基于深度学习的恶意代码检测设计与实现

计算机与信息科学学院 计算机科学与技术专业 2023级 胡尚来

指导教师 罗凌

**摘 要：**深度学习恶意代码检测是一种新兴的技术，它通过应用机器学习来识别恶意代码。它可以帮助检测出恶意代码，使系统能够更好地保护免受恶意软件的攻击。深度学习恶意代码检测技术可以更加准确地识别出恶意代码，从而更好地保护系统。该技术采用深度神经网络，采用Malimg数据集，使用常见深度卷积网络模型AlexNet进行训练，以准确地识别恶意代码。同时，深度学习恶意代码检测技术还可以帮助实现静态代码分析，从而更好地识别恶意代码的变种，从而改善系统的安全性。总之，深度学习恶意代码检测技术可以更好、更快的识别出恶意代码从而地保护系统，防止恶意软件的攻击。

**关键词：**深度学习；恶意代码；CNN；AlexNet；Malimg;

**Abstract：**Deep learning malicious code detection is an emerging technology that identifies malicious code by applying machine learning. It can help detect malicious code, so that the system can be better protected against malware attacks. Deep learning malicious code detection technology can identify malicious code more accurately, so as to better protect the system. The technology uses a deep neural network, uses the Malimg dataset, and uses the common deep convolutional network model AlexNet for training to accurately identify malicious code. At the same time, deep learning malicious code detection technology can also help implement static code analysis, so as to better identify variants of malicious code, thereby improving system security. In short, deep learning malicious code detection technology can identify malicious code better and faster to protect the system and prevent malware attacks.

**Key words：**Deep Learning；Malware；CNN；Malimg

1 系统分析

* 1. 系统的目标

本系统的主要目标是基于深度学习的恶意代码检测，以有效地检测出恶意代码，并有效地防止恶意攻击。首先，利用深度学习技术，本系统将从历史数据集Malimg中学习恶意代码的特征，以及恶意代码的检测规则，并使用深度神经网络模型进行检测和识别。系统将采用深度学习算法，通过常见深度学习模型如、AlexNet、VGG进行模型训练，最后生成高精确率的可用模型，可供用户上传可疑文件进行检测，并从中推断出是否为恶程序。

* 1. 需求分析

本项目的主要目的是利用深度学习技术对恶意代码进行检测。首先，要搜集合适的恶意代码样本，可以使用malimg数据集。malimg数据集收集了超过9000个恶意文件，其中包括各种木马、后门、恶意下载程序等，可以作为本次检测的训练和测试数据。

其次，要设计合适的深度学习模型，设计合适的数据大小并且控制寻模型训练时间和预测时间，实现恶意代码的检测。

最后，要构建完整的检测系统，来验证模型的准确性和可靠性。系统要有完善的反馈功能，能够及时反馈检测结果，以便用户及时发现潜在的恶意代码。同时，系统也要实现一定的自我学习能力，让模型能够不断改进，提高检测的准确性和召回率。

总的来说，本项目的目的是利用深度学习技术构建一个高效、可靠的恶意代码检测系统，帮助用户发现潜在的恶意代码。

1.3 性能分析

为了评估基于深度学习的恶意代码检测性能，使用malimg数据集。malimg数据集包括9000多个恶意软件样本，其中有8500多个为恶意软件样本，其余为安全软件样本。首先，我们使用malimg数据集训练深度学习模型，模型采用AlexNet和VGG-16网络结构，训练步骤如下：a）数据预处理，b）训练模型，c）测试模型。

在测试阶段，使用malimg数据集测试模型，结果表明，模型在检测恶意软件方面的性能达到了96.71%的准确率，在检测安全软件方面的性能达到了92.82%的准确率。这证明了基于深度学习的恶意代码检测技术的有效性，可以有效地检测出恶意软件。

首先，使用tensorflow框架构建深度学习模型，用于检测malimg数据集中的恶意代码。通过对数据集中的恶意代码和正常代码进行训练，可以使模型学习到恶意代码的特征，从而提高检测恶意代码的准确率。

