北京邮电大学
Beijing University of Posts and Telecommunications

http://www.bupt.edu.cn email@bupt.edu.cn

DOI: 10.13700/j.bh.1001-5965.****.****

# 迁移学习实现猫狗图像分类

邓心一，姜百淳，刘姜旺

(北京邮电大学 信息与通信工程学院，北京 100876)

**摘　　要**: 本小组选取了 kaggle 中一道典型的图像分类题目，对数据进行简单预处理，利用迁移学习的思想，采用预训练的卷积神经网络 (Convolutional Neural Network, CNN) 提取图像的特征向量，使用前馈网络对特征向量进行分类，使用 dropout 技术防止过拟合，并在 Xception 模型的基础上引入了 InceptionV3 和 InceptionResNetV2 模型，最后采取加权平均的方式进行模型融合，排名能够达到 top1%。在解决问题的过程中，模型经历了多次迭代，数据处理和模型训练也采取了不同的技巧，文章对不同的方法和技巧做出了分析比较，说明了采用迁移学习和模型融合方法的优势。

**关　键　词**: 迁移学习；图像分类；卷积神经网络；模型融合

**中图分类号**: V221$^{+}$.3；TB553

**文献标识码**: A　　**文章编号**: 1001-5965 （XXXX）XX-XXXX-XX

由于卷积神经网络在机器视觉任务中的优秀表现，Kaggle 竞赛中与图像相关的题目，绝大多数参赛者采用的是深度学习的解决方案。本小组选择了经典的图像分类任务，在问题的解决过程中，熟悉了一般 CNN 网络模型的构建方式，尝试运用一些数据挖掘的技巧，提升模型的表现，最终模型分类的准确率令人满意，也在一定程度上减少了模型训练的代价。挖掘任务的详细解决过程陈述如下。

## 1　任务描述

### 1.1　竞赛题目

竞赛题目选用 Kaggle 竞赛中，类型为 Playground 的 Dogs vs. Cats Redux: Kernels Edition。题目提供已标注的训练集和未标注的测试集，要求参赛者提交测试集中每一张图片为狗的概率。

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

提交的结果用公式 (1)，即对数损失来衡量。式中，$n$ 是测试集的图片数量，$\hat{y}_i$ 是图片为狗的预测概率，图片实际为猫时，$y_i$ 取 0，否则取 1。对数损失越小，预测结果越好。

### 1.2　数据集

数据集分为训练集和测试集，jpg 格式。训练集包括 12500 张猫的图片和 12500 张狗的图片，大小为 543MB，分辨率不等，来自某个宠物领养的网站，推测为程序自动爬取获得，分类标签存在一定程度的错误。测试集包括 12500 张猫与狗的图片，大小为 271MB，来源与训练集相同。

## 2　问题定义

赛题是一个典型的图像分类问题，即根据一定的分类规则将图像自动分到一组预定义类别中。图像分类在多领域有着广泛的实际应用。实际应用中，图像分类任务会受到视角的变化，大小的变化，物体的形变，遮挡，光照条件，背景干扰和多种子类型等各方面因素[1] 的挑战。

收稿日期：2018-01-01；录用日期：2018-01-01
网络出版地址：（无）
基金项目：（无）
通信作者：E-mail：___@bupt.edu.cn

对于传统的数据挖掘算法，如 K 近邻算法 (K-Nearest Neighbor, KNN)，支持向量机 (Support Vector Machine, SVM)，图像分类任务是非常复杂甚至难以胜任的。卷积神经网络，以其在图像识别方面的优异表现，从诞生起就备受瞩目。因此，本小组选择 CNN 作为图像分类算法的核心部分，再利用迁移学习的思想，减少计算量和过拟合的风险。最终将数据挖掘任务划分为以下三个子任务：（1）图像数据的预处理；（2）数据特征的提取；（3）特征处理，对提取出的特征进行分类。

# 3 解决方案

经过多次实验和迭代，本小组最终提出了对图像仅进行简单预处理，以多个卷积神经网络作为特征提取器[2]，训练前馈神经网络 (Feedforward Neural Network, FNN) 作为分类器，最后进行模型融合的解决方案。

