arXiv:2301.02610v1 [cs.NE] 6 Jan 2023

反馈门控整流线性单元

Marco Kemmerling [1]

摘要

反馈连接在人脑中发挥着重要作用，但在人工神经网络研究中却未受到太多关注。本文提出了一种受生物学启发的反馈机制，用于门控整流线性单元。在 MNIST 数据集上，与不带反馈的自编码器相比，带反馈的自编码器收敛更快、性能更好，且抗噪性更强。在 CIFAR-10 数据集上应用带反馈的网络时，也能观察到一些益处，但这些益处不那么明显，也不那么稳定。

1. 引言

数十年来，大脑一直是人工神经网络（人工神经网络，2012）的灵感来源。尽管这些模型通常相较于大脑经过了大幅简化，但近年来，它们在图像识别（Krizhevsky 等人，2012）、语音识别（Hinton 等人，2014年）和机器翻译（Sutskever 等人，2014年）等领域取得了显著成功。

尽管取得了诸多成果，但显而易见，普通人脑远比当今实践中使用的任何模型都更强大、更多才多能，因此，研究大脑与人工神经网络究竟在哪些方面、以何种方式存在差异，或许会有所裨益。

人工神经网络与大脑之间的一个此类差异在于是否存在反馈，即自上而下连接。尽管有明确证据表明大脑中存在显著的反馈连接，但人工神经网络绝大多数仍基于前馈范式设计，不过也存在并非仅遵循前馈原理的网络，这类网络称为循环神经网络（RNN）。如今实际应用中的大多数循环神经网络都侧重于从某一层到其自身的循环连接（例如，长短期记忆网络（Hochreiter & Schmidhuber，1997））。尽管这类连接具有循环性，但严格来说，它们并不构成自上而下连接。这些网络通常用于输入由序列数据构成的问题，循环连接使网络能够记住此前见过的序列元素。

然而，循环连接或反馈的用途并不一定局限于序列数据。如果输入是图像数据，那么首次查看图像（即第一遍处理）便可用于形成对图像内容的粗略认识，并识别出感兴趣的区域，随后再在第二遍处理中对这些区域作进一步检查。

虽然本文所考察的网络架构包含真正的自上而下连接，但重点并非网络拓扑本身，而是这些自上而下连接如何影响单个神经元的行为，即一种整合反馈的机制。

这种反馈机制源自神经科学文献，并从两个广泛的角度进行考察：（1）反馈机制能否以任何方式改进标准方法。相关指标包括收敛速度和训练后的网络的性能质量。（2）考察反馈的特性及其在特定条件（例如信号噪声）下的表现，能否为反馈在大脑中可能发挥的作用提供启示。毋庸置疑，试图根据简化的人工网络推断大脑中机制的功能时，必须谨慎。不过，对人工模型进行实验仍是一个令人感兴趣的机会，因为与真实大脑相比，这些模型自然更容易研究和操控。

本文其余部分将在第 2 节探讨一些神经科学背景知识，为反馈机制提供背景；随后描述反馈机制本身在大脑中的运作方式（第 2.1节）。第 3 节将介绍如何调整该机制以应用于人工神经网络，并在第 3.1节讨论使用该机制的一些实际考量。接下来的章节将介绍一系列实验，旨在回答上述研究问题。

2. 神经科学背景

新皮层是大脑皮层的一部分，也是大脑的一部分，它是哺乳动物近期才演化出的结构。它约占人脑的 80%（Markram 等人，

[1]马斯特里赫特大学，荷兰马斯特里赫特。联系邮箱：Marco Kemmerling<m.kemmerling@student.maastrichtuniversity.nl>。

2004），因此人们常推测它是高级智能出现的原因。

新皮层中最丰富的神经元类型是锥体神经元，占细胞总数的 70% 至 85%。与新皮层中的其他神经元（即所谓的中间神经元）相比，后者大多具有抑制性，而锥体神经元具有兴奋性（DeFelipe& Fari˜nas，1992）。

顾名思义，锥体神经元的细胞体大致呈金字塔形，底部有底面，顶部有顶点。锥体神经元具有两类树突：起源于底部的基底树突，以及一条起源于顶点的顶树突。这条顶树突终止于所谓的顶树突簇，在此处顶树突发生大量分支。（DeFelipe&Fari˜nas，1992）。

这些顶树突和基底树突不仅位置不同，功能也不同。基底树突接收常规的前馈输入，而顶树突簇中的树突接收反馈输入（Larkum，2013）。

新皮层似乎具有一种独特的结构，其特征在于按照层和柱进行组织。柱状组织基于这样的观察：彼此上下堆叠的神经元往往相互连接，并且具有相似的响应特性，而柱与柱之间仅存在少量连接。因此，尽管这一观点在神经科学界仍存在一定争议（Goodhill & Carreira-Perpi˜n´an，2002）。

进一步划分为六层的组织结构由 Brodman 于 1909年提出（Brodmann，1909）。第1层和第6层在这里尤其值得关注。第1层几乎不含细胞体，主要由轴突与锥体神经元的顶树突之间的连接构成（Shipp，2007），也就是说，它充当了反馈信号的连接中枢。第6层向丘脑中的神经元发送信号，而丘脑随后又向同一柱中的第1层神经元发送信号（Shipp，2007）；换言之，第1层和第6层构成了一个回路，反馈从第6层发出，并由第1层接收。

2.1 锥体神经元的远端输入

如上所述，顶树突簇树突接收反馈输入，这似乎会调节相应神经元的增益（Larkum，2004）。人们假设，这是皮层将关于世界的内部表征与外部输入相结合的一种方式，即对神经元的反馈可能会预测该特定神经元是否应当放电；只要反馈信号足够强，即使很小的前馈输入也可能使该神经元放电（Larkum，2013）。

同时考虑前馈输入和反馈输入，

神经元的放电率可以建模如下（Larkum，2004）：

$$f = g(\mu_S + \alpha\mu_D + \sigma + f\beta(\mu_D) - \theta) \tag{1}$$

其中，f 是神经元的放电率，g 是增益，μ_S 是平均胞体电流（即前馈输入），μ_D 是平均远端电流（即反馈输入），α 是衰减因子，σ 表示电流中的波动，θ 是放电阈值，而 $\beta(\mu_D)$ 是树突平均电流的递增函数，当电流值超过某个电流阈值时，该函数会达到饱和。

3. 反馈门控整流线性单元

上一节所述模型可作为推导一种激活函数的基础，该激活函数能够替代常用的整流线性单元（ReLU）（Nair & Hinton，2010 年），即 $f(x) = max(0, x)$。

为了得到一种更实用的激活函数，g 和 θ 从公式1中删去，因为阈值通过偏置单元建模，而 ReLU 的增益（即斜率）按定义为1，因此可以放心删去。删去各项 $\alpha\mu_D$ 和 σ 的合理性较低，但由于它们不影响增益增加这一核心性质，本文将其忽略，从而得到以下简化关系：

$$f = \mu_S + f\beta(\mu_D) \tag{2}$$

从右侧移除 f：

$$f = \frac{1}{1 - \beta(\mu_D)}\mu_S \tag{3}$$

剩下的是对 $\beta(\mu_D)$ 的精确定义。根据（Larkum，2004），它是“树突平均电流 μ 的增函数，在高于 1000pA 的值时达到饱和”。换言之，该函数是有界的，即增益无法增加到任意高的值。因此，需要定义函数能够产生的某个最大值 β_{max}，以及描述何时达到该最大值的阈值 η。假设采用分段线性模型，$\beta(\mu_D)$ 定义如下：

$$\beta(\mu_D) = min\left(\frac{\beta_{max}}{\eta}\mu_D, \beta_{max}\right) \tag{4}$$

