机器学习工程师纳米学位毕业项目

基于 inception-resnet-v2 的猫狗识别

王飞

2017 年 4 月 28 日

目录

1 定义

1.1项目概述

项目训练一个从自然图像中识别猫狗的深度学习模型，模型应用到微信公众号中，从自然图像中，实时判断是猫还是狗。

项目来源于 kaggle 里面的 Dogs vs. Cats 比赛。对于人来说，判断图片中是否包含狗或猫是很容易的事情，但是对于电脑，还是有点难度的。尽管现在有很多图片分类算法，但是能够从自然图片中做出准确分类，还是很有挑战的。

项目使用的算法是深度卷积神经网络。自 2012 年，Alex Krizhevsky 提出卷积神经网络模型 AlexNet 以来，深度卷积神经网络在图像识别大放异彩，准确率不断提高，项目采用的是 Inception-resnet-v2 模型[1]，在 ImageNet 分类（CLS）挑战的测试集上获得 3.08％的 top-5 误差，16.5%的 top-1 误差，在多种分类中表现优异，同样算法适用于猫狗 2 分类的问题。

项目选择的数据集是 kaggle 里面 Dogs vs. Cats 比赛的数据集，包含 12500 张狗的图片和 12500 猫的图片，数据集在图片的名称上标注类别，测试集同样来自 kaggle 的 12500 张图片。项目根据 inception-resnet-v2 模型，用 keras[2]构造了神经网络，用来解决猫和狗分类问题，并且用 ResNet50，Xception 预训练模型提取的特征，重新训练模型，提升准确率，优化项目。

1.2问题陈述

项目解决的问题是自然图片中猫或狗的分类问题。首先是模型的训练，训练完模型之后，用于图片的分类。项目采用微信公众号的方式输入图片，服务端接收到图片之后，运用训练好的模型预测猫或狗的分类，并把分类结果返回给公众号。

项目是一个 2 分类问题。输入是有像素构成的任意像素的图像，目标是将图像分为猫或狗两类。

项目的难点在于，自然图像的质量问题，首先是公众号作为输入源，不同的

手机拍摄的图片的大小不同，不同时间拍摄的光照强度不同，环境和人为因素带来的干扰因素比较多，都会增加识别的难度。

1.3评价指标

模型和公众号应用评价指标是，评价指标一个是分类的准确率（Accuracy），即正确分类的图片/总的图片。另一个是来自 kaggle 的 log loss。

项目限定是图片的 2 分类问题，只是针对单张图片只包含一种类别的分类，项目不涉及同一张图片出现 2 种类别的判断。

2 分析

2.1 数据可视化

数据集分为训练集，测试集，验证集，以及现实生活中拍摄的图片。训练集，验证集和测试集来自 kaggle，部分测试图片来自现实生活中拍摄的图片，数量较少。训练集和验证集 25000 张，测试集 12500 张，公众号测试图片来自手机拍摄的少量图片和互联网上的图片。

训练集和验证集图片的名称前缀表示目标的类别，比如，dog_00001.jpg 表示是狗。图片为 jpg 格式。如图 1，图 2 所示。

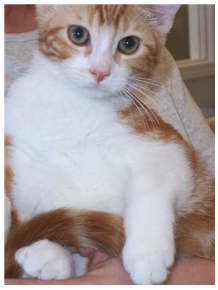

图 1. cat.9.jpg，320 × 425

图 2. dog.134.jpg. 399 × 269

从图片可以看出，每张图片的像素大小不同，由于环境因素，有很多干扰的物体，草地，球场等。

测试集只是包含图片编号，例如 1.jpg，2.jpg 等。如图 3，图 4 所示。

图 3. 1.jpg

图 4. 2.jpg

2.2 算法和技术

2.2.1 神经网络

人工神经网络（Artificial Neural Network，即 ANN）是 20 世纪 80 年代人工智能领域兴起的研究热点。它从信息处理的角度对人脑神经元进行抽象，按照不同的连接方式组成不同的网络。神经网络是一种计算模型，由大量节点之间连接构成，每一种节点代表一种特定的输出函数，成为激励函数。每两个节点的连接有相应的权重，相当于神经网路的记忆。网络的输出由连接的不同方式，权重值和激励函数决定，而网络自身通常对某种算法的函数的逼近，也可能是对一种逻辑策略的表达[3]。

