机器学习工程师纳米学位 – 猫狗大战

-开题报告

-张海鹰
-v1 2018.07.02
-v2 2018.07.06

项目背景

猫狗大战是 kaggle 平台上的一个比赛项目，最终的要求是提供一个模型来识别图片中的对象是猫还是狗，所以抽象来看这是一个图像识别的问题。

图像识别是人工智能的一个重要领域，目前发展十分迅速，应用范围相当广泛，手写数字识别、邮政编码识别、汽车牌号识别、汉字识别、条形码识别，以及如人脸、指纹、虹膜识别等已经在人类日常生活中广泛应用，对经济、军事、文化及人们的日常生活产生重大影响。

支撑其应用实现的重要技术就是深度学习，深度学习被誉为通往人工智能的必经之路，在 2016 年 3 月 Google DeepMind 研发的 AlphaGo 4:1 战胜了世界冠军李世石后，深度学习已经成为现今最为火热的人工智能技术。

问题描述

项目要求的最终输出是辨别图片是猫还是狗，因此属于机器学习领域的分类问题，由于没有要求对狗和猫的品种再次细分，因此是二分类。具体的量化方法就是让模型输出该图片是猫和狗的概率，取值区间为 0-1，1 代表 100%是狗，0 代表 0%是狗，那么自然 100%是猫。具体实施方法可以在训练数据集上使用深度学习来进行模型训练，让模型通过给定的训练数据，不断学习到如何识别猫狗的特征。

数据输入

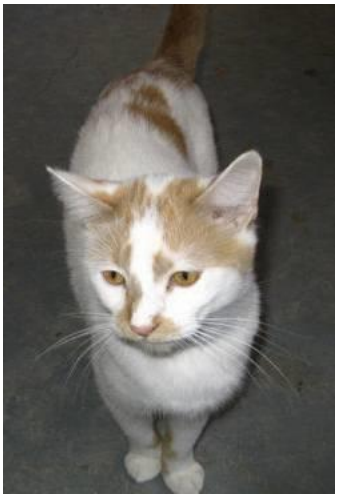
全部数据都由 kaggle 平台提供，分为了训练集和测试集；
训练集共有图片 25000 张，其中猫 12500 张，狗 12500 张；各占一半；
测试集共有图片 12500 张；
训练集中有的图片特别模糊，如下：



cat.4821　　cat.6402　cat.2433

还有的图片的大小存在不一致，如下

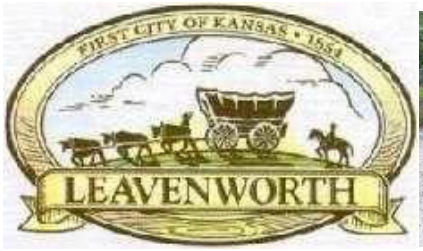

原图大小 299 x 225


原图大小 249 x 368


原图大小 500 x 374

在训练的时候需要 resize;

还有一些明显异常图片（图片中没有猫狗），如下:



在训练时需要删除

我会在训练集中再次划分训练和验证数据集，用来训练和验证模型，最后在测试集上测试模型， 其中要注意的是不能对模型泄露验证集数据。

解决方案

深度学习在图像识别中目前最有代表性的就是卷积神经网络。

普通神经网络处理图像识别问题中，输入层的每一个神经元可能代表一个像素的灰度值。但这种神经网络用于图像识别有几个问题，一是没有考虑图像的空间结构，识别性能会受到限制；二是每相邻两层的神经元都是全连接，参数太多，训练速度受到限制。

而卷积神经网络就可以解决这些问题。卷积神经网络使用了针对图像识别的特殊结构，可以快速训练。因为速度快，使得采用多层神经网络变得容易，而多层结构在识别准确率上又很大优势。

因此本项目会采用卷积神经网络

评估标准

采用 log loss 作为模型评估标准;

其定义如下:

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

评估标准说明:

n 是图片数量;

y_i 是类别标签，1 表示狗，0 表示猫;

$\hat{y_i}$是预测为狗的概率;

Log() 表示自然对数;

基准模型

准备使用 ResNet50, InceptionV3, Xception 作为基准模型，期望能在项目中自己搭建的模型成绩优于这三个单独的基准模型。

另外本项目要求是进入 kaggle 比赛前 10%的水平，该竞赛有 1314 支队伍，即最终结果需要排名在 131 位之前，即成绩要优于 131 名的成绩 0.06127.

项目设计

Step1：数据预处理

- 简单查看图片，思考是否要处理异常值，以及如何处理
- 为了配合 keras 的 ImageDataGenerator，将猫狗图片分别存储到 cat 和 dog 文件夹
- 在训练集中划分出 20%保留为验证集

Step2：模型探索

- 尝试一下 ResNet50, InceptionV3, Xception 这些网络的单独辨识情况

Step3 模型搭建

- 尝试综合利用三种模型的特征向量，然后训练出自己的模型

Step3：模型训练

- 使用 aws 上的 p3.2xlarge 进行模型训练，尝试不同的 epochs

Step4：模型调参

- 尝试调整 learn rate， dropout 等参数;
  learn rate 太大，会出现随着训练代数增加 loss 并不减小，以及 loss 值上下振荡；太小会出现随着训练代数增加 loss 基本不变化，因为无法快速找到好的下降方向；可以先尝试 0.1，0.01，0.001, 0.0001 这些 rate，观察一下 loss 下降的情况，再决定使用哪一个 learn rate。
  如果过拟合, 则添加 dropout, 然后尝试用不同比例的 dropout 进行对比效果。

Step5：模型评估

- 采用 log loss 作为模型评估标准

Step6：可视化

- 把模型训练过程中每一个 epoch 的 acc， loss 用曲线图像展现，这样可以直观的看见模型训练中的进展情况。
- 尝试使用 CAM（类激活图）动态可视化模型的学习过程。