从 ICLR 2019 一览小样本学习最新进展！

笔记本：	ML机器学习		
创建时间：	2019/10/14 14:23	**更新时间：**	2019/10/14 14:26
URL：	https://baijiahao.baidu.com/s?id=1635849659418082412&wfr=spider&for=pc		

从 ICLR 2019 一览小样本学习最新进展！

雷锋网
发布时间：06-0916:06深圳英鹏信息技术股份有限公司
.

雷锋网 AI 科技评论按：通常而言，深度学习是典型的数据驱动型技术，面对数据有限的情况，传统的深度学习技术的性能往往不尽如人意。在本届 ICLR 上，许多研究者们利用元学习、迁移学习等技术对小样本学习问题进行了探究，发表了多篇高质量论文，可谓百家争鸣！深度学习工程师 Isaac Godfried 在 Medium 上发表了一篇文章，基于今年 ICLR 上关于小型数据集深度学习研究的论文，探讨了目前小样本学习的最新进展。雷锋网 AI 科技评论编译如下。

今年的国际表征学习大会（ICLR）于 2019 年 5 月 6 日如期开幕。按照我此前的计划，我会深入研究本届会议发表的一些有趣的 ICLR 论文。其中大多数的论文都与我个人感兴趣的研究领域相关（无监督学习、元学习、注意力机制、自然语言处理），但是我只会选出一些高质量的、并且在其各自的领域有所影响的精品论文进行分析，并更新系列博文。该系列博文的第一篇将介绍在小型数据集上的深度学习研究；第二篇将讨论在自然语言处理和其它类型的序列化数据上取得突破性进展的论文；而第三篇则将分析各类其它的、我认为十分有趣的论文。

迁移学习、元学习和无监督学习

训练数据有限的问题对各行各业都有着广泛的影响，包括医疗卫生、农业、汽车、零售、娱乐，等等。在另外一些情况下，我们拥有大量的数据，但是它们却未被标注。由于收集和标注数据的时间/成本很大，这个问题往往会成为将深度学习技术整合到目标任务中的障碍。

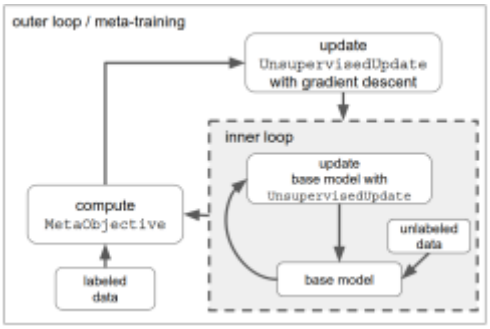
《学习无监督学习规则》

Learning unsupervised learning rules

论文下载地址：https://openreview.net/forum?id=HkNDsiC9KQ

该论文同时建立在元学习和无监督学习（这里指 Metz 等人的工作）的概念之上。具体而言，该论文提出利用元学习以一种无监督学习的方式学习下游任务的有效表征。该论文重点关注「半监督学习」分类问题，但是它之所以有趣是因为：至少在理论上，这种学习规则「可以被优化，从而为任意后续任务生成表征」。这一点十分有用，因为在针对表征的无监督学习的工作中，作者都定义了一个明确的训练算法或损失函数。而这里的模型会「学习创建由元目标确定的有用的表征的算法」。这个自定义的规则往往需要经过大量的实验以及领域知识才能得出，因此并不能很轻易地适用于新的领域。对自编码器的使用就是其中的一个例子，它试着通过先进行编码、再解码出一个与原始数据相同的输出来学习表征。自编码器往往需要一个明确指定的损失函数。

为了理解该方法究竟是如何工作的，我们不妨回想一下：在元学习中，我们通常有一个内层循环和外层循环。在内层循环中，模型会作用于一个具体的任务，例如：在图像分类问题中，这样的任务可能是识别出猫和狗。通常而言，内层循环会在一定数量 n（一般来说，n 在 1 到 10 之间）个示例上运行。然后，外层循环会使用某些内层循环得到的参数（权重本身、累计损失或其它参数）来执行一次「元更新」。这种「元更新」的具体情况随着模型的变化而变化，但是它们通常会遵循如下所示的方法：



