

基于外源特征进行模型所有权验证的方法和装置

5 本申请要求于2021年11月25日提交中国国家知识产权局、申请号为2021114172450、
申请名称为“基于外源特征进行模型所有权验证的方法和装置”的中国专利申请的优先权，
其全部内容通过引用结合在本申请中。

技术领域

10 本说明书实施例涉及人工智能领域，尤其涉及一种基于外源特征进行模型所有权验证
的方法和装置。

背景技术

随着计算机软件和人工智能的不断发展，机器学习模型的应用也越来越广泛。训练
一个性能良好的模型需要收集大量的训练样本和消耗大量的计算资源，因此，机器学习模
15 型是一种重要的资产。为了保护模型不被窃取，模型的拥有者通常对所拥有的模型进行黑
盒保护，即，只为使用者提供模型的使用权限，使用者无法知晓模型的结构、内部参数等，
举例来说，模型的拥有者可以通过提供模型调用接口，允许使用者将数据输入模型，并获
得模型的反馈结果，而对使用者而言，模型调用接口就是黑盒。然而，最近的研究表明，
攻击者即使在只能查询模型反馈结果的情况下，也能进行模型窃取，获得一个和部署模型
20 功能相近的替代模型，给模型拥有者的资产造成巨大威胁。因此，如何保护模型具有重
要的现实意义和价值。

发明内容

25 本说明书的实施例描述了一种基于外源特征进行模型所有权验证的方法和装置，本方
法从所有权验证的角度提出了对模型的保护，首先训练用于识别外源特征的特征知识的元
分类器，而后将可疑模型的相关数据输入元分类器，基于元分类器的输出结果，确定可疑
模型是否为从具有外源特征的特征知识的部署模型窃取的模型，从而实现了基于外源特征
的所有权验证，通过验证可疑模型是否为从部署模型窃取的模型，可以实现对部署模型的
保护。

30 根据第一方面，提供了一种基于外源特征进行模型所有权验证的方法，包括：从初
始样本集中选取部分初始样本构成选中样本集；对上述选中样本集中各选中样本的样本数

据进行处理,得到具有外源特征的转化样本构成的转化样本集,其中,上述外源特征为初始样本的样本数据不具备的特征;基于目标模型、辅助模型和上述转化样本集,训练元分类器,其中,上述辅助模型为使用上述初始样本集训练得到的模型,上述目标模型为使用上述转化样本集和上述初始样本集中除上述选中样本集之外的剩余样本集训练得到的模型,上述元分类器用于识别上述外源特征的特征知识;将可疑模型的相关数据输入上述元分类器,基于上述元分类器的输出结果,确定上述可疑模型是否为从部署模型窃取的模型,其中,上述部署模型具有上述外源特征的特征知识。

在一个实施例中,在上述基于目标模型、辅助模型和上述转化样本集,训练元分类器之前,上述方法还包括:响应于上述可疑模型的模型结构已知,且与上述部署模型的模型结构相同,将上述部署模型确定为上述目标模型,以及基于上述可疑模型的模型结构训练上述辅助模型;响应于上述可疑模型的模型结构已知,且与上述部署模型的模型结构不同,基于上述可疑模型的模型结构训练上述目标模型和上述辅助模型。

在一个实施例中,上述基于目标模型、辅助模型和上述转化样本集,训练元分类器,包括:构造包含正负样本的第一元分类器样本集,其中,正样本的样本数据为上述目标模型针对转化样本的梯度信息;负样本的样本数据为上述辅助模型针对转化样本的梯度信息;使用上述第一元分类器样本集,训练得到第一元分类器。

在一个实施例中,上述梯度信息为,梯度向量中各元素经符号函数计算后的结果向量。

在一个实施例中,上述将可疑模型的相关数据输入上述元分类器,基于上述元分类器的输出结果,确定上述可疑模型是否为从部署模型窃取的模型,包括:从上述转化样本集中选取第一转化样本;确定上述可疑模型针对上述第一转化样本的第一梯度信息;将上述第一梯度信息输入上述第一元分类器,得到第一预测结果;响应于上述第一预测结果指示出正样本,确定上述可疑模型为从上述部署模型窃取的模型。

在一个实施例中,上述将可疑模型的相关数据输入上述元分类器,基于上述元分类器的输出结果,确定上述可疑模型是否为从部署模型窃取的模型,包括:基于从上述转化样本集选取的第一子集、上述第一元分类器和上述辅助模型,使用假设检验对上述可疑模型进行所有权验证。

在一个实施例中,上述使用假设检验对上述可疑模型进行所有权验证,包括:构建第一概率小于等于第二概率的第一原假设,其中,第一概率表示上述第一元分类器针对上述可疑模型的梯度信息的预测结果为正样本的后验概率,第二概率表示上述第一元分类器

针对上述辅助模型的梯度信息的预测结果为正样本的后验概率；基于上述第一原假设和上述第一子集中的样本数据，计算 P 值；响应于确定上述 P 值小于显著性水平 α ，确定上述第一原假设被拒绝；响应于确定上述第一原假设被拒绝，确定上述可疑模型为从上述部署模型窃取的模型。

- 5 在一个实施例中，在上述基于目标模型、辅助模型和上述转化样本集，训练元分类器之前，上述方法还包括：响应于上述可疑模型的模型结构未知，将上述部署模型确定为上述目标模型，以及基于上述部署模型的模型结构训练上述辅助模型。

10 在一个实施例中，上述基于目标模型、辅助模型和上述转化样本集，训练元分类器，包括：构造包含正负样本的第二元分类器样本集，其中，正样本的样本数据为，上述目标模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息；负样本的样本数据为，上述辅助模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息；使用上述第二元分类器样本集，训练第二元分类器。

15 在一个实施例中，上述将可疑模型的相关数据输入上述元分类器，基于上述元分类器的输出结果，确定上述可疑模型是否为从部署模型窃取的模型，包括：分别从上述转化样本集和选中样本集中获取对应的第二转化样本和第二选中样本；确定上述可疑模型针对上述第二选中样本的预测输出与针对上述第二转化样本的预测输出的第二差异信息；将上述第二差异信息输入上述第二元分类器，得到第二预测结果；响应于上述第二预测结果指示出正样本，确定上述可疑模型为从上述部署模型偷取的模型。

