个更紧致的表达空间里进行优化，是不是就可以解决这些问题呢？这个愿望听起来很美好，但实际上做起来非常困难。因为深度神经网络是一个非常复杂的函数，想对它做精确的紧致表达，需要非常强的数学基础和几何表达能力。我们组里的研究员们做了非常多的努力，经过了一年多的时间，才对这个紧致的空间做了一个完整的描述，我们称其为 g-Space。

g-Space 其实是由神经网络中一组线性无关的通路组成的，所谓通路就是从输入到输出所走过的一条不回头的通路，

G-Space Stochastic Gradient Descent



11/29/2018

刘铁岩：推进机器学习发展 @ CSDN

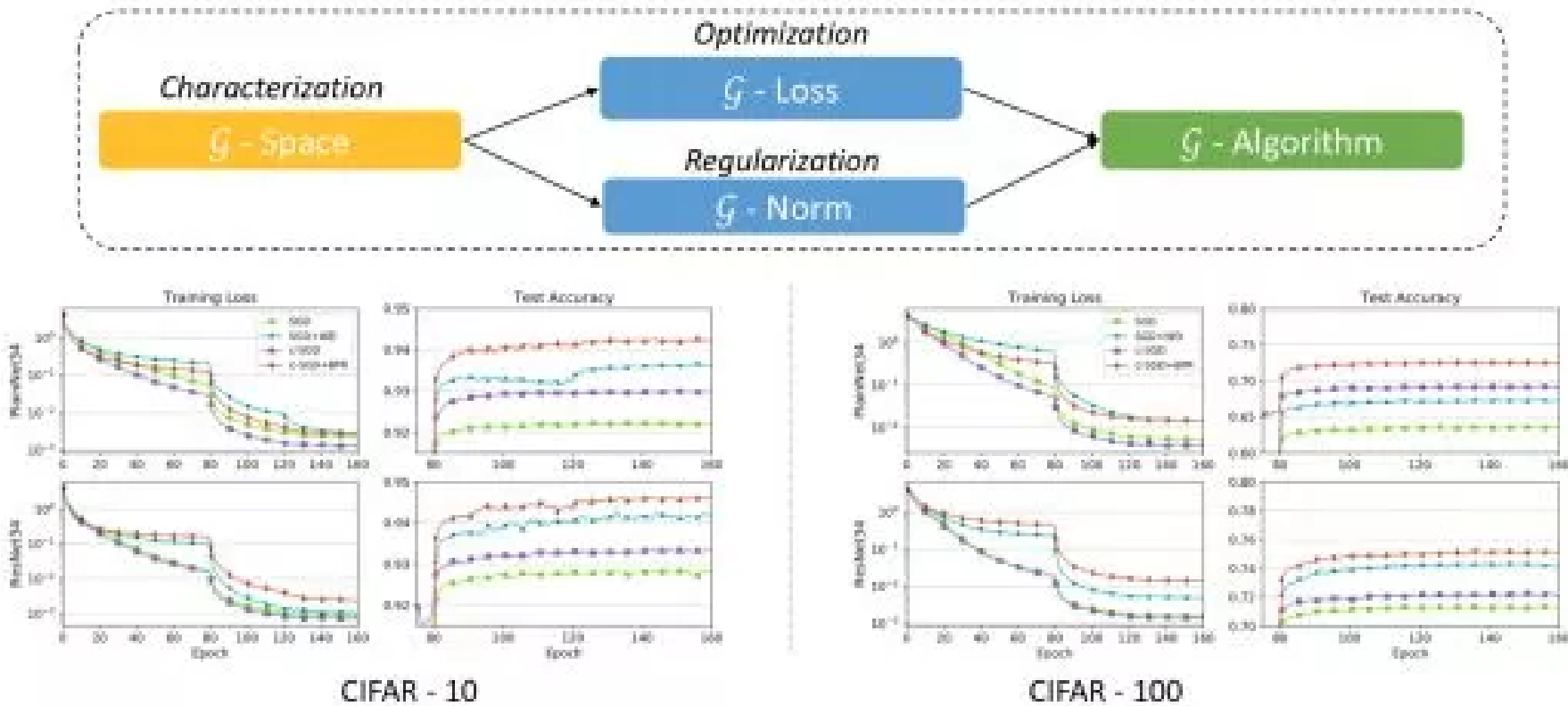
23

有了 g-Space 之后，我们就可以在其中计算梯度，同时也可以可以在 g-Space 里计算距离。有了这个距离之后，我们还可以在 g-Space 里定义一些正则项，防止神经网络过拟合。

我们的论文表明，在新的紧致空间里做梯度的计算复杂度并不高，跟在参数空间里面做典型的 BP 操作复杂度几乎是一样的。换言之，我们设计了一个巧妙的算法，它的复杂度并没有增加，但却回避了原来参数空间里的很多问题，获得了对于 ReLU Network 的紧致表达，并且计算了正确的梯度，实现了更好的模型优化。

有了这些东西之后，我们形成了一套新的深度学习优化框架。这个方法非常 general，它并没有改变目标函数，也没改变神经网络的结构，仅仅是换了一套优化方法，相当于整个机器学习工具包里面只换了底层，就可以训练出效果更好的模型来。

Experimental Results



11/29/2018

刘铁岩：推进机器学习发展 @ CSDN

24

Learning to Teach: Beyond Learning/Meta Learning

$$\omega^* = \arg \min_{\omega \in \Omega} \sum_{(x,y) \in D} L(f_{\omega}(x), y)$$

