

AI算法工程师

毕业项目开题报告

沈捷 2019年3月28日

背景

2013年，kaggle举办了一项趣味性赛事“猫狗大战”，即猫狗图像分类。许多学者及行业人士参与其中，供献了大量的解决方案，机器学习领域的局势由之改变¹。现今图像分类在许多行业已展开深入研究，如医学上热门的癌组织与良性组织的鉴别，甚至有学者尝试利用深度学习方法鉴别出男女大脑MRI图像的差异²，都是类似的问题。但医学领域常由于患者隐私保护政策，难以获得大量的训练数据，所以为了更好地学习、研究新的方法，也有学者会利用猫狗大战数据进行探索尝试和验证³，同理可以延伸至许多应用领域。

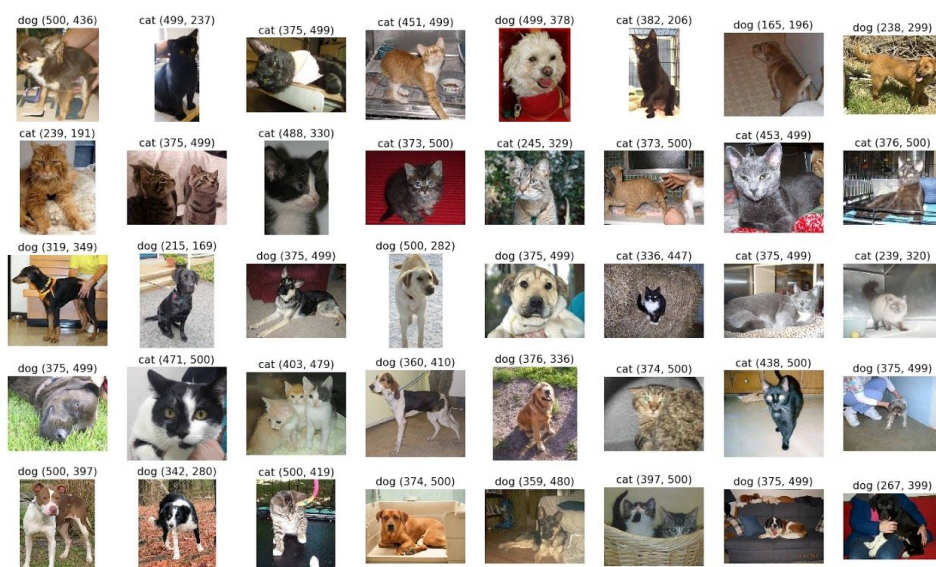
问题描述

区分图片是猫还是狗的照片，显然是一个二分类问题，通过输入图像特征，获得一个概率，通过概率来判断属于哪一个类。

数据输入

该数据集已切分为训练集和测试集。训练集包含25000张图片，猫和狗图片比例为1:1，可以从文件名（标签.编号.jpg）中获得标签；测试集有12500张图片，文件名只有编号。

随机抽取样本，简单探查这些图片：



可见图片长或宽多在300~500像素之间，目标主体基本清晰。只不过有的主体占图片的比例较小，有的图片上有两个或以上的目标，不知是否会影响模型训练，且尚不知是否有分类错误的训练样本。

对于错误分类样本的检测，决定采用预训练的ResNet50模型进行初始预测，根据ImageNet1000的标签⁴，其中有118个狗的品种和7个猫的品种，可以做参考，找到那些预测与标签不符的图片，再做人工确认。如下为ResNet50模型中找到的部分错误标签的图：



可见，这些图中有的目标主体太小，有些有遮挡，有些模糊，有些曝光度太高以至特征消失，甚至一张图片上同时有猫和狗或两者都不是，这都是对模型训练的干扰，这部分图片将被剔除。