元学习过程一览

考虑到这一点，他们的模型架构本质上是通过元学习学到某种在创建表征之后更新内层模型的方法。在创建了某种表征之后，该规则有效地在更新内层模型的过程中替代了随机梯度下降方法。此外，不同于权重本身通过 MAML 方法或注意力模型的权重通过 SNAIL 更新的情况，这种无监督的更新规则是在循环的最后进行更新的。这意味着这种无监督学习规则不仅仅可以被应用于类似的任务，还可以被用于全新的任务、新的基础模型，甚至是新模态的数据（例如从图像数据到文本数据）。

首先，作者通过展现以前方法存在的问题来评价他们的模型的实验结果。例如，一个变分自编码器（VAE）会存在目标函数（即损失）不匹配的问题，随着时间的推移，这会导致模型的性能不佳。尽管可以使用原型网络迁移特征，但如果不同任务的特征维度不同，这种方法就会崩溃。相反，Metz 等人的方法学到了一种在「小样本」分类任务中具备更好的泛化性能的更新规则。他们还展示了训练时的元更新，即使该网络仅仅在图片分类任务上进行训练，它仍然可以泛化到提升文本分类的性能（但同时他们也发现：如果元函数在图片分类任务上训练了太久，会产生明显的性能下降，这是由于该元函数在图片任务上发生了过拟合）。

总而言之，这是一篇非常棒的论文，也是在无监督技术上取得的巨大进步。即使它没有取得任何目前最先进的实验结果，但是它完全可以被应用于许多数据稀疏的领域。本论文官方版本的代码可以通过该链接获取：

https://github.com/tensorflow/models/tree/master/research/learning_unsupervised_learning

通过元学习实现的无监督学习

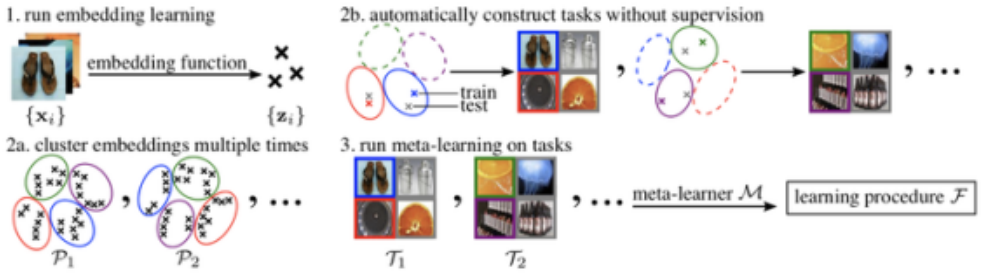


Figure 1: Illustration of the proposed unsupervised meta-learning procedure. Embeddings of raw observations are clustered with k-means to construct partitions, which give rise to classification tasks. Each task involves distinguishing between examples from $N = 2$ clusters, with $K_{\text{m-tr}} = 1$ example from each cluster being a training input. The meta-learner's aim is to produce a learning procedure that successfully solves these tasks.

有趣的是，在今年的 ICLR 上发表了两篇同时提出将元学习和无监督学习结合起来的论文（尽管两篇文章实现的方法完全不同）。在本文中，作者使用无监督学习为元学习划分数据及，而并非使用元学习来学会无监督学习的规则。

本文是我最喜爱的论文之一，因为它开启了无需进行显式任务描述的元学习的大门。元学习存在的某些问题在于：元学习往往需要定义得非常好的任务集合。这就将元学习的适用范畴限制在研究者拥有非常大的已标注元数据集（往往被划分为不同的子数据集）的前提下。本文的方法提出自动地将数据集划分为不同的子集。本文作者发现，即便使用简单的无监督聚类算法（例如 K-means 算法），元学习器仍然能够从这些任务中进行学习，并且在后续人为标记过的任务上比直接利用这些嵌入进行学习的方法（例如在无监督学习后，紧接着进行监督分类的情况）的性能更好。他们使用的两种元学习技术为「ProtoNets」和「MAML」。本文介绍了一种有趣的半监督学习范式，在这里，我们首先进行无监督的预训练，然后进行监督学习。在本例中，「带监督的」部分会进行「小样本学习」（few-shot learning）。

