# 猫狗大战

## 机器学习工程师 纳米学位 毕业项目

陈睿嘉  
2018年3月21日

## **1. 定义**

### **1.1 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，提供了25000张猫狗训练数据以及12500张测试数据，数据主要为真实场景采集的猫狗照片。

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。

### **1.2 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，输入图片像素值向量x，输出样本x表示狗的概率，那么样本表示猫的概率就为。

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。

针对维度灾难，深度神经网络（DNN, Deep Neural Network）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。

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。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。

### **1.3 评价指标**

二分类模型输出单元通常使用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：


损失函数的形式可以通过对模型输出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。

## **2. 分析**

### **2.1 数据描述**

项目使用了Kaggle竞赛提供的数据集，其训练数据集包括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。

需要注意，由于猫狗数据样本比例相当，在进行数据抽样、分割等操作时，必须保证不同子数据集猫狗样本也各占一半。

图片样本特征十分抽象，很难使用数据统计方法剔除异常值，如下对比较直观的特征包括图像宽、高、宽高比进行了箱型图统计：

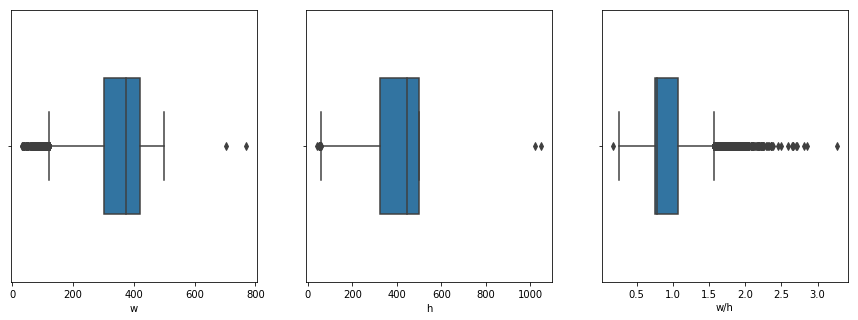


图3. 图像尺寸数据箱型图

如图3所示，虽然大部分数据分布在正常尺寸，但仍然存在大量异常样本，由于数量太大，逐个进行裁剪、填充需要大量时间，并且深度学习对于此类异常样本具有较强鲁棒性。所以，需要使用预训练的分类模型对数据集进行初步分类，识别非猫狗异常样本，然后再对数量较少的异常样本集合进行人为判断，对内容正确但尺寸异常的数据进行裁剪、填充。

### 2.2 算法和技术

### 2.2.1 深度学习

针对传统机器学习方法对于高维度相关性数据（如图片、语音、文字等）上表示能力的不足，深度学习着重解决了由此带来的维度灾难以及计算复杂度问题。

深度神经网络结构提出隐藏层的概念，并且由万能近似定理（Universal Approximation Theorem）证明，若不考虑训练难度，只要给予足够数量的隐藏单元，模型可以以任意精度近似一个从有限维度映射到另一个有限维度的可测函数，并且，模型的表示能力是网络深度的指数级函数[1]。

对于高维度相关性数据，以图片为例，若对每个维度根据均匀分布随机选择像素值，最终有非常大的概率得到噪声图像，说明数据分布通常集中在某个坐标区域，可以使用一个低维流形坐标系表示，深度学习模型可以很好的近似这一坐标映射函数。

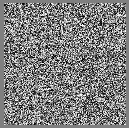


图4. 随机生成的噪声图像

传统机器学习方法如支持向量机（SVM, Support Vector Machine），其核函数也可以将数据隐式映射到合适的维度，但很难针对不同需求找到合适的核函数，深度学习可以通过DNN结构近似表达此种显式映射函数，以解决传统方法的不足。

#### 2.2.2 卷积神经网络

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普通DNN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。

卷积是一种数学运算，其标准定义为：



卷积运算的实际意义需要应用到不同场景，若将t理解为时间，令a<0时w=0，可以理解为从过去到时刻t，x以w为权重的累加和。扩展到离散场景时，可定义为：



在CNN图像应用场景中，卷积可以扩展为更为直观的二维形式：



其中I表示输入图像，i、j为图像像素坐标，K表示卷积核，W、H为卷积核尺寸，即对输入图像中每个像素值（暂不考虑步长）及其附近W、H区域的像素值进行加权累加和，以得到一副新的“图像”。