其余训练集图片会经过随机旋转、缩放、裁剪等预处理，使模型能有更好的泛化能力。这些图像将根据预训练网络的要求，统一为 224×224 大小，并分解为RGB三个颜色通道的色值，再经过标准化处理，即均值为[0.485, 0.456, 0.406]、标准差为 [0.229, 0.224, 0.225]，作为模型的输入。

解决方案

处理好输入数据后，将采用一个预训练的神经网络进行迁移学习，预计将采用ResNet50模型。神经网络的输入节点应为 $224 \times 224 = 50176$ 大小，输出节点数量为1，采用sigmoid激活函数将图像对某一类的线性预测值映射到(0,1)之间，即为概率（狗=1，猫=0）。sigmoid函数公式：

$$S(x) = \frac{1}{1+e^{-x}}$$

训练模型的过程中，将调整epoch、隐藏层节点、学习速率等超参数，以尽量高准确率，降低对数损失（log loss）。

基准模型

本例没有基准模型。直接将迁移学习训练好的模型在测试集上进行预测，提交到kaggle上进行比较判断。模型的训练目标即是达到kaggle的前10%，即LogLoss < 0.05603。

评估指标

模型将采用对数损失函数（log loss）进行评价。对数损失函数需要输入每个分类的预测概率与标签，对错误的分类进行惩罚，从而对准确率（Accuracy）进行量化⁵。损失越少，准确率越高。对数损失函数公式¹：

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

其中：

- n 是样本数量
- \hat{y}_i 是图像 i 为狗的预测概率值
- y_i 是图像 i 的标签， $y_i = 1$ 是狗， $y_i = 0$ 是猫
- $\log()$ 是以自然数 e 为底的对数函数

此外，预测结果将上传到Kaggle进行排名对比，目标是要达到Kaggle排名的前10%。

项目设计

综上，本项目是一个图像的二分类问题。数据集已经被划分为训练集25000张图片和测试集12500张图片。对于训练集图像，首先要进行随机翻转、缩放、裁剪等预处理，并根据预训练模型统一成 224×224 大小，再转化为RGB三个通道的Tensor，与预训练模型的标准参数进行匹配，即均值为 [0.485, 0.456, 0.406]、标准差为 [0.229, 0.224, 0.225]，作为模型的输入。

将采用ResNet50为预训练模型框架进行迁移学习，输入节点根据图像大小设为 224×224 即50176，输出节点为1，用Sigmoid函数输出图像的预测概率（狗=1，猫=0），即预测为狗的概率。当概率小于0.5时，可以认为预测倾向于猫。

在训练过程中，将调整epoch、隐藏层数、隐藏节点数、学习速率、dropout等超参数，尽量提高准确率，降低对数损失。可对logloss的变化绘制折线图，观察logloss收敛情况。

在测试集上进行预测之后，提交到kaggle上进行评价对比。kaggle则采用对数损失函数(log loss) 进行模型准确率的评价。本模型的目标是达到Kaggle排名的前10%，即LogLoss < 0.05603。

参考文献

1. Kaggle. Dogs vs. Cats Redux: Kernels Edition. *Kaggle* Available at:
<https://kaggle.com/c/dogs-vs-cats-redux-kernels-edition>. (Accessed: 5th April 2019)
2. Xin, J., Zhang, Y., Tang, Y. & Yang, Y. Brain Differences Between Men and Women: Evidence From Deep Learning. *Front. Neurosci.* **13**, 185 (2019).
3. Perez, L. & Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. (2017).
4. 262588213843476. text: imagenet 1000 class idx to human readable labels (Fox, E., & Guestrin, C. (n.d.). Coursera Machine Learning Specialization.). *Gist* Available at:
<https://gist.github.com/yrevar/942d3a0ac09ec9e5eb3a>. (Accessed: 16th April 2019)
5. klchang. 对数损失函数(Logarithmic Loss Function)的原理和 Python 实现. 博客园 Available at:
<https://www.cnblogs.com/klchang/p/9217551.html>. (Accessed: 5th April 2019)