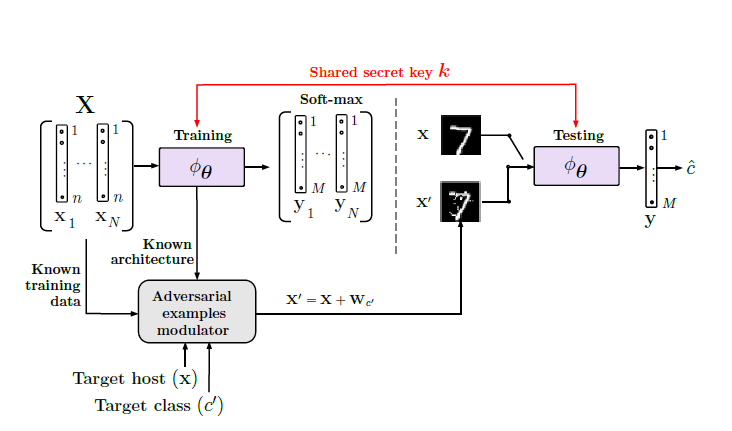
Defending against adversarial attacks by randomized diversification

文章链接：<https://arxiv.org/pdf/1904.00689.pdf>（Accepted by CVPR 2019）

**Abstract**

本文提出了一个随机多样化（randomized diversification）的防御策略。介绍了灰盒场景中的多通道结构（multi-channel）（敌手已知分类器结构和训练集，测试时可以访问密钥和系统内部状态）。防御者在多个通道处理一个输入，每个通道基于训练和测试阶段的共享密钥，在特殊的变换域（transform domain）中引入自己的随机化。这种机制可以保证梯度在防御者的密钥定义子空间内，但会造成梯度反向传播，为攻击者创建各种旁路系统。解决是融合所有通道的soft-outputs，以增加final score的可靠性。密钥共享为防御者创造了优势

**Therat Model**



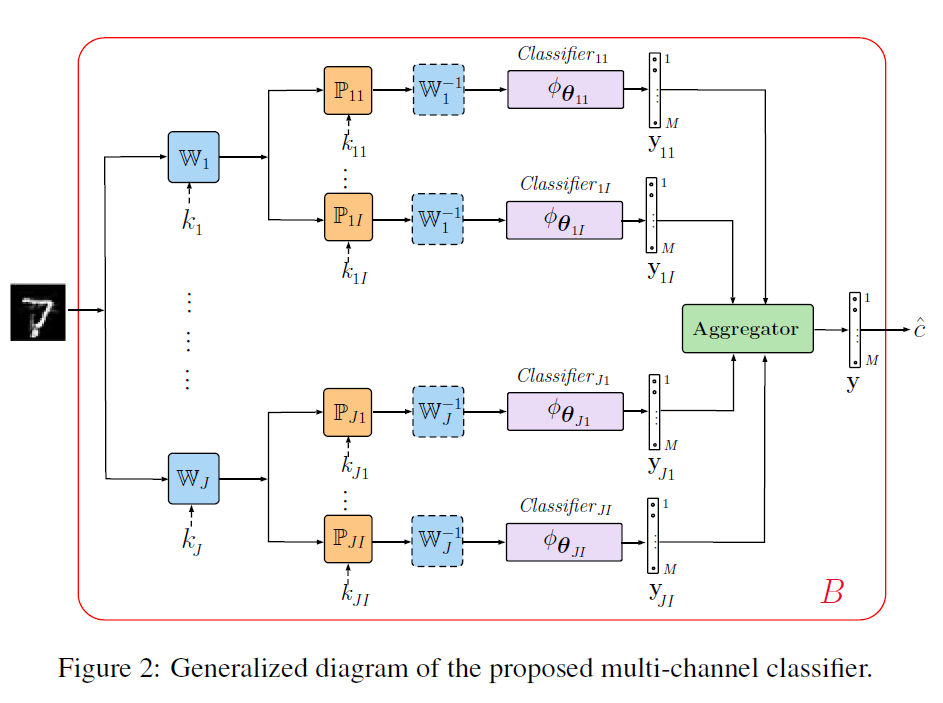
The attacker知道训练集X以及系统结构，但不知道防御者在训练和测试阶段共享的密钥k。防御者可以访问分类器和训练数据集X。分类器输出长度为M的soft-max向量y，其中M为分类的总数，为给定输入x，它属于分类c的概率。是添加扰动后的对抗样本。

