机器学习学位纳米学位

开题报告

黄有维

2018 年 4 月 14 日

猫狗大战

项目背景

猫狗大战项目是基于图片二分类的问题， 1 分类为狗，0 分类为猫。主要涉及机器学习的相关领域，包括计算机视觉的图像处理、卷积神经网络等。

本项目就是利用 kaggle 提供的数据[1]，来进行深度学习算法的建模训练和验证。

随着机器学习领域的快速发展，现在很多大公司都选择在使用深度学习，比如 facebook 的智能标记、亚马逊的推荐算法等。

图 1 各大公司

问题描述

猫狗大战项目实质就是图像的二分类问题，最后的结果不是猫就是狗。

这里的目的就是要让机器能够对输入的图片做出分类。对于我们人类来说，我们能够非常容易分辨出图像的特征从而分辨出猫还是狗，对于计算机看来，眼中的数据只是 0 和 1。因此，在此要解决的问题就是让计算机能够准确地对图片做出 0 和 1 的判断，1 代表狗，0 代表猫，最后输出狗的概率是多少。

要使计算机能够做出分类，就需要构建模型，然后用大量数据对模型进行训练，最后对训练后的模型进行评估，评估通过则问题解决，否则需要继续调参训练模型。

图 1 人类看到的

图二 计算机看到的

输入数据

所使用的数据来自 kaggle 的猫狗大战数据集，分为训练集 train 和测试集 test。训练集总共有 25000 张图片，以 type.num.jpg 的格式命名，都做了分类，狗和猫各占一半，12500 张。测试集总共有 12500 张图片，统一从 1~12500 来命名，没有做猫狗分类。

图三 训练集里猫狗数据

通过观察数据集，其中有一些异常值的，举例如下：

1.图中没有狗却被标记为“狗”

2.被主人盖了风头的猫

这些异常值占的比例很低，没必要再做特殊处理。

这些图片的大小都是不一样的，所以在送入模型之前需要对图像做大小处理，如后面使用的预训练模型为 VGG16，对输入图片(224x224x3)平整化，然后送进模型。

另外，为了防止过拟合，需要从数据中选取验证集。验证集在每次 epoch 后计算的，用于提升模型的 loss 和 accuracy 的，模型不会利用验证集来训练。我们可以从训练集中选取验证集，但要注意的是我们是把原训练集分成新的训练集合测试集，新的训练集不能与验证集有重合。如在 keras 中设置验证集的方法：可以自己建立验证集，然后在模型训练（fit 函数）时，传入 validation_data 中。

