可解释机器学习：以系统为中心的视角的重要性

Manish Narwaria

arXiv:2302.02347v1 [eess.SP] 5 Feb 2023

一、范围

在若干信号处理应用乃至教育领域中，机器学习（ML），尤其是深度学习（DL）的兴起，似乎显著改变了整体格局。其主要原因在于，深度学习能够对信号与相关任务之间复杂且未知的关系进行建模。特别是，监督式深度学习算法在不同应用中识别具有感知或语义价值的信号信息方面取得了相当大的成功（例如，从图像/视频信号中识别物体或感兴趣区域，或从语音信号中识别口语词语，即语音识别等）。在所有这些应用中，训练过程都利用带标注数据来学习一个从信号到所需信息（类别标签或目标标签）的映射函数（通常是隐式学习）。随后，训练好的深度学习模型应能在给定的测试信号中正确识别或分类相关信息。因此，总体而言，基于深度学习的框架非常具有吸引力，因为所需映射的特征和特性几乎完全从数据中学习，无需明确开发模型或系统。

然而，对隐式建模的关注也引发了一个问题：基于深度学习的映射缺乏可解释性/可诠释性，即黑箱问题。因此，可解释机器学习/深度学习已成为一个活跃的研究领域 [2], [3], [4], [5] ，其主要目标是阐明机器学习/深度学习模型如何得出预测结果。不过，我们注意到，尽管人们付出了诸多努力，关于黑箱问题的讨论似乎仍缺乏从以下角度出发的技术探讨：a）其起源和根本原因；以及 b）其对机器学习/深度学习系统设计与部署的实际影响。因此，一个值得提出的问题是：传统的以系统为中心的方法（强调显式

M. Narwaria任职于印度焦特布尔印度理工学院电气工程系，地址：印度拉贾斯坦邦卡尔瓦尔Surpura绕行路62号国道，邮编342030。电子邮箱：narwaria@iitj.ac.in

系统建模）能否为理解黑箱问题的本质提供有益见解，并有助于开发更加透明的机器学习/深度学习系统?

II. 背景与相关性

对于上述问题，答案是肯定的。通过区分以系统为中心的方法和以数据为中心的范式 [1]，可以更好地理解这一点。前者通常依赖先验信息，并从更具分析性的视角出发，旨在对物理过程进行显式建模。例如，将噪声表征为高频分量；利用梯度或边缘（高频）信息分析图像/视频信号的形状；利用信号样本之间的相关性（例如用于信号压缩）；在已知的概率密度函数上定位检验统计量（例如在假设检验中）等。因此，系统设计理念和性能分析在很大程度上仍然便于审查。相比之下，以数据为中心的方法（即机器学习/深度学习）通常侧重于通过学习从输入到期望输出的映射函数，实现隐式系统建模。这尤其得益于深度学习 [6] 强大的建模能力，它能够灵活地逐步形成合适的映射函数，也就是根据训练数据确定一组权重。然而，通常很难直接解释深度学习学到的映射，由此产生了黑箱问题。因此，在机器学习/深度学习的背景下，提出上述问题是合理的：以系统为中心的方法如何帮助我们更好地理解黑箱问题及其实际影响？这对于推动开发更透明、可解释的机器学习系统，预计至关重要。

因此，本讲义的主要目的是阐明上述可解释机器学习的相关方面。我们还尝试从机器学习系统设计的角度，就如何缓解这一问题提供一些思路。为实现这些目标，我们以以系统为中心的哲学为基础展开论述。我们将讨论限制在一个易于理解且富有意义的简单低通滤波器示例上。在我们看来，这不仅便于阐述，也使本讲义能够被不同背景的读者理解。

III. 先决条件

本讲义假定读者熟悉信号与系统的基本概念。

IV. 问题陈述与解决方案

令 $g:\ \mathbb{R}^p \to \mathbb{R}^q$ 表示从输入 $\boldsymbol{x} \in \mathbb{R}^p$ 到实际（期望）输出 $\boldsymbol{y} \in \mathbb{R}^q$的映射函数。例如，考虑图像中的目标检测应用：我们希望识别一幅 100×100 图像中是否存在三个感兴趣物体中的某一个。在这种情况下，我们有 $p = 100 \times 100$ 以及

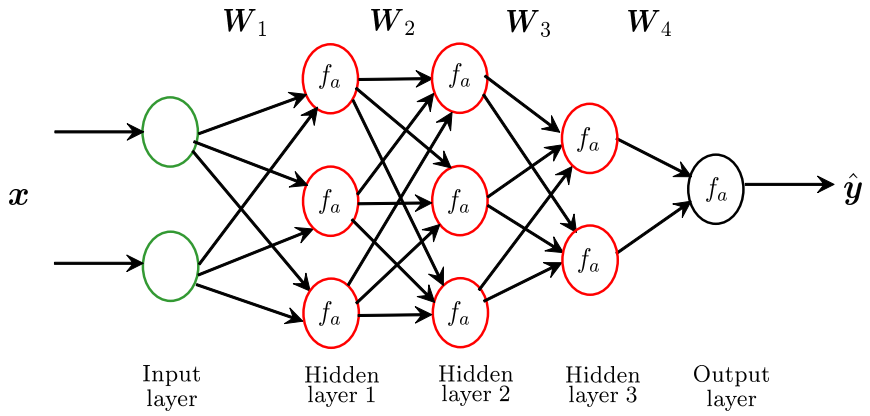


图1：一个包含3 个隐藏层的全连接前馈网络（$L = 3$）。前两个隐藏层各有3个神经元，第三个隐藏层有2个神经元。在本例中，$p = 2$ 和 $q = 1$ 分别表示输入层和输出层，它们分别有2个和1个神经元。f_a 表示激活函数，$\boldsymbol{\theta} = [\boldsymbol{W}_1, ..., \boldsymbol{W}_4]$ 是通过训练学习得到的参数。$\boldsymbol{W}_1, ..., \boldsymbol{W}_4$ 中的各项表示神经元之间连接的权重。示例请参见图2。