注：ReLU （更接近生物神经元）相比 Tanh 能产生相同或者更好的性能，而且能产生真零的稀疏表示，非常适合自然就稀疏的数据。

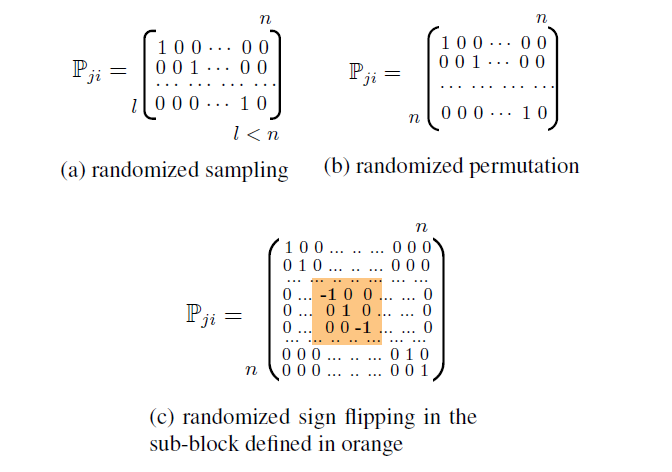
目前已知的防御有三类：non key-based defenses（缺少密码元素，防御者没有信息优势）、Defense via randomization and obfuscation（可避免重复使用系统参数在进行训练，缺点攻击者可以绕过防御块，或者在产生对抗扰动时考虑这种模糊性，只能保证平均性能）、key-based defense（缺点是有用数据的减少会导致性能损失）

**Method**

1. 本文提出的基于密钥的防御系统：

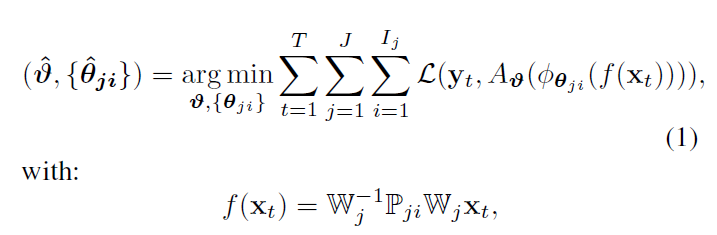


1. 预处理：通过linear mapping ，对输入数据进行转换，可以是随机映射或正交变换（如DFT，DCT，DWT等），也可以是learnable transform。数据的独立变换可以避免从训练数据中泄露密钥。是的密钥。
2. 数据独立处理：，代表随机化的第二层，避免梯度反向传播到direct domain。黄色部分是指基于密钥的符号翻转的密钥定义区域，这个操作对defender来说是可逆的，但对攻击者是不可逆的，因为没有密钥。



1. 分类块：任意分类器都可，但如果是为direct domain中的数据分类所设计的分类器，最好在分类前加上。（？）
2. 聚合块：从简单求和到适应数据或者特定对抗攻击的可学习操作，通过该操作将所有子数据聚合在一起，最后对类进行判断。

算法公式描述为：



是分类损失函数，是样本的分类标签，是参数的聚合操作，是第j个通道的第i个分类器，是分类器参数，T为训练样本数量，J是通道总数，是每个通道的分类器数量（保持不变）。

