



2020年9

月
2003. 14053v2 [cs.CV]Xi
+

1 绪论

联合学习或协作学习[7, 31]是一种分布式学习范式，随着机器学习中的数据要求和隐私问题不断增加，这种范式最近得到了极大的关注[24, 16, 35]。其基本思想是训练机器学习模型，例如神经网络，通过使用损失函数 \mathcal{L} 优化网络的参数 θ ，由输入图像 z 和相应的标签 p 组成的示范性训练数据，以解决

$$\min_{\theta} \sum_{i=1} \mathcal{L}_{\theta}(x_i, y_i). \quad (1)$$

我们考虑一个分布式的环境，在这个环境中，一个服务器想在多个“小伙伴”的帮助下解决 (1)

倒置梯度--打破隐私有多容易 在联合学习中？

乔纳斯-盖平*

Hartmut Bauermeister *

汉娜-德洛格 *

迈克尔-莫勒

塞根大学电子工程和计算机科学系

{jonas.geiping, hartmut.bauermeister, hannah.droege,
michael.moeller}@uni-siegen.de

摘要

联合学习的理念是在服务器上协同训练一个神经网络。每个用户都会收到网络的当前权重，并依次发送基于本地数据的参数更新（梯度）。这个协议的设计不仅是为了高效地训练神经网络，而且还为用户提供了隐私保护，因为他们的输入数据仍然在设备上，只有参数梯度被共享。但共享参数梯度的安全性如何？以前的攻击提供了一种虚假的安全感，因为它只在设计好的环境中成功——甚至是针对单一的图像。然而，通过利用大小不一的损失和基于对抗性攻击的优化策略，我们表明实际上有可能通过对参数梯度的了解忠实地重建高分辨率的图像，并证明这种隐私的打破甚至对于训练有素的深度网络也是可能的。我们分析了架构和参数对重建输入图像的难度的影响，并证明任何输入到全连接层的输入都可以被分析重建，与其余架构无关。最后，我们讨论了在实践中遇到的设置，并表明在联合学习的应用中，即使在几个迭代或几个图像上平均梯度也不能保护用户的隐私。

○

)
。拥有训练数据的用户(z
; , 9
;)。联合学习的理念是只分享梯度
8
8
(
+
"
p,
)
, 而不是原始数据(z
, p
)与服务器, 它随后积累到

*作者贡献相同。

预印本。正在审查中。



图1：从梯度 $W\#6p(z, p)$ 对输入图像 z 进行重建。左图：来自验证数据集的图像。中间：从在 ImageNet 上训练的 ResNet-18 的重建情况。右图：从经过训练的 ResNet-152 重建。在这两种情况下，图像的预期隐私都被破坏了。还请注意，以前的攻击不能恢复 ImageNet 大小的数据 [38]。

更新整体权重。使用梯度下降法，服务器的更新可以，例如，构成

$$\theta^{k+1} = \theta^k - \tau \sum_{i=1}^N \nabla_{\theta} \mathcal{L}_{\theta^k}(x_i, y_i). \quad (2)$$

服务器 用户

更新后的参数 θ^{k+1} 被送回给各个用户。公式 (2) 中的程序被称为 *联合 SGD*。相反，在 *联合平均法* [19, 24] 中，每个用户在本地计算几个梯度下降步骤，并将更新的参数发回给服务器。最后，关于

(z, θ^k) 可以进一步模糊，只共享

几个局部例子的梯度的平均值

$\frac{1}{N} \sum_{i=1}^N \nabla_{\theta} \mathcal{L}_{\theta^k}(x_i, y_i)$ ，我们把它称为 *联邦-图像设置*。

这种分布式学习已经被用于现实世界的应用中，其中用户隐私是、

例如，医院数据 [15] 或移动设备上的文本预测 [3]，有人说 “隐私因 [联合学习] 更新的短暂性和集中性而得到加强” [3]：模型更新被认为比原始数据包含更少的信息，通过聚合多个数据点的更新，原始数据被认为无法恢复。在这项工作中，我们通过分析和经验表明，参数梯度仍然带有关于所谓的私有输入数据的重要信息，正如我们在图 1 中说明的那样。我们的结论是，即使是现实架构上的 *多图像联合平均*，也不能保证所有用户数据的隐私，表明在一批 100 张图像中，仍有几张可以恢复。

威胁模型：我们调查了一个 *诚实但好奇的* 服务器，目的是揭开用户数据：攻击者被允许单独存储和处理单个用户传输的更新，但 *不得干扰协作学习算法*。攻击者不得修改模型结构以更好地适应他们的攻击，也不得发送不代表实际学习的全局模型的恶意全局参数。在第 6 节中，用户被允许在本地积累数据。我们参考补充材料以获得进一步的评论，并提到在对攻击者的较弱约束下，该攻击几乎是琐碎的。

在本文中，我们首先在学术环境中讨论了联合学习的隐私限制，重点讨论了从一个图像进行梯

度反转的案例，并表明

- 对于现实中的深度和非光滑架构，无论是训练过的还是未训练过的参数，都可以通过梯度信息重建输入数据。
- 在正确的攻击下，几乎不存在 "深度防御"--深层网络和浅层网络一样容易受到攻击。
- 我们证明，任何全连接层的输入都可以通过分析来重建，与其余的网络结构无关。

然后我们考虑这些发现对实际情况的影响，发现

- 从其平均梯度重建多个独立的输入图像在实践中是可能的，在多个历时中，使用局部小批处理，甚至对多达100个图像进行局部梯度平均处理。

2 相关工作

之前研究从梯度信息中恢复的相关工作仅限于实际意义不大的浅层网络。对于神经网络，首次讨论从梯度信息中恢复图像数据的是[28, 27]，他们证明对于单个神经元或线性层，恢复是可能的。对于卷积架构，[34]表明，对于4层CNN来说，恢复单个图像是可能的，尽管有一个明显大的全连接（FC）层。他们的工作首先构建了一个输入图像的“表征”，然后用GAN来改进它。[38]对此进行了扩展，表明对于一个4层的CNN（有一个大的FC层，平滑的sigmoid激活，没有跨度，均匀的随机权重），缺失的标签信息也可以被联合重建。他们进一步表明，从它们的平均梯度重建多个图像确实是可能的（对于最大的批次大小为8）。[38]也讨论了更深入的架构，但没有提供具体的结果。一个后续文献[37]指出，标签信息可以从最后一层的梯度中分析计算出来。这些作品对模型架构和模型参数做了强有力的假设，使重建更容易，但违反了我们在这项工作中考虑的威胁模型，并导致不太现实的情况。

在[34, 38, 37]中讨论的中心恢复机制是对一个欧氏匹配项的优化。成本函数

$$\arg \min ||\nabla_{\theta} \mathcal{L}_{\theta}(x, y) - \nabla_{\theta} \mathcal{L}_{\theta}(x^*, y)||^2 \quad (3)$$

