arXiv:2304.02049v2 [cs.CV] 8 Jun 2024

通过权重筛选实现图像分类的多类别遗忘

Samuele Poppi
意大利摩德纳41125，摩德纳-雷焦艾米利亚大学
Sara Sarto
意大利摩德纳41125，摩德纳-雷焦艾米利亚大学
Marcella Cornia意大利雷焦艾米利亚42121，摩德纳-雷焦艾米利亚大学
Lorenzo Baraldi摩德纳-雷焦艾米利亚大学，意大利摩德纳，41125
Rita Cucchiara摩德纳-雷焦艾米利亚大学，意大利摩德纳，41125

摘要—机器遗忘是一种新兴范式，可有选择地消除训练数据点对网络的影响。与针对有限子集或单个类别的现有方法不同，我们的框架能够在单轮中遗忘所有类别。我们通过使用记忆矩阵调节网络组件来实现这一点，使网络在训练后能够针对任意类别表现出选择性遗忘行为。通过发现每个类别特有的权重，我们的方法还以可解释的方式恢复类别表征。我们在小型和中型图像分类数据集上测试了所提出的框架，并采用基于卷积和 Transformer 的骨干网络，展示了通过机器遗忘实现可解释解决方案的潜力。

深度学习模型以及网络数据集都要求保护用户隐私并支持“被遗忘权” [1]。在知识蒸馏、模型剪枝或增量学习等能够改变模型内部知识表征的技术中，机器遗忘 [2] 是唯一能够删除私人训练数据的技术。机器遗忘方法 [3]——[5]，事实上，会从训练好的模型中移除特定数据点留下的痕迹。它们的目标是“反训练”模型，消除不需要的数据点所带来的影响，并使权重达到与……相似的状态

未使用此类数据训练的模型。例如，可以从分类网络中移除整个类别 [6]–[8]，，这在人脸识别等场景中很有用。现有的机器遗忘方法仅限于联合遗忘一个或少数几个类别，而我们的工作进一步实现了以正交方式在所有类别上进行机器遗忘，并且只需一次微调。具体而言，我们的方法能够以正交方式对所有类别进行机器遗忘，从而为用户提供可定制的最终模型，该模型可在运行时按照用户的要求，表现得如同针对用户所选单一类别进行机器遗忘的模型。这是通过利用内部

© 2024

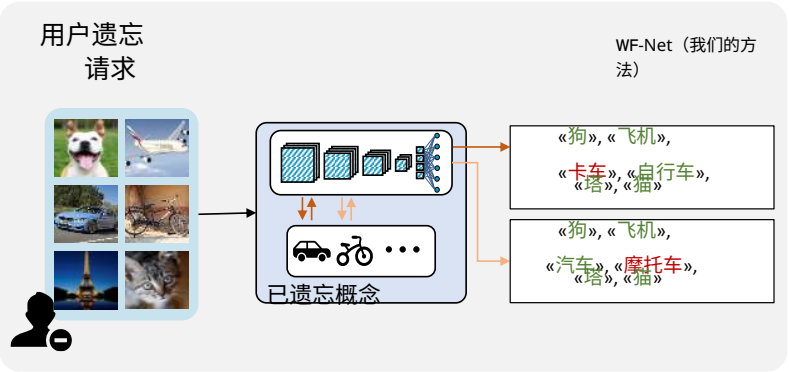

图 1. 所提出的任意 类别单轮反训练轮次中的单轮多类别遗忘设置。WF-Net 所需的计算资源更少，并且具有设计上的可解释性。

网络组件（例如，卷积神经网络（CNN）中的卷积滤波器，或视觉 Transformer（ViT）中的注意力投影）与输出类别之间的映射 [9]，并构建权重过滤层来实现的。该层可以选择性地启用或禁用这些内部组件，从而针对选定类别实现所需的机器遗忘行为（图1）。这种方法通过封装现有网络算子来实现，避免改变网络结构。通过训练所提出的权重过滤层，我们实质上揭示了内部网络组件与输出类别之间的联系。这不仅能够有效实现机器遗忘，还能提供本质上可解释的表征，最终通过机器遗忘获得可解释性。我们在小型和中型图像分类数据集上对所提出的方法进行了实验验证，并将其命名为 WF-Net。实验结果表明，该方法适用于多种图像分类架构，包括卷积神经网络和 ViT，也适用于更具挑战性的情形：无法访问训练集时进行机器遗忘。

贡献。 总而言之，本工作的贡献如下：

- 我们的框架使图像分类网络能够在单轮中同时对多个类别进行机器遗忘。与现有方法相比，这种方法显著降低了机器遗忘和测试阶段的计算需求，并提供了更大的灵活性。

- 我们的方法将网络内部组件（如卷积滤波器或注意力投影）封装到权重过滤层中，从而选择性地激活或停用这些组件，以实现所需的机器遗忘行为。

- 此外，我们的方法能够隐式地发现卷积滤波器或注意力投影与输出类别之间的内在关系，因此可以获得可用于可解释性分析的表示。

- 我们在小型和中等规模的图像分类数据集上进行了实验，采用了基于卷积神经网络和 ViT 的架构。结果表明，我们提出的方法是有效的。

相关工作

机器遗忘旨在从预训练模型中消除特定数据或敏感数据，理想情况下无需进行完整的重新训练。在这一领域，Cao 等人 [3]率先在传统机器学习算法中研究机器遗忘问题。然而，与从头训练相比，他们的方法优势有限。