## 3.1 方案提出

较深的卷积神经网络的训练和计算，需要相当大的计算能力。Google 团队 2016 年提出的 Xception 模型，使用了 60 块 Nvidia K80 GPUs，在 ImageNet 数据集上训练了 3 天[3]，达到 79% 的 Top1 准确率。小组最终采用的模型为 Inception 系列的三个深度均达到 100 层以上的 CNN，而拥有的计算资源仅仅是笔记本上的入门级 GPU GTX 1050m，这对计算资源提出了不小的挑战。

此外，虽然相对于小组拥有的计算资源，题目所给训练集数据量较大，但是对于深度神经网络的训练，提供的数据又稍显不足，容易出现过拟合的情况。针对这些问题，在查阅相关资料后，决定采取迁移学习的方式解决。在深度学习的分类器之外，小组也采用了经典的 SVM 分类器与 FNN 作出对比。由于采用提取特征向量的方式，需要数据集相对固定，因此难以在短时间内尝试多种数据预处理方式。本小组把模型调整的重心放在网络的后半部分，节约出训练网络的时间来对模型做出优化。因此，形成了简单预处理 →CNN 提取特征向量 →FNN 作为分类器 → 模型融合的方案结构。

以下将从 CNN 网络、迁移学习和模型融合三个方面来进行算法的介绍。

## 3.2 Inception 系列网络

方案选取了在 ImageNet 数据集中，表现优秀的 Inception 系列网络，作为特征提取部分的模型。Inception 系列网络是 Google 团队提出的一系列不断改进的 CNN 模型，名字来源于电影盗梦空间 (Inception) 中的一句台词：“We need to go deeper”。经典的 CNN 网络，由负责特征提取的卷积层和负责降采样的池化层交替堆叠而成。Inception 系列网络在经典 CNN 结构的基础上，改进了卷积层的结构，取得了一定的性能提升。

CNN 中的卷积核大小，很大程度上影响着提取到的特征图像。一般来说，卷积核越大，感受野越大，越能发现全局的图像特征；卷积核越小，对图像局部的特征更加敏感[4]。传统的 CNN，每一层采用的是一组大小相同的卷积核。为了加强网络的泛化能力，对不同尺度的目标都能良好地识别，InceptionV1 提出在同一层采用多种大小的卷积核，将卷积得到的特征向量进行拼接，作为下一层的输入。这就给下一层网络提供了不同尺度上的特征向量信息，使网络能够自由选择，提高了网络的泛化能力。

多个卷积核带来了性能的提升，也带来了极大的计算量。为了减少网络的计算，InceptionV1 提出了 1×1 的卷积核。1×1 的卷积核并不会改变特征图像在空间上的相对关系，而是把同一空间位置上的多个通道融合在一起，通过减少输出特征图像的通道数量，达到减少计算量的目的。

为了进一步减少神经网络中的计算，Inception 用两个 3×3 的卷积核代替了一个 5×5 的卷积核，网络参数由 25 个下降到 18 个，还进一步增加了网络的深度。更重要的是，卷积核的感受野是不变的，仍然能够包含 5×5 的特征图像区域。

此外，1×n 和 n×1 的网络也被引入到 Inception 网络结构中。

Xecption 网络则在另一个角度上对传统 CNN 的结构作出了改进。区别于传统的卷积核，Xception 网络使每一个卷积核仅仅进行一个通道的二位卷积操作，这就将空间上的相关性与通道之间的相关性分离开来。

## 3.3 迁移学习

对于本分类任务而言，由于数据量较大，如果是直接在一个巨大的网络后面加全连接，那么每计算一次前向传播就需要 20 分钟以上，而且卷积层都是不可训练的，会造成计算资源的浪费，因此我们应用迁移学习 (transfer learning) 来降低计算开销[5]。下面对这个方法进行一下简单介绍：

