CNN 辅助隐写术——将机器学习与成熟的隐写技术相结合

ANDREW HAVARD，Darwin Deason网络安全研究所，南卫理公会大学，美国 THEODORE MANIKAS，Darwin Deason网络安全研究所，南卫理公会大学，美国 ERIC C. LARSON，Darwin Deason网络安全研究所，南卫理公会大学，美国 MITCHELL A. THORNTON，Darwin Deason网络安全研究所，南卫理公会大学，美国

arXiv:2304.12503v1 [cs.CR] 25 Apr 2023

我们提出一种改进隐写术的方法，通过增强隐写媒体抵御隐写分析发现的能力，提高其隐蔽性。我们的方法通过加入隐写辅助卷积神经网络（SA-CNN），增强了一类隐写方法。先前的研究表明，使用经训练的神经网络作为隐写分析器并将其应用于隐写图像，能够成功发现其中是否存在隐藏信息。我们的结果表明，在生成隐写图像时采用SA-CNN，会降低此类隐写分析器的有效性。我们还探讨了将SA-CNN 的所有可能输出表示在较小的离散空间中，而非连续空间中的利弊。SA-CNN 使某些类别的参数化隐写算法能够根据待嵌入信息的载体媒体特征进行定制。因此，SA-CNN 具有自适应性，因为它能针对每个特定的载体媒体实例对核心隐写算法进行专门配置。我们采用一种近期的隐写技术 S-UNIWARD，在使用和不使用 SA-CNN 的情况下进行了实验，并给出了实验结果。随后，我们将两组隐写图像（分别由使用和不使用 SA-CNN 的方法生成）输入示例隐写分析器 Yedroudj-Net，并比较了结果。我们认为，将神经网络与手工设计的算法相结合，能够提高隐写算法的可靠性和适应性。

CCS概念：•**安全与隐私 → 入侵／异常检测和恶意软件缓解**；**人工免疫系统**；

附加关键词：隐写术、隐写分析、卷积神经网络、生成网络

1 引言

隐写术是将信息隐藏在某种媒介中的过程，例如图像或音频文件。与之相反的过程是隐写分析，即确定某一特定媒体实例是否包含隐藏信息，并因此判断其是否为隐写产物。隐写术的一个主要目标是抵御隐写分析，即使有人获得了包含隐藏信息的媒介，也无法检测出隐藏信息的存在。由于引入了可用作隐写媒介的新型媒体，以及改进了信息嵌入技术，隐写术领域近年来取得了进展。大多数现代隐写术方法都着重于将信息嵌入图像中。为了帮助介绍隐写术和隐写分析的基本原理，我们以数字图像文件作为媒介，或称“载体图像”。当载体图像经过隐写处理后，会转换为相应的“隐写图像”；对于人眼观察者而言，隐写图像通常与相关的载体图像无法区分。

隐写算法随着不断发展，变得日益复杂，也更能抵御隐写分析技术。例如，一种非常简单的隐写方法是：将载体图像像素的最低有效位（LSb）替换为构成隐藏信息的一个比特，从而隐藏信息。更先进的隐写方法与简单方法一样，会将载体图像的最低有效位替换为隐藏消息的对应位，但只替换图像中对应于最不显著像素的最低有效位；因此，在替换最低有效位之前，必须先确定载体图像中恰当的最不显著像素。最不显著像素通常位于图像中高频内容相对丰富的子区域，或者位于颜色和亮度相较于邻近像素有显著变化的区域。人类感知以及某些机器训练感知算法的特性是：与主要由颜色和亮度相同的像素构成的区域相比，对一组空间相邻像素的颜色和亮度稍作修改，会使该区域更难察觉细微差异。这种隐写方法的关键，在于嵌入隐藏信息之前识别载体图像中的此类区域。S-UNIWARD算法 [1] 依赖于一种代价函数：对于每个待考虑用于嵌入隐藏信息的载体图像区域，首先估算其嵌入代价，以此表示使用该特定区域是否容易导致隐写分析方法成功判定图像确实是隐写图像。S-UNIWARD算法的理论基础是使用一种称为UNIWARD 的失真函数，该函数在小波域中对图像区域进行计算，取最高频非抽取子带中小波系数相对变化量之和。因此，算法会评估每个像素的邻域是否存在不连续性，并据此指出载体图像中的边界，在这些位置不应嵌入隐藏信息。如今，隐写术领域的研究仍在继续，尽管目前开发的许多隐写分析方法都使用 S-UNIWARD 作为通用基线，以比较检测隐写内容时的错误率。