最小化，以便从传输的梯度 $\mathbf{g}(x^*, p)$ 中恢复原始输入图像 \mathbf{z}^* 。这个优化问题由L-BFGS求解器[21]来解决。请注意，对 \mathbf{g} 的梯度进行微分对 \mathbf{z} 来说，需要考虑参数化函数的二阶导数，而L-BFGS需要构建一个三阶导数近似，这对具有ReLU单元的神经网络来说是个挑战，因为高阶导数是不连续的。

与输入图像的完全重建相比，一个相关但更容易的问题是从局部更新中检索输入属性[26, 11]，例如，在人脸识别系统中被识别的人是否戴着帽子。即使是与当前任务无关的属性信息也可以从神经网络的更深层中恢复，而这些信息可以从局部更新中恢复。

我们的问题陈述与模型反转[10]进一步相关，其中训练图像是在训练后从网络参数中恢复的。这为我们的设置提供了一个自然的极限案例。如果没有额外的信息，模型反转对于更深层次的神经网络架构[36]通常是具有挑战性的[10, 36]。另一个密切相关的任务是从视觉表征中反转[9, 8, 23]，其中，给定神经网络的某个中间层的输出，重建一个合理的输入图像。这个过程可能会泄露一些信息，例如一般的图像组成，主要的颜色--但是，根据给定的层，它只重建类似的图像--如果神经网络没有被明确选择为（大部分）可反转的话[13]。正如我们在后面所证明的，从视觉表征中反转严格来说比从梯度信息中恢复更难。

3 理论分析：从图像的梯度中恢复图像

为了从理论上理解联合学习中打破隐私的整体问题，让我们首先分析一下数据 $z \in \mathbb{R}^n$ 是否可以从其梯度 $\nabla_{\theta} f_p(z, p) \in \mathbb{R}^m$ 中分析恢复。

由于 z 和 $\nabla_{\theta} f_p(z, p)$ 的维度不同，重建质量肯定是参数 p 的数量与输入像素 n 的问题。如果 $p = n$ ，那么重建至少与从不完整数据中恢复图像一样困难[4, 2]，但即使 $p < n$ ，我们在大多数计算机视觉应用中会期望， $\nabla_{\theta} f_p$ 的规则化 "反转" 的困难与梯度算子的非线性以及它的条件有关。

有趣的是，全连接层在我们的问题中扮演了一个特殊的角色：正如我们在下面证明的那样，全连接层的输入总是可以从参数梯度中分析计算出来，与该层在神经网络中的位置无关（只要有一个技术条件，即

防止零梯度，是满足的）。特别是，分析重建与全连接层之前或之后的具体类型无关，全连接网络的单一输入总是可以通过分析重建，而不需要解决优化问题。下面的陈述是对[27]中例3的概括，适用于具有任意损失函数的任意神经网络的设置：

命题3.1。 考虑一个神经网络包含一个有偏见的全连接层，前面只有（可能是无偏见的）[全连接层]。此外，假设对于任何这些（全连接层），损失函数的导数相对于该层的输出至少包含一个非零条目。那么，网络的输入可以通过网络梯度的唯一性来重建。

证明。下面我们给出证明的概要，更详细的推导请参考补充材料。考虑一个无偏的全连接层将输入 z_l 映射到输出，例如，经过ReLU非线性： $z = \max(Ax, 0)$ ，对于一个兼容的矩阵 A 维度。根据假设，可以认为 $\frac{d}{dx} \max(0, \cdot) = 0$ ，对于某个索引 i ，那么根据连锁规则 z_l 可以被计算为 $\left(\frac{d\max}{d(x_{l+1})_i}\right)^{-1}$ 这样就可以反复计算出

只要知道对某层输出的导数，就可以知道各层的输入。我们最后指出，增加一个偏置可以被解释为一个层将 z_l 映射到 $z_l + b$ ，并指出

鉴于上述考虑，另一个有趣的方面是，许多流行的网络结构使用完全连接的层（或其级联）作为其最后的预测层。因此，这些预测模块的输入是前几层的输出，可以被重构。这些激活通常已经包含了一些关于输入图像的信息，从而使它们暴露给攻击者。在这方面，特别有趣的是，正如[37]中所讨论的那样，可以从最后一个全连接层的梯度中重构地面真实标签信息。最后，提议3.1可以得出结论，对于任何以全连接层结束的分类网络，从参数梯度重建输入严格来说比从其最后一个卷积层倒置视觉表征更容易，如[9, 8, 23]中讨论的那样。

4 数值化的重建方法

由于图像分类网络很少从全连接层开始，让我们转向输入的数字重建：以前的重建算法依赖于两个部分；公式（3）的欧几里得成本函数和通过L-BFGS的优化。我们认为，这些选择对于更现实的架构，特别是任意的参数向量来说不是最佳选择。如果我们将参数梯度分解为其规范大小和方向，我们会发现大小只捕捉到关于训练状态的信息，衡量数据点相对于当前模型的局部最优性。相比之下，梯度的高维方向可以携带重要的信息，因为两个数据点之间的角度可以量化一个数据点在向另一个数据点迈出梯度一步时的预测变化[6, 18]。因此，我们建议使用基于角度的成本函数，即余弦相似度， $\langle z, p \rangle - \langle z, q \rangle / (\|z\| \|p\|)$ 。与公式（3）相比，我们的目标不是找到与观察到的梯度最匹配的图像，而是找到导致模型预测变化与（未观察到的！）地面真相相似的图像。这相当于最小化欧氏成本函数，如果我们另外约束两个梯度向量都被归一化为1的幅度。

我们进一步将我们的搜索空间限制在图像，在 $0, 1$ 之间，并且只添加总变异[30]作为整个问题的简单图像先验，参考[34]：

$$\text{争论 } \frac{\langle \nabla_{\theta} \mathcal{L}_{\theta}(x, y), \nabla_{\theta} \mathcal{L}_{\theta}(x^*, y) \rangle}{\|\nabla_{\theta} \mathcal{L}_{\theta}(x, y)\| \|\nabla_{\theta} \mathcal{L}_{\theta}(x^*, y)\|} - 1 = \text{a TV}(z). \quad (4)$$

其次，我们注意到，我们的目标是通过最小化一个取决于（间接地，通过它们的梯度）中间层输出的数量，在一个给定的区间内找到一些输入 z ，这与为神经网络寻找对抗性扰动的任务有关[32, 22, 1]。因此，我们最小化公式。

