(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107633058 A
(43)申请公布日 2018.01.26

(21)申请号 201710853173.1

(22)申请日 2017.09.20

(71)申请人 武汉虹旭信息技术有限责任公司
地址 430205 湖北省武汉市江夏区藏龙岛谭湖二路1号虹信无线通信产业园

(72)发明人 张成 戴长江

(74)专利代理机构 武汉宇晨专利事务所 42001
代理人 黄瑞棠

(51)Int.Cl.
G06F 17/30(2006.01)

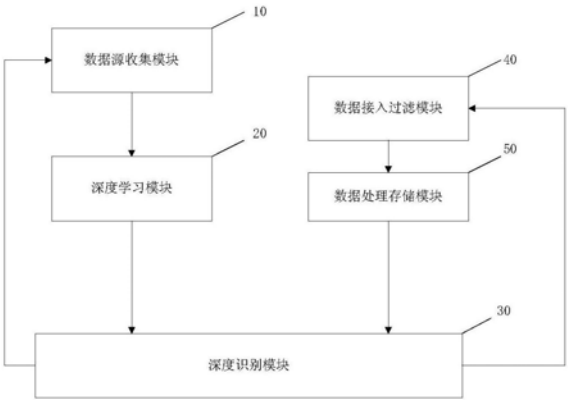
权利要求书3页 说明书6页 附图3页

(54)发明名称

一种基于深度学习的数据动态过滤系统及其方法

(57)摘要

本发明公开了一种基于深度学习的数据动态过滤系统及其方法，涉及网络数据分析领域。本系统包括数据源收集模块(10)、深度学习模块(20)、深度识别模块(30)、数据接入过滤模块(40)和数据处理存储模块(50)；数据源收集模块(10)、深度学习模块(20)和深度识别模块(30)依次循环交互；数据接入过滤模块(40)、数据处理存储模块(50)和深度识别模块(30)次循环交互。本发明较传统的数据分析扩大了网络数据分析的领域，并能倒推出复杂数据基本特征，对其相关数据进行有针对性的全面分析。



CN 107633058 A

1.一种基于深度学习的数据动态过滤系统,其特征在于:

包括数据源收集模块(10)、深度学习模块(20)、深度识别模块(30)、数据接入过滤模块(40)和数据处理存储模块(50);

其交互关系是:

数据源收集模块(10)、深度学习模块(20)和深度识别模块(30)依次循环交互,数据源收集模块(10)收集数据给深度学习模块(20)进行学习,深度学习模块(20)提供计算图和参数给深度识别模块(30)进行识别,深度识别模块(30)将识别出来的复杂数据给数据源收集模块(10)进行收集;

数据接入过滤模块(40)、数据处理存储模块(50)和深度识别模块(30)依次循环交互,数据接入过滤模块(40)根据基本特征过滤网络数据给数据处理存储模块(50),数据处理存储模块(50)将复杂数据提供给深度识别模块(30)进行识别,深度识别模块(30)识别后将复杂数据的基本特征提供给数据接入过滤模块(40)。

2.按权利要求1所述的基于深度学习的数据动态过滤系统,其特征在于:

所述的数据源收集模块(10)的工作流程如下:

a、开始(200)

启动数据源收集模块,要具备基本的数据预处理功能,对图像和频谱图进行裁剪、缩放和旋转功能,还需具备数据集分类功能,另外对噪音数据要进行处理;

b、使用网络获取数据下载经典数据集(201)

网络获取数据和下载经典数据集都是获取数据源的重要方法;

c、人工收集特殊数据集(202)

对于特殊应用场景,需采用人工方法进行数据源的采集;

d、收取深度识别模块发来的复杂数据(203)

深度识别模块发来的复杂数据一般是具备特征的优质数据源,应该收集;

e、读取POSIX、HDFS和/或GCS文件系统中的数据(204)

还有一部分数据是存储在POSIX、HDFS和/或GCS文件系统中的,特别是大数据往往存储在HDFS和GCS文件系统中,也需要收集;

f、预处理并整理成训练、验证和测试三类数据集(205)