近期的研究工作主要集中于将机器遗忘引入深度神经网络。Izzo 等人 [4]提出了一种从线性模型中删除数据点的技术；其他研究则在训练期间进行数据分组，通过限制数据点对模型学习过程的影响，实现无缝机器遗忘 [10]。另一种方法是监控训练数据点对模型参数的影响，从而尽可能减少

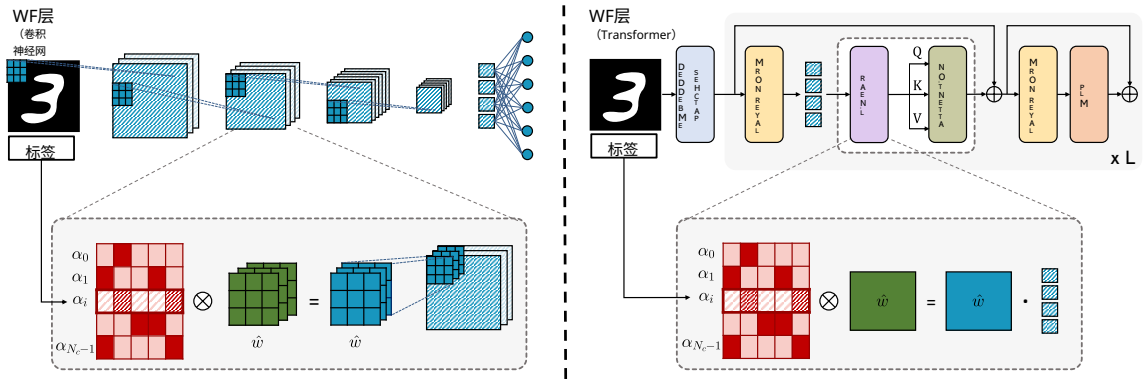

图2. 加权滤波器层在基于 CNN 和 ViT 的网络架构上实现单次多类别机器遗忘的应用。

收到删除请求时所需的重新训练量 [5]。然而，所有上述方法都会导致较高的存储成本。另一项研究提出采用高效的训练数据分区方法。在这些研究中，Ginart 等人 [11] 提出了一种从聚类中删除数据点的方法：将数据划分为独立分区，并为每个分区分别训练一个模型。

近期的机器遗忘范式涉及从预训练模型中移除整个类别 [6]–[8]。例如，Golatkar等人 [12]提出了一种通过随机梯度下降从深度学习网络的中间层移除信息的技术，该技术也可扩展至模型的最终激活 [13]。而 Tarun等人 [14], 则提出了一种有效的逐类别机器遗忘框架：学习遗忘类别的误差最大化噪声矩阵，再结合一部分模型应保留的数据来更新网络参数。

与先前工作不同，我们提出在单个反训练阶段中，以正交方式遗忘预训练模型的所有类别，从而避免保存多个遗忘模型。

提出的方法

预备知识

我们考虑一组输入数据 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$，其中包含 n 个样本；x_i 是第 i个样本，y_i 是其对应的标签，属于类别集合 $\mathcal{C}$。传统训练旨在确定一组权重 $\hat{w} \in \mathcal{W}$，其通过迭代更新规则 $w_{t+1} = w_t - g(w, x')$ 得到，其中 $g(\cdot, x')$ 是固定损失函数的随机梯度。模型收敛后，给定一组从与 $\mathcal{D}$相同的分布中抽取的数据点 $\mathcal{D}_f$，机器遗忘旨在以如下形式确定对 $\hat{w}$ 的更新

$$\hat{w}_{t+1} = \hat{w}_t + g(z'_f), \text{with } \hat{w}_0 := \hat{w} \text{ and } z'_f \sim \mathcal{D}_f \quad (1)$$

使得 $\mathcal{D}_f$ 需要被遗忘。换言之，遗忘模型最终的权重应接近于仅使用 $\mathcal{D} \setminus \mathcal{D}_f$从头训练得到的模型。根据具体应用场景， $\mathcal{D}_f$ 可以是单个数据点（即，样本移除）、一组具有相似特征或标签的数据（即，特征移除），也可以是整个类别（类别移除） [2]。

类别移除。 从训练好的模型中遗忘一个或多个类别的数据颇具挑战。对一个类别进行机器遗忘通常需要多次遍历该类别的数据点，与移除单个数据项或特征相比，计算成本较高。此外，如果需要对不同类别进行机器遗忘，就必须针对每个类别重复执行式1 所述的训练过程，因为遗忘一个类别不可避免地会使$\hat{w}$ 转向一种配置：在这种配置下，恢复先前已遗忘的类别，再遗忘新的类别，既麻烦又低效。若希望模型有选择地遗忘其训练过的所有 N_c 个类别，就需要从$\hat{w}$ 开始重复该过程 N_c 次，并存储同样数量的独立检查点。

