猫狗大战

机器学习工程师纳米学位毕业项目

沈龙飞

2018年3月1号

# 定义

## 1.1项目概述

这是一个Kaggle上的竞赛项目，目标是训练一个深度学习模型，在给定的图像中分辨猫和狗。属于图像分类问题。

本项目中使用的模型是**卷积神经网络**（Convolutional Neural Network, **CNN**）是一种前馈神经网络，它的人工神经元可以响应一部分覆盖范围内的周围单元，对于大型图像处理有出色表现。

卷积神经网络由一个或多个卷积层和顶端的全连通层（对应经典的神经网络）组成，同时也包括关联权重和池化层（pooling layer）。这一结构使得卷积神经网络能够利用输入数据的二维结构。与其他深度学习结构相比，卷积神经网络在图像和语音识别方面能够给出更好的结果。这一模型也可以使用反向传播算法进行训练。相比较其他深度、前馈神经网络，卷积神经网络需要考量的参数更少，使之成为一种颇具吸引力的深度学习结构。

项目选择的数据集是Kaggle上竞赛提供的数据，训练集包括12500张被标记为猫和12500标记为猫的图片。测试机是包含12500未标记的图片。对于每一张测试机中的图像，模型需要预测出是狗图像的概率（1 代表狗，0 代表猫）。

## 问题陈述

Kaggle竞赛数据是从真实数据中采集到的包含猫或狗的图像。图像的分辨率参差不齐，

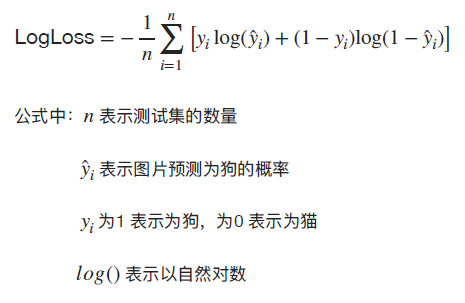
图片中猫和狗的主体也有很大的差异。同时猫和狗的不同品种造成的本身的形体差异。在加上拍摄角度与所处角度环境的关系。照成图片数据中的异常噪音非常大。

在图像分类问题中，需要出现两个过程：体征提取和分类。在特征提取环节，模型需要提取边缘，轮廓，色彩组合等的不同抽象级别特征。在分类部分，需要对提取出的特征进行准确分类。以得出分类结果。

