(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114741322 B
(45)授权公告日 2022.08.16

(21)申请号 202210566644.1
(22)申请日 2022.05.24
(65)同一申请的已公布的文献号
申请公布号 CN 114741322 A
(43)申请公布日 2022.07.12
(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路37号
(72)发明人 吴文峻 杨京波
(74)专利代理机构 北京科迪生专利代理有限责任公司 11251
专利代理师 金怡 邓治平
(51)Int.Cl.
G06F 11/36(2006.01)
G06F 8/75(2018.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

(56)对比文件
CN 113591093 A,2021.11.02
CN 113360906 A,2021.09.07
CN 114253860 A,2022.03.29
EP 0944289 A2,1999.09.22
张睿峰.基于注意力机制的Java API序列推荐方法.《计算机系统应用》.2019,第209-214页.

审查员 刘艳萍

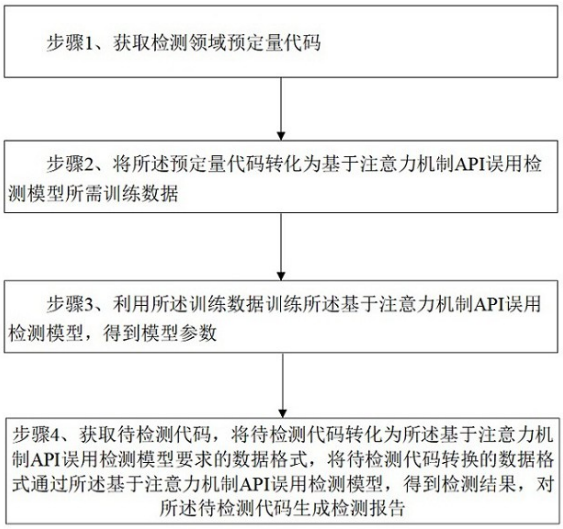
权利要求书3页 说明书7页 附图5页

(54)发明名称

一种基于注意力机制的API误用检测方法及系统

(57)摘要

本发明提出一种基于注意力机制的API误用检测方法及系统，包括如下步骤：步骤1、获取检测领域预定量代码；步骤2、将所述预定量代码转化为基于注意力机制API误用检测模型所需训练数据；步骤3、利用所述训练数据训练所述基于注意力机制API误用检测模型，得到模型参数；步骤4、获取待检测代码，将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式，将待检测代码转换的数据格式通过所述基于注意力机制API误用检测模型，得到检测结果，对所述待检测代码生成检测报告。本发明将自注意力机制应用于API使用规约的学习以及API误用检测，通过采集大量Java代码并通过静态分析构造API使用的样本集，并使用神经网络对这些样本进行训练，构造自动检测API误用的模型。



CN 114741322 B

1.一种基于注意力机制的API误用检测方法,其特征在于,包括如下步骤:

步骤1、获取检测领域预定量代码;

步骤2、将所述预定量代码转化为基于注意力机制API误用检测模型所需训练数据;

步骤3、利用所述训练数据训练所述基于注意力机制API误用检测模型,得到模型参数;所述步骤2将所述预定量代码转化为基于注意力机制API误用检测模型所需训练数据包括:

将代码转化为序列;

将一段代码中的关键信息按照顺序抽取,整理成一个元组,该元组包含该段代码的所有关键信息;所述关键信息包括:方法名称,开发语言关键词,方法中的变量;

将该元组转化为注意力机制API误用检测模型训练所需训练数据,转化方式为:元组中前一个或多个元素和其相邻后一个元素构成一组数据序列,每组数据序列作为一个训练数据,每组数据序列包含以逗号间隔开的两部分,逗号前为Origin部分,逗号后为Target部分;

基于注意力机制API误用检测模型具体如下:

(1)对部分提取的数据序列的Origin,Target部分的每个API进行词嵌入编码,用以增加API的语义信息;

(2)对提取的API数据序列,进行训练,其中根据训练数据的每个序列前半部分即Origin部分作为X,后半部分即Target部分作为Y;

(3)将X作为编码模块的输入,Y作为解码模块的输入,输入到编码模块与解码模块之后,对数据序列中每个词汇根据公式(1)进行计算位置编码PE,位置编码维度与词嵌入编码维度一致:

$$\begin{cases} PE_{(pos,2i)} = \sin(pos / 10000^{2i/d}) \\ PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d}) \end{cases} \quad (1)$$

其中,pos表示单词在句子中的位置,d表示位置编码PE的维度,2i表示偶数的维度,2i+1表示奇数的维度,词嵌入编码和位置编码相加,表示该单词的表示向量;

(4)经过位置编码的向量通过拟合处理,丢弃部分信息,增强模型的健壮性;

(5)将编码后的输入序列embedding(X)通过多头注意力机制层,计算如公式(2)所述,得到多头注意力输出:Multi-Head Attn(X),用以表示X序列中各个API之间依赖的权重关系,之后输入到相加归一化层,即先处理Multi-Head Attn(X)+embedding(X),之后对其和做归一化处理转化为均值为0方差为1的数据,这样起到加快训练速度,加快收敛的作用;

$$\begin{cases} Q_i = QW_i^Q & W_i^Q \in \mathrm{R}^{(d_{model}, d_k)} \\ K_i = KW_i^K & W_i^K \in \mathrm{R}^{(d_{model}, d_k)} \\ V_i = VW_i^V & W_i^V \in \mathrm{R}^{(d_{model}, d_v)} \end{cases} \quad (2)$$

Q,K,V均表示一个输入X,W_i^Q,W_i^K,W_i^V分别表示对Q,K,V进行第i次线性变换的矩阵,Q_i,K_i,V_i表示经过线性变换后的矩阵,d_{model}表示最初序列中每个元素的编码后向

量长度，d_k，d_v分别表示矩阵列宽;R表示实数矩阵；

(6)将步骤(4)的输出最后通过前向神经网络与相加归一化层，得到编码器的输出，即encoderOutput(X)；

(7)将API数据序列的Target部分，按照步骤(4)，进行多头注意力机制时，进行掩码多头注意力操作，即一个序列中的元素仅和该元素前面的信息相关，而和序列中该元素后面的信息无关，用以保持当前输出仅与之前信息相关，得到掩码多头注意力输出Multi-Head Attn(Y)之后，通过相加归一化层进行归一化处理数据；

(8)将编码器的输出 encoderOutput(X)作为解码器的K,V，步骤(7)得到的数据作为Q，进行多头注意力操作，用以计算当前Q对应的API与序列Origin部分API中的权重关系；

(9)最后通过解码模块层，翻译X序列对应的输出概率，此输出即为基于自注意力机制模型的预测下一个调用API；

步骤4、获取待检测代码，将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式，将待检测代码转换的数据格式通过所述基于注意力机制API误用检测模型，得到检测结果，对所述待检测代码生成检测报告。

