我们的灵感来自于一个对恶意程序进行分类的论文，论文中通过将程序反汇编生成汇编指令，将程序的汇编指令进行one-hot编码，并且将编码限制在0-255之间。根据汇编指令对应的编号填充在图片中，形成一张灰度图，然后将灰度图作为样本进行训练和预测。

我们的方法在上述的方法中做了一下改变，因为灰度图的空间有限，所以我们使用了彩色图作为存储样本的载体，从样本中选择了3个特征，并分别编码为0-255的范围中，生成彩色的图片。

CNN是一种深度学习模型，并且并且已经取得很好的效果，尤其是在图像分类领域。 CNN由多层神经网络以及其他深度学习方法组成，并具有称为卷积层的特定结构。 如图2所示，卷积层使用小滤波器对输入图执行卷积滤波。卷积层限制了在计算一个输出元素时使用的输入元素的数量，而所有输入元素都在完全连接的层中使用。 除称为稀疏连通性的技术外，卷积层还实现了参数共享，该技术是将相同的小型滤镜应用于输入的所有元素，以使CNN能够有效地提取图像中的局部特征。

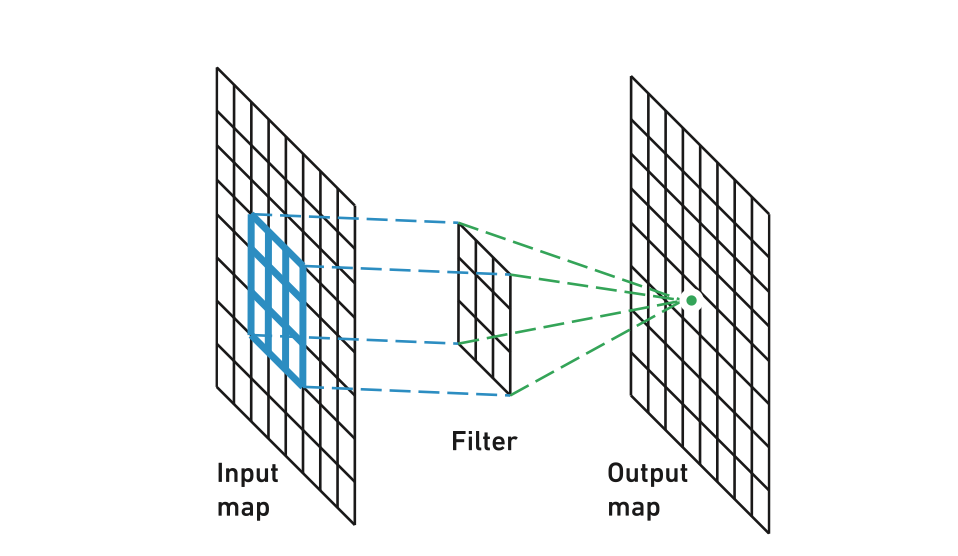


图1 卷积网络结构

我们训练图片分类的模型是一个4层CNN加4个池化层、两个全连接层

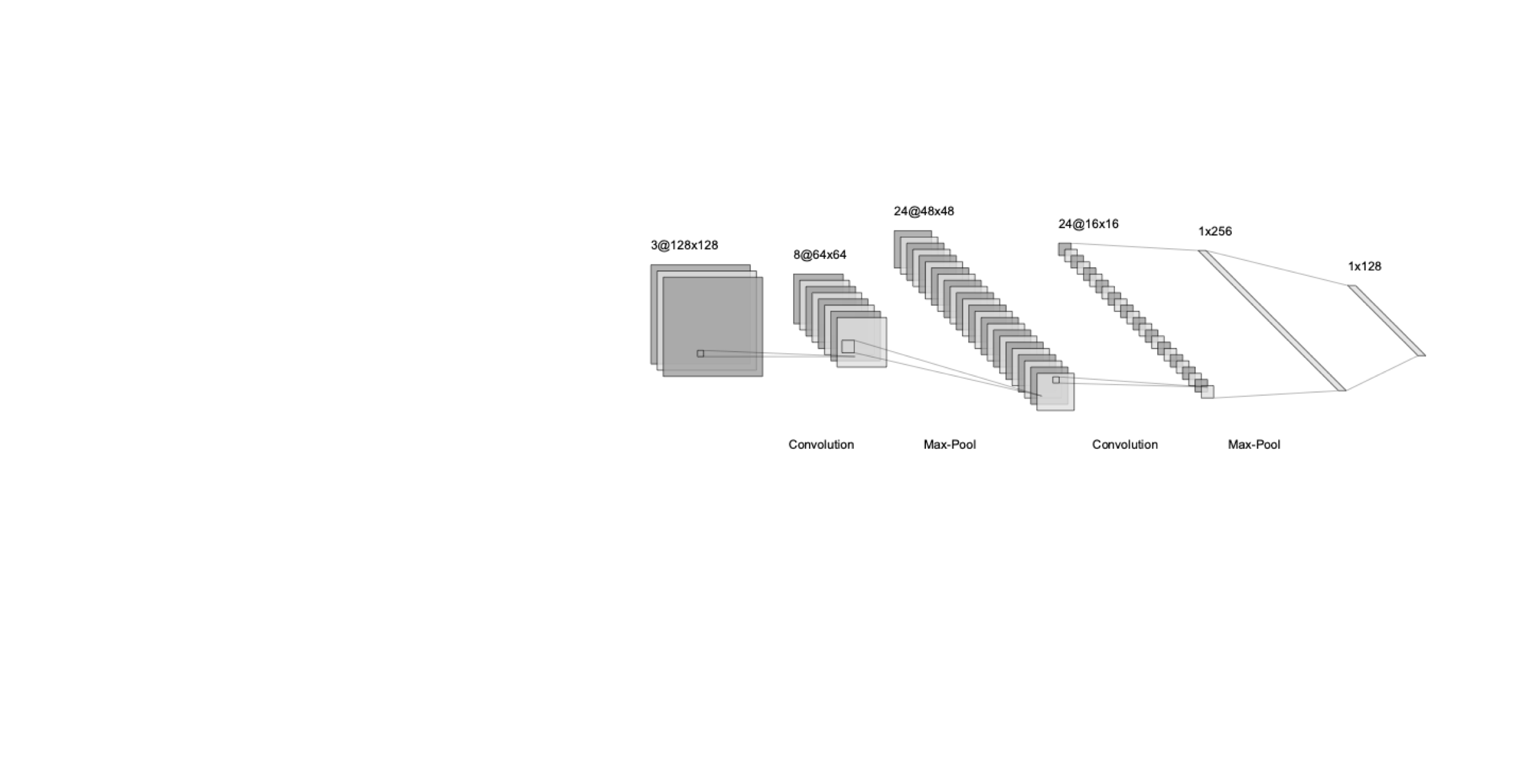
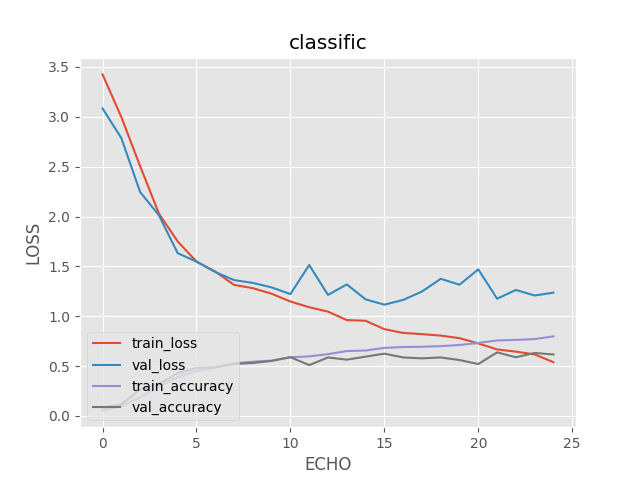


图2 部分CNN网络的结构

在验证时训练集与验证集的比例为4:1，我们的准确率能够稳定在60以上，最好一个达到63%



我们还进行了一下对比实验，在选择图片的边长时，我们通过多组对比从而确定边长为30时取得最好的结果。

|  |  |  |
| --- | --- | --- |
| width=16 | width=25 | width=50 |