(4)只基于其梯度的符号，我们用Adam[17]进行优化，步长衰减。但是请注意，有符号的梯度只影响Adam的一阶和二阶动量，而

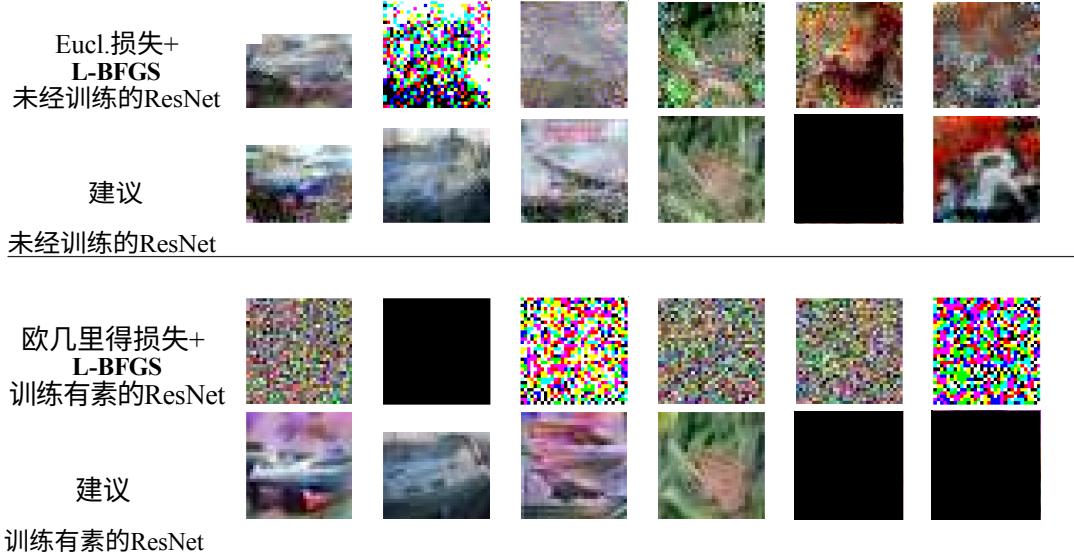


图2：[34, 38]中显示的网络架构的基线比较。我们显示了CIFAR-10验证集的前6幅图像。

实际的更新步骤仍然是基于累积动力的无符号的，因此，图像仍然可以被准确地恢复。

应用这些技术可以得到图1中观察到的重建结果。对所提出的机制的进一步消融可以在附录中找到。我们提供了一个pytorch的实现，

<https://github.com/JonasGeiping/invertinggradients>。

备注（优化标签信息）。虽然我们也可以把公式中的标签 y 看作是未知的。

(4) 和优化 $\text{joint } y \text{ for } (x, y)$ as in [38], we follow [37] who find that label information can be reconstructed analytically for classification tasks. 因此，我们认为标签信息是已知的。

5 从单一梯度重建单一图像

与以前在联合学习环境中打破隐私的工作类似，我们首先关注从梯度 $\nabla_{\theta} \mathcal{L}(z, p) C$ 重建单一输入图像 $z \in C$ 。这个设置是一个概念的证明，也是我们在第6节考虑的多图像分布式学习设置的重建质量的上限。虽然以前的工作已经表明，对于单个图像来说，打破隐私是可能的，但他们的实验仅限于相当浅的、平滑的和未经训练的网络。在下文中，我们将我们提出的方法与先前的工作进行比较，并对架构以及训练相关的选择对重建的影响进行详细的实验。每个实验的所有超参数设置和更直观的结果都在补充材料中提供。

与以前的方法比较。我们首先通过与[34, 38, 37]中考虑的通过L-BFGS优化的欧氏损失(3)进行比较来验证我们的方法。这种方法经常会因为初始化不好而失败，所以我们允许L-BFGS求解器有16次重启的宽松设置。为了进行定量比较，我们使用与[38]相同的浅层平滑CNN（我们将其称为 "LeNet (Zhu)"）以及ResNet架构，在验证集的前100幅图像上测量了重建 32×32 CIFAR-10图像的平均PSNR，其中包括经过训练和未经训练的参数。表1比较了用L-BFGS优化（如[34, 38, 37]）的euclidean loss (3)的重建质量和提议的方法。前者对于未经训练的、平滑的

、浅层的架构效果非常好，但在训练过的ResNet上完全失败。我们注意到[34]应用GAN来提高来自LBFGS重建的图像质量，然而，当代表人物过于扭曲而无法被增强时，GAN就会失效。我们的方法提供了可识别的图像，并且在训练过的ResNet的现实环境中效果特别好，我们可以在图2中看到。有趣的是，根据表1，尽管PSNR值较低，但在训练过的ResNet上的重建比未训练过的ResNet的重建具有更好的视觉质量。让我们在一个更现实的环境中研究训练过的网络参数的效果，即从ResNet-152重建ImageNet图像。

表1：在CIFAR-10验证数据集的第一幅图像上进行的100次实验的PSNR平均值和标准偏差，两个不同的网络有训练过和未训练过的参数。

建筑学 有素 确	LeNet (朱) 假的	真	错	ResNet20-4 训练 准
Eucl.损失+L-BFGS46	.25 -1- 12.66	IJ.24 -1- 0.44	1U.29	5.JS b.9U 2.bU
建议	18.00 -1- 3.33	18.08 -1- 4. 2719	.83 a 2.9613	.95 a 3.38



图3：从经过训练的ResNet-152的参数梯度进行单幅图像重建。最上面一行：地面实况。最下面一行：重构。我们检查了ILSVRC2012验证集的每1000张图像。每张图片泄露的信息量在很大程度上取决于图片内容--虽然有些例子如两个榫头被高度破坏，但黑天鹅几乎没有泄露可用的信息。

训练过的与未训练过的网络。 如果一个网络经过训练，并且有足够的能力使损失函数 Up 的梯度在不同的输入下为零，显然它们的梯度永远无法区分。然而，在实际设置中，由于随机梯度下降、数据增量和有限的训练次数，图像的梯度很少完全为零。虽然我们观察到图像梯度在训练过的网络中比在未训练过的网络中要小得多，但我们(4)中的不分大小的方法仍然可以从训练过的梯度的方向上恢复重要的视觉信息。

我们观察到对训练过的网络有两种普遍的影响，我们用图3中的ImageNet重构来说明：首先，重构似乎隐含地偏向于训练数据中同一类的典型特征，例如，第5张图片中capercaille的羽毛更偏蓝，或者我们预告1中猫头鹰的大眼睛。因此，尽管大多数图像的整体隐私明显被破坏，但这种影响至少阻碍了对细微尺度细节或图像背景的恢复。

第二，我们发现，在神经网络训练过程中使用的数据增强导致训练出来的网络使物体的定位更加困难：请注意，图3中很少有物体保留了它们的原始位置，蛇和壁虎是如何重复的。因此，尽管用数据增强训练的网络进行图像重建仍然是成功的，但一些位置信息已经丢失。

平移不变的卷积。 让我们通过测试传统的卷积神经网络（使用零填充的卷积）与可证明的平移不变的CNN（使用圆形填



充的卷积) 的比较, 来更详细地研究模糊物体位置的能力。

如

如插图所示, 虽然传统的CNN允许恢复相当高质量的图像 (左), 但由于原始物体被分离, 平移不变的网络使得物体的定位不可能 (右)。因此, 我们将常见的零填充确定为隐私风险的来源。

网络深度和宽度。对于分类精度来说, CNN每层的深度和通道数是非常重要的参数, 这就是为什么我们研究它们对重建的影响。图4显示, 重建质量随着通道数的增加而明显提高。

然而, 更大的网络宽度也伴随着实验成功率的增加。然而, 随着实验的多次重启, 更宽的网络可以产生更好的重建, 从而使PSNR值从19增加到近23, 对于

原创	16	ResNet-18与基础宽度： 64	128	卫星网-34卫星网-50
PSNR 平均值。 PSNR 标准	17.24 19.02 2.84	17.37 22.04 5.89	25.25 22.94 6.83	18.62 21.59 4.49
				21.36 20.98 5.57