2.根据权利要求1所述的一种基于注意力机制的API误用检测方法，其特征在于，所述步骤4，获取待检测代码包括：

获取用户需要检测的代码；

获取企业或公司内部新上传的代码。

3.根据权利要求1所述的一种基于注意力机制的API误用检测方法，其特征在于，所述步骤4，对所述待检测代码生成检测报告，检测报告以文本文档形式显示，内容包括API的误用率。

4.根据权利要求1所述的一种基于注意力机制的API误用检测方法，其特征在于，所述步骤3，在训练所述基于注意力机制API误用检测模型时采用交叉熵损失函数，并加入正确率作为训练时的反馈。

5.一种基于注意力机制的API误用检测系统，其特征在于，包括：

代码获取模块，用于获取检测领域预定量代码；

转化模块，将所述预定量代码转化为基于注意力机制API误用检测模型所需训练数据；

训练模块，用于利用所述训练数据训练所述基于注意力机制API误用检测模型，得到模型参数；

所述训练数据包括：

将代码转化为序列；

将一段代码中的关键信息按照顺序抽取，整理成一个元组，该元组包含该段代码的所有关键信息;所述关键信息包括:方法名称，开发语言关键词，方法中的变量；

将该元组转化为注意力机制API误用检测模型训练所需训练数据，转化方式为:元组中前一个或多个元素和其相邻后一个元素构成一组数据序列，每组数据序列作为一个训练数据，每组数据序列包含以逗号间隔开的两部分，逗号前为Origin部分，逗号后为Target部分；

基于注意力机制API误用检测模型具体如下：

(1)对部分提取的数据序列的Origin,Target部分的每个API进行词嵌入编码,用以增加API的语义信息;

(2)对提取的API数据序列,进行训练,其中根据训练数据的每个序列前半部分即Origin部分作为X,后半部分即Target部分作为Y;

(3)将X作为编码模块的输入,Y作为解码模块的输入,输入到编码模块与解码模块之后,对数据序列中每个词汇根据公式(1)进行计算位置编码PE,位置编码维度与词嵌入编码维度一致:

$$\begin{cases} PE_{(pos,2i)} = \sin(pos / 10000^{2i/d}) \\ PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d}) \end{cases} \quad (1)$$

其中,pos表示单词在句子中的位置,d表示位置编码PE的维度,2i表示偶数的维度,2i+1表示奇数的维度,词嵌入编码和位置编码相加,表示该单词的表示向量;

(4)经过位置编码的向量通过拟合处理,丢弃部分信息,增强模型的健壮性;

(5)将编码后的输入序列embedding(X)通过多头注意力机制层,计算如公式(2)所述,得到多头注意力输出:Multi-Head Attn(X),用以表示X序列中各个API之间依赖的权重关系,之后输入到相加归一化层,即先处理Multi-Head Attn(X)+embedding(X),之后对其和做归一化处理转化为均值为0方差为1的数据,这样起到加快训练速度,加快收敛的作用;

$$\begin{cases} Q_i = QW_i^Q & W_i^Q \in \mathrm{R}^{(d_{model}, d_k)} \\ K_i = KW_i^K & W_i^K \in \mathrm{R}^{(d_{model}, d_k)} \\ V_i = VW_i^V & W_i^V \in \mathrm{R}^{(d_{model}, d_v)} \end{cases} \quad (2)$$

Q,K,V均表示一个输入X,W_i^Q,W_i^K,W_i^V分别表示对Q,K,V进行第i次线性变换的矩阵,Q_i,K_i,V_i表示经过线性变换后的矩阵,d_{model}表示最初序列中每个元素的编码后向量长度,d_k,d_v分别表示矩阵列宽;R表示实数矩阵;

(6)将步骤(4)的输出最后通过前向神经网络与相加归一化层,得到编码器的输出,即encoderOutput(X);

(7)将API数据序列的Target部分,按照步骤(4),进行多头注意力机制时,进行掩码多头注意力操作,即一个序列中的元素仅和该元素前面的信息相关,而和序列中该元素后面的信息无关,用以保持当前输出仅与之前信息相关,得到掩码多头注意力输出Multi-Head Attn(Y)之后,通过相加归一化层进行归一化处理数据;

(8)将编码器的输出 encoderOutput(X)作为解码器的K,V,步骤(7)得到的数据作为Q,进行多头注意力操作,用以计算当前Q对应的API与序列Origin部分API中的权重关系;

(9)最后通过解码模块层,翻译X序列对应的输出概率,此输出即为基于自注意力机制模型的预测下一个调用API;

检测模块,用于获取待检测代码,将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式,将待检测代码转换的数据格式通过所述基于注意力机制API误用检测模型,得到检测结果,对所述待检测代码生成检测报告。

一种基于注意力机制的API误用检测方法及系统

技术领域

[0001] 本发明涉及计算机及人工智能技术领域,特别涉及一种基于注意力机制的API误用检测方法及系统。

背景技术

[0002] 随着计算机技术的飞速发展,涌现出越来越多的软件框架、类库等,开发人员只需调用这些框架或类库的应用程序编程接口(Application Programming Interface,API)就能实现多种功能。然而由于编程人员在使用API时忽略文档、或者难以理解API等原因,出现误用API的情况,从而给软件带来缺陷或者隐患。而人为的检测这些API使用一方面维护成本很高,另一方面效率比较低。

[0003] 在目前的API误用研究算法中,基本上可以分为两类:一类是根据API的文档使用自然语言处理技术,提取该API的使用规约,最后根据提取的规约检测API是否误用;另一类是针对现存的API项目集,将API的使用案例转化为API调用情况,根据API调用情况提取为API调用规则,从而根据转化后的规则检测API是否存在误用。在现存的论文研究中,我们发现,无论是从API文档中提取规约,还是在API项目案例中提取API调用规则,很少考虑导致API误用的自注意力机制,即导致API误用的原因可能存在多种情况,但是不同情况的权重不同。API误用的情况可以分为多种,例如空指针误用,即在使用一个API时,应先检测使用API的对象是否为空,在Android开发中设置一个文本框的文字时,textView.setText(“Hello Android”),在使用textView时,应先判断textView控件是否为空,因此这种API误用的检测属于空指针误用的情况所占权重更大;再例如Java在使用文件操作API时,往往会抛出异常,代码中需要加入处理异常的try{}catch(Exception e){}finally{}代码块,例如try{open.file(fileName)}catch(Exception){},对于这种API在检测时,应该更多的关注是否加入try代码块,而对空指针误用可以投入更少的权重。此外,在现存论文研究中,将项目集中的代码转化为序列调用的状态下,一个序列中不同API元素之间的相互依赖程度不完全相同,即在一连串的API调用序列中,可能最后一个元素导致第一个API出现误用,例如在Android开发中给控件设置文字或标题时“getWidgetByName(string.substring(0,10)).setText()”,这一行代码存在多个API,导致 .setText()误用的应为getWidgetByName这个API获取不到对象或者该对象不存在,而不是string.substring()这个API元素,因此API调用序列中不同位置的元素对产生API误用的贡献不一致。此外,有经验的程序员,在检测API是否误用时,针对不同的情况,往往也会关注不同的误用缺陷,比如在检查对象处理时,往往关注对象是否为空,在检查安全API误用时,往往关注API参数等导致产生安全的信息。