作者在 4 个数据集上（MNIST，Omniglot，miniImageNet，以及 CelebA）将他们的方法和无监督学习方法进行了对比。最终，他们发现，他们的方法比所有其它的「无监督+监督学习」方法（包括聚类匹配，多层知机（MLP），线性分类，以及K最近邻）的性能都要好得多。总而言之，

本文朝着「让元学习更容易被应用于各种不同类型的问题」的方向迈出了一大步，而不是让元学习仅仅适用于那些被良好定义的任务切片。

《带有潜在嵌入优化（LEO） 的元学习》

Meta-Learning with Latent Embedding Optimization (LEO)

论文下载地址：https://openreview.net/forum?id=BJgklhAcK7

Algorithm 1 Latent Embedding Optimization

Require: Training meta-set $\mathcal{S}^{tr} \in \mathcal{T}$
Require: Learning rates α, η
Randomly initialize ϕ_e, ϕ_r, ϕ_d
Let $\phi = \{\phi_e, \phi_r, \phi_d, \alpha\}$
while not converged **do**
  for number of tasks in batch **do**
    Sample task instance $\mathcal{T}_i \sim \mathcal{S}^{tr}$
    Let $\left(\mathcal{D}^{tr}, \mathcal{D}^{val}\right) = \mathcal{T}_i$
    Encode $\mathcal{D}^{tr}$ to $\mathbf{z}$ using g_{ϕ_e} and g_{ϕ_r}
    Decode $\mathbf{z}$ to initial params θ_i using g_{ϕ_d}
    Initialize $\mathbf{z}' = \mathbf{z}$, $\theta_i' = \theta_i$
    for number of adaptation steps **do**
      Compute training loss $\mathcal{L}_{\mathcal{T}_i}^{tr}(f_{\theta_i'})$
      Perform gradient step w.r.t. $\mathbf{z}'$:
      $\mathbf{z}' \leftarrow \mathbf{z}' - \alpha \nabla_{\mathbf{z}'} \mathcal{L}_{\mathcal{T}_i}^{tr}(f_{\theta_i'})$
      Decode $\mathbf{z}'$ to obtain θ_i' using g_{ϕ_d}
    end for
    Compute validation loss $\mathcal{L}_{\mathcal{T}_i}^{val}(f_{\theta_i'})$
  end for
  Perform gradient step w.r.t ϕ:
  $\phi \leftarrow \phi - \eta \nabla_{\phi} \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}^{val}(f_{\theta_i'})$
end while

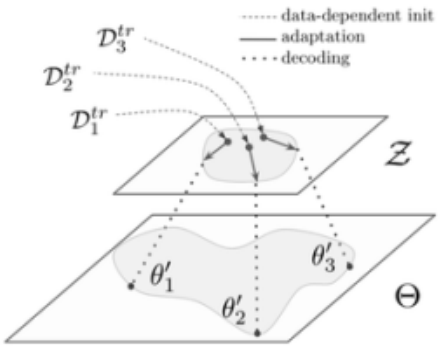


Figure 1: High-level intuition for LEO. While MAML operates directly in a high dimensional parameter space Θ, LEO performs meta-learning within a low-dimensional latent space $\mathcal{Z}$, from which the parameters are generated.

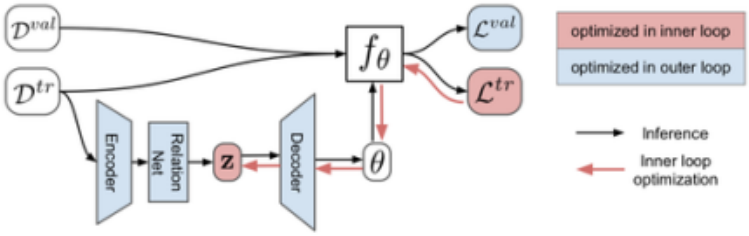


Figure 2: Overview of the architecture of LEO.