## 评价指标

在kaggle的比赛中提出了标准的评价公式，本项目也采用相同的评价指标。

公式如下：

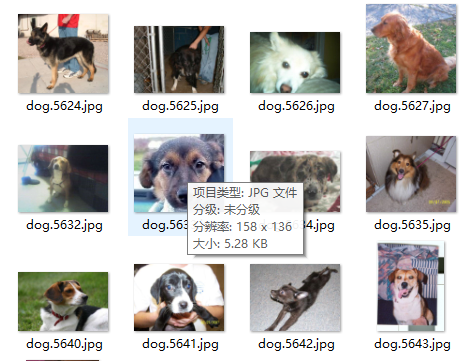


# 分析

## 2.1 数据可视化

竞赛数据集中有两个子集，一个训练数据集，一个是测试数据集，训练数据集中的文件名称即为标签。说明了图片数据猫还是狗。





训练集中共有25000张图片，猫狗各占一半，因此最后训练的时候不需要调整类别权重。

同时猫狗形态各异，要额外注意欠拟合和过拟合现象。

## 算法和技术

### 深度学习简介

**深度学习**（deep learning）是机器学习的分支，是一种试图使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。

深度学习框架，尤其是基于[人工神经网络](https://zh.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C" \o "人工神经网络)的框架可以追溯到1980年福岛邦彦提出的新认知机，而人工神经网络的历史更为久远。1989年，[扬·勒丘恩](https://zh.wikipedia.org/wiki/%E6%89%AC%C2%B7%E5%8B%92%E4%B8%98%E6%81%A9" \o "扬·勒丘恩)（Yann LeCun）等人开始将1974年提出的标准[反向传播算法](https://zh.wikipedia.org/wiki/%E5%8F%8D%E5%90%91%E4%BC%A0%E6%92%AD%E7%AE%97%E6%B3%95" \o "反向传播算法)[[12]](https://zh.wikipedia.org/wiki/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0#cite_note-WERBOS1974-12)应用于深度神经网络，这一网络被用于手写邮政编码识别。尽管算法可以成功执行，但计算代价非常巨大，神经网路的训练时间达到了3天，因而无法投入实际使用。许多因素导致了这一缓慢的训练过程，其中一种是由[于尔根·施密德胡伯](https://zh.wikipedia.org/w/index.php?title=%E4%BA%8E%E5%B0%94%E6%A0%B9%C2%B7%E6%96%BD%E5%AF%86%E5%BE%B7%E8%83%A1%E4%BC%AF&action=edit&redlink=1)的学生[赛普·霍克赖特](https://zh.wikipedia.org/w/index.php?title=%E8%B5%9B%E6%99%AE%C2%B7%E9%9C%8D%E5%85%8B%E8%B5%96%E7%89%B9&action=edit&redlink=1)于1991年提出的梯度消失问题。

最早的进行一般自然杂乱图像中自然物体识别的深度学习网络是翁巨扬（Juyang Weng）等在1991和1992发表的生长网（Cresceptron）。它也是第一个提出了后来很多实验广泛采用的一个方法：现在称为最大汇集（max-pooling)以用于处理大物体的变形等问题。生长网不仅直接从杂乱自然场景中学习老师指定的一般物体，还用网络反向分析的方法把图像内被识别了的物体从背景图像中分区出来。

2007年前后，[杰弗里·辛顿](https://zh.wikipedia.org/wiki/%E6%9D%B0%E5%BC%97%E9%87%8C%C2%B7%E8%BE%9B%E9%A1%BF" \o "杰弗里·辛顿)和鲁斯兰·萨拉赫丁诺夫（Ruslan Salakhutdinov）提出了一种在前馈神经网络中进行有效训练的算法。这一算法将网络中的每一层视为[无监督](https://zh.wikipedia.org/w/index.php?title=%E6%97%A0%E7%9B%91%E7%9D%A3%E5%AD%A6%E4%B9%A0&action=edit&redlink=1" \o "无监督学习（页面不存在）)的[受限玻尔兹曼机](https://zh.wikipedia.org/wiki/%E5%8F%97%E9%99%90%E7%8E%BB%E5%B0%94%E5%85%B9%E6%9B%BC%E6%9C%BA)，再使用有监督的反向传播算法进行调优。在此之前的1992年，在更为普遍的情形下，施密德胡伯也曾在[递归神经网络](https://zh.wikipedia.org/wiki/%E9%80%92%E5%BD%92%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C" \o "递归神经网络)上提出一种类似的训练方法，并在实验中证明这一训练方法能够有效提高有监督学习的执行速度。

自深度学习出现以来，它已成为很多领域，尤其是在计算机视觉和语音识别中，成为各种领先系统的一部分。在通用的用于检验的数据集，例如语音识别中的TIMIT和图像识别中的ImageNet, Cifar10上的实验证明，深度学习能够提高识别的精度。与此同时，神经网络也受到了其他更加简单归类模型的挑战，[支持向量机](https://zh.wikipedia.org/wiki/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA" \o "支持向量机)等模型在20世纪90年代到21世纪初成为过流行的机器学习算法。

硬件的进步也是深度学习重新获得关注的重要因素。高性能[图形处理器](https://zh.wikipedia.org/wiki/%E5%9C%96%E5%BD%A2%E8%99%95%E7%90%86%E5%99%A8" \o "图形处理器)的出现极大地提高了数值和矩阵运算的速度，使得机器学习算法的运行时间得到了显著的缩短。