S-UNIWARD算法需要指定多个参数，以确保该方法能够可靠运行。由于无法从数据中估算参数值，因此通常必须依据一般经验法则或通过反复试验来手动设置这些值。

我们的方法使用隐写辅助卷积神经网络（SA-CNN）来修改S-UNIWARD

作者地址：Andrew Havard，ahavard@smu.edu，Darwin Deason网络安全研究所，南卫理公会大学，达拉斯，得克萨斯州，美国，75275-0001；Theodore Manikas，manikas@lyle.smu.edu，Darwin Deason网络安全研究所，南卫理公会大学，达拉斯，得克萨斯州，美国，75275-0001；Eric C. Larson，eclarson@lyle.smu.edu，Darwin Deason网络安全研究所，南卫理公会大学，达拉斯，得克萨斯州，美国，75275-0001；Mitchell A. Thornton，mitch@lyle.smu.edu，Darwin Deason网络安全研究所，南卫理公会大学，达拉斯，得克萨斯州，美国，75275-0001。

算法，使其更有效，并能根据载体图像进行自适应调整，从而增强其抵御隐写分析器检测的能力。为实现这一改进，我们修改了 S-UNIWARD 算法，使其能够指定算法中某些参数的乘数。系统会自动选择这些参数，目标是最大限度地迷惑预训练的隐写分析网络。这些乘数由卷积神经网络预测。对于连续情形，我们的 SA-CNN 会尝试直接预测最优乘数；对于离散情形，则从人工整理的列表中选择一个乘数。也就是说，我们使用损失函数更新参数，以最大化用于检测隐写图像的算法的错误率。实验结果表明，使用我们的 SA-CNN 处理载体图像后，我们能够调整隐写算法 S-UNIWARD 的适当参数 [1]，使其更能抵御常见隐写分析器（如 Yedroudj-Net [2]）的利用。

2 背景与相关工作

隐写术描述了将信息隐藏在某种媒介中的过程，例如图像或音频文件，从而使截获含有隐藏信息的媒介的人无法轻易察觉其中隐藏了任何信息。隐写分析是判断媒介中是否含有隐藏信息的过程。我们的研究聚焦于使用图像媒介进行隐写术和隐写分析，具体而言，我们使用成熟的隐写分析工具来增强现有的隐写术技术。

在隐写术中，不含任何隐藏信息的图像称为载体图像，含有隐藏信息的图像称为隐写图像。人们已经使用传统图像处理和信息论方法建立了各种方法，用于分类载体图像与隐写图像之间的差异 [3], [2]。

目前，深度学习已成为区分载体图像与隐写图像的最先进方法。在深度学习中，神经网络通常通过使用许多神经网络层进行优化，以降低目标函数中的误差。卷积神经网络（CNN）是常用的一类神经网络，通常用于处理图像，以识别并响应各种空间模式。这些网络通过反向传播优化卷积滤波器的系数，从而为给定的分类任务寻找定制的图像特征提取方式。隐写术和隐写分析方法都采用了卷积神经网络。

凭借这些优势，卷积神经网络因其可靠性和高效性，已成为这些领域事实上的标准。

最初，隐写分析是通过人工设计的算法进行的。然而，在2009年，Oplatková等人 [4] 训练前馈神经网络，以检测图像中的隐写内容。Tan和Li [5] 开始通过在隐写分析数据上训练三个卷积神经网络来解决这些问题。当时，空间富模型（SRM）作为二分类器，是隐写分析器中的前沿方法。然而，Tan认为，卷积神经网络有潜力取代SRM，因为两者的架构存在一定相似性。当时，卷积神经网络要么很快收敛到较差的局部解，要么在试图避开这些局部解时发散到其他解。他们通过先训练堆叠卷积自动编码器，然后以与堆叠卷积自动编码器相同的配置初始化卷积神经网络，成功缓解了这些问题。尽管他们未能达到SRM的效果，但其性能超过了其他卷积神经网络。Qian 等人 [6] 在此方法的基础上，应用卷积神经网络进行隐写分析，将载体图像与隐写图像进行分类。沿着这一研究方向，Ye-Net [7], 、Xu-Net [8], 和 Yedroudj-Net [2] 等多种神经网络，在针对嵌入率为0.4每像素比特数（bpp）的S-UNIWARD进行训练时，错误率分别低至31.2%、27.2%和22.8% [2]。

