

看穿梯度：通过GradInversion进行图像批量恢复

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, Pavlo Molchanov NVIDIA

(Danny, amallya, avahdat, josea, jkautz, pmolchanovj@nvidia.com)

摘要

训练深度神经网络需要从数据批处理中进行梯度估计以更新参数。每个参数的梯度在一组数据上取平均值，这被认为在联合、协作和[编辑]学习应用中对保护隐私的训练是安全的。

之前的工作只显示了恢复输入的可能性

在非常严格的条件下给定梯度的数据--一个

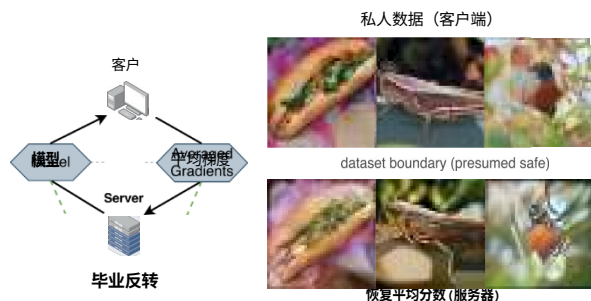
单一输入点，或一个没有非线性的网络，或一个小的 32×32 F*输入批次。因此，人们认为在较大的批次上平均分配颗粒是安全的。在此

在这项工作中，我们引入了GradInversion，利用它，在复杂的数据集如ImageNet（1000类， 224×224 px）上，也可以恢复大批量（8 - 48幅）的input图像，如ResNets（50层）。我们制定了一个优化任务，将随机噪声转换为自然

图像，在规范图像保真度的同时匹配梯度。我们还提出了一种给定梯度的目标类标签恢复算法。我们进一步提出了一个群体一致性正则化框架，其中多个代理从不同的随机种子开始，共同寻找原始数据批的增强重建。我们表明，梯度编码了令人惊讶的大量信息，因此，即使对于复杂的数据集、深度网络和大批量的数据，也可以通过GradInversion高保真地恢复所有的单个图像。

1. 简介

在训练过程中分享权重更新或梯度是深度网络协作、分布式和联合学习的核心思想[1, 22, 24, 25, 28]。在联合随机梯度下降的基本设置中，每个设备在本地数据上进行学习，并分享梯度以更新全局模型。减轻传



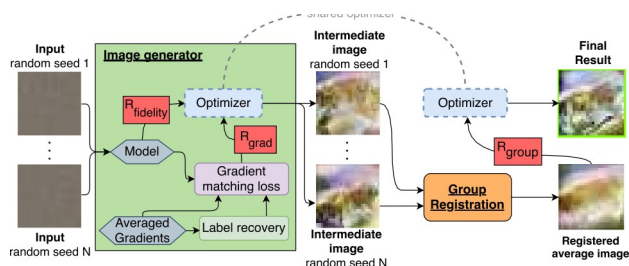
输训练数据的需要提供了几个关键优势。这可以保持用户数据的私密性，缓解与用户隐私、安全和其他所有权相关的担忧。此外，这消除了存储、传输和管理可能的大型数据集的需要。有了这个框架，人们可以在不接触任何个人数据的情况下训练医疗数据模型[32]，或感知模型

(a) 颠倒平均梯度以恢复原始图像的批次

图1: 我们提出 (a) GradInversion, 通过反转 平均梯度, 高保真地恢复隐藏的训练图像批次。GradInversion制定了 (b) 一个优化过程, 将噪声转变成输入图像 (第3.1节)。它从全连接层的梯度开始恢复标签 (第3.2节), 然后在保真度规则化 (第/*.3节) 和基于注册的组一致性规则化 (第3.4节) 下优化输入以匹配目标梯度, 提高重建质量。这使得从 ResNet-50 批次梯度中恢复 224×224 像素的 ImageNet 样本, 这在以前是不可能的 (请放大上面的例子。更多内容见第4章)。

用于自动驾驶, 而没有侵入性的数据收集[41]。虽然这种设置乍看起来很安全, 但最近的一些工作已经开始质疑联合学习的核心前提--梯度是否有可能泄露训练数据的私人信息? 有效地作为训练数据的 "代理", 梯度与数据之间的联系实际上提供了检索信息的潜力:

从揭示原始数据的位置分布[33, 44]、甚至可以从梯度中实现像素级的详细图像重建[13, 55]。尽管取得了显著的进展, 在



(b) Overview of our proposed GradInversion method

通过梯度匹配重建原始图像仍然是一项非常具有挑战性的任务--对于像ImageNet[9]这样的复杂数据集来说，成功重建高分辨率的图像对于批量大小大于1的图像来说仍然难以实现。