由于脑科学方面的大量研究已表明人脑网络不是一个[级联](https://zh.wikipedia.org/wiki/%E7%BA%A7%E8%81%94" \o "级联)的结构，深度学习网络在2001年后正逐渐被更有潜力的基于脑模型的网络所替代。

一部分最成功的深度学习方法涉及到对[人工神经网络](https://zh.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C" \o "人工神经网络)的运用。人工神经网络受到了1959年由诺贝尔奖得主[大卫·休伯尔](https://zh.wikipedia.org/wiki/%E5%A4%A7%E8%A1%9B%C2%B7%E4%BC%91%E4%BC%AF%E7%88%BE" \o "大卫·休伯尔)（David H. Hubel）和[托斯坦·威泽尔](https://zh.wikipedia.org/wiki/%E6%89%98%E6%96%AF%E5%9D%A6%C2%B7%E5%A8%81%E6%B3%BD%E5%B0%94" \o "托斯坦·威泽尔)（Torsten Wiesel）提出的理论启发。休伯尔和威泽尔发现，在大脑的[初级视觉皮层](https://zh.wikipedia.org/wiki/%E5%88%9D%E7%BA%A7%E8%A7%86%E8%A7%89%E7%9A%AE%E5%B1%82" \o "初级视觉皮层)中存在两种细胞：简单细胞和复杂细胞，这两种细胞承担不同层次的视觉感知功能。受此启发，许多神经网络模型也被设计为不同节点之间的分层模型。

[福岛邦彦](https://zh.wikipedia.org/w/index.php?title=%E7%A6%8F%E5%B2%9B%E9%82%A6%E5%BD%A6&action=edit&redlink=1)提出的新认知机引入了使用无监督学习训练的卷积神经网络。[扬·勒丘恩](https://zh.wikipedia.org/wiki/%E6%89%AC%C2%B7%E5%8B%92%E4%B8%98%E6%81%A9)将有监督的[反向传播算法](https://zh.wikipedia.org/wiki/%E5%8F%8D%E5%90%91%E4%BC%A0%E6%92%AD%E7%AE%97%E6%B3%95)应用于这一架构。事实上，从反向传播算法自20世纪70年代提出以来，不少研究者都曾试图将其应用于训练有监督的深度神经网络，但最初的尝试大都失败。[赛普·霍克赖特](https://zh.wikipedia.org/w/index.php?title=%E8%B5%9B%E6%99%AE%C2%B7%E9%9C%8D%E5%85%8B%E8%B5%96%E7%89%B9&action=edit&redlink=1)在其博士论文中将失败的原因归结为梯度消失，这一现象同时在深度前馈神经网络和递归神经网络中出现，后者的训练过程类似深度网络。在分层训练的过程中，本应用于修正模型参数的误差随着层数的增加指数递减，这导致了模型训练的效率低下。

为了解决这一问题，研究者们提出了一些不同的方法。[于尔根·施密德胡伯](https://zh.wikipedia.org/w/index.php?title=%E4%BA%8E%E5%B0%94%E6%A0%B9%C2%B7%E6%96%BD%E5%AF%86%E5%BE%B7%E8%83%A1%E4%BC%AF&action=edit&redlink=1)于1992年提出多层级网络，利用无监督学习训练深度神经网络的每一层，再使用反向传播算法进行调优。在这一模型中，神经网络中的每一层都代表观测变量的一种压缩表示，这一表示也被传递到下一层网络。

另一种方法是[赛普·霍克赖特](https://zh.wikipedia.org/w/index.php?title=%E8%B5%9B%E6%99%AE%C2%B7%E9%9C%8D%E5%85%8B%E8%B5%96%E7%89%B9&action=edit&redlink=1" \o "赛普·霍克赖特（页面不存在）)和[于尔根·施密德胡伯](https://zh.wikipedia.org/w/index.php?title=%E4%BA%8E%E5%B0%94%E6%A0%B9%C2%B7%E6%96%BD%E5%AF%86%E5%BE%B7%E8%83%A1%E4%BC%AF&action=edit&redlink=1)提出的[长短期记忆神经网络](https://zh.wikipedia.org/w/index.php?title=%E9%95%BF%E7%9F%AD%E6%9C%9F%E8%AE%B0%E5%BF%86%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C&action=edit&redlink=1)，LSTM。2009年，在ICDAR 2009举办的连笔手写识别竞赛中，在没有任何先验知识的情况下，深度多维长短期记忆神经网络获取了其中三场比赛的胜利。