D

Learning to select the most appropriate data to train the student model at appropriate time (ICLR 2018)

Ω

Gradually adjust the model structure to better solve the target problem (NIPS 2018)

L

Gradually adjust the loss function adaptive to the current training stages (NIPS 2018)

11/29/2018

刘铁岩：推进机器学习发展 @ CSDN

26

第四个研究方向也非常有趣，我们管它叫 Learning to Teach，中文我没想到特别好的翻译，现在权且叫做“教学相长”。

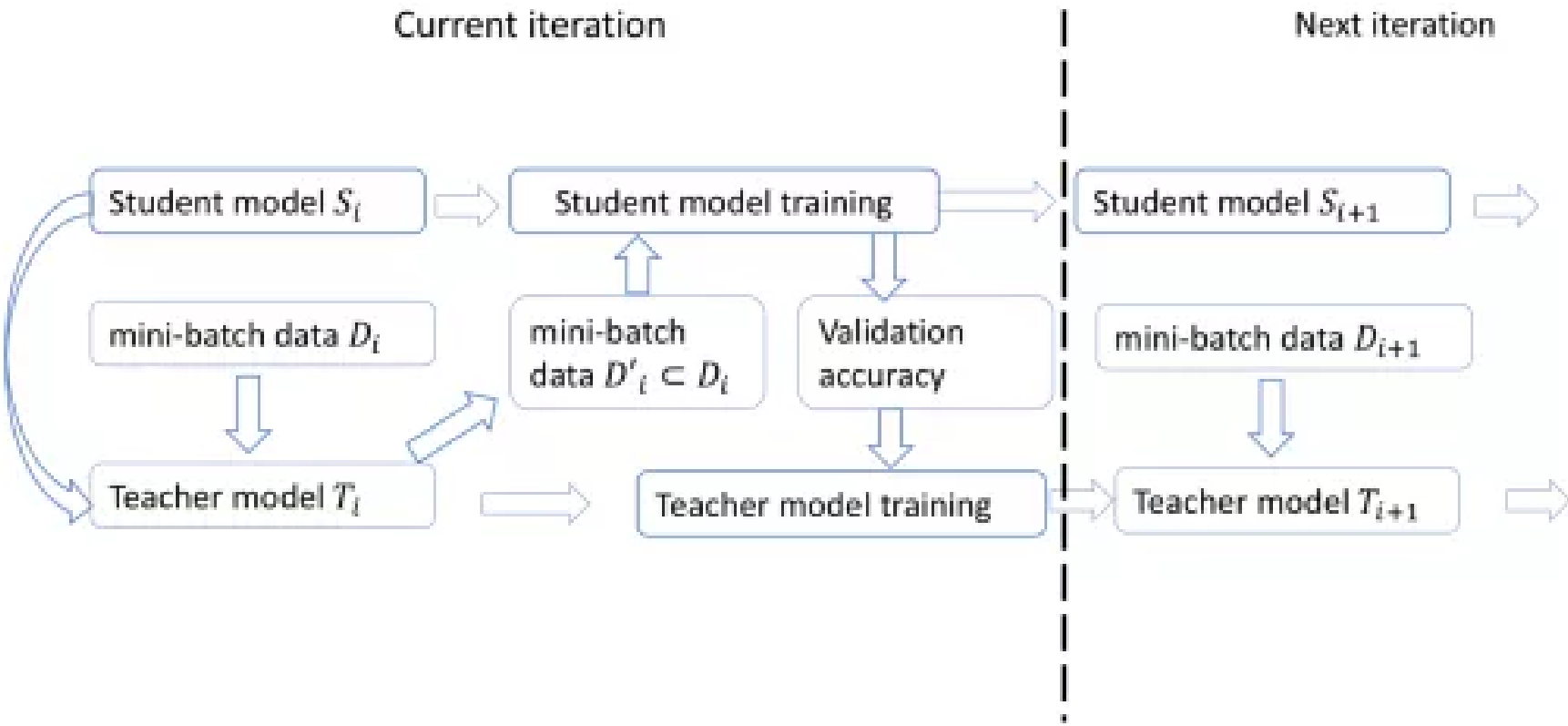
我们提出 Learning to Teach 这个研究方向，是基于对现在机器学习框架的局限性的反思。这个式子虽然看起来很简单，但它可以描述一大类的或者说绝大部分机器学习问题。这个式子是什么意思？首先 (x, y) 是训练样本，它是从训练数据集 D 里采样出来的。f(ω) 是模型，比如它可能代表了某一个神经网络。我们把 f(ω)作用在输入样本 x 上，就会得到一个对输入样本的预测。然后，我们把预测结果跟真值标签 y 进行比较，就可以定义一个损失函数 L。

现在绝大部分机器学习都是在模型空间里最小化损失函数。所以这个式子里有三个量，分别是训练数据 D，损失函数 L，还有模型空间 Ω。这三个量都是超参数，它们是人为设计好的，是不变的。绝大部分机器学习过程，是在这三样给定的情况下去做优化，找到最好的 ω，使得我们在训练数据集上能够最小化人为定义的损失函数。即便是这几年提出的 meta learning 或者 learning2learn，其实也没有跳出这个框架。因为机器学习框架本身从来就没有规定最小化过程只能用梯度下降的方法，你可以用任何方法，都超不出这个这个式子所表达的框架。