新兴的网络反转技术研究为这项任务提供了启示。网络反转通过在适当的损失函数上对可学习的输入进行反向传播梯度，实现噪声到图像的转换。最初的解决方案仅限于浅层网络和低分辨率的合成[11, 34]，或创造一种艺术效果[37]。然而，该领域已迅速发展，使ImageNet上的高保真、高分辨率的图像合成来自普遍训练的分类器，使下游任务的剪枝、量化、持续学习、知识转换等无数据。[5, 17, 42, 48]。其中，DeepInversion[48]在ImageNet的图像合成方面取得了最先进的成果。它通过批量归一化（BN）先验，对特征分布进行正则化处理，使现实数据从一个虚无的预训练的ResNet-50[14]分类器中合成。

在DeepInversion[48]的基础上，我们深入研究了通过梯度反演进行批量恢复的问题。我们将任务表述为对输入数据进行优化，使该数据的梯度与客户提供的梯度相匹配，同时确保输入数据的真实性。然而，由于梯度也是地面真实标签的函数，主要的挑战之一是如何识别批次中每个数据点的地面真实标签。为了解决这个问题，我们提出了一种一次性的批量标签恢复算法，该算法使用最后一个全连接层的梯度。

我们的目标是恢复客户所拥有的准确图像。通过从不同的随机种子产生的嘈杂输入开始，多个优化过程可能会收敛到不同的最小值。由于卷积神经网络（CNN）固有的空间不变性，这些产生的图像共享空间信息，但在确切的位置和排列上有所不同。为了使收敛性更好地接近地面真实图像，我们从所有候选图像中计算出一个注册的平均图像，并在每个优化过程中引入一个组一致性正则化项以减少偏差。我们发现，与之前的优化方法[13, 15]相比，建议的方法和组一致性正则化提供了更好的图像恢复。

与BigGAN[d]等最先进的生成对抗网络（GAN）相比

，我们基于非学习的图像恢复方法能够恢复隐藏输入数据的更多具体细节。更重要的是，我们证明了通过反转批次的梯度，完全恢复具有高保真度和视觉细节的224 x 224 px分辨率的单个图像，现在甚至可以达到48幅图像的批次大小。

我们的主要贡献有以下几点：

- 我们引入GradInversion来恢复隐藏的原始

通过给定批量平均梯度的优化，从随机噪声中获得图像。

- 我们提出了一种标签修复方法，利用最终的全连接层梯度恢复地面真实标签。
- 我们引入了一个基于多种子优化和图像注册的组别一致性正则化项，以提高重建质量。
- 我们证明，对于ResNet-50这样的深度网络来说，从批量平均梯度中完全恢复详细的个体图像现在是可行的。
- 我们介绍了一种新的 *图像识别精度* 方法，以衡量不同批次的反转难易程度，并识别容易被反转的样品。

2. 相关工作

图像合成。 GANs[15, 23, 36, 38, 50]已经为生成性图像建模提供了最先进的结果，例如，ImageNet上的BigGAN-deep[4]。然而，训练GAN的生成器，需要获得原始数据。多项工作也研究了只给一个预先训练好的模型来训练GAN[C, 34]，但结果是图像缺乏细节或感知上与原始数据的相似性。

之前的安全工作研究了来自预训练的单一网络的图像合成。Fredrikson等人的 *模型反转攻击* [11]优化了输入，利用目标模型的梯度获得类图像。后续工作[20, 46, 47]扩展到新的威胁场景，但仍然限于浅层网络。The *Secret Revealer* [12]利用辅助数据集的先验，训练GAN来指导反转，将攻击扩展到现代架构，但在样本种类较少的数据集上，如MNIST和人脸识别。虽然最初的目的是了解网络特性，但可视化技术为从网络中生成图像提供了另一种可行的选择。Mahendran et al.[31]探索反转、激活最大化和漫画化，从训练好的网络中合成“自然的预图像”[30, 31]。Nguyen等人使用全局生成先验来帮助反转训练好的网络[29]的图像，其次是Plug & Play[38]，通过潜在先验提高图像多样性和质量。这些方法仍然依赖于辅助数据集信息、特征嵌入或改变的训练。

最近的努力集中在从预先训练好的网络中生成图像，没有任何辅助信息。Mordvintsev et al.的Deep-Dream[37]暗示了利用输入的梯度在图像上“梦见”新的视觉特征，可扩展到噪声到图像的转换。Saturkar et al.[42]将该方法扩展到更真实的图像。最近的扩展[5, 48]显著提高了现成分类器的图像合成性能，没有辅助信息也没有额外的训练，而是依靠BN统计。

