什么能做，什么不能做.

Model Ratatouille:Recycling Diverse Models for Out-of-Distribution Generalization

摘要

基础模型正在重新定义人工智能系统的构建方式。从业者现在遵循一个标准程序来构建他们的机器学习解决方案:从一个预训练的基础模型，他们微调感兴趣的目标任务的权重。因此，互联网上充斥着为数不多的基础模型，这些模型对许多不同的任务进行了微调:这些单独的微调是孤立存在的，而不是相互受益的。在我们看来，这是一个错失的机会，因为这些专门的模型包含丰富多样的功能。因此，在本文中，我们提出了模型杂烩，这是一种新的策略，可以在不同的辅助任务上循环使用同一基础模型的多次微调。具体来说，我们将这些辅助权重重新用作目标任务上多个并行微调的初始化;然后，对所有微调过的权重求平均值，得到最终模型。这种回收策略旨在通过利用辅助任务的多样性来最大化权重的多样性。从经验上讲，它提高了对分布外泛化的参考DomainBed基准的技术水平。展望未来，这项工作有助于可更新机器学习的新兴范例，类似于开源软件开发，社区合作以可靠地更新机器学习模型。我们的代码在这里发布。

介绍

基础模型框架 **(Bommasani et al.，2021)** 正在推动机器学习解决方案在现实世界应用中的广泛采用:也称为预训练模型，这些机器学习系统是在大型和多样化的数据上训练的(Fang et al.，2022;Nguyen et al.，2022;Abnar et al.， 2022)，易于适应下游任务。在 **抛弃了“从头开始培训”的心态后** ，从业者现在遵循标准化的两步转移学习策略 **(Oquab等人，2014)** 。从一些基础模型出发，他们通常利用有限的内部数据对目标任务进行微调。不幸的是，这些微调中的每一个都有可能锁定从业者训练数据中的特定模式(Arjovsky等人，2019;Milleret al.， 2020;Shah et al.， 2020)。因此，这些目光短浅的模型难以对分布外(OOD)样本进行泛化(Hendrycks & Dietterich, 2019;Taori等，2020;Gulrajani & Lopez-Paz, 2021;hendricks等人，2021)，对人类生活产生负面影响(Taylor等人，2016;Zechet al.， 2018)。增强OOD泛化将能够在鲁棒性和安全

性至关重要的现实应用中负责任地使用机器学习，例如医学成像(DeGrave等人，2021)和自动驾驶(Kuutti等人，2020)。因此，如何对面向对象的泛化的基础模型进行最佳的微调成为研究的中心课题。特别是，最近发现的平均神经网络权重的能力(Izmailov等人，2018;Neyshabur等人，2020)激发了大量现代微调方法。我们在图1中说明了其中的一些，例如移动平均线(Izmailov等人，2018)，WiSE微调(Wortsman等人，2022b)，模型汤(Wortsmanet等人，2022a)和DiWA (Ram等人，2022a)。然而，这些策略不能适应互联网上越来越多的相同基础模型的专门微调。近期的交叉训练(Phang et al.，2018;Pruksachatkun et al.， 2020)和融合(Choshen et al.，2022b;Don-Yehiya et al.， 2022)策略在对目标任务进行微调之前，对辅助任务进行中间微调，以丰富特征。然而，这些回收策略的成功通常取决于辅助任务和目标任务之间的相似性。我们还在第2节中指出，尽管特征多样性提高了ood泛化，但这些策略未能充分利用辅助任务中的多样性(Laakom等人，2021;Nayman et al.，2022;Jain et al.， 2022;Zhang等人，2022)。

问题：我们如何才能最好地回收给定基础模型的各种微调，以实现我们目标任务的强分布外性能?