[斯文·贝克](https://zh.wikipedia.org/w/index.php?title=%E6%96%AF%E6%96%87%C2%B7%E8%B4%9D%E5%85%8B&action=edit&redlink=1)提出了在训练时只依赖梯度符号的神经抽象金字塔模型，用以解决图像重建和人脸定位的问题。

其他方法同样采用了无监督预训练来构建神经网络，用以发现有效的特征，此后再采用有监督的反向传播以区分有标签数据。辛顿等人于2006年提出的深度模型提出了使用多层隐变量学习高层表示的方法。这一方法使用斯摩棱斯基于1986年提出的[受限玻尔兹曼机](https://zh.wikipedia.org/wiki/%E5%8F%97%E9%99%90%E7%8E%BB%E5%B0%94%E5%85%B9%E6%9B%BC%E6%9C%BA" \o "受限玻尔兹曼机)对每一个包含高层特征的层进行建模。 模型保证了数据的对数似然下界随着层数的提升而递增。当足够多的层数被学习完毕，这一深层结构成为一个生成模型，可以通过自上而下的采样重构整个数据集。辛顿声称这一模型在高维结构化数据上能够有效地提取特征。

[吴恩达](https://zh.wikipedia.org/wiki/%E5%90%B4%E6%81%A9%E8%BE%BE)和[杰夫·迪恩](https://zh.wikipedia.org/wiki/%E5%82%91%E5%A4%AB%C2%B7%E8%BF%AA%E6%81%A9)领导的[谷歌大脑](https://zh.wikipedia.org/wiki/%E8%B0%B7%E6%AD%8C%E5%A4%A7%E8%84%91)团队创建了一个仅通过[YouTube](https://zh.wikipedia.org/wiki/YouTube)视频学习高层概念（例如猫）的神经网络。

其他方法依赖了现代电子计算机的强大计算能力，尤其是[GPU](https://zh.wikipedia.org/wiki/GPU)。2010年，在[于尔根·施密德胡伯](https://zh.wikipedia.org/w/index.php?title=%E4%BA%8E%E5%B0%94%E6%A0%B9%C2%B7%E6%96%BD%E5%AF%86%E5%BE%B7%E8%83%A1%E4%BC%AF&action=edit&redlink=1" \o "于尔根·施密德胡伯（页面不存在）)位于[瑞士人工智能实验室](https://zh.wikipedia.org/w/index.php?title=%E7%91%9E%E5%A3%AB%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%AE%9E%E9%AA%8C%E5%AE%A4&action=edit&redlink=1)[IDSIA](https://zh.wikipedia.org/w/index.php?title=IDSIA&action=edit&redlink=1)的研究组中，[丹·奇雷尚](https://zh.wikipedia.org/w/index.php?title=%E4%B8%B9%C2%B7%E5%A5%87%E9%9B%B7%E5%B0%9A&action=edit&redlink=1" \o "丹·奇雷尚（页面不存在）)（Dan Ciresan）和他的同事展示了利用GPU直接执行反向传播算法而忽视梯度消失问题的存在。这一方法在[扬·勒丘恩](https://zh.wikipedia.org/wiki/%E6%89%AC%C2%B7%E5%8B%92%E4%B8%98%E6%81%A9" \o "扬·勒丘恩)等人给出的手写识别MNIST数据集上战胜了已有的其他方法。