但是为什么训练数据集 D、损失函数 L 和模型参数空间 Ω 必须人为预先给定？如果不实现给定，而是在机器学习过程中动态调整，会变成什么样子？这就是所谓的 Learning to Teach。我们希望通过自动化的手段，自动调节训练数据集 D、损失函数 L 和模型参数空间 Ω，以期拓展现有机器学习的边界，帮助我们训练出更加强大的机器学习模型。

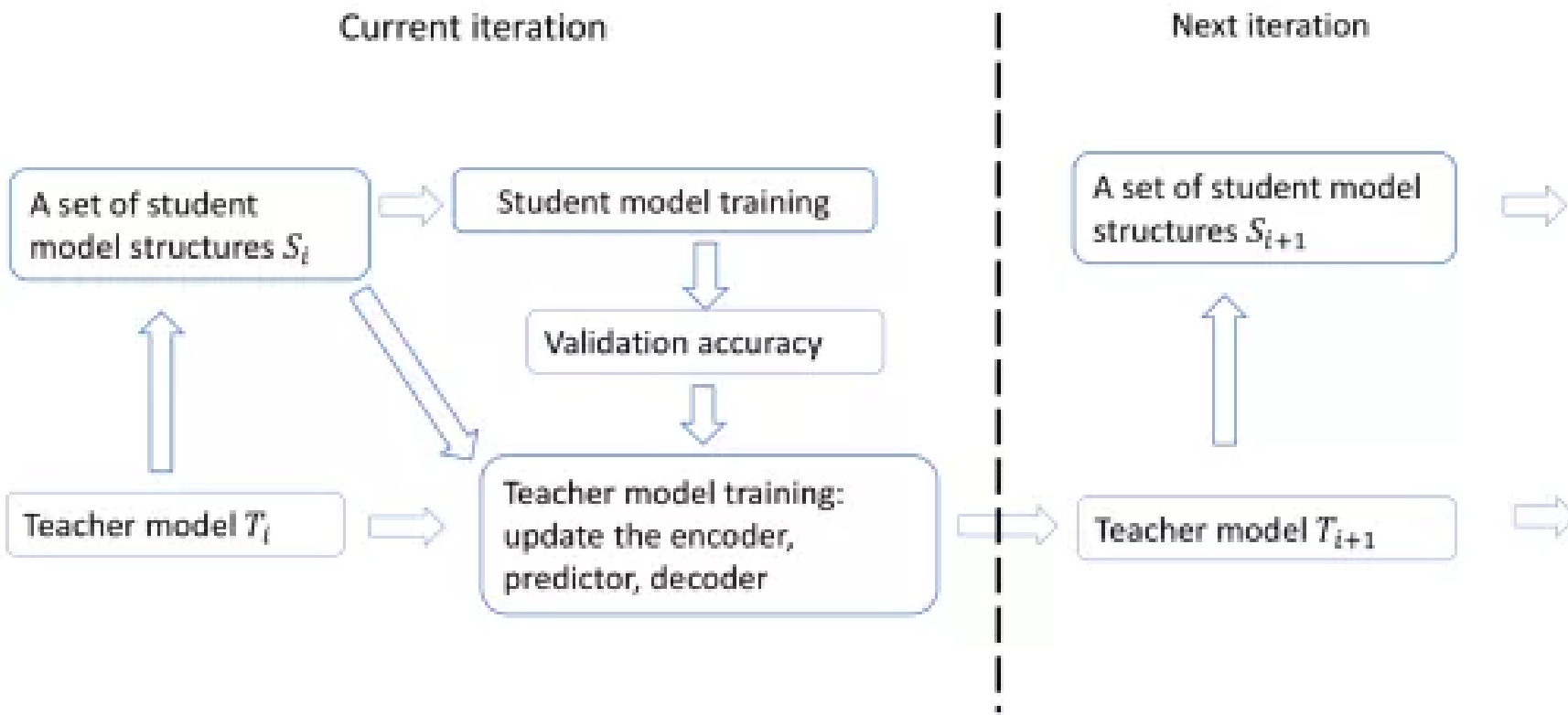
要实现这件事情其实并不简单，我们需要用全新的思路和视角。我们在今年连续发表了三篇文章，对于用自动化的方式去确定训练数据、函数空间和损失函数，做了非常系统的研究。

Data Teaching (ICLR 2018)