本文旨在将基于梯度的元学习和一个潜在的表征网络结合起来。LEO 的操作分为两步：首先，它会学习一个模型参数的低维嵌入；接着它会在模型的低维嵌入空间上执行元学习。具体而言，首先将会为模型给出一个任务 T 以及会被传给编码器的输入。编码器会生成一个潜在编码，该编码随后会被解码成一组参数。该编码器还带有一个关系网络，它有助于将编码变得具有上下文依赖。接着，这些参数会在内层循环中被优化，而编码器、解码器和关系网络则会在外层循环中被优化。作者指出，他们的工作的主要贡献是说明了低维嵌入空间中的元学习会比在类似于 MAML 中使用的那样的高维空间中的元学习的性能好得多。LEO 在「tieredImageNet」和「miniImageNet」数据集上都取得了很好的实验结果（包括在 5 way 1-shot 对比基准测试上实现的准确率为 61%，令人印象深刻，同时还在 5 way 5-shot 任务上取得了77% 的准确率）。和许多其它的论文一样，本文仅仅在图像数据集上进行了测试，因此尚不清楚该模型在其它类型数据上的泛化能力。

《跨程序的迁移学习》

Transferring Learning Across Processes

论文下载地址：https://openreview.net/forum?id=HygBZnRctX

由于本文作者已经在 Medium 上发表了一篇详细介绍其模型工作原理的博文（文章查看地址：https://medium.com/@flnr/transferring-knowledge-across-learning-processes-f6f63e9e6f46），我在这里就不过多赘述技术层面的细节了。相较于其它大量关于元学习的论文，该论文有下面几点值得强调的亮点：首先，本文的模型同时在小样本学习（few-shot learning）和数据规模更大的场景下进行了测试评估。这一点是很重要的，因为元学习算法往往并没有考虑在有更多的数据示例（但数据规模仍然太小，以致于无法从头开始训练模型）的情况下元优化的工作情况。本文还研究了一些尚未被探索的领域。具体而言，本文研究了往往未被探索的「远程迁移」领域，即在明显不同的任务之间实现具有积极效果的知识迁移。

《学习深度多维聚类变分自编码器中的潜在上层结构》

Learning Latent Superstructures in Variational Autoencoders for Deep Multidimensional Clustering

论文下载地址：https://openreview.net/forum?id=SJgNwi09Km

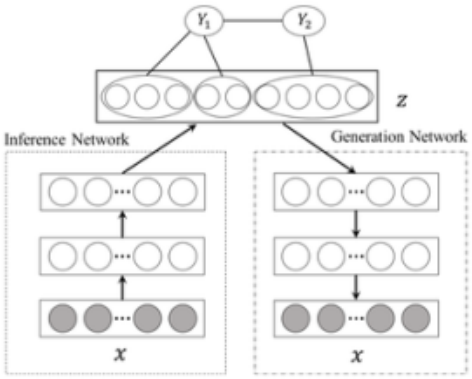


Figure 1: Latent Tree Variational Autoencoder

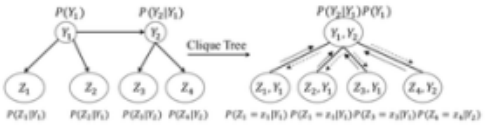


Figure 2: Inference and gradient through message passing. Solid-arrows denote collecting message, and dashed-arrows denote distributing message.

本文讨论了使用一种新型的用于更好地对高维数据进行聚类的变分自编码器（VAE）。在无监督学习中，将数据项聚类到不同的中是一个重要的预处理步骤。本文作者指出，许多种类的数据可以基于其属性的许多不同部分被进行聚类。作者指出「LTVAE 会生成多个数据划分，每个划分都会由一个上层的潜变量导出。」

「LT-VAE 不仅仅会学习每个聚类的位置来更好地表征数据，它还会学习这些簇的编号和底层树形架构的层次结构。这是通过一个三步的学习算法实现的：第一步，训练一个传统的『编码器-解码器』神经网络，从而提升它们对数据的拟合效果。第二步，一种类似于最大期望算法（EM）的优化过程，从而更好地拟合学到的后验概率的潜在先验的参数。第三步，调整潜在先验的结构从而提升其 BIC 得分[3]，这样做在对潜在后验的良好拟合以及潜在先验的参数数量（即复杂度）之间取得了平衡。」