$q = 1$ 也就是说，输出 $g(\boldsymbol{x})$ 要么是0，要么是1，要么是2，分别对应一个类别（物体）标签。同样，在目标定位场景中，我们希望确定图像中某个物体的位置。我们可以用一个由四个坐标点构成的边界框来表示该位置。因此，$p = 100 \times 100$和 $q = 8$。我们注意到，在上述两个应用中，事实上在许多其他应用中，映射函数 g 通常是未知的。这正是深度学习（DL）广受欢迎的原因，因为它有可能从一组带标注数据中学习出一个映射$\hat{}g$。随后，人们希望$\hat{}g$ 从实际角度来看是 g 的准确估计量。

此时，简要介绍深度学习的基本工作原理及其符号表示将有所帮助。如前所述，我们将深度学习模型的输入记为 $\boldsymbol{x} \in \mathbb{R}^p$，即 $\boldsymbol{x}$ 是一个长度为 p的向量。那么，深度学习模型的q维输出$\hat{}\boldsymbol{y} \in \mathbb{R}^q$ 为：

$$\hat{\boldsymbol{y}} = \hat{g}(\boldsymbol{x}; \boldsymbol{\theta}), \tag{1}$$

这里，$\boldsymbol{\theta} = [\boldsymbol{W}_1, ..., \boldsymbol{W}_{L+1}]$ 表示由 $L + 1$ 个权重矩阵构成的参数（为简便起见，我们忽略截距或偏置项）。矩阵 $\boldsymbol{W}_1, ..., \boldsymbol{W}_{L+1}$ 中的元素表示深度学习模型/架构 [6]中不同神经元之间连接的权重。这些神经元通过 L 个隐藏层排列在深度学习模型中。作为示例，我们在图1中展示了一个具有3个隐藏层的深度学习模型，即 $L = 3$。请注意，在此示例中，前两个隐藏层各有3 个神经元，而第三个隐藏层有2 个神经元。此外，假设 $p = 2$ 和 $q = 1$。因此，输入层和

输出层分别有2 个和1 个神经元。读者可以从图1中看到，共有4个权重矩阵： $\boldsymbol{W}_1$、 $\boldsymbol{W}_2$和 $\boldsymbol{W}_3$ 分别对应三个隐藏层，而 $\boldsymbol{W}_4$表示输出层的权重矩阵。为计算输出$\hat{}\boldsymbol{y}$，深度学习模型依次使用 $\boldsymbol{W}_1, ..., \boldsymbol{W}_4$。此外，通过使用函数 $f_a : \mathbb{R}^c \to \mathbb{R}^c$，在每个神经元中引入非线性。该函数是逐元素非线性函数，通常称为激活函数。因此，输出可写为:

$$\hat{\boldsymbol{y}} = f_a\Big(\boldsymbol{W}_4 f_a\Big(\boldsymbol{W}_3 f_a\Big(\boldsymbol{W}_2 f_a\big(\boldsymbol{W}_1 \boldsymbol{x}\big)\Big)\Big)\Big), \tag{2}$$

请注意，由于我们在图1的示例中选择了 $p = 2$ 和 $q = 1$，因此 $\boldsymbol{x}$、$\boldsymbol{W}_1$、 $\boldsymbol{W}_2$、 $\boldsymbol{W}_3$、 $\boldsymbol{W}_3$和$\hat{}\boldsymbol{y}$ 的维度分别为 1×2、 2×3、 3×3、 3×2、 2×1和 1×1。训练深度学习模型的目标是 [6] 通过在一组带标注数据上进行训练来找到参数 $\boldsymbol{\theta} = [\boldsymbol{W}_1, ..., \boldsymbol{W}_{L+1}]$，使$\hat{}\boldsymbol{y}$ 接近 $\boldsymbol{y}$ （实际或期望的输出）。

A. 问题陈述

我们指出，深度学习系统包含两个方面，即设计和验证。前者指选择深度学习架构（例如，激活函数 f_a、层数 L、每层的神经元数量等），并随后进行优化 [6] 以找到参数 $\boldsymbol{\theta}$。后者指针对具体应用，在独立测试集上对训练好的深度学习模型进行基准测试。因此，深度学习模型的这两个方面都高度依赖数据。正如前文所述，我们将此称为以数据为中心的方法。因此，人们通常只依赖隐式建模（即无需进行显式信号分析或使用手工设计的信号特征），并将预测准确率作为显式系统分析的替代指标。这种以数据为中心的方法与以系统为中心的方法形成对比，后者通常基于显式建模和先验知识。因此，我们的问题陈述自然地随之而来，可以表述为：以系统为中心的方法能够揭示哪些额外且具有实际用途的见解，进而帮助设计出更加透明、可解释的机器学习/深度学习系统?