20 在一个实施例中，上述将可疑模型的相关数据输入上述元分类器，基于上述元分类器的输出结果，确定上述可疑模型是否为从部署模型窃取的模型，包括：基于从上述转化样本集选取的第二子集、上述选中样本集中与上述第二子集对应的第三子集、上述第二元分类器和辅助模型，使用假设检验对上述可疑模型进行所有权验证。

25 在一个实施例中，上述使用假设检验对上述可疑模型进行所有权验证，包括：构建第三概率小于等于第四概率的第二原假设，其中，第三概率表示，上述第二元分类器针对上述可疑模型对应的差异信息的预测结果为正样本的后验概率，第四概率表示上述第二元分类器针对上述辅助模型对应的差异信息的预测结果为正样本的后验概率；基于上述第二原假设、上述第二子集的样本数据和第三子集的样本数据，计算 P 值；响应于确定 P 值小于显著性水平 α ，确定上述第二原假设被拒绝；响应于确定上述第二原假设被拒绝，确定上述可疑模型为从上述部署模型窃取的模型。

30 在一个实施例中，上述初始样本集中初始样本的样本数据为样本图像；以及上述对

上述选中样本集中各样本的样本数据进行处理，得到具有外源特征的转化样本构成的转化样本集，包括：使用图像风格转换器，对上述选中样本集中各样本的样本图像进行风格转换，使样本图像具有指定图像风格，其中，上述外源特征为上述指定图像风格相关的特征。

根据第二方面，提供了一种基于外源特征进行模型所有权验证的装置，包括：选取
5 单元，配置为从初始样本集中选取部分初始样本构成选中样本集；转化单元，配置为对上述选中样本集中各选中样本的样本数据进行处理，得到具有外源特征的转化样本构成的转化样本集，其中，上述外源特征为初始样本的样本数据不具备的特征；训练单元，配置为基于目标模型、辅助模型和上述转化样本集，训练元分类器，其中，上述辅助模型为使用上述初始样本集训练得到的模型，上述目标模型为使用上述转化样本集和上述初始样本集
10 中除上述选中样本集之外的剩余样本集训练得到的模型，上述元分类器用于识别上述外源特征的特征知识；验证单元，配置为将可疑模型的相关数据输入上述元分类器，基于上述元分类器的输出结果，确定上述可疑模型是否为从部署模型窃取的模型，其中，上述部署模型具有上述外源特征的特征知识。

根据第三方面，提供了一种计算机可读存储介质，其上存储有计算机程序，当上
15 述计算机程序在计算机中执行时，令计算机执行如第一方面中任一实现方式描述的方法。

根据第四方面，提供了一种计算设备，包括存储器和处理器，其特征在于，上述存储器中存储有可执行代码，上述处理器执行上述可执行代码时，实现如第一方面中任一实现方式描述的方法。

20 根据本说明书实施例提供的基于外源特征进行模型所有权验证的方法和装置，首先将初始样本集中的部分初始样本嵌入外源特征，得到转化样本集。而后，基于目标模型、辅助模型和转化样本集，训练用于识别外源特征的特征知识的元分类器。然后，将可疑模型的相关数据输入元分类器，并基于元分类器的输出结果确定可疑模型是否为从具有外源特征的特征知识的部署模型窃取的模型。由此，实现了基于外源特征对可疑模型进行所有权
25 验证，通过验证可疑模型是否为从部署模型窃取的模型，可以确定是否有攻击者窃取了部署模型，从而实现了部署模型的保护。

附图说明

图 1 示出了本说明书实施例可以应用于其中的应用场景的示意图；

30 图 2 示出了根据一个实施例的基于外源特征进行模型所有权验证的方法的流程示意

图；

图 3 示出了一种根据可疑模型确定目标模型和辅助模型的流程示意图；

图 4 示出了根据一个实施例的基于外源特征进行模型所有权验证的装置的示意性框图。

5 具体实施方式

下面结合附图和实施例，对本说明书提供的技术方案做进一步的详细描述。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。需要说明的是，在不冲突的情况下，本说明书的实施例及实施例中的特征可以相互组合。

- 10 如前所述，攻击者可以通过各种方式在未授权的情况下逆向出一个和部署模型有类似功能的替代模型，实现对部署模型的侵权。现阶段，存在多种对模型进行窃取攻击的方法。例如，在训练数据集可访问的场景下，攻击者可以通过知识蒸馏或者从零开始训练模型等方式获得替代模型。又例如，在模型可访问的场景下，攻击者可以通过零样本知识蒸馏或者使用本地训练样本微调部署模型等方式获得替代模型。再例如，在仅能查询模型的场景
- 15 下，攻击者也可以根据查询模型返回的结果获得替代模型。

- 为了实现模型保护，在一种方案中，模型所有者通过引入扰动/随机性等方式提升部署模型窃取的难度。然而这种方式一般对部署模型的正常精度会有较大的影响，并且可能会被后续的一些适应性攻击完全绕开。在另一种方案中，使用训练数据集的内在特征进行所有权认证，然而这种方式很容易出现误判，尤其是当可疑模型和部署模型的训练集的潜在
- 20 分布存在较大相似度时，即使可疑模型不是从部署模型窃取而来，该方式也会判断为窃取，因此，该方式的准确性较差。在又一种方案中，可以首先使用后门攻击 (backdoor attack) 对部署模型添加水印，然后基于特定的后门进行所有权认证。然而模型后门是一个较为精细的结构，其在偷盗过程中很可能被损毁，导致该防御方法失效。

- 为此，本说明书的实施例提供了一种基于外源特征进行模型所有权验证的方法，从而
- 25 可以实现对部署模型的保护，其中，部署模型具有外源特征的特征知识。以部署模型为图像分类模型，可疑模型的模型结构已知，且与部署模型的模型结构相同，外源特征为指定样式（例如，油画风格）为例，图 1 示出了本说明书实施例可以应用于其中的应用场景的示意图。如图 1 所示，首先，从初始样本集中选取部分初始样本构成选中样本集 101，在本例中，初始样本包括初始样本图像和对应的标签。然后，对选中样本集 101 中各选中样
- 30 本的样本数据进行处理，得到具有外源特征的转化样本构成的转化样本集 102。本例中，