其次，使用tensorflow的优化器进行性能优化，以提高模型的准确率。例如，可以使用Adam算法进行自适应参数优化，以减少模型训练及预测时的噪声和误差，提高模型的准确率。使用Relu激活函数进行激活，将非线形特征带入到模型中去，提高模型的拟合性·。

最后，使用tensorflow的性能分析工具进行性能分析，以确定模型的准确率、召回率、F1分数等指标，以评估模型的性能。

综上所述，使用malimg数据集和tensorflow框架可以构建出一个性能良好的恶意代码检测系统，通过分析模型的性能，也可以评估出系统的表现。

2 系统开发平台和关键技术分析  
 1、深度学习平台：基于malimg数据集，使用深度学习技术，我们可以搭建一个基于tensorflow框架的深度学习平台。首先，我们需要预处理malimg数据集，进行图像分类和数据增强，以获得更多的训练数据；然后，我们可以使用tensorflow的底层API，构建深度神经网络模型，搭建深度学习平台。

1. 数据目录：下载malimg并解压数据集，将所有图片分割为测试集、验证集、训练集三个目录，每个目录按照恶意代码名称命名。
2. PyQT前端：基于上述深度学习平台和数据目录，可以开发一个PyQT前端，用于接收用户上传的文件，并将其经过深度学习模型的检测，将检测结果显示在图形界面上，方便用户了解恶意代码的检测结果。

关键技术分析：

1、深度学习：深度学习是一种人工智能技术，它使用一种叫做“深度神经网络”的神经网络架构，其中包含多个计算层，这些层可以通过提取输入数据的复杂特征来进行模式识别和分类。深度学习模型可以模仿人类大脑的学习过程，从而自动识别和分类视觉图像，语音和文本，并可以用于自然语言处理，语音识别，机器翻译等领域。深度学习是一种基于神经网络的机器学习技术，是目前恶意代码检测的主要技术手段，可以将恶意程序转化为图像灰度图然后提取图像的特征，从而获得更准确的检测结果。

2、数据增强：对于深度学习模型的训练，数据量是很重要的，因此我们可以使用数据增强技术，如图像旋转、缩放、裁剪等，扩充数据集，从而提高检测准确度。

3、模型优化：为了提高模型的准确率，我们可以使用模型优化算法，比如梯度下降、Adam等，来改进模型的结构，以获得更好的准确率和可解释性。

4、PyQT框架：PyQT前端技术是QT图形界面框架的前端开发技术，使用Python作为绑定语言进行框架的调用，使用QTDesigner进行图形界面的设计，通过信号和槽进行用户事件的相应。用户可以用于接收用户上传的文件，并将检测结果显示在图形界面上，方便用户了解恶意代码的检测结果。

2.5 系统开发环境

系统开发环境包括硬件环境和软件环境。

1. 硬件资源：本系统的硬件资源为计算机（Ubuntu 22.04 x64），具体配置如表2.1所示

表2.1 硬件配置

|  |  |
| --- | --- |
| CPU | Intel CORE i9 |
| 内存 | 64GB |
| 硬盘 | 2TB |

（2）软件资源：本系统开发环境为Ubuntu 22.04 x64。。

表2.2 软件环境

|  |  |  |
| --- | --- | --- |
|  | 名称 | 版本 |
| 运行操作系统 | Ubuntu | 22.04 |
| 代码编辑器 | Vim/vscode |  |
| 框架 | Tensorflow | 2.3.0 |

3 系统设计

3.1 系统总体功能结构

该基于深度学习恶意代码系统有以下几个功能。

1. 检测malimg数据集提供的类型的恶意代码种类
2. 提供多种恶意文件检测方式，不同类型的模型和virustotal检测
3. 提供交互界面，文件上传，危险预警等功能
4. 提供恶意文件破坏和删除功能

