小样本学习只是一场学术界自嗨吗

ALme 夕小瑶的卖萌屋 2022-05-07 12:05

文 | ALme@知乎

这两年看见很多人，包括我实习的mentor在内，都在批评few-shot learning，觉得是学术界在自high，思考良久，感觉有必要给这个领域正个名～(注意，此答案仅关注few-shot image classification）

首先，要讨论few-shot learning的价值，咱得先把few-shot learning（FSL）这个问题的定位搞清楚。而要搞清楚few-shot learning的定位，咱得先把FSL和transfer learning的关系捋清楚。

transfer learning大家都知道，一个典型例子就是在Imagenet训练一个backbone，然后在另一个新的数据集上（比如cifar、cub）的训练集微调（fine-tune）backbone，然后在这个新的数据集的测试集上对模型进行测试。那咱为啥不在新数据上从头train一个模型呢？

我们都知道，Imagenet图片量很大，且图片所覆盖域较为全面，可以近似看作对真实世界数据分布的刻画，因此希望在ImageNet上训练的模型能够提取通用的图片特征，而这种通用的特征很可能能迁移到下游一个没有见过的图片域。因此广泛认为，在ImageNet（或者更大的数据集）上训练一个backbone，然后再微调是最好的方式。这也是为什么这两年大家如此钟爱于超大数据预训练模型，有监督半监督自监督应有尽有，就是想着像bert一样造一个万能模型解决一切任务。

transfer learning有一个区别于domain adaptation的非常关键的点，即**训练时的数据集和微调时的数据集的图片不仅domain不同，category也通常是不一样的**。

由于category不同，导致微调时原有的网络分类层不能用了，得重新学一个；而由于domain不一样了，backbone提取的特征也不够discriminative了，因此需要finetune backbone。后面将看到，从这两点将直指few-shot learning核心问题。

重点来了。

transfer learning的setting，是假设我们能够接触到足够多的目标数据集的labeled data的，但在实际应用时，往往目标数据集的labeled data是不足的。

举一个我实习过程遇到的真实案例，当时遇到一个项目，是零件的异常检测，即给定一张工业零件的图片，判断其是否合格。大家都知道，零件造出来往往都是正常的，出错的概率是很低的，因此能够拿到的异常零件图片是很少的，当时的想法是imagenet学到的backbone直接在这些极少量的图片上finetune，最后结果很差很差；另一个例子是医学病情诊断，同样的，某些病情发病率极低，能够拿到的图片十分稀少，如果有机会可以试一试网上公开的ChestX [1]数据集，在labeled data数量给定的情况，从ImageNet finetune的效果也是极差。因此，这种setting在预训练模型十分重要的当下，是极具价值的。

那么这个setting和few-shot learning有啥关系？

其实，这个在transfer learning目标域labeled data不足的setting，就是咱常说的few-shot image classification，或者也可以叫做few-shot transfer [2]。few-shot image classification早期常用的benchmark，比如miniImageNet [5]，满足了few-shot transfer learning中的category gap，而domain gap虽然有，但是不明显。

为弥补这一缺陷，后续提出了cross-domain few-shot learning的benchmark [3] 以及Meta-Dataset [4]，这两年这些benchmark发展迅速，大部分刷传统benchmark的顶会论文也开始把cross-domain的效果放入论文。这些进展使得few-shot learning与实际应用场景的gap迅速缩小。大部分批评FSL的着重点可能都在miniImageNet上，其实，即使是miniImageNet，如果仔细观察，也可以发现其实训练集和测试集类别之间大多数是存在一个较大的gap的，比如测试集出现的微生物、花瓶，在训练集很难找出类似的类。**追溯批评的原因，还是大家在20年之前并没有把few-shot learning和transfer learning的关系搞清楚**，自然会觉得玩miniImageNet这种benchmark的都是在圈地自萌。