神经网络的可以分成 2 种网络结构。一种是前馈神经网络（图 5），一种是递归神经网络。前馈神经网络的输入层源节点提供激活模式的元素（输入向量），组成第二层（第一隐藏层）神经元的输入信号，第二层的输出作为第三层的输入，这样一直传递下去。项目中采用的算法是一种前向神经网络。神经网络的层数又叫深度，多层神经网络就是深度学习。

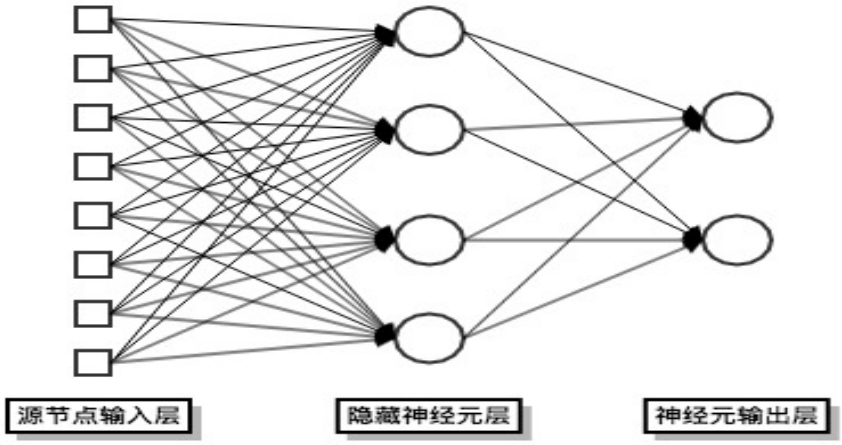

图 5.具有一个隐藏层和输出层的全连接前馈网络

2.2.2 卷积神经网络

卷积神经网络是一种前馈网络。它的提出所隐含的思想受到了神经生物学的启发，可以回溯到 Hulel and Wiesel（1962，1977）的开创性研究，该研究是关于猫视觉皮质上局部传感和方位选择神经元。

卷积神经网络在特征提取时，每一个神经元从上一层的局部接受域得到突触输入，因而一旦一个特征被提取出来，只要他相对于其他特征的位置被近似的保留下来，它的精确位置就变得不那么重要了。在特征映射时，有平移不变性，通过权重共享缩小自由参数数量。在子抽样时，每个卷积层跟着一个实现局部平均和子抽样的计算层，由此特征映射的分辨率降低，这种操作具有特征映射输出对平移和其他形式的变形的敏感度下降的作用。

卷积，7×7 的输入图像，经过 3×3，步长为 2 的卷积，转化为 3×3 的特征图，卷积的目的提取输入的不同单元的特征。算法使用二维卷积层。如图 6 所示。

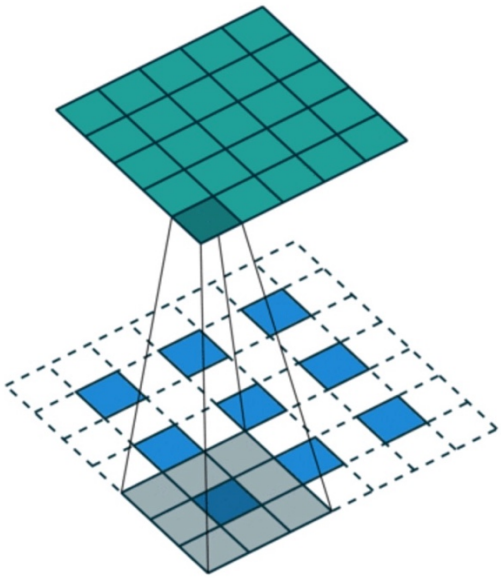

图 6.卷积[4]

池化，输入 4×4 经过每隔 2 个元素最大池化，得到 2×2 的特征。最大池化将输入图像划分若干个矩形区域，对每个区域输出最大值。池化层会减小数据空间的大小，参数和计算量也会下降。在一定程度上控制过拟合。当然也有其他的池化，取决于非线性池化函数，项目算法采用平均池化和最大池化。如图 7 所示。

