猫狗大战

机器学习工程师 纳米学位 开题报告

陈睿嘉

2018 年 1 月 17 日

项目概述

猫狗大战源自 Kaggle 于 2013 年举办的一个娱乐竞赛项目，要求分辨给定测试图片是猫还是狗，属于计算机视觉领域图像分类问题，适合使用以卷积神经网络（Convolutional Neural Network，CNN）构建模型、最小化对数损失函数作为策略、梯度下降与反向传播作为算法的深度学习方法。

项目背景

卷积神经网络根据生物视觉神经中感受野（Receptive Field）概念的启发，使用卷积核与上层的稀疏连接取代传统 DNN（Deep Neural Network）层与层之间的全连接权重，极大降低了网络参数数量，而理论上相比全连接网络表示能力只是略微降低，并且更容易训练。由于计算机视觉相关问题数据结构的特殊性（图像、视频等），往往每个像素只与附近区域的像素高度相关，非常适合使用 CNN 模型。

图像分类问题的代表性数据集为 ImageNet，需要从超过 100,0000 张图片的训练数据中学习识别 1000 种图像分类（包括猫狗等动物）。通常在 ImageNet 问题上表现优秀的模型，对其他图像分类问题甚至不限于分类问题具有良好的泛

化性能。ImageNet 的分类评价指标一般为 top-5 test error，即根据概率从大到小排名前 5 的分类结果中没有包括正确分类的百分比错误率。

VggNet 使用最深为 19 层的传统 CNN 结构，在 ImageNet 中达到单模型 7.0%，组合模型 6.8%的 top-5 error[1]。

GoogLeNet 即第一版 Inception 模型，增加网络宽度，减少了选择卷积核尺寸等参数的人为因素，直接将不同尺寸卷积核以及池化层放在同一层作为一个 block，并拼接为一个输出层，然后为了降低计算参数，使用 1×1 卷积核进行降维，达到单模型 7.9%，组合模型 6.7%的 top-5 error[2]。

ResNet 则着重解决网络深度增加难以训练的问题。以解决输入到输出的直接映射为核心思想，将输出函数改为源目标函数 H(x)与输入的残差（Residual）F(x)，再将输入层无参数连接（或维度映射）到输出层，使最终输出保持为 H(x)=F(x)+x，使网络能够达到极高的深度而几乎不增加额外参数。ResNet50 与 ResNet152 分别达到 6.7%与 5.7%，组合模型甚至达到 3.7%的 top-5 error[3]。

Inception v3 在 GoogLeNet 的基础上，将大型卷积核分解为多个小型甚至一维卷积核，并加上非线性激活单元，在降低参数数量的基础上提高了模型表现，组合模型最好表现达到 3.58%的 top-5 error[4]。

Inception v4 将 block 结构设计的更加复杂，使单模型 top-5 error 达到 3.8%，同时提出结合 ResNet 的 Inception-ResNet v2，达到 3.7%的 top-5 error 并且训练更快[5]。

Xception 将图像空间信息与 channel 信息分开处理，使用 depthwise separable convolution 方法对每个 channel 分开卷积，然后使用 1×1 卷积核进行合并，进一步降低参数数量，提高模型表现能力，在 ImageNet 上单模型表现接近 Inception v4，作者将此归为 Inception 的复杂设计对 ImageNet 存在过拟合[6]。

ResNeXt 针对 Inception 的 block 内部结构太过复杂而引入过多先验设计，设计一种子结构相同的 block，在复杂度降低一半的情况下达到比 Res200 更好的表现[7]。

猫狗大战作为二分类问题相对更简单，能够以更小的数据量达到更高的分类正确率，更适合初学者理解和实现图像分类甚至目标检测（Object Localization and Detection）问题的解决方案。

问题陈述

猫狗大战本质上属于监督学习，假设样本数据服从伯努利分布，学习一个条件概率分布模型 $p(y|x)$，输入图片像素值向量 x，输出样本 x 表示狗的概率 $\hat{y} = p(y = dog|x)$，那么样本表示猫的概率就为 $1-\hat{y}$。

然而，图像数据的特征不易提取，甚至不易理解，人类本身也无法准确描述分辨猫狗的核心特征。但若使用整个像素向量作为特征向量，以像素向量的高维度性，使用传统机器学习方法将带来维度灾难（Curse of dimensionality），需求像素维度指数级别的均匀样本，以及指数增长的模型规模与计算量。

数据描述

训练数据集包括 25000 张图片，并在文件名中标注了图片为猫还是狗，猫狗数量各占一半，测试数据集包括 12500 张图片，没有标注类别，文件以数值 ID 命名。

