小尺寸图片的图像多分类问题

及零样本学习的简单实践

wss321

一、 摘要

图像物体分类与检测是计算机视觉研究中的两个重要的基本问题，也是图像分割、物体跟踪、行为分析等其他高层视觉任务的基础。图像分类问题目前已经能够达到了很高的准确率，但是对于尺寸较小的图片来说，由于包含细节少，因而在分类上会比大图像更难。本文通过超分辨和迁移学习的方法对小尺寸图片的数据集内的图片进行处理，并通过卷积神经网络进行的迁移学习，在样本较少但种类多的数据集上进行了实验，相比未使用超分辨和迁移学习，我们的方法有更好的分类效果。

另外本文还对于图像分类进行了更深一层的拓展，对零样本学习进行了一个简单的实践。零样本学习(zero-shot learning，简称 ZSL) 识别从未见过的数据类别，即训练的分类器不仅仅能够识别出训练集中已有的数据类别，还可以对于来自未见过的类别的数据进行区分。这是一个很有用的功能，使得计算机能够具有知识迁移的能力，并无需任何训练数据，很符合现实生活中海量类别的存在形式。对图像理解、(从已知类别到未知类别的) 知识迁移具有重要意义，是目前研究的一个重点和难点。本文采用了卷积神经网络特征提取与属性和词向量联合嵌入结合的方法进行了实验，最终取得了不错效果。

本文实验代码: https://github.com/wss321/dl_course

二、 背景

1. 图像分类

图像分类算法主要有 KNN、SVM、BPNN、CNN 和迁移学习等，目前效果最好的是 CNN 和迁移学习算法。CNN 图像分类对于数据量要求高，对于图片尺寸较大种类较少的数据集分类来说，CNN 能达到很好的效果。然而在很多实际应用中，图片例如在车牌识别的分类任务中，需要从图片中提取出车牌进行识别，但是分割出来的图片尺寸都比较小而且比较模糊，对于尺寸小、种类多且样本少的数据集来说，单纯的卷积神经网络和迁移学习效果并不理想。因此本文通过超分辨的方法将图片进行放大处理，然后再进行分类。

2. 零样本学习

在传统图像识别任务中，训练阶段和测试阶段的类别是相同的，但每次为了识别新类别的样本需要在训练集中加入这种类别的数据。一些类别的样本收集代价大，即使收集到足够的训练样本，也需要对整个模型进行重新训练。这都会加大识别系统的成本，零样本学习方法便能很好的解决这个问题，而且有着很好的应用前景，如：未知物体识别、未知语言翻译、未知类别图像合成和图像哈希等，而且零样本学习也成为了迁移学习领域的热门研究方向之一。

在 ZSL 中，某一类别在训练样本中未出现，但是我们知道这个类别的特征，然后通过语料知识库，便可以将这个类别识别出来。零样本学习的一个重要理论基础就是利用高维语义特征代替样本的低维特征，使得训练出来的模型具有迁移性。语义向量就是高维语义特征，比如一个物体的高维属性语义为“四条腿，有尾巴，宠物

的一种”，那我们就可以判断它是狗。以上是通过属性特征嵌入进行零样本学习，目前主流的零样本方法还通过词嵌入(word embedding)用词向量(word vector)来替代属性特征进行训练，效果一般会比属性嵌入较好。本文采用两者的结合进行训练，达到了比二者单独训练更好的准确率。

三、 方法

1. 图像分类的方法

本文的图像分类主要运用了数据增强、超分辨和迁移学习的方法，模型图如图一所示。

(1) 迁移学习：预训练模型

迁移学习是一种机器学习方法，就是把为任务 A 开发的模型作为初始点，重新使用在为任务 B 开发模型的过程中。

深度学习中在计算机视觉任务和自然语言处理任务中将预训练的模型作为新模型的起点是一种常用的方法，通常这些预训练的模型在开发神经网络的时候已经消耗了巨大的时间资源和计算资源，迁移学习可以将已习得的强大技能迁移到相关的的问题上。深度学习中的这种迁移被称作归纳迁移。就是通过使用一个适用于不同但是相关的任务的模型，以一种有利的方式缩小可能模型的搜索范围。