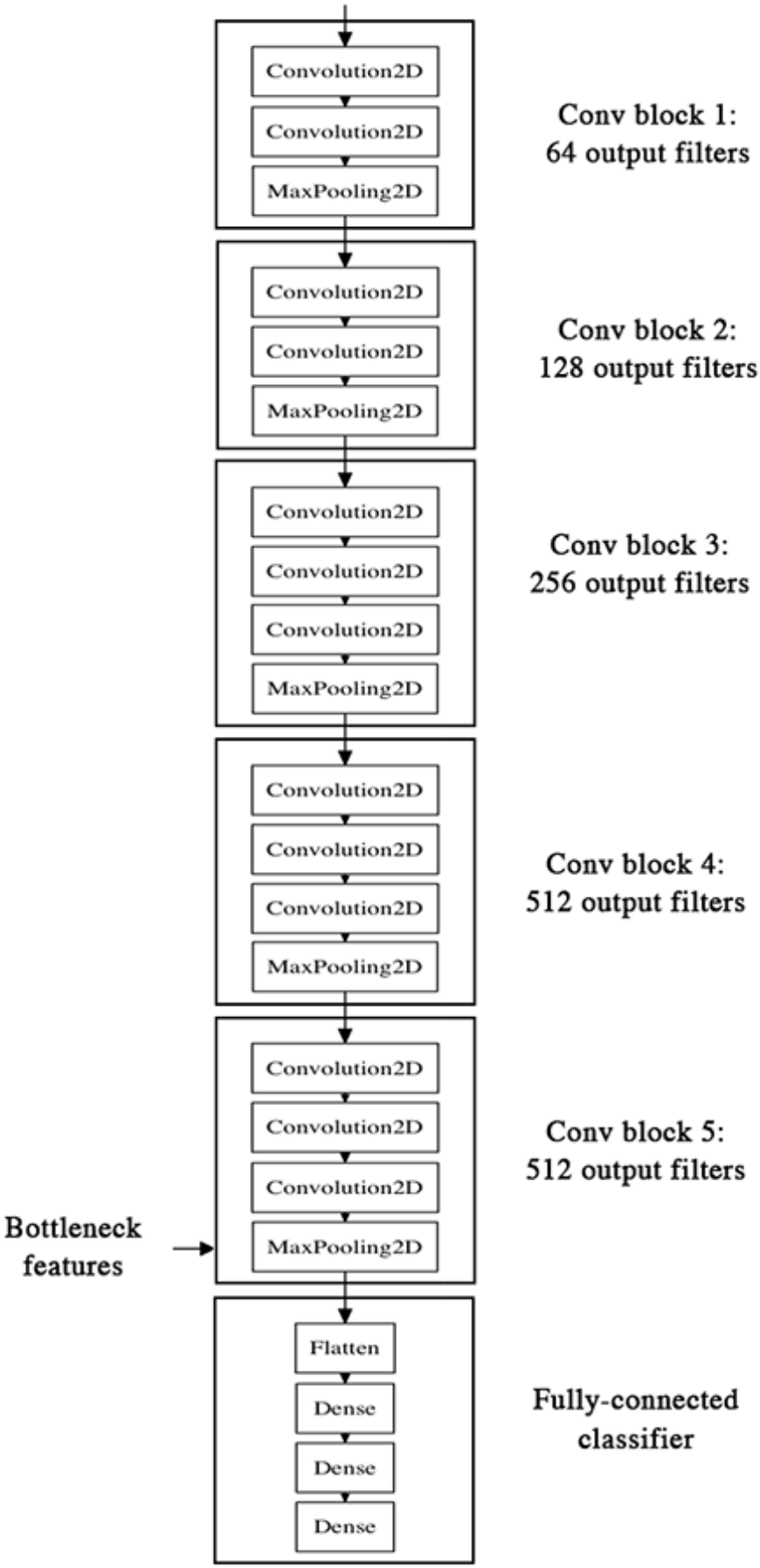
解决办法

本项目使用迁移学习技术[2]，顾名思义就是就是把已学训练好的模型参数迁移到新的模型来帮助新模型训练。考虑到大部分数据或任务是存在相关性的，所以通过迁移学习我们可以将已经学到的模型参数（也可理解为模型学到的知识）通过某种方式来分享给新模型从而加快并优化模型的学习效率不用像大多数网络那样从 0 学习。

本项目选择了 keras[3]工具，VGG16[4]作为预训练模型，因为该模型已预先在 ImageNet 数据集上进行训练，而 ImageNet 数据集已经包含了 1000 个类中的几个”猫”类和”狗”类，所以这个模型已经学到了与我们的分类问题相关的特征，作为本项目的猫狗分类效果应该会不错。

对于 VGG16 模型，首先去掉全连接层的网络，得到数据集的 bottleneck feature，然后自己设计几层全连接层，对其进行训练，最后再微调得到最终模型。

VGG16 的基本架构如下：


Convolution2D
Convolution2D
MaxPooling2D
Conv block 1:
64 output filters
Convolution2D
Convolution2D
MaxPooling2D
Conv block 2:
128 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 3:
256 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 4:
512 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 5:
512 output filters
Bottleneck
features
Flatten
Dense
Dense
Dense
Fully-connected
classifier

基准模型

本项目最后预测值的基准的最低要求是 kaggle Public Leaderboard 前 10%，也就是 logloss 要低于 0.06127。

评估指标

本项目评估指标是预测概率值的$loss\ log$，

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

其中，n为测试集的图片数量，$\hat{y}_i$为预测图片为狗的概率值，y_i为 1 则图片为狗，y_i为 0 则图片为猫，log()是以e为底的自然对数。$loss\ log$值越小越好。

设计大纲

1. **数据预处理**
2. **模型构建**
3. **模型训练**
4. **模型调参**
5. **模型评估**
6. **可视化**

参考文献：

[1] Kaggle, https://www.kaggle.com/c/dogs-vs-cats-redux-kernels-edition
[2] Transfer learning, https://en.wikipedia.org/wiki/Transfer_learning
[3] Keras Doc, https://keras.io/
[4] Karen, Simonyan, Andrew, Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition，2015