算法1 WF模型的一般前向步骤

要求： 一个由一系列层组成的 WF 模型 M 层 $\mathcal{L}$、一个输入张量 X_{in}、一个标签 Y_{unl}

确保： 一个预测标签 $Y_{\text{out}} \leftarrow M(X_{\text{in}}, Y_{\text{unl}})$

初始化 $X_0 \leftarrow X_{\text{in}}$

对于 $i \in \mathbb{N}_{|\mathcal{L}|}$ **执行**

$l \leftarrow \mathcal{L}[i]$

$\hat{w}_l \leftarrow \boldsymbol{\alpha}_l[Y_{\text{unl}}] \odot \hat{w}_l$

$X_{i+1} \leftarrow l(X_i, \hat{w}_l)$

结束循环

返回 $\text{argmax}(X_{|\mathcal{L}|})$

单次多类别机器遗忘

利用内部组件与类别之间已知的关联 [9]，我们的机器遗忘过程促使模型恢复这些关联，并将其存储在可学习记忆矩阵中。为此，我们将网络内部组件封装在权重过滤（WF）层中，从而能够在单个反训练步骤中正交移除每个类别，并最终只需一个模型检查点。在测试时，只需选择其中一个 N_c 类别，并指示模型表现得仿佛已遗忘所选类别。概览见图 2。

直观理解与单类别场景。我们通过介绍仅针对单个类别进行机器遗忘的情形来引入 WF-Net。给定一个模型 $M: \mathcal{X} \to \mathcal{Y}$、一层 l 及其已学习参数集 $\hat{w}_l$，我们将l 封装到WF层中。随后，WF将原始层与一个形状为 $1 \times \mathcal{K}$的可学习记忆向量$\boldsymbol{\alpha}_l$ 相结合，其中 $\mathcal{K}$ 是原始可学习权重的基数。该记忆向量旨在屏蔽已学习权重 $\hat{w}_l$，并过滤传递至后续更深层的信息。 $\boldsymbol{\alpha}_l$ 的参数为实数，通过应用Sigmoid函数将其压缩至区间$[0, 1]$ 内。

具体而言， $\boldsymbol{\alpha}_l$ 调节层内权重的贡献，从而得到一个$\hat{}$经过修改的层 $l(\cdot)$ ，其行为如下

$$\hat{l}(\cdot) = l(\cdot, \alpha_l \odot \hat{w}_l). \tag{2}$$

当 $\boldsymbol{\alpha}_l$ 的所有元素都设为 1 时，修改后的层等同于原始层。同理，$\boldsymbol{\alpha}_l$ 可以通过删除部分 α 分数并将其设为 0，从而充当遗忘层。在此过程中，模型的原始参数保持冻结状态（仅训练 $\boldsymbol{\alpha}_l$ ）。回想内部组件与类别之间的依赖关系，一个良好的机器遗忘过程应当删除极小但最优的一组 α 分数，并将其设为 0，从而彻底遗忘目标遗忘类别。

多类别场景。 虽然上述方法可作为单类别情况下经典遗忘的一种替代方案，但当需要处理更多类别时，它无法解决由此产生的时间和存储问题。然而，由于我们的方法仅依赖可学习记忆向量，因此可以轻松扩展到多类别场景。将 $\boldsymbol{\alpha}_l$ 扩展为 $N_c \times \mathcal{K}$ 矩阵后，每个行向量都可以记忆类别特定信息。在前向传播时，给定一个遗忘类别标签，只需索引 $\boldsymbol{\alpha}_l$ 中对应的行并应用该层（见算法1）。最后，结合适当的反训练方案，即可通过单一流程分别对所有类别执行机器遗忘。最后值得一提的是，在计算机器遗忘损失时，也可以通过随机选择一个真实标签作为 α_l 矩阵的行选择器，将我们的方法扩展到多标签数据集。

图像架构的应用

为证明其泛化能力，我们将权重过滤应用于基于 CNN 和 Transformer 的架构。

卷积神经网络。 在基于 CNN 的架构中，我们屏蔽卷积核和偏置，并以输出通道为屏蔽粒度。对于一个包含卷积核 l $\hat{}$ $w_l^c \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times kH \times kW}$ 和偏置 $\hat{}w_l^b \in \mathbb{R}^{C_{\text{out}}}$的卷积层，我们对每个输出通道分别应用两个加权矩阵： $\boldsymbol{\alpha}_l^c \in \mathbb{R}^{N_{\text{classes}} \times C_{\text{out}}}$ 作用于 $\hat{}w_l^c$ ，$\boldsymbol{\alpha}_l^b \in \mathbb{R}^{N_{\text{classes}} \times C_{\text{out}}}$ 作用于偏置。这样便对 C_{out} 卷积滤波器进行屏蔽，其形状为 $C_{\text{in}} \times kH \times kW$ 且具有 C_{out} 个激活值，并由 $\boldsymbol{\alpha}_l^c$中选定的一行进行加权。此外，每个 C_{out} 偏置都由 $\boldsymbol{\alpha}_l^b$中选定的一行进行加权。