具体指使用一个已训练的样式转换器 103, 基于一张指定风格图像 104 (例如, 油画风格) 对选中样本集 101 中的初始样本图像进行样式转化, 将选中样本集 101 中的初始样本图像转化指定风格的图像。如此, 转换样本集 102 中的各个转化样本, 也具有了该指定风格, 例如, 油画风格。本例中, 可以将部署模型确定为目标模型 106, 并基于可疑模型的模型结构训练辅助模型 107, 其中, 目标模型 106 为使用转化样本集 102 和初始样本集中除选中样本集 101 之外的剩余样本集 105 训练得到的模型, 辅助模型 107 为使用初始样本集训练得到的模型。可以理解, 由于训练过程使用了转化样本集 102, 而转化样本具有例如油画风格的外源特征, 如此训练的目标模型 106 相应具有上述外源特征的特征知识, 即处理上述外源特征的能力。而辅助模型 107 基于初始样本集训练, 因此不具有上述外源特征的特征知识。基于该核心区别, 在本说明书的技术构思中, 基于目标模型 106、辅助模型 107 和转化样本集 102, 训练用于识别外源特征的特征知识的元分类器 108。最后, 将可疑模型的相关数据输入元分类器 108, 基于元分类器 108 的输出结果, 确定可疑模型是否为从部署模型窃取的模型。由此, 实现了基于外源特征对可疑模型进行所有权验证, 通过验证可疑模型是否为从部署模型窃取的模型, 可以确定是否有攻击者窃取了部署模型, 从而实现了部署模型的保护。

继续参见图 2, 图 2 示出了根据一个实施例的基于外源特征进行模型所有权验证的方法的流程示意图。可以理解, 该方法可以通过任何具有计算、处理能力的装置、设备、平台、设备集群来执行。如图 2 所示, 该基于外源特征进行模型所有权验证的方法, 可以包括以下步骤:

步骤 201, 从初始样本集中选取部分初始样本构成选中样本集。

在本实施例中, 执行基于外源特征进行模型所有权验证的方法的执行主体, 可以从初始样本集中选取部分初始样本构成选中样本集。例如, 可以预先设定选取样本的数量, 并根据该数量随机从初始样本集中选取初始样本构成选中样本集。又例如, 可以预先设定比例 $\gamma\%$, 并根据比例 $\gamma\%$ 随机从初始样本集中选取初始样本构成选中样本集。这里, 初始样本集中的初始样本可以包括样本数据和标签。

步骤 202, 对选中样本集中各选中样本的样本数据进行处理, 得到具有外源特征的转化样本构成的转化样本集。

在本实施例中, 可以对步骤 201 中得到的选中样本集中的各个选中样本的样本数据进行处理, 得到具有外源特征的转化样本构成的转化样本集。这里, 外源特征可以为初始样本集中的初始样本的样本数据不具备的特征。对于样本集的内在特征和外源特征, 简单来

说，一个样本如果来源于这个数据集，那么它一定具有的特征被定义为内在特征；如果一个样本有外源特征，那么它一定不来源于这个样本集。具体的，一个特征 f 被称为数据集 D 中的内在特征，当且仅当从数据集 D 中任取样本数据，其中均包含该特征 f 。同样的，任取样本数据 $(\mathbf{x}, \mathbf{y})$ ，如果该样本数据包含特征 f ，就可得出该样本数据不属于数据集 D ，那么可以将该特征 f 称为数据集 D 的外源特征。

这里，基于模型所能够实现的功能，初始样本集中的初始样本的样本数据可以是各种数据。例如，当模型实现的功能为文本分类时，初始样本的样本数据可以是文本信息。此时，外源特征可以是同语言的预设词、句子等，也可以是另一种语言的预设的词、句子等，此时，可以通过在文本信息中插入外源特征来得到具有外源特征的转化样本。又例如，当模型实现的功能与语音相关（例如，语音识别）时，初始样本的样本数据可以是语音信息，此时，外源特征可以是特定的噪音等非自然声音，此时，可以通过在语音信息中插入外源特征来得到具有外源特征的转化样本。

在一些可选的实现方式中，本实施例的模型可以是图像分类模型，初始样本集中的初始样本的样本数据可以是样本图像，以及上述步骤 202 可以具体如下实现：使用图像风格转换器，对选中样本集中各样本的样本图像进行风格转换，使样本图像具有指定图像风格，其中，外源特征为指定图像风格相关的特征。

在本实现方式中，图像风格转换器可以是预先训练的机器学习模型，用于将图像转化为指定图像风格。作为示例，指定图像风格可以是各种风格，例如，油画风格，水墨画风格，滤镜效果，马赛克展示，等等。

举例来说，对于预先设定的一张指定风格图像 $\mathbf{x}_s$ ，图像风格转换器 T 可以将选中样本集 $\mathcal{D}_s$ 中的各选中样本进行风格转换，使得选中样本中的样本图像具有与指定风格图像 $\mathbf{x}_s$ 相同的图像风格，得到转化样本集。即， $\mathcal{D}_t = \{(\mathbf{x}', \mathbf{y}) | \mathbf{x}' = T(\mathbf{x}, \mathbf{x}_s), (\mathbf{x}, \mathbf{y}) \in \mathcal{D}_s\}$ ，其中， $\mathcal{D}_t$ 可以表示转化样本集； $\mathbf{x}, \mathbf{y}$ 分别表示选中样本的样本数据和标签； $\mathbf{x}'$ 表示选中样本中的样本图像经图像风格转换器 T 进行风格转换后的、与指定风格图像 $\mathbf{x}_s$ 的图像风格相同的图像。可以理解，在本实现方式中，仅对选中样本的样本图像的风格进行转换，而不对样本图像的内容进行改变。例如，如图 1 所显示的，样本图像中原来显示一条狗，在进行风格转换后依然显示一条狗，因此无需对选中样本的标签进行改变。

需要理解，在本说明书的实施例中，要求被保护的部署模型所使用的训练数据集包含上述转化样本集，从而在部署模型中引入外源特征的特征知识。另外，需要理解，通过上述实现方式嵌入的外源特征没有明确的特征表达，也不会对基于转化样本集训练得到的部

署模型的预测造成很大的影响。可以理解，在部署模型的训练中转化样本集的转化样本仅占总样本的一小部分。举例来说，可以通过以下公式 $\min_{\theta} \sum_{(\mathbf{x}, y) \in \mathcal{D}_b \cup \mathcal{D}_t} \mathcal{L}(V_{\theta}(\mathbf{x}), y)$ 训练得到部署模型，其中， V_{θ} 可以表示部署模型， $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ 可以表示初始样本集，其中， N 可以表示样本数量，样本集 $\mathcal{D}_b \triangleq \mathcal{D} \setminus \mathcal{D}_s$ 可以表示初始样本集 $\mathcal{D}$ 中除选中样本集 $\mathcal{D}_s$ 之外的
 5 剩余样本集。 $\mathcal{L}(\cdot)$ 可以表示损失函数（例如，交叉熵）。由此，可以使部署模型具有外源特征的特征知识。