我们的答案是一个简单的微调策略，我们将其命名为ratatouille模型，如图1所示，并在第3节中进行了描述。与将废物转换为新用途的可重用材料类似，我们在不同的辅助任务上对相同的基础模型进行微调，并将其重新用作初始化，以启动目标下游任务上的多个微调。具体来说，我们(i)在每个辅助任务上对基础模型的副本进行微调，(ii)在目标任务上对每个辅助模型进行微调，以及(iii)将所有目标微调权重的平均值作为最终模型返回。简而言之，虽然模型汤(Wortsman等人，2022a)平均了从共享初始化中微调的多个权重，但模型杂烩(ratatouille)在不同的辅助任务上平均了从不同初始化中微调的多个权重(Phang等人，2018)。正如我们将看到的，料理杂烩之所以有效，是因为微调保持线性连接(Frankle et al.， 2020;Mirzadeh等人，2021)在损失景观(尽管有不同的初始化)

图注释

本文中讨论的不同微调策略:香草微调(Oquab等人，2014)，移动平均(Izmailov等人，2018)和变体(Wortsman等人，2022b)，模型汤(Wortsman等人，2022a)和DiWA (Ram ' e等人，2022a)，内部训练(Phang等人，2018)，融合(Choshen等人，2022b)和我们提出的模型ratatouille。它们都是从一个预先训练好的基础模型开始的。有些策略在辅助任务(细实线箭头)上对预训练模型进行微调:这些辅助微调可以由社区的不同贡献者在他们自己的数据上执行。然后，所有策略对感兴趣的目标任务(粗实箭头)执行微调。最后，对目标任务进行微调的权重按原样使用，或者在最终模型中取平均值(虚线箭头)。Ratatouille (i)在整个训练过程中实现了计算并行性，(ii)最大化了模型预测的多样性，(iii)**在DomainBed (Gulrajani & Lopez-Paz, 2021)中实现了最先进的性能，这是OOD泛化的标准计算机视觉基准**，(iv)与**传统的超参数搜索**相比，不会产生任何推理或

训练开销。

名字由来

我们以这道传统的法国菜命名我们的方法主要有两个原因。首先，炖菜通常被用作回收剩菜的一种方式。其次，炖菜最好是将每种食材分别烹饪，然后再将它们混合在一起:正如厨师joel Robuchon(摩纳哥，2020年)所指出的那样，这种技术确保每种食材“都能真正品尝到自己的味道”。

我们在第4节中展示了ratatouille模型的有效性，其中我们在**DomainBed (Gulrajani & Lopez-Paz, 2021)上设置了新的技术状态，这是评估OOD泛化的参考基准**。我们将展示如何利用辅助任务之间的多样性来构建一个最终模型，以减少对特定于任务的模式的过度拟合。正如我们在结语第5节中讨论的那样，这项工作有助于新兴的**可更新机器学习范式(Raffel, 2023)**，在这种范式中，从业者通过协作，逐步可靠地更新机器学习模型的功能。最近的作品 **(Matena & Raffel, 2022;Li等人，2022a)**，我们设想了一个未来，深度神经网络通过与版本控制系统的开源开发中类似的管道进行训练。

我们的模型ratatouille是对同一预训练模型的各种辅助微调进行循环的建议;将其与图1中的其他微调策略进行比较，并在图2b中详细列出。Ratatouille将这些微调作为不同的初始化进行循环，以在目标任务上并行微调。与融合相比，我们延迟了加权平均，反过来又破坏了多样性。《料理杂烩》遵循这五步食谱。

经历过初步人工智能或者机器学习的实践者会发现：人工智能其实并不智能。

我们的数值实验支持四种主要主张，按粒度排序。首先，4.1节展示了在DomainBed (Gulrajani & Lopez-Paz, 2021)中ratatouille的最先进(SoTA)结果。其次，第4.2节说明了这些增益是如何从平均模型的多样性增加中产生的。第三，第4.3节从经验上支持假设1和假设2，即使加权平均成功的技术条件。最后，第4.4节讨论了ratatouille对域内任务的影响。我们邀请好奇的读者查阅我们的补充材料。在其他实验中，我们在附录B中删除了ratatouille过程的不同组成部分，例如辅助任务的数量，并在附录E中提出了一个鲁棒的ratatouille来进一步提高性能。我们的代码发布在https://github.com/facebookres search /ModelRatatouille。