视觉Transformer。对于基于Transformer的网络，我们则对权重$\hat{}w_l^c \in \mathbb{R}^{F_{\text{in}} \times F_{\text{out}}}$ 和偏置$\hat{}$ $w_l^b \in \mathbb{R}^{F_{\text{out}}}$进行掩码处理

这些权重和偏置属于生成每个注意力层的查询、键和值的线性投影 l。

训练

借鉴先前工作 [15],，我们采用两个分类损失函数，即机器遗忘损失 L_f 和保留损失L_r。第一个损失鼓励模型遗忘特定类别 c_f，而第二个损失用于衡量保留有关其他所有类别信息的能力（$c_r \in \mathcal{C} \setminus c_f$）。我们将二者都实现为交叉熵损失函数。

为避免通过反转梯度符号得到的负梯度方法 [12],所带来的缺点，我们建议最小化以下目标函数，其中我们采用遗忘损失的倒数：

$$L = \lambda_0 \sum_{(x,y)\in\mathcal{D}_r} L_r\left(M(x), y\right) + \lambda_1 \sum_{(x,y)\in\mathcal{D}_f} \frac{1}{L_f\left(M(x), y\right)}. \quad (3)$$

因此，将公式 3 最小化至零，意味着应最大化 $L_f(\cdot)$，并最小化 $L_r(\cdot)$。

正则化。为确保在反训练过程中 $\boldsymbol{\alpha}_l$ 中只有少数元素被置零，我们添加一个正则化项，使 $\boldsymbol{\alpha}_l$ 中的元素几乎全部保持激活状态，只有数量有限的参数除外，这些参数将用作门控。因此，我们在最终损失函数中加入正则项 $R\left(\hat{}\alpha\right)$。

标签扩展。为了同时对所有类别进行反训练，我们不会在物理上将数据集 $\mathcal{D}$ 拆分为保留子集和机器遗忘子集 $\mathcal{D}_r$ 和 $\mathcal{D}_f$，因为数据集中的每个样本既可用于遗忘其指定类别，也可用于保留所有其他类别。相反，给定一个随机采样的小批量（长度为 B），我们将其分成两半，得到$B/2$ 图像和标签，用于机器遗忘；另一半 $B/2$ 则用于保留。第一半中的样本连同其真实标签用作 $\boldsymbol{\alpha}_l$ 矩阵的行选择器，并用于计算遗忘损失。另一半则与从 $\boldsymbol{\alpha}_l$中随机选择的行配对，并结合其真实标签用于计算保留损失。

需要注意的是，在优化过程中，类别为 c 的每张图像都会在 $\boldsymbol{\alpha}_l$ 的同一行上进行机器遗忘（例如，类别 0 会在第 0 行上进行机器遗忘）。在保留时，由于在 $\boldsymbol{\alpha}_l$中进行了随机选择，每个类别都会随机保留在其他行之一上。这意味着每个类别都会对其对应行产生全部影响，但对用于保留的各行产生的影响则小得多。为弥补这种差异，我们将小批量中的“保留”部分复制 χ 次，并与不同采样的标签配对，使每张用于保留的图像产生的影响增大 χ 倍。因此，保留损失的形状为 $(\chi,\ B/2)$，而反训练损失的形状为 $(B/2,\ )$。最后一步是对两个损失分别求平均。

实验评估

数据集

我们在三个知名的图像分类数据集上进行实验：MNIST、CIFAR-10 和 ImageNet。MNIST 包含 60,000 张训练图像和 10,000 张测试图像，每张图像对应 10 个类别之一，分别代表手写数字。CIFAR-10 包含 60,000 张图像，分为 10 个类别，每类 6,000 张图像，并划分为训练集（50,000 张）和验证集（10,000 张）。ImageNet 包含 128 万张训练图像和 50,000 张验证图像，每张图像对应数据集 1,000 个类别之一。在 ImageNet 上进行评估时，我们从验证集中随机选取 5,000 张图像，每个类别选取 5 张。

评估指标

我们从评估指标的角度评估 WF-Net，这些指标既衡量模型的机器遗忘能力，也用于估计其可解释性水平。

保留集和遗忘集上的准确率。我们使用准确率衡量模型对保留集和遗忘集的预测是否正确。理想情况下，保留集上的准确率应接近机器遗忘前原始模型的准确率，而遗忘集上的准确率应接近零，或等于一个在重新训练时未使用遗忘类别样本的模型的准确率。