图4：多种ResNet架构对原始图像的重建（左）。PSNR值指的是显示的图像，而平均PSNR是通过前10张CIFAR-10图像计算的。标准差是指在给定架构下一个实验的平均标准差。ResNet-18架构显示了三种不同的宽度。

因此，更大的网络宽度增加了攻击者的计算努力，但并没有提供更大的安全性。

从我们从不同深度的ResNets获得的重建结果来看，所提出的攻击随着网络深度的增加而退化的程度非常小。特别是--如图3所示，通过ResNet-152甚至可以进行忠实的ImageNet重构。

6 带有联合平均法和多个图像的分布式学习

到目前为止，我们只考虑了从单幅图像的梯度恢复，并讨论了这种情况下的局限性和可能性。现在，我们转向严格意义上的更困难的泛化设置--联合平均法[24, 25, 29]和多图像重建，以表明所提出的改进也能转化为这种更实际的情况，讨论这种应用中的可能性和局限性。

联合平均法不是只根据本地数据计算网络参数的梯度，而是在将更新的参数送回服务器之前，对本地数据执行多个更新步骤。按照[24]的说法，我们让用户一方的本地数据由n张图像组成。在一定数量的本地历时中，用户在每个历时中执行g个随机梯度更新步骤，其中B表示本地小批量大小，导致总的本地更新数量为 ϵg

步骤。然后每个用户i将本地更新的参数 \hat{a}^i 送回服务器，而服务器又通过对所有用户的平均数来更新全局参数 B^i 。

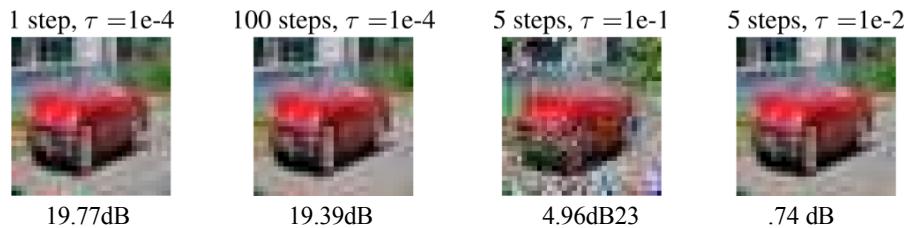


图5：说明局部更新步骤的数量和学习率对重建的影响：左边两幅图比较了固定学习率 $\tau=1e-4$ 时梯度下降步骤数的影响。右边的两幅图像是在固定的5个梯度下降步骤的情况下改

变学习率的结果。图像下方显示了PSNR值。

我们的经验表明，即使是具有 $n=1$ 张图像的联合平均的设置也有可能受到攻击。为此，我们试图通过对本地更新的了解来重建 n 个图像的本地批处理 $\hat{A} + \hat{B}$ 。在下文中，我们对 n 、 f_i 和 B 的不同选择评估了重建图像的质量。我们注意到，前几节研究的设置对应于 $n=1$ ， $A=1$ ， $\&=1$ 。在我们所有的实验中，我们使用了一个未经训练的ConvNet。



图6：ResNet32-10在CIFAR-100上的一批100张图像的信息泄露。显示的是整个批次中最容易识别的5张图像。尽管大多数图像是不可识别的，但即使在大批量的设置中，隐私也会被破坏。我们参考补充材料中的所有图像。

表2：各种联合平均设置的PSNR统计数据，在CIFAR-10验证数据集的前100张图像上实验的平均值。

1个纪元		5个纪元	
4个图像	图像	图像	图像
batchsize 2	batchsize 2	batchsize 8	batchsize 1
16.92i 2.10 14.66 1.12 1%.49 1.U2 25.U5i 3.2b			16.55i 0.56

多个梯度下降步骤， $B \cdot n \cdot l$, $f_i \cdot 1$:

图5显示了在不同数量的局部epochs f_i 和不同选择的学习率 r 下 $n=1$ 图像的重建情况。即使是在高数量的100个局部梯度下降步骤中，重建质量也没有受到影响。我们能够举例说明的唯一失败案例是由选择 $l=1$ 的高学习率引起的。然而，这种设置相当于一个会导致分歧训练更新的步骤大小，因此不能提供有用的模型更新。

多图像恢复， $\& = n \cdot 1$ ， $E \cdot 1$:

到目前为止，我们只考虑了单个图像的恢复，似乎有理由相信，在向服务器发送更新之前，对多个（本地）图像的梯度进行平均，可以恢复联合学习的隐私。虽然这样的多图像恢复在[38]中已经考虑到了 $B \cdot 8$ ，但我们证明了所提出的方法能够从100个平均梯度的批次中恢复一些信息：虽然大多数恢复的图像无法识别（如补充材料所示），但图6显示了5张最容易识别的图像，说明即使对100张图像的梯度进行平均也不能完全保证私人数据。最令人惊讶的是，批处理所产生的失真并不均匀。我们可以预期所有的图像都会有同样的扭曲，而且几乎无法恢复，然而有些图像的扭曲程度很高，而其他图像的扭曲程度只达到可以轻易识别出图片中的物体。

一般案例

我们还考虑了在每个小批量梯度步骤中使用整个局部数据的一个子集的多个局部更新步骤的一

般情况。表2提供了所有进行的实验的概述。对于每个设置，我们在CIFAR-10验证集上进行了100次实验。对于一个小批量中的多张图像，我们只使用不同标签的图像，以避免同一标签的重建图像的互换模糊性。正如预期的那样，从PSNRs值来看，单幅图像重建结果是最容易受到攻击的。尽管在PSNR方面表现较差，我们仍然观察到所有多图像重建任务的隐私泄漏，包括那些在随机小批次中获取梯度的任务。比较1和5个历时的全批8幅图像的例子，我们看到我们之前的观察，即多个历时不会使重建问题变得更困难，也延伸到多幅图像。对于表2中所有实验设置的重建图像的定性评估，我们参考了补充材料。

7 结论

联合学习是分布式计算中的一个现代范式转变，然而它对隐私的好处还没有被很好地理解。我们揭示了可能的攻击途径，分析了分析重建任何全连接层的输入的能力，提出了一个基于优化的一般攻击，并讨论了其对不同类型的架构和网络参数的有效性。我们的实验结果是用现代计算机视觉架构获得的，用于图像分类。它们清楚地表明，可证明的差分隐私仍然是保证安全的唯一途径，甚至可能是对于较大的数据点批次。

更广泛的影响 - 联合学习不保证隐私

最近关于联合学习设置中的隐私攻击的工作 ([28, 27, 34, 38, 37]) 暗示了这样一个事实，即以前希望的 "隐私因[联合学习]更新的短暂性和集中性而得到加强"[3]在一般情况下并不真实。在这项工作中，我们证明了改进的优化策略，如余弦相似性损失和有符号的亚当优化器，可以在计算机视觉的工业现实设置中，在联合学习设置中进行图像恢复：与之前的工作中的理想化架构相反，我们证明了在优化器的多个联合平均步骤中，甚至在100张图像的批次中，图像恢复在深度非光滑架构中是可能的。