发明内容

[0004] 为了解决上述技术问题,本发明提供一种基于注意力机制的API误用检测方法及系统,省去人工检测API代码带来的时间成本以及人力成本。

[0005] 本发明的技术方案为:一种基于注意力机制的API误用检测方法,包括如下步骤:

[0006] 步骤1、获取检测领域预定量代码;

[0007] 步骤2、将所述预定量代码转化为基于注意力机制API误用检测模型所需训练数据;

[0008] 步骤3、利用所述训练数据训练所述基于注意力机制API误用检测模型,得到所述模型参数;

[0009] 步骤4、获取待检测代码,将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式,将待检测代码转换的数据格式通过所述模型,得到检测结果,对所述待检测代码生成检测报告。

[0010] 进一步的,获取检测领域预定量代码包括:

[0011] 如果检测为开源代码,则开源代码质量评价为领域内多星级好评;

[0012] 如果检测代码为企业内部平台或产品代码,则该代码质量应满足所在单位专家审核通过或者该代码所产生的产品或平台为高可用;

[0013] 所述检测领域预定量代码指至少包含3万条代码片段。

[0014] 进一步的,将代码转化为注意力机制模型所需训练数据包括:

[0015] 将代码转化为序列;

[0016] 将一段代码中的关键信息按照顺序抽取,整理成一个元组,该元组包含该段代码的所有关键信息;所述关键信息包括:方法名称,开发语言关键词,方法中的变量;

[0017] 将该元组转化为模型训练所需训练数据,转化方式为:元组中前一个或多个元素和其相邻后一个元素构成一组数据序列,每组数据序列作为一个训练数据,每组数据序列包含两部分,逗号前为Origin部分,逗号后为Target部分。

[0018] 进一步的,基于注意力机制API误用检测模型具体如下:

[0019] (1)对部分提取的数组序列的Origin,Target部分的每个API进行词嵌入编码,用以增加API的语义信息;

[0020] (2)对提取的API数据序列,进行训练,其中根据训练数据的每个序列前半部分即Origin部分作为X,后半部分即Target部分作为Y;

[0021] (3)将X作为编码模块的输入,Y作为解码模块的输入,输入到编码模块与解码模块之后,对数组序列中每个词汇根据公式(1)进行计算位置编码PE,位置编码维度与词嵌入编码维度一致:

[0022]
$$\begin{cases} PE_{(pos,2i)} = \sin(pos / 10000^{2i/d}) \\ PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d}) \end{cases} \quad (1)$$

[0023] 其中,pos表示单词在句子中的位置,d表示位置编码PE的维度,2i表示偶数的维度,2i+1表示奇数的维度,单词编码和位置编码相加,表示该单词的表示向量。

[0024] (4)经过位置编码的向量通过拟合处理,丢弃部分信息,增强模型的健壮性;

[0025] (5)将编码后的输入序列embedding(X)通过多头注意力机制层,计算如公式(2)所述,得到多头注意力输出:Multi-Head Attn(X),用以表示X序列中各个API之间依赖的权重关系,之后输入到相加归一化层,即先处理Multi-Head Attn(X)+embedding(X),之后对其和做归一化处理转化为均值为0方差为1的数据,这样起到加快训练速度,加快收敛的作用;

[0026] $$\begin{cases} Q_i = QW_i^Q & W_i^Q \in \mathbb{R}^{(d_{model},d_k)} \\ K_i = KW_i^K & W_i^K \in \mathbb{R}^{(d_{model},d_k)} \\ V_i = VW_i^V & W_i^V \in \mathbb{R}^{(d_{model},d_v)} \end{cases} \quad (2)$$

[0027] Q,K,V均表示图3中的一个输入X，W_i^Q，W_i^K，W_i^V分别表示对Q,K,V进行第i次线性变换的矩阵，Q_i，K_i，V_i表示经过线性变换后的矩阵，d_{model}表示最初序列中每个元素的编码后向量长度，d_k，d_v分别表示矩阵列宽;R表示实数矩阵；

[0028] (6)将步骤(4)的输出最后通过前向神经网络与相加归一化层，得到编码器的输出，即encoderOutput(X)；

[0029] (7)将API数据序列的Target部分，按照步骤(4)，进行多头注意力机制时，所示进行掩码多头注意力操作，即一个序列中的元素仅和该元素前面的信息相关，而和序列中该元素后面的信息无关，用以保持当前输出仅与之前信息相关，得到掩码多头注意力输出Multi-Head Attn(Y)之后，通过相加归一化层进行归一化处理数据；

[0030] (8)将编码器的输出 encoderOutput(X)作为解码器的K,V，步骤(7)得到的数据作为Q，进行多头注意力操作，用以计算当前Q对应的API与序列Origin部分API中的权重关系；

[0031] (9)最后通过解码模块层，翻译X序列对应的输出(Output)概率，此输出即为基于自注意力机制模型的预测下一个调用API。

[0032] 进一步的，所述步骤4，获待检测代码包括：

[0033] 用户需要检测的代码；

[0034] 企业或公司内部最新上传的代码，所述最新指新产品或软件或平台形成前所需的新代码。

[0035] 进一步的，模型的结果对所述待检测代码生成检测报告指：

[0036] 报告以文本文档形式显示。

[0037] 进一步的，在训练模型时采用交叉熵损失函数，并加入正确率作为训练时的反馈。

[0038] 根据本发明发另一方面，还提出一种基于注意力机制的API误用检测系统，包括：

[0039] 代码获取模块，用于获取检测领域预定量代码；

[0040] 转化模块，用于根据获取的代码，将所述代码转化为基于注意力机制API误用检测模型所需训练数据；

[0041] 训练模块，用于利用所述训练数据训练所述基于注意力机制API误用检测模型，得到所述模型参数；

[0042] 检测模块，用于获取待检测代码，将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式，将待检测代码转换的数据格式通过所述模型，得到检测结果，对所述待检测代码生成检测报告。

[0043] 有益效果：

[0044] 本发明将自注意力机制应用于API使用规约的学习以及API误用检测，通过采集大量Java代码并通过静态分析构造API使用的样本集，并使用神经网络对这些样本进行训练，构造自动检测API误用的模型。本发明提供的检测API误用的方法，在训练数据时，准确率以