在 DNN 中采用预训练模型的方法进行迁移学习：使用现有的已经训练好的卷积神经网络，去掉其顶端的全连接层并接上自己设计的全连接层，基于预训练模型的权重进行重新训练或者冻结掉全连接层之前的网络的权重，只训练自定义的全连接层。本文采用的是重新训练的方法， 使用 Gao Huang, Zhuang Liu 等人提出的 DenseNet [1]的 121 层模型在 imagenet 上训练的权重进行重新训练。

DenseNet 是一种具有密集连接的卷积神经网络。在该网络中，任何两层之间都有直接的连接，即网络每一层的输入都是前面所有层输出的并集，而该层所学习的特征图也会被直接传给其后面所有层作为输入。DenseNet 缓解梯度消失问题，加强特征传播，通过特征复用，极大的减少了参数量，而且具有非常好的抗过拟合性能，尤其适合于训练数据相对匮乏的应用。基于 DenseNet 的种种优点，本文采用 DenseNet 预训练。

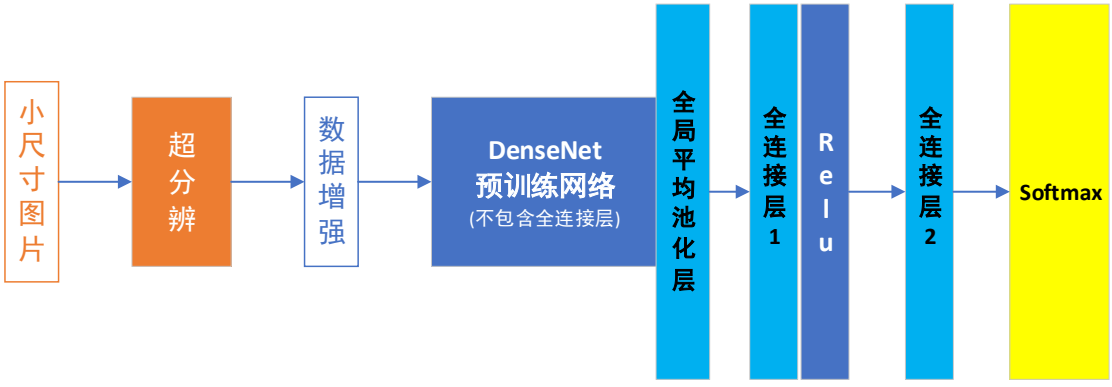

图 1 分类网络模型

(2) 超分辨

超分辨率技术（Super-Resolution）是指从观测到的低分辨率图像重建出相应的高分辨率图像，在监控设备、卫星图像和医学影像等领域都有重要的应用价值。SR 可分为两类:从多张低分辨率图像重建出高分辨率图像和从单张低分辨率图像重建出高分辨率图像。基于深度学习的 SR，主要是基于单张低分辨率的重建

方法，即 Single Image Super-Resolution (SISR)。

SISR 是一个逆问题，对于一个低分辨率图像，可能存在许多不同的高分辨率图像与之对应，因此通常在求解高分辨率图像时会加一个先验信息进行规范化约束。在传统的方法中，这个先验信息可以通过若干成对出现的低-高分辨率图像的实例中学到。而基于深度学习的 SR 通过神经网络直接学习分辨率图像到高分辨率图像的端到端的映射函数。

基于深度学习的 SR 方法主要包括 SRCNN，DRCN, ESPCN，VESPCN，SRGAN 和 DCSCN 等。

本文采用的是 Jin Yamanaka 等人于 2017 年提出的 DCSCN[3]模型来对小尺寸图片进行超分辨处理。DCSCN 通过残差网络、多跳连接(Skip Connection)和网络中的网络(Network in Network) 大大降低了计算量。DCSCN 的模型图如图二所示，残差网络用于特征提取，并行的 1*1 卷积层和上采样组成的网络用于图像的重构(reconstruction)， 输入低分辨率的图片，输出高分辨的图片。相比于其他的超分辨网络，DCSCN 显著的降低了计算成本，大大提高了运算速度，而且达到了当时最好的效果。