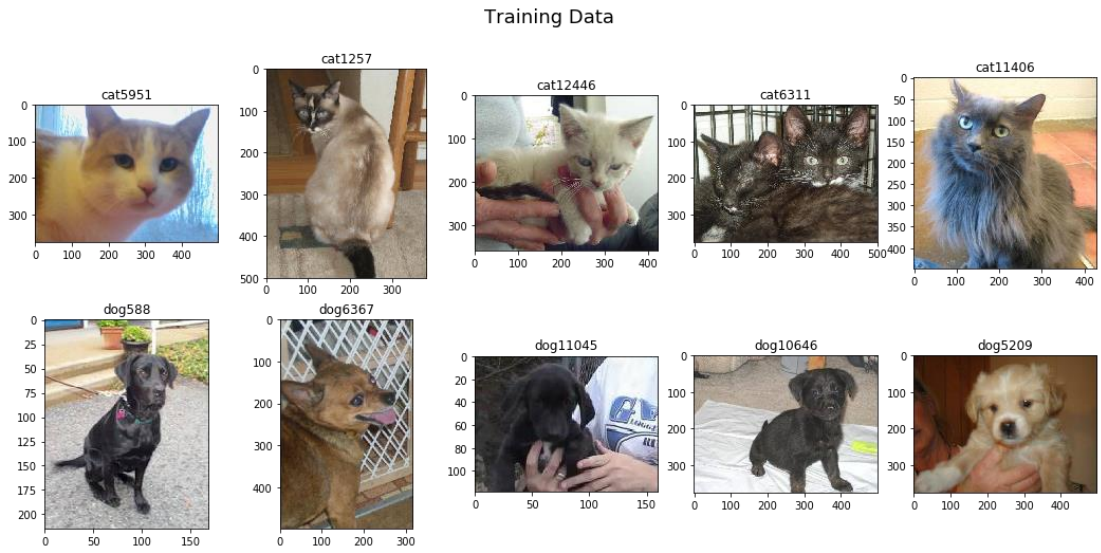

图 1. 训练数据按类别随机抽样

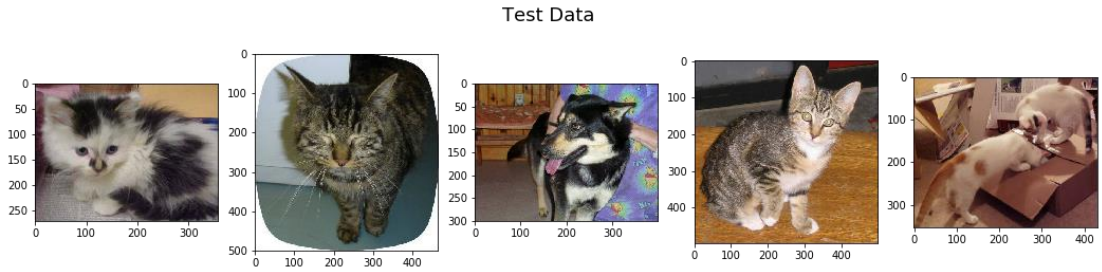

图 2. 训练数据随机抽样

如上图所示，图 1 为从训练数据集中随机抽样猫狗各 5 个样本，图 2 为从测试数据集中随机抽样 5 个样本，可见数据内容具备多样性与复杂性。如：图片尺寸参差不齐，不同图片包括了猫/狗身体不同角度，不同的猫/狗数量，有的图片没有包含全身，甚至没有正面脸部，图片背景元素也具备充分的干扰性。

解决方案

针对维度灾难，深度网络的层次结构可以将模型容量（Capacity）增长为样本维度的指数级别，而没有增加太多的模型参数，并且，高维度的图像数据向量并不是每种组合都有意义，模型需要学习的模式通常为数据向量的某种低维流形表示，随机像素组合大都没有意义，能表示猫狗甚至有意义的图像的像素向量只占极少数，并不需要样本维度指数级别的训练样本，使用深度网络模型满足解决此问题的理论基础。然而，对于图像数据，DNN 模型引入的参数太多，随着网络规模增加，几乎无法训练，由此，根据像素值往往只与附近区域的像素高度相关这一先验特征，CNN 模型成为了当前最适合进行图像分类的先验模型。

由于猫狗分类属于 ImageNet 的子问题，适合使用迁移学习（Transfer Learning）方法，使用预训练模型在 ImageNet 中学习到的经验参数表示猫狗图片的低维特征向量，并学习对这些特征向量进行分类。