步骤 203，基于目标模型、辅助模型和转化样本集，训练元分类器。

在本实施例中，可以基于目标模型、辅助模型和转化样本集，训练元分类器。其中，辅助模型可以为使用初始样本集 $\mathcal{D}$ 训练得到的模型，目标模型可以为使用转化样本集 $\mathcal{D}_t$
 10 和初始样本集中除选中样本集之外的剩余样本集 $\mathcal{D}_b$ 训练得到的模型。元分类器可以用于识别外源特征的特征知识。实践中，元分类器可以是二分类器。

步骤 204，将可疑模型的相关数据输入元分类器，基于元分类器的输出结果，确定可疑模型是否为从部署模型窃取的模式。

在本实施例中，可以将可疑模型的相关数据输入步骤 203 训练得到的元分类器，并基
 15 于元分类器的输出结果，确定可疑模型是否为从部署模型窃取的模式。这里，部署模型可以具有外源特征的特征知识。如前所述，可以使用嵌入外源特征的转化样本和未嵌入外源特征的初始样本共同来训练模型，得到部署模型，由此，部署模型可以学习到外源特征的特征知识。可以理解，部署模型可以是模型所有者部署到线上，供用户使用的模型。如前面所描述的，外源特征不会对部署模型的预测造成很大的影响，因此，部署模型不会影响
 20 用户的正常使用。同时，由于部署模型具有外源特征的特征知识，因此，如果攻击者通过窃取获得一个与部署模型的功能相近的替代模型，则该替代模型也将具有外源特征的特征知识。基于此，如果怀疑某个模型为从部署模型窃取的替代模型，可以将该模型作为可疑模型进行所有权验证。举例来说，如果该模型也具有外源特征的特征知识，则可以确定该模型为从部署模型窃取的模式。

25 实践中，不同结构的机器学习模型也可以实现相同的功能，因此，攻击者通过窃取部署模型得到的替代模型的模型结构可以与部署模型的模型结构相同，也可以不同。即，可疑模型的模型结构可以与部署模型的模型结构相同，也可以不同。

在一些可选的实现方式中，在基于目标模型、辅助模型和转化样本集，训练元分类器之前，上述基于外源特征进行模型所有权验证的方法，还可以包括确定目标模型和辅助模
 30 型的过程。举例来说，可以根据可疑模型的模型结构是否已知，以及是否与部署模型的模

型结构相同，分为多种场景。如图 3 所示，图 3 示出了一种根据可疑模型确定目标模型和辅助模型的流程示意图。可以包括以下步骤：

步骤 301，确定可疑模型的模型结构是否已知。

步骤 302，响应于确定可疑模型的模型结构已知，进一步确定可疑模型的模型结构是否 5 与部署模型的模型结构相同。

步骤 303，响应于确定可疑模型的模型结构已知，且与部署模型的模型结构相同，将部署模型确定为目标模型，并基于可疑模型的模型结构训练辅助模型。

在本实现方式中，在可疑模型与部署模型的模型结构相同的情况下，可以将部署模型作为前述目标模型，由此，可以节省目标模型的训练时间。此外，可以根据初始样本集中的 10 的初始样本，训练得到与目标模型（部署模型）和可疑模型的模型结构均相同的辅助模型。由于初始样本集中的初始样本未嵌入外源特征，所以也可以将初始样本集称为良性样本集，而辅助模型是根据未嵌入外源特征的初始样本训练得到的，由此也可以将辅助模型称为良性模型或正常模型。辅助模型不具有外源特征的特征知识。

步骤 304，响应于确定可疑模型的模型结构已知，且与部署模型的模型结构不同，基 15 于可疑模型的模型结构训练目标模型和辅助模型。

在本实现方式中，在可疑模型与部署模型的模型结构不相同的情况下，可以根据转化样本集和初始样本集中除选中样本集之外的剩余样本集，以及可疑模型的模型结构训练得到目标模型。在目标模型的训练过程中，目标模型可以学习到外源特征的特征知识，并具有与可疑模型相同的模型结构。此外，还可以根据初始样本集训练得到与可疑模型结构相 20 同的辅助模型。

由步骤 303 和步骤 304 可知，在可疑模型的模型结构已知的情况下，目标模型和辅助模型的模型结构与可疑模型的模型结构相同。

步骤 305，响应于确定可疑模型的模型结构未知，将部署模型确定为目标模型，以及基于部署模型的模型结构训练辅助模型。

25 在本实现方式中，在可疑模型的模型结构未知的情况下，可以将部署模型确定为目标模型，并可以根据初始样本集和部署模型的模型结构训练得到辅助模型。也就是说，在可疑模型的模型结构未知的情况下，目标模型和辅助模型的模型结构与部署模型的模型结构相同。

30 在一些可选的实现方式中，在可疑模型的模型结构已知的情况下，上述步骤 203，基于目标模型、辅助模型和转化样本集，训练元分类器，可以具体如下进行：

首先,构造包含正负样本的第一元分类器样本集。

在本实现方式中,为了训练第一元分类器,首先需要构造包含正负样本的第一元分类器样本集。这里,正样本的样本数据可以为目标模型针对转化样本的梯度信息。负样本的样本数据可以为辅助模型针对转化样本的梯度信息。例如,可以将梯度向量作为梯度信息。

5 可选的,梯度信息还可以为梯度向量中各元素经符号函数计算后的结果向量。梯度向量经符号函数计算后的结果向量,更简单且仍能体现梯度的方向特点,因此,可以作为梯度信息。