图3 不同边长的准确率

由于数据较少，我们用到了数据增强的方法，希望能够将100个样本组合在一起，让后每隔一个间隔就去一段样本的方法来增强数据。这样的方法确实能够增强准确率，但是由于图片太小，会造成前一个图片与后一个图片有很大的重复，所以这个方法被我们放弃。我们在训练时对输入的图片进行了小幅度的旋转、平移和镜像变化来提高样本量。

我们在选择模型时对模型的架构也进行了比较发现4层CNN网络比2层CNN网络的机构更好，4层CNN网络可以达到63%，而2层CNN网络只能达到53%的准确率。当CNN网络的机构更复杂时会让准确率不稳定。

|  |  |
| --- | --- |
| 2层CNN | 6层CNN |

图4 不同网络结构下的准确率

不足：

该试验中我们最大的不足就是数据集的分类种类很多，但是每一类的样本都很少，这也是导致准确度低的主要原因之一。数据量小导致每个图片只有30\*30的大小，这样的图片直接进行分类效果并不是很好

我们在提取特征的过程中，对于特征使用了比较原始的方法手动进行提取，这样的做法可能导致其中有一些重要的特征没有被捕捉到。

对于数据增强的手段也有很多不足，直接对图片进行平移和反转对于数据的增强效果并不明显。

未来的发展

现在将self-attention机制应用于CNN网络中进行图片的分类能够提取出那个样本中值得关注的像素点，这样做法对于图片分类的精确度又很大的提升。我们可以将这个机制利用到我们的这个分类模型中，提高模型的准确率。

需要设计新的数据增强算法，可以对同一个网站的样本进行组合和切分，这样也能够生成更多的样本。

使用迁移学习的方法，试验模型的基本策略就是使用预训练模型的权重作为特征提取器，将预训练的权重进行冻结，只训练全连接层。这样也能够解决训练样本量小的问题。