## 1 INTRODUCTION

近年来，深度神经网络已经显示出构建安全应用程序的巨大潜力。到目前为止，研究人员已成功应用深度神经网络训练分类器进行恶意软件分类[2,16,21,48,68]，二进制逆向工程[15,52,71]和网络入侵检测[24,62]，都达到了极高的准确度。虽然高度准确，但安全从业者对深度学习模型缺乏透明度感到担忧，因此在安全和安全关键领域广泛采用深度学习分类器犹豫不决。更具体地说，深度神经网络可以容易地包含数十万甚至数百万个神经元。该网络一旦使用大量数据集进行培训，就可以提供高分类精度。然而，网络的高复杂性也导致模型的“可解释性”低。很难理解神经网络有多深入做出某些决定。缺乏透明度为建立模型信任或有效排除分类错误带来了主要障碍。为了提高深度神经网络的透明度，研究人员开始研究解释分类结果的解释方法。大多数现有工作集中在非安全应用程序，如图像分析或自然语言处理（NLP）。图1a显示了一个示例。给定输入图像，解释方法通过精确定位最具影响力的特征来解释分类结果。常见方法涉及在网络中运行前向传播[17,19,32,76]或向后传播[3,50,53]以推断重要特征。更高级的方法[34,45]在“黑盒”设置下产生解释，其中不知道分类器细节。基本思想是使用线性模型近似局部决策边界来推断重要特征。遗憾的是，现有的解释方法不能直接应用于安全应用程序。首先，大多数现有方法被设计用于图像分析，其优选使用卷积神经网络（CNN）。但是，CNN模型在安全领域并不是很受欢迎。诸如二进制反向工程和恶意软件分析之类的安全应用程序具有高级特征依赖性（例如，二进制代码序列），或者需要高可伸缩性。因此，回归神经网络（RNN）或多层感知器模型（MLP）被更广泛地使用[15,21,52,68]。到目前为止，还没有解释方法在RNN上运行良好。其次，现有方法仍然存在低解释保真度，正如我们在§5中的实验所证实的那样。这对于图像分析可能是可接受的，但可能会在安全应用程序中造成严重问题。例如，在图1a中，突出显示的像素并不完全准确（特别是在边缘区域），但足以提供直观的理解。但是，对于安全应用程序，例如二进制分析，错误地突出显示一个字节的代码可能会导致严重的误解或解释错误。我们的设计。在本文中，我们寻求开发一种专用于安全应用的新颖，高保真解释方法。我们的方法在黑盒环境下工作，并引入专门的设计来解决上述挑战。给定输入数据实例x和诸如RNN的分类器，我们的方法旨在识别对x的分类具有关键贡献的一小组特征。这是通过在x附近生成目标分类器的决策边界的局部近似来完成的。为了显着提高近似的保真度，我们的方法不再假设局部检测边界是线性的，也不假设特征是独立的。这些是现有模型[34,45]做出的两个关键假设，这些假设在安全应用中经常被违反，导致解释保真度差。相反，我们引入了一种新的方法来近似非线性局部边界，基于融合回归模型[27]增强融合套索[64]。我们的设计基于两个关键见解。首先，理论上，混合回归模型可以在给定足够数据的情况下逼近线性和非线性决策边界[35]。这使我们能够灵活地优化非线性边界的局部近似并避免大的拟合误差。其次，“融合套索”是通常用于捕获特征依赖性的惩罚术语。通过将融合套索添加到学习过程中，混合回归模型可以将特征作为一组进行捕获，从而捕获相邻特征之间的依赖关系。通过这种方式，我们的方法通过同时保持深度学习模型的局部非线性和特征依赖性来产生高保真度的解释结果。为方便起见，我们将我们的方法称为“使用非线性近似的局部解释方法”或LEMNA。评估。为了证明我们的解释模型的有效性，我们将LEMNA应用于两个有前景的安全应用程序：分类PDF恶意软件[55]，并检测函数开始逆向工程二进制代码[52]。分类器分别对10,000个PDF文件和2,200个二进制文件进行训练，两者的准确度均达到98.6％或更高。我们应用LEMNA来解释他们的分类结果并开发一系列保真度量指标来评估解释的正确性。通过直接比较近似检测边界和实际检测边界，或运行端到端特征测试来计算保真度度量。结果表明，LEMNA在所有不同的分类器和应用程序设置中明显优于现有方法。除了有效性评估之外，我们还展示了安全分析师和机器学习开发人员如何从解释结果中受益。首先，我们通过解释分类器如何做出正确的决定来证明LEMNA可以帮助建立信任。特别是，对于二进制和恶意软件分析，我们证明分类器已经成功地在相应的领域中学习了许多众所周知的启发式和“黄金规则”。其次，我们说明LEMNA可以从分类器中提取“新知识”。这些新的启发式方法难以直接手动汇总，但一旦领域专家被LEMNA提取，就会对其有直观的意义。最后，利用LEMNA的能力，分析师可以解释为什么分类器会产生错误。这允许分析师自动生成目标补丁通过增加每个可解释错误的训练样本，并通过有针对性的再训练提高分类器的性能。贡献。我们的论文做出了三项重要贡献