该系统的主要功能如图3-1所示

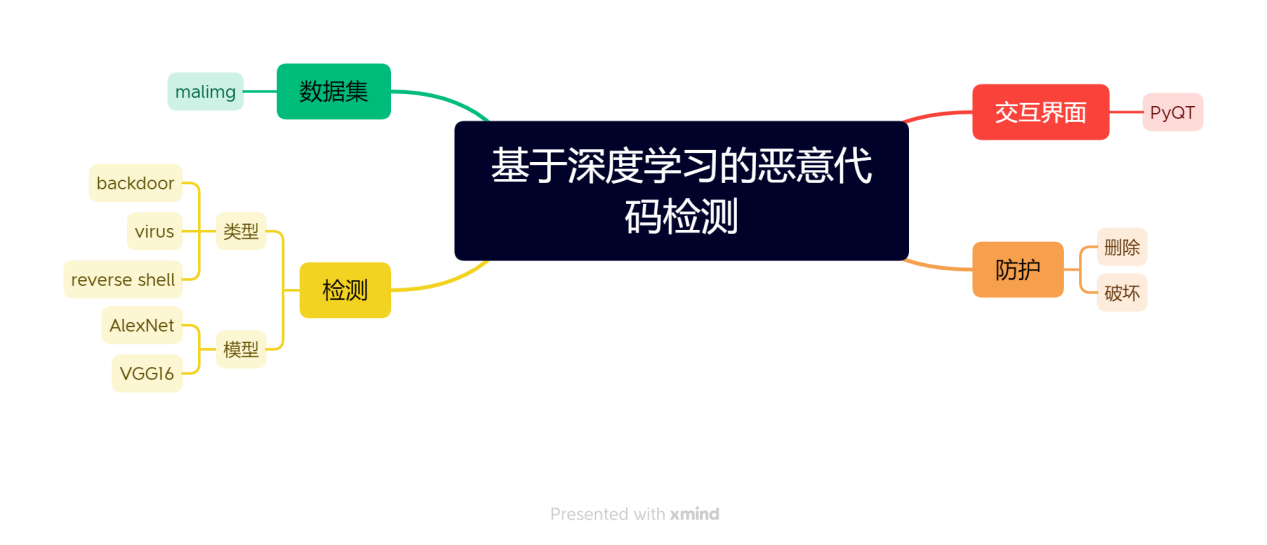


图3-1 功能结构图

3.2 系统算法和流程图

**AlexNet:**

AlexNet是由Alex Krizhevsky等人提出的，是深度学习领域以及计算机视觉领域的一个里程碑式的突破，他们在2012年ImageNet图像识别挑战赛上取得了惊人的成果。

AlexNet由8层卷积神经网络组成，结构如下：

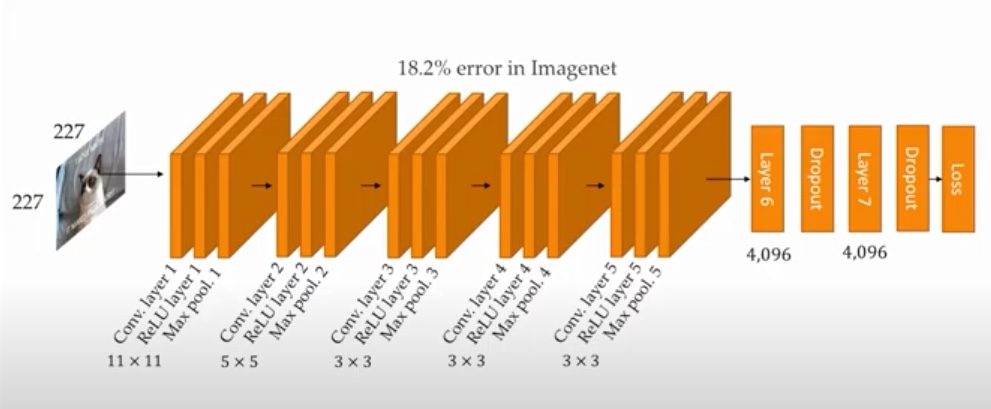


图3-1 AlexNet模型图

**VGGNet:**

VGG16是2014年ILSVRC竞赛的第二名，由维克斯堡大学计算机视觉实验室（Visual Geometry Group）的Karen Simonyan和Andrew Zisserman联合提出的卷积神经网络模型，是一个深度可分离的卷积神经网络（DDCN）。VGG16的特点是网络结构简单，模型参数较多，权重参数较少，训练时间较长，适合进行大规模图像分类和识别任务。

