计算机视觉基础

第三章 图像分类

- 3.1 概述
- 3.2 基于词袋表示的图像分类
- 3.3 基于Fisher向量的图像表示方法
- 3.4 基于深度学习的图像分类

3.1 概述

图像分类问题，指的是通过学习标注图像与其标签的对应关系，对未标注图像进行自动标注的问题。

- 图像分类问题可以针对通用的类别
  - Pascal2012挑战赛中的图像分类挑战包含了20个类别：人、鸟、猫、牛、狗、马、羊、飞机、自行车、船、公交车、轿车、摩托车、火车、瓶子、椅子、餐桌、植物、沙发、电视/显示器
  - 海量数据库ImageNet2010挑战赛中的图像分类挑战包含了27大类：动物、家用电器、鸟、鱼、花、食物、水果、乐器、树、蔬菜、人等
- 也可以是针对某个领域的图像
  - 场景分类问题针对图像中所在的场景进行分类
  - 人体行为分类问题针对图像中包含的人进行的动作进行分类

3.1 概 述

图像分类问题，指的是通过学习标注图像与其标签的对应关系，对未标注图像进行自动标注的问题。

- 例如：假设我们的标签集合是{猫、狗}，通过标注好的训练数据对分类器进行训练，然后给定测试图片，分类器需要估计其标签。

图3-1 训练好的图像分类器能够将该图像自动标注为“狗”

3.1 概 述

图像分类问题的类别（按照图像类别的粒度）

1. 类别级图像分类
2. 子类别级图像分类
3. 实例级图像分类

3.1 概 述

图像分类的种类（按照图像类别的粒度）

1. 类别级图像分类

- 不会跟其他类同属于一个类别，即使是同类别物体其类间差别也可能比较大
- 一般来说，类别之间具有较大的差别，而类内具有较小的差别。 例如：
  - ImageNet2010数据库

图3-2 类别级图像分类示例：ImageNet数据库中的15中类别的图像示例

3.1 概述

图像分类的种类（按照图像类别的粒度）

2. 子类别级图像分类

- 类别标签一般都从属于同一个更大的类别
- 与类别级图像分类相比，子类别级图像分类类间差别更小，因此准确估计标签更难
- 子类别级图像分类一般更具有领域针对性。 例如：
  - 鸟类数据库-2011为生物学家提供研究数据
  - 中文路标数据库应用在智能交通领域训练无人驾驶系统
  - 手写字体识别应用在银行手写支票自动识别

图3-3 中文路标数据库示例

3.1 概述

图像分类的种类（按照图像类别的粒度）

3. 实例级图像分类

- 实例级图像分类是对物理世界中个体进行分类

➢ 例如：

- 对人脸进行识别，可以使用在门禁系统中进行人脸打卡

图3-4 LFW数据库部分人脸图像及其对应身份标注

3.1 概 述

图像分类的发展

- 图像识别领域大量的研究成果都是建立在PASCAL VOC、ImageNet等公开的数据集上

1. PASCAL VOC是2005年发起的一个视觉挑战赛，到2010年结束。主要包含三个赛道：图像分类比赛、物体检测比赛、图像分割比赛，后来增加了人类行为识别比赛和人体部分检测比赛
   - 词袋模型是Pascal竞赛中分类算法的基本框架。基于词袋模型的图像识别方法一般包括底层特征提取、特征编码、分类器设计、模型融合等几个阶段。
2. 1994年，Yan LeCunn提出LeNet-5网络来进行手写字体识别的，在MNIST数据库上进行验证
3. ImageNet是2010年发起的大规模视觉识别竞赛的数据集，总共有1400多万幅图片，涵盖2万多个类别
   - 2012年，Alex Krizhevsky提出AlexNet，夺得2012年ILSVRC比赛的冠军，top5预测的错误率为16.4%，远超第二名。
   - 2013年，ILSVRC分类任务冠军网络是Clarifai，不过更为我们熟知的是ZFNet：Hinton的学生Zeiler和Fergus在研究中利用反卷积技术引入了神经网络的可视化，对网络的中间特征层进行了可视化。
   - 2014年的冠亚军网络分别是GoogLeNet和VGGNet。
   - 2015年，何凯明提出ResNet获得了分类任务冠军。