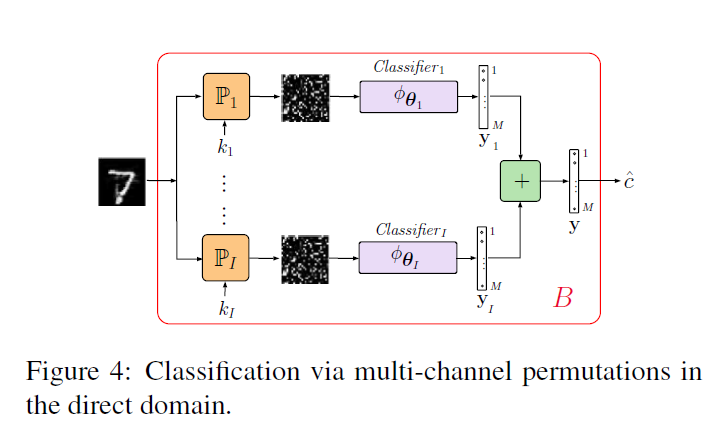
为了防止敌手获取到密钥或，or利用后向传播可微逼近技术（Backward Pass Differentiable Approximation technique，参考0327\_ICML\_2018\_best）或通过旁路替换器替换掉key-based blocks，使得整个系统端到端变为可微的，本文限制敌手无法访问模块B内的内部结果。

该方法优势：

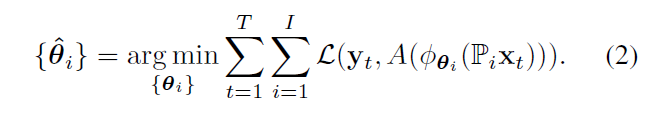
1. 在聚合之前，防御者可以单独训练每个通道。与此同时，攻击者在所有通道中同时训练和反向传播梯度，或者至少要保证聚合后大部分的分类错误。
2. 使用多通道结构和聚合操作可以稳定结果偏差（由于随机化使用或者有损变换导致）。
3. 正确的聚合操作可以提高自由度，并通过可能适应特定攻击类型来增加系统安全性。
4. 每个通道独立随机化可以提高整体安全等级，每个通道可以根据所需的防御等级对随机量调整。

**简单情况**

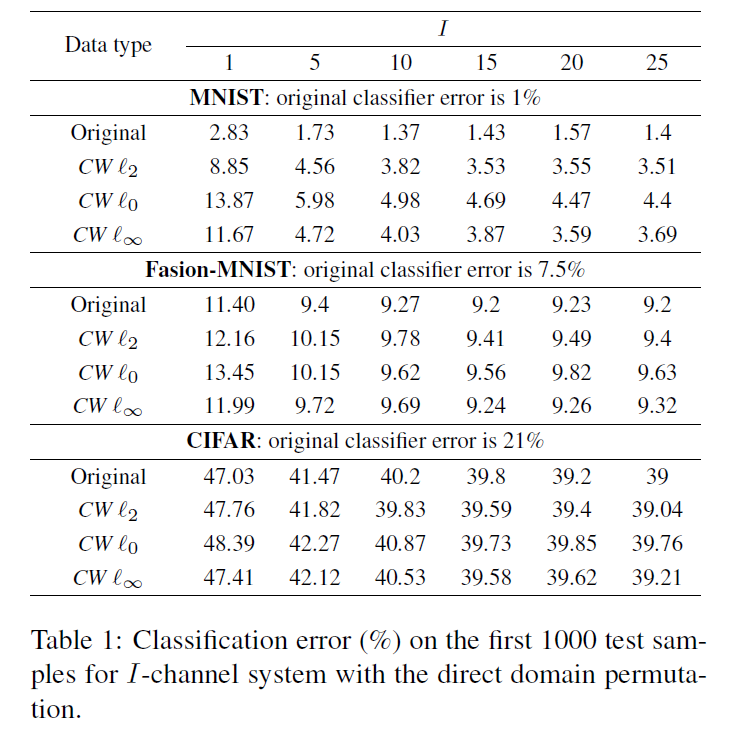
**Classification with multi-channel permutations in the direct domain**