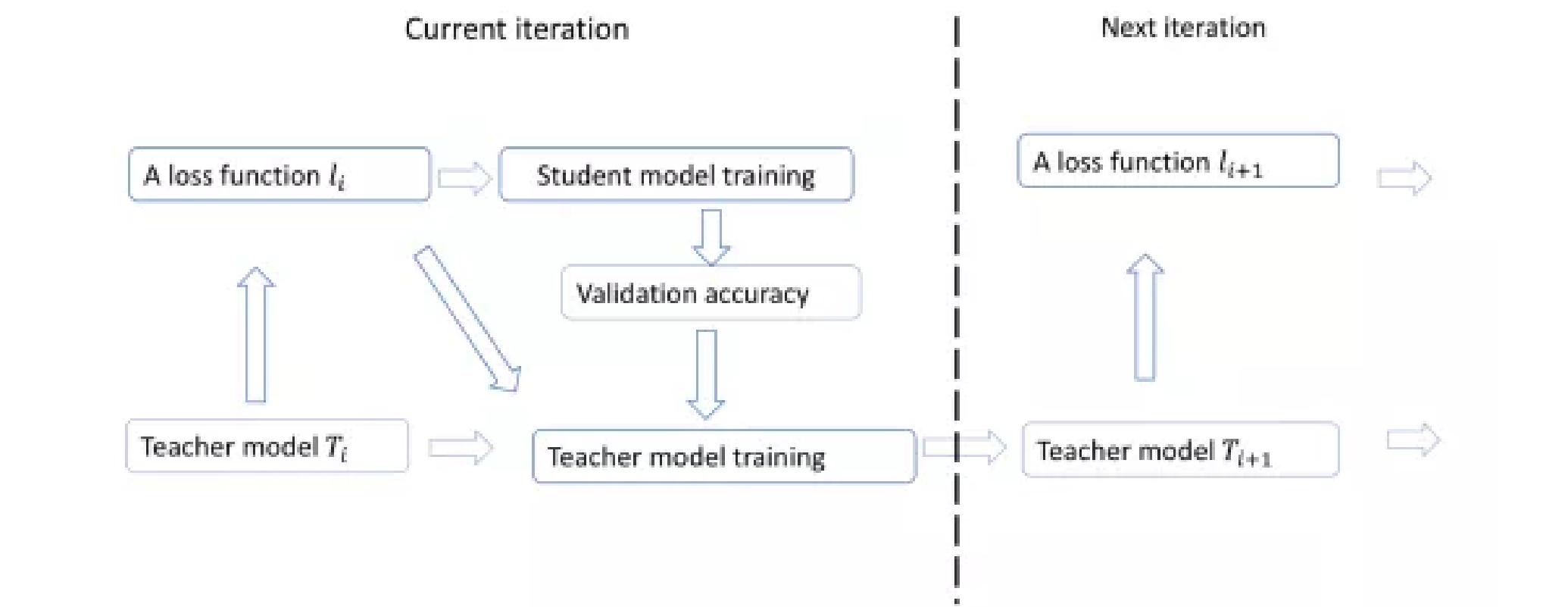


我给大家形象地描述一下我们的研究。比如我们怎么用自动化的方式去选择合适的数据？其实很简单。除了原来的机器学习模型以外，我们还有一个教学模型 teacher model。这个模型会把原来的机器学习的过程、所处的阶段、效果好坏等作为输入，输出对下一阶段训练数据的选择。这个 teacher model 会根据原来的机器学习模型的进展过程，动态选择最适合的训练数据，最大限度提高性能。同时teacher model也会把机器学习在交叉验证集上的效果作为反馈，自我学习，自我提高。

Model Teaching (NIPS 2018)



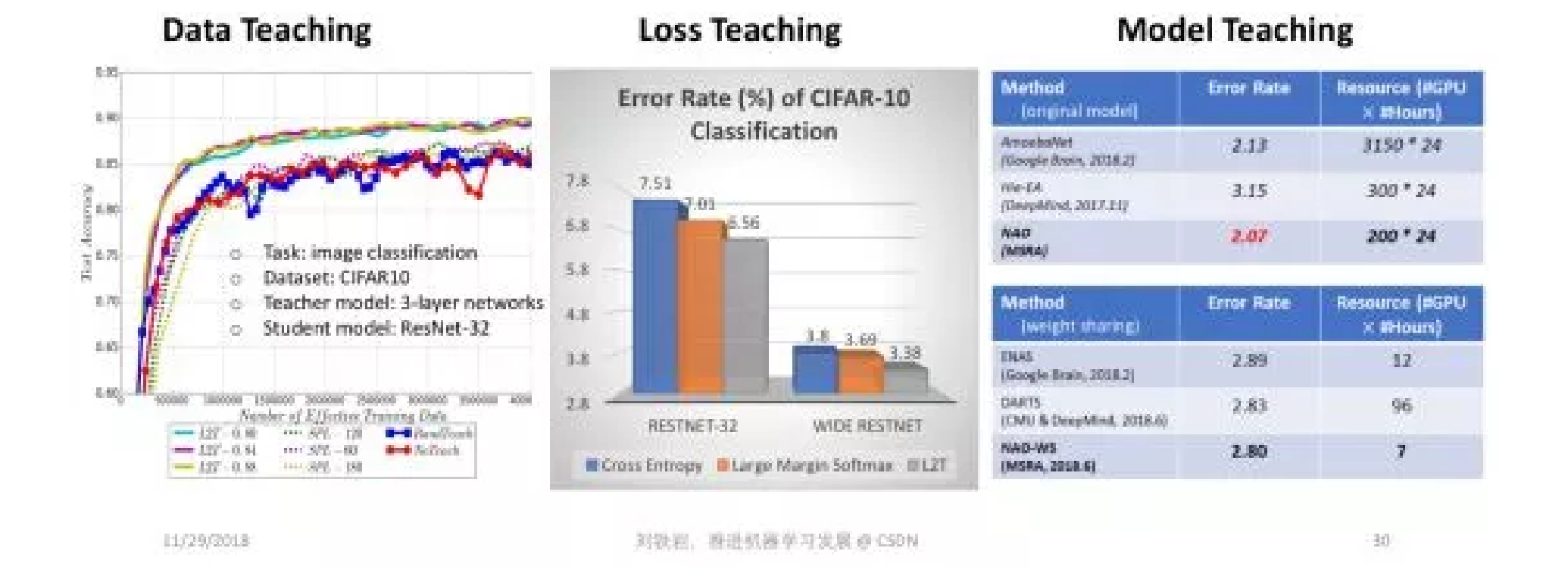
同样 model teaching 的环路中也存在一个 teacher model，它会根据原来的机器学习过程所处的阶段、训练的效果，选择合适的函数空间，让原来的机器学习扩大自己的搜索范围，这个过程也是自适应的、动态的。原来的机器学习模型我们叫 student model，和我们引入的教学模型 teacher model 之间进行互动，就可以将学习过程推向一个新的高度。