由于没有明显的值可以赋给 β_{max} 和 η，因此将它们视为超参数。由于将 β_{max} 设为 1 会导致除以 0，而 $\beta_{max} > 1$ 的值会造成负斜率，因此 β_{max} 应小于 1。

将方程 4代入方程3可得：

$$f = \frac{1}{1 - min(\frac{\beta_{max}}{\eta}\ \mu_D, \beta_{max})}\ \mu_S \tag{5}$$

由于上面的方程未考虑 μ_S 的负值，因此将 μ_S 替换为 $max(0,\ \mu_S)$，即经典的 ReLU 函数：

$$f = \frac{max(0, \mu_S)}{1 - min(\frac{\beta_{max}}{\eta}\ \mu_D, \beta_{max})} \tag{6}$$

3.1. 实际应用中的反馈门控 ReLU

反馈通路旨在模拟大脑中的自上而下通路。因此，终止于某一层的反馈应源自（前馈）层级中更高的一层。

由于只有在更高层事先接收到前馈输入后，才能计算来自这些更高层的反馈，因此至少需要两个时间步才能将经过修改的 ReLU 纳入网络。具体而言，某些数据（例如一张图像）会被输入网络两次：第一次输入用于计算反馈，随后在第二次输入时利用这些反馈。虽然不需要超过两个时间步，但也可以使用任意数量的时间步，这一点将在第 4.1.1 节中探讨。

任何接收反馈的层都需要一组额外的权重来计算 μ_D。具体而言，每个大小为 n 的层 h_i 从大小为 m 的层 h_j 接收反馈时，都会引入 $n \times m$ 个额外参数。

随后可以将所得网络展开，创建一个前馈网络，因此对于 t 个时间步，每一层都会出现 t 次，同时在每个时间步使用相同的权重（见图 1）。由于展开后的网络纯粹是前馈的，因此标准反向传播是一种合适的学习规则。

在卷积神经网络（LeCun，1989年）中，反馈以逐滤波器方式实现，即每个神经元不会接收各自唯一的反馈信号，而是每个滤波器接收一个唯一的反馈信号，该信号由属于该滤波器的所有单元共享。

随机失活（Srivastava 等人， 2014年）应当在所有遍次中随机失活相同的单元。否则，例如，如果仅在最后一次遍次中应用随机失活，那么剩余单元仍会接收到来自前几次遍次中被随机失活单元的信号，这就违背了随机失活的目的。

4. 实验结果

前述各节介绍了反馈机制，以及如何在实践中实现该机制。这里将开展一系列实-

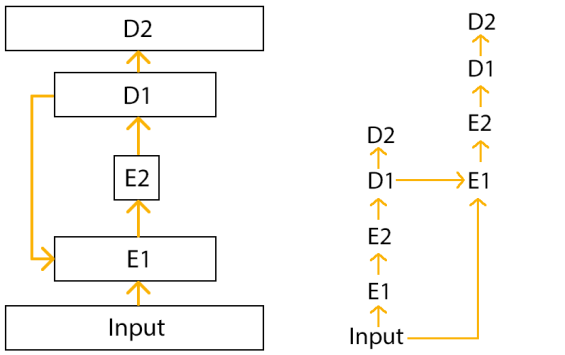

图 *1*。左：带（部分）反馈的自编码器。右：展开的自编码器。

验，以观察这种反馈机-制如何改变人工神经网络的行为。我们在两个数据集 MNIST（LeCun 等人，2010年）和 CIFAR-10（Krizhevsky 等人，2014年）上应用了多个网络。具体而言，这些实验旨在回答引言中提出的研究问题：（1）反馈能否改善人工神经网络的性能；（2）观察反馈在人工模型中的作用方式，能否揭示反馈在大脑中的功能线索。第4.1.3、4.1.4和4.2.2 节用于回答后一个问题，其中第4.1.3 节主要对反-馈进行一般性分析，而第4.1.4 节和第4.2.2 节则检验反馈是否能够增强网络的抗噪性。其余各节主要围绕问题（1）展开，测试不同配置下的收敛速度和性能质量。

4.1. MNIST

MNIST 数据集由 28×28 像素的手写数字二值图像组成，分为 60000 个训练样本和10000 个测试实例（LeCun 等人，2010年）。每张图像对应十个类别中的一个，分别表示介于0 和 9之间的数字。

以下实验中使用的模型基于一个具有两个编码层和两个解码层的（非卷积）自编码器。输入层的维度为（1×784），第一个编码层（E1）的输出数据维度为（1×392），第二个编码层（E2）的维度为（1×196），第一个解码层（D1）的维度为（1×392），而第二个解码层（D2）则将数据恢复为其原始维度。除最终层外，每一层后面都接一个 ReLU 激活。最终层使用 Sigmoid 激活函数。

最初的实验仅在第一个解码器与第一个编码器之间设置了一条反馈连接（见图 1）。

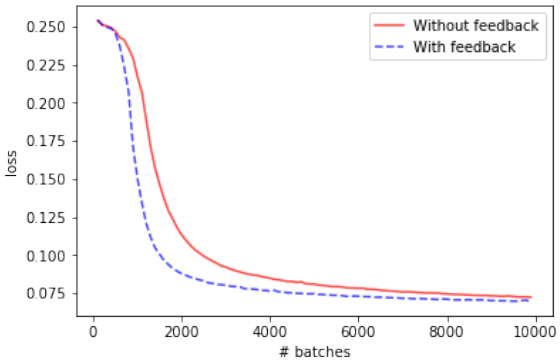

图 *2.* 有无反馈的自编码器在测试集上的损失。第二个编码层的维度为 196。

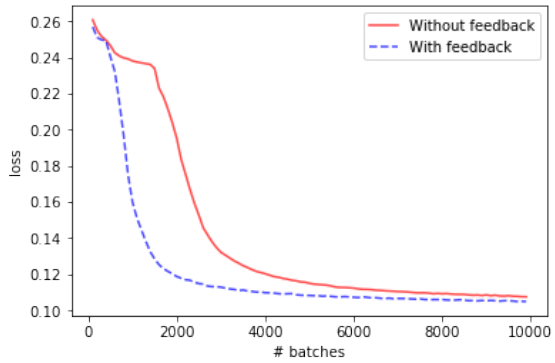

图 *3.* 有无反馈的自编码器在测试集上的损失。第二个编码层的维度为 10。

通过网格搜索确定了 η 和 β_{max} 的最优值（$\beta_{max} = 0.95$, $\eta = 5$）。

图2展示了有无反馈的自编码器损失曲线。带反馈的自编码器收敛明显更快，但差异相对较小。可以想见，如果任务难度增加，反馈可能会产生更大的影响。虽然“难度”并非一个定义明确的术语，但降低第二个编码层的维度（即瓶颈）可以说是增加了难度。