•我们设计和开发LEMNA，这是一种基于深度学习的安全应用的专业解释方法。使用融合套索增强的混合回归模型，LEMNA为包括RNN在内的一系列深度学习模型生成高保真解释结果。

•我们使用两种流行的安全应用程序评估LEMNA，包括PDF恶意软件分类和二进制逆向工程中的功能启动检测。我们提出了一系列“保真度”指标来量化解释结果的准确性。我们的实验表明，LEMNA的表现优于现有的解释方法。

•我们演示了解释方法的实际应用。对于二进制分析和恶意软件检测，LEMNA揭示了分类器为何做出正确和错误决策的原因。我们提出了一种简单的方法来自动将洞察转换为可操作的步骤，以修补分类器的目标错误。据我们所知，这是第一个专门为安全应用和RNN定制的解释系统。我们的工作只是提高模型透明度的第一步，以便更有效地测试和调试深度学习模型。通过使决策过程可以解释，我们的努力可以为关键应用构建可靠的深度学习系统做出积极贡献

## 2、EXPLAINABLE MACHINE LEARNING

在本节中，我们首先介绍可解释的机器学习的背景，然后讨论现有的解释技术。 接下来，在第3节中，我们使用深度学习模型介绍关键安全应用程序，并讨论为什么现有解释技术不适用于安全应用程序。

## 2.1、Problem Definition

可解释的机器学习旨在为分类结果提供可解释的解释。更具体地，给定输入实例x和分类器C，分类器将在测试时间期间为x分配标签y。然后，解释技术旨在说明为什么实例x被归类为y。这通常涉及识别一组重要特征，这些特征对分类过程（或结果）做出了重要贡献。如果所选特征可由人类分析师解释，则这些特征可以提供“解释”。图1显示了图像分类和情感分析的示例。分类器决策可以通过选择的特征（例如，突出显示的像素和关键字）来解释。在本文中，我们关注深度神经网络以开发用于安全应用的示例方法。到目前为止，大多数现有的解释方法都是为图像分析或NLP而设计的。我们将它们分类为“whitebox”和“blackbox”方法，并描述它们的工作方式。

## 2.2、Whitebox Explanation Methods

大多数现有的解释技术在白盒设置下工作，其中模型体系结构，参数和训练数据是已知的。这些技术也称为深度解释方法，主要为CNN设计。他们利用两个主要策略来推断特征重要性：（1）基于前向传播的输入或结构遮挡; （2）基于梯度的反向传播。我们将在下面讨论这些技术。基于前向传播的方法。给定输入样本，关键思想是扰乱输入（或隐藏的网络层）并观察相应的变化。背后的直觉是，扰乱重要特征更有可能导致网络结构和分类输出的重大变化。现有方法要么取消特征子集，要么删除网络的中间部分[17,32,74,76]。最近的一项工作[19]将这一想法扩展到检测对抗性的例子（即，旨在引起分类错误的恶意输入）。基于向后传播的方法。基于反向传播的方法利用深度神经网络的梯度来推断特征重要性。梯度可以是关于输入或隐藏层的分类器输出的偏导数。通过将输出传播回输入，这些方法直接计算输入要素的权重。对于图像分类器，基本方法是使用相对于图像[54,57]或视频帧[18]中的输入像素的输出梯度来计算特征“显着图”。后来的工作通过逐层应用显着图[3]或映射像素组[50]来改进这一想法。基于向后传播的方法面临“零梯度”的挑战。在神经网络内部，激活函数通常具有饱和部分，并且相应的梯度将变为零。零梯度使“显着图”难以（如果不是不可能）回溯重要特征。最近的作品[53,59]试图通过近似来解决这个问题。然而，这牺牲了解释的保真度[34]