此文似乎是在调参。让人听起来很没水平。

In Search of Lost Domain Generalization （DomainBed）

摘要

领域泛化算法的目标是对不同于训练中看到的分布进行很好的预测。虽然存在无数的领域泛化算法，但实验条件(数据集、架构和模型选择标准)的不一致性使得公平和现实的比较变得困难。在本文中，**我们感兴趣的是了解领域泛化算法在现实环境中有多有用**。作为第一步，我们意识到**模型选择对于领域泛化任务是非常重要的**。与先前的工作相反，我们认为**没有模型选择策略的领域泛化算法应该被认为是不完整的**。接下来，**我们实现了DOMAINBED，一个包含7个多域数据集、9个基线算法和3个模型选择标准的域泛化测试平台**。我们使用DOMAINBED进行了广泛的实验，并发现，当仔细实施时，**经验风险最小化**在所有数据集中显示出最先进的性能。展望未来，我们希望**DOMAINBED的发布**，以及其他研究人员的贡献，将简化领域泛化中可重复和严格的研究。

介绍

机器学习系统通常不能泛化分布外，当在训练样本领域之外进行测试时，会以惊人的方式崩溃[Torralba和Efros, 2011]。学习系统对训练分布的过度依赖表现得很普遍。例如，**自动驾驶汽车系统**很难在与训练条件不同的条件下执行，包括光线变化[Dai和Van Gool, 2018]，天气[Volk等人，2019]和物体姿势[Alcorn等人，2019]。另一个例子是，**在一家医院收集的医疗数据上训练的系统不能推广到其他医疗中心**[Castro等人，2019,AlBadawy等人，2018,Perone等人，2019,Heaven, 2020]。**Arjovsky等人[2019]认为**，未能推广分布外就是未能捕获数据变化的因果因素，而是坚持更容易拟合的虚假相关性，这些虚假相关性很容易从训练域变化到测试域。学习机器通常吸收的虚假相关性的例子**包括种族偏见[Stock和Cisse, 2018]、纹理统计[Geirhos等人，2018]和物体背景[Beery等人，2018]**。可惜，分布外的机器学习系统的反复无常行为是它们在关键应用中部署的障碍。

意识到这个问题，**研究团体在过去十年中花费了大量的精力来开发能够泛化分布外的算法**。特别是，领域泛化的文献假设在训练期间**访问多个数据集**，每个数据集包含关于相同任务的示例，但在不同的领域或环境下收集[Blanchard et al.， 2011, Muandet et al.， 2013]。领域泛化算法的目标是将这些训练数据集的**不变性**合并到分类器中，希望这些不变性也适用于新的测试领域。**尽管领域泛化非常重要，但文献是分散的:每年出现大量不同的算法，这些算法在不同的数据集和模型选择标准下进行评估**。借鉴ImageNet等标准计算机视觉基准的成功经验 **[Russakovsky等人，2015]**，本研究的目的是对领域泛化算法进行标准化、严格的比较。特别是，我们会问:领域

泛化算法在现实环境中有多有用?为了回答这个问题，我们首先研究了 **领域泛化方法的模型选择标准**，得出以下建议:

领域泛化算法应该负责指定模型选择方法。

然后，我们在**7个多领域数据集和3个模型选择**标准上仔细实施了9种领域泛化算法，得出了表1和表4所示的结论: **当配备现代神经网络架构和数据增强技术时，经验风险最小化在领域泛化中实现了最先进的性能**。作为我们研究的结果，我们发布了DOMAINBED，一个在领域泛化中简化严格和可重复实验的框架。使用DOMAINBED，**添加新算法或数据集只是几行代码的问题(我需要这么方便吗？)**;一个单一的命令运行所有的实验，执行所有的模型选择，并自动生成这项工作中包含的所有表。此外，我们的动机是保持DOMAINBED的活力，欢迎来自同事的拉取请求来更新可用的算法、数据集、模型选择标准和结果表。第2节首先回顾域泛化设置。第3节讨论了领域泛化中模型选择的困难，并提出了前进的道路建议。第4节介绍了DOMAINBED，描述了初始版本中包含的算法和数据集。第5节讨论了运行整个DOMAINBED套件的实验结果;这些都说明了ERM的力量和模型选择标准的重要性。最后，**第六节对领域泛化的未来研究方向提出了展望**。我们的附录回顾了该主题十年来研究的**一百篇文章**，收集了**三十多个**已发表算法的实验性能。