因此，第二个编码层的维度降至 10（此项修改将在之后的所有实验中沿用），并重新进行实验。确实，图 3 显示，有反馈的自编码器与无反馈的自编码器之间的差距大得多，这支持了以下假设：反馈在难度更高的任务中可能更有益。

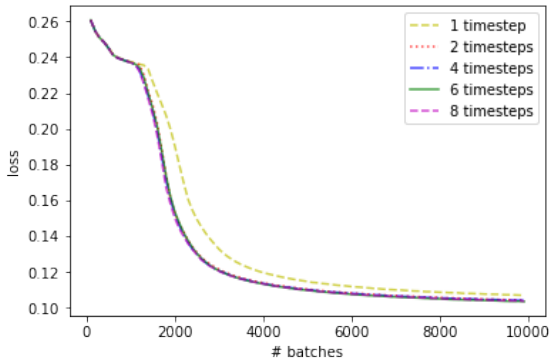

图 *4.* 不同时间步数下自编码器的性能。每种配置均训练并评估了 10 次。图中曲线显示的是测试集上的平均损失。

4.1.1. 多于两个时间步

虽然至少需要两个时间步才能纳入反馈，但尚不清楚是否应恰好使用两个时间步，或者使用 > 2 个时间步是否会更有利。为此，研究人员训练了分别使用 1、2、4、6 和 8 个时间步的自编码器。

图 4所示结果表明，超过两个时间步并未带来任何改进，或改进微乎其微。当然，这可能取决于数据和／或任务。由于 MNIST 是一个相当简单的数据集（二值图像、背景与前景界限分明等），其他数据集上的任务可能会从超过两个时间步中获益，这也并非不可想象。

4.1.2. 全面反馈

在先前的实验中，反馈仅从一个解码层发送到一个编码层。自然，还有许多其他可能的配置可以纳入更多反馈连接。在接下来的实验中，每一层都会接收来自其上方所有层的反馈，也就是说，网络中包含所有可能的自上而下连接。这种方式称为全面反馈，而先前的方法则称为部分反馈。

如图 5所示，上述配置不仅比标准自编码器收敛得更快，而且最终的损失值也更低；而仅采用部分反馈时，情况并非如此。

4.1.3. 反馈与恒定增益

为了解反馈究竟如何帮助提升性能，我们考察了不同反馈值的出现频率。这里对反馈和增益进行了区分：反馈指的是

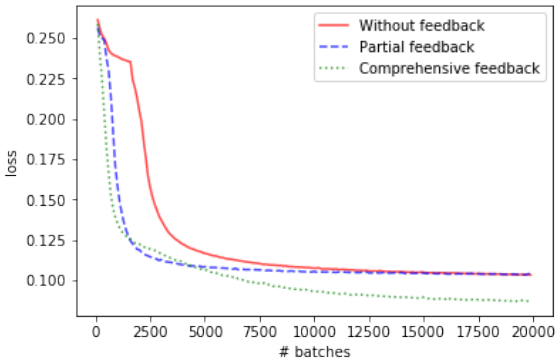

图*5*。 无反馈、部分反馈和全面反馈的自编码器在测试集上的损失。请注意，横轴与前面的图不同，也就是说，训练时间更长。

到 μ_D，增益指的是 $1-\frac{1}{(\frac{\beta_{max}}{\eta}(\mu_D)\ \beta_{max})}$。

图6 展示了在具有单个反馈连接的网络中收集的数据。

虽然也有一些较小的增益值，但绝大多数值都是网络能够产生的最大增益。这引出了一个问题：学习反馈是否有很大益处，还是简单地将所有激活值乘以一个常数也能取得类似的效果。

只需将受影响层中每个 ReLU 的增益设为常数 10，即可轻松检验这一点。

如图7所示，这确实使损失曲线比标准自编码器更陡，尽管还不及采用实际学习得到的反馈的自编码器。不过，训练完成后的性能比标准自编码器更差。

对多个反馈连接重复进行同样的实验，即针对具有全面反馈的自编码器进行实验，所得结果如图8所示。

在这种设置下，起初，仅通过乘以一个常数进行调节的收敛速度甚至比采用学习得到的反馈的自编码器更快。虽然在训练后期，其性能不及反馈自编码器，但与标准自编码器的性能相当。