基准模型

使用 ResNet50、Inception v3、Xception 组合模型输出特征向量进行逻辑回归（Logistic Regression），可以在验证集上达到 99.6%的正确率，在测试集的损失函数值达到 0.4141，提交 Kaggle 排名为 20/1314，要求项目结果不低于此标准。

评价指标

模型输出单元使用 Sigmoid 激活函数，损失函数采用对数损失函数，即伯努利分布下的交叉熵：

$$L(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$Cost = \frac{1}{m}\sum_{i=1}^{m} L(y_i,\hat{y}_i)$$

损失函数的形式可以通过对模型输出 $\hat{y} = p(y = dog \mid x)$ 取极大似然估计导出，并且其对数形式可以解决 Sigmoid 输出单元两端饱和导致梯度消失的缺陷。

模型在测试数据的分类结果可以上传到 Kaggle 进行评价、排名。Kaggle 对于测试结果的评价指标也使用对数损失函数，表示不仅要求分类准确率尽量高，还要求正确分类置信度尽量高。

项目设计

项目使用以 Tensorflow 为框架封装的 Keras API 进行模型实现，详细步骤如下：

数据预处理

剔除异常样本，异常值判定使用四分位数法，判定的统计指标可以是图片尺寸、像素均值等，也可以使用预训练模型输出图像特征向量，根据特征向量到均值向量的欧氏距离作为判定值。异常样本删除需慎重，轻微异常的样本有益于模

型的泛化性能，最好对判定为异常的样本图片进行人为判断是否保留。

图像尺寸统一，一般统一为所使用模型需要的输入尺寸，对于放大的图像需要进行插值处理。

像素值规范化（Normalization），一般将像素值映射到[0,1]或[-1,1]的闭区间，不同模型需求的映射区间不同。

数据增强（Data Augmentation），方法包括如翻转、平移、旋转、缩放、错切（Shear）、高光、低光等方式处理训练样本，必要时进行插值处理。数据增强可以增加训练集、验证集样本数量，有益于减少过拟合，提高模型泛化能力，Keras 的 ImageDataGenerator 接口已经封装了大部分所需要用到的方法。

数据混洗、切分，将 25000 张训练样本随机打乱，并按猫狗各一半的比例切分出 8%的样本作为验证集，得到训练集数量 23000，验证集数量 2000。

基本模型

由于使用迁移学习方法，可以选择一个在 ImageNet 数据集中预训练的模型（包括但不限于：VggNet、ResNet、Inception、Xception 等），其卷积层已经学习到常见图像数据中的关键特征。一般做法是通过预训练模型的卷积层前馈输出训练数据的特征向量，并使用如 Adam、RMSprop 等自适应学习方法其进行逻辑回归，快速得到一个基本模型，并记录模型在训练集和验证集的损失函数取值 Loss 以及预测准确率，绘制学习曲线。

改进模型

观察基本模型的学习曲线，若训练集和验证集 Loss 都较高，说明存在欠拟合，需要增加模型容量，可以尝试使用多个预训练模型，拼接输出特征向量，或增加全连接层；若训练集 Loss 很低而验证集较高，说明存在过拟合，需要使用正则化（Regularization）方法，由于猫狗分类问题由于输入数据（包括测试数据）的局限性，需要模型对猫狗二类图片数据特征产生一定的过拟合，某些关键特征权重会明显高于其他特征，使用 L1/L2 Regularization 方法将抑制权重，故而一般使用 Dropout 与 Early Stop 方法。

得到一个满意的模型后，使用全部训练数据投入训练，调节 batch_size、epoch、momentum 等超参数使模型收敛的更快、更好。

Fine-tune

为了进一步提高模型表现，需要对预训练模型与输出向量相连的卷积 block 权重参数进行微调，使用随机梯度下降（Stochastic Gradient Decent，SGD）方法重新训练模型。由于适用于 ImageNet 问题的模型应对猫狗问题时，本身就存在严重过拟合，过度修改参数可能破坏其学习到的具备较强泛化能力的关键特征，所以不需要对整个网络参数进行调整，并且在训练模型时需要保持较低的学习率。

最后，使用最终模型对测试数据进行预测，并将结果上传至 Kaggle。

参考文献

[1] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[2] Christian Szegedy, Wei Liu, Yangqing Jia. Going Deeper with Convolutions, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren. Deep Residual Learning for Image Recognition, 2015.

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe. Rethinking the Inception Architecture for Computer Vision, 2015.

[5] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 2016.

[6] François Chollet. Xception: Deep Learning with Depthwise Separable Convolutions, 2017.

[7] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He. Aggregated Residual Transformations for Deep Neural Networks, 2017.