B. 解决方案

为了识别有关深度学习系统可解释性方面具有实际意义的见解，我们依托于利用显式先验知识这一理念，而这正是以系统为中心的方法的基础。例如，通信系统中的信道建模可以利用信道滤波器抽头的概率模型知识（如使用瑞利分布）。类似地，在不同应用中（如纹理分析），通过表征不同频率和方向上的视觉信号（如使用 Gabor滤波器），试图明确模拟人类视觉系统对频率和方向的选择性。因此,

以系统为中心的方法通常试图通过明确表征所研究的物理过程/应用的不同子系统（组件），来近似 g 。这进而使我们能够从应用背景下某一既定知识库的角度，对每个子系统的优势和弱点进行明确的多维分析。受这一理念启发，我们提出使用一个显式系统 S （其 g已知）作为参照。随后，我们尝试使用深度学习对 S建模，即训练一个深度学习网络，使其学得的函数$\hat{}g \approx g$。由于 S 是通过选择明确给出的，因此，对基于深度学习的 S 模型进行分析，可以在问题陈述的背景下揭示更多见解。

有多种 S 选择。但在本讲义中，我们选择一种简单的滤波应用场景，旨在衰减感知上不太重要的信号信息。这是传统应用领域（例如信号去噪、平滑、抗混叠等）和新兴应用领域（例如增强现实与虚拟现实中的空间音频渲染、物联网智能摄像机设计、医学成像等）中的常见应用场景。因此，我们令 S 为低通（移动平均）滤波器，并以此为例，滤除超过2千赫的信号信息。除了概念上简单之外，选择这种 S还有助于从生成训练数据和后续优化的角度，实现相对直接的深度学习方法。现在，从以系统为中心的视角来看， S 可以用以下差分方程方便地描述：

$$\boldsymbol{y}[n]=\frac{1}{M}\sum_{k=0}^{M-1}\boldsymbol{x}[n-k]=\sum_{k=0}^{M-1}\boldsymbol{w}[k]\boldsymbol{x}[n-k]\,, \tag{3}$$

其中 $k=0,1,...,\ M-1$。请注意，式（3）所表示的系统 S是显式的，也可以在频域中进行可视化，以便更清晰地进行物理解释。令 $M=2$，即可得到一个简单的低通滤波器，其滤波器系数为 $\boldsymbol{w}[0]=\boldsymbol{w}[1]=0.5$。因此，输出只是输入 $\boldsymbol{x}[n]$ 中当前及过去采样值的平均值，即：

$$\boldsymbol{y}[n]=0.5\boldsymbol{x}[n]+0.5\boldsymbol{x}[n-1]\,, \tag{4}$$

现在，我们尝试使用基于深度学习的回归来建模 S 。为此，我们将训练数据表示为$\{(\mathbf{z}_i,\ \mathbf{t}_i)\}_{i=1,\ ...,\ T}$。由于 $M=2$，$\mathbf{z}$是一个由$\mathbf{z}_i$作为第 i^{th} 行的矩阵。$\mathbf{t}$ 表示目标值（期望值）构成的 $T\times 1$向量。根据我们的问题设定，$\mathbf{t}_i$ 仅表示$\mathbf{z}_i$中两个数字的平均值。 T 表示训练数据的规模。训练过程旨在估计函数$\hat{}g$，使得$\mathbf{t}_i=\hat{}g\left(\mathbf{z}_i,\ \boldsymbol{\theta}^*\right)$，其中参数 $\boldsymbol{\theta}^*$ 使选定的损失函数 ℓ 最小化，即。

$$\boldsymbol{\theta}^*=\arg\min_{\boldsymbol{\theta}}\ell\left(\mathbf{z},\mathbf{t},\boldsymbol{\theta}\right)\,, \tag{5}$$

我们采用了广泛使用的均方误差作为损失函数，即。

$$\ell(\mathbf{z},\mathbf{t},\boldsymbol{\theta})=\frac{1}{T}\sum_{i=1}^{T}||\mathbf{t}_i-\hat{g}(\mathbf{z}_i,\boldsymbol{\theta})||^2, \tag{6}$$

因此，我们要寻找一个映射函数，使预测值 $\hat{g}(\mathbf{z}_i,\boldsymbol{\theta})$ 与相应目标值 $\mathbf{t}_i$ 之间的平方差之和最小化，即 $\ell(\mathbf{z},\mathbf{t},\boldsymbol{\theta})\leq\epsilon$。这里，$\epsilon$ 表示容差，即模型训练期间允许的最大误差。设置合适的 ϵ 至关重要，通常取决于应用以及训练数据的大小 T。在本讲义中，我们并不特别关注这些方面，而是直接为实验选择 $\epsilon=10^{-4}$ 和 $T=1000$。此外，要完整地指定基于深度学习的模型/架构，还需要确定若干超参数 [6]，包括隐藏层的数量 L、每个隐藏层中的神经元数量、激活函数的类型 f_a以及损失函数 ℓ。因此，在训练深度学习模型之前，需要先选定这些参数。深度学习模型经过适当训练后，人们希望它能够对训练集未出现过的数据实现良好的泛化。换言之，我们希望对于任意测试信号 $\mathbf{z}_{test}$，预测值 $\hat{g}(\mathbf{z}_{test},\boldsymbol{\theta})$ 都接近未知的目标值 y_{test}（也就是说，训练后的深度学习模型对其处理方式与训练数据相似）。在本讲义中，我们使用了三种常用的激活函数，即 sigmoid$f_a(x)=\frac{1}{1+e^{-x}}$、修正线性单元（ReLU）$f_a(x)=max(0,x)$ 和带泄漏的ReLU $f_a(x)=\begin{cases}\alpha x & x<0\\ x & x\geq 0\end{cases}$。由此，我们得到了三个训练好的深度学习网络，分别为 $\mathbf{N}_s$、$\mathbf{N}_r$ 和 $\mathbf{N}_{lr}$（$f_a=$ sigmoid、$f_a=$ ReLU 和 $f_a=$带泄漏的ReLU），其对应的函数近似分别为 $\hat{g}_s$、$\hat{g}_r$ 和 $\hat{g}_{lr}$。这些网络只有一个隐藏层（$L=1$），如图2所示。我们还使用 $L=3$ 和 $f_a=$ 带泄漏的ReLU训练了另一个网络 $\mathbf{N}_{lr}^{(3)}$（其近似函数为 $\hat{g}_{lr}^{(3)}$）。如图4所示。此外，我们使用 $\ell=$ 均方误差训练了全部4个深度学习网络。请注意，所有这些深度学习模型/网络在训练数据集上都实现了 $\epsilon\approx 10^{-4}$。此外，还需要确保缓解过拟合（或记忆化）问题，即深度学习模型在训练数据上表现良好，但在测试集上的预测误差相对较高。为此，一种实用且广受认可的解决方案是检查训练好的深度学习模型在独立于训练数据的测试集上的性能[6]。因此，我们使用包含200个数据点的独立测试数据集，对训练好的深度学习模型的性能进行了交叉验证。我们发现，全部4个训练好的深度学习模型产生的误差为 $\approx 10^{-4}$，也就是说，其性能与在训练数据集上的表现相似。因此，$\mathbf{N}_s$、$\mathbf{N}_r$、$\mathbf{N}_{lr}$ 和 $\mathbf{N}_{lr}^{(3)}$ 代表了 S 的深度学习模型。