最近，Sharifzadeh 等人推导出了一种更具统计学依据的隐写算法，称为自适应批大小图像合并（AdaBIM） [9]。他们将载体和隐写消息建模为高斯变量，并最大化三个基于统计的隐写分析器的误差。他们还将其方案扩展到了基于代价的隐写术领域，将该方法应用于基于代价的隐写算法之上，其中包括S-UNIWARD。然而，他们的研究重点是批量隐写术，即在多幅载体图像中隐藏信息，这与我们的工作不同。此外，他们使用了不同类型的隐写分析器，采用集成分类器而非深度神经网络。

3 方法与结果

3.1 数据集收集

为了评估模型并为其设计提供依据，我们使用了 BOSS-Base 1.01 [10] 图像数据库。该图像数据库包含 10,000 张 PGM图像格式的灰度图像。我们使用该数据库，是因为它可免费获取，其中的图像采用无损格式（因此与 S-UNIWARD 兼容），并且该数据库用于训练我们研究中所使用的隐写分析器 Yedroudj-Net。由于 Yedroudj-Net 最初使用尺寸为 256x256 的图像进行发布，而 BOSSBase 1.01 图像的原始尺寸为 512x512，因此我们遵循 Yedroudj-Net 的示例，使用双线性重采样将每张图像缩放至其原始宽度和高度的一半。

3.2 隐写方法操控

对于我们的隐写嵌入算法，我们使用了免费提供的 S-UNIWARD算法。S-UNIWARD算法及其变体具有内容自适应特性，这实际上意味着它们会将信息隐藏在经过计算后最不容易暴露隐写内容的区域中。S-UNIWARD及其变体被广泛用于隐写术研究，因为它们是先进的隐写算法，能够以指定的嵌入率嵌入虚拟数据，并使隐写分析器产生较高的错误率，尤其是在嵌入率较低时。在我们的研究中，我们使用了每像素0.4比特（bpp）的嵌入率。S-UNIWARD还需要三个可调参数。S-UNIWARD的用户通常会设置这些参数以实现信息隐藏。然而，我们选择使用SA-CNN分析载体图像，自动控制这些参数。通过这种方式，SA-CNN会根据所关注的载体图像对 S-UNIWARD的参数进行微调。

我们优化的S-UNIWARD的三个参数是 σ、ϵ σ和湿代价。该值决定了S-UNIWARD算法应在多大程度上重视平滑地嵌入数据，但代价是将数据嵌入到更容易被察觉的位置。ϵ值并不严格属于S-UNIWARD方法的参数。相反，它是所使用的MATLAB实现为S-UNIWARD算法的预处理逻辑设置的常量。该值用作容差值，用于将“足够接近”0或1的值舍入到最近的整数。我们将该值作为参数处理：向S-UNIWARD算法传入一个值，使其与 ϵ相乘，并在S-UNIWARD算法中使用所得乘积。湿代价是分配给不应被修改的像素的值。实际上，该值应足够大，以确保只有被检测概率较低的像素会被修改。

3.3 针对性隐写分析器

对于我们的示例隐写分析器，我们选择了 Yedroudj-Net [2]。在其研究中，Yedroudj等人展示了其相较于其他类似神经网络隐写分析器的改进，其中针对 S-UNIWARD 的错误率，在以 0.2 bpp 和 0.4 bpp 编码的图像上分别低至 36.7% 和 22.8%。其相较于先前技术的部分改进包括使用更多的全连接层、每个卷积层包含更多特征，以及在预处理中使用包含 30 个滤波器的滤波器组。Yedroudj-Net 还提供了一个开源实现，使我们能够轻松加入自定义功能。我们将 Yedroudj-Net 用作 BOSSBase 1.01 数据集上的隐写分析器，并通过复现 Yedroudj-Net 的结果验证了我们的结果 [2]。