基于梯度的反演。早期曾有过一些尝试

为了追求原始数据的代理信息，例如，存在某些训练样本[-33, -14]或数据集的样本属性[21, 44]，对梯度进行反转。这些方法主要是针对非常浅的网络。

一个更具挑战性的任务旨在从梯度重建前行为图像。Phong *et al.*[21]的早期尝试通过展示单神经元或单层网络上可证明的重建可行性，为这项任务带来了理论上的见解。Wang 等人[46]根据经验从4层网络的梯度中反转出单幅图像的表示。沿着同一思路，Zhu 等人[55]通过联合优化“伪”标签和输入以匹配目标梯度，将梯度反演推向了更深的架构。该方法导致了精确到像素级的重建，同时仍然限于连续模型（例如，用sigmoid代替ReLU的模型）而没有任何进展，并且可以扩展到低分辨率的CIFAR数据集。Zhao *et al.*[3]用标签修复步骤扩展了该方法，从而提高了单幅图像重建的速度。Geiping 等人最近的工作[13]首次将边界推向了ImageNet级别的梯度反演--它从梯度中重建了单幅图像。尽管取得了显著的进展，但当梯度被平均化时，该领域在ImageNet上对任何大于1的批次大小都很吃力。

3. 毕业反转

在本节中，我们将详细解释GradInversion。我们首先将从梯度中重建输入的问题设定为一个优化过程。然后，我们解释了我们的批量标签恢复方法，接着是用于确保真实性和群体一致性正则化的辅助损失。

3.1. 目标函数

给出一个具有权重 \mathbf{W} 的网络和一个从基础事实批次中计算出的批次平均梯度 $\Delta \mathbf{W}$ ，该梯度为图像 \mathbf{x}^* 和标签 \mathbf{y}^* ，我们的优化求解为

$$\hat{\mathbf{x}}^* = \operatorname{argmin}_{\mathbf{x}} \mathcal{L}_{\text{grad}}(\hat{\mathbf{x}}; \mathbf{W}, \Delta \mathbf{W}) + \mathcal{R}_{\text{aux}}(\hat{\mathbf{x}}), \quad (1)$$

其中 $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ (K, U, H, W 为批次大小、颜色通道数、高度、宽度) 是一个“合成”的输入批次，初始化为随机噪声，并向地面真相 \mathbf{x}^* 优化。 $\text{grad}()$ 强制匹配该合成数据的梯度(对网络的原始损失)。

$\mathcal{L}_{\text{grad}}(\hat{\mathbf{x}}; \mathbf{W}, \Delta \mathbf{W}) = \alpha_G \sum_l \|\nabla_{\mathbf{W}^{(l)}} \mathcal{L}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) - \Delta \mathbf{W}^{(l)}\|_2$, (3)
与权重 \mathbf{W} 与提供的梯度 $\Delta \mathbf{W}$ 。这是由基于图像保真度和

其中 $\nabla_{\mathbf{W}^{(l)}} \mathcal{L}(\mathbf{x}^*, \mathbf{y}^*)$ 指的是第 l 层的地面真相梯度，而以 G 为尺度的求和则贯穿所有层。这里缺少的一个关键部分是

\mathbf{y} ，启动反向传播。我们接下来解释一种有效的算法，从全连接分类层的梯度中恢复批判性的标签。

3.2. 批量修复标签

考虑到分类任务的交叉熵损失，批量大小为 N 的 $\mathbf{x}^* = [\mathbf{x}_1^*, \dots, \mathbf{x}_N^*]$ 的地面真相梯度可以被分解为：

$$\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{x}^*, \mathbf{y}^*) = \frac{1}{K} \sum_k \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{x}_k, \mathbf{y}_k), \quad (4)$$

其中 $\mathbf{z} = (\mathbf{p}, \mathbf{c})$ 表示一个原始图像/标签对。对于每个图像 \mathbf{z} ，在索引 n 处的网络最终对数 c 的梯度是 $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{z}, \mathbf{p}) = \mathbf{p}_n - \mathbf{p}$ ，其中 \mathbf{p}_n 是范围为(0, 1)的后最大概率， \mathbf{p}_k 是 \mathbf{p}_k 在索引 n 处在 N 个总类中的二元呈现。因此，这使得符号 $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{z}, \mathbf{p}_n)$ 为负数 *iff* $n = n_d$ at the ground truth index, and positive other-wise. 然而，我们无法获得 $\nabla_{\mathbf{W}} \mathcal{L}(\mathbf{z}, \mathbf{p}_n)$ ，因为梯度只在模型参数上给出。

用 \mathbf{W}^{FC} 表示最终全连接分类层的参数， M 是嵌入特征的数量， N 是目标特征的数量。

定义 $\nabla_{\mathbf{W}^{\text{FC}}} \mathcal{L}(\mathbf{z}, \mathbf{p})$ 为梯度。

在 \mathbf{W}^{FC} 中，图像 \mathbf{z} 的训练损失，连接特征 \mathbf{r}_n 和logits \mathbf{p}_n 。我们只得到张量 $\nabla_{\mathbf{W}^{\text{FC}}} \mathcal{L}$ 沿批次维度 k 的平均数。使用连锁规则，我们有：

$$\Delta \mathbf{W}_{m,n,k}^{(\text{FC})} = \nabla_{z_{n,k}} \mathcal{L}(x_k, y_k) \frac{\partial z_{n,k}}{\partial w_{m,n}}. \quad (5)$$

$$\frac{\partial z_{n,k}}{\partial w_{m,n}} = o_{m,k}$$

请注意 其中 $o_{m,k}$ 是指在一个叫 $\mathbf{R}^{n \times h}$ 的输入中

群体一致性正则化的辅助正则化 \mathcal{R}_{aux} 所增强的、

$$\mathcal{R}_{\text{aux}}(\mathbf{x}) = \lambda \sum_l \|\mathbf{x}^{(l)} - \mathbf{x}^{(l-1)}\|_2 \quad (2)$$