然后,使用第一元分类器样本集,训练得到一个二分类的分类器作为第一元分类器。

10 在本实现方式中,可以使用第一元分类器样本集,训练得到第一元分类器。以第一元分类器样本集中正样本的标签为+1,负样本的标签为-1,梯度信息为梯度向量中各元素经符号函数计算后的结果向量为例,第一元分类器样本集 $\mathcal{D}_c$ 可以表示为 $\mathcal{D}_c = \mathcal{D}_{positive} \cup \mathcal{D}_{negative}$, 其中,正样本 $\mathcal{D}_{positive} = \{(g_V(\mathbf{x}'), +1) | (\mathbf{x}', y) \in \mathcal{D}_t\}$, 该正样本中标签为+1; $\mathcal{D}_t$ 可以表示转化样本集, $\mathbf{x}'$ 表示转化样本。这里, $g_V(\mathbf{x}') = \text{sign}(\nabla_{\theta} \mathcal{L}(V(\mathbf{x}'), y))$, 其中, V 可以表示目标模型, $g_V(\mathbf{x}')$ 表示目标模型针对转化样本的梯度信息, $\nabla_{\theta} \mathcal{L}(V(\mathbf{x}'), y)$ 表示目标模型针对转化样本的损失函数梯度向量, $\text{sign}(\cdot)$ 表示 sign 函数, sign 函数为符号函数。负样本 $\mathcal{D}_{negative} = \{(g_B(\mathbf{x}'), -1) | (\mathbf{x}', y) \in \mathcal{D}_t\}$, 该负样本中标签为-1, 这里, $g_B(\mathbf{x}') = \text{sign}(\nabla_{\theta} \mathcal{L}(B(\mathbf{x}'), y))$, 其中, B 表示辅助模型, $g_B(\mathbf{x}')$ 表示辅助模型针对转化样本的梯度信息, $\nabla_{\theta} \mathcal{L}(B(\mathbf{x}'), y)$ 表示辅助模型针对转化样本的损失函数梯度向量。本例中,第一元分类器 C 可以通过以下公式 $\min_{\mathbf{w}} \sum_{(\mathbf{x}, y) \in \mathcal{D}_c} \mathcal{L}(C_{\mathbf{w}}(\mathbf{x}), y)$ 训练, 其中, $\mathbf{w}$ 可以表示分类器中的模型参数。

20 在一些可选的实现方式中,在可疑模型的模型结构已知的情况下,上述步骤 204,将可疑模型的相关数据输入元分类器,基于元分类器的输出结果,确定可疑模型是否为从部署模型窃取的模型,可以具体包括以下步骤 1)~4):

步骤 1), 从转化样本集中选取转化样本作为第一转化样本。

25 步骤 2), 确定可疑模型针对第一转化样本的第一梯度信息。

步骤 3), 将第一梯度信息输入第一元分类器,得到第一预测结果。

步骤 4), 响应于确定第一预测结果指示出正样本,确定可疑模型为从部署模型窃取的模型。

30 举例来说,还是以第一元分类器样本集中正样本的标签为+1,负样本的标签为-1,梯度信息为梯度向量中各元素经符号函数计算后的结果向量为例,假设可疑模型为 S , 第一

元分类器为 C ，第一转化样本为标签为 y 的转化图像 x' ，可以通过 $g_S(x') = \text{sign}(\nabla_{\theta} \mathcal{L}(S(x'), y))$ 确定可疑模型针对第一转化样本的第一梯度信息。之后，将第一梯度信息输入第一元分类器 C ，即， $C(g_S(x'))$ ，得到第一预测结果。如果第一预测结果指示出正样本，即 $C(g_S(x')) = 1$ ，则可以确定可疑模型为从部署模型窃取的模型。

- 5 本例中， $C(g_S(x')) = 1$ 可以表示可疑模型与部署模型一样，都具有外源特征的特征知识，由此，可以确定可疑模型为从部署模型窃取的模型。通过本实现方式，可以实现对可疑模型的所有权验证。

- 在另一些可选的实现方式中，在可疑模型的模型结构已知的情况下，上述步骤 204，将可疑模型的相关数据输入元分类器，基于元分类器的输出结果，确定可疑模型是否为从部署模型窃取的模型，还可以具体包括：基于从转化样本集选取的第一子集、第一元分类器和辅助模型，使用假设检验对可疑模型进行所有权验证。
- 10 部署模型窃取的模型，还可以具体包括：基于从转化样本集选取的第一子集、第一元分类器和辅助模型，使用假设检验对可疑模型进行所有权验证。

在本实现方式中，首先可以从转化样本集 $\mathcal{D}_t$ 中选取（例如，随机抽取）多个转化样本构成第一子集，然后根据第一子集、第一元分类器和辅助模型，使用多种假设检验对可疑模型进行所有验证。例如，可以使用 Z 检验对可疑模型进行所有权验证。

- 15 可选的，上述使用假设检验对可疑模型进行所有权验证，可以包括：使用单边配对样本 T 检验对可疑模型进行所有权验证，具体可以包括以下内容：

首先，构建第一概率小于等于第二概率的第一原假设。

- 在本实现方式中，针对第一子集，第一概率 μ_S 可以表示第一元分类器针对可疑模型的梯度信息的预测结果为正样本的后验概率，第二概率 μ_B 可以表示第一元分类器针对辅助模型的梯度信息的预测结果为正样本的后验概率。举例来说，以 $\mathbf{X}'$ 表示第一子集中转化样本的样本数据，正样本的标签为 +1 为例，第一概率 μ_S 和第二概率 μ_B 分别表示事件 $C(g_S(\mathbf{X}')) = 1$ 和 $C(g_B(\mathbf{X}')) = 1$ 的后验概率，可以对此构建原假设 $H_0: \mu_S \leq \mu_B$ ，其中， S 表示可疑模型， B 表示辅助模型。
- 20 模型的梯度信息的预测结果为正样本的后验概率。举例来说，以 $\mathbf{X}'$ 表示第一子集中转化样本的样本数据，正样本的标签为 +1 为例，第一概率 μ_S 和第二概率 μ_B 分别表示事件 $C(g_S(\mathbf{X}')) = 1$ 和 $C(g_B(\mathbf{X}')) = 1$ 的后验概率，可以对此构建原假设 $H_0: \mu_S \leq \mu_B$ ，其中， S 表示可疑模型， B 表示辅助模型。

- 其次，基于上述第一原假设和第一子集中的样本数据，计算 P 值。可以理解，在单边配对样本 T 检验中， P 值的计算是本领域技术人员所公知的，此处不再赘述。
- 25 配对样本 T 检验中， P 值的计算是本领域技术人员所公知的，此处不再赘述。

然后，响应于确定 P 值小于显著性水平 α ，确定第一原假设被拒绝。这里，显著性水平 α 可以是技术人员根据实际需要确定的值。

- 最后，响应于确定第一原假设被拒绝，确定可疑模型为从部署模型窃取的模型。实践中，由于辅助模型不具有外源特征的特征知识，因此， μ_B 应该为一个较小的值，而如
- 30 果 μ_S 小于等于 μ_B 成立，则可以表示可疑模型也不具有外源特征的特征知识，即，可疑模