3.4 卷积神经网络设计与训练

对于我们的 SA-CNN，我们采用了图 1 所示的结构：3 个带有批归一化和步幅下采样的卷积层，后接全连接层，并分别使用 40%、60% 和 80% 的丢弃法。随后，我们使用 softplus 激活函数将神经网络输出映射为一个因子，通过该因子可以对 S-UNIWARD 的标准参数进行乘法调整。我们使用的损失函数如下所述。此外，我们还为 AdaM 优化器使用了指数学习率调度器。我们的训练集和验证集按照 90%/10% 的比例划分。

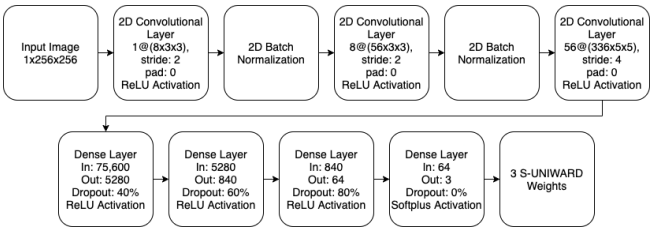

图 1. 卷积神经网络结构

我们的训练过程遵循图 2 所示的序列。我们使用 SA-CNN，为数据集中的每幅图像初步估计用于微调 S-UNIWARD 参数的值。随后，我们将这些经过“调整”的参数及其对应的载体图像输入 S-UNIWARD，以生成隐写图像。接着，我们将这些载体图像-隐写图像对作为输入传入 Yedroudj-Net 模型，并使用其输出结果，针对每幅图像评估我们的 SA-CNN 对 Yedroudj-Net 的效果。随后，我们利用这些结果进行反向传播并更新神经网络，直至其完成 140 个训练轮次（在我们的高性能计算集群 SMU Maneframe II 上耗时约七天）。然后，我们将训练后 Yedroudj-Net 的错误率与 [2] 在同一数据集上报告的基线报告错误率进行比较。Yedroudj-Net 模型的错误率越高，意味着我们的模型输入使 S-UNIWARD 更加有效，因为隐写嵌入更难被检测。因此，我们的目标函数是最大化 Yedroudj-Net 模型的错误率。

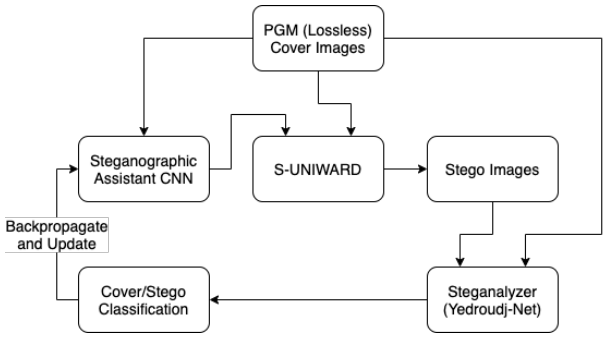

图 2. 神经网络训练方法

3.5 实验结果

在训练 SA-CNN 时，我们跟踪训练集和验证集的错误率，以了解神经网络是否出现过拟合。在训练过程中，隐写分析器的错误率越高，表示隐写分析器越经常受到误导，也就是说，我们的 SA-CNN 提高了 S-UNIWARD 的有效性。每当验证集上的隐写分析器错误率达到新的全局最高值时，我们就保存模型。我们在整个训练周期内跟踪隐写分析器的训练错误率。随后，我们遍历整个验证集并跟踪隐写分析器分类的准确性，以测量验证准确率。在评估中，我们使用来自其他图像数据库的数据来了解 SA-CNN 的性能。

3.5.1 相似数据集。在初步评估中，我们提出以下问题：使用与训练数据集相似的图像时，我们的 *SA-CNN* 是否也会出现错误率上升？为回答这个问题，我们对 BOSSbase 数据集进行了处理：不再将图像缩小到原始宽度和高度的一半，而是裁剪出图像右下角的四分之一区域，使图像与训练数据相似但并不相同。随后，我们将整个数据集通过 S-UNIWARD 和熟悉的 Yedroudj-Net 模型，以计算基准错误率；接着使用 SA-CNN、S-UNIWARD 和 Yedroudj-Net，计算各模型的混淆矩阵和错误率。从混淆矩阵（表1中 SA-CNN 和熟悉模型两列）可以看出，当分类器判断载体图像是否为隐写图像时，新的错误率约为 36.3%，比报告的基准错误率 28.2% 高出 8.1 个百分点。这意味着 Yedroudj-Net