接下来，我们将对每个术语进行单独阐述。对于梯度匹配，我们最小化合成图像 \mathbf{x} 上的梯度和地面真实梯度之间的 $2z$ 距离：

全连接层，也是前一层的 m^{th} 输出。如果前一层有常用的激活函数，如ReLU或sigmoid， σ ，"总是非负的。这就通过新的信息指标的迹象暗示了目标标签的存在：

$$S_{n,k} := \sum_m \Delta \mathbf{W}_{m,n,k}^{(\text{FC})} = \sum_m \underbrace{\nabla_{z_{n,k}} \mathcal{L}(x_k, y_k)}_{\text{neg. iff } n = n_k^*} \times$$
(6)

其中， $S = (S_q)$ 是一个 $N \times N$ 矩阵，通过沿特征维度对张量 $\text{fiW}^{(\text{FC})}$ 进行求和而构建的

有趣的是， S 包含每个实例的地面真实标签的负值。因此， S 的 k^{th} 列可以用来恢复 k^{th} 的地面真实标签图像，只需确定负数条目的索引。Zhao et al.[53]探讨了这一规则用于单幅图像的标签修复。然而，在我们的研究中，我们没有机会接触到 S

多样本批量设置，因为给定的梯度是所有图像的平均数。

受此启发，我们通过对S的列进行平均来定义N维的批处理水平向量 $s = (s^n)$ ：

$$s_n = \frac{1}{K} \sum_k \sum_m \underbrace{\Delta w_{m,n,k}^{(PC)}}_{\text{在 } AW^{m \times n \times c} \text{ 中给出}}. \quad (7)$$

s 的吸引人的特性是，它可以通过沿特征维度求和，轻松计算出全连接层的给定梯度，如公式7右侧所示。

如上所述， S 中的每一列都是一个矢量，包含在标签索引处有一个单一的负峰值，否则就是正值。由于矢量 s 是 S 列的线性叠加，来自批次中的所有单个图像 $z, 's$ ，这种信息在求和时可能会丢失。然而，我们从经验上观察到，编码的位置往往是正... 较大的量级 $|Sq^*_k|$ $Sqyq^*_k$ 。当求和过程中从其他图像中引入正值时，这就使负号基本保持不变。

$\mathcal{R}_{\text{fidelity}}(\hat{x}) = \alpha_{TV} \mathcal{R}_{TV}(\hat{x}) + \alpha_{\ell_2} \mathcal{R}_{\ell_2}(\hat{x}) + \alpha_{BN} \mathcal{R}_{BN}(\hat{x}), (9)$ 列的最小值，而不是沿特征维度的求和来计算 s ：要使沿特征维度的求和为负，至少要有有一个位置为负，反之则不然。这进一步提高了标签恢复的准确性，尤其是在批量大的时候。因此，我们将批量大小为 K 的最终标签修复算法制定为：

$$\hat{y} = \arg \text{sort} \left(\min_m \nabla_{w_{m,n}^{(FC)}} \mathcal{L}(x^*, y^*) \right) [: K], \quad (8)$$

与 m 对应的是全连接层之前的特征嵌入维度。得到的 y 支持公式，在随后的 x 优化中追求 x^* 。所提方法的一个局限性是，它假定批次中的标签不重复，这对于随机抽样的 N 大小的批次来说通常是成立的，而 N 大小远远小于 ImageNet 的 1000 类的数量。

即使有正确的 y^* ，找到全局最小值的 $\text{sgd}()$ 仍然具有挑战性。该任务受限不足，由于非线性和集合层而遭受信息损失，并且只有一个正确的解决方案 [3, 5n]。我们接下来介绍基于保真度和群体一致性正则化的 $\text{aux}(-)$ ，以帮助进行这种优化。

3.3. 逼真度（现实主义）正则化

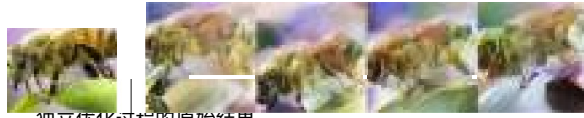


图2：单路径优化中的重建变化，集中在8个大小的批次中的一个目标。优化遵循完全相同的损失超参数，只给了不同的随机种子用于 x 的像素级初始化。

其中 J_{TV} 和 R_t 表示标准图像先验 [3, 7, 40]，对 x 的总方差和 l_2 规范进行惩罚，与缩放系数 α_{TV} 相同。DeepInversion 的关键见解在于利用 BN 统计中的强先验：

$$BN() = || (x) - BNC(\text{平均值}) ||^2 \quad (10)$$

$$\sum_l v^2(x) - BNC(\text{方差}) z,$$

其中 (x) 和 $\sigma_l^2(x)$ 是分批的平均数和方差

我们使用 DeepInversion [4S] 中提出的强先验来指导对自然图像的优化。具体来说，我们在损失函数中加入保真度 (\cdot) ，以引导 x 远离没有可辨识的视觉信息的不真实的图像：