型不是从部署模型窃取的模型。反之，如果 μ_S 小于等于 μ_B 不成立（即，被拒绝），则可以表示可疑模型具有外源特征的特征知识，即可疑模型是从部署模型窃取的模型。本实现方式，通过统计学中的假设检验对可疑模型进行所有权验证，可以避免所有权验证过程中转化样本选择的随机性对所有权验证的准确性的影响，从而使验证更加准确。

- 5 如前述图 3 所示，在一些可选的实现方式中，可疑模型的模型结构未知，因此难以获得模型的梯度信息，构建元分类器的训练样本。在这样的情况下，上述步骤 203，基于目标模型、辅助模型和转化样本集，训练元分类器，可以具体如下进行：

首先，构造包含正负样本的第二元分类器样本集。

- 10 在本实现方式中，为了训练第二元分类器，首先需要构造包含正负样本的第二元分类器样本集。这里，正样本的样本数据为，目标模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息。负样本的样本数据为，辅助模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息。实践中，如果目标模型和辅助模型为分类模型，那么目标模型和辅助模型的预测输出可以是分别针对多个类别标签的多个预测概率形成的概率向量。作为一个示例，差异信息可以指差值向量。
- 15 作为另一示例，差异信息还可以为差值向量经符号函数计算后的结果，比如，正样本的样本数据为 $\text{sign}(V(\mathbf{x}) - V(\mathbf{x}'))$ ，其中， $V(\mathbf{x})$ 表示目标模型针对选中样本的预测输出（体现为一个概率向量）， $V(\mathbf{x}')$ 表示目标模型针对该选中样本对应的转化样本的预测输出。负样本的样本数据为 $\text{sign}(B(\mathbf{x}) - B(\mathbf{x}'))$ ，其中， $B(\mathbf{x})$ 表示辅助模型针对选中样本的预测输出， $B(\mathbf{x}')$ 表示辅助模型针对该选中样本对应的转化样本的预测输出。

- 20 然后，使用第二元分类器样本集，训练第二元分类器。

在本实现方式中，可以使用第二元分类器样本集，训练第二元分类器。通过本实现方式，可以在可疑模型的模型结构未知的情况下，训练元分类器，以便后续的模型所有权验证。

- 25 在一些可选的实现方式中，在可疑模型的模型结构未知的情况下，上述步骤 204，将可疑模型的相关数据输入元分类器，基于元分类器的输出结果，确定可疑模型是否为从部署模型窃取的模型，可以具体包括以下步骤一～步骤四：

步骤一，分别从转化样本集和选中样本集中获取对应的第二转化样本和第二选中样本。这里，某一个第二转化样本与某一个选中样本相对应，可以是指该第二转化样本是由该选中样本通过嵌入外源特征得到的。

- 30 步骤二，确定可疑模型针对第二选中样本的预测输出与针对第二转化样本的预测输出

的第二差异信息。

步骤三，将第二差异信息输入第二元分类器，得到第二预测结果。

步骤四，判断第二预测结果是否指示出正样本，响应于确定第二预测结果指示出正样本，确定可疑模型为从部署模型偷取的模型。通过本实现方式，可以在可疑模型的模型结构未知的情况下，实现对可疑模型的所有权验证。

在另一些可选的实现方式中，在可疑模型的模型结构未知的情况下，上述步骤 204，将可疑模型的相关数据输入元分类器，基于元分类器的输出结果，确定可疑模型是否为从部署模型窃取的模型，还可以具体包括：基于从转化样本集选取的第二子集、选中样本集中与第二子集对应的第三子集、第二元分类器和辅助模型，使用假设检验对可疑模型进行所有权验证。比如，可以使用 Z 检验对可疑模型进行所有权验证。

可选的，上述使用假设检验对可疑模型进行所有权验证，可以包括：使用单边配对样本 T 检验对可疑模型进行所有权验证，具体可以包括以下内容：

首先，构建第三概率小于等于第四概率的第二原假设。

在本实现方式中，针对第二子集和第三子集，第三概率可以表示第二元分类器针对可疑模型对应的差异信息的预测结果为正样本的后验概率。第四概率可以表示第二元分类器针对辅助模型对应的差异信息的预测结果为正样本的后验概率。

其次，基于第二原假设、第二子集的样本数据和第三子集的样本数据，计算 P 值。可以理解，在单边配对样本 T 检验中，P 值的计算是本领域技术人员公知的，此处不再赘述。

然后，响应于确定 P 值小于显著性水平 α ，确定第二原假设被拒绝。这里，显著性水平 α 可以是技术人员根据实际需要确定的值。

最后，响应于确定第二原假设被拒绝，确定可疑模型为从部署模型窃取的模型。实践中，由于辅助模型不具有外源特征的特征知识，因此，第四概率应该为一个较小的值，而如果第三概率小于等于第四概率成立，则可以表示可疑模型也不具有外源特征的特征知识，即，可疑模型不是从部署模型窃取的模型。反之，如果第三概率小于等于第四概率不成立（即，被拒绝），则可以表示可疑模型具有外源特征的特征知识，即可疑模型是从部署模型窃取的模型。本实现方式，通过统计学中的假设检验对可疑模型进行所有权验证，可以避免所有权验证过程中转化样本选择的随机性对所有权验证的准确性的影响，从而使验证更加准确。

根据另一方面的实施例，提供了一种基于外源特征进行模型所有权验证的装置。上述基于外源特征进行模型所有权验证的装置可以部署在任何具有计算、处理能力的设备、

平台或设备集群中。

图 4 示出了根据一个实施例的基于外源特征进行模型所有权验证的装置的示意性框图。如图 4 所示，该基于外源特征进行模型所有权验证的装置 400 包括：选取单元 401，配置为从初始样本集中选取部分初始样本构成选中样本集；转化单元 402，配置为对上述选中样本集中各选中样本的样本数据进行处理，得到具有外源特征的转化样本构成的转化样本集，其中，上述外源特征为初始样本的样本数据不具备的特征；训练单元 403，配置为基于目标模型、辅助模型和上述转化样本集，训练元分类器，其中，上述辅助模型为使用上述初始样本集训练得到的模型，上述目标模型为使用上述转化样本集和上述初始样本集中除上述选中样本集之外的剩余样本集训练得到的模型，上述元分类器用于识别上述外源特征的特征知识；验证单元 404，配置为将可疑模型的相关数据输入上述元分类器，基于上述元分类器的输出结果，确定上述可疑模型是否为从部署模型窃取的模型，其中，上述部署模型具有上述外源特征的特征知识。

