ADVERSARIALLY ROBUST TRANSFER LEARNING

文章链接：<https://openreview.net/forum?id=ryebG04YvB>

**1 Abstract**

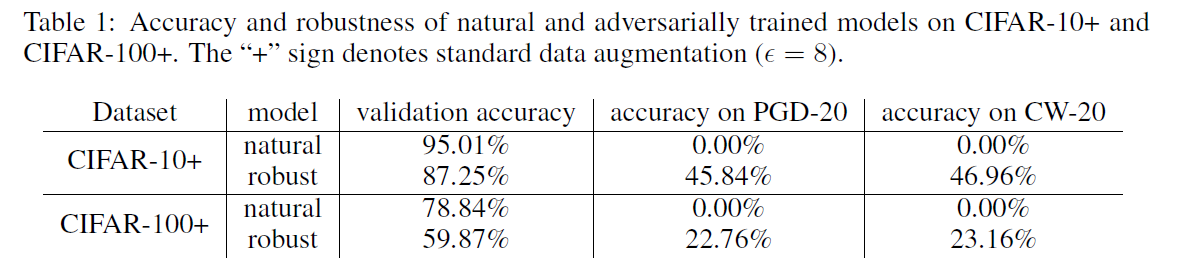
本文提出robust transfer learning，同时考虑performance和robustness，只需要少量training data，不需要耗时的adversarial training。作者发现robust networks具有robust feature extractors。实验显示1）target 模型重用 source模型的robust feature extractor（冻结，不训练），只re-training最后的全连接层可以继承robust source 模型的robustness。2）target 模型re-training或fine-tuning整个模型，robustness性质会丢失，因为robust feature extractor忘记了开始学到的robustness。进而，作者利用lifelong learning，解决了robustness无法继承的问题。

lifelong learning：即一个model可以先学习task1，再学习task2。得到的模型同时可以解决这两个tasks，解决forgetting问题

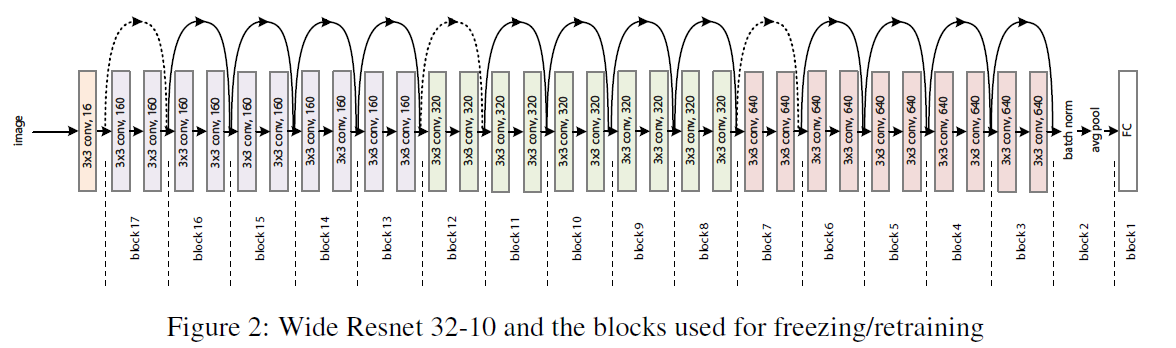
**2 Introduction**

adversarial training需要大量的计算开销和训练数据，为了避免这个问题，本文利用transfer learning引入robustness，既不需要耗时的adversarial training，又只需少量的clean dataset。