我们注意到，考虑到图像数据的固有结构、图像分类网络的规模，以及相对于其他个人信息而言，单个用户可能拥有的图像数量较少，图像分类可能特别容易受到这些类型的攻击。另一方面，这种攻击可能只是向更强大的攻击迈出的第一步。因此，这项工作指出，如何在协作训练高度精确的机器学习方法时保护我们的数据隐私，这个问题在很大程度上还没有解决：虽然差分隐私提供了可证明的保证，但它也大大降低了所产生的模型的准确性 [14]。因此，差分隐私和安全聚合的实现成本很高，所以数据公司有一些经济上的动机，只使用基本的联合学习。关于更普遍的讨论，见[33]。因此，人们对进一步研究使本文提出的攻击无效的保护隐私的学习技术有着强烈的兴趣。这可能是通过防御机制或通过可计算的保证发生的，这些保证允许从业者验证他们的应用程序是否容易受到这种攻击。

参考文献

- [1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *arXiv:1802.00420 [cs.J]*, February 2018.
- [2] Martin Benning 和 Martin Burger. 逆问题的现代正则化方法. *Acta Numerica*, 27:1-111, May 2018.
- [3] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingberman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecny, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roslander. 迈向规模化的联合学习：系统设计。 *arXiv:1902.01046 [cs, stat.J]*, March 2019.
- [4] E.J. Candes, J. Romberg, and T. Tao. 稳健的不确定性原则：来自高度不完整频率信息的精确信号重建。 *IEEE Transactions on Information Theory*, 52(2):489-509, February 2006.
- [5] Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, and Elliot Holtham. *arXiv:1709.03698*. Reversible Architectures for Arbitrarily Deep Residual Neural Networks.

lcs, stat J, September 2017.

- [6] Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka.从神经网络的角度看输入相似性。In *Advances in Neural Information Processing Systems 32*, pages 5342-5351.Curran Associates, Inc., 2019.
- [7] Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman.Adam项目：建立一个高效和可扩展的深度学习训练系统。在第11届[USENIX J Symposium on Operating Systems Design und Implementation (OSDI J 14)]，第571-582页，2014年。

- [8] Alexey Dosovitskiy和Thomas Brox. 基于深度网络的感知相似度指标生成图像。在《神经信息处理系统研究进展》第29期, 第658-666页。Curran Associates, Inc., 2016.
- [9] Alexey Dosovitskiy和Thomas Brox. 用卷积网络颠倒视觉表征。在IEEE计算机视觉和模式识别会议论文集, 第4829-4837页, 2016年。
- [10] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart.利用信心信息的模型反转攻击和基本对策。在第22届ACM SIGSAC Citt F-ter und Communications Security会议论文集, CCS '15, 第1322-1333页, 美国科罗拉多州丹佛, 2015年10月。计算机械协会。
- [11] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov.使用互换不变表示对全连接神经网络的属性推断攻击。在2017年ACM SIGSAC计算机和通信安全会议上, 第619-633页, 加拿大多伦多, 2018年1月。ACM.
- [12] Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, and Tom Goldstein.真相还是逆向宣传? An Empirical Investigation of Deep Learning Theory. *arXiv:1910.00359 [cs, math, stat J]*, October 2019.
- [13] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon.I-RevNet: Deep Invertible Networks. *arXiv:1802.07088 [cs, stat]*, 2018年2月。
- [14] Bargav Jayaraman和David Evans.Evaluating Differentially Private Machine Learning in Practice. *urXiv:1902.08874 [cs, statJ]*, August 2019.
- [15] Arthur Jochems, Timo M. Deist, Issam El Naqa, Marc Kessler, Chuck Mayo, Jackson Reeves, Shruti Jolly, Martha Matuszak, Randall Ten Haken, Johan van Soest, Cary Oberije, Corinne Faivre-Finn, Gareth Price, Dirk de Ruysscher, Philippe Lambin, and Andre Dekker.通过在三个国家的分布式学习, 开发和验证NSCLC患者的生存预测模型。*International Journal of Radiation Oncology Biology Physics*, 99(2):344-352, October 2017.
- [16] Arthur Jochens, Timo M. Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker.分布式学习: 在不离开医院的情况下, 基于多家医院的数据开发预测模型--一个现实生活中的概念证明。*Radiotherapy and Oncology*, 121(3):459-467, December 2016.
- [17] Diederik P. Kingma 和 Jimmy Ba.Adam: A Method for Stochastic Optimization.在国际学习表征会议(*ICLR*)上, 圣地亚哥, 2015年5月。
- [18] Pang Wei Koh和Percy Liang.通过影响函数理解黑盒预测。在国际机器学习会议, 第1885-1894页, 2017年7月。
- [19] Jakub Konečny, Brendan McMahan, and Daniel Ramage.Federated Optimization:Distributed Optimization Beyond the Datacenter. *arXiv:1511.03575 tc.i, math J*, November 2015.
- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton.用深度卷积神经网络进行图像分类。In *Advances in Neural Information Processing Systems*, pages 1097-1105, 2012.
- [21] Dong C. Liu and Jorge Nocedal.On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503-528, August 1989.
- [22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]*, June 2017.
- [23] Aravindh Mahendran 和 Andrea Vedaldi.使用自然预设图像实现深度卷积神经网络的可视化。*International Journal of Computer Vision*, 120(3):233-255, December 2016.
- [24] H.Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas.*arXiv:1602.05629 [csJ*, February 2017].
- [25] H.Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang.Learning Differentially Private Recurrent Language Models. *urXiv:1710.06963 [cs]*, February 2018.
- [26] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov.利用协作学习中的非故

- 意特征泄漏。In *2019 IEEE Symposium on Security and Privacy (SP)* , pages 691-706, May 2019.
- [27] Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai.保护隐私的深度学习：重新审视和加强.In *Applications and Techniques in Information Security*, Communications in Computer and Information Science, pages 100-110, Singapore, 2017.Springer.
- [28] Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai.通过加法同态加密保护隐私的深度学习。技术报告715, 2017。

- [29] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečny, Sanjiv Kumar, and H. Brendan McMahan. Adaptive Federated Optimization. *arXiv.2003.00295 [cs, stat]*, 2020年2月。
- [30] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. 基于非线性总变异的噪声去除算法. *Physica D: Nonlinear Phenomena*, 60(1): 259-268, November 1992.
- [31] Reza Shokri and Vitaly Shmatikov. 保护隐私的深度学习。在第22届ACM SIGSAC计算机和通信安全会议论文集-CCS '15, 第1310-1321页, 美国科罗拉多州丹佛市, 2015年。ACM出版社。
- [32] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 神经网络的诱人特性。In *arXiv. '1312.6199 [CsJ]*, December 2013.
- [33] Michael Veale, Reuben Binns, and Lilian Edwards. 记忆中的算法：模型反转攻击和数据保护法。皇家学会哲学论文集A.数学、物理和工程科学, 376 (2133) : 20180083, 2018年11月。
- [34] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond Inferring Class Representatives: *arXiv: 1812.00535 [csJ]*, December 2018.
- [35] 杨强, 刘洋, 陈天健, 童永新. Federated Machine Learning: *arXiv. '1902.04885 [cs]*, February 2019.
- [36] 张宇恒, 贾若曦, 裴恒志, 王文晓, 李波, 宋黎明。The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. *arXiv. '1911.07135 [cs, statJ]*, November 2019.
- [37] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. iDLG: Improved Deep Leakage from Gradients. *arXiv:2001.02610 [cs, statJ]*, 2020年1月。
- [38] Ligeng Zhu, Zhijian Liu, and Song Han. 梯度的深度泄漏. In *Advances in Neural Information Processing Systems* 32, pages 14774-14784. Curran Associates, Inc., 2019.