在本实施例的一些可选的实现方式中，上述装置 400 还包括：第一模型训练单元（图中未示出），配置为响应于上述可疑模型的模型结构已知，且与上述部署模型的模型结构相同，将上述部署模型确定为上述目标模型，以及基于上述可疑模型的模型结构训练上述辅助模型；第二模型训练单元（图中未示出），配置为响应于上述可疑模型的模型结构已知，且与上述部署模型的模型结构不同，基于上述可疑模型的模型结构训练上述目标模型和上述辅助模型。

在本实施例的一些可选的实现方式中，上述训练单元 403 进一步配置为：构造包含正负样本的第一元分类器样本集，其中，正样本的样本数据为上述目标模型针对转化样本的梯度信息；负样本的样本数据为上述辅助模型针对转化样本的梯度信息；使用上述第一元分类器样本集，训练得到第一元分类器。

在本实施例的一些可选的实现方式中，上述梯度信息为，梯度向量中各元素经符号函数计算后的结果向量。

在本实施例的一些可选的实现方式中，上述验证单元 404 进一步配置为：从上述转化样本集中选取第一转化样本；确定上述可疑模型针对上述第一转化样本的第一梯度信息；将上述第一梯度信息输入上述第一元分类器，得到第一预测结果；响应于上述第一预测结果指示出正样本，确定上述可疑模型为从上述部署模型窃取的模型。

在本实施例的一些可选的实现方式中，上述验证单元 404 进一步配置为：基于从上述转化样本集选取的第一子集、上述第一元分类器和上述辅助模型，使用假设检验对上述可

疑模型进行所有权验证。

在本实施例的一些可选的实现方式中，上述使用假设检验对上述可疑模型进行所有权验证，包括：构建第一概率小于等于第二概率的第一原假设，其中，第一概率表示上述第一元分类器针对上述可疑模型的梯度信息的预测结果为正样本的后验概率，第二概率表示上述第一元分类器针对上述辅助模型的梯度信息的预测结果为正样本的后验概率；基于上述第一原假设和上述第一子集中的样本数据，计算 P 值；响应于确定上述 P 值小于显著性水平 α ，确定上述第一原假设被拒绝；响应于确定上述第一原假设被拒绝，确定上述可疑模型为从上述部署模型窃取的模式。

在本实施例的一些可选的实现方式中，上述装置 400 还包括：第三模型训练单元（图中未示出），响应于上述可疑模型的模型结构未知，将上述部署模型确定为上述目标模型，以及基于上述部署模型的模型结构训练上述辅助模型。

在本实施例的一些可选的实现方式中，上述训练单元 403 进一步配置为：构造包含正负样本的第二元分类器样本集，其中，正样本的样本数据为，上述目标模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息；负样本的样本数据为，上述辅助模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息；使用上述第二元分类器样本集，训练第二元分类器。

在本实施例的一些可选的实现方式中，上述验证单元 404 进一步配置为：分别从上述转化样本集和选中样本集中获取对应的第二转化样本和第二选中样本；确定上述可疑模型针对上述第二选中样本的预测输出与针对上述第二转化样本的预测输出的第二差异信息；将上述第二差异信息输入上述第二元分类器，得到第二预测结果；响应于上述第二预测结果指示出正样本，确定上述可疑模型为从上述部署模型偷取的模型。

在本实施例的一些可选的实现方式中，上述验证单元 404 进一步配置为：基于从上述转化样本集选取的第二子集、上述选中样本集中与上述第二子集对应的第三子集、上述第二元分类器和辅助模型，使用假设检验对上述可疑模型进行所有权验证。

在本实施例的一些可选的实现方式中，上述使用假设检验对上述可疑模型进行所有权验证，包括：构建第三概率小于等于第四概率的第二原假设，其中，第三概率表示，上述第二元分类器针对上述可疑模型对应的差异信息的预测结果为正样本的后验概率，第四概率表示上述第二元分类器针对上述辅助模型对应的差异信息的预测结果为正样本的后验概率；基于上述第二原假设、上述第二子集的样本数据和第三子集的样本数据，计算 P 值；响应于确定 P 值小于显著性水平 α ，确定上述第二原假设被拒绝；响应于确定上述第二原

假设被拒绝，确定上述可疑模型为从上述部署模型窃取模型。

在本实施例的一些可选的实现方式中，述初始样本集中初始样本的样本数据为样本图像；以及上述转化单元 402 进一步配置为：使用图像风格转换器，对上述选中样本集中各样本的样本图像进行风格转换，使样本图像具有指定图像风格，其中，上述外源特征为上述指定图像风格相关的特征。

根据另一方面的实施例，还提供一种计算机可读存储介质，其上存储有计算机程序，当上述计算机程序在计算机中执行时，令计算机执行图 2 所描述的方法。

根据再一方面的实施例，还提供一种计算设备，包括存储器和处理器，其特征在于，上述存储器中存储有可执行代码，上述处理器执行上述可执行代码时，实现图 2 所描述的方法。

本领域普通技术人员应该还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执轨道，取决于技术方案的特定应用和设计约束条件。本领域普通技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

结合本文中所公开的实施例描述的方法或算法的步骤可以用硬件、处理器执轨道的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程 ROM、电可擦除可编程 ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1. 一种基于外源特征进行模型所有权验证的方法，包括：

从初始样本集中选取部分初始样本构成选中样本集；

对所述选中样本集中各选中样本的样本数据进行处理，得到具有外源特征的转化样本构成的转化样本集，其中，所述外源特征为初始样本的样本数据不具备的特征；

基于目标模型、辅助模型和所述转化样本集，训练元分类器，其中，所述辅助模型为使用所述初始样本集训练得到的模型，所述目标模型为使用所述转化样本集和所述初始样本集中除所述选中样本集之外的剩余样本集训练得到的模型，所述元分类器用于识别所述外源特征的特征知识；