显然，反馈的影响无法完全由这一恒定增益解释，但恒定增益这一想法似乎确有一定道理。

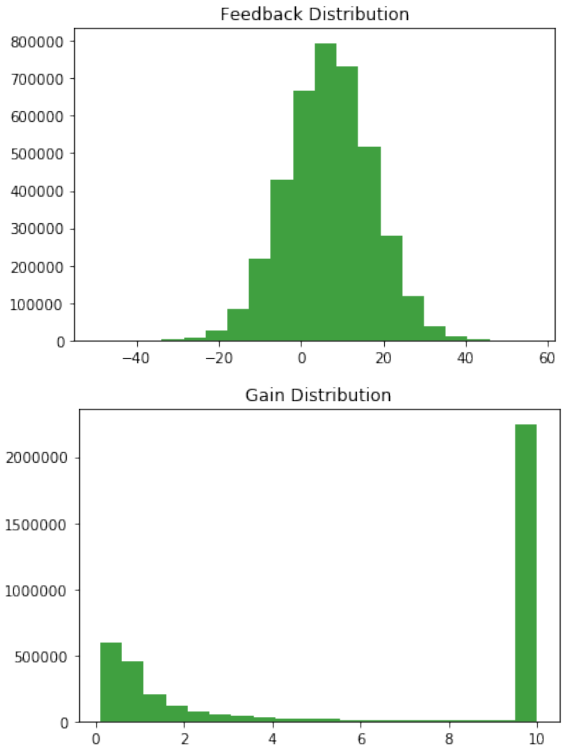

图*6*。在整个 MNIST 测试集上，具有部分反馈的网络中反馈值（上）和增益值（下）的分布。

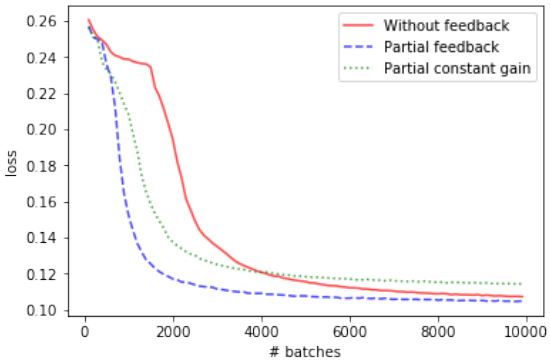

图7。标准自编码器、具有部分反馈的自编码器和具有部分恒定增益的自编码器的比较（第二个编码层中所有单元的增益均设为 10）

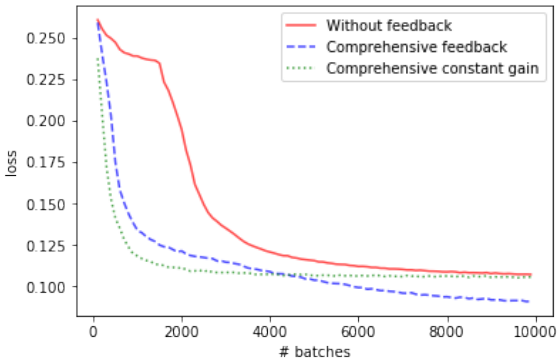

图*8*。标准自编码器、具有全面反馈的自编码器，以及具有全面增益的自编码器（所有层的增益均设为 10）的比较。



图*10*。在有反馈和无反馈的网络中加入均值为零、标准差（横轴）各不相同的高斯噪声。图中展示了两个网络的重建质量（以损失函数衡量，纵轴）随标准差大小的变化。

4.1.4 含噪激活

虽然噪声信号通常不是人工网络中的问题，但大脑中的噪声却非常普遍（Faisal 等人，2008）。为了了解反馈是否能使模型对噪声更加鲁棒，我们在有反馈和无反馈的网络的（预）激活上都加入了均值为零、标准差各不相同的高斯噪声。网络仅在加入噪声的情况下进行评估，训练过程中不加入噪声。需要注意的是，在有反馈的网络中，两个遍次的激活都会加入噪声。

$$h = f(W^T x + b + \mathcal{N}(0, \sigma^2)) \tag{7}$$

如图 9 所示，使用反馈显著增强了网络的抗噪性。虽然这对机器学习模型并非特别有用，但这可能是大脑中存在反馈通路的原因之一。

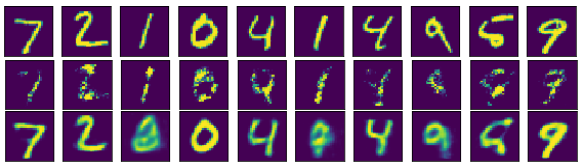

图*9*。将均值为零、标准差为 $\sigma = 2.0$ 的高斯噪声添加到有反馈和无反馈的网络中。顶排显示网络的输入样本，中排和底排分别显示无反馈和有反馈网络的重建结果。

4.2 CIFAR-10

CIFAR-10 数据集由 32×32 像素彩色图像组成，图像内容为各种物体，分为 50000 个训练样本和 10000 个测试样本。每张图像属于以下类别之一：飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车（Krizhevsky 等人，2014年）。

4.2.1 自编码器

与 MNIST 实验类似，我们在 CIFAR-10 数据集上训练了一个自编码器。同样，网络架构由两个编码层和两个解码层组成。与 MNIST 不同，此处使用的编码层/解码层是带有 165 $\times$ 5 个滤波器的卷积层/转置卷积层。

如图 11 所示，带反馈的自编码器明显优于不带反馈的自编码器，尽管两者之间的差异不像 MNIST 实验中那么显著。

有趣的是，如果在激活函数之后使用批归一化（Ioffe & Szegedy，2015），反馈便无法提升标准自编码器的性能。这或许表明，反馈与批归一化之间存在某种相互作用，导致反馈失去作用。

4.2.2. 噪声激活

在 CIFAR-10 数据集上重复了第 4.1.4 节 中的实验。所使用的网络是前一项实验中采用的、未进行批归一化的自编码器。

由于反馈提高了 MNIST 自编码器的抗噪性，因此预计这里也会出现相同行为。然而，正如图13所示，带有反馈的网络对噪声（即使是少量噪声）比不带反馈的网络敏感得多。

这可能表明，所学习到的反馈

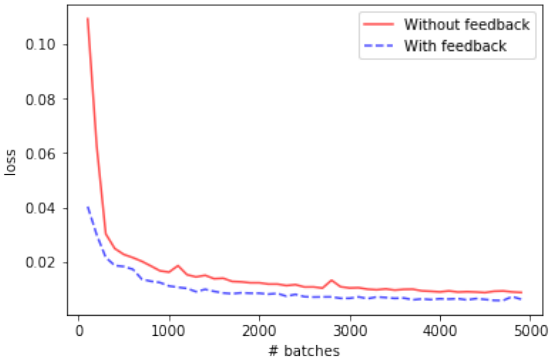

图*11*。 CIFAR-10 数据集上使用和未使用反馈的自编码器在测试集上的损失。两种模型均未使用批归一化。

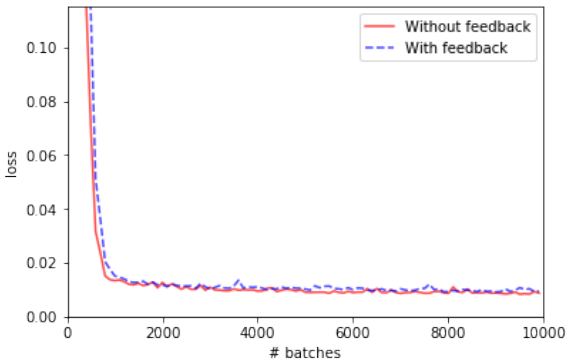

图*12*。CIFAR-10 数据集上使用和未使用反馈的自编码器在测试集上的损失。两种模型均使用了批归一化。

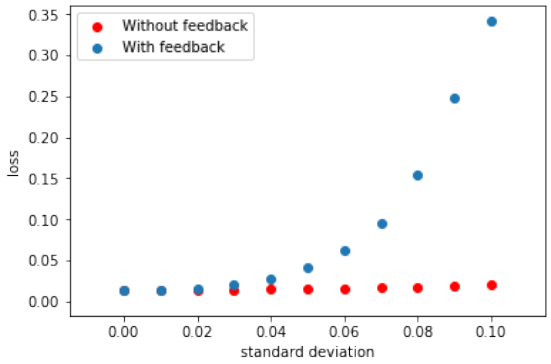

图*13*。向使用和未使用反馈的 CIFAR-10 自编码器添加均值为零、标准差各异的高斯噪声。尽管由于图表的比例而不明显，但未使用反馈的网络的数据呈现出与使用反馈的网络相似的形状。

网络中的反馈与 MNIST 实验中学习到的反馈存在根本差异，因此它会放大噪声，而非起到整流作用。

4.2.3. 分类

使用卷积神经网络对 CIFAR-10 数据集进行分类。该网络由两个卷积层组成，每层包含 64 个大小为 5×5 的滤波器，每个卷积层后接一个最大池化（ Zhou& Chellappa，1988）层，其窗口大小为 2×2，步幅为 2。卷积层和池化层后接一个全连接层（200 个单元）和一个 Softmax（Bridle，1990）层。在池化层后应用批归一化，并在池化层和全连接层后应用失活率为 0。5 的随机失活。

为了检验反馈能否提升分类性能，研究人员对网络进行了训练，分别采用（全面）反馈和不采用反馈。图 14 显示，两种网络的性能差异仅为边际程度，采用反馈的网络略优。训练结束时，采用反馈的网络在整个测试集上的分类准确率约高出0.7%。

请注意，此处采用的网络使用了批归一化；如前一节所示，批归一化与反馈结合使用可能会产生问题。目前尚不清楚这里是否存在这种情况，因为禁用批归一化后，该特定网络无法收敛（无论是否使用反馈）。