## 2.3 Blackbox Explanation Methods

Blackbox解释方法不需要关于分类器内部的知识，例如网络体系结构和参数。相反，他们将分类器视为“黑盒子”，并通过发送输入和观察输出（即模型归纳方法）对其进行分析。这一类别中最具代表性的系统是LIME [45]。给定输入x（例如，图像），LIME系统地扰动x以从特征空间中的x的附近区域获得一组人造图像（参见图2中的x'和x“）。然后，LIME将人工图像馈送到目标分类器f（x）以获得标签，并使用标记数据拟合线性回归模型д（x）。该д（x）旨在近似特征空间中输入图像附近的f（x）的小部分。 LIME假设输入实例附近的分类边界的局部区域是线性的，因此使用线性回归模型局部表示由f（x）做出的分类决策是合理的。线性回归是自我解释的，因此LIME可以根据回归系数确定重要特征。最近的工作SHAP [34]试图通过向人工生成的数据样本添加权重来扩展LIME。其他工作建议使用其他线性模型（例如，决策树[6]和决策集[31]）来递增地近似目标检测边界。作为旁注，我们想澄清机器学习解释与主成分分析（PCA）[26]，稀疏编码[39]或卡方统计[49]等特征选择方法完全不同。解释方法旨在识别特定输入实例x的关键特征，以具体解释实例x如何被分类。另一方面，诸如PCA的特征选择方法通常在训练之前应用于整个训练数据以减小特征维度（以加速训练或减少过度拟合），这无法解释如何做出特定的分类决策。

## 3、解释安全应用

虽然深度学习显示了构建安全应用程序的巨大潜力，但相应的解释方法却大都落后了。 因此，缺乏透明度会降低信任度。 首先，如果安全从业者不了解如何做出关键决策，他们可能不会信任深度学习模型。 其次，如果安全从业者不能解决分类错误（例如，由有偏见的训练数据引入的错误），则担心的是这些错误可能在实践中稍后被放大。 在下文中，我们介绍了两个关键的安全应用程序，其中深度学习最近取得了成功。 然后我们讨论为什么现有的解释方法不适用于安全应用程序

## 3.1、Deep Learning in Security Applications

在本文中，我们关注两类重要的安全应用程序：二进制逆向工程和恶意软件分类。

二元逆向工程。深度学习在二元分析中的应用包括识别函数边界[52]，精确定位函数类型签名[15]和跟踪类似的二进制代码[71]。更具体地说，使用双向RNN，Shin等人。改进功能边界识别，实现近乎完美的性能[52]。 Chua等人。还使用RNN准确地追踪二进制中的参数和函数类型[15]。最近，徐等人。使用MLP编码控制流图以查明易受攻击的代码片段[71]。

恶意软件分类。现有工作主要使用MLP模型进行大规模恶意软件分类。例如，研究人员已经训练MLP在二进制代码级别检测恶意软件[48]并对Android恶意软件进行分类[2,21]。最近，Wang等人。 [68]提出了一种抗对抗的神经网络，用于根据审计日志检测恶意软件[7]。一个关键的观察是，与CNN相比，这些安全应用更广泛地采用RNN和MLP。原因是RNN旨在处理顺序数据，这在处理长二进制代码序列时表现非常出色。特别地，双向RNN可以捕获每个十六进制之间的输入序列中的双向依赖性[52]。对于恶意软件分类，MLP因其高效率而被广泛使用。另一方面，CNN在图像上表现良好，因为它可以利用2D图像上特征的分组效果[30]。这些安全应用程序没有这样的“矩阵式”数据结构，无法从使用CNN中受益