1) 深入探讨黑箱问题：如前所述，可解释机器学习/深度学习旨在通过阐明训练好的机器学习/深度学习模型为何得出某个特定预测，来深入了解黑箱（即训练好的机器学习/深度学习模型）。然而，似乎鲜少有人讨论其真正含义究竟是

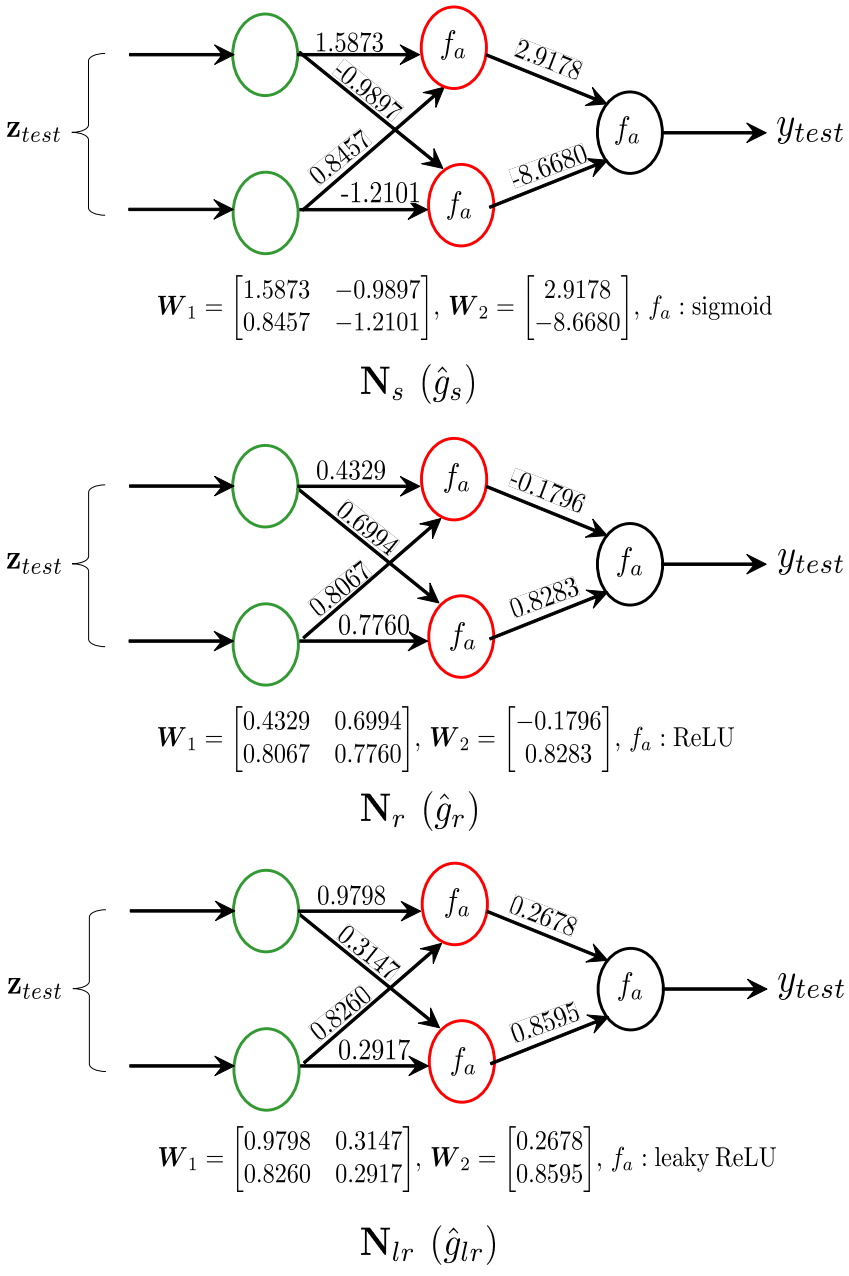


图 2：三个训练好的深度学习网络 $\mathbf{N}_s$、$\mathbf{N}_r$ 和$\mathbf{N}_{lr}$ 分别生成近似函数$\hat{}g_s$、$\hat{}g_r$ 和$\hat{}g_{lr}$，并通过设计 $\hat{}g_s = \hat{}g_r = \hat{}g_{lr} = g$。为清晰起见，输入层、隐藏层和输出层分别以绿色、红色和黑色显示。各网络下方列出了训练权重和激活函数的信息。