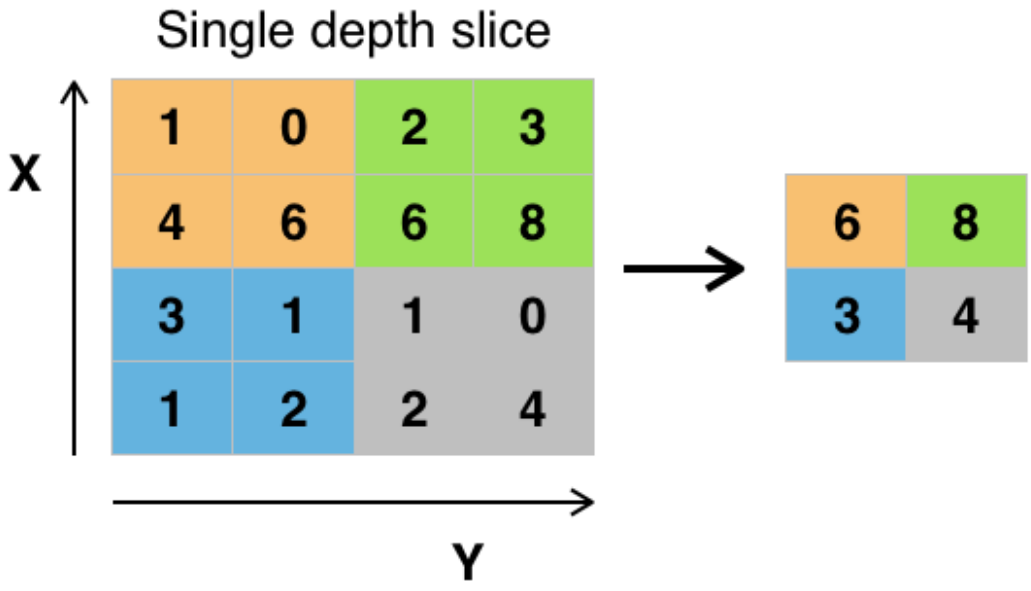

图 7.池化[5]

损失函数，categorical_crossentropy，多类的对数损失，评估真实值和预测

值的差异。使用时需要将标签转化为二值序列（samples，classes）。

优化器 Adam，实现了 Adam 算法的优化器， 它利用梯度的一阶矩估计和二阶矩估计动态调整每个参数的学习率。Adam 的优点主要在于经过偏置校正后，每一次迭代学习率都有个确定范围，使得参数比较平稳，适用于大数据集和高维空间。

dropout，每次参数更新时,保留一些输入神经元，参数 rate 是输入神经元保留的可能性。防止因所有特征选择器公共作用一直放大或缩小某些特征。防止过拟合，增强泛化能力。

2.2.3 Inception-resnet-v2 模型

结合三个残差和一个 Inception-v4， 在 ImageNet 分类（CLS）挑战的测试集上达到 3.08％的前 5 个错误。数据表明，具有残差连接的训练可以显著加速 Inception 网络的训练,而且有更好的性能表现。

