猫狗大战

项目背景

图像识别技术是信息时代的一门重要技术，其目的是为了让计算机替代人类入处理大量的物理信息。比如说让计算机来从人群中识别出罪犯，对癌症的准确侦测等。由于理论基础的积累已经硬件性能的大发展，在2012年后，机器学习迎来了爆发期，因此，对于通过机器学习进行图像识别，也成为了可能。**2012年，Hinton课题组为了证明深度学习的潜力，首次参加ImageNet图像识别比赛，其通过构建的CNN网络AlexNet一举夺得冠军，且碾压第二名（SVM方法）的分类性能。也正是由于该比赛，CNN吸引到了众多研究者的注意[1]。**

问题描述

猫狗大战，该题目，即通过针对一定数量的已经加过标签的图片的训练，验证，测试，得出可靠的特征模型，再根据模型识别新的图片。期望结果为为测试集的每一张图片，识别出是狗的图片的概率。

数据或输入

项目中，所使用的数据是来自于kaggle的图片数据：https://www.kaggle.com/c/dogs-vs-cats-redux-kernels-edition/data，其数据形式为jpg格式的图片，分为训练用数据和测试用数据。训练图片文件名称已经写明是cat或者dog，其中训练图片有25000张，包含猫的图片12500, 狗的图片12500张。测试图片文件名并未写明cat或者dog，测试图片有12500张。可以看出，训练数据猫与狗的图片数量是一样的，而且猫的图片排列在前；而训练的数据并未通过文件名来告知是猫或者狗。

除此之外，无论是训练图片还是测试图片，虽然都是jpg格式，但是没有一个统一的图片尺寸，好在图片之间并没有特别大的差距，宽高都保持在500像素以内；按照内容来看，多数图片中的场景，都是一只猫或者狗，有少数是有多只猫或者狗，并没有猫和狗同时出现的情况。

鉴于以上情况，需要对图片做以下预处理：

1. resize操作，把图片调整到一个统一的合适尺寸；
2. 打乱训练集数据，破坏原有的猫图片都在前边，狗图片都在后边的情况；
3. 将打乱后的数据，分为训练集和验证集。

解决方法描述

1. 先获取数据，从kaggle上下载两个图片包，并解压在项目目录的data目录下
2. 读取数据，将数据读入程序，并识别标签，cat为0，dog为1。
3. 拆分数据，分成训练集和验证集。
4. 用训练集训练，生成模型
5. 用验证集验证模型有效性
6. 使用测试数据，测试从未见过的新数据

评估标准

kaggle提供的模型的评分公式为：LogLoss=−1n∑i=1n[yilog(ŷ i)+(1−yi)log(1−ŷi)]

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

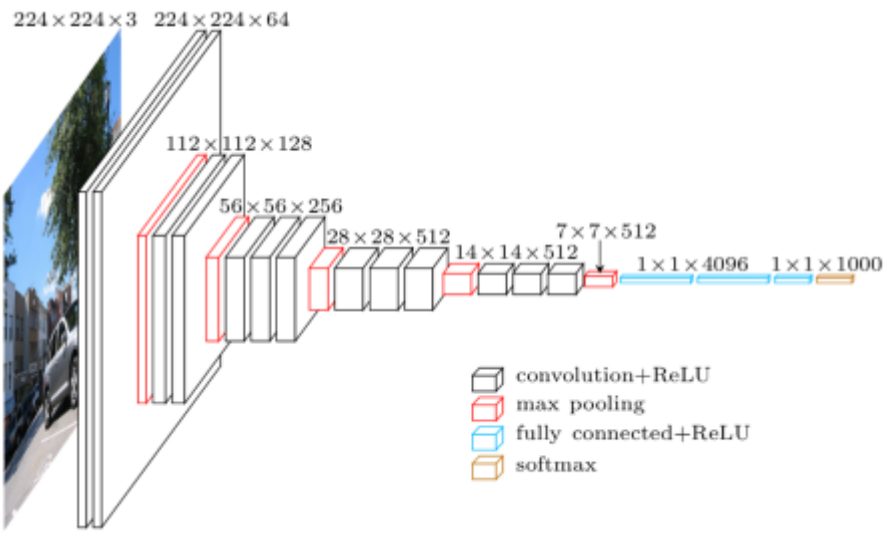
其中参数含义如下：

- n 测试数据大小
- $\hat{y}_i$ 预测第i张图片为狗的概率
- y_i 第i张图片若为狗则值为1，否则为0
- $log()$ 自然对数函数

在验证集上通过此函数，得出模型评估分数，该分数 约小越好。当模型分数足够优秀，则可以在测试集上使用此模型，生成评估概率的csv文件，提交给kaggle。

基准模型

卷积神经网络图像识别基准模型VGGNet，其流程图如下：



从图中，我们可以看到对图片的处理过程，标准化，最大池化，完全链接，softMax激活。

特别注意的是，在构建keras中的VGG16网络时候，需要特别指定一个参数classes为1。

项目设计

1. 获取训练和测试的图片，并解压到项目目录的train和test文件夹下。
2. 图片预处理：将图片数据进行标准化处理，利用图片的文件名包含dog或者cat字符，进行one_hot编码。将预处理过的数据进行保存。避免每次都重复做预处理动作。
3. 构建神经网络：实现卷积，最大池化，扁平化，全连接层，最后输出。
4. 执行训练，生成模型，看准确率，对参数进行调优
5. 执行验证，看能否对验证数据进行有效分类
6. 若验证通过，则执行测试

参考文献

[1] CSDN博客-遍地流金 https://blog.csdn.net/u012177034/article/details/52252851