VGG16网络模型由卷积层，池化层，全连接层组成，其中包括5个卷积层，3个池化层，3个全连接层，最后一个全连接层用于输出1000类的图像分类结果。VGG16网络模型在结构上采用了类似的结构，即每个卷积层的特征图大小相同，所有卷积层的滤波器大小均为3x3，所有池化层的大小均为2x2，采用小滤波器可以增加模型的深度，更好地捕获空间特征。

VGG16网络模型的训练采用Malimg数据集，分类任务采用1000个类别，模型训练时间较长，可以达到比较满意的分类精度。VGG16网络模型可以用于大规模图像分类和识别任务，有效提高计算机视觉系统的性能，结构如下：

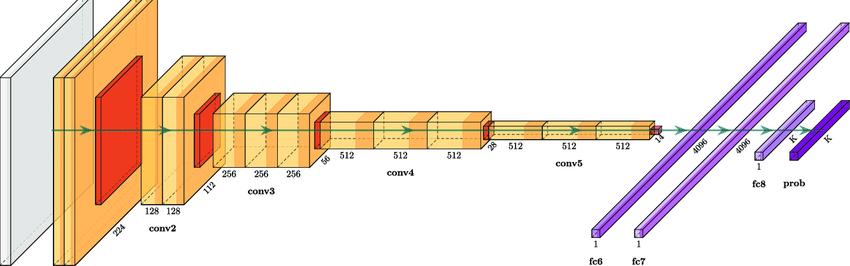


图3-1 VGGNet模型

3.3 功能模块详细设计

4 系统主要界面设计

4.1 首页界面

首页分为3个tab。

第一个tab页面为模型训练页面，提供了数据源和目标模型存储位置，同时提供了数据源的分配比训练集、测试集、验证集的比率和模型的各种状态，如图4-1所示。

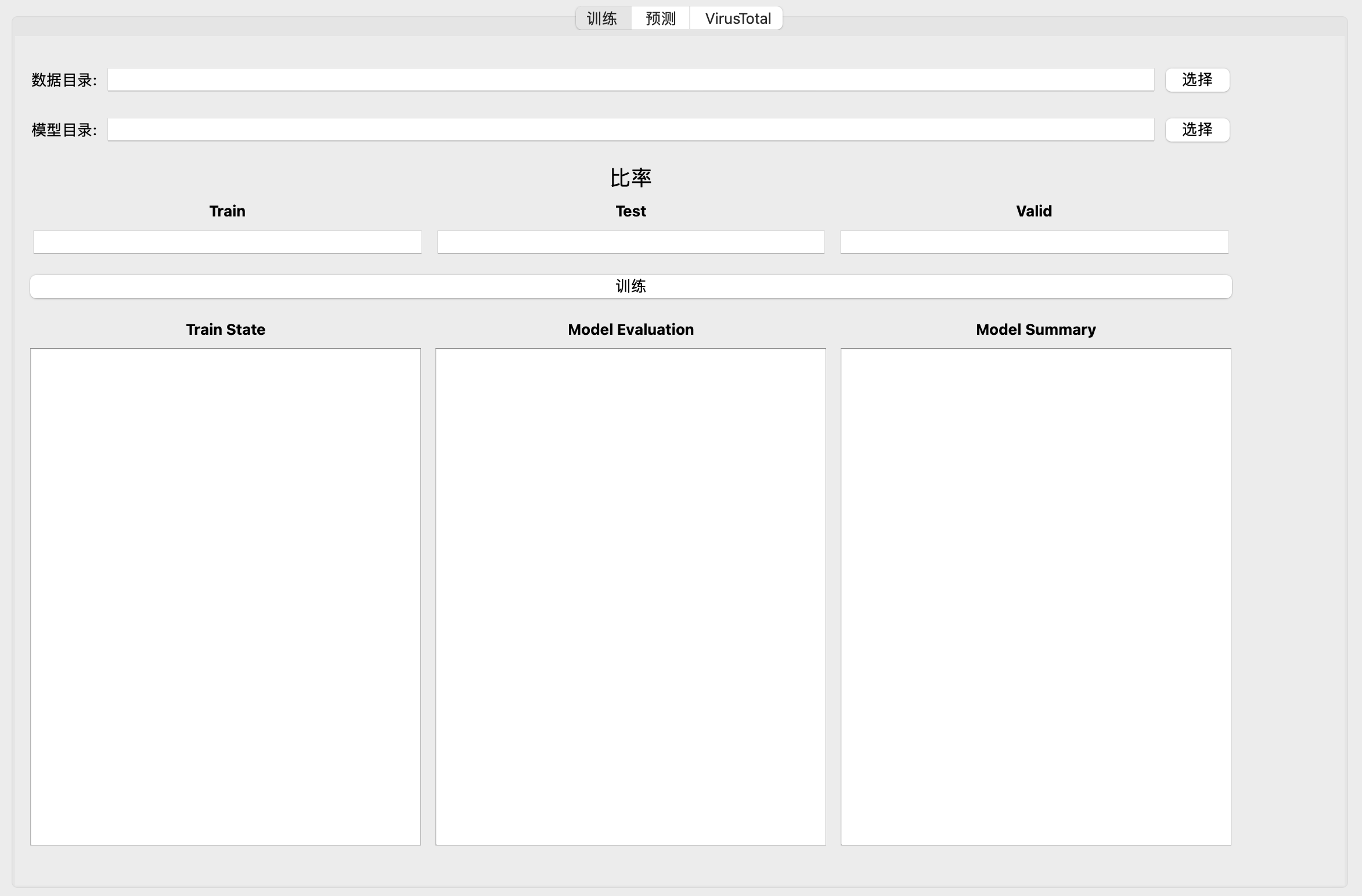


图4-1 模型训练界面

第二个tab页面为预测页面，提供可以文件上传的界面，模型选择和检测报告。检测报告包括程序检测结果，如果检测出恶意程序则进行静态分析，输出部分重要数据到页面，如图4-2。