重新学习时间（RT）。它衡量模型从遗忘集中的某个类别恢复至完整准确率所需的轮次数。遵循

先前工作 [14],我们每轮使用训练集中的 500 个随机样本训练模型；如果模型在 100 轮后仍未达到原始准确率，则停止训练。在我们的实验中，我们报告数据集中所有类别的平均重新学习时间，并在括号中注明模型在 100 轮后仍无法重新学习的类别数。

零重训练遗忘（ZRF）分数。通过与随机初始化的网络进行比较，它估计遗忘模型的随机性。具体而言，该指标使用 Jensen-Shannon 散度（JS）比较两个模型的输出概率。按照以往文献 [15]，，我们按如下方式计算 ZRF 分数：

$$\text{ZRF} = 1 - \frac{1}{N_f} \sum_{i=0}^{N_f} JS\left(M(x_i), M^*(x_i)\right), \quad (4)$$

其中，M 是遗忘模型，M^* 是其随机初始化版本，x_i 是遗忘集中的第 i 个样本，而 N_f 是遗忘集中的元素数量。最终分数介于 0 和 1 之间，接近 1 的分数表示 M的行为完全随机。

插入分数。 该指标用于衡量可解释性图的质量 [17]。简而言之，它通过相关性从基线（通常为 0 张量）开始，逐步重新激活图像像素，并计算置信度分数的上升幅度。在我们的案例中，$\boldsymbol{\alpha}_l$ 矩阵包含有关已遗忘类别的关键信息。与原始实现不同，在每次迭代中，我们都会按照相关性对每个 $\boldsymbol{\alpha}_l$ 中的元素排序，并重新激活其中一定比例的元素；随后评估网络的置信度分数，并将其归一化为基线模型的置信度分数，即即反训练前的对应模型。最后，取平均 AUC 分数。分数越接近 1，$\boldsymbol{\alpha}_l$ 矩阵所选特征就越重要。

删除分数。 它是插入分数的对应指标，用于衡量在每次迭代中按相关性将图像像素置零时，分数的下降幅度，直至初始图像变为基线张量。在我们的案例中，我们再次调整 $\boldsymbol{\alpha}$ 分数，以展示其值与已遗忘类别相关信息之间的相关性。为计算该分数，我们从一个未遗忘类别的 WF-Net 开始，并逐步将其中一定比例的值置为0，从相关性最高的值开始，依次到相关性最低的值。每次迭代都会记录 WF-Net 对已遗忘类别的归一化置信度分数，并将其与基线模型的置信度分数进行比较。所得曲线的平均 AUC 即为最终分数。

实现与训练细节

我们使用多种骨干网络开展研究，其模型大小适用于中等规模数据集，包括 VGG-16、ResNet-18，以及微型（ViT-T）和小型（ViT-S）版本的 ViT。

在反训练过程中，我们将所有 $\boldsymbol{\alpha}_l$ 初始化为 3，使得 $\sigma(\boldsymbol{\alpha}) \simeq 1$ 在学习初期具有非零梯度。我们采用固定的（反）学习率 100，以便摆脱预训练期间达到的极小值，并将小批量大小固定为128。对于卷积神经网络，我们采用 $\lambda_0 = 1$, λ_1 =10, $\lambda_2 = 1$。对于 ViT，我们将 λ_1 提高至100。所有情况下的标签扩展因子 $\chi = 3$，最终，我们以10的耐心值提前停止反训练过程，每轮次评估5 次验证损失。

实验结果

遗忘性能。 我们首先评估所提出方法的遗忘性能。表 1 展示了在 MNIST、CIFAR-10 和 ImageNet 训练集上的结果。我们将其与从头开始使用完整训练集训练的模型（称为“原始模型”）的性能，以及分别移除其中一个类别后训练得到的一组模型的性能进行比较。由于 ImageNet 数据集的类别数量较多，我们仅针对 MNIST 和 CIFAR-10 训练这些模型。这些模型的平均性能称为“重新训练的模型”。此外，我们还实现了两种单类别遗忘方法，即 随机标签 [16] 和负梯度 [12]。前一种方法中，我们使用为遗忘集中的样本随机分配的标签对模型进行微调；后一种方法中，我们使用负梯度在遗忘集上对模型进行微调（即，沿梯度上升方向进行微调）。值得注意的是，这两种方法的遗忘阶段都是针对每个类别分别执行的。因此，我们不报告这两种方法在 ImageNet 数据集上的结果。