同样，teacher model也可以动态调整原来student model 所要优化的目标。比如，我们的学习目标可以从简到难，最开始的时候，一个简单的学习目标会让我们很快学到一些东西，但是这个学习目标可能和我们最终问题的评价准则相差很远。我们不断把简单平滑的目标，向着问题评价的复杂的非连续函数逼近，就会引导 student model 不断提高自己的能力，最后实现很好的学习效果。

总结一下，当我们有一个 teacher model，它可以动态地设计训练数据集、改变模型空间、调整目标函数时，就会使得原来“student model”的训练更宽泛、更有效，它的边界就会被放大。我们在三篇论文里面分别展示了很多不同数据集上的实验结果。

Experimental Results



我自己认为 Learning to Teach 非常有潜力，它扩大了传统机器学习的边界。我们的三篇论文仅仅是抛砖引玉，告诉大家这件事情可以做，但前面路还很长。

现在机器学习领域的会议越来越膨胀，有一点点不理智。每一年那么多论文，甚至都不知道该读哪些。人们在写论文、做研究的时候，有时也不知道重点该放在哪里。比如，如果整个学术界都在做 learning2learn，是不是我应该做一篇 learning2learn 的论文？大家都在用自动化的方式做 neural architecture search，我是不是也要做一篇呢？现在这种随波逐流、人云亦云的心态非常多。

我们其实应该反思：现在大家关注的热点是不是涵盖了所有值得研究的问题？有哪些重要的方向其实是被忽略的？我举个例子，比如轻量级的机器学习，比如 Learning to Teach，比如对于深度学习的一些理论探索，这些方面在如今火热的研究领域里面涉及的并不多，但这些方向其实非常重要。只有对这些方向有非常深刻的认识，我们才能真正推动机器学习的发展。希望大家能够把心思放到那些你坚信重要的研究方向上，即便当下它还不是学术界关注的主流。

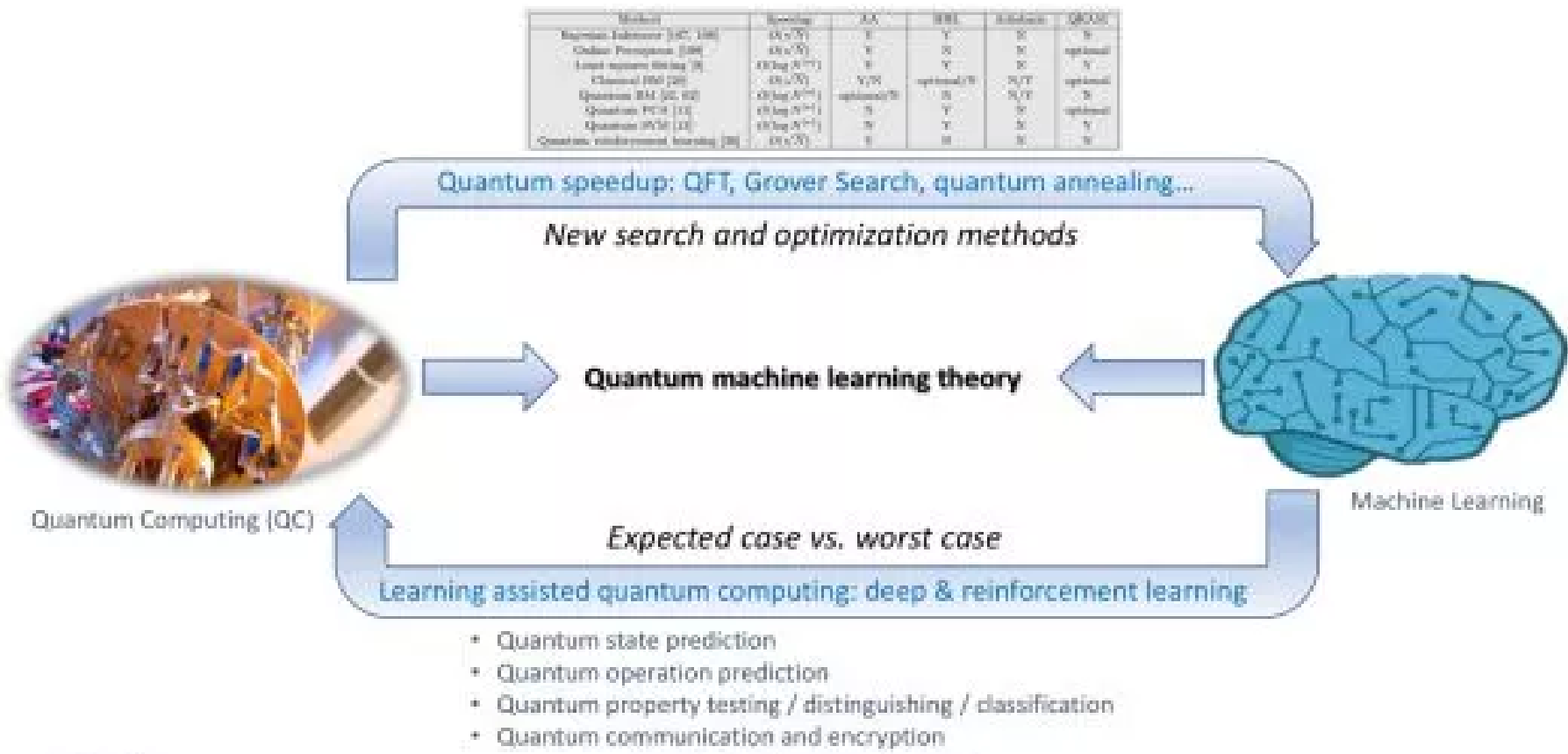
接下来我们对机器学习未来的发展做一些展望，这些展望可能有些天马行空，但是却包含了一些有意义的哲学思考，希望对大家有所帮助。