领域泛化不同于无监督的领域自适应。在后一种方法中，假设在训练过程中有来自测试域的未标记数据可用[Pan and Yang, 2009, Patel et al.， 2015, Wilson and Cook, 2018]。表2比较了不同的机器学习设置，以突出领域泛化问题的本质。因果关系文献将领域泛化称为从多个环境中学习[Peters等人，2016,Arjovsky等人，2019]。虽然具有挑战性，但领域泛化是对实际预测问题的最佳逼近，其中训练和测试数据之间不可预见的分布差异肯定是可以预料到的。

Remark

看了这么多formulation，我觉得问题的本质是introduction里的第一句“机器学习系统通常不能泛化分布外，当在训练样本领域之外进行测试时，会以惊人的方式崩溃[Torralba和Efros, 2011]”.

```
Machine learning systems often fail to generalize out-of-distribution, crashing in spectacular ways
```

在这里，我们讨论了在领域泛化中围绕模型选择(选择超参数、训练检查点、架构变体)的问题，并为前进的道路提出了具体的建议。因为我们**无法访问与测试数据相同分布的验证集**，所以域泛化中的模型选择不像监督学习中的模型选择那么简单。一些作品采用启发式策略，**其行为没有得到很好的研究**，而另一些作品只是**忽略了如何选择超参数的描述**。这留下了使用测试数据选择超参数的可能性，这在方法上不合理。由于不一致的调优实践而产生的结果差异可能被错误地归因于所研究的算法，从而使公平评估复杂化。我们认为，在领域泛化中围绕模型选择的许多困惑源

于将其视为实验设计问题。**实际上，选择超参数是一个学习问题，至少和拟合模型一样困难(因为我们可以将任何模型参数解释为超参数)**。像所有的学习问题一样，模型选择需要假设测试数据与训练数据之间的关系。不同的领域泛化算法做出了不同的假设，并且先验地不清楚哪些假设是正确的，或者这些假设如何影响模型选择准则。事实上，选择合理的假设是领域泛化研究的核心。因此，一个没有超参数选择策略的领域泛化算法是不完整的。建议1领域泛化算法应该负责指定模型选择方法。虽然没有合理的模型选择方法的算法是不完整的，但它们可能是研究议程中的踏脚石。在这种情况下，我们可以通过考虑oracle模型选择方法来评估不完整的算法，而不是使用特别的模型选择方法，其中我们在测试域中选择超参数。当然，避免在oracle结果和没有使用oracle方法调优的基线之间进行无效比较是很重要的。同样，除非我们以某种方式限制对测试领域数据的访问，否则我们可能会获得无意义的结果。例如，我们可以使用监督学习在这样的测试领域数据上进行训练。建议2研究人员应该放弃任何oracle选择结果，并指定策略来限制对测试域的访问。

DOMAINBED包括7个多域图像分类任务的下载和加载器:Colored MNIST [Arjovsky等，2019]、rotrotmnist [Ghifary等，2015]、PACS [Li等，2017]、VLCS [Fang等，2013]、Office-Home [Venkateswara等，2017]、Terra Incognita [Beery等，2018]和DomainNet [Peng等，2019]。我们在表3中列出并展示了每个数据集的示例图像，并在附录c中提供了它们的完整细节。数据集在许多方面有所不同，但有两个特别重要。**第一个区别是合成数据集和真实数据集之间的区别**。在旋转MNIST和有色MNIST中，域是综合构建的，这样我们就知道哪些特征会先验地泛化，因此使用太多的先验知识(例如通过旋转增加)是禁止的，而其他数据集包含由自然过程产生的域，因此使用先验知识是明智的。第二个区别是跨领域的变化。一方面，在有色MNIST以外的数据集中，域改变了图像的分布，但可能没有关于真正的图像到标签映射的信息。另一方面，在Colored MNIST中，该域影响真正的图像到标签映射，使试图直接估计该函数的算法产生偏差。