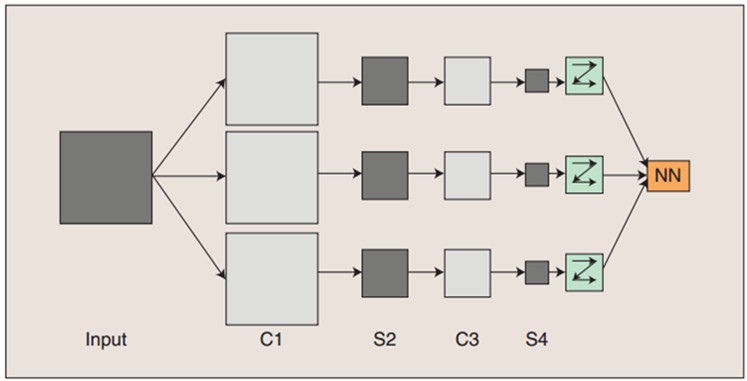
截止2011年，前馈神经网络深度学习中最新的方法是交替使用卷积层（convolutional layers）和最大值池化层（max-pooling layers）并加入单纯的分类层作为顶端。训练过程也无需引入无监督的预训练。从2011年起，这一方法的GPU实现多次赢得了各类模式识别竞赛的胜利，包括IJCNN 2011交通标志识别竞赛和其他比赛。

这些深度学习算法也是最先在某些识别任务上达到和人类表现具备同等竞争力的算法。

### 卷积神经网络

卷积神经网络（Convolutional Neural Network, CNN）是一种[前馈神经网络](https://zh.wikipedia.org/wiki/%E5%89%8D%E9%A6%88%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C" \o "前馈神经网络)，它的人工神经元可以响应一部分覆盖范围内的周围单元，对于大型图像处理有出色表现。

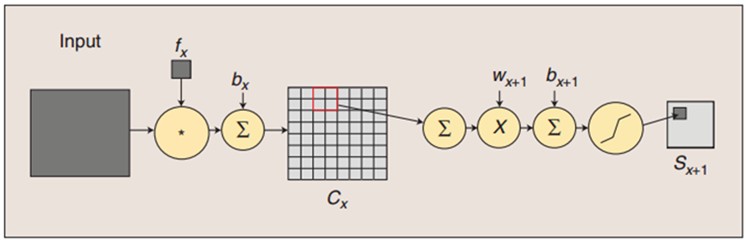
卷积神经网络由一个或多个卷积层和顶端的全连通层（对应经典的神经网络）组成，同时也包括关联权重和[池化](https://zh.wikipedia.org/w/index.php?title=%E6%B1%A0%E5%8C%96&action=edit&redlink=1" \o "池化（页面不存在）)层（pooling layer）。这一结构使得卷积神经网络能够利用输入数据的二维结构。与其他深度学习结构相比，卷积神经网络在图像和[语音识别](https://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3%E8%AF%86%E5%88%AB" \o "语音识别)方面能够给出更好的结果。这一模型也可以使用[反向传播算法](https://zh.wikipedia.org/wiki/%E5%8F%8D%E5%90%91%E4%BC%A0%E6%92%AD%E7%AE%97%E6%B3%95)进行训练。相比较其他深度、前馈神经网络，卷积神经网络需要考量的参数更少，使之成为一种颇具吸引力的深度学习结构。



*图：卷积神经网络的概念示范：输入图像通过和三个可训练的滤波器和可加偏置进行卷积，滤波过程如图一，卷积后在C1层产生三个特征映射图，然后特征映射图中每组的四个像素再进行求和，加权值，加偏置，通过一个Sigmoid函数得到三个S2层的特征映射图。这些映射图再进过滤波得到C3层。这个层级结构再和S2一样产生S4。最终，这些像素值被光栅化，并连接成一个向量输入到传统的神经网络，得到输出。*