将可疑模型的相关数据输入所述元分类器，基于所述元分类器的输出结果，确定所述可疑模型是否为从部署模型窃取的模型，其中，所述部署模型具有所述外源特征的特征知识。

2. 根据权利要求 1 所述的方法，其中，在所述基于目标模型、辅助模型和所述转化样本集，训练元分类器之前，所述方法还包括：

响应于所述可疑模型的模型结构已知，且与所述部署模型的模型结构相同，将所述部署模型确定为所述目标模型，以及基于所述可疑模型的模型结构训练所述辅助模型；

响应于所述可疑模型的模型结构已知，且与所述部署模型的模型结构不同，基于所述可疑模型的模型结构训练所述目标模型和所述辅助模型。

3. 根据权利要求 2 所述的方法，其中，所述基于目标模型、辅助模型和所述转化样本集，训练元分类器，包括：

构造包含正负样本的第一元分类器样本集，其中，正样本的样本数据为所述目标模型针对转化样本的梯度信息；负样本的样本数据为所述辅助模型针对转化样本的梯度信息；

使用所述第一元分类器样本集，训练得到第一元分类器。

4. 根据权利要求 3 所述的方法，其中，所述梯度信息为，梯度向量中各元素经符号函数计算后的结果向量。

5. 根据权利要求 3 所述的方法，其中，所述将可疑模型的相关数据输入所述元分类器，

基于所述元分类器的输出结果，确定所述可疑模型是否为从部署模型窃取模型，包括：

- 从所述转化样本集中选取第一转化样本；
- 确定所述可疑模型针对所述第一转化样本的第一梯度信息；
- 将所述第一梯度信息输入所述第一元分类器，得到第一预测结果；
- 响应于所述第一预测结果指示出正样本，确定所述可疑模型为从所述部署模型窃取的模型。

6. 根据权利要求3所述的方法，其中，所述将可疑模型的相关数据输入所述元分类器，基于所述元分类器的输出结果，确定所述可疑模型是否为从部署模型窃取的模型，包括：

- 基于从所述转化样本集选取的第一子集、所述第一元分类器和所述辅助模型，使用假设检验对所述可疑模型进行所有权验证。

7. 根据权利要求6所述的方法，其中，所述使用假设检验对所述可疑模型进行所有权验证，包括：

- 构建第一概率小于等于第二概率的第一原假设，其中，第一概率表示所述第一元分类器针对所述可疑模型的梯度信息的预测结果为正样本的后验概率，第二概率表示所述第一元分类器针对所述辅助模型的梯度信息的预测结果为正样本的后验概率；

- 基于所述第一原假设和所述第一子集中的样本数据，计算P值；

- 响应于确定所述P值小于显著性水平 α ，确定所述第一原假设被拒绝；

- 响应于确定所述第一原假设被拒绝，确定所述可疑模型为从所述部署模型窃取的模型。

8. 根据权利要求1所述的方法，其中，在所述基于目标模型、辅助模型和所述转化样本集，训练元分类器之前，所述方法还包括：

- 响应于所述可疑模型的模型结构未知，将所述部署模型确定为所述目标模型，以及基于所述部署模型的模型结构训练所述辅助模型。

9. 根据权利要求8所述的方法，其中，所述基于目标模型、辅助模型和所述转化样本集，训练元分类器，包括：

- 构造包含正负样本的第二元分类器样本集，其中，正样本的样本数据为，所述目标模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息；

负样本的样本数据为，所述辅助模型针对某选中样本的预测输出与针对该选中样本对应的转化样本的预测输出的差异信息；

使用所述第二元分类器样本集，训练第二元分类器。

10. 根据权利要求 9 所述的方法，其中，所述将可疑模型的相关数据输入所述元分类器，基于所述元分类器的输出结果，确定所述可疑模型是否为从部署模型窃取模型，包括：

分别从所述转化样本集和选中样本集中获取对应的第二转化样本和第二选中样本；

确定所述可疑模型针对所述第二选中样本的预测输出与针对所述第二转化样本的预测输出的第二差异信息；

将所述第二差异信息输入所述第二元分类器，得到第二预测结果；

响应于所述第二预测结果指示出正样本，确定所述可疑模型为从所述部署模型偷取的模型。

11. 根据权利要求 9 所述的方法，其中，所述将可疑模型的相关数据输入所述元分类器，基于所述元分类器的输出结果，确定所述可疑模型是否为从部署模型窃取的模型，包括：

基于从所述转化样本集选取的第二子集、所述选中样本集中与所述第二子集对应的第三子集、所述第二元分类器和辅助模型，使用假设检验对所述可疑模型进行所有权验证。

12. 根据权利要求 11 所述的方法，其中，所述使用假设检验对所述可疑模型进行所有权验证，包括：

构建第三概率小于等于第四概率的第二原假设，其中，第三概率表示，所述第二元分类器针对所述可疑模型对应的差异信息的预测结果为正样本的后验概率，第四概率表示所述第二元分类器针对所述辅助模型对应的差异信息的预测结果为正样本的后验概率；

基于所述第二原假设、所述第二子集的样本数据和第三子集的样本数据，计算 P 值；

响应于确定 P 值小于显著性水平 α ，确定所述第二原假设被拒绝；

响应于确定所述第二原假设被拒绝，确定所述可疑模型为从所述部署模型窃取的模型。

13. 根据权利要求 1 所述的方法，其中，所述初始样本集中初始样本的样本数据为样本图像；以及

所述对所述选中样本集中各样本的样本数据进行处理，得到具有外源特征的转化样本构成的转化样本集，包括：

使用图像风格转换器，对所述选中样本集中各样本的样本图像进行风格转换，使样本图像具有指定图像风格，其中，所述外源特征为所述指定图像风格相关的特征。

14. 一种基于外源特征进行模型所有权验证的装置，包括：

选取单元，配置为从初始样本集中选取部分初始样本构成选中样本集；

转化单元，配置为对所述选中样本集中各选中样本的样本数据进行处理，得到具有外源特征的转化样本构成的转化样本集，其中，所述外源特征为初始样本的样本数据不具备的特征；

训练单元，配置为基于目标模型、辅助模型和所述转化样本集，训练元分类器，其中，所述辅助模型为使用所述初始样本集训练得到的模型，所述目标模型为使用所述转化样本集和所述初始样本集中除所述选中样本集之外的剩余样本集训练得到的模型，所述元分类器用于识别所述外源特征的特征知识；

验证单元，配置为将可疑模型的相关数据输入所述元分类器，基于所述元分类器的输出结果，确定所述可疑模型是否为从部署模型窃取的模型，其中，所述部署模型具有所述外源特征的特征知识。

15. 一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行权利要求 1-13 中任一项所述的方法。

16. 一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-13 中任一项所述的方法。