数据集需要整理成训练、验证和测试三类,以便深度学习模块(20)使用。

3.按权利要求1所述的基于深度学习的数据动态过滤系统,其特征在于:

所述的深度学习模块(20)的工作流程如下:

A、开始(300)

tensorflow开始初始化,迁移学习情况下先加载模型;

B、定义前向传播算法(301)

前向传播算法,在深度学习中需要激活输入数据,一般情况下采用Relu算法,对于复杂数据采用卷积神经网络或循环神经网络算法进行深层次的特征提取,加入隐藏层、卷积层和池化层进行计算;

C、定义反向传播算法(302)

反向传播算法对深度学习的模型进行优化,通过损失函数即交叉熵或均方差的计算,对模型进行收敛,优化函数则根据参数的调优情况进行选择;

D、定义多线程、队列与GPU设备(303)

使用多线程、队列和GPU设备提高训练的速度,即多线程采用coordinator和start_queue_runners函数,队列采用string_input_producer函数,GPU安装CUDA以便使用;

E、开始训练并验证(304)

使用数据源收集模块(10)提供的数据进行训练,从输入层到输出层进行优化,即学习率、随机采样和正则化,对部分数据生成batch进行验证,通过梯度下降得到一个优化的结果;

F、保存训练结果(305)

训练结果保存后,发给深度识别模块(30)对复杂数据进行识别。

4.按权利要求1所述的基于深度学习的数据动态过滤系统,其特征在于:

所述的深度识别模块(30)的工作流程如下:

I、开始(400)

模块开始初始化,不同的识别需求对应不同的进程,并需要提供不同的计算图和参数;

Ⅱ、加载并初始化模型(401)

加载计算图与参数,读取protobuf格式的数据,在andriod系统上数据大小一般不超过64M,硬盘上大小一般不超过512M,为计算数与参数创建会话;

Ⅲ、读取复杂数据并进行预处理(402)

将音频与信号转换成相应的频谱图流,然后对图像进行预处理,包括大小,翻转、亮度、对比度、色相、饱和度和标注,行成一个张量;

Ⅳ、在模型上运行复杂数据(403)

在模型上运行复杂数据,行成一个输出,这个输出也是一个张量,在分类问题中,张量包含每个类别的概率大小;

Ⅴ、给出TOP5的结果(404)

对于分类问题,给出概率最大的前五名的类别名称,并呈现出来供分析记录;

Ⅵ、对识别的复杂数据及基本特征进行处理(405)

为数据源收集模块10提供复杂数据,为数据接入过滤模块40提供复杂数据的基本特征。

5.按权利要求1所述的基于深度学习的数据动态过滤系统,其特征在于:

所述的数据接入过滤模块(40)的工作流程如下:

i、开始(500)

模块开始初始化,Cavium公司的HFA硬件初始化,分配相应的内存并加载基本特征,为每个数据包创建TCP或UDP数据流;

ⅱ、从网络数据包中提取基本特征(501)

从网络数据包中提取基本特征,即数据的ip、五元组、URL或HOST,以及特殊字段;

ⅲ、从深度识别模块获取基本特征并与数据包匹配(502)

深度识别模块30会把复杂数据的基本特征发给本模块,此时由HFA协处理器进行五元组匹配或字符串的模糊匹配;

ⅳ、按流过滤并向后发送数据(503)

匹配基本特征的数据包,按照其关联上的流数据进行过滤并发给数据处理存储模块

50。

6.按权利要求1所述的基于深度学习的数据动态过滤系统，其特征在于：

所述的数据处理存储模块(50)的工作流程如下：

α、开始(600)

模块开始初始化，传统的深度报文检测启动，数据库启动；

β、对网络数据进行处理(601)

将网络数据包的基本特征与复杂数据关联起来，即数据的ip，五元组，URL或HOST，以及特殊字段与复杂数据对应起来；

γ、将相关基本特征与复杂数据信息存储起来(602)