项目使用 Inception-resnet-v2 模型结构训练了分类模型。结构如图 8 所示。

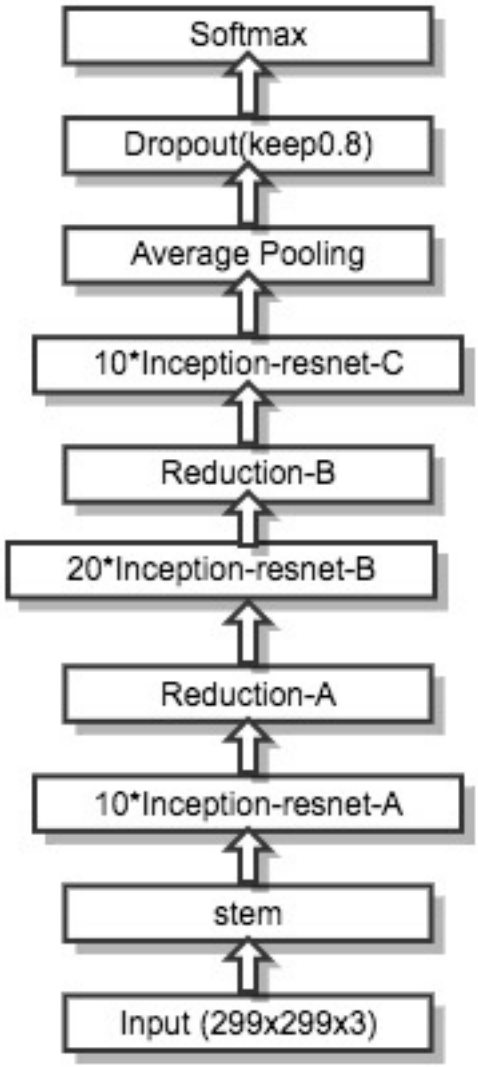

图 8.Inception-resnet-v2 模型结构

项目按照参考论文的算法以及 Inception-resnet-v2[6]项目的部分实现，具体处理流程如下：

预处理：输入为三维张量，预处理之后为（299，299，3）。

stem：过滤特征，经过 stem 输出（35，35，384）。

Inception-resnet-A：残差网络，加速训练，提升性能。输出（35，35，384）。

Reduction-A：调整各个 filter 的大小，过滤特征。输出（17，17，1152）。

Inception-resnet-B：残差网络，同时降维分解。输出（17，17，1152）。

Reduction-B：多个不同步长的卷积和最大池化连接，发现特征。输出（8，8，2144）。

Inception-resnet-C：残差网络，同时降维分解。输出（8，8，2144）。

Average Pooling：均值化特征，防止过拟合。输出(1， 1，2144)。

Dropout：保留一定比例的卷积层，防止过拟合。输出(1， 1，2144)。

Softmax：分类结果，输出 2 类。

Inception-resnet-A，Inception-resnet-B，Inception-resnet-C 结构设计借鉴了 ResNet 网络设计的思想。ResNet 网络有 2 中形式，一种是主 path 和 shortcut 都有卷积，一种是主 path 有卷积，shortcut 没有卷积。以上三种结构使用的是第二种设计形式。Inception-resnet-B，Inception-resnet-C 主 path 的卷积层设计有降维的思想。Inception-resnet-B 使用了 1×7 和 7×1 的卷积，Inception-resnet-C 使用 1×3 和 3×1 的卷积，降维可以降低参数，增加深度，加速计算。

2.2.4 技术

项目中采用的开源框架 keras，它是一个高层神经网络 API，由 python 编程，并且基于 Tesorflow 或 Theano，能够快速构建模型。

keras 具有简易和快速的原型设计，对 Tesorflow 透明可见，支持 cnn 和 rnn，或者两者的结合，无缝切换 cpu 和 gpu 等特点，支持快速实验，能够快速的把想法转化为结果，降低了开发人员的编程难度。后端采用的 Tesorflow 框架，是 google 开源的神经学习框架，能够利用多核 cpu 和 gpu，提升训练速度。

2.3基准指标

Inception-resnet-v2 在 ImageNet 分类（CLS）使用 144 层隐藏层的卷积神经网络实现了 83.5%的 top1 的准确率，实验表明，隐藏层数越多，对分类结果越有利，但是层数增加将会带来训练时间的增加。项目同样采用 144 层隐藏层的卷积神经网络，但是类别只有两种，在 AWS 的 px2.large 上，用 GPU 训练 34 个小时。

考虑到卷积神经网络的训练难度，以及算法论文中的分类类别和本项目有很大差异，因此本文设定准确率目标是 90%，也就是说，模型在测试集上的表现是正确率大于 90%，公众号的应用体现的分类结果真确率也是 90%以上，模型 logloss 值小于 0.3。

3 具体方法

3.1数据预处理

把所有训练集图片缩放为 299×299 大小的图片，按照图片的命名规则，提取猫狗的分类标签。狗标记为 1，猫标记为 0。

包含 25000 张图片，猫和狗各有 12500 张，训练模型时，采用 20%的图片作为验证集，因此，训练集 20000 张，验证集 5000 张。由于采用的是三维张量，颜色不需要特别处理。