及损失值优于已存在的基于LSTM(长短时记忆网络)的方法,同时训练收敛速度较快。发明还将预测结果与现有的API误用检测技术对比,结果表明,采用本文自注意力机制的方法在一定程度上优于其他算法。

附图说明

[0045] 图1为本发明的一种基于注意力机制的API误用检测方法流程图;

[0046] 图2为本发明的基于注意力机制API误用检测模型示意框图;

[0047] 图3 为损失值和准确率曲线示意图;

[0048] 图4为 F1值效果示意图;

[0049] 图5 为精确率曲线示意图;

[0050] 图6 为召回率示意图;

[0051] 图7 为本发明的系统操作示意图。

具体实施方式

[0052] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整的描述,显然,所描述的实施例仅为本发明的一部分实施例,而不是全部的实施例,基于本发明中的实施例,本领域的普通技术人员在不付出创造性劳动的前提下所获得的所有其他实施例,都属于本发明的保护范围。

[0053] 实施例一

[0054] 根据本发明的实施例,如图1所示,提出一种基于注意力机制的API误用检测方法,其中包括如下步骤:

[0055] 步骤1、获取检测领域预定量代码;

[0056] 步骤2、将所述预定量代码转化为基于注意力机制API误用检测模型所需训练数据;

[0057] 根据本发明的一个实施例,代码包含如下关键信息:方法名称(例如开源框架中的可被调用的方法或函数),开发语言关键词(例如while,for,do,foreach,try,catch等),方法中的变量(例如整形变量,全局变量等),将一段代码中的关键信息按照顺序抽取,整理成一个元组,该元组包含该段代码的所有关键信息。

[0058] 例如如下代码:

[0059] “ while(int i =0; i < 10; i++){

[0060] s = Java.util.Text.subString(0, 1);

[0061] }”

[0062] 转化为如下元组:[while, Java.util.Text.subString(0, 1)]

[0063] 代码中若含有4个关键信息,则转化的元组为:[A,B,C,D],进一步将该元组转化为模型训练所需数据,转化方式为:前一个或多个元素和其相邻后一个元素构成一组或多组数据序列(即API数据序列),例如上述元组转化为[A,B],[AB,C],[ABC,D]这3个训练数据,每个数据都是一个训练序列,可以看出每个训练序列包含两部分,逗号前为Origin部分,逗号后为Target部分;本发明优选采用高质量代码进行转换,此种转化目的在于,使用某一个API或者某一个开源框架的一套高质量代码写法一般是固定的,通过利用转化的数据进行

训练,可以从该代码中获取使用的规则或者方式。

[0064] 步骤3、利用所述数据训练所述基于注意力机制API误用检测模型,得到模型参数;

[0065] 其中,所述步骤3中的基于注意力机制的API误用检测模型,如图2所示:

[0066] (1)对部分提取的数组序列的Origin,Target部分的每个API进行词嵌入编码,用以增加API的语义信息;

[0067] (2)对所述部分提取的API数据序列,进行训练,其中根据训练数据的每个序列前半部分即Origin部分作为图3中的X,后半部分即Target部分作为Y;

[0068] (3)将X作为编码模块的输入,Y作为解码模块的输入,输入到编码模块与解码模块之后,对数组序列中每个单词根据公式(1)进行计算位置编码PE,位置编码维度与词嵌入编码维度一致:

[0069] $$\begin{cases} PE_{(pos,2i)} = \sin(pos / 10000^{2i/d}) \\ PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d}) \end{cases} \quad (1)$$

[0070] 其中,pos表示单词在句子中的位置,d表示位置编码PE的维度,2i表示偶数的维度,2i+1表示奇数的维度,单词编码和位置编码相加,表示该单词的表示向量。

[0071] (4)经过位置编码的向量通过拟合处理,丢弃部分信息,增强模型的健壮性;

[0072] (5)将位置编码后的输入序列embedding(X)通过多头注意力机制层,计算如公式(2)所述,得到多头注意力输出:Multi-Head Attn(X),用以表示X序列中各个API之间依赖的权重关系,之后输入到相加归一化层,即先处理得到Multi-Head Attn(X)+embedding(X),之后对其和做归一化处理转化为均值为0方差为1的数据,这样起到加快训练速度,加快收敛的作用;

[0073] $$\begin{cases} Q_i = QW_i^Q & W_i^Q \in \mathbb{R}^{(d_{model},d_k)} \\ K_i = KW_i^K & W_i^K \in \mathbb{R}^{(d_{model},d_k)} \\ V_i = VW_i^V & W_i^V \in \mathbb{R}^{(d_{model},d_v)} \end{cases} \quad (2)$$

[0074] Q,K,V均表示图3中的一个输入X,W_i^Q,W_i^K,W_i^V分别表示对Q,K,V进行第i次线性变换的矩阵,Q_i,K_i,V_i表示经过线性变换后的矩阵,d_{model}表示最初序列中每个元素的编码后向量长度,d_k,d_v分别表示矩阵列宽;R表示实数矩阵;

[0075] (6)将步骤(4)的输出最后通过前向神经网络与相加归一化层,得到编码器的输出,即encoderOutput(X);

[0076] (7)将API数据序列的Target部分,按照步骤(4),进行多头注意力机制时,所述进行掩码多头注意力操作,即一个序列中的元素仅和该元素前面的信息相关,而和序列中该元素后面的信息无关,用以保持当前输出仅与之前信息相关,得到掩码多头注意力输出Multi-Head Attn(Y)之后,通过相加归一化层进行归一化处理数据;

[0077] (8)将编码器的输出 encoderOutput(X)作为解码器的K,V,步骤(7)得到的数据作为Q,进行多头注意力操作,用以计算当前Q对应的API与序列Origin部分API中的权重关系;

[0078] (9)最后通过解码模块层,翻译X序列对应的输出(Output)概率,此输出即为基于

自注意力机制模型的预测下一个调用API；

[0079] 进一步的，在训练模型时采用交叉熵损失函数，并加入正确率作为训练时的反馈；

[0080] 至此，本发明搭建了一个基于自注意力机制的API误用检测模型。

[0081] 步骤4、获取待检测代码，将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式，将待检测代码转换的数据格式通过所述基于注意力机制API误用检测模型，得到检测结果，对所述待检测代码生成检测报告。

[0082] 本步骤中，获取待检测代码，将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式，将待检测代码转换的数据格式通过所述模型，得到检测结果，例如第一个API，模型检测到该API后面应该使用M这个API的概率为 60%，但是却使用了N或其他的API，使用其他的API就是误用率，如果第一个API后面可以使用A和B两个API，可以使用A的概率为33%，可使用B的概率为22%，当使用其他API时，就表示API误用，对所述待检测代码生成检测报告。