将相关基本特征与复杂数据信息存储起来，目前对于大数据采用流行的HDFS文件系统进行存储，由深度识别模块30读取文件进行识别。

7.基于权利要求1-6所述系统的数据动态过滤方法，其特征在于包括下列步骤：

①首先启动数据源收集模块，方式为人工收集、下载、或读取POSIX、HDFS、GCS三种文件系统中的数据，并收集在深度识别模块中经过识别的数据，这些数据用来给深度学习模块进行学习；

②数据收集后，发给深度学习模块进行卷积神经网络或循环神经网络计算，将训练的计算图和参数结果进行保存，并发送给深度识别模块用来对复杂数据进行匹配；

③此时，启动数据接入过滤模块，对接入的网络数据根据基本特征进行过滤，将过滤匹配的网络数据发给数据处理存储模块；

④数据处理存储模块会进行传统的数据分析并存储待识别的复杂数据；

⑤深度识别模块此时读取复杂数据进行识别：将识别出的复杂数据发给数据源收集模块用来训练，此处又经第①步骤循环进行；将识别出的复杂数据的基本特征进行提取并发给数据接入过滤系统进行过滤，以进行针对性的全面分析，此处又经第③步骤循环进行。

一种基于深度学习的数据动态过滤系统及其方法

技术领域

[0001] 本发明涉及网络数据分析领域，尤其涉及一种基于深度学习的数据动态过滤系统及其方法。

背景技术

[0002] 自2016年3月AlphaGo在围棋上战胜李世石以来，人工智能掀起一波热潮，智能驾驶，百度大脑紧跟而上，其核心技术在于深度学习。深度学习在图像处理、自然语言处理和音频处理等领域与传统的人工智能相比，大幅提高了准确率与智能化，在计算机性能不断提高以及GPU不断发展的今天，通过深度学习从海量数据中识别并提取特征已经可以实现，目前比较热门的软件平台有Google公司的tensorflow。

[0003] 在网络数据分析领域，前端数据接入由Cavium公司出产的OCTEON网络处理芯片进行处理，其吞吐性能与协处理器功能比较先进，在海量网络数据的分析与过滤上性能较高。

发明内容

[0004] 本发明的目的就在于针对网络数据分析领域，通过复杂网络数据即图像、音频和信号数据倒推出其基本特征，对相关源数据进行过滤，提供一种基于深度学习的数据动态过滤系统及其方法。

[0005] 实现本发明目的技术方案是：

[0006] 首先启动数据源收集模块，人工收集、下载、和/或读取linux、HDFS和/或GCS等文件系统中的数据，并收集在深度识别模块中经过识别的数据；数据收集后，发给深度学习模块tensorflow进行训练，深度学习将训练的计算图和参数结果进行保存，并发送给深度识别模块用来进行匹配；此时，启动数据接入过滤模块，将数据发给数据处理存储模块；数据处理存储模块会进行传统的数据分析并存储数据；深度识别模块此时读取复杂数据进行识别，将识别出的复杂数据发给数据源收集模块用来训练，将识别出的复杂数据的基本特征进行提取并发给数据接入过滤系统进行过滤，以进行专门的分析。

[0007] 具体地说：

[0008] 一、基于深度学习的数据动态过滤系统(简称系统)

[0009] 本系统包括数据源收集模块、深度学习模块、深度识别模块、数据接入过滤模块和数据处理存储模块；

[0010] 其交互关系是：

[0011] 数据源收集模块、深度学习模块和深度识别模块依次循环交互，数据源收集模块收集数据给深度学习模块进行学习，深度学习模块提供计算图和参数给深度识别模块进行识别，深度识别模块将识别出来的复杂数据给数据源收集模块进行收集；

[0012] 数据接入过滤模块、数据处理存储模块和深度识别模块依次循环交互，数据接入过滤模块根据基本特征过滤网络数据给数据处理存储模块，数据处理存储模块50将复杂数据提供给深度识别模块进行识别，深度识别模块30识别后将复杂数据的基本特征提供给数

据接入过滤模块。

[0013] 二、基于深度学习的数据动态过滤方法(简称方法)

[0014] 本方法包括下列步骤：

[0015] ①首先启动数据源收集模块，方式为人工收集、下载、或读取POSIX、HDFS、GCS三种文件系统中的数据，并收集在深度识别模块中经过识别的数据，这些数据用来给深度学习模块进行学习；