对应于/"卷积层的特征图的估计。通过在所有层面强制执行有效的中间分布，<ndelity(')产生了向现实的解决方案的收敛。

3.4. 群体一致性正则化

基于梯度的反转的另一个挑战在于目标对象的准确定位，这是由于CNN的跨区域不变性。与理想情况不同的是，在优化收敛到一个基本事实时，我们观察到，当用不同的种子重复优化时，例如在图2中，每个优化过程都会出现一个局部最小值，在所有层面上分配语义正确的图像特征，但与其他不同--图像围绕基本事实移动，关注的细节略有不同。在前向传递过程中，池化层、分层卷积和零填充的存在，共同导致了修复后图像的空间等值，Geiping 等人也观察到了这一点[1 3]。然而，来自不同种子的修复图像的组合，暗示着有可能更好地修复出更接近地面真相的最终图像。

我们引入了一个群体一致性正则化术语，以联合优化的方式同时利用多个种子，如图6所示。直观地说，在梯度下降过程中，多路径的联合探索可以扩展和扩大搜索空间。然而，鉴于对单一目标的搜索，我们必须对其进行规范化处理，以防止出现过多的分歧，至少在最后阶段是如此。我们使用目标公式1来优化每个输入。为了促进信息交流，我们用一个新的群体一致性正则化项同时正则化所有的输入：

$$\mathcal{R}_{\text{group}}(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{g \in G}) = \alpha_{\text{group}} \|\hat{\mathbf{x}} - \mathbb{E}(\hat{\mathbf{x}}_{g \in G})\|_2, \quad (11)$$

共同考虑所有的图像候选者，在所有的

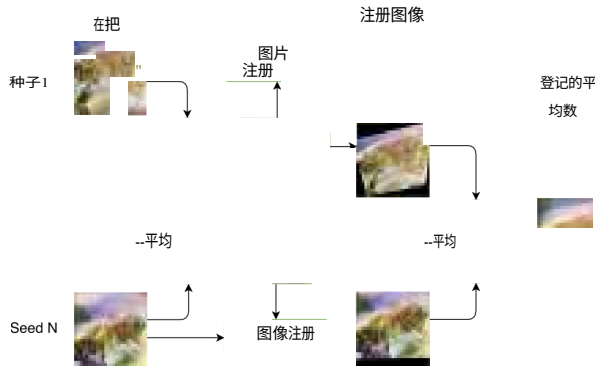


图3：群体一致性正则化的概述。

种子，并惩罚任何候选人 x ，一旦它偏离了小组的 "共识 " 图像 $E(xq,p)$ 。

$IE(xyty)$ 的一个快速而直观的选择是像素平均法。虽然看起来很 "懒"，但像素平均法已经通过混合组内所有种子的信息和反馈，带来了视觉上的改善，我们将在后面展示。为了进一步探索种子间的基本转换，创造更好的共识，我们加入了图像注册，以改善 $IN(xq,p)$ ：

$$\mathbb{E}(\hat{\mathbf{x}}_{g \in G}) = \frac{1}{|G|} \sum_g \mathbf{F}_{\hat{\mathbf{x}}_g \rightarrow \frac{1}{|G|} \sum_g \hat{\mathbf{x}}_g}(\hat{\mathbf{x}}_g). \quad (12)$$

这导致了我们的最终群体一致性正则化，如图3所示。我们(i)首先计算大小为 $|G|$ 的候选集中的像素平均数作为粗略的注册目标，(ii)通过 $F()$ 将每个单独的图像注册到target上，(iii)获得注册后的平均数作为正则化的目标。我们使用RANSAC-flow[43]进行 $F()$ 。正如我们将在后面展示的那样，分组一致性正则化能够在各种评估指标中实现恢复的持续改进，进一步缩小重建批次和原始批次之间的差距。

3.5. 最后的更新

利用所有上述损失，我们以迭代的方式更新输入。为了进一步鼓励探索和diversity，我们在每次更新中加入像素级的随机高斯噪声，灵感来自于基于能量的模型中的朗文更新[10, 12, 15]。我们的最终优化步骤是：

$$\begin{aligned} \Delta_{\hat{\mathbf{x}}^{(t)}} &\leftarrow \nabla_{\hat{\mathbf{x}}} (\mathcal{L}_{\text{grad}}(\hat{\mathbf{x}}^{(t-1)}, \nabla \mathbf{W}) + \mathcal{R}_{\text{aux}}(\hat{\mathbf{x}}^{(t-1)})) \\ \mathbf{q} &\leftarrow \mathcal{M}(0, 1) \\ \hat{\mathbf{x}}^{(t)} &\leftarrow \hat{\mathbf{x}}^{(t-1)} + (\hat{\mathbf{x}}^{(t-1)} - \hat{\mathbf{x}}^{(t-1)}) \mathbf{a}^{(t)} + (\hat{\mathbf{x}}^{(t-1)} - \hat{\mathbf{x}}^{(t-1)}) \mathbf{q} \end{aligned}$$