模型更容易受到欺骗，这表明对于该数据集，SA-CNN 与 S-UNIWARD 结合使用比在原始测试集上单独使用 S-UNIWARD 更有效。

为进一步验证我们的结果，我们按照之前的相同步骤训练了第二个 Yedroudj-Net 模型，但保持已训练的 SA-CNN 模型不变。进行第二次试验旨在帮助我们了解：SA-CNN 是否只是在利用某一个已训练 Yedroudj-Net 模型的具体特征，还是确实能够针对多个隐写分析器改进 S-UNIWARD 算法。从表1（右）可以看出，错误率相近：错误率为 37.2%（比基线高 14.4 个百分点），而原始错误率为 36.3%。这表明，SA-CNN 并非只是在利用某个特定 Yedroudj-Net 模型中的伪影。

3.5.2 不相似数据集。 我们现在提出以下问题：我们的 *SA-CNN* 模型能否泛化到与训练数据不相似的数据集？为回答这一问题，我们从 ImageNet ILSVRC 2010 数据集随机选取了 10,000 张图像 [11]（这些图像均未出现在 BOSSBase 数据库中）。我们将这些图像调整为相同的 256x256 像素大小，并将图像从 RGB JPEG 图像转换为 Yedroudj-Net 和 SA-CNN 的训练所使用的灰度 PGM 图像。我们使用这些图像对 Yedroudj-Net 进行了两次测试：一次使用 S-UNIWARD，另一次使用我们的 SA-CNN+S-UNIWARD。我们没有使用这些数据重新训练 Yedroudj-Net 或 SA-CNN 模型；这些数据仅用于评估。此外，我们使用了第二个 BOSSBase Yedroudj-Net 模型，而我们的 SA-CNN 并未针对该模型进行训练。根据我们的结果（表2），我们发现，在一组不相关的图像上，我们的 SA-CNN 将 S-UNIWARD 的性能从 46.0% 的错误率提升至 49.1%。这表示错误率提高了 3.1 个百分点，而此前观察到的提升幅度分别为 13.5 和 14.4 个百分点。尽管从数值上看提升较小，但这一错误率增幅仍然十分有意义，因为对于隐写算法而言，理想性能是使隐写分析器达到 50.0% 的错误率；在这一情况下，隐写分析器在检测隐写图像方面不会比随机猜测更加有效。Yedroudj-Net 模型在这些新数据上的较差表现表明，它对 BOSSBase 1.01 数据集过拟合，但我们仍然能够将其性能提升至接近理想水平。

3.5.3 σ 与 $\sigma + \epsilon +$ 湿代价。接下来，我们要探究：我们的 ϵ *SA-CNN* 模型设定的值和湿代价值是否会影响模型的有效性？为回答这一问题，我们仅允许 SA-CNN 修改 S-UNIWARD 在嵌入过程中的 σ 值，而将 ϵ 和湿代价值保留为默认值。对于隐写分析器，我们使用了第二个 Yedroudj-Net 模型，该模型在训练期间未与我们的 SA-CNN 模型进行交互。对于数据集，我们使用了裁剪后的 BOSSBase 数据集。我们发现，修改 ϵ 和湿代价值确实会对 SA-CNN 的有效性产生显著影响（见表 3）。我们认为，这可能是因为 SA-CNN 利用 ϵ 和湿代价值，迫使 S-UNIWARD 将少量载荷嵌入到计算得出的代价极高的区域中；嵌入的信息不足以被可靠检测，但足以减少嵌入到传统嵌入区域中的载荷量，从而使载荷更难以被辨识。

为验证这一假设，我们展示了一些使用 SA-CNN+S-UNIWARD 生成的图像的可视化结果。图3展示了来自 ImageNet 的载体图像，以及分别由 S-UNIWARD 和 SA-CNN+S-UNIWARD 生成的隐写图像，同时还展示了将变化放大八倍后的图像。每组图像的载体图像位于第一列，随后第二列和第三列分别为 S-UNIWARD 图像及其放大后的变化，最后第四列和第五列分别为 SA-CNN+S-UNIWARD 图像及其放大后的变化。