迁移学习，顾名思义是把 B 域中的知识迁移到 A 域中，以提高 A 域的分类效果。它作为一种新的学习范式，吸引了众多研究者的注意。对迁

移学习的研究来源于一个观测：人类可以将以前的学到的知识应用于处理新的问题，以获得更好的效果。换句话说，迁移学习被赋予这样一个任务：从一个或多个源任务中抽取知识、经验，然后应用到一个新的目标领域当中，从而更快更好地解决问题。

```
 (/gpu:0) -> (device: 0, name: GeForce GTX 1050, pci bus id: 0000:01:00.0)
 - 10s - loss: 0.0767 - acc: 0.9828 - val_loss: 0.0324 - val_acc: 0.9897
Epoch 2/12
 - 4s - loss: 0.0266 - acc: 0.9916 - val_loss: 0.0218 - val_acc: 0.9935
Epoch 3/12
 - 4s - loss: 0.0239 - acc: 0.9922 - val_loss: 0.0171 - val_acc: 0.9942
Epoch 4/12
 - 4s - loss: 0.0204 - acc: 0.9939 - val_loss: 0.0223 - val_acc: 0.9925
Epoch 5/12
 - 4s - loss: 0.0185 - acc: 0.9937 - val_loss: 0.0180 - val_acc: 0.9948
Epoch 6/12
 - 4s - loss: 0.0158 - acc: 0.9952 - val_loss: 0.0179 - val_acc: 0.9932
Epoch 7/12
 - 4s - loss: 0.0146 - acc: 0.9950 - val_loss: 0.0145 - val_acc: 0.9947
Epoch 8/12
 - 4s - loss: 0.0145 - acc: 0.9950 - val_loss: 0.0133 - val_acc: 0.9952
Epoch 9/12
 - 4s - loss: 0.0160 - acc: 0.9943 - val_loss: 0.0164 - val_acc: 0.9941
Epoch 10/12
 - 4s - loss: 0.0133 - acc: 0.9956 - val_loss: 0.0124 - val_acc: 0.9961
Epoch 11/12
 - 4s - loss: 0.0130 - acc: 0.9956 - val_loss: 0.0134 - val_acc: 0.9948
Epoch 12/12
 - 4s - loss: 0.0115 - acc: 0.9956 - val_loss: 0.0128 - val_acc: 0.9963
Train on 39956 samples, validate on 9990 samples
```

图 1 训练过程

Fig. 1 training process

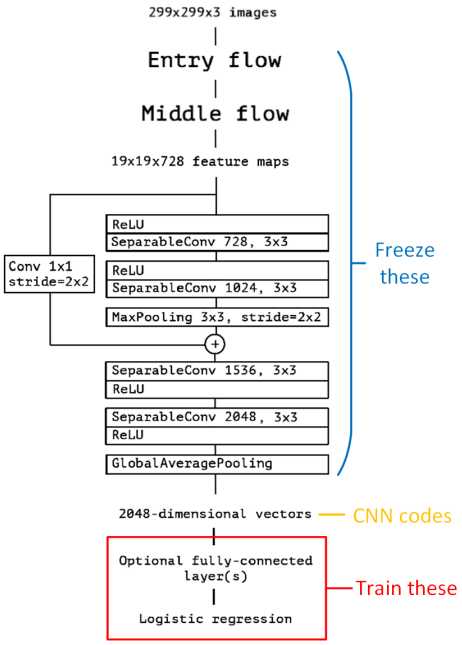


图 2 迁移学习示例

Fig. 2 example of transfer learning

基于迁移的内容，迁移学习可分为如下几类——基于实例的迁移学习：源领中的数据的某一部分可以通过调整权重的方法重用，用于目标域的学习；基于特征表示的迁移学习：通过源域学习一个好的特征表示，把知识通过特征的形式进行编码，并从源域传递到目标域，提升目标域任务效果；基于参数的迁移学习：目标域和源域的任务之间共享相同的模型参数或者是服从相同的先验分布；基于关系知识的迁移学习：假设源域和目标域中，数据之间联系关系是相同的，从而进行相关领域之间的知识迁移。

