猫狗大战

开题报告

背景介绍

猫狗大战（Dogs vs Cats）是源自 kaggle 上 2013 年的一个竞赛项目。这个项目主要是为了使用机器学习算法进行图片分类，尤其是深度学习。项目中提供了一个数据集，包括 25000 个已标定的数据和 12500 个未标定数据。通过这 25000 个已标定数据来构建相应的预测算法模型，并使用这个算法模型对 12500 个未标定的数据进行预测，并将预测的结果提交到 kaggle 上，以得到一个综合评分，以评价预测算法模型的优良水平。

问题描述

猫狗大战提供的图片源自真实拍摄，分辨率差异较大。图片中猫和狗的颜色丰富，类型多样，姿态迥异，同时还有复杂的背景，极大的增加了识别的难度。究其本质，本项目属于一个监督学习二分类问题，这里将使用基于深度学习的卷积神经网络（CNN）来构建预测算法模型，通过 25000 个已标定数据（train data）来训练神经网络，然后使用训练好的神经网络模型来预测未标注的数据（test data），来完成预测准确度排名进入到 kaggle 上猫狗大战 Public Leaderboard 排名前 10%的项目目标。

数据集

从 kaggle 的 dogs_vs_cats 的数据集中可以发现，数据集的 train 和 test 是进行了分类的，但是 train 数据集中的验证集是没有单独分类的，需要在训练模型的时候自行生成。另外，train 数据集的数据标注也是包含在文件名中的，需要进行特征提取。

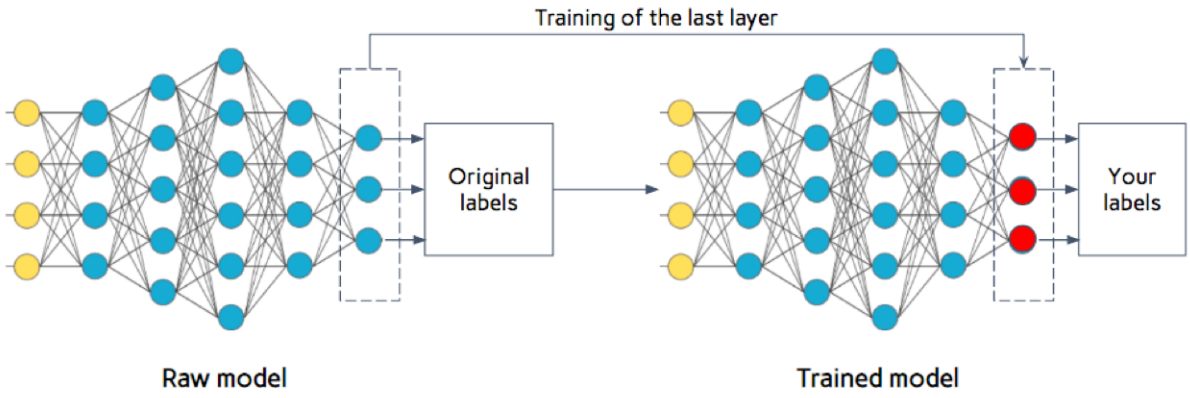
文件	标注（cat=0.0，dog=1.0）
datas/train/cat.6938.jpg	0.0
datas/train/dog.11432.jpg	1.0
datas/train/cat.433.jpg	0.0
datas/train/cat.11305.jpg	0.0