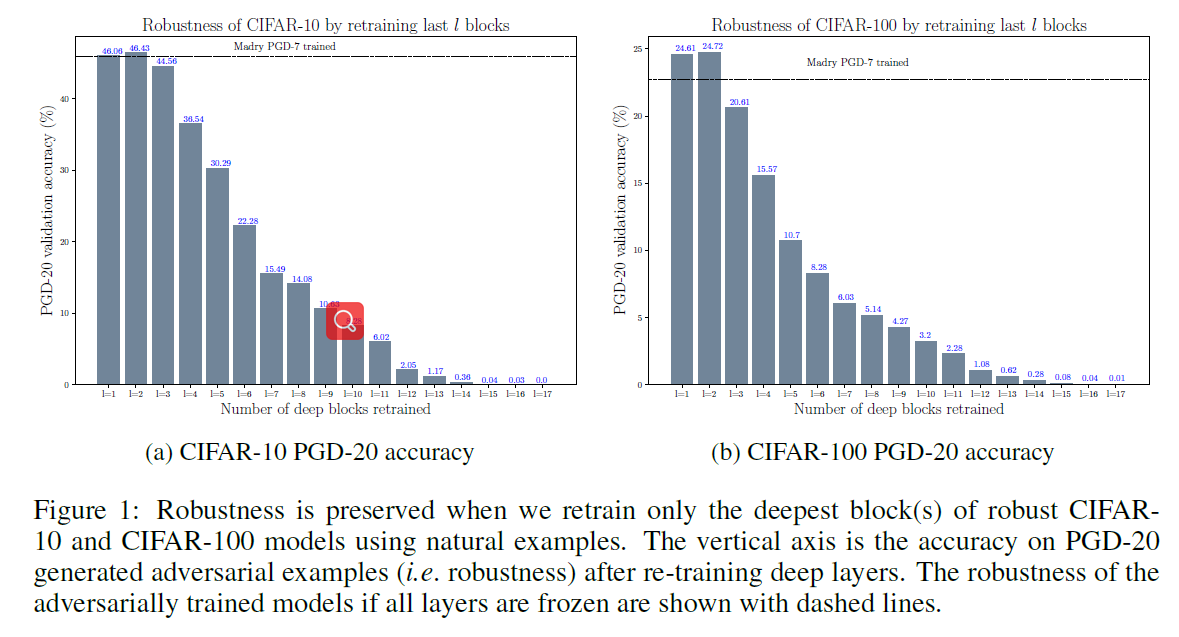
**3 Approach**



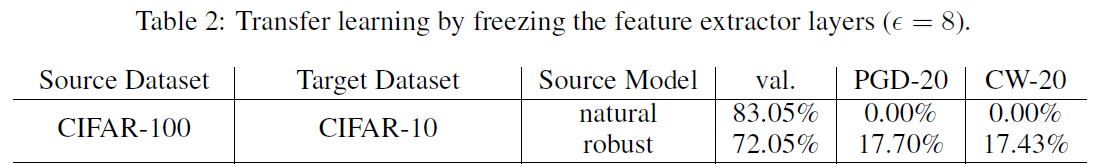
如上表，作者给出在cifar-10和cifar-100上，没有transfer learning时，对抗训练后accuracy和robustness，做为一个baseline。



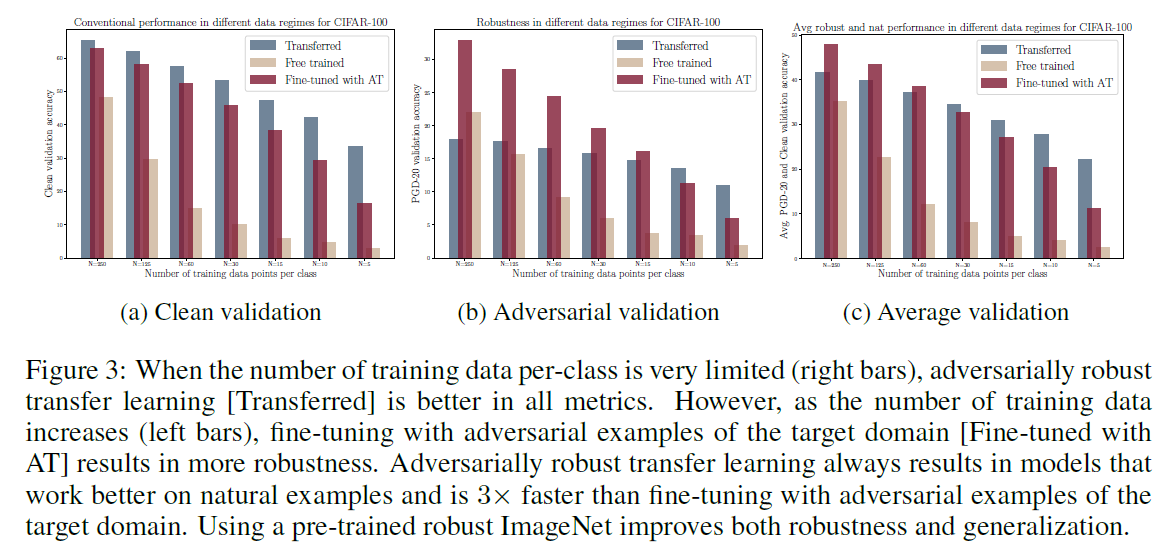
为了研究transfer learning对robustness的影响，作者将上图Resnet 32网络分块，在相同的source and target domains下，用相同的数据re-training不同层，验证其对robustness的影响。



如上图，作者发现only re-training last layer对robustness没有影响，甚至会提高。但是，re-training 更多的层，robustness严重下降。因此，robust network依赖robust deep features。