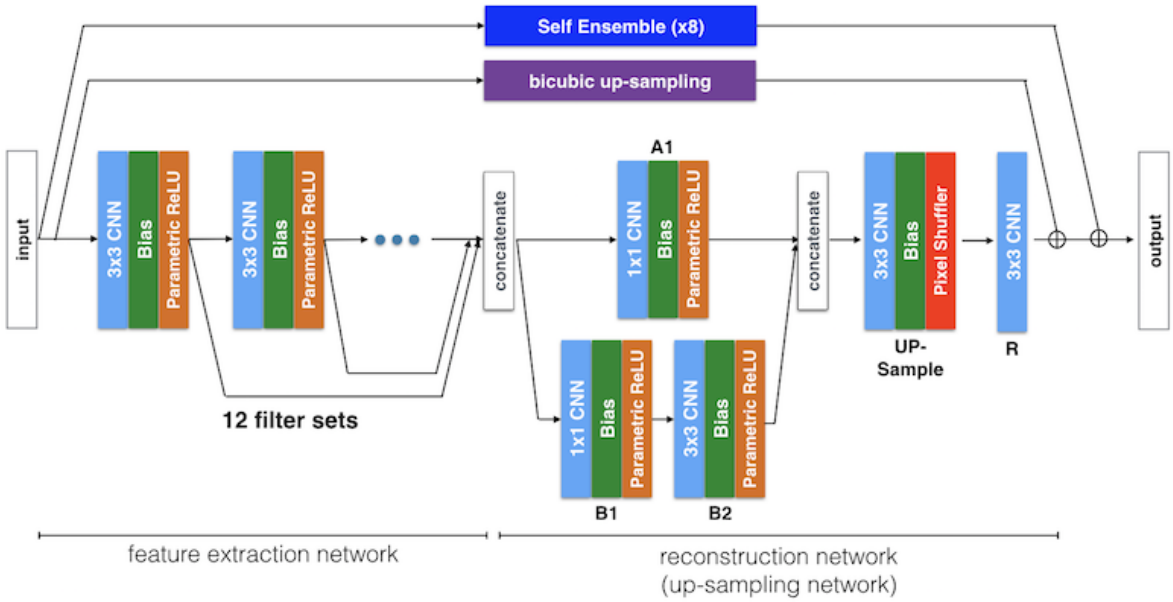

图 2 DCSCN 模型图

(3) 数据增强

随着神经网络的加深，需要学习的参数也会随之增加，这样就会更容易导致过拟合，当数据集较小的时候，过多的参数会拟合数据集的所有特点，而非数据之间的共性。而过拟合就是神经网络可以高度拟合训练数据的分布情况，但是对于测试数据来说准确率很低，缺乏泛化能力。

因此在这种情况下，为了防止过拟合现象， 提高模型的泛化性能，数据增强应运而生。当然除了数据增强，还有正则项或者 dropout 等方式可以防止过拟合。

本文采用的数据增强主要有：随机旋转、左右平移、上下翻转、缩放、加入高斯白噪声和弹性变换[4]。

2. 零样本学习的方法

(1) 属性与词向量标签

目前的研究工作提出了很多基于属性或词向量的零样本学习方法，具有代

表性的有 DAP(Direct Attribute Prediction)、ALE(Attribute Label Embedding)、DEVISE(Deep Visual Semantic Embedding)、CONSE(Convex Combination of Semantic Embeddings)、SAE(Semantic Autoencoder)等等。

其中属性指的是某个类含有的一些特征，如：颜色、是否有毛发、是否可食用等等，是由人们手动标注的，具有直观性和可理解性。而词向量是通过维基百科的大量文本数据训练出来的，将一类将词的语义映射到向量空间中去的自然语言处理技术。即将一个词用特定的向量来表示，向量之间的距离（例如，任意两个向量之间的 L2 范式距离或更常用的余弦距离）一定程度上表征了的词之间的语义关系。目前常用的词向量模型有 Word2Vec 和 GloVe。本文采用的是 GloVe 训练好的词向量作为标签进行实验。

(2) 特征提取

特征提取指的是从图像数据中提取出特征，目前常用的特征提取方法主要有传统的 SIFT（尺度不变特征变换和 HOG（方向梯度直方图）以及不具可解释性的 CNN（卷积神经网络）特征提取。本文采用 CNN 特征提取的方法，将图片输入已经训练好的卷积神经网络，从全连接层提取特征，模型图如图三。

本次实验采用了两种特征提取模式，原始图片经过一次超分辨，进行一次放大操作（由于预训练模型的输入图片长宽不可低于 221*221，因此进行放大），通过训练好的 CNN 分类网络，在第一个全连接层提取出特征，如图三。