本文提出的方法的主要优点在于，它提高了聚类的可解释性（即使从对数似然方面来说，它整体的效果并没有那么好）。此外，针对特定的因素进行聚类使其在许多真实世界的应用中变得十分具有吸引力。尽管本文与许多其它的文章有所不同，并且没有显式地研究小样本学习问题，我认为将这种聚类方法与小样本方法相结合可能会很有用。例如，它可能可以在「基于元学习环境的无监督学习」问题中被用于任务划分。

《基于元学习的深度在线学习》

Deep online learning via meta-learning

论文下载地址：https://sites.google.com/berkeley.edu/onlineviameta

本文聚焦于使用元学习和一个「Chinese Restaurant Proces」，在强化学习模型在线运行时（即在生产过程中）快速地更新它们。该工作受启发于这一事实：人类常常面临之前从未（真正地）经历过的新状况；然而我们可以利用过去的经验，并将其与我们从新的经历中获得的反馈相结合，从而迅速适应新的状况。

本文提出的方法首次使用了 MAML 来初步训练模型。在 MAML 给出有效的先验后会使用在线学习算法。该在线学习算法使用了「中餐馆程序」来生成新的带有合适的初始化设置的新模型或选择一个已经存在的模型。接着，作者会基于在线学习的结果，使用随机梯度下降（SGD）算法更新模型参数。作者将本文提出的方法命名为「用于在线学习的元学习」（或简称 MoLE）。

作者在一些强化学习环境中测试评估了他们提出的方法。第一个环境是穿越不同难度的斜坡的仿真猎豹。第二个环境是一个腿部有残缺的六足履带机器人。实验结果表明，MoLE 比基于模型的强化学习、使用元学习的k-shot 自适应技术、以及使用元学习的连续梯度步技术的性能要好（尽管有趣的是，它仅仅略微优于使用元学习的梯度步）。

《学习通过最大化迁移和最小化干扰进行不会遗忘的学习》

Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference

论文下载地址：https://arxiv.org/pdf/1810.11910.pdf

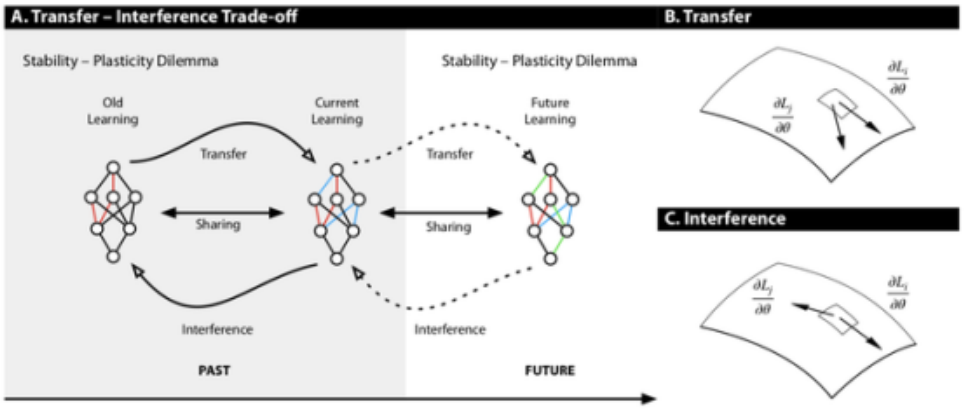


Figure 1: A) The stability-plasticity dilemma considers plasticity with respect to the current learning and how it degrades old learning. The transfer-interference trade-off considers the stability-plasticity dilemma and its dependence on weight sharing in both forward and backward directions. This symmetric view is crucial as solutions that purely focus on reducing the degree of weight-sharing are unlikely to produce transfer in the future. B) A depiction of transfer in weight space. C) A depiction of interference in weight space.