The datasets differ in many ways but two are particularly important. The first difference is between **synthetic and real datasets**. In Rotated MNIST and Colored MNIST, domains are synthetically constructed such that we know what features will generalize a priori, so using too much prior knowledge (e.g. by augmenting with rotations) is off-limits, whereas the other datasets contain domains arising from natural processes, making it sensible to use prior knowledge. **The second difference is about what changes across domains**. On one hand, in datasets other than Colored MNIST, the domain changes the distribution of images, **but likely bears no information about the true image-to-label mapping**. On the other hand, in Colored MNIST, the domain influences the true image-to-label mapping, biasing algorithms that try to estimate this function directly. (不懂这里在说什么.)

3.1三种模型选择方法在提出了广泛的建议之后，我们回顾并证明了三种在领域泛化中经常使用

但很少被识别的模型选择方法。**我们将每个训练域划分为训练子集和验证子集**。然后，我们将每个训练域的验证子集集合起来，**创建一个整体的验证集**。最后，**我们选择在整个验证集上精度最大化的模型**。**该策略假设训练和测试示例遵循相似的分布**。例如，Ben-David等人[2010]通过训练域误差加上训练域和测试域之间的发散度量来绑定分类器的测试域误差。给定dtr训练域，我们用相等的超参数训练dtr模型，每个超参数保留一个训练域。我们在其hold -out域上评估每个模型，并在其hold -out域上平均这些模型的精度。最后，我们选择最大平均精度的模型，在所有dtr域上重新训练。该策略假设训练和测试域是从域上的元分布中绘制的，并且我们的目标是在该元分布下最大化预期性能。测试域验证集(oracle)我们选择在遵循测试域分布的验证集上最大化准确性的模型。按照我们之前限制测试域访问的建议，我们允许每个算法有20个查询(在随机搜索中，每个超参数选择一个查询)。这意味着我们不允许基于验证集的提前停止。相反，我们用相同的固定步数训练所有模型，并且只考虑最后的检查点。回想一下，我们并不认为这是一个有效的基准测试方法，因为它需要访问测试域。oracle选择结果可以是乐观的，因为我们访问了测试分布;也可以是悲观的，因为查询限制减少了考虑的超参数组合的数量。作为限制查询数量的替代方案，我们可以从差分隐私中借用工具，该工具以前用于在标准监督学习中实现验证集的多次重用[Dwork等人，2015]。简而言之，差分隐私工具在将其报告给从业者之前，将拉普拉斯噪声添加到算法的准确性统计中。

先前的一些文献讨论了在领域泛化问题中选择超参数的附加策略。例如，Krueger等人[2020，附录B.1]建议选择超参数来最大化外部数据集所有领域的性能。该策略的有效性取决于数据集之间的相关性。Albuquerque等人[2019，第5.3.2节]建议基于损失函数(通常包含特定于算法的正则器)进行模型选择，DInnocente和Caputo[2018，第3节]推导出特定于其算法的策略。

算法

DOMAINBED的初始版本包括**九个基线算法的实现**:经验风险最小化(ERM, Vapnik[1998])最小化跨域和示例的错误总和。群体分布鲁棒优化(DRO, Sagawa等人[2019])在提高误差较大的域的重要性的同时执行ERM。interdomain Mixup (Mixup, Xu et al. [2019]， Yan et al. [2020]， Wang et al.[2020])对来自随机域对及其标签的示例进行线性插值，执行ERM。元学习领域泛化(MLDG, Li等人[2018a])利用MAML [Finn等人，2017]来元学习如何跨领域泛化。Ganin等人[2016]使用跨域匹配的分布来学习特征(Xd)的流行算法的不同变体:—域对抗神经网络(DANN, Ganin等人[2016])使用对抗网络来匹配特征分布。—类条件DANN (C-DANN, Li et al. [2018d])是DANN的一种变体，匹配所有标签y跨域的条件分布P((Xd)jY d = y)。—CORAL [Sun and Saenko，， 2016]匹配特征分布的均值和协方差。- MMD [Li et al.， 2018b]匹配特征分布的MMD [Gretton et al.， 2012]。不变风险最小化(IRM [Arjovsky等人，2019])学习一个特征表示(Xdd)，使得该表示之上的最优线性分类器跨域匹配。附录D描述了所有算法的网络架构和超参数搜索空间