表 1. MNIST、CIFAR-10 和 ImageNet-1k 上的机器遗忘结果，与原始模型和重新训练的模型进行比较。† 标记表示单类别遗忘方法。

		MNIST				CIFAR-10				ImageNet-1k		
		准确率$_r$ ↑	准确率$_f$ ↓	ZRF ↑	RT ↑	准确率$_r$ ↑	准确率$_f$ ↓	ZRF ↑	RT ↑	准确率$_r$ ↑	准确率$_f$ ↓	零遗忘率 ↑
VGG-16	原始模型	99.6	99.6	48.0	-	93.0	93.0	48.3	-	71.2	71.3	0.35
	重新训练的模型	99.4	0.0	48.7	74.5 (5)	89.9	0.0	50.1	65.3 (5)	-	-	-
	随机标签†	99.6	0.0	53.3	40.3 (0)	88.7	0.0	52.5	56.5 (5)	-	-	-
	负梯度†	99.4	0.0	48.3	24.6 (0)	89.4	0.0	51.6	76.7 (7)	-	-	-
	WF-Net	73.2	0.0	79.2	27.6 (1)	80.2	18.3	57.4	54.2 (5)	64.1	1.89	0.53
ResNet-18	原始模型	99.6	99.6	47.0	-	93.9	94.0	48.0	-	70.5	70.3	0.35
	重新训练的模型	99.4	0.0	48.7	43.6 (0)	90.5	0.0	51.4	$>$ 100 (10)	-	-	-
	随机标签†	99.4	0.0	57.1	15.7 (0)	90.6	0.0	59.4	21.6 (0)	-	-	-
	负梯度†	99.2	0.0	52.1	20.1 (0)	92.4	0.0	55.2	31.3 (3)	-	-	-
	WF-Net	94.0	9.68	63.1	22.7 (0)	79.7	9.25	63.9	94.3 (9)	64.4	1.47	0.40
ViT-T	原始模型	98.9	98.9	47.2	-	78.0	78.0	49.8	-	75.6	75.5	0.34
	重新训练的模型	99.0	0.0	50.2	71.4 (7)	71.2	0.0	67.6	52.2 (4)	-	-	-
	随机标签†	98.3	0.0	48.7	81.9 (8)	73.2	0.0	57.7	83.5 (8)	-	-	-
	负梯度†	98.4	0.0	54.6	71.6 (7)	74.4	0.0	50.1	81.7 (8)	-	-	-
	WF-Net	93.5	0.0	47.4	$>$ 100 (10)	73.5	0.0	59.8	$>$ 100 (10)	68.0	2.51	0.44
ViT-S	原始模型	99.0	98.9	47.1	-	85.2	85.2	53.6	-	82.3	82.2	0.34
	重新训练的模型	99.0	0.0	49.5	91.5 (8)	75.5	0.0	61.3	51.1 (4)	-	-	-
	随机标签†	99.5	0.0	50.4	74.1 (7)	74.8	0.0	50.8	75.0 (6)	-	-	-
	负梯度†	99.6	0.0	52.3	44.1 (4)	75.5	0.0	51.7	61.7 (5)	-	-	-
	WF-Net	94.0	0.0	48.0	91.0 (9)	74.7	0.0	59.7	81.1 (8)	69.2	9.46	0.45

可以看出，WF-Net 在所有类别和模型上都表现出良好的遗忘效果。具体而言，在 MNIST 和 CIFAR-10 上，使用 VGG-16、ViT-S 和 ViT-T 训练时，遗忘集上的准确率均为 0.0，这表明所提出的策略能够有效移除指定类别的知识，即使不直接更新模型权重，并且能够同时处理多个类别的遗忘。在保留集准确率方面，与完整模型和重新训练的模型相比，我们观察到的性能损失较小。例如，在 ImageNet-1k 数据集上，与原始模型的准确率相比，基于卷积神经网络的 WF-Net 版本在保留集上的准确率仅下降了几个百分点（例如，使用 ResNet-18 时从 70.5 降至 64.4）[1]。相比之下，基于 Transformer 网络架构的模型在保留集准确率方面的损失略为明显（例如，使用 ViT-S 时从 82.3 降至 69.2）。考察在已遗忘类别上计算的 ZRF 分数时，可以发现我们的方法所达到的性能始终高于原始模型的表现总体上与重新训练的模型相当或更好，这再次表明我们的方法能够有效实现机器遗忘。

此外，尽管单类别机器遗忘方法需要针对每个待遗忘类别分别训练，但我们的方法同样表现出色。具体而言，虽然 WF-Net 在准确率方面略逊于随机标签和负梯度，但在大多数情况下，其 ZRF 分数显著更高，进一步证实了我们方法的机器遗忘能力。至于重新学习时间，与其他所有模型相比，我们的方法重新学习已遗忘类别所需的训练轮次相当或更多，尤其是在使用基于 ViT 的模型作为骨干网络时。

插入和删除分数。在表2中，我们根据四种模型网络架构报告了插入和删除分数。修改α值会影响对属于已遗忘类别的图像所给出的分数，但不会影响属于其他类别的图像。这表明，所提出的方法能够识别出负责识别各个类别的滤波器和注意力权重。

滤波器可视化。最后，在图 3 中，我们对滤波器与……之间的关联进行了定性可视化

[1]我们还在 ResNet-50 和 ResNet-101 等更大型骨干网络上观察到了类似的结果，其保留准确率分别为 72.6 和 74.4，遗忘准确率则为 0。