评估我们方法的每个组成部分的贡献。然后，我们展示了GradInversion的成功，并与现有技术进行比较。最后，我们增加了批量大小，以探索梯度反演的极限。

其中 fiqt' 对应于优化器的更新， p 表示随机抽样的噪声以鼓励探索， $A(t)$ 是学习率， aq 对最后加入的噪声进行了重新调整。

4. 实验

我们在 224×224 像素的大规模 1000 级 ImageNet ILSVRC 2012数据集[?]上评估了我们的分类任务。我们首先进行了一些消融，以

实施细节。在所有情况下，图像像素是由高斯噪声 $\epsilon=0$ 和 $w=1$ 的*i.i.d.*初始化的。我们主要关注用于分类任务的ResNet-50架构，用MOCO V2预训练，只对分类层进行了微调，在ImageNet上取得了71.0%的最高准确率。我们观察到，与默认的预训练PyTorch模型和较浅的网络结构（ResNet-18）相比，更强的特征提取导致了更好的修复效果。我们使用Adam进行优化，学习率为0.1，余弦学习率衰减，50次迭代作为热身。我们使用 $\eta_{\text{base}} = 1 \cdot 10^{-4}$, $\eta_{\text{lr}} = 1 \cdot 10^{-5}$, BN'。

0-1y G' 0-001) 组 ' 0- 01s oq = 0.2作为损失比例常数。对于特征分布正则化，我们主要关注目标批次的BN统计数据与梯度共同提供的情况，这在分布式学习中通常需要全局BN更新[29, 40, 54]。我们还分析了对网络BN均值和方差的正则化--对数据集的平均化，它们提供了单一批次统计数据的代理。我们使用NVIDIA V100 GPU和自动混合精度（AMP）[35]加速器合成分辨率为224 x 224的图像批次。每个批次的优化需要消耗20K次优化迭代。

评价指标。我们展示了在不同设置下获得的图像的视觉比较，并评估了图像相似性的三个定量指标。为了说明像素上的不匹配，我们计算了：（i）FFT D频率响应的余弦相似度，（ii）登记后的PSNR，以及（iii）重建图像和原始图像之间的LPIPS感知相似度得分[51]。

4.1. 消融研究

4.1.1 标签修复

我们首先从完全连接的层的梯度中恢复标签。表1总结了在ImageNet训练集和验证集上的平均标签恢复准确率，给定10K个随机抽取的样本，分成不同的批次大小。在一个零点的方法中，GradInversion准确地恢复了原始标签，改进了现有技术[53]。

4.1.2 批量重建

接下来，我们逐渐将每个提议的损失添加到优化过程中

。在这里，我们专注于一批8张图像的算法消融，然后再向更大的批次规模扩展。我们在表5中总结了结果，并在接下来讨论见解：添加 E_{and} 我们发现22个损失比余弦相似[13]的梯度匹配要好--详见附录。

¹ Based on hL ps : / MOCO V2 (Chen et al. [8]) 用SimCLR (Chen et al. [7])增强了MOCO (He et al. [11])，并报告ImageNet top-1准确率为71.1dc [8]。

Batch 尺寸	标签修复的准确性 (%)			
	训练集 [53]	我们的 的	验证集 [53]	我们的
1	100.0	100.0	100.0	100.0
8	90.89	99.06	96.08	99.47
32	89.88	99.29	90.32	99.19
64	84.51	98.79	82.27	98.21
96	80.53	97.88	82.13	98.11

表1：来自ImageNet训练/验证集的不同批次大小的10K个随机样本的平均修复精度，没有标签重复。': 原始方法[53]只适用于单幅图像--我们通过对公式7采用其总和规则来扩展它，然后显示改进。

意思是说。功能 "g	" (ñ- q)	与原始图像的距离		
		FFT2D	PSNR }	辽宁省
$\mathcal{N}(0, \mathcal{I})$	8.625	0.706	9.964	1.351
	4.190	0.404	10.753	0.919
	3.206	0.279	12.058	0.655
$\mathcal{L}_{\text{grad}}$	2.918	0.233	12.261	0578
+ $\mathcal{R}_{\text{fidelity}}$	2.685	0175	12.929	0484
+ $\mathcal{R}_{\text{group.lazy}}$				
+ $\mathcal{R}_{\text{group.reg}}$				