量子计算

第一个方面涉及机器学习和量子计算之间的关系。量子计算也是一个非常火的研究热点，但是当机器学习碰到量子计算，会产生什么样的火花？其实这是一个非常值得我们思考的问题。

目前学术界关注的问题之一是如何利用量子计算的算力去加速机器学习的优化过程，这就是所谓的quantum speedup。但是，这是否是故事的全部呢？大家应该想一想，反过来作为一名机器学习的学者，我们是不是有可能帮助量子计算呢？或者当机器学习和量子计算各自往前走，碰到一起的时候会迸发出怎样的新火花？

Machine Learning vs. Quantum Computing



11/29/2018

刘铁岩：推进机器学习发展 @ CSDN

33

其实存量子计算里有一些非常重要的核心问题，比如我们要去评估或者或者预测 quantum state（量子态），然后才

quantum state 做比较好的估计。但这件事情会带来负面影响，量子计算虽然很快，但是如果探测量子态耗费了大量时间来做采样，就会拖垮原来的加速效果，最后合在一起，并没有实现任何加速。

我们知道很多最坏情况下非常复杂的问题，比如 NP Complete问题，用机器学习的方法去解，其实可以在平均意义上取得非常好的效果。我们今年在ACML上获得最佳论文的工作就是用机器学习的方法来解travelling salesman问题，取得了比传统组合优化更高效的结果。沿着这个思路，我们是不是可以用机器学习帮助处理量子计算里的问题，比如 quantum state prediction，是不是根本不需要指数级的采样，就可以得到一个相当好的估计？在线学习、强化学习等都能在这方面有所帮助。

同时，量子和机器学习理论相互碰撞时，会发生一些非常有趣的现象。我们知道，量子有不确定性，这种不确定性有的时候不见得是件坏事，因为在机器学习领域，我们通常希望有不确定性，甚至有时我们还会故意在数据里加噪声，在模型训练的过程中加噪声，以期获得更好的泛化性能。


从这个意义上讲，量子计算的不确定性是不是反而可以帮助机器学习获得更好的泛化性能？如果我们把量子计算的不确定性和机器学习的泛化放在一起，形成一个统一的理论框架，是不是可以告诉我们它的 Trade-off 在哪里？是不是我们对量子态的探测就不需要那么狠？因为探测得越狠可能越容易 overfit。是不是有一个比较好的折中？其实这些都是非常有趣的问题，也值得量子计算的研究人员和机器学习的研究人员共同花很多年的时间去探索。

以简治繁

第二个方向也很有趣，它涉及到我们应该以何种方式来看待训练数据。深度学习是一个以繁治繁的过程，为了去处理非常复杂的训练数据，它使用了一个几乎更复杂的模型。但这样做真的值得吗？跟我们过去几十年甚至上百年做基础科学的思路是不是一致的？


Simple & Elegant Laws vs. Complex Model

photon



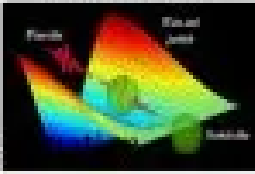
Maxwell equations

small particles



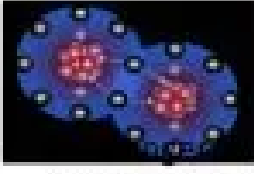
Schrödinger equation

any matter



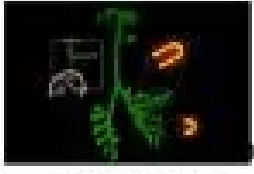
Einstein's field equations

chemical bonds




Molecular structure

morphogen



Morphogen pattern

economics



Economic model

"It turns out that almost all the traditional mathematical models that have been used in physics and other areas of science are ultimately based on partial differential equations."

— Stephen Wolfram

Learning Laws vs. Fitting Data

以简治繁

以繁治繁

- It was shown that natural laws can be automatically discovered by evolutionary algorithms (Science 2009)
- How about automatically learning simple & elegant laws behind complicated data we have?
 - Data is just the phenomenon
 - Laws that govern the generation of the data is the essence
 - New machine learning models are needed, such as dynamic systems and partial equations

11/29/2018

刘铁岩：推进机器学习发展 @ CSDN

14

在物理、化学、生物这些领域，人们追求的是世界简单而美的规律。不管是量子物理，还是化学键，甚至经济学、遗传学 很多复杂的现象背后其实都是一个二阶偏微分方程 比如薛定谔方程 比如麦克斯韦方程组 等等 这些方程