## 3.2、Why Not Existing Explanation Methods

将现有解释方法直接应用于安全应用程序存在关键挑战。在表1中，我们总结了所需的属性，以及现有方法无法提供它们的原因。支持RNN和MLP。上述安全应用程序的模型选择与现有解释方法之间存在明显的不匹配。大多数现有的解释方法是为CNN设计的，以便与图像分类器一起使用。但是，如第3.1节所述，我们利益的安全应用主要采用RNN或MLP。由于模型不匹配，现有的解释方法不太适用。例如，包括“显着图”[3,18,54,57]和激活差异传播[53]的反向传播方法需要对CNN的卷积层和池化层进行特殊操作，这在RNN或MLP中不存在1。诸如LIME之类的Blackbox方法也不能很好地支持RNN（稍后通过我们的实验验证）。像LIME这样的方法假设特征是独立的，但RNN违反了这一假设，RNN明确地模拟了顺序数据的依赖性。

支持局部非线性决策边界。大多数现有方法（例如，LIME）假设决策边界的局部线性。但是，当局部决策边界是非线性的时，大多数复杂网络都是如此，那些解释方法会产生严重错误。图3a示出了围绕x的决策边界是高度非线性的示例。换句话说，线性部分严格限制在非常小的区域。典型的采样方法很容易击中线性区域之外的人工数据点，使得线性模型难以逼近x附近的决策边界。稍后在我们的实验（第5节）中，我们确认简单的线性近似将显着降低解释保真度。

支持Blackbox设置。虽然whitebox和blackbox方法都有其应用场景，但黑盒方法仍然更适用于安全应用程序。值得注意的是，人们使用预先训练的模型并不少见（例如，Dyninst [5]中的“双向RNN”[52]，“前缀树”），其中详细的网络架构，参数或训练数据并非全部可用。尽管可以强制一些前向传播方法在黑盒设置下工作（通过放弃对中间层的观察），但这将不可避免地导致性能下降。

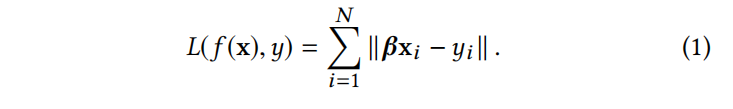
摘要。在本文中，我们旨在通过开发安全应用程序的专用解释方法来弥合差距。我们的方法旨在在黑盒设置下工作，并有效支持流行的深度学习模型，如RNN，MLP和CNN。更重要的是，该方法需要实现更高的解释保真度以支持安全应用程序

4

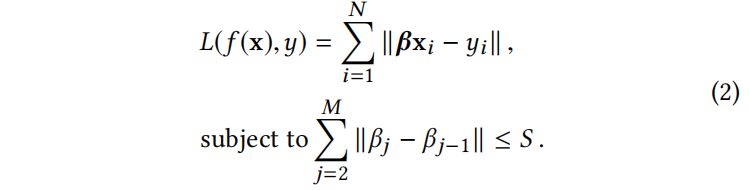
为实现上述目标，我们设计并开发了LEMNA。在高级别，我们将目标深度学习分类器视为黑盒，并通过模型近似得出解释。为了提供高保真度的解释，LEMNA需要采用与现有方法截然不同的设计路径。首先，我们引入融合套索[64]来处理安全应用程序和RNN中经常遇到的特征依赖性问题（例如，时间序列分析，二进制代码序列分析）。然后，我们将融合套索集成到混合回归模型[28]中，以近似局部非线性决策边界，以支持复杂的安全应用。在下文中，我们首先讨论使用融合套索和混合回归模型的设计选择背后的见解。然后，我们描述技术细节，将它们集成到单个模型中，同时处理特征依赖性和局部非线性。最后，我们引入了额外的步骤来利用LEMNA来获得高保真度的解释

## 4.1、Insights behind Our Designs

融合的套索。融合套索是通常用于捕获特征依赖性的惩罚术语，并且对于处理诸如RNN的深度学习模型中的依赖特征是有用的。在高级别，“融合套索”迫使LEMNA将相关/相邻特征组合在一起以产生有意义的解释。下面，我们介绍这种直觉的技术细节。为了从一组数据样本中学习模型，机器学习算法需要最小化损失函数L（f（x），y​​），其定义模型的真实标签和预测标签之间的不相似性。例如，为了从具有N个样本的数据集中学习线性回归模型f（x）=βx+ε，学习算法需要使用最大似然估计（MLE）来最小化关于参数β的以下等式[38]。