[0016] ②数据收集后，发给深度学习模块进行卷积神经网络或循环神经网络计算，将训练的计算图和参数结果进行保存，并发送给深度识别模块用来对复杂数据进行匹配；

[0017] ③此时，启动数据接入过滤模块，对接入的网络数据根据基本特征进行过滤，将过滤匹配的网络数据发给数据处理存储模块；

[0018] ④数据处理存储模块会进行传统的数据分析并存储待识别的复杂数据；

[0019] ⑤深度识别模块此时读取复杂数据进行识别：将识别出的复杂数据发给数据源收集模块用来训练，此处又经第①步骤循环进行；将识别出的复杂数据的基本特征进行提取并发给数据接入过滤系统进行过滤，以进行针对性的全面分析，此处又经第③步骤循环进行。

[0020] 本发明具有以下优点和积极效果：

[0021] ①使用tensorflow框架，其源码由Google公司提供，稳定，更新速度快，支持GPU计算，易于维护和移植；

[0022] ②深度学习扩大了网络数据分析的领域，并提高了准确率；

[0023] ③可自动识别出的复杂数据的基本特征进行提取并发给数据接入过滤系统进行过滤，以进行专门的分析，加强了数据分析的针对性与全面性；

[0024] ④运用Cavium公司出产的OCTEON网络处理器进行协处理器过滤，性能较高；

[0025] ⑤动态过滤，可实时应对情况的变化。

[0026] 总之，本发明较传统的数据分析扩大了网络数据分析的领域，并能倒推出复杂数据基本特征，对其相关数据进行有针对性的全面分析。

附图说明

[0027] 图1为本系统的结构方框图；

[0028] 图2为数据源收集模块10的工作流程图；

[0029] 图3为深度学习模块20的工作流程图；

[0030] 图4为深度识别模块30的工作流程图；

[0031] 图5为数据接入过滤模块40的工作流程图；

[0032] 图6为数据处理存储模块50的工作流程图。

[0033] 图中：

[0034] 10—数据源收集模块；

[0035] 20—深度学习模块；

[0036] 30—深度识别模块；

[0037] 40—数据接入过滤模块；

[0038] 50—数据处理存储模块。

[0039] 英译汉：

[0040] 1、AlphaGo：由Google公司设计的围棋智能机器人；

[0041] 2、HDFS：；运行在通用硬件上的分布式文件系统；

[0042] 3、GCS：Google公司的分布式文件系统；

[0043] 4、tensorflow：Google公司的深度学习软件平台；

[0044] 5、Cavium：全球领先的多核MIPS和ARM处理器提供商；

[0045] 6、OCTEON：Cavium公司提供的网络处理芯片；

[0046] 7、POSIX：UNIX的可移植操作系统接口；

[0047] 8、batch：深度学习的批处理对象，检验数据集；

[0048] 9、URL：统一资源定位符，互联网上资源位置和访问方法的简结表示；

[0049] 10、HOST：互联网中的主机名称；

[0050] 11、GPU：显卡芯片；

[0051] 12、Relu：一种神经元激活函数算法；

[0052] 13、CUDA：NVIDIA英伟达公司提供的运算平台，实现GPU协同处理；

[0053] 14、HFA：OCTEON芯片的模糊匹配协处理器；

[0054] 15、protobuf：Google公司定义的一种数据交换格式。

具体实施方式

[0055] 以下结合附图和实施例详细说明。

[0056] 一、系统

[0057] 1、总体

[0058] 如图1，本系统包括数据源收集模块10、深度学习模块20、深度识别模块30、数据接入过滤模块40和数据处理存储模块50；

[0059] 其交互关系是：

[0060] 数据源收集模块10、深度学习模块20和深度识别模块30依次循环交互，数据源收集模块10收集数据给深度学习模块20进行学习，深度学习模块20提供计算图和参数给深度识别模块30进行识别，深度识别模块30将识别出来的复杂数据给数据源收集模块10进行收集；