这之中基于关系知识和参数的迁移学习适用于本项目分类任务，如图 2具体为固定预训练模型的前半部分网络参数，只训练 CNN 网络的后半部分，这样无论迭代次数为何，前半部分网络对相同的图像只需经过一次前向传播的计算，大大减少了训练时间。在提取出特征向量之后，仅需要对 FNN 分类器进行短时间的训练就可以达到良好的训练效果。如图 1所示，Xception 网络提取出的特征向量，仅仅需要 54s 就能完成 12epochs 的迭代，并且在验证集上达到了 99.63% 的准确率。

### 3.4 模型融合

集成学习 (ensemble learning) 通过构建并结合多个学习器来完成分类任务，其通过将多个学习器进行结合，常可获得比单一学习器显著优越的泛化性能[6]。在本项目中，提高模型表现的一种有效方法是综合各个不同的模型（学习器），从而兼听则明，得到更为出色的效果。学习器的结合策略包括平均法、投票法和学习法，本小组采用最常见的平均法来对不同模型的数值型输出进行融合。平均法中的加权平均法 (weighted averaging) 为：

$$H(x) = \sum_{i=1}^{T} \omega_i h_i(x) \tag{2}$$

其中 $\omega_i$ 是个体学习器 $h_i$ 的权重，通常要求 $\omega_i \geq 0$ 且 $\sum_{i=1}^{T} \omega_i = 1$。而简单平均法是加权平均法令 $\omega_i = 1/T$ 的特例。加权平均法在集成学习中具有特别的意义，集成学习中的各种结合方法都可视为其特例或变体：对给定的基学习器，不同的集成学习方法均可等效为通过不同的方式来确定加权平均法中的基学习器权重。这里的权重来自于训练数据，可由正比于各训练模型的准确率或反比于差错率两种方式得到，本小组采用第一种方法来得到各模型的加权权重，实践证明，此种方式取得了较为满意的性能提升。小组最终采用的模型如图 3所示。

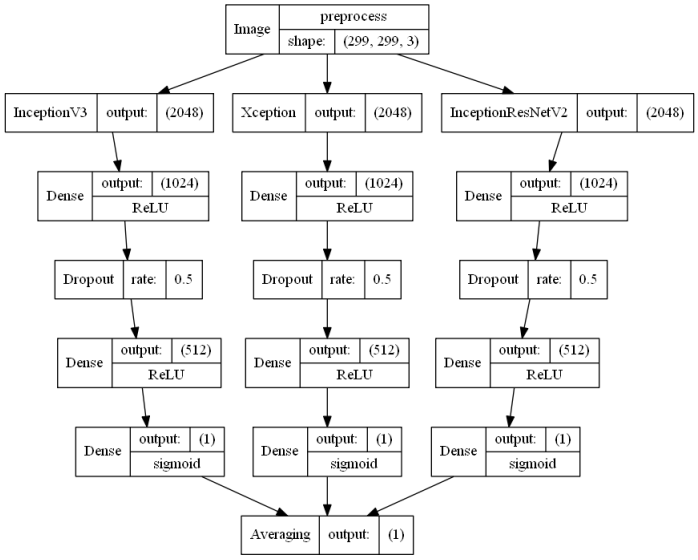


图 3 最终方案

Fig. 3 final solutuon


图 4 预处理结果

Fig. 4 result of preprocessing

# 4 数据处理

## 4.1 数据预处理

作为特征提取器的 CNN 网络需要 299×299 像素、RGB 通道的图像输入。此外，为了加强神经网络的性能，防止过拟合，对图像进行一定程度的数据增强是相当有必要的。为此，需要先对数据做一些预处理工作。

训练集中的图像比例不均，若简单将其拉伸为正方形，则会使物体发生形变。因此需要为图像填充背景后拉伸，以保持图像的比例。背景颜色对图像来说是人为引入的噪声，需要将其影响降到最低。Inception 系列网络在输入图像时，会利用公式 $x = (x/255 - 0.5) \times 2$ 对每个像素进行正则化，映射到 $[-1, +1]$ 范围内。因此，将背景色填充为 $(127, 127, 127)$，使其正则化之后最接近 0，能在最大程度上减少背景色对数据的影响。