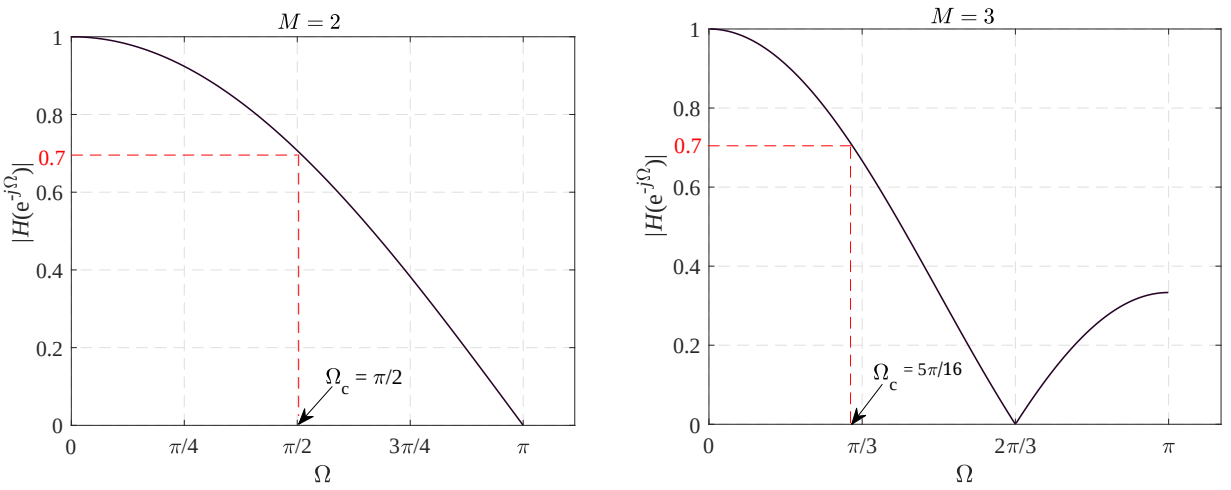


图3：绘制了 $[0,\pi]$ 范围内、 M取两个值时的幅频响应。(a) $M=2$，截止频率选为$\Omega_c=\frac{\pi}{2}$（即2千赫）；(b) $M=3$，截止频率选为$\Omega_c=\frac{5\pi}{16}$（即1.25 千赫）。

首先，训练好的深度学习模型具有黑箱性质。因此，为了更深入地理解这个问题，我们不妨先分析为何由式（4）表示的 S 并非黑箱。将式（4）写成频域形式，可得：

$$Y\left(e^{j\Omega}\right)=0.5\left(1+e^{-j\Omega}\right)X\left(e^{j\Omega}\right)=H\left(e^{j\Omega}\right)X\left(e^{j\Omega}\right), \tag{7}$$

其中我们用大写字母表示相应信号的离散时间傅里叶变换（DTFT）。符号Ω 表示离散信号的频率（因此，原则上$\Omega\in[-\pi,\pi]$）。表达式（7）告诉我们， S 的输出$Y\left(e^{j\Omega}\right)$ 只是输入$X\left(e^{j\Omega}\right)$ 与系统传递函数 $H\left(e^{j\Omega}\right)$逐元素相乘的结果。因此，观察图 3a 所示的 $H\left(e^{j\Omega}\right)$ 幅度，就能了解 S的具体特性。也就是说，系统 S 本质上会缩小（减弱）输入中较高频率分量的强度。从实用角度来看，这种分析有以下好处。首先，无需实际实现系统，就能对系统输出获得有意义的先验洞见并进行分析。例如，如果输入中不存在频率Ω_0，那么它就不可能出现在输出中。其次，这有助于我们构想并最终设计用于信号去噪或平滑、抗混叠、均衡等任务的系统（算法）。这是因为可以适当地表征输入中不需要的信号（例如噪声或感知上无关的分量），然后将其衰减。举例来说，如图 3a 所示，我们将截止频率$\Omega_c=\frac{\pi}{2}$设为弧度/采样，即对于 $0\le\Omega\le\frac{\pi}{2}$， $0.7\le|H\left(e^{j\Omega}\right)|\le1$（或等价地

$|H\left(e^{j\Omega}\right)| < 0.7$ 对于$\frac{\pi}{2} < \Omega \leq \pi$）。假设采样频率 f_s 为8 千赫，我们可以看到，从实际角度而言， S会衰减超过2千赫的分量。当然，也可以根据应用场景适当调整滤波特性（例如选择另一个截止频率）。因此，（7）和图 3a 为 S的工作方式提供了可量化的洞见。这些洞见也有助于将 S 实用地推广到其他应用场景。

考虑到以系统为中心的方法所提及的各个方面，现在可以方便地考察与 S 对应的基于深度学习的模型是否也适合进行类似分析。为此，我们仔细研究图2所示的训练后的深度学习网络 $\mathbf{N}_r$ ，并从写出明确的输入输出关系开始。令 $\mathbf{z}_{test} = [x_1 \quad x_2]$ ，其中 $x_1, x_2 \in \mathbb{R}$。然后，利用训练后的模型 $\mathbf{N}_r$的权重矩阵 W_1 和 W_2 ，可以将 y_{test} 表示为:

$$y_{test} = max\Big\{0, \Big(0.8283\ max\left(0, 0.6994\ x_1 + 0.7760\ x_2\right) - 0.1796\ max\left(0, 0.4329\ x_1 + 0.8067\ x_2\right)\Big)\Big\}, \tag{8}$$

现在，根据设计，$\mathbf{N}_r$ 在实际效果上等同于 S。这反过来意味着，由式（8）定义的相应函数$\hat{}$ g_r ，本质上是通过非线性$max(.)$ 函数的加权组合来近似低通滤波器。然而，通过分析式（8）来确立$\mathbf{N}_r$ 的低通特性可能很困难，而这在实践中会产生严重影响。具体而言，这意味着我们并不清楚$\mathbf{N}_r$ 中的训练权重（即矩阵 W_1 和 W_2）究竟如何与其低通滤波特性相关联。训练权重与其对输入信号的影响之间缺乏具有实际意义的联系，正是深度学习黑箱性质的体现。我们也可以类似地分析模型$\mathbf{N}_s$、$\mathbf{N}_{lr}$ 和$\mathbf{N}_{lr}^{(3)}$ ，并得出与$\mathbf{N}_r$类似的结论。