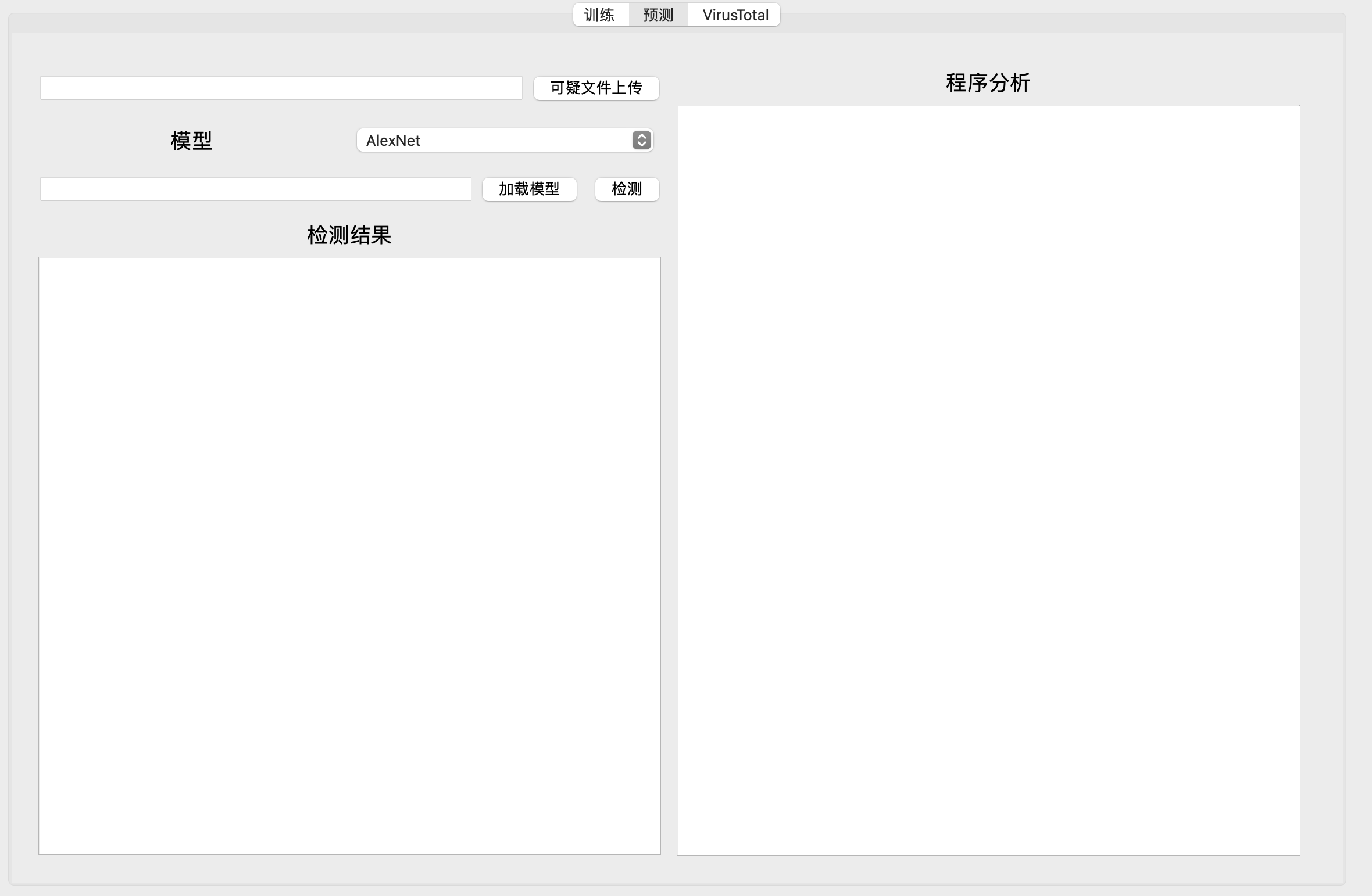


图4-2 模型训练界面

# 5 运行效果

## 主程序

## 性能测试

7 结束语

整个系统的开发实现了用户闯关、用户练习、错题重做、在线测试、下载试卷等功能。本系统主要采用的软件工程的研究方法，具体步骤包括：系统的总体设计、详细设计、编写代码、单元测试的方法。系统基于B/S模式，采用ASP.NET为编程语言，后台数据库采用SQL Server进行后台数据库的设计和维护，使用html,css,vs2017等实现各个界面的设计。本系统根据不同类型的题目针对练习，专攻薄弱板块；根据小朋友的做题情况逐步提升难度，从中提升自身的能力；反复重做错题，线上线下同时操作，提高正确率。本系统提高了小学生的学习效率，因此具有一定的实用价值。因为时间和设计经验不足的原因，这个系统还存在一些不完善的地方，但是这个系统的开发让我对软件开发的流程有了深刻的认识，相信在将来的日子中，我会设计出更加完善，更具实用价值的系统。

参考文献：

[1] 陈伟,查道贵,罗颖. 基于Blackboard的ASP.NET课程翻转课堂教学模式[J]. 内江师范学院学报,2018,33(02):100-105.

[2] 韩道军,贾培艳. 新建构主义在ASP.NET MVC Web程序设计课程中的应用[J]. 软件导刊,2018,17(02):224-226.

[3] 韩道军,贾培艳. 新建构主义在ASP.NET MVC Web程序设计课程中的应用[J]. 软件导刊,2018,17(02):224-226.

[4]佘东.ASP.NET精品资源共享课程的设计与实现[J].黑河学院学报,2017,8(12):106-107.

[5] 于雷. ASP.NET中页面之间值的传递方法探讨[J]. 黑龙江科学,2017,8(21):116-117.

[6] 潘天恒,周方. 《ASP.NET网站设计》课程教学改革研究[J]. 软件导刊(教育技术),2018,17(01):68-69.

[7] 郭海智,郭亮. 基于项目实践的ASP.NET课程教学改革探索[J]. 信息记录材料,2018,19(02):247-248.

[8] 支易. 基于ASP.NET网站的安全性研究与实现[J]. 现代信息科技,2017,1(06):100-101+104.