4 结果

4.1模型评价与验证

从训练结果看，30 个 epoch 后，训练集 acc 达到 97.86%，并趋于稳定，后面 20 个 epoch 提升大约 1%。训练集 loss 在 30 个 epoch 之后 0.0567，趋于稳定，在后面 20 个 epoch 减少 0.02，验证集 acc 在 22 个 epoch 达到 90.34%，后面 28 个 epoch 提升 2%左右。验证集 loss 在 22 个 epoch 达到 0.2231，后面 28 个 epoch 趋于稳定。训练模型时，回调函数只是保存最高的验证集时的模型数据。

在验证集上，当 epoch=44 时，准确率是 0.92820，测试集在 kaggle 上的得分是 0.25167。在 50 个 epoch 中，当验证集的准确率得到提升时，模型参数记录下来，也就是说，当训练过程中产生过拟合，验证集的准确率下降时，模型参数并不会记录下来。训练过程如图 9 所示。

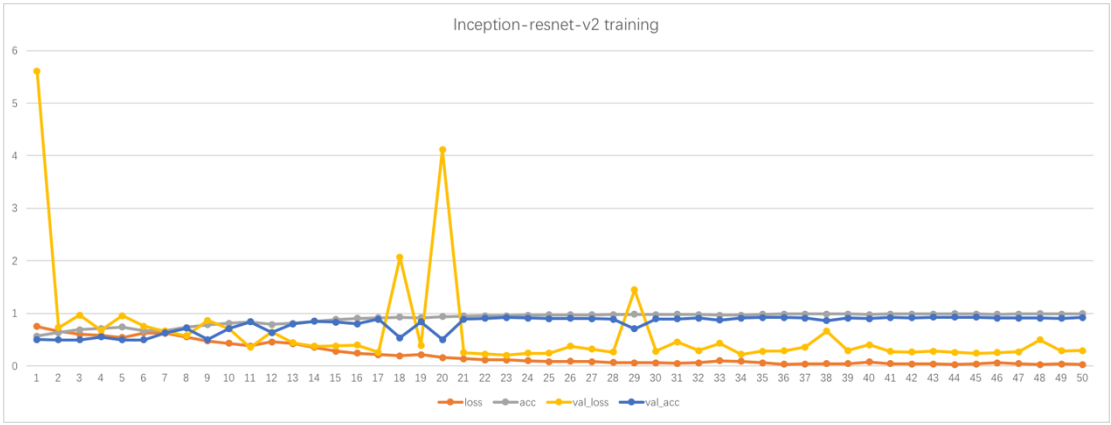

图 9.模型训练曲线

4.2结果分析

4.2.1 模型分析

训练完成模型之后，我们从验证集中选取 5 张分类正确的图片和 5 张分类错误的图片对比。如图 10 所示。

图 10.验证结果分析

第一排的五张照片分类错误，中间的那一张外形和毛发特征比较像猫，图像比较模糊，第一张同时出现猫和狗，无法分辨到底是哪一种，第二张狗的面部特征比较多，其他特征较少，第四张猫所占比例较大，特征不完整，第五张几乎和狗的一致。

第二排的五张照片分类都正确，这些图片比较清晰，特征比较明显，人眼很容易识别出来。

错误分类，说明卷积神经网络把目标的大部分特征都识别出来，权重比较平均，相对于人眼识别，我们把目标的面部特征权重设置的比较大，也就是通过面部特征就能判断是哪一类，所以对于算法，还有优化的空间。

4.2.2 应用识别分析

公众号对图像的识别，整体来说，从拍摄到结果返回，大约 1-2 秒左右，时间较短，准确性由训练的模型决定，因此，模型的准确率对公众号的服务能力来说至关重要。如图 11 所示。

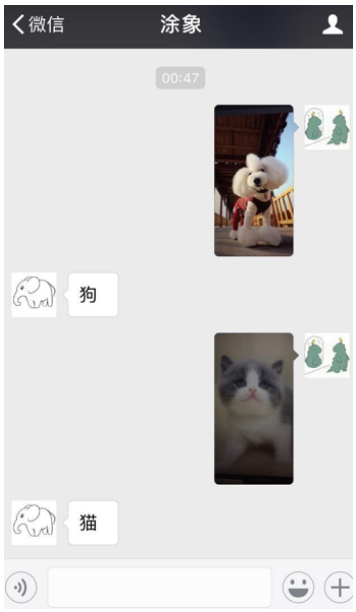

图 11.公众号识别猫狗

4.3改进

4.3.1 迁移学习

迁移学习是把训练好的模型参数迁移到新的模型来帮助新模型训练数据集。用组合特征向量，重新构建模型。新的模型的输入就是特征向量的组合。不同模型导出的特征向量从不同角度描述了分类特征的特点，重新组合之后，能够加速训练速度，很快收敛到稳定的状态。组合就是增加特征向量的维度。