此外，本小组还尝试了几种数据增强的方式：图像翻转，椒盐噪声，直方图均衡。

图像的翻转、旋转是 CNN 图像处理中最常见的数据增强方式，简单高效，能够有效减少神经网络的过拟合风险。但随着数据量的增大，所需的计算量也随之增加。考虑到现实中（以及测试集中）猫狗图像一般不会有上下颠倒的情况，仅仅采取了水平翻转的方式，实际训练中，在验证集上取得了一定的准确率提升。

添加椒盐噪声，是指按一定的信噪比随机在图片中添加纯白或纯黑像素点。小组尝试了 0.8 和 0.9 的信噪比，发现预测准确率均存在一定程度的下降。推测是椒盐噪声虽然增加了网络的鲁棒性，然而也破坏了图片中携带的信息，使得网络难以辨认图像，从而影响预测的结果。此外，CNN 网络是固定不变的，椒盐噪声带来的鲁棒性提升很难被网络所利用。因此，此方案最后并未被采用。

直方图均衡是为了解决部分图像亮度条件较差的问题而采用的方案。它的原理是统计彩色图像中 RGB 值的累积概率函数，再将其线性化，使得像素点的取值在 $[0, 255]$ 之间接近均匀分布，从而增加图像全局的对比度。自适应直方图均衡克服了直方图均衡的缺陷，将图片划分为多个子块，在子块上平衡累积概率分布，从而避免了图像分布被整体移动、背景噪声被加强的缺陷。小组尝试了直方图均衡和自适应直方图均衡两种方式，在实际训练中，并未发现此方法对预测结果有提升。由于预处理之后需要重新进行特征向量的提取，在直方图均衡效果不明显的情况下，将此方案舍弃。

图像随机切割也是一种较好的数据增强方式，考虑随机切割会极大增加数据量，并且极有可能将目标物体排除（一些图片中猫狗的体积非常小），因此并未采用。随机调整对比度、亮度的方案也因为极大地增加了计算量，并未实际采用。

## 4.2 特征向量提取

小组采用在 ImageNet 数据集中，预训练的 InceptionV3，Xception，InceptionResNetV2 网络，

取最后一层卷积层的输出，经过全局平均池化（即对输出特征图像的每个通道取平均值，三个网络都是 2048 个通道）后，作为提取出的特征向量。

ImageNet 数据集包含至少 1000 类物体的图像，并且三种 Inception 网络为了迁移学习的需要，在网络最后一层均保留了 1000 类的全连接层，所以最后的卷积层得到的特征向量具有良好的泛化能力。网络将同时提取训练集和测试集的特征向量，训练集的特征向量用于训练 FNN 分类器，测试集的特征向量则用于给出最终的预测结果。

利用本组现有的计算资源，每一个网络提取特征向量的时间平均 30 分钟，相当于在网络中计算一遍所有图片的前向传播所花费的时间。如果按照通常的训练策略，将所有图片在网络中迭代十次，每个网络，前向传播加上反向传播的时间，可以预计将超过 $10 \times 2 \times 30 = 10h$，远远不如单独提取特征向量的运算效率。

### 4.3 FNN 分类器

实际使用的 FNN 分类器是一个两层的前馈神经网络，为了防止过拟合，第一层的输出经过 Dropout 之后，再输出到第二层网络。理论上，取得特征向量之后，所采用的分类器可以自由选择，为了与 FNN 做对比，小组采用了传统的 SVM 分类器进行分类，效果不如 FNN。鉴于 FNN 能够自由调节网络结构，并且在数据量较多时，FNN 的分类结果比 SVM 更为准确[7]，采用 FNN 作为模型的分类器。

通过迭代次数和网络参数的调节，可以有效地防止 FNN 出现过拟合，另外，采用验证集，结合 early stopping 技术，也可以使网络达到局部最优的训练结果。经过多次训练和调试，选择第一层 1024 个神经元，第二层 512 个神经元，50% 的 Dropout 比例为参数，20% 的数据划分为验证集，为每个网络分别构建 FNN 分类器，迭代次数根据 CNN 表达能力的强弱进行了相应调整，梯度下降算法是 adam，可以自适应地改变学习率。