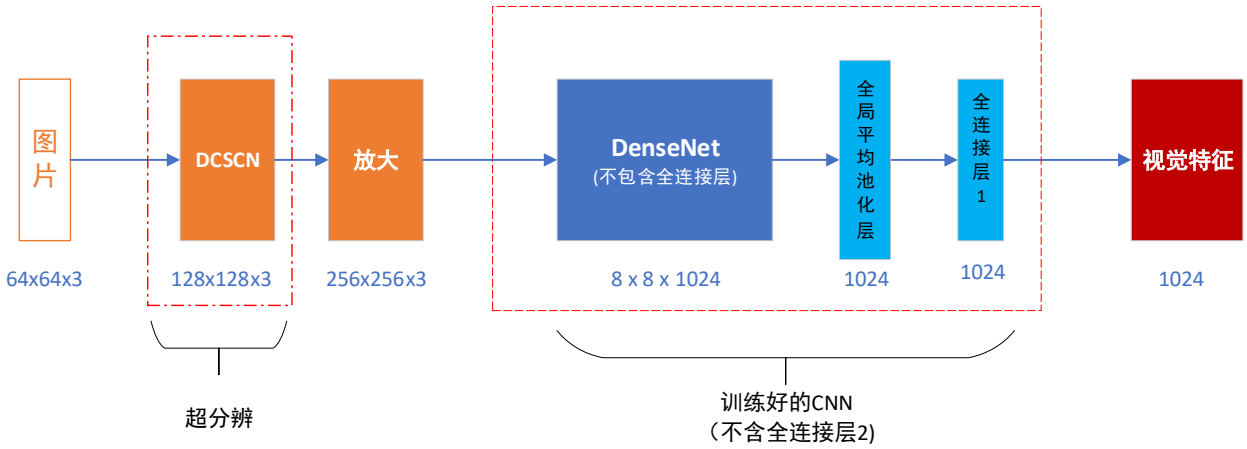

图 3 CNN 特征提取

(3) 训练和测试网络

在第二部分完成了特征提取之后，下一步就是训练一个 (1).从视觉特征到语义嵌入特征或者 (2).从语义嵌入特征到视觉特征 之间的映射网络了。本文对两种转换情况都进行了实验，并对结果进行了对比，结果显示：语义嵌入特征到视觉特征的网络准确率会略高于视觉特征到语义嵌入特征网络。两种网络分别如图四和图五所示。

在训练阶段，模型 1 输入图片对应的由特征提取网络提取的视觉特征，输出与其对应语义嵌入向量进行比较，采用负余弦相似度作为损失函数进行训练；模型 2 输入图片语义嵌入向量，输出与其对应视觉特征进行比较，同样采用负余弦相似度作为损失函数进行训练；

在测试阶段，对于模型 1，将测试图片的视觉特征输入网络中，得到的输

出与所有测试的类别对应的语义嵌入向量进行比较，取余弦相似度最高的向量对应的类作为该图片的预测结果；对于模型 2，将所有测试的类对应的语义嵌入向量作为输入，得到对应每个类别的视觉特征，然后将测试集图片的视觉特征与得到的语义嵌入向量求余弦相似度，同样取相似度最高的向量对应的类作为该图片的预测结果；