4.3现实评估的实现选择我们的目标是对领域泛化算法进行现实评估。为此，我们做出了几个与前面工作不同的实现选择，如下所述。大型模型先前关于VLCS和PACS的大多数工作都借鉴了ResNet-18模型的特征或对其进行了微调[He等，2016]。由于**已知较大的resnet泛化效果更好**，我们选择对除旋转MNIST和彩色MNIST外的所有数据集微调ResNet-50模型，其中我们使用较小的CNN架构(见附录D)。数据增强数据增强是训练图像分类模型的标准成分。在领域泛化中，当数据增强可以近似领域之间的一些变化时，数据增强可以发挥特别重要的作用。因此，对于所有非mnist数据集，我们使用以下数据增强进行训练:随机大小和宽高比的裁剪，调整大小到224 224像素，随机水平翻转，随机颜色抖动，以10%的概率对图像进行灰度化，并使用ImageNet通道均值和标准差进行归一化。对于MNIST数据集，我们不使用数据扩增。使用旋转MNIST中所有可用的数据，而通常版本的数据集从相同的1000位数字集构建所有域，我们将所有MNIST数字均匀地划分到域中。我们偏离标准实践有两个原因:我们相信在训练和测试领域使用相同的数字会泄漏测试数据，并且我们相信人为地限制可用的训练领域数据会以一种不切实际的方式使任务复杂化。

我们对DOMAINBED中提供的所有算法(第4.2节)、数据集(第4.1节)和模型选择标准(第3节)运行实验。我们考虑数据集的所有配置，其中我们隐藏一个域进行测试，并在其余的域上进行训练。对于每个算法和测试环境，我们在**超参数分布上对20个试验进行随机搜索[Bergstra和Bengio, 2012]**(见附录D)。我们使用第3节中的每种模型选择方法从随机搜索的20个模型中进行选择。我们将每个领域的数据分成80%和20%的部分。我们使用较大的分割用于训练和最终评估，较小的分割用于选择超参数。虽然一些领域泛化文献报告了种子间的误差条，但模型选择产生的随机性往往被忽略。如果目标是最佳对最佳的比较，这是可以接受的，但它禁止细致入微的分析。例如，方法A优于方法B仅仅是因为随机搜索A很幸运吗?因此，我们将整个研究重复三次，重新进行每个随机选择:超参数、权重初始化和数据集分割。我们报告的每个数字都是这些重复的平均值，以及它们的估计标准误差。这个实验协议**总共训练了45900个神经网络**。

5.1结果表4总结了我们的实验结果。对于每个数据集和模型，我们在测试域中平均最佳结果(根据每个模型选择标准)。然后，我们报告在整个扫描的三个独立运行中该数字的平均值，以及相应的标准误差。对于每个数据集和领域的结果，我们请读者参阅附录b。我们从我们的结果中得出三个主要结论:**我们的ERM基线优于之前发表的所有结果**，表1总结了使用训练域验证集执行模型选择时的结果。**是什么导致了这种强劲的表现**?我们怀疑有四个因素:**更大的网络架构(ResNet-50)，强大的数据增强，仔细的超参数调优，以及在旋转的MNIST中，使用完整的训练数据来构建我们的域(而不是使用1000个图像子集)**。虽然我们不是第一个单独使用这些技术的人，**但我们可能是第一个将所有这些技术结合起来的人**。有趣的是，这些结果表明，**提高分布内泛化的标准技术在提高分布外泛化方面非常有效**。我们的结果并没有反驳先前的工作:有可能使用类似的技术，一些竞争的方法可能会改进ERM。相反，我们的结果强调了将领域泛化算法与强大且现实的基线进行比较的重要性。将新的算法合并到DOMAINBED中是一种简单的方法。