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

- 词袋模型（Bag of Words）是自然语言处理与信息检索中的一种简化表示，这种方法不考虑语法和单词顺序，只计算单词的出现频率，将文本表示成以单词为容器（bin）值的直方图

3.2.1 基于词袋的句子检索

- 例子：假设数据库中存储了3个句子，使用基于词袋的表示方法，在数据库中搜索与“海很宽阔任凭鱼儿游来游去，天空很高任凭鸟儿飞来飞去”这句话最相近的话。

1. “鸟以为把鱼举在空中是一种慈善的举动。”
2. “人们可以争取自由，但却永远不能恢复自由。”
3. “生命漫长也短暂。”

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.1 基于词袋的句子检索

- 例子：假设数据库中存储了3个句子，使用基于词袋的表示方法，在数据库中搜索与“海很宽阔任凭鱼儿游来游去，天空很高任凭鸟儿飞来飞去”这句话最相近的话。

1. “鸟以为把鱼举在空中是一种慈善的举动。”
2. “人们可以争取自由，但却永远不能恢复自由。”
3. “生命漫长也短暂。”

- 解决方案：

1. 计算词典。去掉没有意义的助词和连词，把所有的单词的集合作为词典。数据库中所有的句子得到的词典为：

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
鸟	以为	鱼	举	在空中	是	慈善	举动	人们	可以	争取	自由	不能	恢复	生命	漫长	短暂

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.1 基于词袋的句子检索

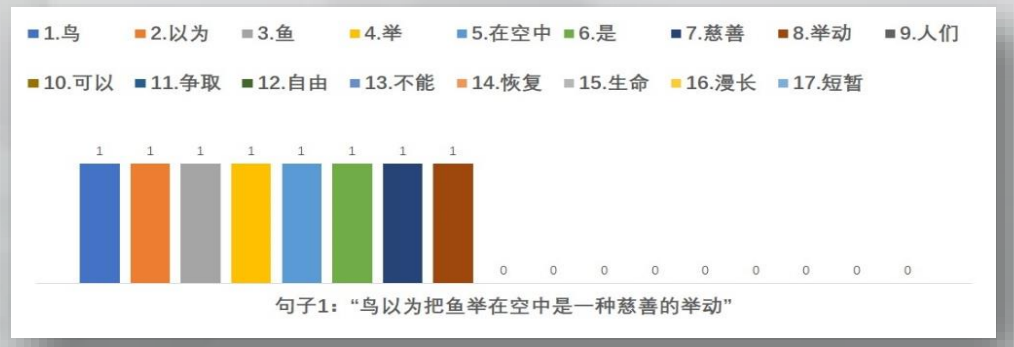
- 例子：假设数据库中存储了3个句子，使用基于词袋的表示方法，在数据库中搜索与“海很宽阔任凭鱼儿游来游去，天空很高任凭鸟儿飞来飞去”这句话最相近的话。

1. “鸟以为把鱼举在空中是一种慈善的举动。”
2. “人们可以争取自由，但却永远不能恢复自由。”
3. “生命漫长也短暂。”

- 解决方案：

1. 计算词典。
2. 每个句子都可以表示成词典的直方图。数据库中的句子1包含：“鸟”“以为”“鱼”“举”“在空中”“是”“慈善”“举动”，所以句子1的直方图为：

111111110000000000



3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.1 基于词袋的句子检索

- 例子：假设数据库中存储了3个句子，使用基于词袋的表示方法，在数据库中搜索与“海很宽阔任凭鱼儿游来游去，天空很高任凭鸟儿飞来飞去”这句话最相近的话。

1. “鸟以为把鱼举在空中是一种慈善的举动。”
2. “人们可以争取自由，但却永远不能恢复自由。”
3. “生命漫长也短暂。”

- 解决方案：

1. 计算词典。
2. 每个句子都可以表示成词典的直方图。
3. 同样的句子2的直方图为：0000000011121000；句子3的直方图为：0000000000000111。检索句子的直方图为：1010100000000000