5. 结论

这里提出的反馈机制能够提升传统网络的性能，体现在

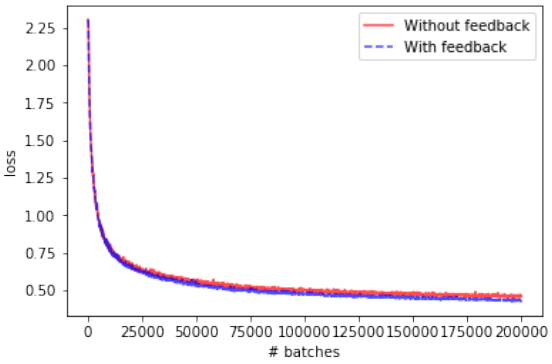

图*14.* CIFAR-10 测试集上的分类损失。200000个批次的训练时间相当于512个训练轮次。

收敛速度以及训练后的网络应用于 MNIST 数据集时的性能上。然而，反馈在 CIFAR-10 数据集上的益处则不那么明显。原则上，带反馈的自编码器可以略微优于相应的不带反馈的自编码器，但在自编码器中使用批归一化时，反馈带来的这一积极效果便会被抵消。理解反馈与批归一化之间这种不利的相互作用，或许能帮助我们更深入地了解反馈的工作原理及其发挥的作用。

在 CIFAR-10 上进行分类时，反馈似乎有一些积极作用，尽管这种作用很小，因此似乎不宜据此得出任何确定的结论。

考察网络的抗噪性时，可以观察到 MNIST 与 CIFAR-10 在性能上的差距更大。在 CIFAR-10 上，反馈不仅没有帮助，反而大幅增加了网络对噪声的敏感性；而在存在反馈时，MNIST 自编码器的抗噪性则有所提升。

造成不同数据集之间这种差异的一个可能解释是，反馈机制的有效性取决于数据集；也就是说，它可能利用了 MNIST 数据集高度规律的结构，因此在结构规律性较低的 CIFAR-10 数据集上不那么有用。

两个不同数据集上的实验之间，另一个普遍差异是卷积层的使用情况：所有 CIFAR-10 实验都使用了卷积层，而 MNIST 实验则完全没有使用。提供逐滤波器方式的反馈可能过于简单，或者与卷积相关的某些其他方面不利于反馈机制。进一步研究反馈与卷积网络的结合，或许能找到一种配置，使反馈的优势更加明显。

当然，MNIST 上的结果也可能只是一个离群值，以某种方式掩盖了当前人工神经网络中使用反馈所面临的一个更根本的问题。例如，反向传播可能并非适用于反馈的理想学习算法，或者反馈可能依赖于更贴近现实的模型，例如脉冲神经网络（Ghosh-Dastidar & Adeli，2009）。

如果有明确证据表明反馈在 MNIST 之外也有用，那么未来一个有趣的研究方向是创建基于反馈的多模态模型，将来自多个不同来源的感觉输入结合起来，以执行例如分类任务。比如，如果网络同时接收视觉输入和听觉输入，狗叫声可能会通过反馈机制提高网络对视觉输入中出现狗的预期。

致谢

我要感谢 Kurt Driessens、Mario Senden 和 Alexander Kroner 在本项目期间给予的指导。

参考文献

Bridle, John S. 前馈分类网络输出的概率解释及其与统计模式识别的关系。载于*Neurocomputing*，第 227–236 页。Springer，1990年。Brodmann, Korbinian. 《大脑皮层比较定位学：基于细胞结构阐述其原理》。Barth，1909年。DeFelipe, Javier 和 Fari˜nas, Isabel. 大脑皮层锥体神经元：突触输入的形态学和化学特征。*Progress in Neu-robiology*，39(6):563–607，1992年。Faisal, A. Aldo, Selen, Luc P. J., 和 Wolpert, Daniel M. 神经系统中的噪声。*NatureReviews Neuro-science*，9(4):292–303，2008年。Ghosh-Dastidar, Samanwoy 和 Adeli, Hojjat. 脉冲神经网络。*Internationaljournalofneuralsystems*，19(04):295–308，2009年。´Goodhill, Geoffrey J 和 Carreira-Perpi˜n´an, Miguel A. 皮层柱。*Encyclopediaofcognitive science*，2002年。Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, An- drew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, 等。用于语音识别声学建模的深度神经网络：四个研究团队的共同观点。*IEEESignalProcessing Magazine*，29 (6):82–97，2012年。

Hochreiter, Seppand Schmidhuber, J¨urgen. 长短期记忆。*Neuralcomputation*，9(8):1735–1780，1997年。

Ioffe, Sergey 和 Szegedy, Christian. 批量归一化：通过减少内部协变量偏移加速深度网络训练。载于国际机器学习会议，第 448–456 页，2015年。

Krizhevsky, Alex、Sutskever, Ilya 和 Hinton, Geoffrey E. 使用深度卷积神经网络进行 ImageNet 分类。载于神经信息处理系统进展，第 1097–1105 页，2012年。

Krizhevsky, Alex、Nair, Vinod 和 Hinton, Geoffrey. CIFAR-10 数据集。在线地址：*http://www. cs. toronto. edu/kriz/cifar. html*，2014年。

Larkum, M. E. 自上而下的树突输入会增加第5层锥体神经元的增益。大脑皮层，14(10): 1059–1070，2004年。

Larkum, Matthew. 皮层联想的细胞机制：大脑皮层的一项组织原则。*Trends in Neurosciences*，36(3):141–151，2013年。

LeCun, Yann. 泛化与网络设计策略。*Connectionism in Perspective*，第 143–155 页，1989 年。LeCun, Yann、Cortes, Corinna 和 Burges, Christopher JC。MNIST 手写数字数据库。*AT&T 实验室[在线]*。网址：*http://yann. lecun. com/exdb/mnist*，2，2010年。Markram, Henry、Toledo-Rodriguez, Maria、Wang, Yun、Gupta, Anirudh、Silberberg, Gilad 和 Wu, Caizhi。新皮层抑制系统中的中间神经元。*NatureReviews Neuroscience*，5(10)：793–807，2004 年。

Nair, Vinod 和 Hinton, Geoffrey E。整流线性单元改进受限玻尔兹曼机。载于第 *27* 届国际机器学习会议（*ICML-10*）论文集，第 807–814 页，2010年。

Shipp, Stewart. 大脑皮层的结构与功能。*CurrentBiology*, 17(12):R443–R449, 2007. Srivastava, Nitish, Hinton, Geoffrey E, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. 随机失活：一种防止神经网络过拟合的简单方法。*Journal of machine learningresearch*, 15(1):1929–1958, 2014年。Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. 使用神经网络进行序列到序列学习。In*Advances inneural information processing systems*, pp. 3104–3112, 2014年。

Zhou, YT 和 Chellappa, R. 使用神经网络计算光流。载于*IEEE* 国际神经网络会议，第 1998 卷，第 71–78 页，1988 年。

6. 附录

6.1. 超参数调优

如4.1节所述， β_{max} 和 η的最优值通过网格搜索确定。初始网格由 $\eta = [5, 10, 15, \ldots, 50]$ 和 $\beta_{max} = [0.1, 0.2, \ldots, 0.8]$定义。