根据 ResNet50，重新训练猫狗图片，导出猫和狗的特征向量，同理 Xception 也导出猫和狗的特征向量，然后结合 Inception-resnet-v2 导出的特征向量，建立全连接网络。ResNet50 直接预测准确率在 70.1%，Xception 直接预测在准确率在 86.4%。经过重新训练数据之后，ResNet50 特征向量直接 dropout，准确率 97%，Xception 特征向量直接 dropout，准确率 97%，ResNet50 和 Xception 特征向量全连接网络，准确率在 99.6%。再结合 Inception-resnet-v2 导出的特征向量，准确率有略微提升。结构如图 12 所示。

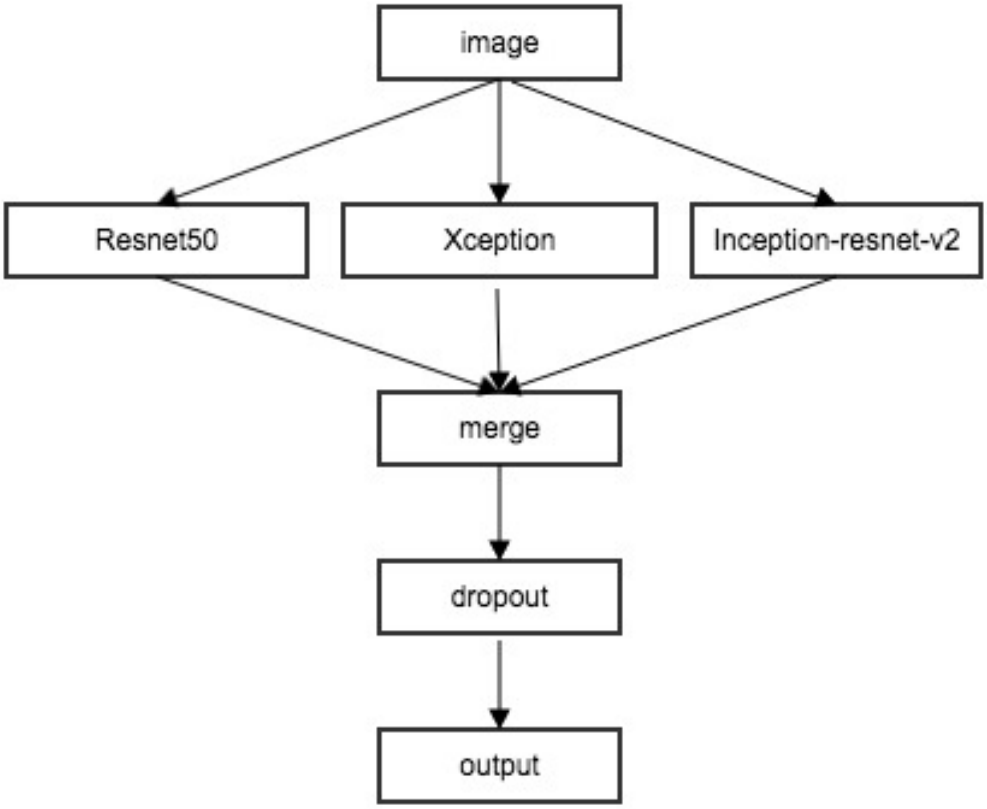

图 12.预训练模型全连接网络

训练全连接网络，经过 8 个 epoch 之后，准确率达到了 99.76%，提升了 7 个百分点。而且训练时间大大降低，在 kaggle 的得分是 0.04948，比之前降低了 0.2，是一个不错的成绩。训练过程如图 13 所示。