[0061] 数据接入过滤模块40、数据处理存储模块50和深度识别模块30依次循环交互，数据接入过滤模块40根据基本特征过滤网络数据给数据处理存储模块50，数据处理存储模块50将复杂数据提供给深度识别模块30进行识别，深度识别模块30识别后将复杂数据的基本特征提供给数据接入过滤模块40。

[0062] 工作机理是：

[0063] 数据源收集模块10和深度学习模块20交互，将数据集传给深度学习模块20，深度学习模块20和深度识别模块30交互，将计算图和参数传给深度识别模块30；

[0064] 深度识别模块30和数据源收集模块10交互，将识别的复杂数据传给数据源收集模块10组成数据集的一部分；

[0065] 数据接入过滤模块40和数据处理存储模块50交互，将网络数据传给数据处理存储模块50进行传统的分析，并进行存储，数据处理存储模块50与深度识别模块30交互，将复杂

数据传给深度识别模块30,进行识别;

[0066] 深度识别模块30和数据接入过滤模块40,将复杂数据的基本特征传给数据接入过滤模块40,进行动态过滤。

[0067] 2、功能模块

[0068] 1)数据源收集模块10

[0069] 人工收集、下载、网络获取和/或读取POSIX、HDFS和/或GCS三种文件系统中的数据,并收集在深度识别模块30中经过识别的复杂数据;

[0070] 具体地说,如图2,数据源收集模块10的工作流程如下:

[0071] a、开始-200

[0072] 启动数据源收集模块,要具备基本的数据预处理功能,对图像和频谱图进行裁剪、缩放和旋转功能,还需具备数据集分类功能,另外对噪音数据要进行处理;

[0073] b、使用网络获取数据下载经典数据集-201

[0074] 网络获取数据和下载经典数据集都是获取数据源的重要方法;

[0075] c、人工收集特殊数据集-202

[0076] 对于特殊应用场景,需采用人工方法进行数据源的采集;

[0077] d、收取深度识别模块发来的复杂数据-203

[0078] 深度识别模块发来的复杂数据一般是具备特征的优质数据源,应该收集;

[0079] e、读取POSIX、HDFS和/或GCS文件系统中的数据-204

[0080] 还有一部分数据是存储在POSIX、HDFS和/或GCS文件系统中的,特别是大数据往往存储在HDFS和GCS文件系统中,也需要收集;

[0081] f、预处理并整理成训练、验证和测试三类数据集-205

[0082] 数据集需要整理成训练、验证和测试三类,以便深度学习模块20使用。

[0083] 2)深度学习模块20

[0084] 采用tensorflow框架,通过前向传播算法激活,反向传播算法优化,对不同复杂数据采用卷积神经网络或循环神经网络算法进行深层次的特征提取,最终训练出高准确率的计算图和参数;

[0085] 具体地说,如图3,深度学习模块20的工作流程如下:

[0086] A、开始-300

[0087] tensorflow开始初始化,迁移学习情况下先加载模型;

[0088] B、定义前向传播算法-301

[0089] 前向传播算法,在深度学习中需要激活输入数据,一般情况下采用Relu算法,对于复杂数据采用卷积神经网络或循环神经网络算法进行深层次的特征提取,加入隐藏层、卷积层和池化层进行计算;

[0090] C、定义反向传播算法-302

[0091] 反向传播算法对深度学习的模型进行优化,通过损失函数即交叉熵或均方差的计算,对模型进行收敛,优化函数则根据参数的调优情况进行选择;

[0092] D、定义多线程、队列与GPU设备-303

[0093] 使用多线程、队列和GPU设备提高训练的速度,即多线程采用coordinator和start_queue_runners函数,队列采用string_input_producer函数,GPU安装CUDA以便使

用；

[0094] E、开始训练并验证-304

[0095] 使用数据源收集模块10提供的数据进行训练，从输入层到输出层进行优化，即学习率、随机采样和正则化，对部分数据生成batch进行验证，通过梯度下降得到一个优化的结果；