对于在文献中发表的关于三十多种算法的结果的广泛回顾，我们建议读者参阅附录A.5。**当所有条件都相等时，没有任何算法比ERM表现得更好**。我们在表4中观察到这一结果，它是通过从头开始运行DOMAINBED中包含的数据集、算法和模型选择标准的每个组合获得的。在给定任何模型选择标准的情况下，没有任何一种方法可以提高ERM在多个点上的平均性能。我们并不是说这些算法中的任何一种都不可能改进ERM，但是在这些数据集上对ERM进行实质性的领域泛化改进证明是具有挑战性的。我们观察到，使用训练域验证集的模型选择优于跨多个数据集和算法的leave-one- out交叉验证。这并不意味着使用训练域验证集是调优超参数的正确方法。毕竟，它没有使任何算法显著优于ERM基线。此外，oracle-selection更强的性能(+2%)表明可能有改进的空间。

我们对领域泛化算法进行了广泛的经验评估。我们的研究结果**得出了两个主要结论**。首先，与八种流行的领域泛化替代方案相比，**经验风险最小化实现了最先进的性能**，也改进了先前文献中报道的所有数字。其次，**模型选择对领域泛化有重要影响**，它应该被视为任何提出的方法的组成部分。最后，我们将进行一系列小型讨论，回答一些问题，但也会提出更多问题。我们如何进一步推动数据增强?在进行实验的过程中，**我们意识到了数据增强的力量**。**Zhang等人[2019]表明，强数据增强可以提高分布外泛化，同时不影响分布内泛化**。我们将数据增强视为特征去除:我们对训练示例的增强越多，我们的预测器相对于应用的转换就越不变性。如果执行者足够幸运，并且执行了数据增强，**消除了域与域之间的虚假相关性**，那么分布外性能应该得到改善。给定一个特定的领域泛化问题，我们应该实现哪种类型的数据增强管道?

这就够好了吗?我们质疑在考虑的数据集中是否期望域泛化。为什么我们认为神经网络应该能够在只给出逼真的训练数据的情况下对漫画进行分类?**在旋转MNIST的情况下，是否存在数字类的真正旋转不变特征?这些特征可以用神经网络表达吗**?即使在正确选择模型的情况下，现代ERM实现的分布外性能是否达到了应有的水平?还是说它和其他选择一样糟糕?我们如何通过域泛化技术建立分布外可实现的性能的上限?

这些是正确的数据集吗?领域泛化文献中考虑的一些数据集不能反映现实情况。实际上，如果想对漫画进行分类，最简单的选择是收集一个小的标记漫画数据集。我们是否应该考虑更现实、更有影响力的任务来更好地研究领域泛化?有吸引力的替代方案包括**不同医院的医疗成像**和**不同城市的自动驾驶汽车**。**每次我们使用ERM时，我们都假设训练和测试示例来自相同的分布。而且每一次，这都是一个无法验证的假设**。这同样适用于域泛化:每个算法假定跨域的不变性类型不同(不可测试)。因此，领域泛化算法的性能取决于手头的问题，只有时间才能告诉我们是否做出了好的选择。这类似于科学理论的一般化，比如牛顿的万有引力理论，它无法被证明，但迄今为止还没有被证伪。我们相信在测试期间具有自适应能力的算法是有希望的。虽然限制使用现代技术**降低了实验的成本，但它也使实验偏离了更现实的场景**，这是我们研究的重点。我们的观点是，基准设计师应该平衡这些因素，以促进一套游戏规则，这些规则不仅定义明确，而且现实且

动机良好。合成数据集是有用的工具，但我们不能忽视我们的目标，那就是人工智能能够在现实世界中进行推广。用马塞尔·普鲁斯特的话来说:也许我们周围的事物之所以是不动的，是因为我们相信它们是它们自己，而不是其他任何东西，也因为我们对它们的观念是不动的。