表 2. WF-Net 模型的插入（Ins）和删除（Del）分数。

	MNIST		CIFAR-10		ImageNet	
	插入 ↑	删除 ↓	插入 ↑	删除 ↓	插入 ↑	删除 ↓
VGG-16	0.78	0.09	0.79	0.10	0.73	0.05
ResNet-18	0.79	0.05	0.69	0.06	0.75	0.04
ViT-T	0.90	0.07	0.88	0.05	0.60	0.04
ViT-S	0.90	0.08	0.88	0.14	0.75	0.04

滤波器（或注意力投影）与输出类别，如 α 随机选取的层在机器遗忘后的值。接下来，我们为每个类别选出排名前 10 的滤波器，并移除未至少出现在两个类别中的滤波器，以便发现类别之间的关系。可以看出，所提出的方法能够发现滤波器与类别之间的恰当关系。例如，滤波器 #332 对两个不同类别（即，狗和马）有贡献，而滤波器 #3 对类别 猫、鹿和狗有贡献。这些模式有助于更好地理解模型的行为，并揭示网络中的潜在关系。图4中还展示了其他可视化结果。

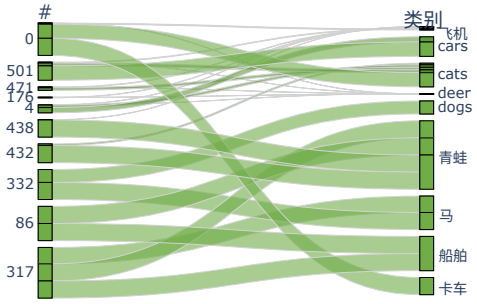

图 3. WF-Net 所揭示的 VGG-16 某一层的权重与 CIFAR-10 类别之间的关系。

结论

我们提出了一种用于单轮多类别机器遗忘的新方法。对于给定的图像分类网络，我们的方法能够在单个遗忘轮次中同时遗忘所有类别。训练后，可以指示生成的网络表现得仿佛它从未在任何一个类别上训练过。我们通过实验验证了该方法在经典图像数据集和不同网络架构上的有效性，还表明该方法能够发现卷积滤波器或注意力投影与输出类别之间的关系。

致谢

本研究得到欧盟委员会资助的欧盟地平线项目“ELIAS——面向可持续发展的欧洲人工智能灯塔”（编号：101120237）的支持。

参考文献

1. Quang-Vinh Dang. 机器学习时代的被遗忘权。载于 *ICADS*，2021年。

2. Thanh Tam Nguyen、Thanh Trung Huynh、Phi Le Nguyen、Alan Wee-Chung Liew、Hongzhi Yin 和 Quoc Viet Hung Nguyen。《机器遗忘综述》。*arXiv* 预印本 *arXiv:2209.02299*，2022年。3. Yinzhi Cao 和 Junfeng Yang。《通过机器遗忘让系统学会遗忘》。载于 *IEEE*安全与隐私研讨会，2015年。4. Zachary Izzo、Mary Anne Smart、Kamalika Chaudhuri 和 James Zou。《从机器学习模型中近似删除数据》。载于 *AISTATS*，2021年。5. Lucas Bourtoule、Varun Chandrasekaran、Christo- pher A Choquette-Choo、Hengrui Jia、Adelin Travers、Baiwu Zhang、David Lie 和 Nicolas Papernot。《机器遗忘》。载于*IEEE*安全与隐私研讨会，2021年。6. Quoc Phong Nguyen、Bryan Kian Hsiang Low 和 Patrick Jaillet。《变分贝叶斯遗忘》。载于*NeurIPS*，2020年。7. Seth Neel、Aaron Roth 和 Saeed Sharifi-Malvajerdi。《下降式删除：机器遗忘的基于梯度的方法》。载于*ALT*，2021年。8. Thomas Baumhauer、Pascal Sch¨ottle 和 Matthias Zeppelzauer。《机器遗忘：基于 logit 的分类器的线性过滤》。机器学习，111(9):3203–3226，2022年。9. Andong Wang、Wei-Ning Lee 和 Xiaojuan Qi。《HINT：分层神经元概念解释器》。载于*CVPR*，2022年。10. Yinjun Wu、Edgar Dobriban 和 Susan Davidson。《Delt-agrad：快速重新训练机器学习模型》。载于*ICML*，2020年。11. Antonio Ginart、Melody Guan、Gregory Valiant 和 James Y Zou。《让人工智能忘记你：机器学习中的数据删除》。载于*NeurIPS*，2019年。12. Aditya Golatkar、Alessandro Achille 和 Stefano Soatto。《纯净网络的永恒阳光：深度网络中的选择性遗忘》。载于 *CVPR*，2020年。

