|  |  |
| --- | --- |
| **文章信息** | **内容** |
| **阅读程度**：  精读  **文章标题：**  TWINS: A Fine-Tuning Framework for Improved Transferability of Adversarial Robustness and Generalization  **中文标题：**  TWINS：一个用于提高（对抗鲁棒性与泛化性）的迁移性的微调框架  **发表于：**  CVPR2023  **作者：**  Ziquan Liu1, Yi Xu2\*  **单位：**  一作来自香港城市大学  通信作者来自大连理工大学  我对本文的理解是，双BN层中冻结BN始终保持着从Imagenet数据集中提取到的批归一化均值方差，相当于Imagenet数据集的一些统计信息，因此Cifar10的样本经过这些BN层后相当于使用Imagenet数据集的信息对Cifar10的样本进行了数据增强。  就像首次使用双BN层的AdvProp方法（使用自然样本BN与对抗样本BN，实现利用对抗样本作为数据增强来提高自然样本精度，此处对抗样本BN就是一个数据增强BN层）。 | 本文旨在研究如何从预训练模型开始实现在其他模数据集上的对抗鲁棒性问题：即已有一个基于Imagenet的对抗训练模型，如何对其进行使用，实现在其他数据集上的鲁棒性模型。  在之前的微调问题中，已有基于数据与基于模型的微调策略，当然这些策略在正常训练中是有效的，但是在对抗训练中效果不佳，作者选择了两种典型的策略，即基于模型的正则化方法和基于数据的正则化方法。    首先损失函数的第一项都是标准对抗训练的损失函数，基于模型的方案中，正则化项旨在促进当前模型与原始预训练模型对当前数据集对抗样本提取到的特征向量保持一致，基于数据的方案的正则化项旨在促进当前模型对原始数据集依然保持较好的特征提取能力。两个正则化项都旨在缩小微调后网络与原始预训练网络之间的距离，旨在希望尽可能地保留原始预训练网络已经取得的鲁棒性特征提取能力。  本文提出的方法：  双bn层结构，第一个bn层参数冻结，保持为原预训练网络提取到的均值方差信息，第二个bn层就正常更新。训练时，生成对抗样本的过程使用第二个bn层，损失函数如下：    简单来说就是一半样本经过冻结bn层然后反向传递提出，一半样本经过adaptive bn层正常传递梯度。  在4.2部分作者进行了一些理论上的分析，没看后续可以看看  **本文提供了两个思路：**  **第一对比学习的预训练网络能否在后续的微调过程中使用此策略，保留对比学习过程中学习到的鲁棒性特征提取能力。**  **第二本文是基于baseline进行改进的方法，所有对比实验也是仅与baseline进行对比，还有较大的改进空间，可以考虑复现并改进。** |
| **摘要** |  |
| 近年来，预训练模型及其下游培训在深度学习研究和应用中的重要性越来越大。（大背景）  同时，对抗样本的防御主要是在对简单分类任务的随机初始化训练的背景下进行研究的。（小背景下的局限性）  为了更好地利用预训练模型在对抗性鲁棒性方面的潜力，本文重点研究了反向预训练模型在各种分类任务中的微调。（本文的研究方向）  现有的研究表明，由于鲁棒预训练模型已经学习了鲁棒特征提取器，关键问题是在学习下游任务时如何保持预训练模型的鲁棒性。（已有的研究内容）  我们研究了基于模型和基于数据的方法，发现这两种常见的方法都不能同时达到提高泛化和对抗性鲁棒性的目标。（已有研究内容的问题）  因此，我们提出了一种新的基于统计的方法，Two-WIng NormliSation (TWINS)微调框架，该框架由两个神经网络组成，其中一个在批处理归一化层中保持训练前数据的总体均值和方差。（本文提出的方法）  除了鲁棒信息传递外，由于在标准批归一化层中权值范数与其梯度范数之间的关系被打破，双胞胎在不影响训练稳定性的情况下提高了有效学习率，从而更快地摆脱了次优初始化，减轻了鲁棒过拟合。（简单解释本文提出的方法）  **最后，双胞胎在泛化和图像分类数据集的鲁棒性方面都是有效的。（效果）**  **Tips:对于没有达到sota的方法应该这样写效果。** | 最主要的性能展示：    可以看到基线是使用AT与TRADES直接在预训练模型上进行微调，  然后是本文提出的方法对基线产生的改进，评估策略就是简单的PGD10与AA策略。  后续还有关于可行性的消融实验。 |