的最高值 β_{max} （0。8）无论 η的取值如何，始终表现出最佳性能，如图 η15所示。请注意，当 η 取较高的常数值且 β_{max} 取不同值时，损失曲线之间的分散程度通常会更小，因为激活函数在 β_{max} 取 η 较小时会更加敏感。

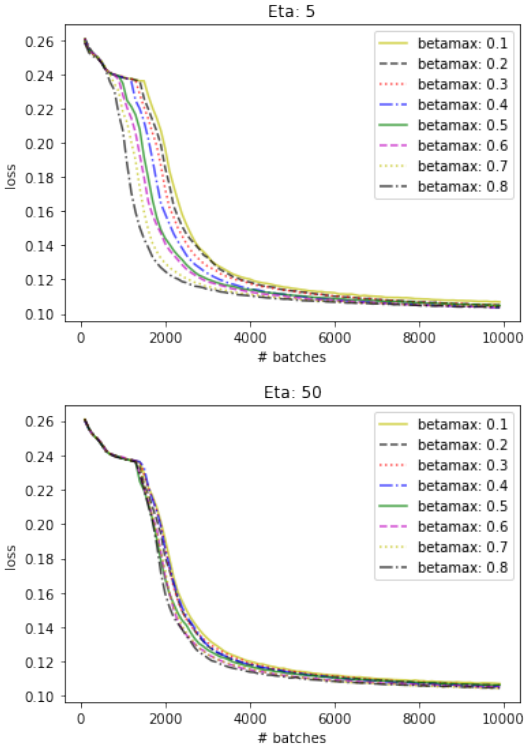

图*15*。不同超参数下的自编码器性能。上图：固定 η 为5，改变 β_{max} ；下图：固定 η 为50，改变 β_{max} 。

虽然较高的 β_{max} 会带来更好的性能，但 η则呈现相反的关系，即较低的 η 会带来更好的性能。这一点如图16所示。

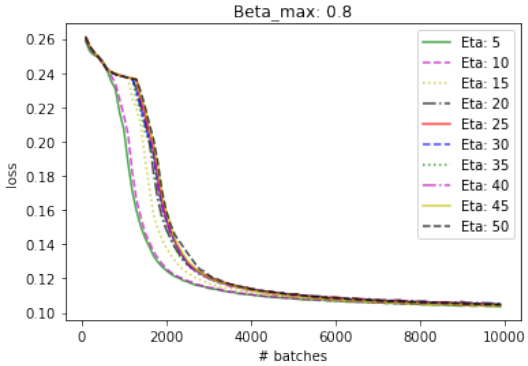

图*16*。固定 β_{max} 为0。8并改变 η 时的自编码器性能。

我们使用 $\eta = [1, 2, 3, 4, 5]$， $\beta_{max} = [0.8, 0.85, 0.9, 0.95]$ 进行第二次网格搜索，以确定进一步降低或提高取值能否改善性能。事实上，将 β_{max} 提高到 0。95 可以提升性能，但进一步降低 η 并无益处。

6.2. 反馈控制阈值

公式1 不仅描述了通过反馈进行的增益调制，还描述了对激活函数阈值的调整，即 $\alpha\mu_D$ 是求和中的一项。虽然增益调制是本文关注的主要特性，但可以设想，阈值的变化也在这一机制中发挥着重要作用。

将这一阈值机制纳入公式 6可得：

$$f = \frac{max(0, \mu_S + \alpha\mu_D)}{1 - min(\frac{\beta_{max}}{\eta}\ \mu_D, \beta_{max})} \tag{8}$$

其中 α 是网络需要学习的一个参数。虽然α 也可以设为一个常数（经调优的）值，但先前的实验表明，让网络在训练过程中调整$alpha$是有益的。

如图所示17，增加阈值机制并未能提升采用增益机制的网络的性能。尽管采用反馈控制阈值的模型表现均优于标准自编码器，但仅使用增益而不采用阈值机制的模型总体性能仍然最佳。

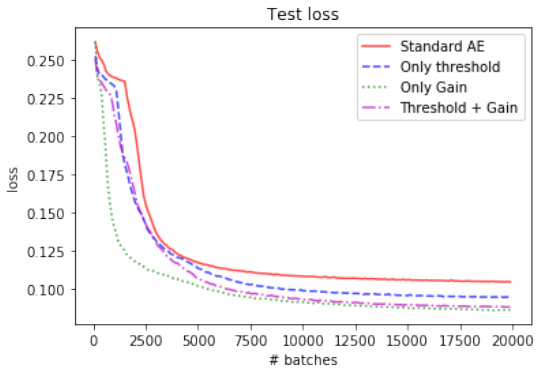

图*17*。标准自编码器、带有反馈控制阈值的自编码器、带有反馈控制增益的自编码器，以及同时带有反馈控制阈值和增益的自编码器在 MNIST 测试集上的性能。

6.3. 低对比度输入

将对比度降低的图像输入经过训练（使用常规对比度图像训练）的网络，以观察第二次传递能否重建出更接近常规对比度图像的图像。为降低对比度，将图像中的每个像素都乘以某个对比度因子 $0 \leq c \leq 1$。

图19 显示了不同对比度因子下，第一次与第二次重建结果之间平均像素值的绝对差异。高对比度输入图像会导致平均像素值差异较大，而低对比度图像则会使第一次与第二次重建结果之间的差异较小。

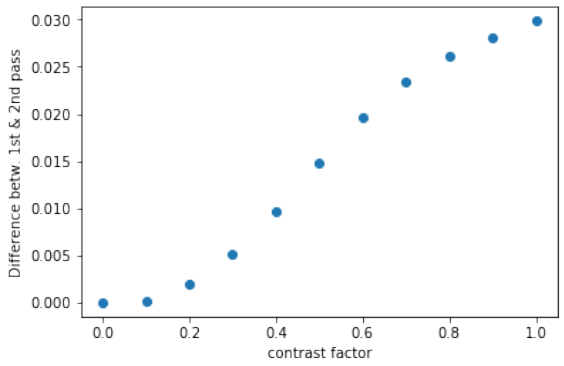

图*18*。第一遍和第二遍重建结果的平均像素值绝对差，随不同对比度因子变化的情况（从 0.0 到 1.0，步长为 0.1）。对比度因子为 1.0 时表示对比度未降低，而对比度因子为 0.0 则表示输入图像全黑。

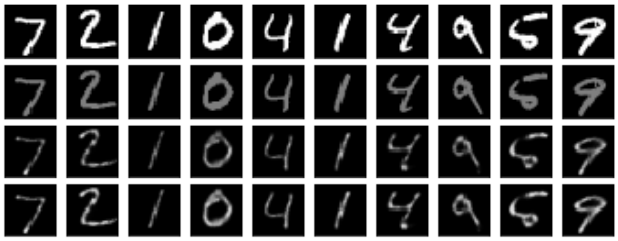

图*19*。从上到下依次为：原始图像、对比度降低后的图像、第一遍重建结果、第二遍重建结果。对比度降低后的图像由原始图像乘以 0.5 的对比度因子得到，即对比度降低后的图像中每个像素的值都在 $[0.0, 0.5]$ 范围内，而非 $[0.0, 1.0]$