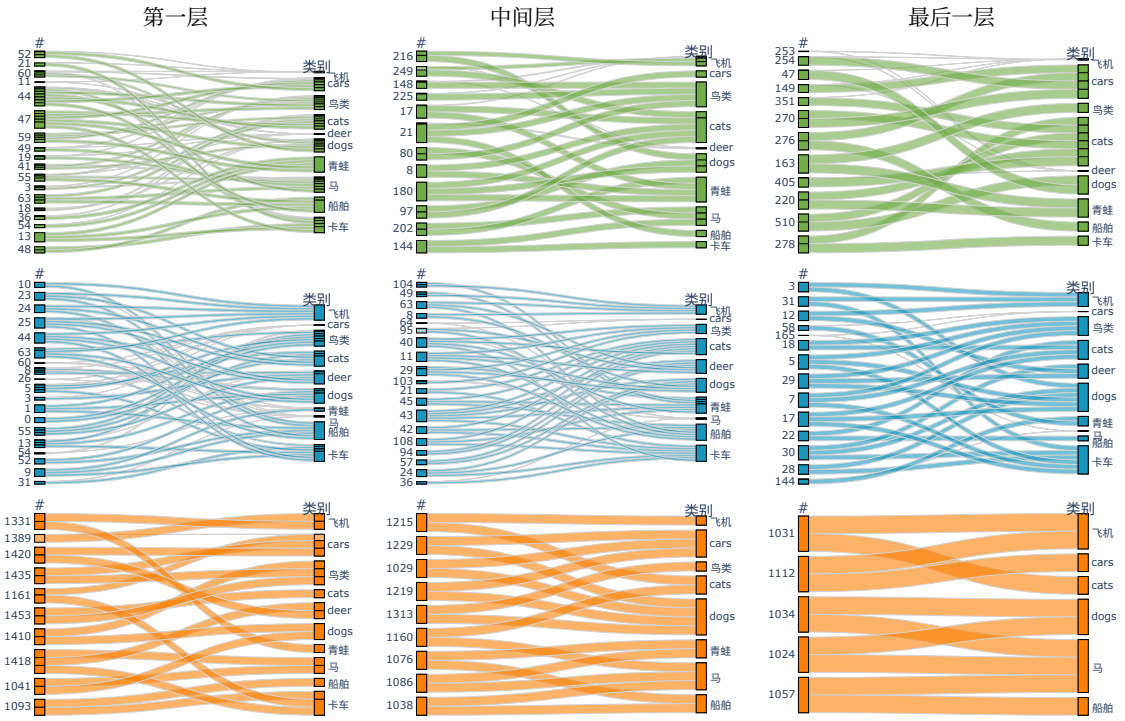

图 4. 在 CIFAR-10 数据集上，使用 VGG-16（第一行）、ResNet-18（第二行）和 ViT-T（第三行），在不同层应用加权滤波器层进行单轮多类别机器遗忘。

13. Aditya Golatkar、Alessandro Achille 和 Stefano Soatto。《框外遗忘：清除深度网络中可通过输入输出观测获取的信息》。载于*ECCV*，2020年。14. Ayush K Tarun、Vikram S Chundawat、Murari Mandal 和 Mohan Kankanhalli。《快速而有效的机器遗-忘》。*IEEE*神经网络与学习系统汇刊，2023年。15. Vikram S Chundawat、Ayush K Tarun、Murari Mandal 和 Mohan Kankanhalli。《糟糕的教学能诱发遗忘吗？使用不称职教师进行深度网络遗忘》。载于*AAAI*，2023年。16. Tomohiro Hayase、Suguru Yasutomi 和 Takashi Katoh。《在比样本更细的粒度上选择性遗忘深度网络》。*arXiv* 预印本 *arXiv:2012.11849*，2020年。17. Vitali Petsiuk、Abir Das 和 Kate Saenko。《RISE：用于解释黑箱模型的随机输入采样》。载于*BMVC*，2018年。

Samuele Poppi 是摩德纳-雷焦艾米利亚大学意大利人工智能国家博士项目的一名博士生，地址为意大利摩德纳 41125。他的研究兴趣包括可解释人工智能和机器遗忘。可通过 samuele.poppi@unimore.it 联系他。

Sara Sarto 目前正在攻读摩德纳-雷焦艾米利亚大学的信息与通信技术博士学位，意大利摩德纳 41125。她的研究兴趣包括图像描述生成、跨模态检索和注意力模型。联系邮箱：sara.sarto@unimore.it。

Marcella Cornia 是摩德纳-雷焦艾米利亚大学的一名长聘轨助理教授，地址为意大利雷焦艾米利亚 42121。她已合作发表 70 余篇论文，研究主要涉及视觉与语言融合、注意力模型和显著性模型。可通过 marcella.cornia@unimore.it 联系她。

Lorenzo Baraldi 是摩德纳-雷焦艾米利亚大学的一名长聘轨助理教授，地址为意大利摩德纳 41125。他已合作发表 100 余篇论文，研究主要涉及视频理解、深度学习和多媒体。可通过 lorenzo.baraldi@unimore.it 联系他。

Rita Cucchiara 是摩德纳-雷焦艾米利亚大学的正教授，地址为意大利摩德纳41125。她担任人工智能研究与创新中心主任，负责协调 AImageLab 研究实验室和 Modena ELLISUnit。她的研究兴趣涉及计算机视觉、模式识别和多媒体。联系方式：rita.cucchiara@unimore.it。