现在，我们将从实践角度分析上述黑箱问题的局限性。在此之前，需要指出，训练数据中存在未知随机噪声 $\{(\mathbf{z}_i, \mathbf{t}_i)\}_{i=1,\ldots,T}$以及损失函数的选择（例如本例中的均方误差）也会影响优化。因此，权重矩阵 $\boldsymbol{W}_1$ 至 $\boldsymbol{W}_4$ （见图 2 和图 4）可能会发生变化。不过，只要最终的深度学习模型能够近似 S，这就不会影响分析和结论。

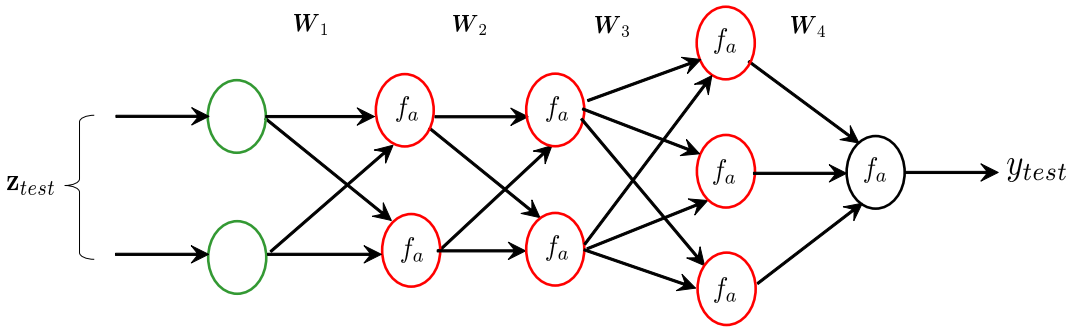
2）黑箱问题的实际影响：尽管黑箱问题看似有着理论基础，但从实际角度来看，深入理解机器学习/深度学习的黑箱性质同样十分重要。为进一步说明这一点，我们不妨考虑另一个应用场景：此时，感知上无关的信号信息位于例如1.25 千赫以上（而非2 千赫）。显然， S 在这种情况下无法直接发挥作用。不过，从以系统为中心的视角来看，很容易看出可以利用式（3）中的 $M = 3$ 构造所需的滤波器。之后，可以进行与式（7）和图 3a 类似的分析。将截止频率设为$\Omega_c = \frac{5\pi}{16}$ 弧度/采样，如图所示

图3b所示，便可得到所需的系统 S'。因此，尽管 S 和 S' 是两个不同的系统（滤波器），但它们本质上通过一种可解释的分析理念统一起来了。然而，对于$\mathbf{N}_r$，情况则大不相同。具体而言，可能不存在一种通用且可解释的过程，能够通过修改$\mathbf{N}_r$ 的权重，使所得模型（例如$\mathbf{N}'_r$）具有1.25千赫的截止频率。再强调一次，原因在于训练权重与$\mathbf{N}_r$的滤波特性之间缺乏有意义的联系。因此，式（7）和图 3a 所提供的 S 分析，不仅能让我们清楚了解 S，还提供了一种系统化的方法，将其拓展到相关的实际应用场景。同样，通带波纹控制、过渡带增益控制等其他实际问题，也可以在 S中得到明确处理。相比之下，基于深度学习的建模由于具有黑箱性质，缺乏这种具有实际意义的功能。

黑箱问题的另一个值得注意的实际局限，在机器学习/深度学习需要部署于受限环境时可能会显现出来；在这些环境中，延迟、隐私或通信带宽不足等因素可能至关重要。在此应用场景下，深度学习训练模型的预测和更新必须在计算资源有限的条件下进行。例如，在传感器附近的本地嵌入式处理设备上，或在边缘服务器上 [7]。因此，深度学习模型的参数/权重数量（这是衡量模型复杂度的指标之一*）通常必须减少，同时维持实际可接受的预测性能。在端侧/边缘计算的场景中，黑箱性质可能成为深度学习模型设计和压缩（减少权重数量）的瓶颈。原因在于，训练权重与深度学习模型功能之间缺乏清晰且可量化的关联，这可能妨碍人们系统地理解各权重的重要性。因此，要对深度学习模型的缩减作出有意义的决策可能会变得困难，也容易沦为反复试错。例如，哪些权重应进行更多（或更少）的量化、量化程度如何，以及应剪除哪些连接（即将某些权重值设为 0）等，都可能难以确定。另一方面，以系统为中心的方法能够更灵活地明确分析 S 的滤波特性是否因任何原因发生变化 $\boldsymbol{w}$ 。

黑箱性质通常也会妨碍人们从弱点（如果存在）的角度对深度学习模型进行有意义的分析。因此，在这方面，人们可能只能分析深度学习的输出，看看它是否符合期望的输出（也就是说，几乎完全聚焦于预测准确率）。这可能会为直接进行深度学习系统分析留下空间。相比之下，以系统为中心的方法可以更直接地进行系统分析。举例来说，我们可以考虑

*模型复杂度还包括其他组成部分 [7]，例如所需的数学运算次数、内存需求、表达能力等。



$$\boldsymbol{W}_1 = \begin{bmatrix} 0.9535 & 0.2735 \\ 0.0939 & 0.8767 \end{bmatrix}, \boldsymbol{W}_2 = \begin{bmatrix} 0.5287 & 0.5002 \\ 0.7715 & 0.4054 \end{bmatrix}$$

$$\boldsymbol{W}_3 = \begin{bmatrix} 0.5053 & 0.3826 & 0.7080 \\ -0.0784 & 0.6882 & 0.1601 \end{bmatrix}, \boldsymbol{W}_4 = \begin{bmatrix} 0.2209 \\ 0.4241 \\ 0.2827 \end{bmatrix}, f_a : \text{leaky ReLU}$$

$$\mathbf{N}_{lr}^{(3)} \left(\hat{g}_{lr}^{(3)} \right)$$

图4：另一个经过训练的深度学习网络$\mathbf{N}_{lr}^{(3)}$（其中 $L = 3$），它生成近似函数$\hat{g}_{lr}^{(3)}$，使得$\hat{g}_{lr}^{(3)} = g$。输入层、隐藏层和输出层分别以绿色、红色和黑色显示。网络下方还注明了训练权重和激活函数的信息。为便于直观观察，神经元之间的连接上未显示权重。