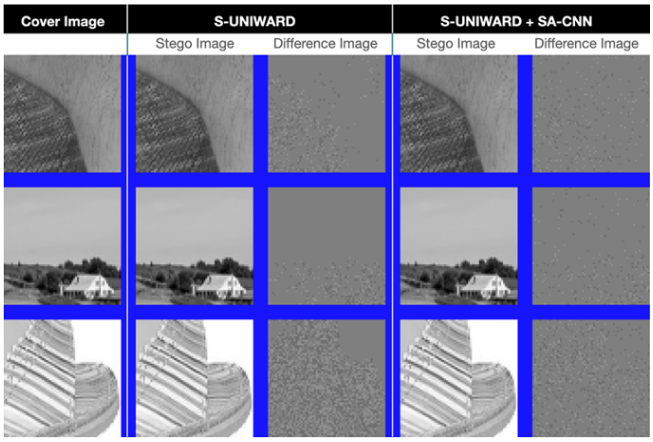

图3。我们 SA-CNN 方法的示例图像结果

总体而言，我们发现，SA-CNN+S-UNIWARD 生成的隐写图像比仅使用 S-UNIWARD 生成的图像嵌入数据更加平滑。SA-CNN 的输出结果支持这一点，因为它始终建议使用高于默认值的 σ 值。虽然可以在不被检测到的情况下将更多信息隐藏在高频区域中，但我们认为，默认的 S-UNIWARD 在载体图像的高频区域中嵌入了过多数据。隐写分析器因此更有可能检测到这些多余数据，并将图像正确分类为隐写图像。第三行可能看起来意味着，SA-CNN 鼓励 S-UNIWARD 将信息嵌入通常不会被选作信息隐藏区域的位置。这一结果很有趣，因为它意味着，由于隐写分析器往往会忽略隐藏在平滑区域中的信息，因此在这些区域中隐藏信息是一种欺骗现有隐写分析器的有效方法。然而，根据表1中的结果，我们认为，训练出的 Yedroudj-Net 模型对饱和区域中的变化不太敏感。我们认为，当针对更加关注高度饱和区域中嵌入载荷的隐写分析器进行训练时，SA-CNN 将适应该隐写分析器，并更加倾向于将载荷嵌入高频区域。

3.5.4 离散输出空间。 最后，我们提出这样一个问题：我们能否通过离散输出空间实现相当的结果？ 这种方法的优势如下。首先，我们可以预先计算离散空间中的每个可能输出，从而节省神经网络训练期间的时间。其次，将神经网络的任务从回归任务简化为分类任务，可能使神经网络在更少的训练轮次内收敛。然而，这种方法也存在一些缺点。首先，需要在训练神经网络之前生成每个可能的输出；尽管这项任务很容易并行化，但它确实需要大量存储空间，尤其是对于 S-UNIWARD算法所使用的无损图像而言。然而，如果所需存储空间过大，以至于超过了训练时间缩短所带来的收益，那么就不需要预先计算所有可能的输出。

表1。针对裁剪后的 BOSSbase 数据集上两个不同 Yedroudj-Net 模型的混淆矩阵

	仅使用 S-UNIWARD		SA-CNN 与熟悉模型 SA-CNN 与不熟悉模型			
	载体类	隐写类	载体类	隐写类	载体类	隐写类
载体图像	35.7%	14.3%	32.2%	17.8%	32.4%	17.6%
隐写图像	13.9%	36.1%	18.5%	31.5%	19.6%	30.4%
错误率	28.2%		36.3%		37.2%	

表 2。ILSVRC 数据集的混淆矩阵

	仅 S-UNIWARD		S-UNIWARD + SA-CNN	
	载体类	隐写类	载体类	隐写类
载体图像	25.0%	25.0%	23.8%	26.2%
隐写图像	21.0%	29.0%	22.9%	27.1%
错误率	46.0%		49.1%	

表 3. 不同 S-UNIWARD 参数的混淆矩阵

参数

	$\sigma + \epsilon +$ 湿代价		σ	
	载体类	隐写类	载体类	隐写类
载体图像	32.4%	17.6%	42.8%	7.2%
隐写图像	19.6%	30.4%	8.3%	41.7%
错误率	37.2%		15.5%	