### 卷积层的基本原理

卷积层（Convolutional layer），卷积神经网络中每层卷积层由若干卷积单元组成，每个卷积单元的参数都是通过[反向传播算法](https://zh.wikipedia.org/wiki/%E5%8F%8D%E5%90%91%E4%BC%A0%E6%92%AD%E7%AE%97%E6%B3%95" \o "反向传播算法)最佳化得到的。卷积运算的目的是提取输入的不同特征，第一层卷积层可能只能提取一些低级的特征如边缘、线条和角等层级，更多层的网路能从低级特征中迭代提取更复杂的特征。



*图：卷积和子采样过程：卷积过程包括：用一个可训练的滤波器fx去卷积一个输入的图像（第一阶段是输入的图像，后面的阶段就是卷积特征map了），然后加一个偏置bx，得到卷积层Cx。子采样过程包括：每邻域四个像素求和变为一个像素，然后通过标量Wx+1加权，再增加偏置bx+1，然后通过一个sigmoid激活函数，产生一个大概缩小四倍的特征映射图Sx+1。*

通过对CNN和卷积的基本了解，大致可以认为在整个模型构建过程中，可以对结果产生影响的方面包括：

- 卷积的层数

- 滤波器的大小

- 隐藏层的数据

### 2.2.4 技术实现

项目实现CNN使用的是Facebook 开源的PyTorch。由于使用CNN处理分辨率全彩色图片所需要的计算量非常大，项目使用了GPU进行计算加速。在安装PyTorch的同时还安装了CUDA和cuDNN。

在图片处理的CNN问题中，主要包含的步骤如下：

1. 读取图片
2. 标准化图片
3. 卷积和池化（可以出现多次卷积和池化）
4. 归一化（隐藏层）
5. 输出结果

整个神经网络中由于面对具体问题的不一样可能会在不同的步骤呈现不同的特点。在猫狗问题中，我们面对的主要挑战来自，图片像素尺寸，以及猫狗动作的不一致性非常高。还包含又猫和狗的品种的不一致产生的数据噪音问题。

用深度学习中的术语来描述，则是说图片中提取的特征需要尽可能多，同时又要避免产生过拟合。

在深度学习的整个流程中，特征提取主要是在图片标准化，卷积和池化环境产生。换言之，这几个环节提取的特征数量，对后续的结果输出影响会很大。

## 基准指标

依据本题的考核标准。需要获得kaggle比赛前10%的成绩，在Kaggle的竞赛页面中，截止目前提交成绩人数为1314名。前10%则意味着分数要超过131名，在页面中可以看到第131名的得分为0.06127

基准模型指标得分需要小于0.06127

# 具体方法

## 3.1 迁移学习

### 3.1.1 迁移学习介绍

迁移学习，简单的说，就是能让现有的模型算法稍加调整即可应用于一个新的领域和功能的一项技术。这个概念目前在机器学习中其实比较少见，但其实它的潜力可以相当巨大。杨强教授在CCF-GAIR上的演讲中曾提到一个愿景——利用迁移学习，即使是自身没有条件获得大量训练数据的小公司也可以按照自己的需要应用大公司训练出来的模型，从而普及AI的应用。

本题中从零开始训练一个卷积神经网络，需要进行细致的网格结果设计，并调整大量的参数进行优化。耗费时间和精力非常惊人。因此我们采用迁移学习，通过微调其中参数以得到较好的结果。

### 3.1.2 模型选择

依据之前得出的论断，我们需要的是对特征尽可能多的提取，以保证在后面的全连接层中可以得到较好的分类结果。

而在PyTorch中提供有多种模型，最符合上面筛选条件的是ResNet模型。

两种主要的迁移学习方法如下：

**微调convnet**：我们使用预训练的网络进行参数初始化而不是参数随机初始化。之后的训练就和平常的训练一样了。

**convnet作为特征提取器**：这里，我们将冻结所有网络层的权重，除了最后一层全连接层。最后一层是一个参数随机初始化的新层，并且只有这一层被训练。

### 3.1.3 数据预处理

由于我们使用的迁移学习，所有数据预处理必须要能符合模型的输入结构。