该系统 S'。从图3b可以看出， S' 的传递函数存在一个旁瓣，即在频率$\frac{2\pi}{3}$处，幅值为0，而在频率 π处略微增至约0.35 。显然，考虑到其应用场景是衰减超过1.25千赫的信息，旁瓣应尽可能小。因此，这是 S' 的一项局限，也促使人们采取措施加以缓解。对于相应的深度学习系统，例如$\mathbf{N}_r$，由于其黑箱性质，可能无法实现类似的功能。

最后，读者会从图2注意到，$\mathbf{N}_s$、$\mathbf{N}_r$ 和$\mathbf{N}_{lr}$ 都只有6个可训练权重。因此，从训练权重的维度来看，这些深度学习模型并不算复杂，尤其是与视觉领域中若干知名的深度学习架构相比（例如，一个拥有50层的ResNet[8] 含有超过2500万个训练权重）。然而，就当前的用途而言，它们本质上都是黑箱。因此，黑箱问题未必仅仅归因于训练权重的高维度。相反，正如我们所阐明的，这本质上是一个实际问题，涉及训练权重与深度学习模型相应功能之间缺乏有意义的关联。

3）为什么可解释性是深度学习系统设计的基础？：我们注意到，对可解释机器学习/深度学习的需求，很大程度上源于医疗保健、自动驾驶、法律执法、金融等高风险应用 [2]

领域。事实上，不可否认，在这些应用中，黑箱模型所犯的未能解释的错误或作出的错误决策可能会造成严重后果。然而，同样值得注意的是，可解释性实际上是机器学习/深度学习系统设计的一个基本概念，而不仅仅是设计完成后的要求。为了理解这一点，再次从以系统为中心的视角进行思考，并考虑我们的 S示例，会很有帮助。请注意， S 由（4）唯一确定，或者等价地由（7）确定。因此，可以明确地从可解释性、性能（优势和/或弱点）、实现问题等角度，对 S （甚至 S'）进行分析。然而，基于深度学习的 S建模则并非如此，因为这四个模型，即$\mathbf{N}_s$、$\mathbf{N}_r$、$\mathbf{N}_{lr}$ 和$\mathbf{N}_{lr}^{(3)}$ ，都能准确模拟 S的功能。也就是说，我们有 $g = \hat{}g_s = \hat{}g_r = \hat{}g_{lr} = \hat{}g_{lr}^{(3)}$ 。事实上，通过选择不同的 L、每层的神经元数量、 f_a 等，还可以训练出其他多个深度学习模型。因此， g 可能由大量深度学习模型近似表示。乍看之下，这似乎很有用。然而，仔细审视便会发现，这种非唯一的基于深度学习的近似，对于未知函数 g带来的问题多于答案。具体而言，我们可以提出以下问题：

（i）应当依赖哪些深度学习模型的解释，才能准确洞察 g 所代表的底层系统？为什么？
（ii）既然所有待考察的深度学习模型都能很好地近似 g ，它们最终是否应该得出相同或相似的解释？

关于第一个问题，我们可能缺乏系统的方法，无法从多个候选模型中（例如 $\mathbf{N}_s$、$\mathbf{N}_r$、$\mathbf{N}_{lr}$ 和 $\mathbf{N}_{lr}^{(3)}$ ）选定一个深度学习模型，并将其相应的解释作为 g的替代模型。第二个问题也揭示了深度学习建模中一个有趣的方面。如果答案是肯定的，那么我们仍可能难以确立这些架构各异的深度学习模型之间的等价性（差异体现在 L、隐藏层中的神经元数量以及激活函数上）。另一方面，如果上述第二个问题的答案是否定的，那么我们可能面临这样的情况：针对同一个物理过程（在我们的示例中由 g 定义），却有不同的、基于深度学习的解释。从实践角度看，这或许意义不大。因此，上述两个问题都强调了为什么可解释性应当成为机器学习/深度学习系统设计中内在且优先考虑的设计原则之一，而不应仅仅是一种事后分析程序。

五、支持基于以系统为中心理念的可解释机器学习

我们已经说明，显式建模的理念可以用于理解深度学习系统中的黑箱性质及其实际影响。因此，同样的以系统为中心的哲学也有望使深度学习系统设计受益，这一想法合乎逻辑。事实上，我们可以利用

先验领域知识和显式工具来进行深度学习系统分析与设计。更具体地说，从深度学习系统设计的角度来看，这可能带来若干具有实际价值的启示：

1) 对输入数据进行预处理，以克服深度学习模型本身的局限性（例如，利用傅里叶特征映射的思路来弥补标准多层感知机的弱点 [9]）。2) 将显式系统的某些方面融入基于学习的框架中（例如，使用深度展开进行图像超分辨率 [10]，或用于通信系统 [11]）。3) 改进深度学习模型本身的某些方面，例如性能（例如，在卷积神经网络中使用离散傅里叶变换来改进池化 [12]，或理解深度学习与图信号处理之间的连接 [13]）。4) 避免从头开始训练深度学习模型，转而基于已知工具中更简单且可解释的方面进行构建（例如，可微分数字信号处理（DDSP）框架 [14] 通过学习将可解释的数字信号处理工具适配到多种数据，以用于音频应用）。5) 可视化深度学习层中的学习过程（例如，使用主成分分析检测对抗样本 [15]）；在可解释深度学习领域开发交互式可视分析框架（例如， [16], [17]）；或为医学应用创建可解释系统的设计空间（例如，[18]）。