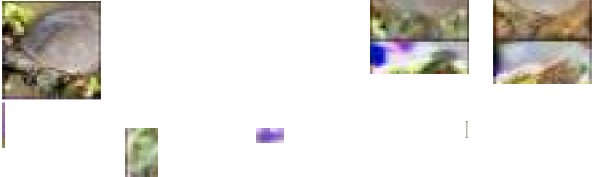


表2：消融研究当每个建议的损失来优化目标函数--定量（上）和定性（下）的比较。原始批次包含8个样本--由于空间限制，我们在此展示4个样本，整个批次见附录。

模型	与原始图像的距离		
	FFT2DI	PSNR 1	LPIPS }
ResNet-50(MOCO V2)	0.17a	12.529	0.484
ResNet-50(标准)	0.204	11.771	0.684
ResNet-18 (标准)	0.218	10.729	0.693

表3：在不同的特征提取强度下的重建。

重建的图像仍然有噪音。部分原始特征出现了，但在批次内的图像之间会有泄漏。

增加保真度--增加保真度正则化可立即提高图像质量。在图像先验的条件下，梯度反转开始将视觉细节分配给个体。

视觉上和数量上都有了很大的提高，在表S中。

添加Jgppppp-组一致性正则化进一步改善了重建。在这个分析中，我们使用了8个随机种子，每个种子确定一个高斯初始化的输入和相关的像素级扰动。所有的种子都是

联合优化，与标准的多节点训练管道兼容，该管道只支持 $E(J_q)$ 的同步计算。为了更好地了解情况，我们接下来比较了(i) "懒惰的"像素平均数和(ii)注册增强的平均数作为正则化目标的选择。

a) 懒惰的正则化。我们观察到 "懒惰的"像素平均数作为正则化的目标已经带来了性能上的提高。尽管还没有考虑到种子间的变化，但像素平均数暗示了目标物体的正确 "感知"位置。物体开始出现在正确的位置，方向也得到了改善。

b) 登记增强。然后我们加入注册，以利用候选人之间的共识。我们在SK初始优化迭代后开始注册，以允许足够的特征出现，然后每100次迭代。理想情况下，每个候选者都应被注册到其原始图像上以获得最佳的空间调整。虽然没有这样的机会，但注册到像素级的平均值是有效的。在这个阶段，GradInversion从平均梯度中准确地将详细的原始内容分配给各个图像。**反转不同的网络。**我们观察到，来自更强的特征提取器的梯度会泄露更多信息--见表3的快速比较。与相同的ResNet-50架构的标准训练方案和较弱的ResNet-18相比，ResNet-50的自我监督预训练导致了最好的图像重建。我们继续用ResNet-50 MOCO V2进行分析，研究梯度反演下的批量重建的极限。

4.2. 与最先进的技术比较

接下来，我们与现有技术在224x224px的8张图像的批量大小上进行比较。我们总结了定性（图4）和定量的结果（表4）。我们与三种可行的图像合成方法进行比较：

(i) 梯度反演[13, 55]：我们首先与之前的梯度匹配的模型反演方法进行比较：（i）Zhu等人的深度梯度泄漏方法[55]和（ii）Geiping等人的联合梯度反演[13]。我们首先按照作者的公共开源资源库[14, 56]将这两种技术扩展到ImageNet批量修复。为了进行额外的公平比较，我们还在图5中与这两种方法在批量大小为1时进行了比较，并显示出明显的保真度和定位改进。

(ii) DeepInversion[48]：我们还分析了与基线DeepInversion方法相比的性能提升，该方法以地面真实标签为条件合成图像。

(iii) GAN潜伏空间投影[23]：我们最后与基于GAN的潜伏代码优化方法进行比较。我们应用了StyleGAN2[23]中的潜代码投影，用于分辨率为256 x 256的BigGAN-deep generator[4]。鉴于无法获得原始图像的投影损失[23]，我们



图4: ResNet-50的ImageNet批量梯度反演与最先进方法的视觉对比。GradInversion的标签按升序重新排列, 与标签修复后的地面真相相匹配, 准确率为100%。最好以彩色观看。

目标损失建立在合成的和地面真相梯度。

GradInversion在视觉上(图4)和数字上(表I)都优于现有技术。在没有标签修复的情况下, 寻求图像-标签对的联合优化[55]在ImageNet上很难收敛, [13]也观察到了这一点, 即使在批量大小为1的情况下。如[13]中的总变异先验和大小不一的损失有助于改善重建, 但仍然太弱, 无法指导优化走向地面真相。深度反演[48]的基线如期改善了图像的保真度, 但反演的图像与原始批次几乎没有可观察的联系。投射到BigGAN的潜空间提供了图像保真度和恢复的细节之间的平衡, 但

[27]之间的距离, 并错过了视觉细节[23]