[0083] 其中图3中单词对应关系为：Transformer表示本发明方法模型，S-LSTM表示堆栈化长短时序列模型，D-LSTM表示深度长短时序列模型，横坐标epoch表示迭代次数，纵坐标loss表示训练过程中的损失函数，acc表示训练过程中的准确率，由图3可以看出，本发明提出的方法从损失函数收敛速度优于其他的两个方法，准确率高于其他的两个方法，说明了本发明的有效性。

[0084] 图4，图5，图6中的Top-K表示预测结果的前K个，3-gram表示n-gram模型中n取3时候的模型，查准率表示precision，Recall表示召回率，F1-score表示查准率和召回率的均值，从图4-6可以看出召回率，查准率，F1-Score均比已有的方法要好。

[0085] 实施例2：

[0086] 根据本发明的一个实施例，提出一种基于注意力机制的API误用检测系统，包括：

[0087] 代码获取模块，用于获取检测领域预定量代码；

[0088] 转化模块，用于根据获取的代码，将所述代码转化为基于注意力机制API误用检测模型所需训练数据；

[0089] 训练模块，用于利用所述训练数据训练所述基于注意力机制API误用检测模型，得到所述模型参数；

[0090] 检测模块，用于获取待检测代码，将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式，将待检测代码转换的数据格式通过所述模型，得到检测结果，对所述待检测代码生成检测报告。

[0091] 根据本发明的一个实施例，所述的系统采用实施例1的模型，在计算机上实现，如图7所示。当有新程序或者新代码出现时，操作人员可以在上图中的“更新程序片段界面”输入相关代码，这样会把相关代码存入数据库，有新代码进入数据库后，需要对模型重新进行训练，操作人员点击“模型训练操作界面”，便可以采用实施例1中训练模型的步骤，将实验更新后的代码进行训练，训练完毕模型后，操作人员可以在“待检测程序输入界面”输入自己要检测的程序，等检测完毕，便可以在“输出报告界面”查看待检测程序中API误用情况。

[0092] 根据本发明的一个实施例，对API误用检测系统可以设置定期(例如1天或者3天)爬取指定地址的代码，以及设置定期训练模型，例如github网站，或者公司自己网站指定的代码，这样便实现系统自动更新代码的功能，将爬取的代码自动更新到步骤1所述的数据

库,定期训练模型设置后,系统便根据最新的数据库里面的代码进行模型训练,这样起到实时更新模型的作用。

[0093] 根据本发明的一个实施例,API误用检测系统还可以获取设备的地理位置信息,同时系统里面存储不同的地理位置信息对应的公司或者其他信息,根据地理位置信息自动判断需要爬取何种类型的代码,例如如果获取的地理位置信息为 北京五彩城,因为小米在五彩城,那么系统根据该位置信息判断使用本系统的单位为小米,可能小米采用某种技术(例如Android里面的通信安全的API)比较多,那么该系统会自动爬取相关通信安全的API代码,存储到系统的数据库,并自动进行模型训练,这样起到模型根据用户或者公司自身特点自动训练模型的作用。

[0094] 根据本发明的一个实施例,API误用检测系统还可以根据用户经常存储的代码判断用户需要更新何种类型的代码库,例如用户经常更新数据存储API使用的代码,那么系统可以定期自动爬取相关网站(这些网站系统里面预先由提供商或者用户定义,比如数据存储对应A网站地址,通信安全对应B网站地址)数据存储API的代码,爬取到代码后,便可以自动的重新训练模型。

[0095] 根据本发明的一个实施例,API误用检测系统还可以根据用户个人信息判断所需更新何种类型代码库(登陆系统时用户可以输入个人信息),例如某单位女生偏重js等前端界面开发,男生偏重java等后台服务器开发,如果一个经常操作该误用检测系统的是女生,那么该系统可以自动爬取前端开发相关网页的代码片段,并更新到数据库,更新完毕后,自动训练模型。

[0096] 尽管上面对本发明说明性的具体实施方式进行了描述,以便于本技术领域的技术人员理解本发明,且应该清楚,本发明不限于具体实施方式的范围,对本技术领域的普通技术人员来讲,只要各种变化在所附的权利要求限定和确定的本发明的精神和范围内,这些变化是显而易见的,一切利用本发明构思的发明创造均在保护之列。