[0096] F、保存训练结果-305

[0097] 训练结果保存后，发给深度识别模块30对复杂数据进行识别。

[0098] 3)深度识别模块30

[0099] 采用深度学习模块20提供的计算图和参数，对复杂数据即图像、音频、视频和信号数据进行预处理和识别，为数据源收集模块10提供数据，为数据接入过滤模块40提供复杂数据的基本特征；

[0100] 具体地说，如图4，深度识别模块30的工作流程如下：

[0101] Ⅰ、开始-400

[0102] 模块开始初始化，不同的识别需求对应不同的进程，并需要提供不同的计算图和参数；

[0103] Ⅱ、加载并初始化模型-401

[0104] 加载计算图与参数，读取protobuf格式的数据，在andriod系统上数据大小一般不超过64M，硬盘上大小一般不超过512M，为计算数与参数创建会话；

[0105] Ⅲ、读取复杂数据并进行预处理-402

[0106] 将音频与信号转换成相应的频谱图流，然后对图像进行预处理，包括大小，翻转、亮度、对比度、色相、饱和度和标注，行成一个张量；

[0107] Ⅳ、在模型上运行复杂数据-403

[0108] 在模型上运行复杂数据，行成一个输出，这个输出也是一个张量，在分类问题中，张量包含每个类别的概率大小；

[0109] Ⅴ、给出TOP5的结果-404

[0110] 对于分类问题，给出概率最大的前五名的类别名称，并呈现出来供分析记录；

[0111] Ⅵ、对识别的复杂数据及基本特征进行处理-405

[0112] 为数据源收集模块10提供复杂数据，为数据接入过滤模块40提供复杂数据的基本特征。

[0113] 4)数据接入过滤模块40

[0114] 对网络数据进行接入，按复杂数据的基本特征进行动态过滤，过滤其相关的数据；

[0115] 具体地说，如图5，数据接入过滤模块40的工作流程如下：

[0116] ⅰ、开始-500

[0117] 模块开始初始化。Cavium公司的HFA硬件初始化，分配相应的内存并加载基本特征，为每个数据包创建TCP或UDP数据流；

[0118] ⅱ、从网络数据包中提取基本特征-501

[0119] 从网络数据包中提取基本特征，即数据的ip、五元组、URL或HOST，以及特殊字段；

[0120] ⅲ、从深度识别模块获取基本特征并与数据包匹配-502

[0121] 深度识别模块30会把复杂数据的基本特征发给本模块，此时由HFA协处理器进行

五元组匹配或字符串的模糊匹配；

[0122] ⅳ、按流过滤并向后发送数据-503

[0123] 匹配基本特征的数据包，按照其关联上的流数据进行过滤并发给数据处理存储模块50。

[0124] 5)数据处理存储模块50

[0125] 传统的数据分析，将复杂的数据进行存储，以供深度模块30处理。

[0126] 具体地说，如图6，数据处理存储模块50的工作流程如下：

[0127] α、开始-600

[0128] 模块开始初始化，传统的深度报文检测启动，数据库启动；

[0129] β、对网络数据进行处理-601

[0130] 将网络数据包的基本特征与复杂数据关联起来，即数据的ip，五元组，URL或HOST，以及特殊字段与复杂数据对应起来；

[0131] γ、将相关基本特征与复杂数据信息存储起来-602

[0132] 将相关基本特征与复杂数据信息存储起来，目前对于大数据采用流行的HDFS文件系统进行存储，由深度识别模块30读取文件进行识别。

[0133] 3、本动态过滤系统的工作机理

[0134] 数据源收集模块和深度学习模块连接，将数据集传给深度学习模块，深度学习模块和深度识别模块连接，将计算图与参数传给深度识别模块；

[0135] 深度识别模块与数据源收集模块，将识别的复杂数据传给数据源收集模块组成数据集的一部分；

[0136] 数据接入过滤模块和数据处理存储模块连接，将网络数据传给数据处理存储模块进行传统的分析，并进行存储，数据处理存储模块与深度识别模块连接，将复杂数据传给深度识别模块，进行识别；