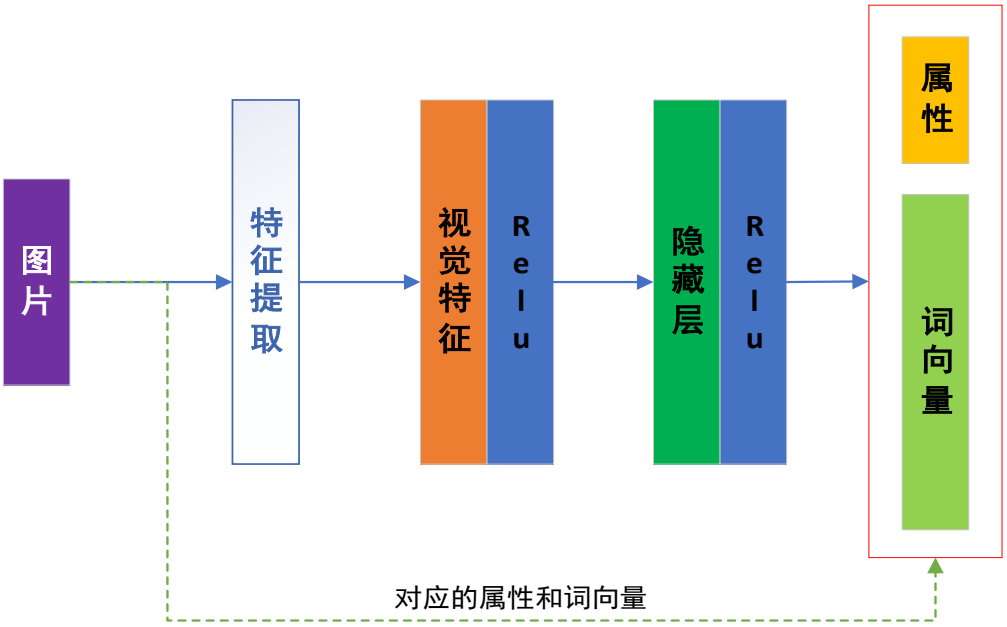

图 4 零样本学习模型 1：视觉特征到语义嵌入特征映射

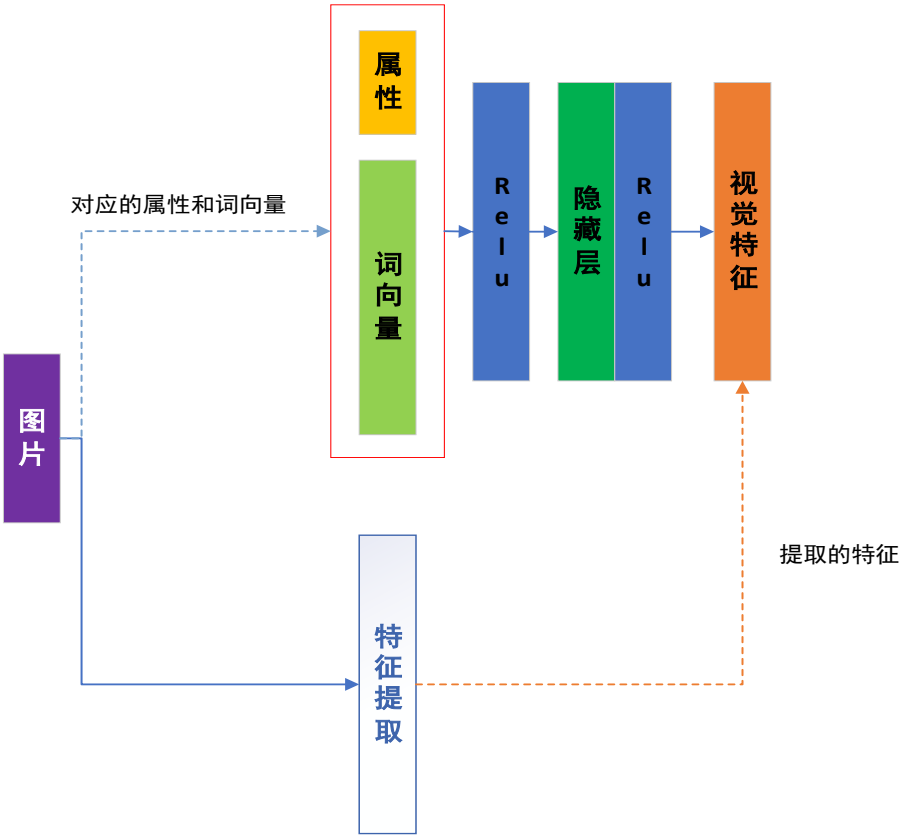

图 5 零样本学习模型 1：语义嵌入特征到视觉特征映射

四、 实验

为了验证我们提出的方法的有效性，本文分别对图片分类和零样本学习在 keras 平台上进行了实验。其中单纯的图像分类只在阿里云的一个数据集上进行实验，而对于零样本学习，本文不仅在阿里云的数据集上进行了实验，还在目前很流行的零样本学习数据集之一的 AwA 数据集进行的实验。整个实验的代码已经托管至 github: https://github.com/wss321/dl_course