当神经网络对一系列任务进行学习时，它往往会遭遇被称作「灾难性遗忘」的问题。由于灾难性遗忘，神经网络无法再在之前训练的任务上取得好的性能。灾难性遗忘可以被认为是存在明显的消极负向迁移的迁移学习的特例。迁移学习（正如大多数人们所指的）以及元学习通常寻求最大化在最终的任务上的正向积极迁移，但是一般来说并不会关注它对于源任务的影响。本文试图在仍然能够实现积极迁移但不以灾难性遗忘（干扰）为代价的情况下取得更大的平衡。

为了解决这个问题，Riemer 等人提出了一种被称为元经验回放（MER）的方法。MER 采用了标准的经验回放，交叉存取过去的训练示例与当前的训练示例，从而防止发生灾难性遗忘。作者假设过去的训练示例学习率较低；其次，MER 采用流行的 REPTILE 元学习算法在新数据上进行训练。不过，MER 也将内存缓存器中的过去的训练示例与新的示例交错在一起，输入给由 REPTILE 驱动的内层训练循环，从而防止灾难性遗忘的发生。

我非常喜欢这篇论文，因为它同时探究了积极迁移和消极迁移的想法。本文的方法在 Omniglot 和强化学习环境中取得的实验结果似乎相当不错。然而，作者只在小型「玩具」数据集上进行了测试，尤其是在监督分类问题中。他们本应该也在 CIFAR-10 对比基准、CALTech-Birds 或 CORRE50 上进行测试。从这一点上说，由于还存在许多更加真实的 CL 数据集，他们没有理由仅仅在稍微修改过的 MNIST 或 Omniglot 数据集上进行测试。此外，我发现由于作者「重复命名」了一些之前命名过的概念，文中的一些术语令人困惑。而且，在理想情况下，当我们连续进行学习时，我们不必再在任何之前的数据上重新进行训练（重新训练会带来额外的计算开销）。然而，所有的这一切都是朝着正确的方向迈进了，我希望有更多的论文同时关注正向和负向迁移。更多关于该论文的信息，请参阅 IBM 的博文：「Unifying Continual Learning and Meta-Learning with Meta-Experience Replay」（https://www.ibm.com/blogs/research/2019/05/meta-experience-replay/）；论文代码地址：https://github.com/mattriemer/MER

《文本转语音的高效自适应采样》

Sample Efficient Adaptive Text-to-Speech

论文下载地址：https://openreview.net/forum?id=rkzjUoAcFX

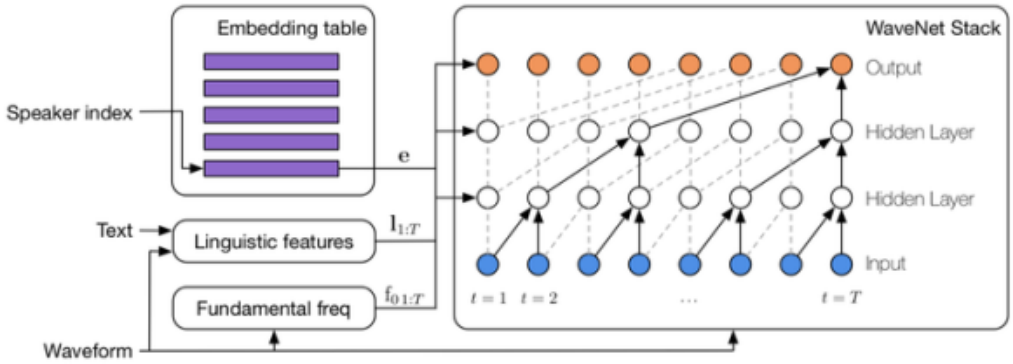


Figure 1: Architecture of the WaveNet model for few-shot voice adaptation.