以上提及的观点（绝非详尽无遗）强调了在设计、分析更高性能且更具可解释性的机器学习/深度学习系统时，采用以系统为中心的哲学所带来的潜在益处。从本质上说，设计以系统为中心的深度学习模型可以分为两个步骤：a）开发一个可能更多依赖先验领域知识的基础模型；b）通过从特定应用数据中学习，对所得的基础模型进行改进。这应能打造出这样的深度学习系统：不仅性能更佳（无论在泛化能力还是准确率方面），而且足够透明、易于审查，从而支持具有实际意义的审视。

VI. 我们的经验总结

机器学习/深度学习的黑箱性质是一个根本性问题。在此背景下，以系统为中心的视角有助于从问题的起源及实际影响出发，更深入地理解这一问题。它也有助于认识到，为什么可解释性是机器学习/深度学习系统设计的一个基本方面。这样的理解不仅可以改善学习成果，也能为改进机器学习/深度学习系统的实际部署提供有意义的见解。此类部署涵盖范围广泛，从设计方面的考量（包括增强可解释性、提升泛化能力、恰当的模型初始化与训练等），到硬件实现（设备端及受限环境下的实现）均包括在内。

VII. 致谢

作者感谢科学与工程研究委员会（SERB）、印度科学技术部及印度政府通过编号为 SRG/2020/000849 的资助项目提供资金支持。

参考文献

[1] M. Narwaria，“从白盒到黑箱的转变：信号处理教育中的挑战与机遇”，IEEE《信号处理杂志》，第38卷，第3期，第163–173页，2021年。[2] N. Xie、G. Ras、M. van Gerven 和 D. Doran，“可解释深度学习：新手指南”，arXiv，2020年。[3] X. Bai、X. Wang、X. Liu、Q. Liu、J. Song、N. Sebe 和 B. Kim，“面向高效、稳健模式识别的可解释深度学习：近期进展综述”，《模式识别》，第120卷，文章编号108102，2021年。 [在线]。网址：https://www.sciencedirect.com/science/article/pii/S0031320321002892[4] M. Zhang、H. You、P. Kadam、S. Liu 和 C.-C. J. Kuo，“Pointhop：一种用于点云分类的可解释机器学习方法”，IEEE《多媒体汇刊》，第22卷，第7期，第1744–1755页，2020年。[5] P. Angelov 和 E. Soares，“迈向可解释深度神经网络（xdnn）”，arXiv，2019年。[6] I. Goodfellow、Y. Bengio 和 A. Courville，《深度学习》。麻省理工学院出版社，2016年。[7] J. Chen 和 X. Ran，“边缘计算中的深度学习：综述”，《IEEE会刊》，第107卷，第8期，第1655–1674页，2019年。[8] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于《2016年IEEE计算机视觉与模式识别会议（CVPR）论文集》，2016年，第770–778页。[9] M. Tancik、P. Srinivasan、B. Mildenhall、S. Fridovich-Keil、N. Raghavan、U. Singhal、R. Ramamoorthi、J. Barron 和 R. Ng，“傅里叶特征使网络能够学习低维域中的高频函数”，载于《第34届神经信息处理系统国际会议论文集》，NIPS’20系列。美国纽约州雷德胡克：Curran Associates Inc.，2020年。[10] K. Zhang、L. Van Gool 和 R. Timofte，“用于图像超分辨率的深度展开网络”，载于《2020年IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，2020年，第3214–3223页。[11] E. Bjornson 和 P. Giselsson，“深度学习在通信系统物理层中的两种应用 [讲义],”，IEEE《信号处理杂志》，第37卷，第5期，第134–140页，2020年。[12] J. Ryu、M.-H. Yang 和 J. Lim，“用于视觉分类的基于离散傅里叶变换的变换不变池化层”，载于《欧洲计算机视觉会议（ECCV）论文集》，2018年9月。[13] M. Cheung、J. Shi、O. Wright、L. Y. Jiang、X. Liu 和 J. M. F. Moura，“ 图信号处理与深度学习：卷积、池化和拓扑 ”，IEEE 《信号处理杂志》，第 37 卷，第 6 期，第 139–149页，2020年。[14] J. Engel、L. Hantrakul、C. Gu 和 A. Roberts，“DDSP：可微分数字信号处理”，CoRR，2020年。 [在线]。网址：http://arxiv.org/abs/1907.07374[15] X. Li 和 F. Li，“利用卷积滤波器统计量检测深度网络中的对抗样本”，载于《2017年IEEE国际计算机视觉会议（ICCV）论文集》，2017年，第5775–5783页。[16] T. Spinner、U. Schlegel、H. Schfer 和 M. El-Assady，“explainer：用于交互式可解释机器学习的可视分析框架”，IEEE《可视化与计算机图形学汇刊》，第26卷，第1期，第1064–1074页，2020年。[17] J. Choo 和 S. Liu，“可解释深度学习的可视分析”，IEEE《计算机图形学与应用》，第38卷，第4期，第84–92页，2018年。

[18]Y. Xie、X. A. Chen 和 G. Gao，“为医学诊断中的可解释智能系统勾勒设计空间”，载于《与第24届ACM智能用户界面大会（ACM IUI 2019）同期举办的ACM IUI 2019研讨会联合论文集》，美国洛杉矶，2019年3月20日，CEUR研讨会论文集系列，C. Trattner、D. Parra 和 N. Riche 编，第2327卷。CEUR-WS.org，2019年。 [在线]。网址：http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-18.pdf