这里，xi是训练样本，由M维度特征向量（x1，x2，...，xM）T表示。 xi的标记表示为yi。矢量β=（β1，β2，...·βM）包含线性模型的系数。 ∥·∥是衡量模型预测与真实标签之间差异的L2范数。 融合套索是一个惩罚术语，可以引入学习算法使用的任何损失函数。以线性回归为例。融合套索表现为对系数施加的约束，即，

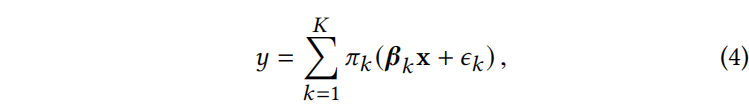


当学习算法使损失函数最小化时，融合套索限制分配给小阈值S（即，超参数）内的相邻特征的系数的不相似性。结果，惩罚项迫使学习算法为相邻特征分配相等的权重。直观地，这可以被解释为强制学习算法将特征作为组，然后基于特征组学习目标模型。安全应用程序（例如时间序列分析和代码序列分析）通常需要使用RNN明确建模顺序数据的特征依赖性。得到的分类器基于特征的共现来做出分类决策。如果我们使用标准线性回归模型（例如，LIME）来推导解释，我们就无法正确地近似局部决策边界。这是因为线性回归模型无法捕获特征依赖性并独立处理它们。通过在近似局部决策边界的过程中引入融合套索，我们期望得到的线性模型具有以下形式



将功能组合在一起，因此重要功能可能被选为一组或多组。在LEMNA中对此过程进行明确建模有助于获得更准确的解释，特别是对于RNN做出的决策。我们使用图1b中的情绪分析示例进一步解释了这个想法。在融合套索的帮助下，回归模型将共同考虑相邻特征（例如，句子中彼此相邻的单词）。在得出解释时，我们的模型不仅仅产生一个单词“not”2，而是可以准确地捕获“不值得”的短语作为情绪分析结果的解释

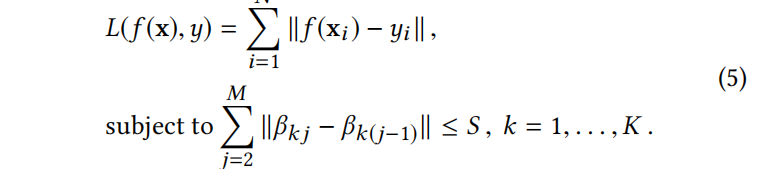
混合回归模型允许我们更准确地近似局部非线性决策边界。如图3b所示，混合回归模型是多个线性回归模型的组合，这使得它更具表现力来执行近似



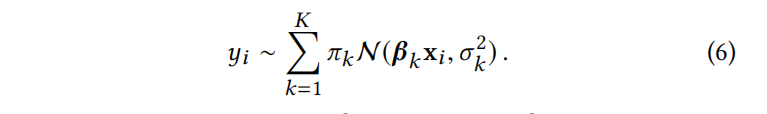
其中K是一个超参数，表示混合模型中组合的线性分量的总数; πk表示分配给该相应组件的权重。给定足够的数据样本，无论分类器是否具有线性或非线性决策边界，混合回归模型几乎可以完全逼近决策边界（使用有限的线性模型集）[35]。因此，在深度学习解释的背景下，混合回归模型可以帮助避免上述非线性问题并得出更准确的解释。为了说明这个想法，我们使用图3中的示例。如图3a所示，标准线性近似不能保证在输入x周围采样的数据仍然保留在局部线性区域中。这很容易导致不精确的近似和低保真度的解释。我们在图3b中的方法用多边形边界近似局部决策边界，其中每条蓝线代表一个独立的线性回归模型。产生解释的最佳线性模型应该是通过数据点x的红线。通过这种方式，近似过程可以产生最佳线性回归模型，用于精确定位重要特征作为解释

## 4.2、Model Development

接下来，我们将这些设计见解转换为功能性解释系统。我们介绍了将融合套索集成到混合回归模型学习过程中的技术步骤，以便我们可以同时处理特征依赖和决策边界非线性。从技术上讲，我们需要通过最小化以下等式得出混合回归模型