6.4. 补充图

以下图表包含作为第 4 节实验一部分收集的补充数据。



图*20.* MNIST 自编码器中某个特定测试实例的激活可视化。最左侧一列对应输入层，其余各列依次对应第一编码层、第二编码层、第一解码层和第二解码层。每列中的矩形数量对应相应层的单元数。较大的值用绿色矩形表示，较小的值用白色矩形表示。上：第一次传递；下：第二次传递。

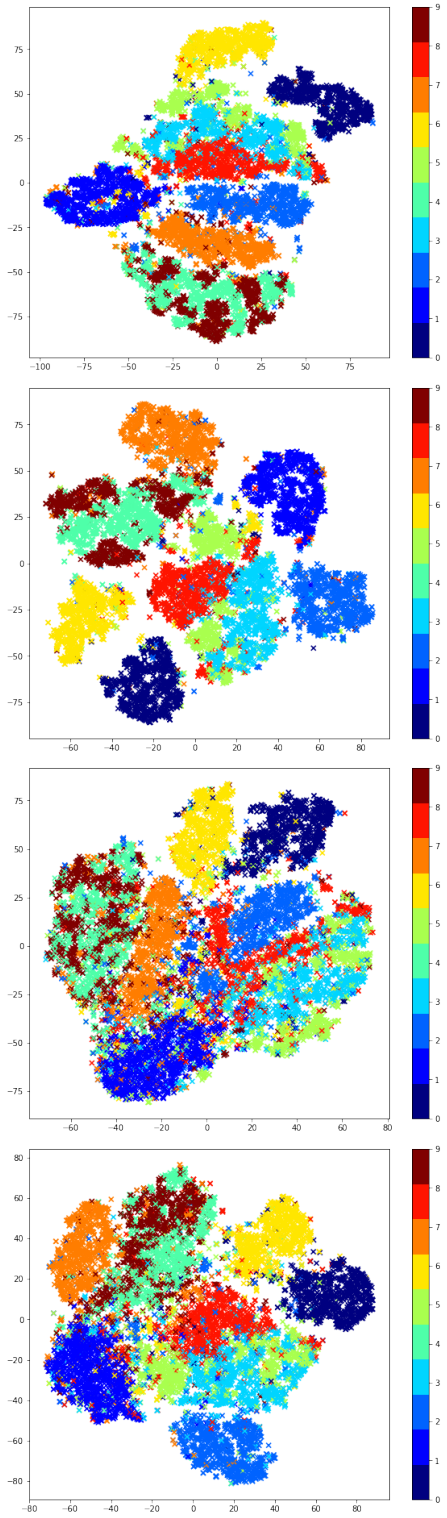

图*21.*整个 MNIST 测试集上自编码器第二编码层的 t-SNE 可视化。从上到下依次为：第一次传递、第二次传递、加入噪声后的第一次传递（如第 4.1.4 节所述）、加入噪声后的第二次传递。