A 威胁模式的变化

在这项工作中, 我们考虑了导言中讨论的诚实-小屋-好奇的威胁模式。偏离这种情况主要可以通过两种方式进行: 首先是改变架构, 其次是保持架构的非恶意性, 但改变发送给用户的全局参数。

A.1 不诚实的架构

到目前为止, 我们假设服务器是在一个诚实而好奇的模型下运行, 因此不会恶意修改模型以使重建更容易。如果我们允许这一点, 那么重建就变得几乎微不足道: 可以使用几种机制: 根据提议1, 服务器可以, 例如, 在第一层放置一个完全连接的层, 或者甚至通过串联直接将输入连接到网络的末端。稍微不那么明显的是, 该模型可以被修改为包含可逆块[5, 13]。这些块允许从它们的输出中恢复输入。从提议1中我们知道, 我们可以重建分类层的输入, 所以这允许立即访问输入图像。如果服务器恶意地为每个批次的例子引入单独的权重或子模型, 那么这也允许恢复一个任意大的批次的数据。在这样的环境下操作, 这种行为是可能的, 这就要求用户 (或用户信任的提供者) 以手动或编程方式审查任何传入的模型。

A.2 不诚实的参数载体

然而，即使有一个固定的诚实架构，对全局参数的恶意选择也会显著影响重建质量。例如，考虑到[38]中的网络架构不包含步长，并将卷积特征扁平化，不诚实的服务器可以将所有的卷积层设置为代表身份[12]，将输入通过网络不变地移动到分类层，从那里可以分析计算出输入，如在提议1中。同样，对于一个包含可识别的较低分辨率的架构[34]，当正确的参数向量被发送给用户时，输入可以立即恢复，尽管分辨率较低。

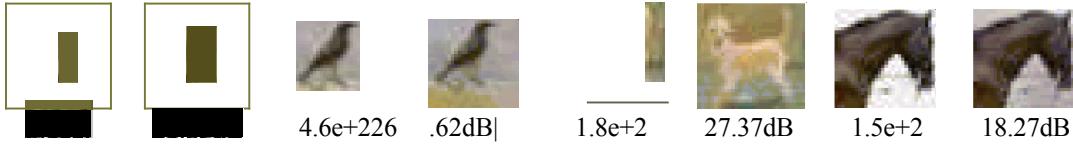


图7：标签翻转。当最后分类层的参数中的两行被翻转时，图像可以很容易地被重建。每张输入图像下面都有梯度大小，每张输出图像下面都有PSNR。将这些结果与图9中的其他例子进行比较

然而，这种具体的参数选择很可能是可以检测到的。一个更微妙的方法，至少在理论上是可能的，那就是优化发送给用户的网络参数本身，使这些参数的重建质量达到最大化。虽然这样的攻击在用户端可能很难被发现，但它也将是非常密集的计算。

标签翻转。甚至还有一个更便宜的选择。根据第5章，非常小的梯度向量可能包含较少的信息。一个不诚实的服务器提高这些梯度的简单方法是在分类层的权重矩阵和偏置中替换两行，有效地翻转标签的语义。这种攻击对用户来说很难察觉（只要梯度大小保持在通常的范围内），但却有效地欺骗了他，让他在错误的标签上区分他的网络。图7显示，这种机制可以通过提高PSNR分数来实现可靠的重建，因为训练过的模型的效果被否定了。

B 实验细节

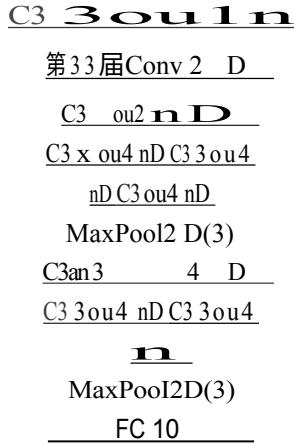


图8：网络结构ConvNet，由8个卷积层组成，指定了相关的输出通道数量。每个卷积层后面都有一个批处理归一化层和一个ReLU层。 D 代表输出通道的数量，默认设置为 $D=64$ 。

B.1 联邦平均法

将公式（4）扩展到联合平均的情况（在这种情况下，采取多个本地更新步骤并将其送回服务器）是很简单的。首先注意，给定旧参数 B' ，本地更新 $f_i +'$ ，学习率 r ，以及关于更新步数的知识²的知识，更新可以被改写为更新梯度的平均值。

$$\theta^{k+l} = \theta^k - \tau \sum^l \nabla_{\theta^{k+m}} \mathcal{L}_{\theta^{k+m}}(x, y) \quad (5)$$

²我们假设服务器知道本地更新的数量，然而这也可以通过暴力手段找到，因为/是一个小整数。

表3：在CIFAR-10上训练的ResNet-18架构的拟议方法的消融研究。重建的PSNR分数是CIFAR-10验证集前10张图像的平均值（括号内为标准误差）。

基本设置	20.12 dB (-1-1.02)
L2损失而不是余弦相似度	15.13 dB (A0.70)
没有总变数	19.96 dB (-1-0.75)
用L-BFGS而不是亚当	5.13 dB (A0.50)

从 ${}^{8k} \text{中减去} {}^l$ ，我们只需将提议的方法应用于所得到的更新的平均值：

$$\arg \min_{\in [0,1]^n} 1 / \left\| \sum_{m=1}^l \nabla_{\theta^{k+m}} \mathcal{L}_{\theta^{k+m}}(x, y) \right\| \left\| \sum_{m=1}^l \nabla_{\theta^{k+m}} \mathcal{L}_{\theta^{k+m}}(x^*, y) \right\| + \alpha \text{TV}(x). \quad (6)$$

利用自动分化，我们从更新步骤的平均值中反向传播梯度w.r.t到z。

B.2 沟通网络(ConvNet)

我们使用ConvNet架构作为实验的基线，因为它的优化速度相对较快，在CIFAR-10上达到了909c以上的精度，并且包括两个最大池层。它是AlexNet[20]的一个粗略类似物。该架构在图8中描述。