在原始梯度而不是原始图像的较弱指导下显得不足, 因为投影到潜空间是

4.3. 扩大批次规模的效果

我们接下来增加批次大小。我们目前的分析使用32GB的NVIDIA V100 GPU扩展到批次大小48。如图6所示，随着批量大小的增加，可恢复的图像内容的数量逐渐减少。如图所示，在一个批次中，更多的梯度信息平均化能更好地保护单个图像的隐私。令人惊讶的是，GradInversion在批次大小为48的情况下，仍能揭示出相当数量的原始梯度信息，有时还能进行可行的完整重建，如图7所示。

图像可识别性精度（IIP）。我们制定了一个新的分数，衡量梯度反转所揭示的 "图像特定 "特征的数量。直观地讲，这可以衡量

方法	要求	与原始图像的距离			
	y*	GAN	FFTzn i	PSNR I	LPIPS t
噪声N(0.I)(起始点)			0.706	9.964	1.3ñ1
BigGAN-deep的潜伏投影 (Karras ct al. CVPR'20 [23]) (Brock ct al. ICLR'19 [4]) 。			0.275	10.149	0.722
DeepInversion (Yin ct al. CVPR'20 [4S])			0.238	10.131	0.728
倒置梯度 (Geiping ct al. NeurIPS'20 [1?])			0 365	11 703	0 749
深度梯度泄漏 (Zhu ct al. NeurIPS'19 [ii]) 。			0 602	10 2ñ2	1 319
GradInversion - BN2ppp%oj (我们的)			0.232	11.235	0.633
GradInversion - BN""t (我们的) 。			0.175	12.929	0.484

表4：GradInversion与ImageNetK上ResNet-50梯度反演的最先进方法的比较。BNqpppgg表示从原始数据集学习的网络中的正则化BN统计量；BN "表示在分布式设置中共享（或 泄露）的目标批次的BN统计量，用于全局BN更新，*例如*同步 Batch Normalization [4

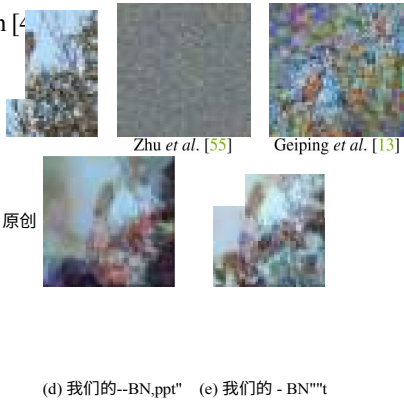


图5：与现有技术中ResNet-50 (ImageNet)在批量大小为1时对 [1?]中的 "挑战性 "样本进行重力反演的比较。

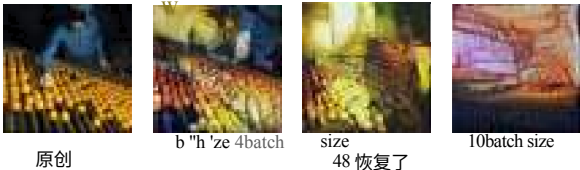


图6：还原的原始视觉特征的数量减少，因为批量大小增加。

在原始数据集中的所有类似图像中，仅给定其重建，识别一个特定的图像有多容易。我们以量化的方式计算原始图像和与其重建的最近的邻居之间完全匹配的部分。由此产生的指标，被称为*图像识别精度 (IIP)*，评估了不同批次的梯度反转强度。图S绘制了GradInversion的IIP曲线。正如预期的那样，随着批次大小的增加，重构效率逐渐下降，这在图6中也可以看到。我们观察到一个令人惊讶的现象，许多样本（28%）甚至在平均了48张图像的梯度后也能被正确识别。

脆弱的人群。我们从经验上观察到重建效果和梯度大小之间的正相关关系。深入研究这一观察结果，我们确定了一组新的图像，这些图像在GradInversion下更 "容易 "泄露



信息。为此，我们在ImageNet的每个类别中找出一张图像，其梯度规范是该类别文件夹中最大的。与图8中的随机图像相比，从这种 "易受攻击的群体 "中取样的批次大大增加了IIP，在批次大小为48时几乎翻了一番。这就要求我们仔细关注



图7：在批量大小为48的情况下，信息泄露的程度不同在ImageNet验证集上。每个区块包含一对（左）原始样本和其（右）通过GradInversion进行的重建。

图8：Grad-Inversion在ImageNet验证集上的图像可识别性精度（IIP）曲线，作为增加批次大小的函数。每一点都是按256个随机选择的不同批次大小的样本计算的平均数（批次大小为48的样本为240）。在avgpool特征空间余弦相似度中测量的最近的邻居。

在梯度共享之前，对这种脆弱的样本。

结论

我们引入了GradInversion，在给定平均梯度的情况下，在一个批次中重建单个图像。我们表明，在复杂的数据集上分享深度网络的梯度时，即使在大批量的情况下，隐私的假设也不成立。这为保护隐私的深度学习框架的发展提供了新的见解。

研究使原始数据从梯度中恢复的信息传输的基本机制也是富有成效的。我们希望未来的工作可以研究基于aggregation的联合学习[2]的脆弱性，以及进一步加强它们以防止反转。