0. 数据集介绍：

(1). 本次分类实验采用了阿里云的天池大数据的一个小尺寸图像数据集，其中包含了 203 个类，及其每个类对应的属性和 GloVe 词向量。其中每张图片大小为 64*64*3，每个类大约 400 张图片，每类属性有 30 维，每个词向量有 300 维。其中词向量和属性数据仅在零样本实验中使用。

(2). AwA 数据集(Animals with Attributes)：本次实验未使用 AwA 数据集的图片数据，而是使用了已经经过 googlenet 模型进行特征提取的 1024 维视觉特征数据，以及每类对应的 85 维属性和 1000 维和词向量。该数据集包含了 30475 个图片的视觉特征，共 50 个类。

1. 图像分类实验

由于天池的数据集类别多，图片总数量太大，难以进行训练，同时为了给零样本学习实验留一部分类别作为测试集，因此我们只使用了该数据集中的 164 类进行训练。

我们分别对未使用预训练和超分辨、使用了超分辨但未使用预训练和既用了预训练又进行了超分辨进行了实验。

其中的超分辨网络 DCSCN 是在 DIV2k 数据集上训练的模型。超分辨效果如图 6 所示，原图是 64*64*3 的图像，超分辨之后为 128*128*3。由于没有重新训练，因此在本数据集上超分辨的效果并不是很理想，存在花纹的现象。

图 6

实验结果如表 1，既用了预训练又进行了超分辨的分类效果相比于只用了数据增强的模型提升了 28.7%，预训练但未进行超分辨提升了 23.2%。

模型	loss	准确率
未预训练+未超分辨	2.02	45.6%
预训练+未超分辨	1.47	68.8%
预训练+超分辨	**1.24**	**74.3%**

表 1 图像分类结果

2. 零样本学习实验

在零样本学习方面，我们对天池大数据的数据集以及 AwA 数据集都进行了实验，同时采用了两种不同的模型，如图 4 和图 5。在天池大数据的数据集上，我们挑选了 50 类动物的图片进行实验，训练集包含其中的 40 类，测试集则为剩下的 10 类，训练集与测试集的类别不重复，即测试集的类不参与训练过程。在 AwA 数据集上同样如此，训练集和测试集分别为 40 类和 10 类。实验结果如表 2，在天池大数据的数据集上两种模型的准确率都低于在 AwA 上的准确率，本文认为原因在于语义嵌入向量的质量差别，即天池大数据的属性仅有 30 维，而且不够准确，而 AwA 则有 85 维，比相比之下维数更多更有泛化性。

序号	模型	数据集	语义嵌入	训练集准确率(%)	测试集准确率(%)
1	模型 1	天池	属性	68	23.3
2	模型 1	天池	词向量	86	33.5
3	**模型 1**	**天池**	**属性+词向量**	**86**	**35.4**
4	模型 1	AwA	属性	_	57.1
5	模型 1	AwA	词向量	_	52.8
6	**模型 1**	**AwA**	**属性+词向量**	_	**62.1**
7	模型 2	天池	属性	68	28.7
8	模型 2	天池	词向量	81	25.3
9	**模型 2**	**天池**	**属性+词向量**	**81**	**26.7**
10	模型 2	AwA	属性	_	74.2
11	模型 2	AwA	词向量	_	74.3
12	**模型 2**	**AwA**	**属性+词向量**	_	**78.4**

表 2 零样本学习实验结果

五、 结论

迁移学习、超分辨和数据增强能够很大程度的提高分类的准确性，并且可以很大程度上缩减训练的时间。其中数据增强和迁移学习得到广泛的应用。

另外本文提出的两个简单的零样本模型在 AwA 数据集上取得了不错的效果，其中训练一个从语义特征到视觉特征的映射网络的效果比一个从视觉特征到语义特征的映射网络更优。

六、 参考文献

[1]. D Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger,"Densely Connected Convolutional Networks," in CVPR, 2017.

[2]. Xi Cheng, Xiang Li, Ying Tai, Jian Yang, “SESR: Single Image Super Resolution with Recursive Squeeze and Excitation Networks”.

[3]. Jin Yamanaka, Shigesumi Kuwashima, Takio Kurita , “Fast and Accurate Image Super Resolution by Deep CNN with Skip Connection and Network in Network”.

[4]. P.Y. Simard，D. Steinkraus，J.C. Platt, “Best practices for convolutional neural networks applied to visual document analysis”, 2003.