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.1 基于词袋的句子检索

- 例子：假设数据库中存储了3个句子，使用基于词袋的表示方法，在数据库中搜索与“海很宽阔任凭鱼儿游来游去，天空很高任凭鸟儿飞来飞去”这句话最相近的话。

1. “鸟以为把鱼举在空中是一种慈善的举动。”
2. “人们可以争取自由，但却永远不能恢复自由。”
3. “生命漫长也短暂。”

- 解决方案：

1. 计算词典。
2. 每个句子都可以表示成词典的直方图。
3. 同样的句子2的直方图为：0000000011121l000；句子3的直方图为：0000000000000111。检索句子的直方图为：1010100000000000
4. n维空间点a($x_{11},x_{12},\ldots,x_{1n}$)与b($x_{21},x_{22},\ldots,x_{2n}$)间的欧氏距离（两个n维向量）：

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2}$$

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.1 基于词袋的句子检索

- 例子：假设数据库中存储了3个句子，使用基于词袋的表示方法，在数据库中搜索与“海很宽阔任凭鱼儿游来游去，天空很高任凭鸟儿飞来飞去”这句话最相近的话。

1. “鸟以为把鱼举在空中是一种慈善的举动。”
2. “人们可以争取自由，但却永远不能恢复自由。”
3. “生命漫长也短暂。”

- 解决方案：

1. 计算词典。
2. 每个句子都可以表示成词典的直方图。
3. 同样的句子2的直方图为：0000000011121l000；句子3的直方图为：0000000000000111。检索句子的直方图为：1010100000000000
4. 计算句子间的欧几里得距离，那么查询句子跟句子1、2、3之间的距离分别为$\sqrt{5}$、$\sqrt{12}$、$\sqrt{6}$，因此跟库中第一个句子最相近。返回句子1。

3.2 基于词袋的图像分类方法

3.2.2 基于词袋的图像分类

基于词袋的图像分类与基于词袋的句子检索在流程上基本相同，一般都包含以下几个步骤：

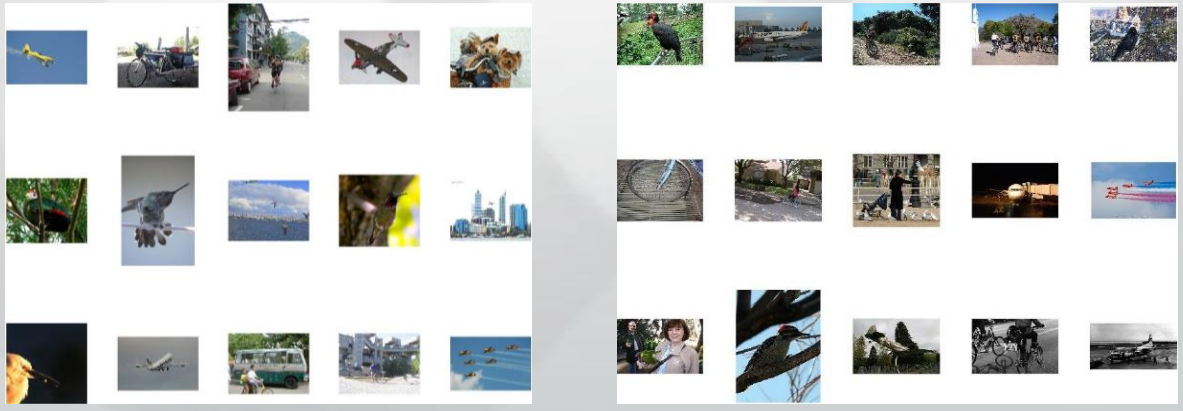
1. 对数据进行特征提取
2. 多特征融合
3. 通过数据库中数据计算词典
4. 将所有的数据表示成词典的直方图
5. 基于机器学习模型的回归或者分类

3.2 基于词袋的图像分类方法

3.2.2 基于词袋的图像分类[1]

- 数据集介绍：
  - 数据采用Pascal Challenge VOC 2007中的数据，为了方便演示，我们选取了一个有15幅训练图像和15幅测试图像的小型实例。任务是判断图像是否包含飞机。

