机器学习（进阶）纳米学位

毕业项目开题报告

一、项目背景

“猫狗大战”是 Kaggle 上最为著名的娱乐型竞赛项目之一，至今已经举办过多次。本项目是 2017 年 3 月举办的“Dogs vs. Cats Redux: Kernels Edition”。该项目的目标是在测试数据集中分辨出猫和狗的图片。

这是一个经典的卷积神经网络（Convolutional Neural Network，CNN）图片分类项目。CNN 是深度学习领域的一种重要方法。它运用卷积运算大幅度减少训练神经网络所需的参数数量，使得训练更深的神经网络成为可能，从而提高神经网络的性能。图片分类正是 CNN 的主要应用领域之一。本项目就将运用 CNN 技术进行图片内容分类。

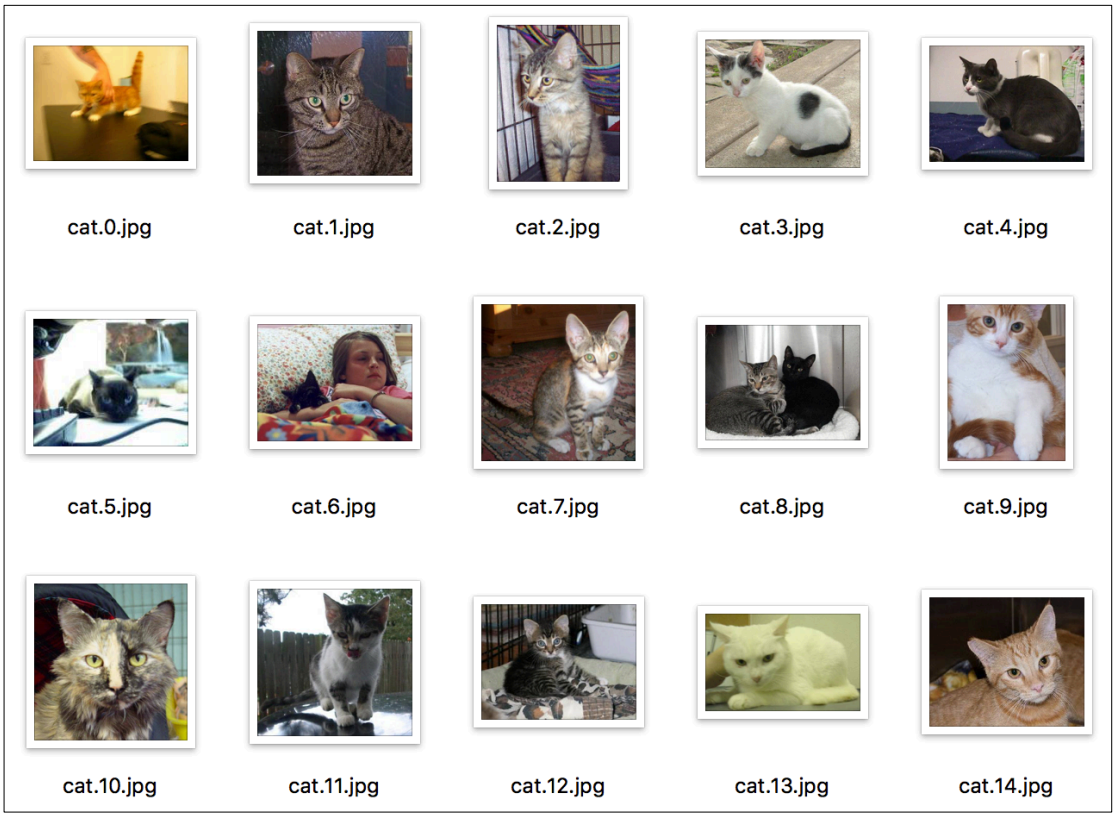
二、问题描述

该项目要求对一组未标记的图片进行“猫”和“狗”的分类，以“图片中是狗的几率”（狗是 1，猫是 0）来描述分类结果。

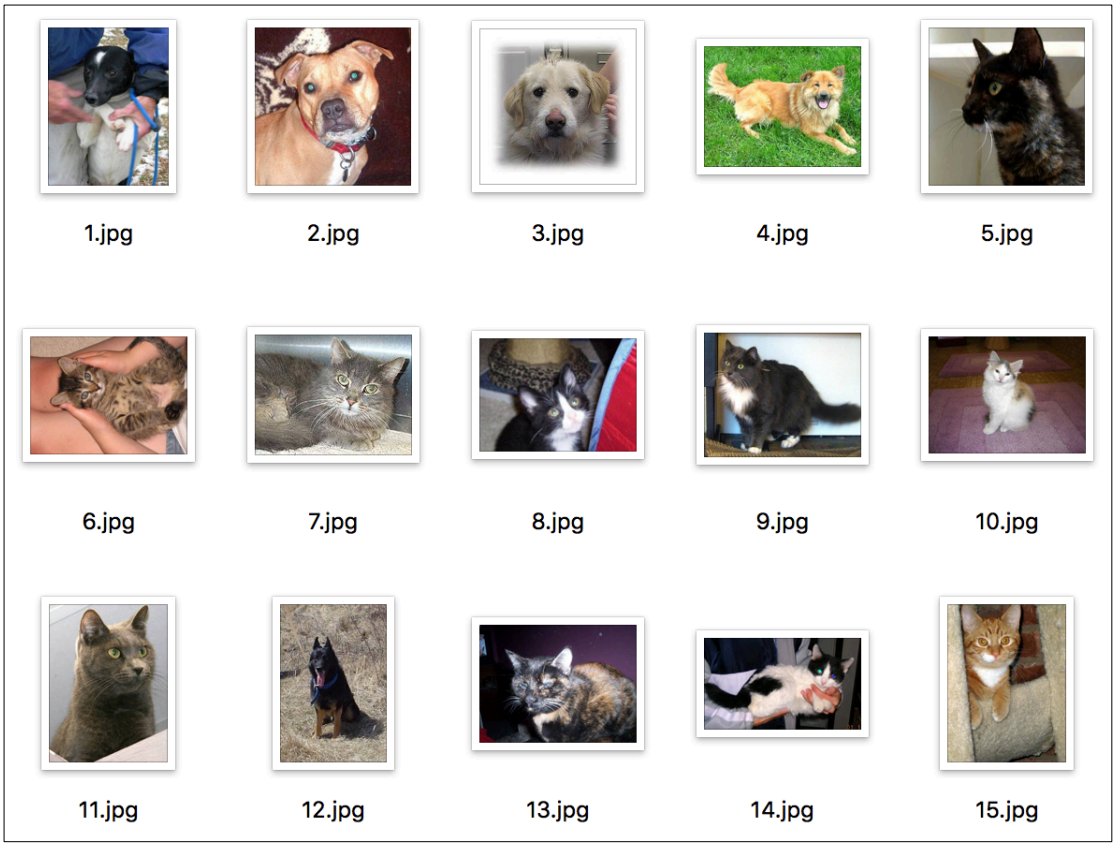
为解决这个问题，本项目将训练一个用于图片分类的 CNN 模型，对图片进行特征提取和识别，从而分辨出图片中的猫和狗。

三、数据或输入

该项目的数据来自 Kaggle。完整的数据集包括 25,000 张已标记的图片——其中猫和狗各 12,500 张，和 12,500 张未标记的测试图片。所有图片均来自日常拍摄的猫或狗的照片，其中包括一些像素质量较差或经过特殊处理的照片，还有一些照片中有不止一只动物（但猫和狗不会出现在同一张照片中），或有人等干扰性内容。



训练集中的图片用文件名进行了标记。猫的图片文件名为“cat.<number>.jpg”，狗的图片文件名为“dog.<number>.jpg”。



测试集中的图片以数字序列命名，没有分类标记。

可以看到数据集中的图片有着不同的分辨率和长宽比，在输入模型前需要进行预处理，将其调整为统一的尺寸。Keras 库内置了图片处理函数，可用于完成这一工作。具体预处理方法将根据模型需求确定。

四、解决方法

本项目将训练一个 CNN 模型解决这个图片分类问题，并使用迁移学习的方法提高模型效率和简化训练工作量。具体来说，将使用 Keras 的内置模型，及成熟的 ImageNet 数据集权重作为预训练，在此基础上加入自己的模型输出最终分类结果。

ImageNet 是目前最流行和最庞大的公开图片识别数据集之一，