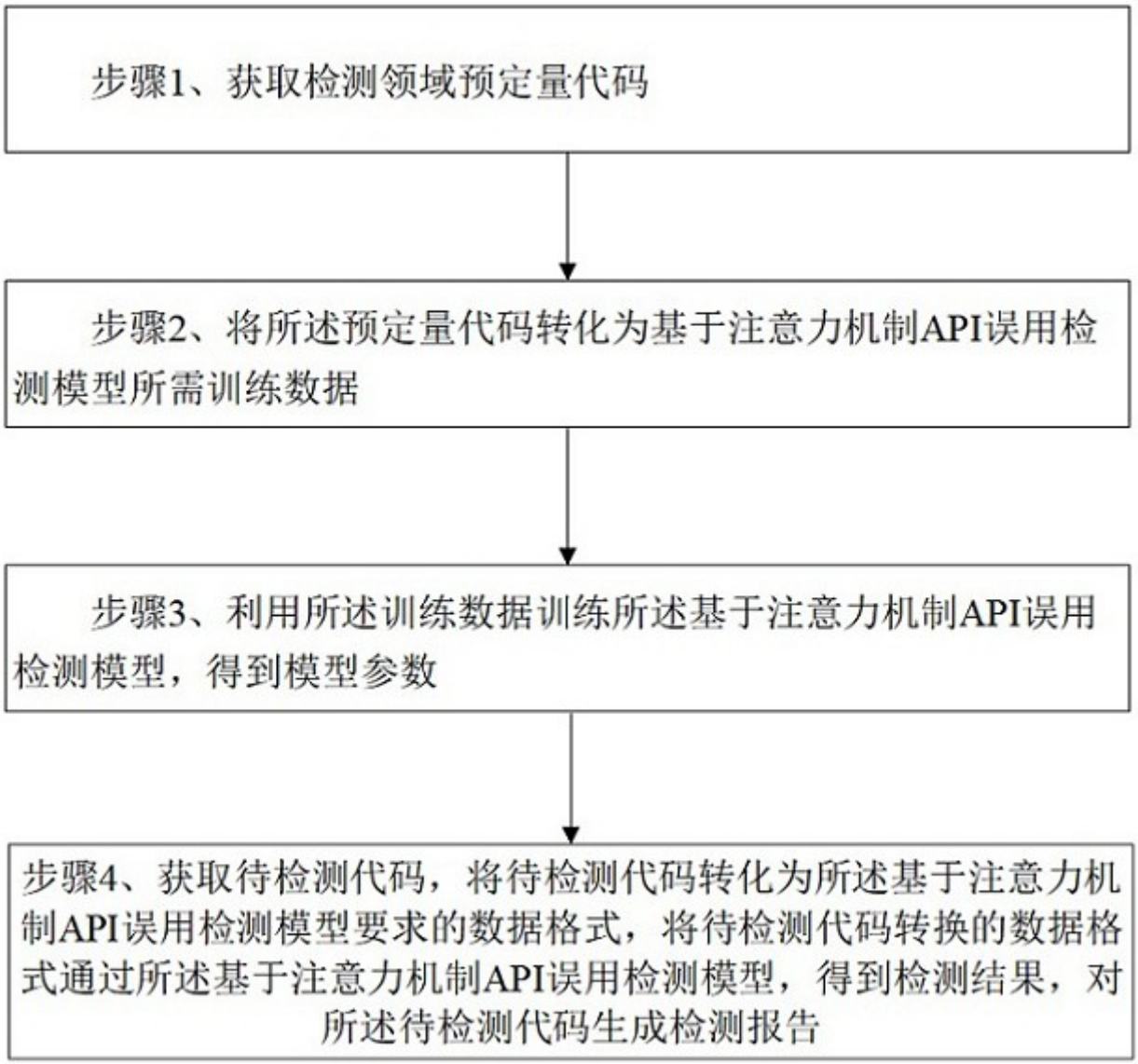

步骤1、获取检测领域预定量代码
步骤2、将所述预定量代码转化为基于注意力机制API误用检测模型所需训练数据
步骤3、利用所述训练数据训练所述基于注意力机制API误用检测模型，得到模型参数
步骤4、获取待检测代码，将待检测代码转化为所述基于注意力机制API误用检测模型要求的数据格式，将待检测代码转换的数据格式通过所述基于注意力机制API误用检测模型，得到检测结果，对所述待检测代码生成检测报告

图1

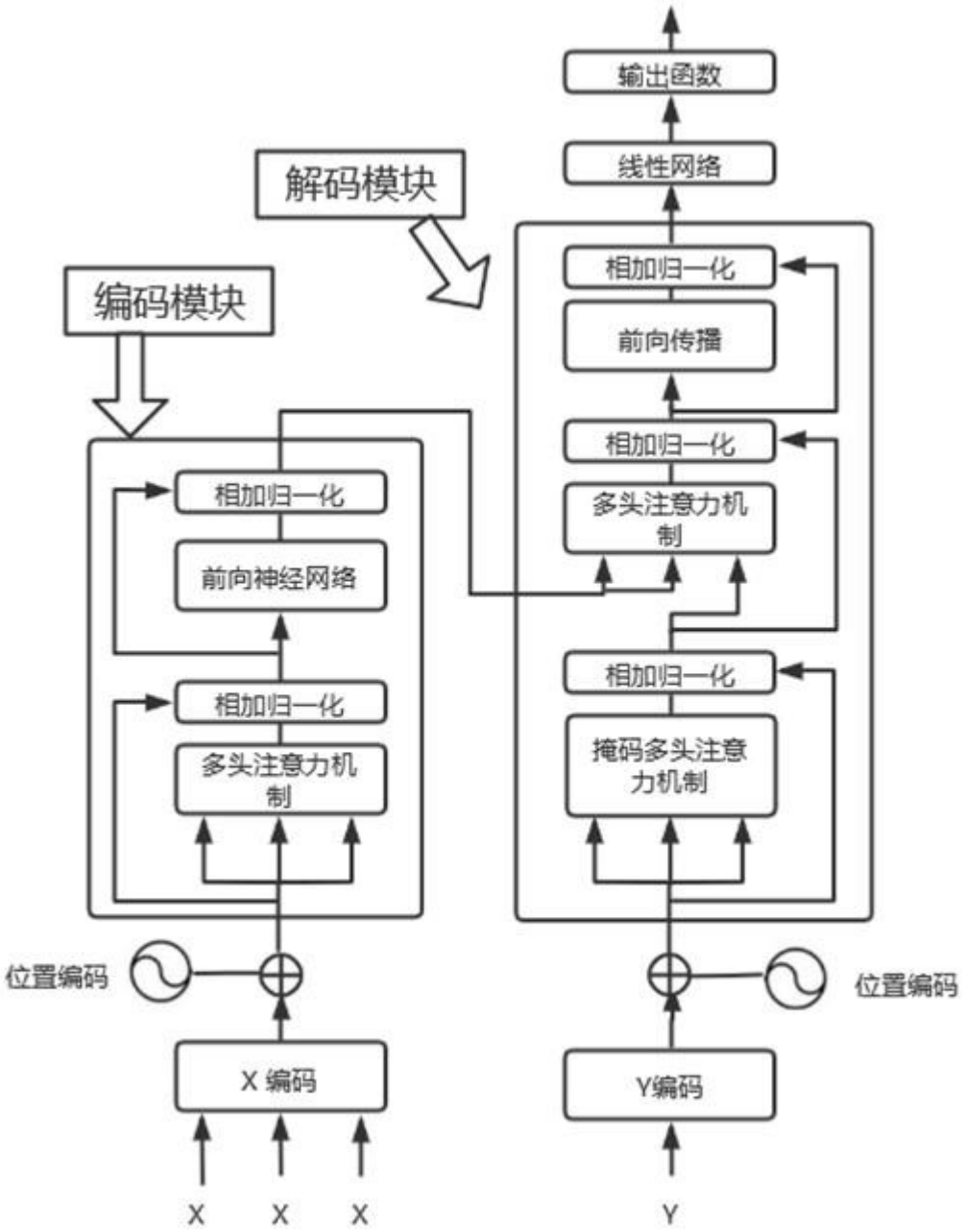

输出函数
线性网络
解码模块
编码模块
相加归一化
前向传播
相加归一化
多头注意力机制
相加归一化
前向神经网络
相加归一化
多头注意力机制
相加归一化
掩码多头注意力机制
位置编码
位置编码
X 编码
Y编码
X
X
X
Y