这是一项容易并行化的任务，但确实需要大量存储空间，尤其是对于 S-UNIWARD算法所使用的无损图像而言。然而，如果所需存储空间过大，以至于超过了训练时间缩短所带来的收益，那么就不需要预先计算所有可能的输出。此外，为离散输出空间训练的神经网络，其结果可能不如为连续输出空间训练的神经网络强，因为神经网络的最优输出可能无法在离散空间中表示。

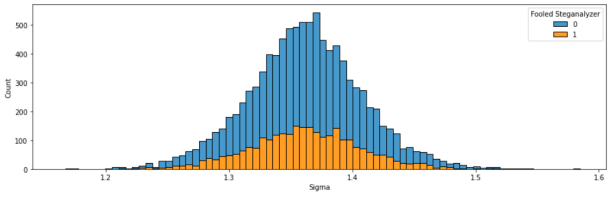

图 4. 连续模型 σ 分布

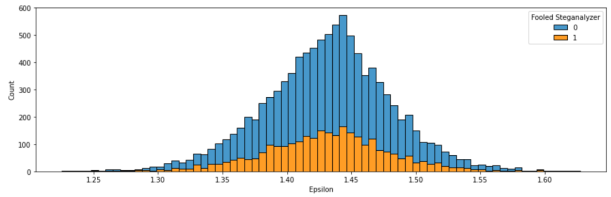

图 5. 连续模型 ϵ 分布

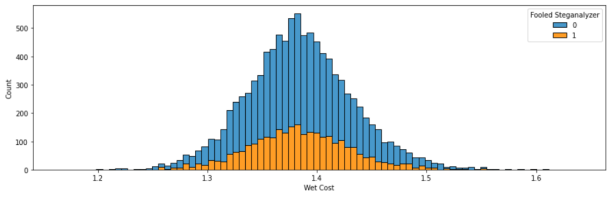

图 6. 连续模型湿代价分布

为了确定离散空间模型应能够从哪些可能的输出中进行选择，我们研究了连续空间模型的输出分布。每种输出类型（ σ、ϵ和湿代价）均呈正态分布，都以正常值的 1.4 倍为中心，范围介于正常值的 1.2 倍和 1.6 倍之间。基于这一信息，我们为离散空间选择并微调了 1.3 到 1.5 之间的值，最终得到以下 13 个值的分布：1.3875、1.1、1.3、1.325、1.35、1.3625、1.375、1.4、1.4125、1.425、1.4375、1.45、1.475、1.5.....。由于 σ、 ϵ和湿代价值使用相同的分布，因此输出离散空间的大小为 $13 \times 13 \times 13$。下面给出了连续模型各项输出分布的若干直方图，我们正是以此为依据确定离散模型的分布。

我们使用三项指标评估了模型在 BOSSBase 数据集上的离散空间输出：存储空间、每个训练周期的训练时间和验证准确率。我们将每项指标与连续空间中的对应指标进行了比较。

首先，我们可以看到，连续空间模型所需的存储空间约为数百兆字节，而

表4. 连续与离散空间输出的比较

	存储空间	训练时间	验证错误率
连续	704 MB	3796秒 ± 20秒	37.2%
隐写图像	2.8 TB	723秒 ± 130秒	18.1%

离散空间模型所需的存储空间约为数太字节。不过，训练时间缩短了 5 倍以上，这意味着通常需要半周时间训练的连续模型，若采用离散模型，只需约 16 小时即可完成训练。

我们发现，经过多次试验后，离散模型仍无法收敛到优于基础 S-UNIWARD 实现的验证错误率。虽然连续模型使 YedroudjNet 模型的错误率提高了 14.4 个百分点，但离散模型却使 YedroudjNet 模型的错误率降低了 4.1 个百分点，这与我们的目标相反。为了解决这一问题，我们尝试通过移除部分卷积层和全连接层来降低模型的复杂度。我们尝试改变离散空间的大小和分布，从最初的 $7 \times 7 \times 7$ 调整到最终的 $13 \times 13 \times 13$。我们决定不再继续尝试大小为 $17 \times 17 \times 17$ 或 $26 \times 26 \times 26$ 的分布，因为我们估计，对这些离散空间进行预处理分别需要约 6.3 TB 和 22.4 TB 的存储空间，这被认为是不可行的。我们还尝试修改输出分布的形状，分别测试了均匀分布和我们最终采用的非均匀分布。