[0137] 深度识别模块和数据接入过滤模块，将复杂数据的基本特征传给数据接入过滤模块，进行动态过滤。

[0138] 本发明应用举例，如在网络数据中发现一张图片，经过深度识别模块发现是张三的，而该图片对应的账号是为“hacker”的用户的，于是在数据接入过滤模块中将账号为“hacker”的网络数据过滤出来进行分析，达到了通过图片查找目标及过滤分析目标相关网络数据的结果。

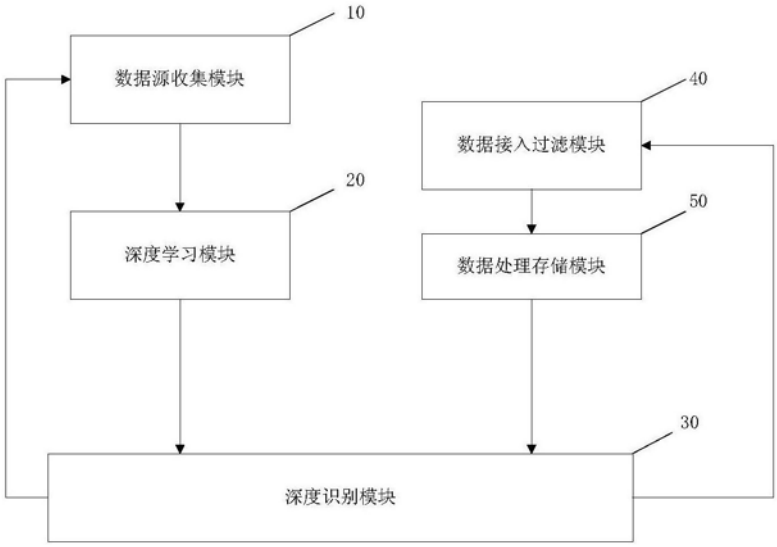

10
数据源收集模块
20
深度学习模块
40
数据接入过滤模块
50
数据处理存储模块
30
深度识别模块

图1

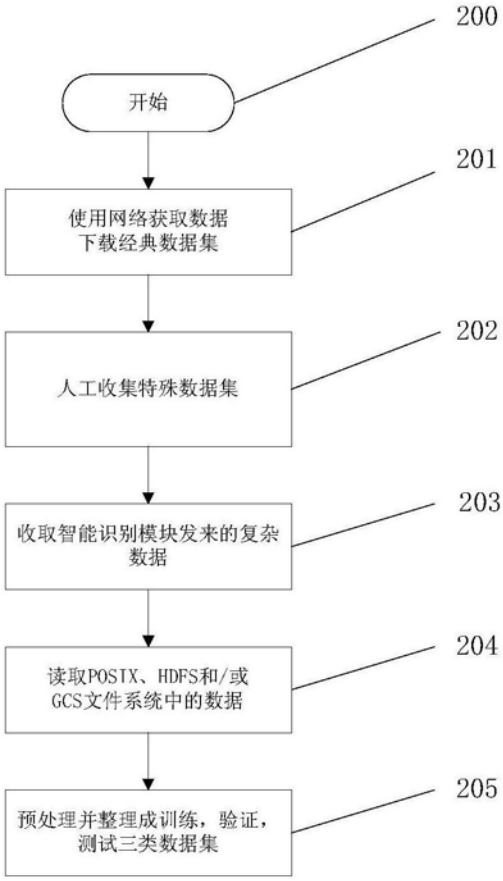

200
开始
201
使用网络获取数据
下载经典数据集
202
人工收集特殊数据集
203
收取智能识别模块发来的复杂
数据
204
读取POSIX、HDFS和/或
GCS文件系统中的数据
205
预处理并整理成训练，验证，
测试三类数据集