通过卷积计算得到的新“图像”，可以表示出原图像的某些特征，比如传统的边缘检测算子就是一种人为设计权重的卷积核，将如3×3大小的卷积核分别对原图像每个像素进行卷积运算，从而得到如图5所示的边缘检测效果，而CNN模型则是根据代价函数和梯度下降让卷积核自己学习需要的权重，让模型自己学习需要提取的图像特征。并且，随着模型深度的增加，卷积核的视野逐渐增大，从而可以提取更为高级、抽象的特征，如从边缘到毛发，最终到身体轮廓。



图5. 边缘检测算子

相对于DNN结构，CNN最大的优势就是通过稀疏交互和参数共享极大降低了网络参数数量，并且保留了特征的平移不变性。对于图像数据的维度规模，大型DNN结构带来的参数量与计算量使模型几乎无法训练，并且将会严重过拟合，相对的，CNN的缺点是增加了额外的需要人为调整的超参数，如卷积核大小、步长等。

#### 2.2.3 解决方案

项目使用Python3.6开发，通过对Tensorflow进一步封装的Keras API进行模型实现与训练。Tensorflow是Google提供的深度学习开源框架，其主要功能是自动完成了计算图的梯度计算与反向传播过程，使开发者可以专注于模型搭建与训练，并且将前馈计算、反向传播等密集浮点运算交由效率更高的语言完成，支持GPU加速计算。由于GPU的构架特点，其浮点运算性能远超CPU，使用NVIDIA GPU加速计算，还需要安装CUDA以及cuDNN。Keras对Tensorflow进行了进一步封装，实现了常用的层次结构，使调用接口更加人性化，并且提供了一些搭建好的预训练模型，以提高开发效率，避免重复造轮子。

项目主要使用Keras提供的ImageNet数据集预训练模型进行迁移学习，通过fine-tune卷积层、多模型拼接等方法提高模型性能，以及通过Dropout、图像增强等方法解决过拟合。

### **2.3 基准指标**

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[2]。

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[3]。

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[4]。

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[5]。

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[6]。

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[7]。

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[8]。

使用ResNet50、Inception v3、Xception组合模型进行迁移学习，对输出的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。

## 3. 方法

### **3.1 模型选择**

在Keras目前提供的所有预训练模型中，[InceptionResNetV2](https://keras.io/applications/" \l "inceptionresnetv2)、Xception、InceptionV3模型在ImageNet验证集表现位列前三，所以选用此三个模型分别进行实验。

在将图片数据输入模型之前，需要通过重采样或插值方法将尺寸映射到模型需要的尺寸，所选择的三个模型输入尺寸都为299×299，并需要将像素值从[0, 255]的取值范围映射到[-1, 1]。

### **3.2 数据清洗**

由于ImageNet分类中已包含118种狗类别和7种猫类别，使用预训练模型分别对各12500个猫狗训练样本进行预测，并输出前n个预测结果分类，若猫样本的n个预测结果中没有7种猫类别，或狗样本的n个预测结果中没有118种狗类别，即判定该样本为异常样本，n值的大小经过反复实验设定为猫样本60，狗样本20。最终使用InceptionResNetV2、Xception、InceptionV3三个模型分别预测后，对三种结果取并集得到共116个异常样本。