B.3 消融研究

我们在表3中提供了建议的选择的消融情况。我们注意到，有两件事是核心，亚当优化器和相似性损失。总变化是一个小的好处，而使用有符号的梯度是一个小的好处。

C 超参数设置

在我们的实验中，我们使用基于签名梯度的Adam作为优化算法，使用余弦相似度作为成本函数来重建网络的输入，如第4章所述。值得注意的是，攻击的最佳超参数取决于具体的攻击场景--默认参数下的攻击失败并不能保证安全。我们总是从均值为0、方差为1的高斯分布中初始化我们的重建（注意，对于所有考虑的数据集来说，输入数据都是正常的），并将优化算法的步长设置在 $/0.01, 1$ 之内。我们在第5.2节中使用较小的步长为0.1，用于更深的网络，在第6节中使用较大的步长为1的联合平均实验，0.1是默认选择。优化运行的迭代次数高达24000次。步长的衰减始终是固定的。

在 3 和 7 迭代之后发生，并且每次都将学习率降低0.1倍。迭代次数是一般的保守估计，隐私往往可以更早被打破。

我们根据具体的攻击场景来调整总变化参数，但要注意的是，如表3所示，它对平均PSNR的影响主要是很小。如果没有其他说明，我们默认为0.01的值。

备注（重启）。一般来说，从不同的随机初始化中多次重启攻击可以

+为了能够对多个图像进行定量的实验评估，我们在这项工作中不考虑重启（除了第5节，在那里我们改善了竞争性LBFGS求解器的结果）--但强调一个有足够的资源的攻击者可以进一步改善他

的攻击，并在多次重启后运行。

C.1 第5节中实验的设置

与以往方法的比较 为了与第5节中的基线进行比较，我们重新实现了[38]中的网络，在下文中我们称之为LeNet（Zhu），并额外运行了ResNet20-4架构的所有体验。我们的网络和方法都是基于以下的代码

建筑乐网(Zhu) 资源网20-4				
训练有素	错误	准确	假的	准确
电视	10	10	0	10

表4：在第5节的基线实验中，所提出的方法使用的电视正则化值。

编号为 "Numberofepochsfor 本地图片的数量	1	1	1	5	5
迷你批处理	4	S	T	1	S
	2	2	8	1	8

表5：重建TO网络的总加权数。

输入 在第4.2节的实验中

38]的作者，*。对于LBFGS-L2的优化，我们使用1e 4的学习率和300次迭代。对于ResNet的实验，我们使用了慷慨的8次重启，对于更快地优化LeNet (Zhu) 架构，我们使用了更高的16次重启。所有用建议的方法进行的实验只使用一次重启，4800次迭代，学习率为0.1，电视正则化参数详见表4。请注意，在描述的设置中，所提出的方法的优化时间明显少于LBFGS的优化。

空间信息 对空间信息的实验是在具有 D 64个通道的ConvNet结构上进行的。

C.2 第6节中的实验设置

对于表2中的五种情况，我们考虑了一个未经训练的ConvNet，学习率为1，4800次迭代，一次重启和表5中给出的TV正则化参数。这100个实验中的每一个都使用了不同的图像，即每个实验都使用了CIFAR-10验证集的图像，紧随前一个实验中使用的图像。由于在一个小批次中出现多个相同标签的图像会导致图像排序的模糊性，我们不考虑这种情况。如果一个已经遇到的标签的图像即将被添加到相应的小批量中，我们就跳过该图像，使用验证集中下一个具有不同标签的图像。

D 第3.1节的证明

下面我们将对提议3.1进行更详细的证明，它直接来自下面的两个提议：

命题D.1。 让一个神经网络在某一点上包含一个有偏的全连接层，即对于该层的 $lnFut x$ " 其输出 $y_l = A_l x_l + b_l$ ，那么计算 J 为 $\max\{y_l, 0\}$ 或

$$y_l = A_l x_l + b_l, \quad (7)$$

对于 $A \in \mathbb{R}^{n \times m}$ 和 $b \in \mathbb{R}^n$ ，那么输入的 x_l 可以从以下方面进行重建 A_l 和 b_l 。如果有

存在一个指数 i s.t. $\frac{dL}{dA_l} \neq 0$ 。

证明。认为 $\frac{dL}{d(b_l)_i} = \frac{dL}{d(y_l)_i} \cdot \frac{d(y_l)_i}{d(A_l)_{i,:}}$ 和 $= z^*$ 。因此

$$\frac{d\mathcal{L}}{d(b_l)_i} \cdot x_t^T \quad (8) \quad (9)$$

表示 A 的第*i*行。因此， Z_t 可以被唯一地确定，只要

$$\frac{d\mathcal{L}}{d(b_l)_i} = 0. \quad D$$

³<https://github.com/mit-han-lab/dlg>



图9：训练后的ConvNet模型（顶部）和ResNet20-4（中间）的图像重建。我们展示了来自CIFAR-10的**最坏情况**图像和**最佳情况**图像的重建，基于训练和验证集的梯度大小。每张输出图像的下面是梯度大小，每张输出图像的下面是PSNR。底部一行显示的是未经训练的模型对最坏情况下的重建。

命题D.2。考虑一个 $/u//y$ 连接的层（不一定包括偏置）后有一个ReLU激活函数，即对于一个 $\max(F^*, 0)$ 的“ottippt + i + i o”计算为 $zt + j - mcx[y_l | 0]$ ；或

$$y_l = A_l x_l, \quad (10)$$

其中最大值是按元素计算的。现在，假设我们有关于输出 $\frac{d\mathcal{L}}{dx_{l+1}}$ 的导数的额外知识，此外，假设存在一个指数 i s.t. $d(s_i + i)$ ；

Thin the inFut v可以从Proof的知识中推导出来。如同 $\frac{d\mathcal{E}}{dA_l}$ ，

$$\frac{d(x_{l+1})_i}{d(A_l)_i} = \frac{d\mathcal{L}}{d(y_l)_i} = \frac{d\mathcal{L}}{d(x_{l+1})_i} \text{ 由此可见, } d(A_l)_i = 0, \text{ 则认为}$$

$$\frac{d\mathcal{L}}{d(A_l)_{i,:}} = \frac{d\mathcal{L}}{d(y_l)_i} \cdot \frac{d(y_l)_i}{d(A_l)_{i,:}} \quad (11)$$

$$= \frac{\omega}{d(x_{l+1})_i} \cdot x_l^T. \quad (12)$$

E 其他例子

E.1 其他CIFAR-10的例子

图9显示了CIFAR-10的其他 "极端 "例子，重建了CIFAR-10的训练和验证集的训练和未训练的ConvNet和ResNet20-4模型的最低和最大梯度的图像。