只有看清楚了这层关系，才能脱离出few-shot learning原本的范围，站在一个更高的维度思考问题本质。令人庆幸的是，虽然水论文在这个领域占比较大，但仍有一部分人正在朝着正确的方向前进，这就够了。

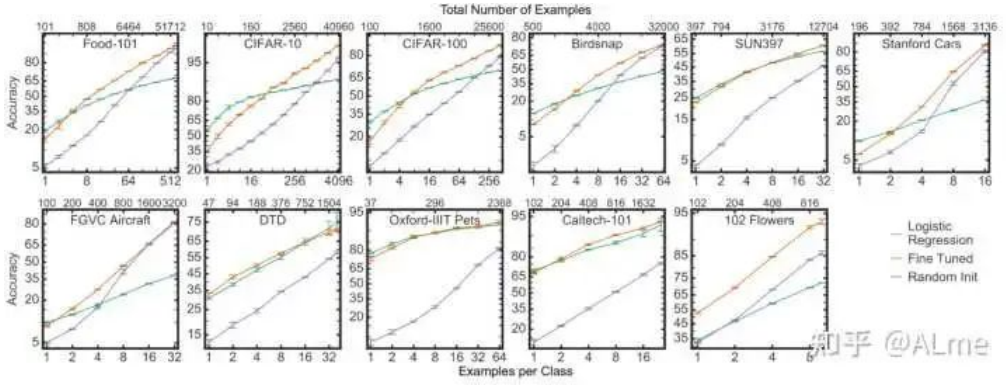
我们现在清楚了，few-shot image classification其实等价于限制目标域labeled data数量的transfer learning，那么问题来了，transfer learning基本就finetune一条路，玩不出花，为啥一旦把目标域数据量限制到很小，就出现了各种百花齐放的方法呢？

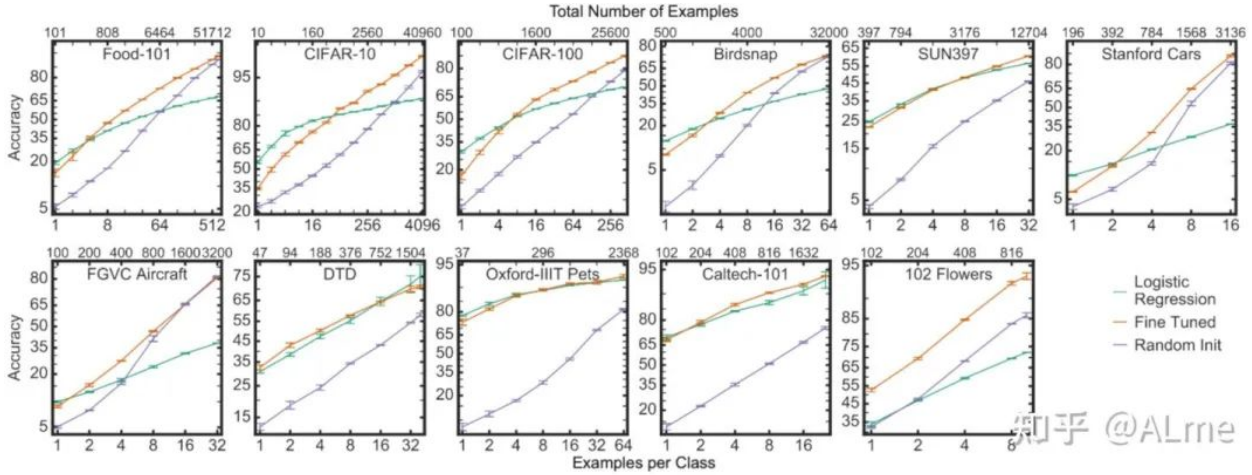
这些方法包括但不仅限于meta-learning、conventional training、metric-based method、generation-based method、linear classification、dense-feature based method。

其实，这一问题的答案可以追溯到19年谷歌发布于CVPR的一篇论文：

Do Better ImageNet Models Transfer Better?