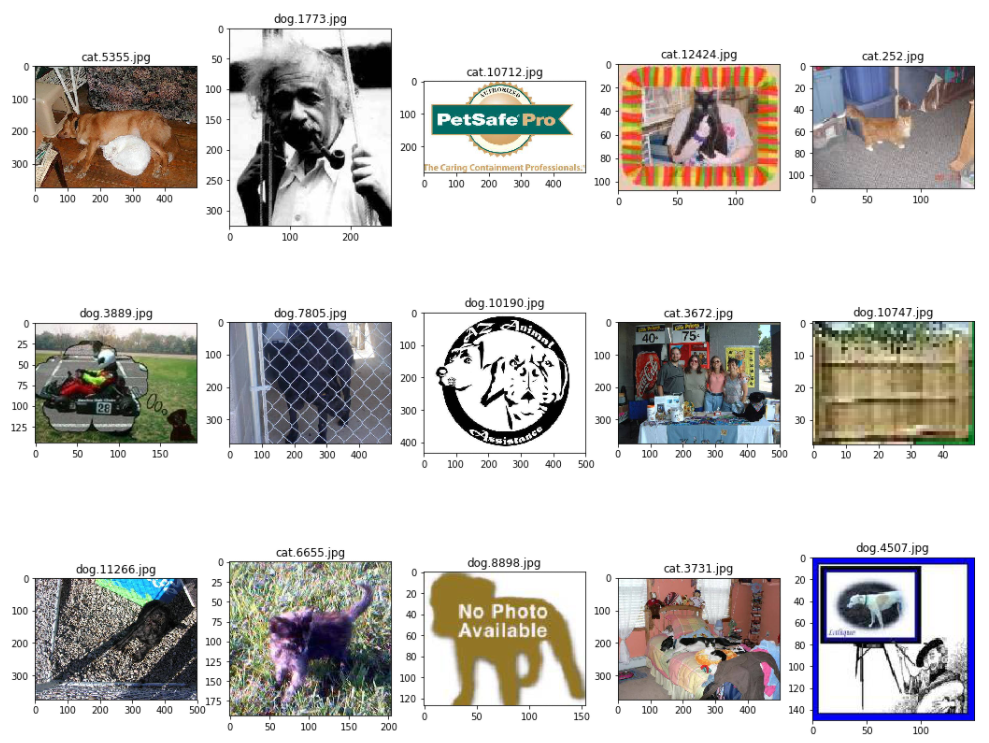


图6. 部分异常样本

可以发现，其中大部分图片如“dog.1773.jpg”、“cat.10712.jpg”判定为异常样本毫无疑问，必须剔除，而如“cat.12424.jpg”、“cat.252.jpg”属于分辨率过低，也可以剔除，又如“dog.7805.jpg”此种类型，背景干扰太大。为了进一步筛选异常样本，在将此116个异常样本剔除后，本次项目从中进一步人为判断筛选出11个轻度异常样本，并将其裁剪、填充至模型需要的尺寸，项目选用的三个模型输入尺寸都为299×299，如图6中“cat.6655.jpg”，将其填充为图7所示：

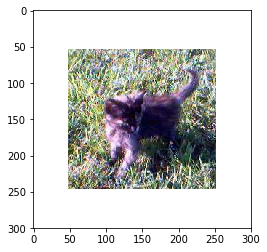


图7. 裁剪填充结果

### **3.3 迁移学习**

迁移学习利用预训练模型学习到的具有较强泛化性能的高级抽象特征，使数据通过这些特征表示几乎线性可分，从而节省训练时间。一般方法为

首先选择一个预训练模型，

需要注意，由于Kaggle评估指标为损失函数，进行多模型拼接时，不宜使用投票表决法，该方法通常对准确度指标提升较大，同时，选择解决过拟合方法时，也不宜使用L1/L2 Regularization，前者会导致权重稀疏，后者会导致权重衰减，两者都会使预测时输出层Sigmoid函数接收的激活绝对值变小，从而输出概率不够自信。Dropout方法在每次训练时随机丢掉一部分特征，使当前某些强特征随时会被在下次训练时丢弃掉，从而让模型不会过分依赖某几个强特征，虽然也会降低训练时的输出层激活绝对值，但预测过程中将不会对特征进行丢弃，非常适用于此项目；图像增强方法是通过让模型见到更多人造数据以提高泛化性能，但会相应增加每代训练时间。

## 参考文献

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

**[2] Christian Szegedy, [Wei Liu](https://arxiv.org/find/cs/1/au:+Liu_W/0/1/0/all/0/1), [Yangqing Jia](https://arxiv.org/find/cs/1/au:+Jia_Y/0/1/0/all/0/1). 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](https://arxiv.org/find/cs/1/au:+Ioffe_S/0/1/0/all/0/1). 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.**