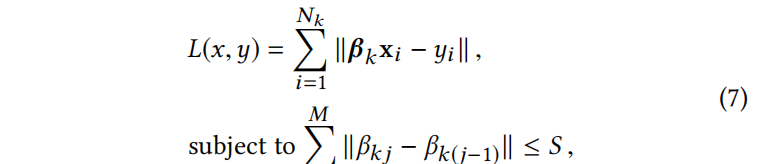


其中f（·）表示等式（4）中所示的混合回归模型，βkj表示与第j个特征相关的第k个线性回归模型中的参数。与标准线性回归不同，我们的优化目标是难以处理的，我们不能简单地利用MLE来执行最小化。为了有效地估计混合回归模型的参数，我们使用另一种方法。首先，我们以概率分布的形式表示混合回归模型



然后，我们将π1：K，β1：K和σ21：K视为参数3。通过猜测这些参数，我们初始化它们的值，从而通过使用期望最大化（EM）[37]执行参数估计，这是一种通过重复执行两个步骤（E-Step和M-Step）来估计参数的算法。在下文中，我们将简要介绍如何在我们的问题中使用此EM算法。更多细节可以在附录-A中找到。

在等式（6）中，yi遵循组合K个高斯分布的分布，并且这些分布中的每一个具有平均值βkxi和方差σ2k。在E-Step中，我们通过遵循学习普通混合回归模型中应用的标准程序将每个数据样本分配给一个高斯分布。基于先前E步骤中分配的数据样本，我们然后重新计算参数π1：K，β1：K和σ21：K。对于参数π1：K和σ21：K，重新计算仍然遵循普通混合模型学习所使用的标准程序。但是，对于β1：K中的每个参数，重新计算遵循定制过程。即通过最小化关于βk的以下等式来计算βk



其中Nk是指分配给第k个分量的样本数。这里，这种重新计算定制背后的原因是必须对参数β1：K施加融合套索，以便给混合回归模型赋予处理特征依赖性的能力。我们可以观察到，上面的等式与等式（2）中所示的形式具有相同的形式。因此，我们可以通过MLE最小化方程，从而计算参数β1：K的值。按照EM算法的标准程序，我们重复执行E步和M步。在达到稳定性之前（即，从E步骤到M步骤的高斯分布变化不大），我们输出混合回归模型。注意，我们通过遵循普通混合模型学习中应用的标准方法将σ21：K转换为模型参数ε1：K。

## 4.3、应用解释模型

通过增强的混合回归模型，我们现在讨论如何推导出深度学习分类器的高保真解释。近似局部决策边界。给定输入实例x，生成解释的关键是近似目标分类器的局部决策边界。最终产品是一个“可解释的”线性模型，允许我们选择一小组顶部特征作为解释。为此，我们首先按照[45]中描述的方法在本地（大约x）合成一组数据样本。这个想法是随机地使x的特征子集无效。使用合成数据样本的语料库，我们然后估计局部决策边界。有两种可能的方案：一种是训练单一混合回归模型以执行多类分类;另一种方案是训练多个混合回归模型，每个模型执行二元分类。出于效率考虑，我们选择第二种方案并对附录B进行更严格的分析。得出解释。给定输入数据实例x及其分类结果y，我们现在可以生成解释作为x分类的一小组重要特征。更具体地说，我们获得了由融合套索增强的混合回归模型。然后，我们从该混合模型中识别出具有局部决策边界最佳近似的线性分量。线性模型中的权重（或系数）可用于对特征进行排名。选择一小组顶部特征作为解释结果。请注意，LEMNA旨在同时处理非线性和特征依赖性，但这并不意味着LEMNA无法使用相对独立的功能（例如，MLP或CNN）在深度学习模型上工作。事实上，LEMNA的设计提供了根据目标深度学习模型调整解释方法的灵活性。例如，通过增加超参数S（它是融合套索的阈值），我们可以放宽对参数β1：K施加的约束，并允许LEMNA更好地处理较少依赖的特征。在第5节中，我们通过将LEMNA应用于基于RNN和MLP的安全应用程序来演示通用性水平