为了更好的对数据有直观的了解，将内存中的图片可视化出来。如下：

通过上图可以看出，训练所用的数据，尽管有各种复杂的背景，尽管猫和狗的姿势和形态也比较多样性，但整体来看图片都比较清晰，算是比较理想的数据集。

解决方案

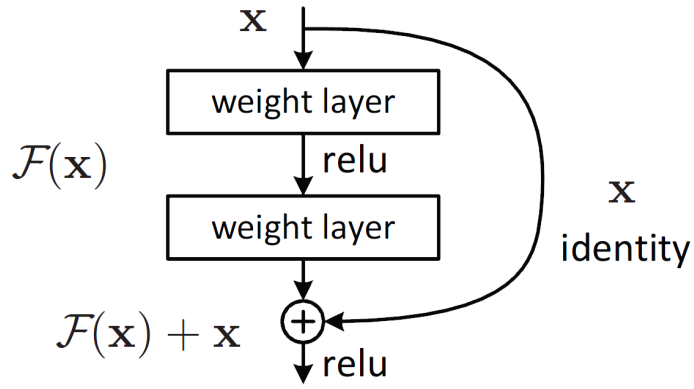
为了完成项目的分类需求，最常用的就是使用卷积神经网络（CNN）来进行分类，较为著名的 CNN 网络，有 AlexNet、VGGNet、ResNet、Xception 和 Inception 等。这里我们使用 Resnet50 作为我们的神经网络，基本上可以得到不错的评分结果。鉴于项目的时间要求和算力方面考虑，这里通过迁移学习来达成项目的目标。



本项目中使用的 dogs_vs_cats 与 ImageNet 整体上来看，相似度还是比较高的，且 dogs_vs_cats 只提供了 25000 个训练数据（含验证数据）和 12500 个测试数据，样本偏小，非常适合迁移学习。

基准模型

ResNet（Residual Neural Network）由微软研究院的 Kaiming He 等四名华人提出，通过使用 ResNet Unit 成功训练出了 152 层的神经网络，并在 ILSVRC2015 比赛中取得冠军，在 top5 上的错误率为 3.75%，同时参数量比 VGGNet 低，效果非常突出。ResNet 网络上由若干个 ResidualBlock 和 IdentityBlock 构成的，比较常用的分别是 Resnet50、Resnet101 和 Resnet152。



借助于 Resnet50 的模型，通过迁移学习，只需要构建最后的完全连接层的模型。模型由一个 GAP 层，加一个 Dropout 层，最后通过 Dense 层输出预测结果。

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
global_average_pooling2d_1 ( (None, 2048)              0
_________________________________________________________________
dropout_1 (Dropout)          (None, 2048)              0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 2049
=================================================================
Total params: 2,049.0
Trainable params: 2,049.0
Non-trainable params: 0.0
_________________________________________________________________
```

评价指标

在机器学习领域，通常会对算法模型设定损失函数作为评价指标，以便更好的评价算法模型的优良。常用的分类问题，都会使用交叉熵作为损失函数，猫狗大战属于二分类图像识别问题，需要使用了二分类交叉熵（binary_crossentropy）作为损失函数。具体公式如下：

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

项目设计

数据处理

通常数据集中会包含一些错误的数据，会极大的影响模型的训练，由于项目本身是用来训练猫和狗的分类，所以对于非猫非狗的图片，就是错误的数据，需要剔除掉。下面是通过算法筛选出来的错误图片。

需要对数据进行特征分析，这里考虑到图片数据难以使用常规的数据分析方法，这里对图片的分辨率的情况进行散点分布分析。

提取 **bottleneck** 特征集

将清洗过的数据提取数据标注信息，然后将训练集数据进行分拆，形成新的训练集和验证集数据。使用去掉了完全连接层的 Resnet50 模型进行预测，将得到的数据保存为 bottleneck 特征集（包含训练集数据、验证集数据和测试集数据），同时也要将训练集和验证集的数据标注信息也对应保存起来，以方便后续的数据使用。

训练神经网络及参数调优

搭建新的完全层神经网络，包含一个 GAP 层，一个 Dropout 层和一个 Dense 层（可以根据评价的反馈进行优化）。使用上一步提取的 bottleneck 特征集，对新构建的完全连接层进行神经网络模型训练。这里需要对搭建的模型进行不断的参数调优，直到得到可观的评价结果。

预测测试集并得到评分

使用上一步训练的神经网络模型，对测试集的数据进行预测，并将预测的结果提交到 kaggle 上获取准确度评分结果。为了方便对预测结果做初步的评价，可以通过可视化的方式，将预测结果展现出来。

引用

[1] Sumit Saha, A Comprehensive Guide to Convolutional Neural Networks.

[2] CS231n Convolutional Neural Networks for Visual Recognition.

[3] Jason Yosinski, Jeff Clune, Yoshua Benjio, and Hod Lipson, How transferable are features in deep neural networks.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition.

[5] Jason Brownlee, How to Check-Point Deep Learning Models in Keras.

[6] Aaditya Prakash, One by One [1 x 1] Convolution - counter-intuitively useful.

[7] Francois Chollet, Building powerful image classification models using very little data.