该文探究了ImageNet上训练的模型的transfer learning效果。论文中的图9给出了transfer learning随着目标域labeled data数量增长时的效果变化，图片如下：





红色的线为finetune方法效果，绿色的线为冻住backbone，仅在特征之上用目标域数据训练一个线性logistic分类器的效果，黑色为在目标数据集上从头训练一个模型。

首先，黑色线效果不行，说明transfer的必要性。其次，更为有趣的是finetune和线性分类的performance在给定不同目标域数据量的差异。在目标域labeled data数据量较大情况下，finetune通常占据压倒性优势，但在few-shot场景下，finetune方法往往比不过冻住backbone的线性分类方法，注意到，该论文虽然降低了每类数目，但没有降低类别数目，而这些数据集上类别数目都很大，后来我自己做了实验，发现当类别数目变小时两种方法差异更大，这表示finetune效果与labeled data数据总量正相关。

这种现象，仔细思考其实很好理解，就是finetune backbone调整的参数量过多，在few-shot下很容易使得模型过拟合。这也解释了为什么MAML这类基于finetune的方法在few-shot learning下表现明显不如metric-based method等其他冻住backbone的方法。

既然不能finetune，那么理所当然地，在源域所学得的network backbone质量就至关重要。

换句话说，从backbone引导出的feature space必须足够general，足够generalizable。这一目标正是19-21年整个few-shot community关注的重点之一 [2, 6-8]，而该目标又恰好和这两年基于linear protocol evaluation的对比学习一致，好的，few-shot learning本质问题至此来到了vision problem的深水区：

怎么学得一个泛化能力极强的visual representation，使得在遇到下游极端奇异且少量labeled data时仍表现良好？

或者说，现有学得的visual representation在很奇怪的图片上时仍然存在怎样的问题？这些问题都是finetune打遍天下的传统transfer learning不具有的，也是few-shot learning的核心问题之一。从早期的元学习，到后来metric-based pretraining（cosine classifier）以及加各种自监督学习、蒸馏学习的loss，目标都是学一个更好的特征表示。

如果看过Big Transfer（BiT）[9]那篇文章，可能会问一个问题：

是不是只要数据量足够大，特征表示就足够好，小样本分类问题就解决了？

回答应该是，**partially solved**。

首先小样本分类效果和源域数据集大小在绝大部分目标数据集上是正相关关系，因此增大训练数据量是一个非常好的途径；但是，实验发现，这一增长在某些domain gap差距较大的数据集上，**特别是**实际遇到的真实应用场景中，是有上限的，如果不能从根本探究清楚pretrained visual representation在小样本下存在的问题，或者不使用除finetune之外的目标数据集adaptation方法，这一瓶颈看上去将无法解决。因此，few-shot image classfication这一问题有其独特价值，与image representation learning的核心问题紧密相关。

训练从源域学得general image representation之后，在测试时，目标域few-shot任务的所有图片，不管是support（训练）图片还是query（测试）图片，大部分方法均会先将其转为representation再进行下一步操作。这导向另一个问题，即在给定的representation下，如何最大化利用support set少量图片的representation构造一个分类器，使该分类器具有良好泛化能力？

把图像represention的潜力发挥到极致的方法很多，而这直接导致了few-shot learning方法的百花齐放。比如元学习方法，从训练开始就target这一问题，但这些元学习方法忽略了一个重要问题：训练源数据分布和测试时的目标数据分布是不同的，而这直接导致元学习的任务同分布假设不成立，这是元学习效果不佳的重要原因之一。

这里再举另外一个例子，由于

1. 目标域labeled data少
2. 目标域类别在训练时没见过

因此backbone网络会不知道在纷繁复杂的图片应该关注什么信息。

比如一张图，一个人牵着一只狗，标签为人，但由于网络在训练时可能只把狗作为标签（比如imagenet），因此提取特征时便关注狗去了，而不是人。为解决这类问题，dense-feature based方法应运而生，其核心思想是backbone出来的feature不过global pooling，保留spatial信息，对比不同图片的spatial feature map，从中找出对应关系，这样如果有两张图，其共性是人而不是狗，那通过这种人和人的对应关系就能把狗这一confounding factor给去除。这一类方法论文如：CAN[16]、CTX[2]、DeepEMD [10]、LDAMF[17]、MCL[18]。