除了为每个网络独立分配 FNN 分类器之外，如图 5，小组也尝试了将三个网络提取出的特征向量直接拼接，再用同一个 FNN 分类器进行分类的方案，发现方案的效果不及模型融合，为了得到模型融合的优势，采用了多个分类器的方式。

### 4.4 模型融合

简单起见，小组采用的是加权求平均的方式，将三个网络的预测值，按照其在 ImageNet 数据集中的 Top1 准确率进行加权平均。相比仅仅拼接特征向量的方式，模型融合对预测结果有可观的提升。最终网络在本地验证集上的准确率，甚至可以达到 99.8% 左右（不排除出现了过拟合现象）。

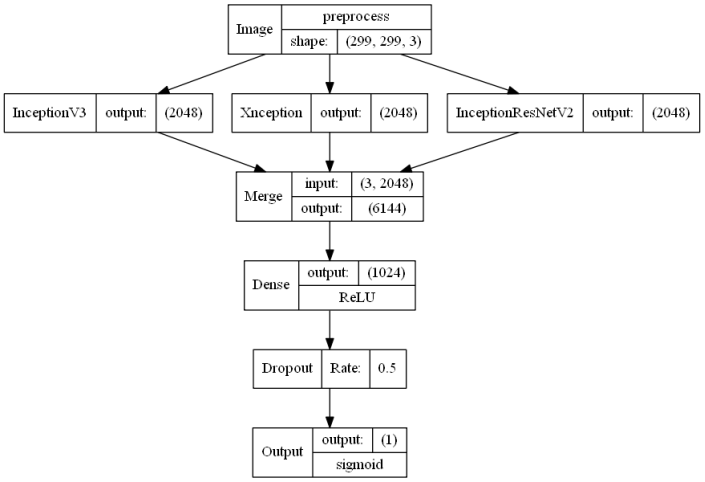


图 5 合并特征向量的方案

Fig. 5 solutuon of combining feature vectors

## 5 结果分析

### 5.1 预测结果

表 1是各个模型预测结果提交后的损失得分和最终排名（由于比赛已经结束，排名参考的是 public leaderboard 中的 LogLoss 值）。分析不同模型之间的差距，可以简单地得到一些结论：

**表 1 预测结果**

**Table 1 result of prediction**

| 模型 | 对数损失 | 排名 |
|---|---|---|
| Xception+SVM | 0.05709 | 102 |
| Xception+FNN | 0.04475 | 28 |
| 双模型 +FNN | 0.04099 | 18 |
| 三模型 +FNN | 0.03941 | 15 |
| 最终方案 | 0.03842 | 12 |

1)SVM 与 FNN 的差距，在于模型所采用的 SVM 是硬分类器，而 FNN 是软分类器，因此在分类出错时，SVM 受到损失函数的惩罚会更严重；此外，FNN 在表达能力和灵活性方面都要优于 SVM 分类器，样本数量较大时 FNN 也占优势；

2) 表 1 第三和第四行中采用的模型，是用拼接特征向量的方式组合到一起的，也就是说，每增加一个模型，特征向量的维度都会增加 2048，从而获得更全面的特征信息。然而，拼接的特征向量仍然是采用同一个 FNN 分类器进行训练的，难以回避过拟合的风险；

3) 模型加权平均在多模型的基础上为每个 CNN 特征提取器训练不同的 FNN 分类器，降低了过拟合的风险，还可以对不同的 FNN 采取不同的训练策略，小组采用了 12epochs、20epochs 和 12epochs 的迭代次数分别对 Xception，InceptionV3 和 InceptionResNetV2 进行了训练，弥补了 InceptionV3 准确率不及另外两个网络的缺陷。