通过在直接域内的多通道置换分类如图，直接域是指：。因此每个通道只有1个数据独立置换块Pi，这个置换块由输入信号的无损置换表示（n为输入图像尺寸，m是图片通道数目），因此置换矩阵Pi的大小为，由密钥产生，这个矩阵每行除了一个1外，都是0，如图3b。实验假定Pi都相同，聚合操作为求和，则优化问题可以归约到：



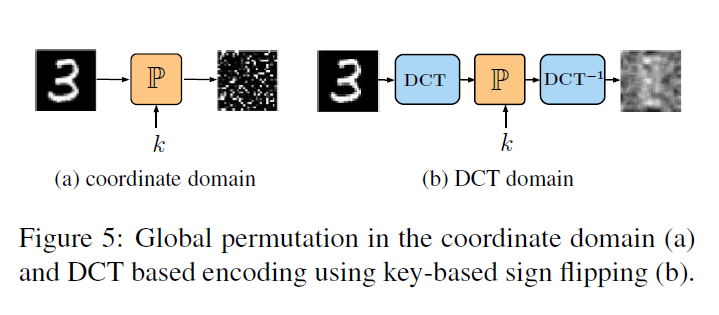
实验结果TABLE1所示：



将本文提出的方法应用到MINIST、Fashion-MINIST、CIFAR，**可以看到，增加通道数可以降低分类错误**，多通道算法可以将分类错误降低2-2.5倍。但在CIFAR-10中，应用本文方法破坏了本地相关性，影响了分类器性能。

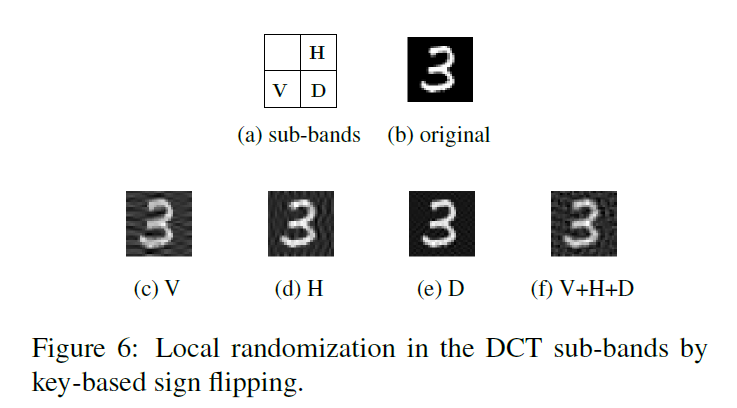
**改进2**

**Classification with multi-channel sign permutation in the DCT domain**



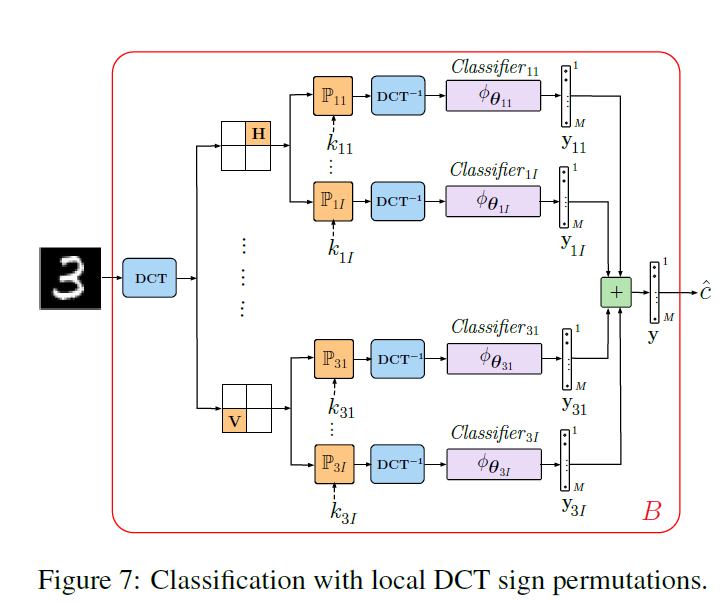
坐标系域内置换和DCT域内置换的结果如图，坐标域的置换会干扰图像的局部相关，从而影响局部梯度。而全局DCT符号置换可以随机化的同时，可以保证局部相关性。将直接域内的多通道中的是DCT，，保密部分包括基于密钥的DCT系数符号翻转。