可以看到，**训练学得一个good representation**，和**测试时从有限labeled data建立一个好的分类器**在一般的任务中是可以统一起来的。但在few-shot learning中，随着元学习方法的缺点不断被挖掘，这两点割裂开来，成为两个独立的问题。

前者涉及vision representation的本质问题，若为了涨效果可以照搬cv近期各自提升feature质量的trick，比如对比学习、蒸馏等等，成为了各大cv顶会刷点必备，这些方法水一水是可以的，但要真正解决问题，还是要探究visual representation在目标域labeled data是few-shot时所存在的核心问题，这样的研究最近是有[11-13]，但很少；后者涉及如何给定pretrained feature，做到快速task adaptation，核心点是

1. 取pretrained feature之精华，去其糟粕
2. 从support set feature及目标query feature中最大化可用信息，比如从support set中找类内共性，或者找support feature和query feature之间的对应关系，或者从训练集中找寻并利用和support set的相似图片，这第二点可以统称为task adaptation。

最后安利一下**meta-dataset**，这个benchmark非常接近真实场景，其中multi-domain FSL的setting从根本上解决了训练集单一domain泛化差的问题，根除了元学习方法的泛化障碍，可能能够使得task adaptation方法更加自然、有效，是一种可能的真正解决few-shot learning的方法途径。

这里提一嘴meta-dataset存在的一个bias，即测试时shot和way普遍偏高，这导致partial fine-tune[14,15]方法重现江湖，但实验后发现这些方法在1-shot和5-shot表现不佳，是值得注意的点。

最后的最后，吐槽一下**transductive few-shot learning**，我是真的不理解这种setting能有什么价值，如果有人知道，请告诉我：）

References:

[1] ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. CVPR 2017.

[2] Crosstransformers: Spatially-aware Few-shot Transfer. NeurIPS 2020.

[3] A Broader Study of Cross-Domain Few-Shot Learning. ECCV 2020.

[4] Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. ICLR 2020.

[5] Matching Networks for One Shot Learning. NeurIPS 2016.

[6] Rapid learning or feature reuse? towards understanding the effectiveness of MAML. ICLR 2020.

[7] A baseline for few-shot image classification. ICLR 2020.

[8] Rethinking few-shot image classification: A good embedding is all you need? ECCV 2020.

[9] Big Transfer (BiT): General Visual Representation Learning. ECCV 2020.

[10] DeepEMD: Few-Shot Image Classification with Differentiable Earth Mover's Distance and Structured Classifiers. CVPR 2020.

[11] Interventional Few-Shot Learning. NeurIPS 2020.

[12] Powering Finetuning in Few-Shot Learning: Domain-Agnostic Bias Reduction with Selected Sampling. AAAI 2022.

[13] Z-Score Normalization, Hubness, and Few-Shot Learning. ICCV 2021.

[14] Learning a Universal Template for Few-shot Dataset Generalization. ICML 2021.

[15] Cross-domain Few-shot Learning with Task-specific Adapters. CVPR 2022.

[16] Cross Attention Network for Few-shot Classification. NeurIPS 2019.

[17] Learning Dynamic Alignment via Meta-filter for Few-shot Learning. CVPR 2021.

[18] Learning to Affiliate: Mutual Centralized Learning for Few-shot Classification. CVPR 2022.

发表于北京

阅读原文

喜欢此内容的人还喜欢

贝叶斯深度学习：一个统一深度学习和概率图模型的框架

AI科技评论

号称最强深度学习笔记本电脑，雷蛇与Lambda公司推出，售价超2万

机器之心

我是吴恩达：人在美国，刚上知乎，先答个「如何系统学习机器学习」

量子位