都告诉我们，看起来很复杂的世界，其实背后的数学模型都是简单而美的。这些以简治繁的思路，跟深度学习是大相径庭的。

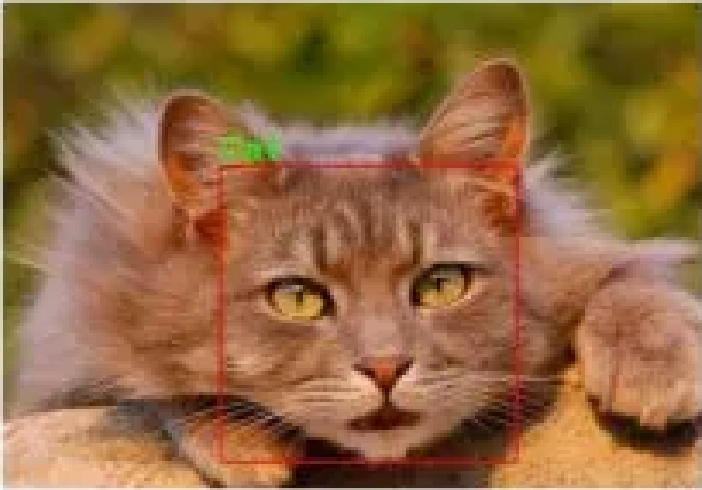
机器学习的学者也要思考一下，以繁治繁的深度学习真的是对的吗？我们把数据看成上帝，用那么复杂的模型去拟合它，这样的思路真的对吗？是不是有一点舍本逐末了？以前的这种以简治繁的思路，从来都不认为数据是上帝，他们认为背后的规律是上帝，数据只是一个表象。

我们要学的是生成数据的规律，而不是数据本身，这个方向其实非常值得大家去思考。要想沿着这个方向做很好的研究，我们需要机器学习的学者扩大自己的知识面，更多地去了解动态系统或者是偏微分方程等，以及传统科学里的各种数学工具，而不是简单地使用一个非线性的模型去做数据拟合。

Improvisational Learning



Pattern Cognition vs. Prediction

Pattern Recognition



Predictive Learning

- Build world model + predict the future
 - Infer the state of the world from partial information
 - Infer the future from the past and present
 - Infer past events from the present state
- Filling in the visual field at the retinal blind spot
- Filling in occluded images
- Filling in missing segments in text, missing words in speech
- Predicting the consequences of our actions
- Predicting the sequence of actions leading to a result
- Predicting any part of the past, present or future percepts from whatever information is available
- That's what predictive learning is
- But really, that's what many people mean by unsupervised learning



11/29/2018

刘铁岩，推进机器学习发展 @ CSDN

35

第三个方向关乎的是我们人类到底是如何学习的。到今天为止，深度学习在很多领域的成功，其实都是做模式识别。模式识别听起来很神奇，其实是很简单的一件事情。几乎所有的动物都会模式识别。人之所以有高的智能，并不是因为我们会做模式识别，而是因为我们有知识，有常识。基于这个理念，Yann LeCun 一个新的研究方向叫 Predictive Learning（预测学习）。它的思想是什么？就是即便我们没有看到事物的全貌，因为我们有常识，有知识，我们仍然可以做一定程度的预测，并且基于这个预测去做决策。这件事情已经比传统的模式识别高明很多，它会涉及到人利用知识和常识去做预测的问题。

但是，反过来想一想，我们的世界真的是可以预测的吗？可能一些平凡的规律是可以预测的，但是我们每个人都可以体会到，我们的生活、我们的生命、我们的世界大部分都是不可预测的。所以这句名言很好，The only thing predictable about life is its unpredictability（人生中唯一能预测的就是其不可预测性）。

Prediction vs. Improvisation

Predictive Learning

- Build world model + predict the future
 - Infer the state of the world from partial information
 - Infer the future from the past and present
 - Infer past events from the present state
 - Filling in the visual field at the retinal blind spot
 - Filling in occluded images
 - Filling in missing segments in text, missing words in speech.
 - Predicting the consequences of our actions
 - Predicting the sequence of actions leading to a result
 - Predicting any part of the past, present or future percepts from whatever information is available.
 - That's what predictive learning is
 - But really, that's what many people mean by unsupervised learning



Improvisational Learning

- Challenge: Is the world predictable?

"The only thing predictable about life is its unpredictability." — Remy in Ratatouille
- Improvisational learning:

The world is full of exceptions and one needs to improvise to survive when unexpected things happen.