图2

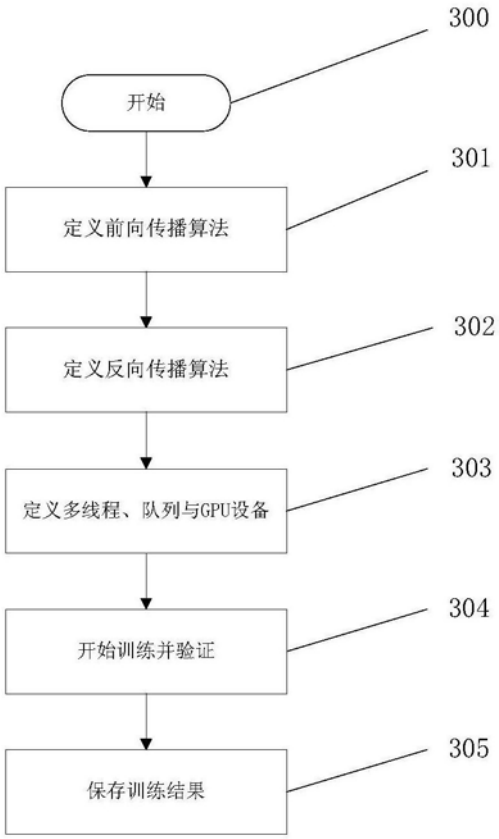

300
开始
301
定义前向传播算法
302
定义反向传播算法
303
定义多线程、队列与GPU设备
304
开始训练并验证
305
保存训练结果

图3

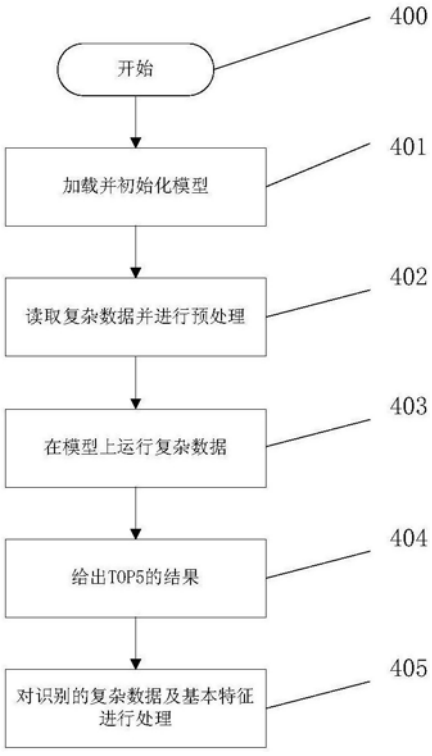

400
开始
401
加载并初始化模型
402
读取复杂数据并进行预处理
403
在模型上运行复杂数据
404
给出TOP5的结果
405
对识别的复杂数据及基本特征进行处理

图4

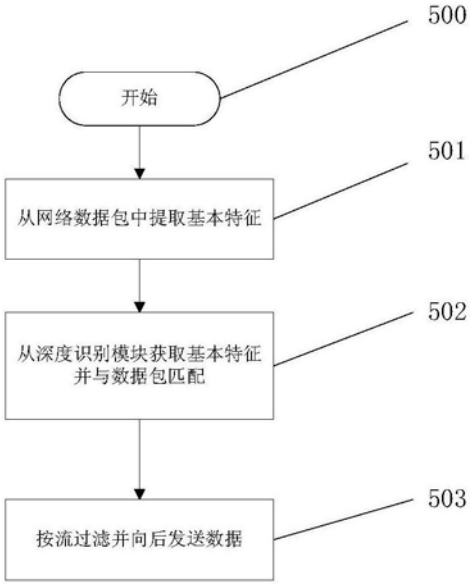

500
开始
501
从网络数据包中提取基本特征
502
从深度识别模块获取基本特征并与数据包匹配
503
按流过滤并向后发送数据

图5

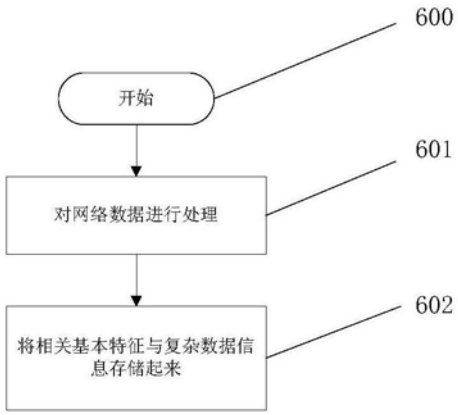

600
开始
601
对网络数据进行处理
602
将相关基本特征与复杂数据信息存储起来

图6