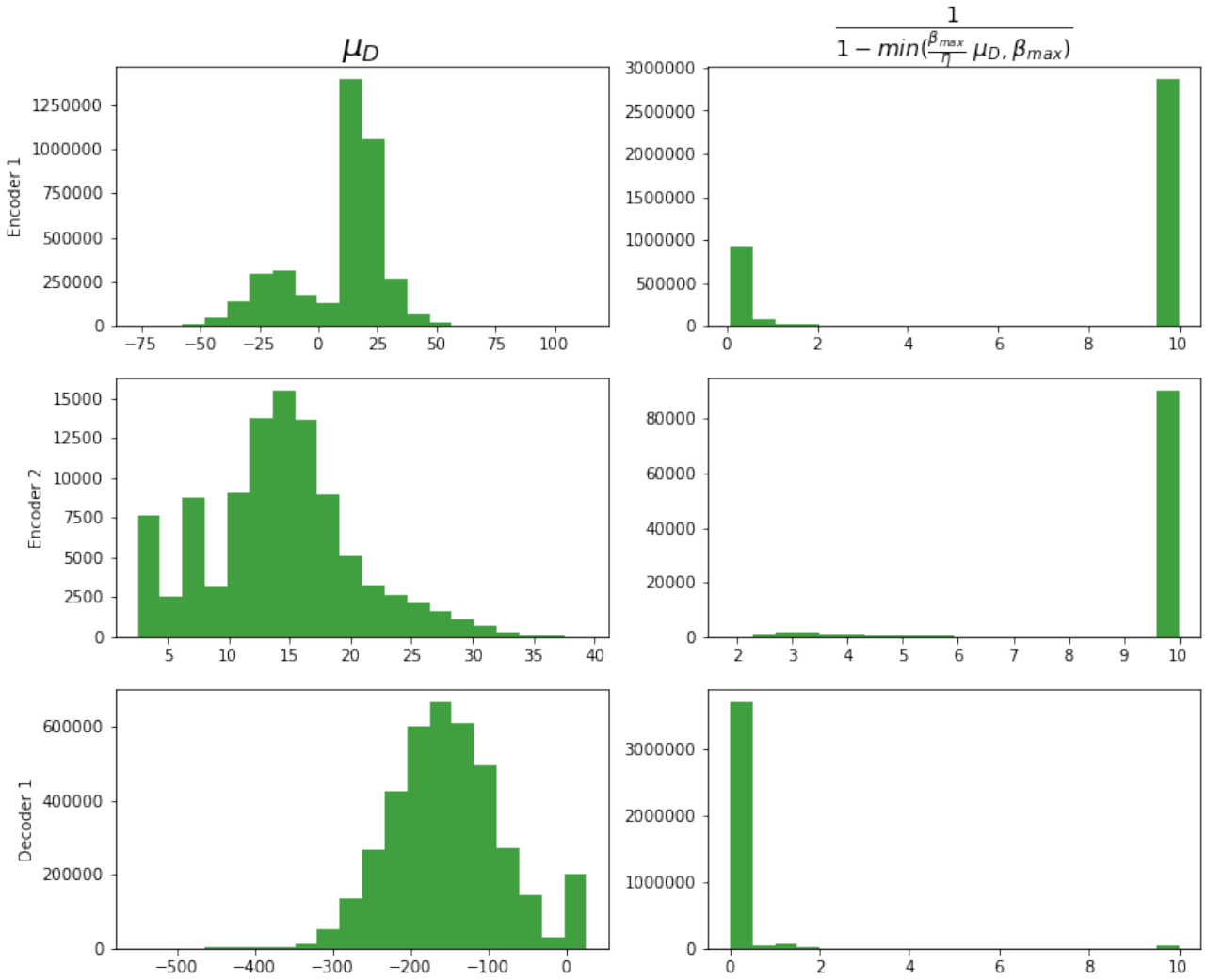

图 *22.* 第 4.1.3 节中的直方图，但针对具有全面反馈的自编码器。

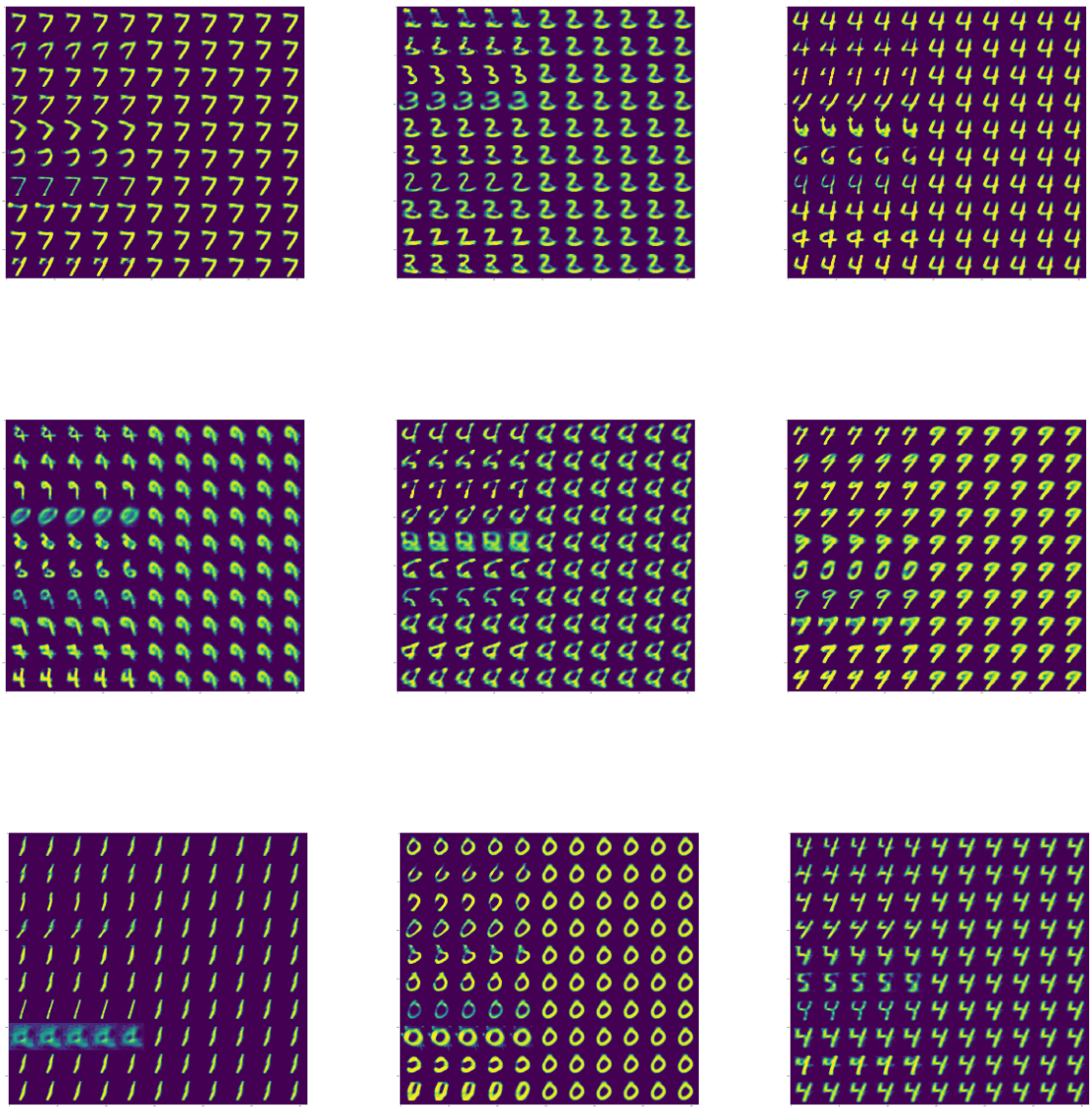

图*23*。手动将不同的增益值输入网络的第二个编码层，并将得到的重建结果可视化。在上方的每张图中，向网络输入一张特定的图像，但增益有所变化。每张图的第 i 行中，第二个编码层的每个单元的增益均为 10，只有单元 i 的增益介于 0 到 10 之间，具体取决于它所在的列。使用具有全面反馈的 MNIST 自编码器时，可以观察到，第二个编码层中只有一个单元的增益会发生变化（无论输入是什么，其余单元的增益始终为 10）。这个单元对应于每张图中自下而上的第四行，似乎负责设定重建结果的“强度”。

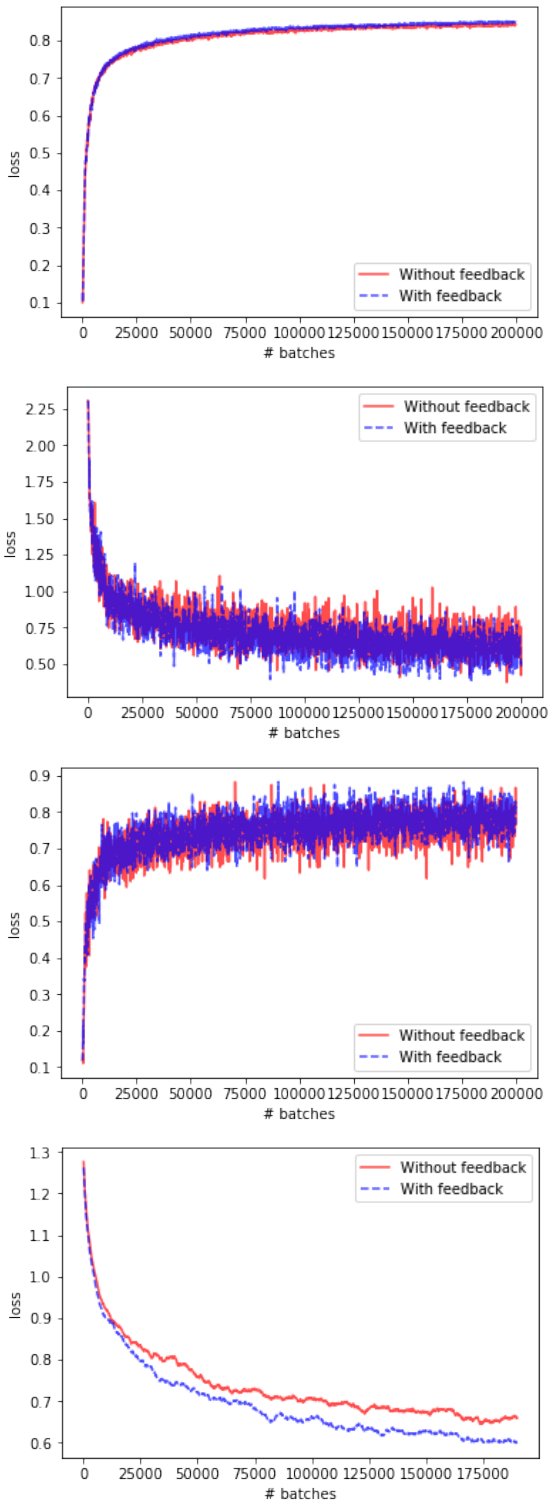

图 *24*。 第 4.2.3 节中的 CIFAR-10 分类。从上到下依次为：测试集准确率、训练集损失、训练集准确率，以及应用移动平均滤波器后的训练集损失（窗口大小为 100）。