图2

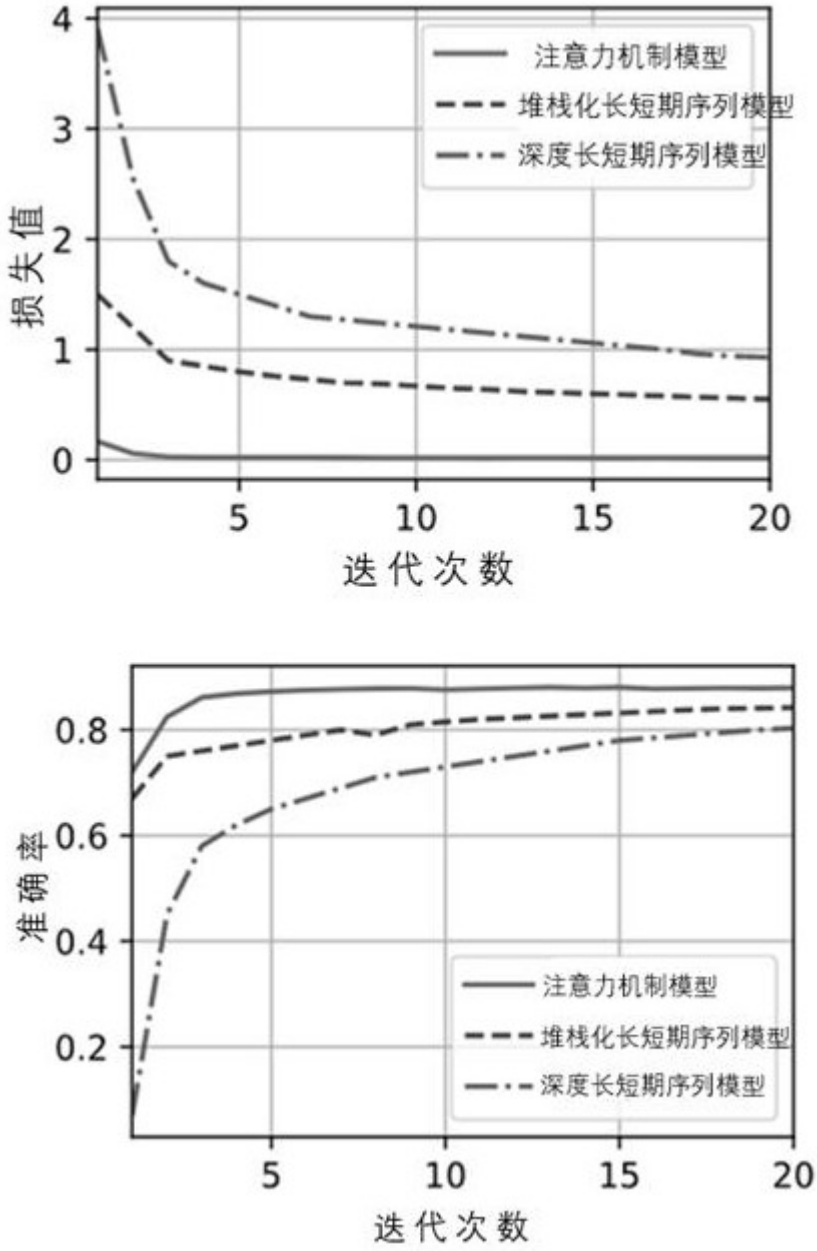


图3

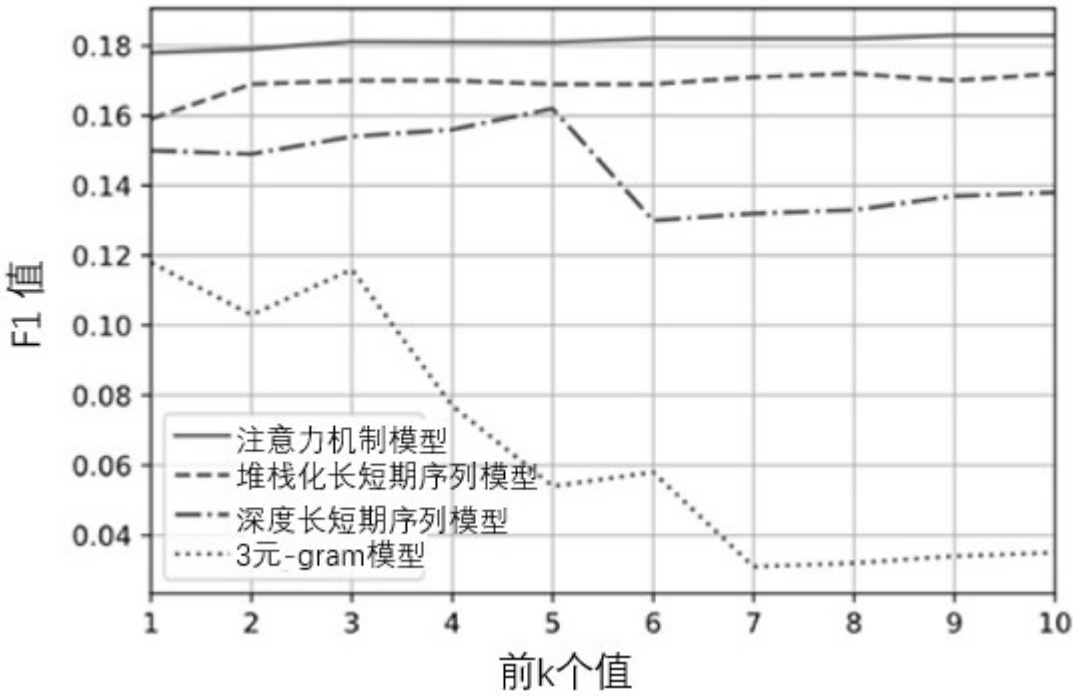

0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
F1 值
注意力机制模型
堆栈化长短期序列模型
深度长短期序列模型
3元-gram模型
1
2
3
4
5
6
7
8
9
10
前k个值

图4

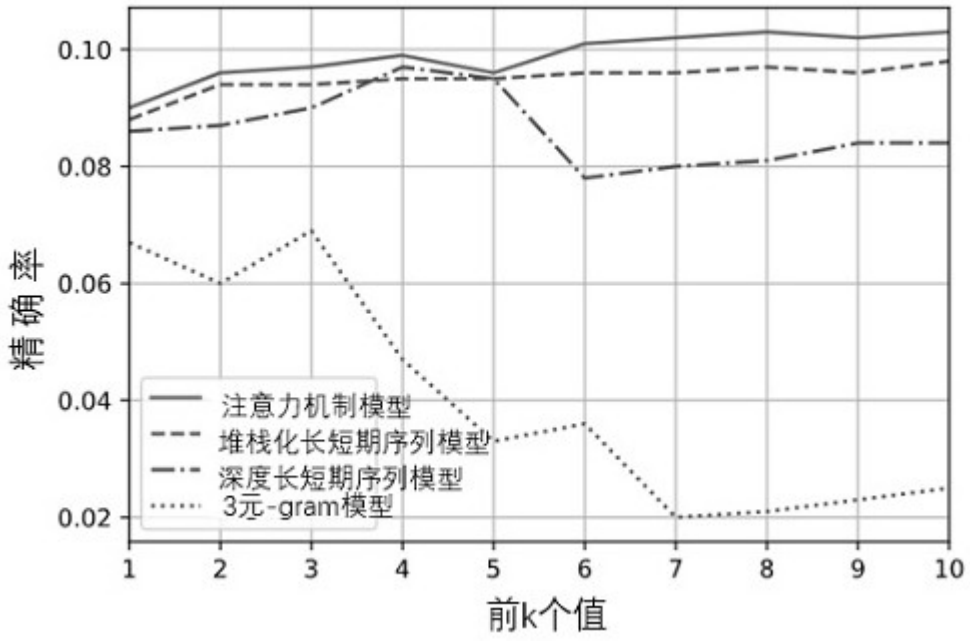

0.10
0.08
0.06
0.04
0.02
精确率
注意力机制模型
堆栈化长短期序列模型
深度长短期序列模型
3元-gram模型
1
2
3
4
5
6
7
8
9
10
前k个值