已经有成熟的训练权重数据。使用该数据进行迁移学习，则不必从头开始训练，极大地简化了训练难度和时间，是本项目合适的解决方法。

五、评估标准

Dogs vs. Cats Redux: Kernels Edition 竞赛使用对数损失（LogLoss）作为评估标准。本项目采用同样的标准。损失函数$\mathrm{L}(y,\hat{y})$用于评价预测值$\hat{y}$与真实值y之间的差异程度。损失函数的值是梯度下降法的核心，模型训练的过程在数学上就是（寻找参数）令损失函数最小化的过程。

对数损失函数是最常见的用于分类问题的损失函数。其的公式为:

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

对数损失函数指标是一个连续值，与正确率（正确样本数/总样本数）指标相比，其对预测能力的评价更加细致，对判断错误的样本会给予更严重的惩罚，因此对模型的要求也更高。

目前 Kaggle 上公开的竞赛结果排名中，TOP10%的 LogLoss 值在 0.05629 以下。本项目的目标就是模型的 LogLoss 值低于 0.05629。

六、基准模型

本项目采用成熟的 ResNet50、InceptionV3 和 Xception 三个模型作为基准模型。先分别采用这三个模型进行迁移学习训练，并以最优的结果作为评分基准。

七、项目设计

首先，手动对数据集进行浏览，理解和明确数据集的数量、内容和特点。

第二，确定评估指标和基准模型。在本项目中，使用对数损失作为评估指标，目标是 Kaggle 公开排名的 TOP100。选择 ResNet50、InceptionV3、Xception 作为基准模型。

第三，根据模型要求，确定需要对数据进行的预处理方法。由于选择的三个模型都是 Keras 内置模型，因此将采用 Keras 提供的预处理方法，以简化工作量。

第四，分别使用三个模型进行训练和预测，并分别将三次预测结果上传到 Kaggle 进行评分，选择最优者作为基准评分。因为采用了迁移学习的方法，训练将分为两步进行：一，使用 Keras 内置模型并去掉最后的全连接层，对输入数据进行预训练；二，对预训练的数据进行自定义模型的训练。采用这种方法，预训练只进行一代，自定义模型进行多代训练，极大地节约了计算量和时间。

第五，将三个模型的预训练结果综合起来，进行自定义模型的训练。由于结合了三个成熟模型的训练结果，预期模型性能会有所提高，预测结果更加准确。

第六，在第五步的基础上，考虑采用扩充数据集等手段，进一步提高模型性能。