### 5.2 误差分析

CNN 在图像识别任务上有出色的表现，但是其运行机理更像是一个黑箱，经过多重网络的乘加等运算，很难定量分析误差的情况。为此，小组从预测结果出发，试图从一个更为直观的角度分析误差产生的原因。


图 6 准确率较低的预测集图像

Fig. 6 images in test set with low accuracy

编写脚本，从预测结果中挑选出概率在 $[0.2, 0.8]$ 区间内的图像，部分图像如图 6所示，共得到 134 张，通过手工筛查，发现图像有以下特征：

1) 错误样本：在训练集和测试集中，均存在无法进行分类的图像，如 “Photo Unavailable”、纯色背景、猫狗在同一场景中，非猫狗物体等；2) 目标过小：存在目标物体占图像整体比例少于 10% 的情况，导致网络难以识别；3）分辨率低：部分测试集图像是由极低分辨率 (如 $[60, 44]$) 拉伸而成，与 $[299, 299]$ 的网络输入差别较大，无法准确预测；4) 部分图片中的猫狗物体有非正常的拍摄角度，如上下倒置，这类角度在训练集中极少出现，导致网络难以预测这部分图片的分类情况。

### 5.3 改进方案

由于时间原因，模型及算法尚有改善空间，具体可在以下几个方面进行优化：

1) 使用其他预训练模型（如 VGG）进行微调；2) 网络调优，可以在已有数据的基础上训练网络的后几个卷积层甚至整个网络；3) 使用异常检测[8] 的方法检测出错误及非常规图像，而后单独标定；4) 在得到的特征向量的基础上尝试聚类、随机森林等神经网络之外的数据挖掘方式……这些待完善的地方还会在后续进行探索。

此外，针对误差分析中发现的情况，可以采用相应的数据增强方式（如上下倒置，随机放缩），来提升网络在测试集上的预测效果。

## 参考文献（References）

[1] Image Classification. Retrieved January 1, 2018, from CS231n Convolutional Neural Networks for Visual Recognition: http://cs231n.github.io/classification/.

[2] Transfer Learning. Retrieved January 1, 2018, from CS231n Convolutional Neural Networks for Visual Recognition: http://cs231n.github.io/transfer-learning/.

[3] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[J]. arXiv preprint arXiv:1610.02357, 2016.

[4] 周俊宇, 赵艳明. 卷积神经网络在图像分类和目标检测应用综述 [J]. 计算机工程与应用,2017,53(13):34-41.

[5] Pan, S. J. and Q. Yang. (2010). “A survey of transfer learning.” IEEE Transactions on Knowledge and Data Engineering, 22(10):1345-1359.

[6] 周志华著，机器学习，清华大学出版社，2017.3.

[7] Yang, Shaomei, and Qian Zhu. “Research on comparison and application of SVM and FNN Algorithm.” Wireless Communications, Networking and Mobile Computing, 2008. WiCOM’08. 4th International Conference on. IEEE, 2008.

[8] Chandola, V., A. Banerjee, and V. Kumar. (2009). “Anomaly detection: A survey.” ACM Computing Surveys，41(3): Article 15.

# Transfer Learning for Image Classification of Cats and Dogs

DENG Xinyi, JIANG Baichun, LIU Jiangwang

(School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

**Abstract:** The team selected a typical image classification problem in kaggle, firstly, preprocessing the data , then with the idea of transfer learning, extracted the feature vector of the image using pretrained convolutional neural network. Feedforward neural network is used for classification of the feature vector. What's more, InceptionV3 and InceptionResNetV2 are introduced on the basis of Xception. Finally, the three models are merged by weighted averaging, therefore our ranking can reach top 1%. Through out the process of problem solving, the model has undergone many iterations, and a variety of methods are applied to data preprocessing. Different methods and techniques have been analyzed and compared, to illustrate the advantages of transfer learning and model ensembling.

**Key words:** transfer learning; image classification; convolutional neural network; model ensembling

**Received:** 2018-01-01; **Accepted:** 2018-01-01
**URL:** (None)
**Foundation item:** (None)
**Corresponding author.** Tel.: 010-8231xxxx E-mail: ___@bupt.edu.cn