我们认为，离散空间模型失效可能是由以下一个或多个因素造成的。首先，离散空间的稀疏性可能大幅减少那些具有有效的 σ、ϵ 和湿代价组合、能够欺骗隐写分析器的图像数量；由于可选方案不足，最佳错误率因此受到限制。其次，离散空间的稀疏性可能使模型难以进行精细调整，尤其是对于隐写术这类对细微变化十分敏感的任务。最后，我们选择的离散空间可能未能充分表示最优输出分布，导致模型无法可靠地欺骗隐写分析器。我们不认为模型过于复杂而无法收敛到最优解，因为我们测试了从 4 层到 7 层的模型，结果都相似。我们也不认为问题对于模型而言过于复杂，因为离散情形和连续情形使用的模型架构几乎相同，只对离散模型的输出层进行了修改。离散问题的难度理应与连续问题相当，甚至更低，因为可能输出值的范围从无限缩小到了 $13 \times 13 \times 13$。

4 结论与未来工作

我们的结果表明，可以利用神经网络修改既有方法的参数，从而改进既有隐写算法。具体而言，我们以 S-UNIWARD 为例，提高了它在多个数据集上对抗隐写分析器时的错误率。尤其是，由于这些结果建立在一种成熟方法之上，因此更易于解读。未来仍有多个可探索的研究方向，包括：

- 同时针对多个隐写分析器训练 SA-CNN。
- 将当前模型扩展至适用于其他隐写方法，并采用可泛化至不同隐写方法的输出。
- 研究采用 SA-CNN 输出的 S-UNIWARD 参数在不同嵌入率下的影响。
- 使用生成对抗网络的对抗训练方法，让 SA-CNN 与 Yedroudj-Net 相互训练。

参考文献

[1] V. Holub、J. Fridrich 和 T. Denemark，“适用于任意域隐写术的通用失真函数”，*EURASIP* 信息安全期刊，第 1 卷，2014 年 12 月。[2] M. Yedroudj、F. Comby 和 M. Chaumont，“Yedroudj-net：一种用于空间域隐写分析的高效 CNN”，载于*2018 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2018 年，第 2092–2096 页。[3] T. H. Thai、R. Cogranne 和 F. Retraint，“基于精确 DCT 系数模型的 OutGuess 最优检测”，载于*2014 IEEE* 信息取证与安全国际研讨会（*WIFS*），2014 年，第 179–184 页。[4] Z. Oplatková、J. Holoska、I. Zelinka 和 R. Senkerik，“利用神经网络检测由 OutGuess 和 Steghide 插入的隐写信息”，载于*2009*年第三届亚洲建模与仿真国际会议，2009 年，第 7–12 页。[5] S. Tan 和 B. Li，“用于数字图像隐写分析的堆叠卷积自动编码器”，载于信号与信息处理协会年度峰会与会议（*APSIPA*），*2014* 年亚太会议，2014 年，第 1–4 页。[6] Y. Qian、J. Dong、W. Wang 和 T. Tan，“通过卷积神经网络进行隐写分析的深度学习”，*SPIE* 会议论文集——国际光学工程学会，第 9409 卷，2015 年 3 月。[7] J. Ye、J. Ni 和 Y. Yi，“用于图像隐写分析的深度学习层次化表示”，*IEEE* 信息取证与安全汇刊，第 12 卷，第 11 期，第 2545–2557 页，2017 年。[8] G. Xu、H.-Z. Wu 和 Y.-Q. Shi，“面向隐写分析的卷积神经网络结构设计”，*IEEE* 信号处理快报，第 23 卷，第 5 期，第 708–712 页，2016 年。[9] M. Sharifzadeh、M. Aloraini 和 D. Schonfeld，“自适应批大小图像合并隐写术与量化高斯图像隐写术”，*IEEE* 信息取证与安全汇刊，第 15 卷，第 867–879 页，2020 年。[10] S. Binghamton。（2021 年）数字数据嵌入实验室。 [在线]。网址：http://dde.binghamton.edu/download/ [11] O. Russakovsky、J. Deng、H. Su、J. Krause、S. Satheesh、S. Ma、Z. Huang、A. Karpathy、A. Khosla、M. Bernstein等，“ImageNet 大规模视觉识别挑战赛”，国际计算机视觉杂志，第 115 卷，第 211–252 页，2015 年。