11/29/2018

刘铁岩，推进机器学习发展 @ CSDN

35

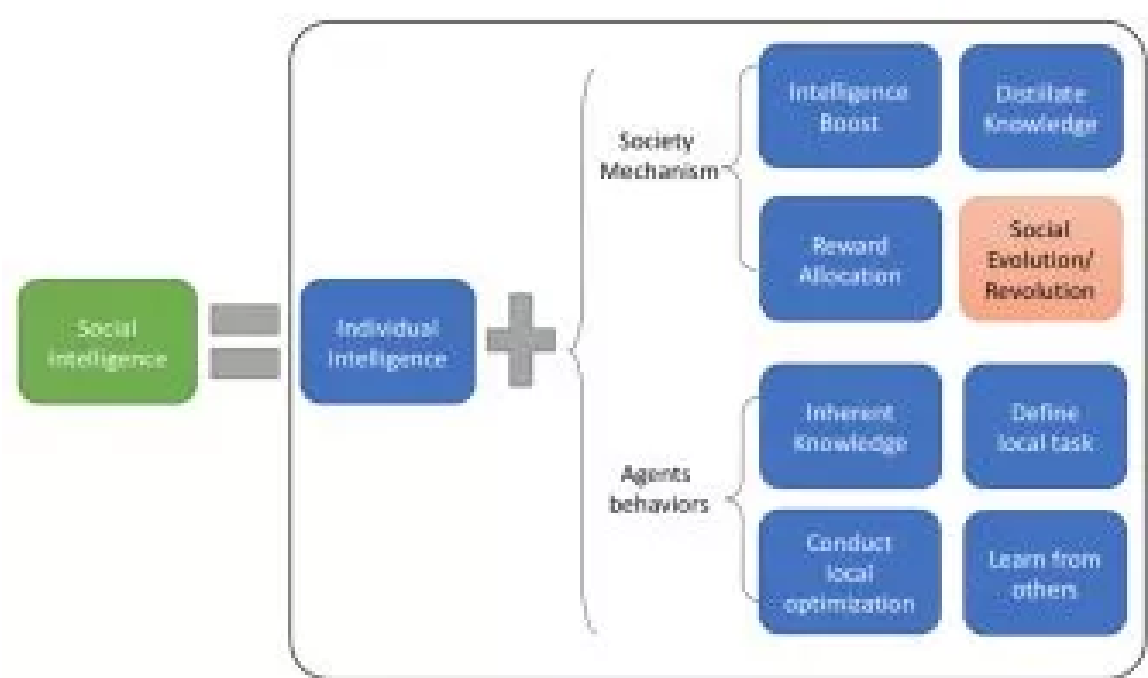
我们既然活在一个不可预测的世界里，那么我们到底是怎样从这个世界里学习，并且越来越强大？以下只是一家之言，我们猜测人类其实在做一件事情，叫 Improvisation，什么意思？就是我们每个人其实是为了生存在跟这个世界抗争。我们每天从世界里面学习的东西，都是为了应付将来未知的异常。当一件不幸的事情发生的时候，我们如何才能生存下来？其实是因为我们对这个世界有足够的了解，于是会利用已有的知识，即兴制定出一个方案，让我们规避风险，走过这个坎。

我们希望在我们的眼里，世界的熵在降低。我们对它了解越多，它在我们的眼里的熵越低。同时，我们希望当环境发生变化时，比如意外发生时，我们有能力即兴地去处理。这张PPT 里面描述的即兴学习框架就是我们在跟环境互动，以及在做各种思想实验，通过无监督的方式自我学习应对未知异常的能力。

从这个意义上讲，这个过程其实跟 Predictive Learning 不一样，跟强化学习也不一样，因为它没有既定的学习规律和学习目标，并且它是跟环境做交互，希望能够处理未来的未知环境。这其实就跟我们每个人积累一身本事一样，为的就是养兵千日用兵一时。当某件事情发生时，我怎么能够把一身的本事使出来，活下去。这个过程能不能用数学的语言描述？ Improvisational Learning 能不能变成一个新的机器学习研究方向？非常值得我们思考。

群体智慧

Social Intelligence vs. Individual Intelligence



- Social Competition
 - Multiple layers of sub-societies with different mechanisms.
 - Local agents are coopetiting (collaborating and competing) with each other, given the structure of sub-societies.
- Society Evolution/Revolution:
 - Diversity and E-E tradeoff play an important role in evolution process
 - If a sub-society always has low performance, it will be replaced by another sub-society and its mechanism.
 - With the competition among sub-societies, the whole society is evolving towards higher performance.

最后一个展望涉及到一个更哲学的思辨：人类的智能之所以这么高，到底是因为我们个体非常强大，还是因为我们群体非常强大？今天绝大部分的人工智能研究，包括深度学习，其实都在模仿人类个体的大脑，希望学会人类个体的学习能力。可是扪心自问，人类个体的学习能力真的比大猩猩等人类近亲高几个数量级吗？答案显然不是，但是今天人类文明发展的程度，跟猴子、跟大猩猩他们所处社区的文明的发展程度相比却有天壤之别。

所以我们坚信人类除了个体聪明以外，还有一些更加特殊的东西，那就是社会结构和社会机制，使得我们的智能突飞猛进。比如文字的产生，书籍的产生，它变成了知识的载体，使得某一个人获得的对世界的认知，可以迅速传播给全世界其他人，这个社会机制非常重要，会加速我们的进化。

再者，社会分工不同会使得每个人只要优化自己的目标，让自己变强大就可以了。各个领域里有各自的大师，而这些大师的互补作用，使得我们社会蓬勃发展。

所以社会的多样性，社会竞争、进化、革命、革新，这些可能都是人类有今天这种高智能的原因。而这些东西在今天的机器学习领域，鲜有人去做非常好的建模。我们坚信只有对这些事情做了非常深入的研究，我们才能真正了解了人的智能，真的了解了机器学习，把我们的研究推向新的高度。



微软研究院AI头条

专注科研19年，盛产黑科技

相关数据

- 刘铁岩
 - 人物
- 深度学习
 - 技术
- 半监督学习
 - 技术

展开全部数据

登录后评论



暂无评论~



关于我们 寻求报道 商务合作 加入我们 服务条款

©2019 机器之心（北京）科技有限公司
京 ICP 备 12027496

加入合作计划

提交应用案例 | 寻求技术提供方 | 垂直领域合作



联系电话：+86 010-57150141
联系邮箱：contact@jiqizhixin.com

全球人工智能信息服务

友情链接：[Synced Global](#)

[机器之心 Medium 博客](#)

[PaperWeekly](#) [网易智能](#)
[动脉网](#) [硬蛋网](#) [达观数据](#)
[品途商业评论](#) [艾耕科技](#)