这是一个将元学习运用到「序列到序列」建模任务中的有趣应用。在本例中，作者使用元学习来实现对说话者声音的小样本自适应。该应用十分重要，因为大多数情况下，你可能并不能获取某个特定说话者持续 100 秒或 1000 秒的声音。具体而言，作者拓展了 WaveNet 架构，从而引入了元学习技术。有趣的是，根据作者的说法，在他们初步的试验中，MAML 并没有生成有意义的先验。因此，他们不得不开发他们自己的架构。

该架构的工作流程分为三步：（1）在一个包含多名说话者的「文本-语音」对的大型语料库上训练模型；（2）根据某个特定说话者的少量「文本-语音」对调整模型；（3）最终在纯文本上进行推理，并将其转化为合适的语音。作者研究了两种小样本学习场景：带有一个嵌入编码器（SEA-ENC）的参数化 few-soht 自适应，以及带有调优过程的非参数化 few-shot 自适应（SEA-ALL）。在 SEA-ENC 的情况下，作者训练一个辅助嵌入网络，该网络会在给定新数据的情况下预测出一个说话者的嵌入向量。相比之下，对于 SEA-ALL 来说，作者同时训练网路和嵌入。在测试评估阶段，SEA-ALL 似乎性能更好，尽管作者声称模型在 SEA-ALL 的情况下会发生过拟合。因此，他们推荐使用早停法（early stopping）防止过拟合。（他们的模型仅仅在 10 秒内的 Librispeech 任务上比早先的论文所提出的模型表现更好）。

本文是一个很好的范例，它将小样本学习应用于典型的图像分类领域之外的棘手问题，并对其进行必要的调整使其能够真正有效。希望我们能够在未来看到有更多的研究者尝试将小样本学习应用于通用模型。作者提供了一个网站，你可以在上面测试他们的 TTS（Text to speaking）模型的demo。然而，遗憾的是，他们似乎没有公开他们的代码。

ICLR 其它相关论文概述

《K for the Price of 1：参数高效的多任务和迁移学习》

K for the Price of 1: Parameter-efficient Multi-task and Transfer Learning

论文下载地址：https://openreview.net/pdf?id=BJxvEh0cFQ

Mudrarkarta 等人提出了一个由少量可学习的参数组成的模型补丁包，这些参数专门针对各个任务。这种方法替代了对网络的最后一层进行调优的通常做法。作者发现这种方法不仅可以减少参数的数量（从超过 100 万减少到 3.5 万），还可以在迁移学习和多任务学习的环境下提升调优的准确率。唯一的缺点是，该补丁包似乎针对的只是相当具体的架构。

《用于距离度量学习的无监督域自适应方法》

Unsupervised Domain Adaptation for Distance Metric Learning

论文下载地址：https://openreview.net/forum?id=BklhAj09K7

尽管本论文第一部分的标题为「无监督域自适应」，它实际上研究的是迁移学习问题。回想一下，通常目标域会通过域自适应获得一组相同的标签。然而，在本例中，作者假设了一个无标签的目标域——正如一些审稿人提到的，本论文因此也变得有些令人困惑；不过，本文仍然有一些值得关注的地方：为了分离源域和目标域的调整空间，作者提出了一种特征迁移网络 FTN。并且，该作者在跨种族人脸识别任务上取得了目前最先进的性能。

《学习用于语法引导的程序合成的元解算器》

Learning a Meta-Solver for Syntax-Guided Program Synthesis

论文下载地址：https://openreview.net/forum?id=Syl8Sn0cK7&noteId=BJlUkwHxeV

本文讨论如何将元学习应用到程序合成任务中。在本文中，作者构建了一个语法引导程序，它遵循一个逻辑公式和语法，然后生成一个程序。本文是一个将元学习用于典型的小样本图像数据集之外的应用中的很好的范例。

《深度线性网络中泛化动态和迁移学习的分析理论》

An analytic theory of generalization dynamics and transfer learning in deep linear networks

论文下载地址：https://arxiv.org/abs/1809.10374
本文研究了学习和迁移学习的理论。作者声称「我们的理论解释了知识迁移敏感但可计算依赖于『信噪比』和任务对的输入特征对齐」。总而言之，对于那些喜欢深入研究理论的人来说，这篇文章非常有趣。

结语

我希望本文很好地概述了本届 ICLR 上有关小样本学习的大多数论文（尽管我可能会漏掉一些）。如你所见，本届 ICLR 上出现了各种各样有趣的新技术，它们开启了将深度学习用于数据有限的情况的大门。

via https://towardsdatascience.com/iclr-2019-overcoming-limited-data-382cd19db6d2雷锋网

.