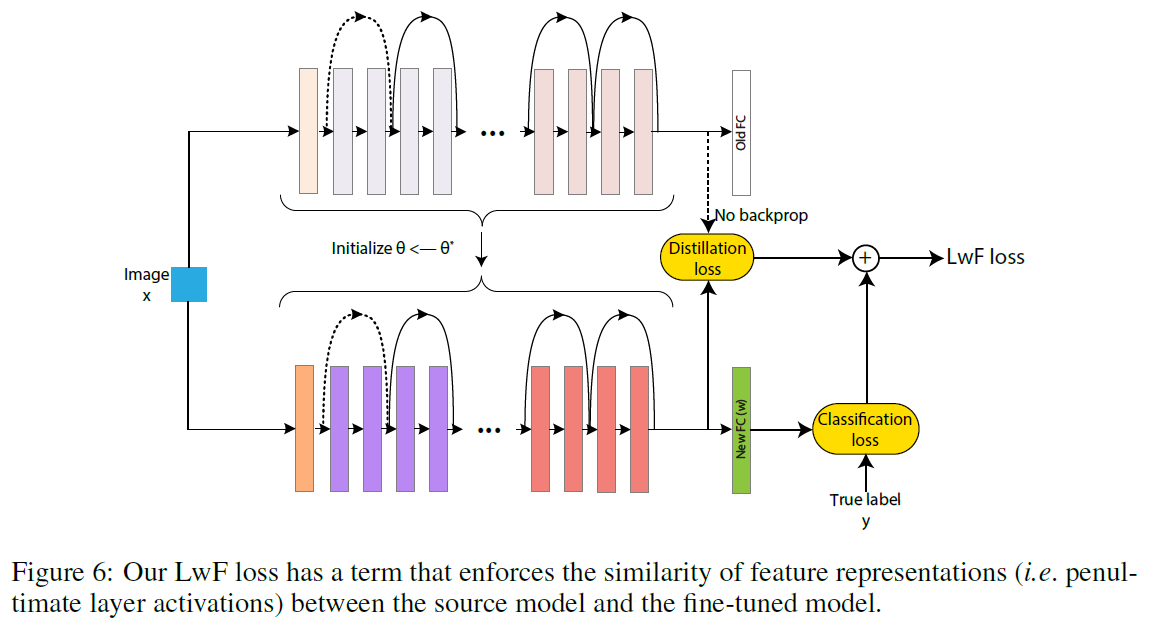


如上表，在different source and target domains下进行验证。重用robust feature extractor，只retraining最后一层，相比于从natural training的source model上迁移，从robust training的source mode上迁移，accuracy略有下降，但是robustness增加了。



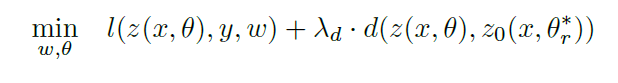
如上图，作者又验证了训练样本的数量对accuracy和robustness的影响。少量样本时，利用从robust source model上迁移，再retraining最后一层，可以得到更好的结果。数据样本多时，利用从robust source model上迁移，再利用adversarial training，fine-turning整个模型效果更好。

由图1可知，target 模型re-training或fine-tuning整个模型，robustness性质也会丢失。为了解决这个问题，作者利用lifelong learning。

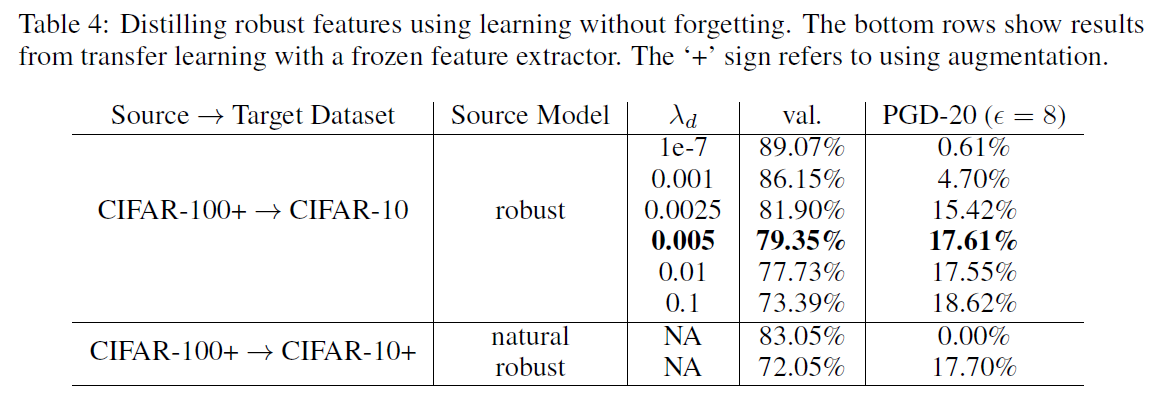


如上图，作者首先从robust source model上迁移，再fine-turning整个模型，同时强制target model的feature extractor层的参数，与source model的robust feature extractor层的参数差别不大，因此target model产生的deep features变化不大，保持robustness。

需要解决以下目标函数求解：



其中，z（）表示feature extractor得到的特征，表示transfer的feature extractor的参数，w表示最后一层的参数。后一项做为penalty function，限制变化不能过大。



如上图，作者发现，利用lifelong learning + transfer learning产生的target model，比只利用transfer learning产生的模型，有更好的accuracy和robustness。

**4 Conclusion**

个人观点：已有文章研究adversarial sample的transferability，本篇文章利用相同思想研究model robustness的transferability。本文全文几乎没有公式，因此利用实验结果探索未研究的领域，仍然可以发表很好的顶会。