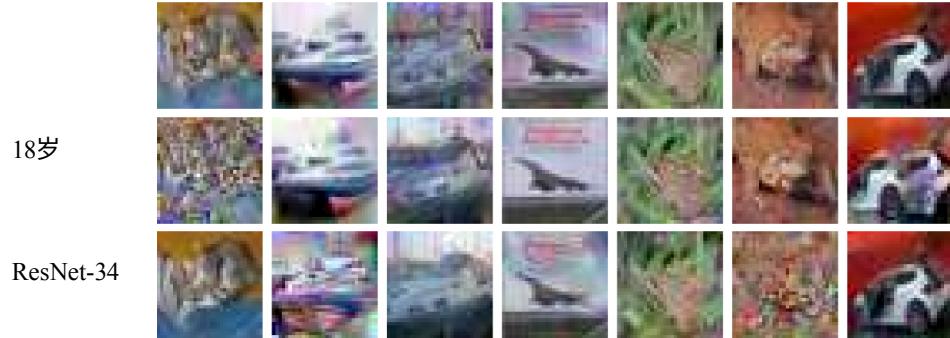
E.2 第5节中实验的可视化

网络宽度 不同宽度的ResNet-18架构对前六幅CIFAR图像的重建情况见图10。



图10：使用不同宽度的ResNet-18架构进行的重建。

网络深度 关于网络深度的实验是针对不同的深度ResNet架构进行的。图11显示了不同深度网络的多重重建结果。



共和国网-50

图11：使用不同的深度ResNet架构进行的重建。

E.3 第5节的更多图像网实例

图12显示了经过训练的ResNet-18对ImageNet验证图像进行重建的进一步指导性例子（与本文中的图3设置相同）。我们展示了一个非常好的重建（德国牧羊犬），一个好的，但被翻译的重建（大熊猫）和两个失败的案例（救护车和花）。例如，对于救护车，救护车上的实际文字仍被隐藏起来。对于花来说，花瓣的确切数量被隐藏了。另外，请注意大熊猫的重建比图像中共同出现的树桩的重建要清晰得多，我们认为这是第5节中描述的自我规范化效应的指标。

图13和14显示了更多的例子。我们注意到，这些图和图3中的例子不是手工挑选的，而是根据它们在ILSVRC2012, ImageNet, 验证集中的ID中立地选择的。每张图片的ID是通过对组成数据集的synset按其synset ID递增排序，并对每个synset中的图片按其synset ID递增排序而得到的。这是Tor chvision中的默认顺序。

E.4 第6节的多图像恢复

对于多图像的恢复，我们在图20中展示了全套的100张图像，我们建议放大到数字版本的图中。独立图像的成功率是半随机的，取决于初始化的情况。

E.5 第6节的一般情况

我们在图15、16、17、18、19中展示了前十个实验的结果。在图15中，我们甚至显示了所有的100个实验，因为每个实验只使用一个图像。



图12：额外的定性ImageNet例子，失败案例和正面案例，用于训练ResNet-18。图片取自ILSVRC2012验证集。



图13：从训练好的ResNet-的参数梯度中进行额外的单幅图像重建。

152.最上面一排：地面实况。底排：地面实况：重构。该论文展示了来自ILSVRC2012验证集的0000、1000、2000、3000、4000、5000、6000、7000的图像。这些图像是8000-12000.



图14：从训练过的ResNet-的参数梯度中进行额外的单幅图像重建。
152.最上面一排：地面实况。底排：地面实况：重建。这些是图像500，1500，2500，3500、

4500.

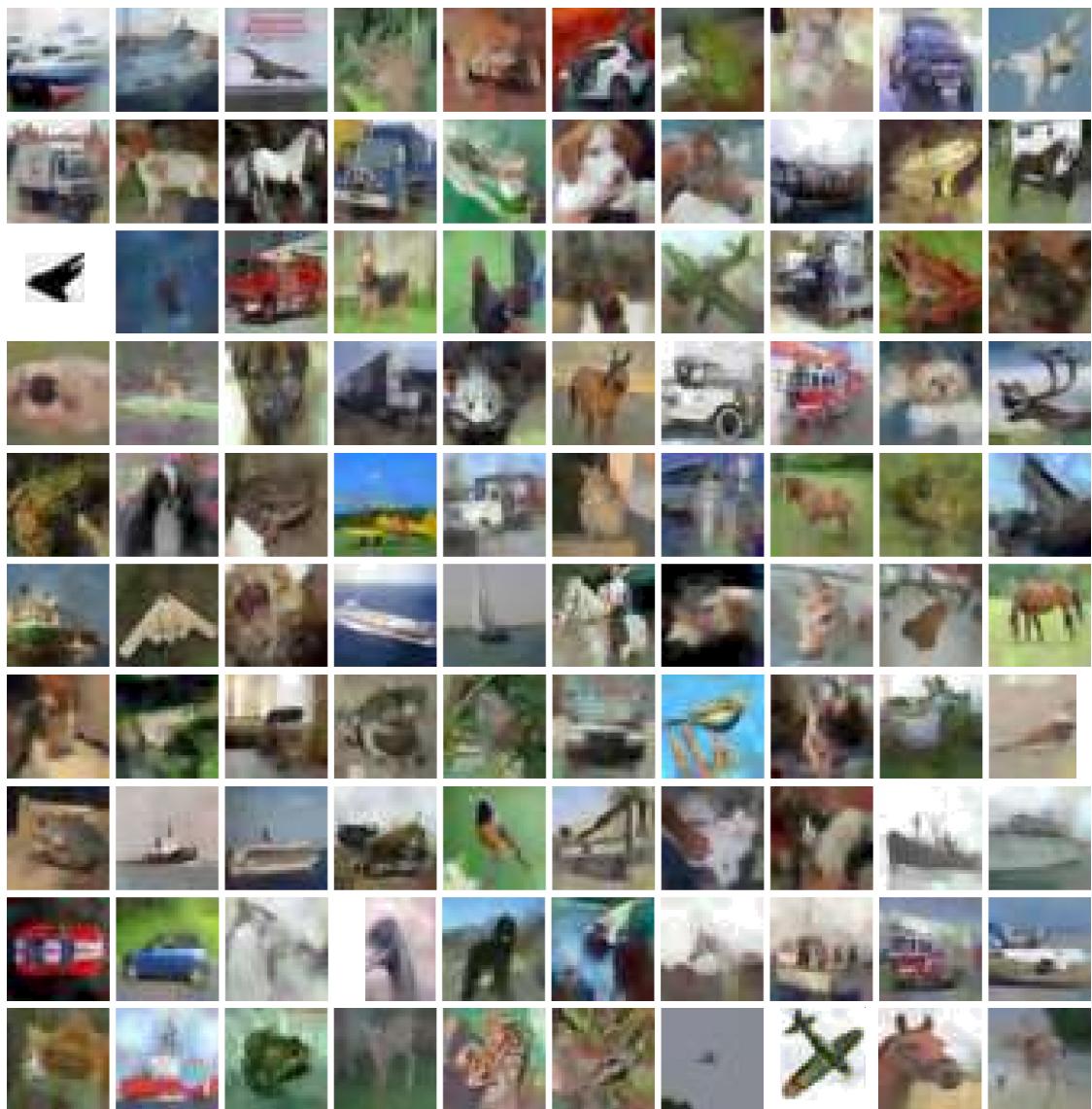


图15：前100次实验的结果， $\epsilon=5$, $n=1$, $B=1$ 。

更多图片见下页。



图16： $E=1$, $n=4$, $B=2$ 的前十个实验结果。



图17: $E=1$, $n=8$, $B=2$ 的前十个实验结果。



图18: $E=1$, $n=8$, $B=8$ 的前十个实验结果。



图19： $E=5$, $n=8$, $B=8$ 的前十个实验结果。



图20：CIFAR-100图像批次的全部结果。与本文图6的实验相同。