图5

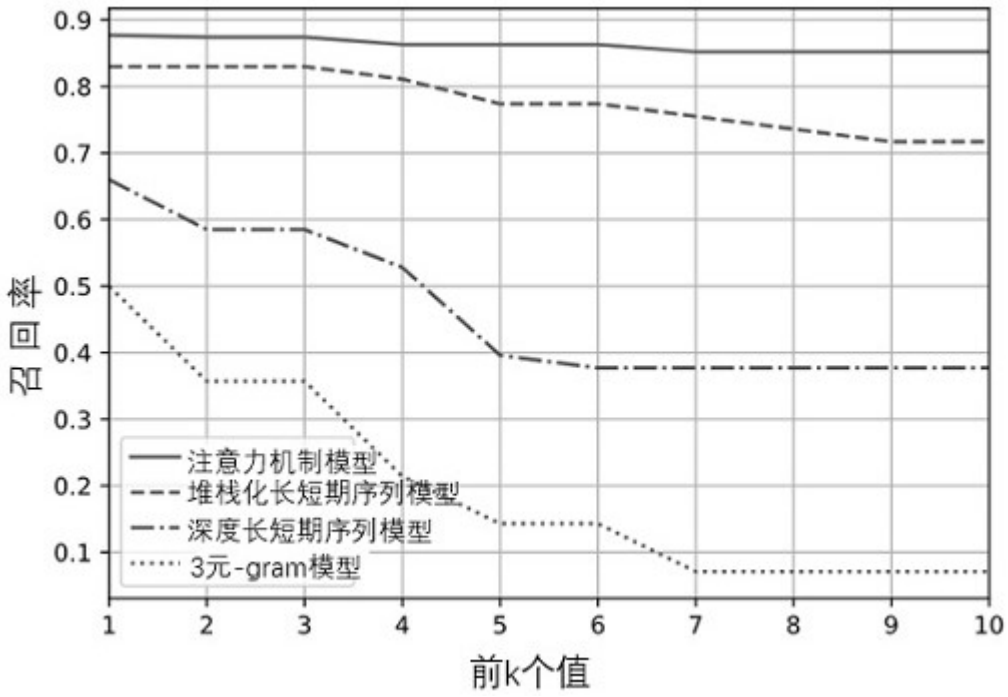

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
召回率
注意力机制模型
堆栈化长短期序列模型
深度长短期序列模型
3元-gram模型
1
2
3
4
5
6
7
8
9
10
前k个值

图6

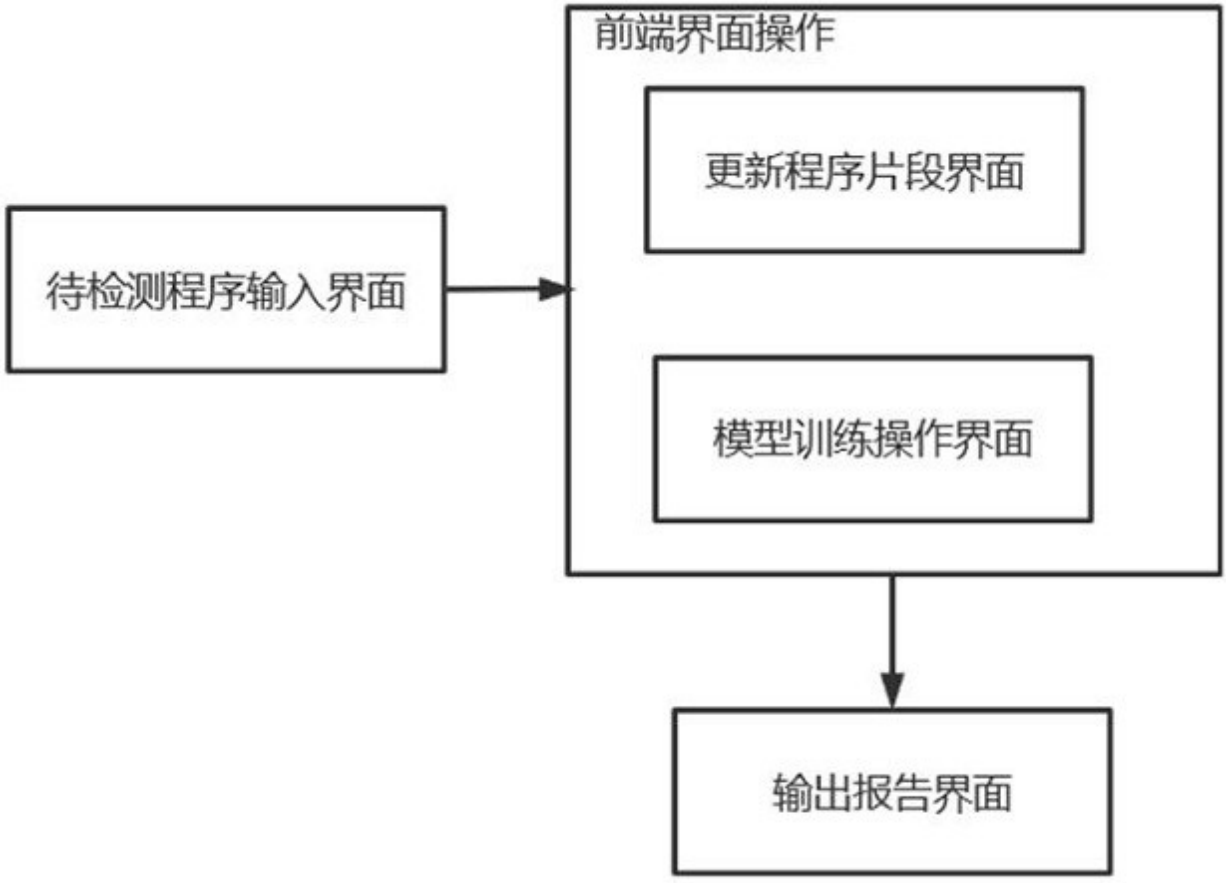

前端界面操作
更新程序片段界面
待检测程序输入界面
模型训练操作界面
输出报告界面

图7