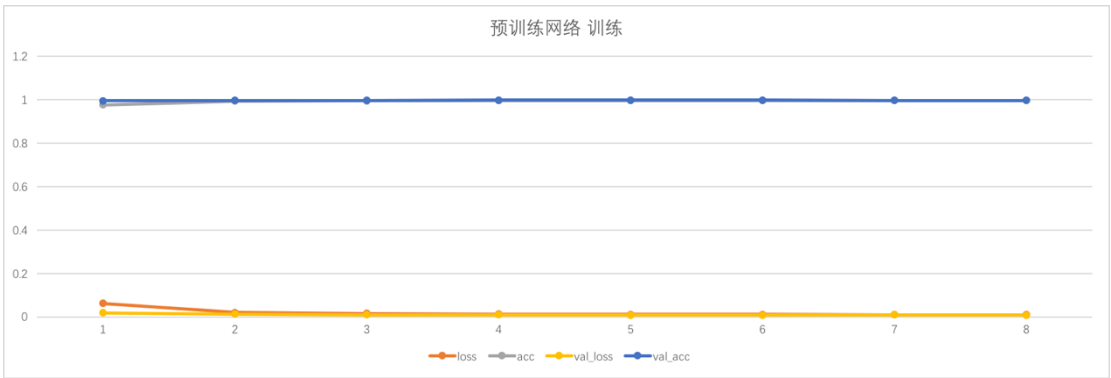

图 13.预训练网络训练过程

5 结论

5.1新采集的图片测试

在 AWS 上训练完成后，把模型加载到微信后台服务，用手机拍摄一些猫和狗的图片，或者在互联网上下载一些图片，发送到微信公众后之后，微信公众号返回类别判断。

预测结果如图 14 所示，选取了 4 张图片。都正确预测，目前来看，除了一些非猫狗的图片无法预测或预测错误，模型在不同的像素和环境下基本预测正确。抗干扰能力比较强。

图 14.公众号识别图片

从第一张图片可以看出，图片颜色比较灰暗，整体不太清晰，但是仍然能够准确分类；第二张图片环境和清晰度都很好，可以正确分类；第三张是手绘图片，非自然图片，目标的特征比较明显，可以正确分类；第三张是合成图片，环境影响因素比较大，目标显示部分特征，可以正确分类。对于没有目标的图片，公众号仍然能识别猫狗，推测是环境因素的权重比较大，属于误判，但是模型只有 2 类，出现 unknown 的类别时，应用确实无法识别。

5.2总结

以往的物体识别，需要分割出物体的形状，然后再训练，项目模型直接缩放训练图像，虽然在模型复杂度上比较高，训练时间比较长，但是，在环境干扰，像素模糊度上等等都表现出了良好的性能，已经达到了人类识别物体的分类准确率。

识别物体的神经网络模型，充分体现了其灵活性，图片处理简单，只要缩放图片即可，符合人类识别物体的思路。模型对环境抗干扰能力比较强，在一些亮度不足，模糊，环境背景复杂的图片中，同样有很高的分类准确率。虽然部分图

片只能体现待识别物体的部分特征，但是仍然能够很好的识别出来。

项目成功达成了设定的目标，即模型的准确识别率达到了 90%，优化之后达到了 99.76%，在 kaggle 上的得分排名在 50 左右，效果不错。

5.3后续改进

项目对训练集图片进行训练，基于算法的优越性，针对猫狗两种类别，我们达到了训练的分辨精度。但是从产品的角度，需要再增加一种类别（unknown），公众号服务更加友好。对于未知类别，简单的做法就是把预测值在[0.45，0.55]的图片设置为未知，也就是模型不是很确定的分类。

为了更好的提升模型对类别的多种形态的辨别，在训练模型的时候可以提升图片。在样本有限的情况下，增加训练样本的不同形态，对模型训练有所帮助。图片提升有助于提升模型的适应性。

公众号需要持续的优化，不仅在性能上，能够给用户带来更顺畅的体验。比如在图片上增加物体的标识，把分类的标注添加到图片上。

参考文献

[1].Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi.(2016).Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning.

[2].http://keras-cn.readthedocs.io/en/latest/

[3].Simon Haykin.Neural Networks and Learning Machines,Third Edition.

[4].https://github.com/vdumoulin/conv_arithmetic

[5].https://en.wikipedia.org/wiki/Convolutional_neural_network

[6].https://github.com/titu1994/Inception-v4/blob/master/inception_resnet_v2.py