但在DCT域内进行全局符号置换并没有改善上一节的问题。因而提出在使用全局DCT变换时仅翻转选定的系数的符号，如图6c。

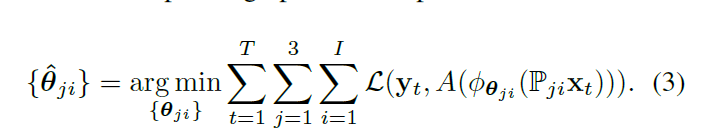


为了简单，本文将DCT域分为4个子带，左上角为图片的低频，左下角为垂直，右上角为水平，右下角为对角线分解后，然后，将DCT符号翻转应用于每个子带的随机化，保持所有其他子带不变，并应用反DCT变换。最后应用在3个子带得到6f。很容易看出，局部DCT符号翻转应用于一个单独的子带时很好地保留了局部图像内容。3个子带同时应用时，会造成失真。

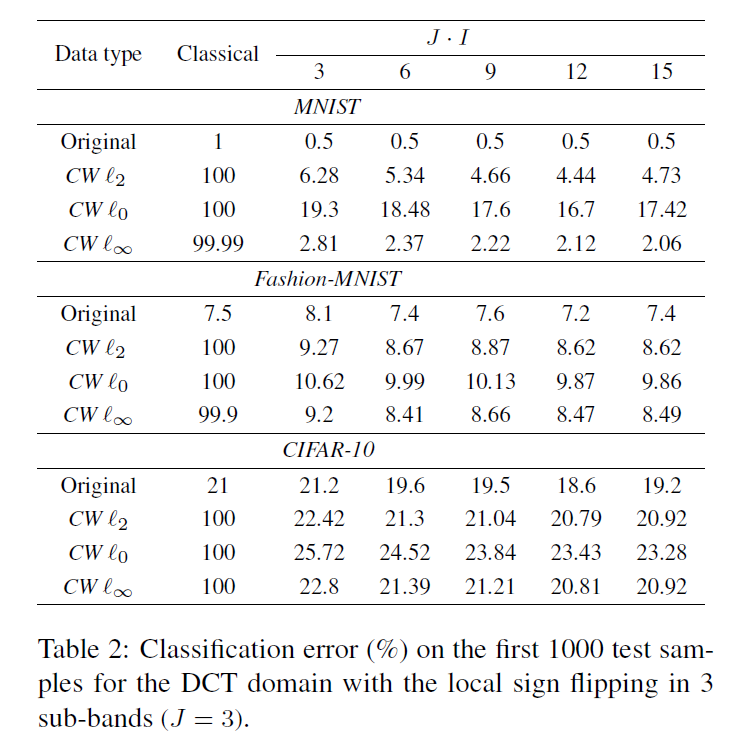
采用分子带的DCT进行随机变换来对数据进行处理如图：



利用I子通道的3通道模型，使用标准的DCT变换，对数据进行处理，只需要使用3个不重叠的等大小的子带就达到了section2的效果，公式描述为：



实验结果：



分类错误率相比于table1，明显得到降低，且在CIFAR-10中，分类器性能不会受到影响。攻击成功率仅为2%

**Conclusion**

未来的工作：考虑基于梯度的稀疏攻击和非梯度下降攻击的防御。

（细节上关于域内的变换还没完全搞懂，好像没看到是针对的什么类型攻击）