当前的机器学习系统在面对新的示例分布时会反复无常地失败。这种不可靠性阻碍了机器学习系统在交通、安全和医疗保健等关键应用中的应用。在这里，我们努力寻找强大的机器学习模型，摒弃虚假的相关性，因为我们期望不变模式可以推广分布外。这将带来更公平、更安全、更可靠的机器学习系统。但权力越大，责任越大:领域泛化研究人员必须坚持最严格的模型选择和评价标准。我们希望我们的结果和DOMAINBED的发布是朝着这个方向迈出的一小步，我们期待着与其他研究人员合作，以简化可重复和严格的研究，实现真正的泛化能力。

Muandet等人[2013]使用核方法找到了一种特征变换，该变换可以(i)最小化变换后的特征分布跨域之间的距离，(ii)不破坏原始特征与目标之间的任何信息。在他们的开创性工作中，Ganin等人[2016]提出了领域对抗神经网络(DANN)，这是一种使用生成对抗网络(gan, Goodfellow等人[2014])的领域自适应技术，以学习跨训练域匹配的特征表示。Akuzawa等人[2019]通过考虑域和类标签变量之间存在统计依赖的情况来扩展DANN。Albuquerque等人[2019]通过考虑单对全的对手来扩展DANN，这些对手试图预测每个示例属于哪个训练域。Li等人[2018b]使用gan和最大平均差异标准[Gretton等人，2012]来对齐跨域的特征分布。Matsuura和Harada[2019]利用聚类技术来学习域不变特征，即使没有给出训练域之间的分离。Li等人[2018c,d]学习了一种特征变换，使得条件分布P((Xd) j j Y d = Y)与所有训练域d和标签值Y相匹配。Shankar等人[2018]使用域分类器为标签分类器构建对抗示例，并使用标签分类器为域分类器构建对抗示例。这将产生具有更好的领域泛化的标签分类器。Li等。[2019a]训练一种鲁棒特征提取器和分类器。鲁棒性来自(i)要求特征提取器产生特征，使得在域d上训练的分类器可以对域d06 = d的实例进行分类，以及(ii)要求分类器使用在域d06 = d上训练的特征提取器产生的特征来预测域d上的标签。Li等人[2020]采用终身学习策略来解决域泛化问题。Motiian等人[2017]学习了一种特征表示，使得(i)来自不同领域但同一类的示例距离近，(ii)来自不同领域和类的示例距离远，以及(iii)训练样例可以正确分类。Ilse等人[2019]训练了一个变分自编码器[Kingma and Welling, 2014]，其中瓶颈表示分解了输入空间中关于域、类标签和残差变化的知识。Fang等人[2013]学习了一种结构性SVM度量，使得每个示例的邻域包含来自同一类别和所有训练域的示例。Sun和Saenko [2016]，Sun等人[2016]，Rahman等人[2019a]的算法在一定程度上表示跨训练域的特征协方差(二阶统计量)匹配。Ghifary等人[2016]，Hu等人[2019]的算法使用基于核的多元成分分析来最小化训练域之间的不匹配，同时最大化类可分性。虽然很流行，但学习域不变特征也受到了一些批评[Zhao等人，2019,Johansson等人，2019]。下面我们将回顾一些替代方案。Peters等人[2016]，Rojas-Carulla等人[2018]认为应该在训练域中搜索导致相同最优分类器的特征。在他们的开创性工作中，Peters等人[2016]将这种不变性与数据的因果结构联系起来，并提供了一种基于特征选择学习不变线性模型的基本算法。Arjovsky等人[2019]在其不变风险最小化(IRM)原则

中将之前的方法扩展到一般基于梯度的模型，包括神经网络。Teney等人[2020]在IRM的基础上学习了一种特征转换，该转换可以最小化训练数据集上分类器权重的相对方差。作者应用他们的方法来减少视觉问答(VQA)任务中虚假相关性的学习。Ahuja等[2020]从博弈论的角度分析IRM，以开发替代算法。Krueger等人[2020]提出了IRM问题的近似方法，包括减少跨域误差平均的方差。Bouvier等人[2019]通过重新加权数据样本来解决与IRM相同的问题。