➢ 图3-7显示了训练和测试数据集。每个训练数据图像，都包含一个类别标签的标注，如果图像中包含飞机标签为1，否则标签为-1。在训练判别器的过程中，训练数据集中的图像和标签都可以使用。在测试过程中，我们通过测试数据图像特征推测其标签。



（a）训练数据集　　（b）测试数据集

图3-7 演示实例训练数据集合

1. 本教材按照Pascal Challenge中给出的程序框架 的基本流程进行介绍，感兴趣的同学可以下载Development Kit和数据库进行算法和程序开发：http://host.robots.ox.ac.uk/pascal/VOC/voc2012/#devkit

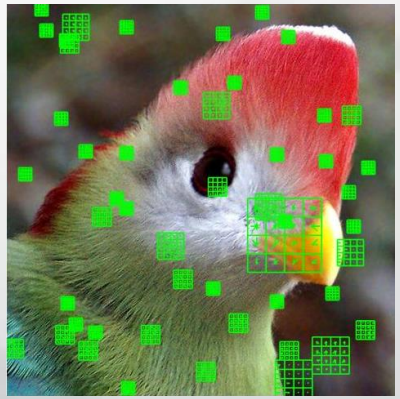
3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.2 基于词袋的图像分类[1]

1. 对数据进行特征提取（使用VLFeat的denseSIFT[2]）

- denseSIFT对图像进行特征提取和表示，该特征在SIFT特征提取的基础上进行了加速处理。



（a）一幅鸟类图像　　（b）denseSIFT提取特征示例

图3-8 denseSIFT提取特征示例

2. https://www.vlfeat.org/index.html

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.2 基于词袋的图像分类

1. 对数据进行特征提取（使用VLFeat的denseSIFT）
2. 多特征融合
   - 一种特征通常只能表示有限的图像特性
     - 例如：SIFT提取的是边角特征，HOG提取的是边缘特征
   - 多特征融合有很多种方法。最常用的方法有两类：早期融合（early fusion）和后期融合（late fusion）。
     - 早期融合一般指特征级融合，例如我们可以把对图像提取的SIFT和HOG特征向量合并成一个向量
     - 后期融合一般指判别器融合，即将多个特征对图像进行的估计按照不同的权重进行融合

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.2 基于词袋的图像分类

1. 对数据进行特征提取（使用VLFeat的denseSIFT）
2. 多特征融合
3. 利用K-Means算法构造词典

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.2 基于词袋的图像分类

1. 对数据进行特征提取（使用VLFeat的denseSIFT）
2. 多特征融合
3. 利用K-Means算法构造词典

- K-means方法是机器学习模型中非常经典也是非常实用的一个方法，通常被用来进行无监督学习。
- K代表的是聚类个数，在这个实验中设置为4096。
- 下面给出了通过K-means方法计算得到的属于不同聚类中心的图像块。图中显示了随机选择的2个类，每个类中显示了被划分到这个聚类的所有图像特征。

（a）聚类1中的图像块

（b）聚类2中的图像

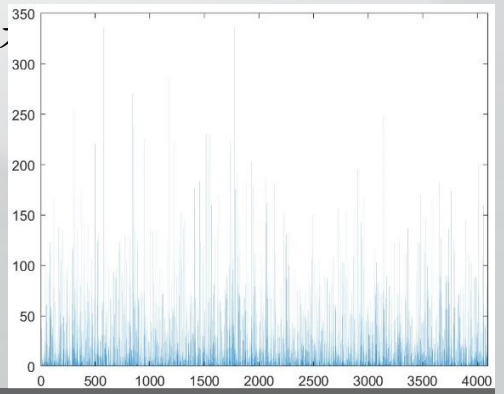
图3-9 聚类结果中不同的类中包含的图片块

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.2 基于词袋的图像分类

1. 对数据进行特征提取（使用VLFeat的denseSIFT）
2. 多特征融合
3. 利用K-Means算法构造词典
4. 图像量化
   - 图像中的每个特征都与词典进行比较，并将其近似表示成与该特征最相似的词典项
   - 计算所有特征中每个词典项出现的次数可以得到该词典的直方图
   - 每幅图像都被表示成词典的直方图分



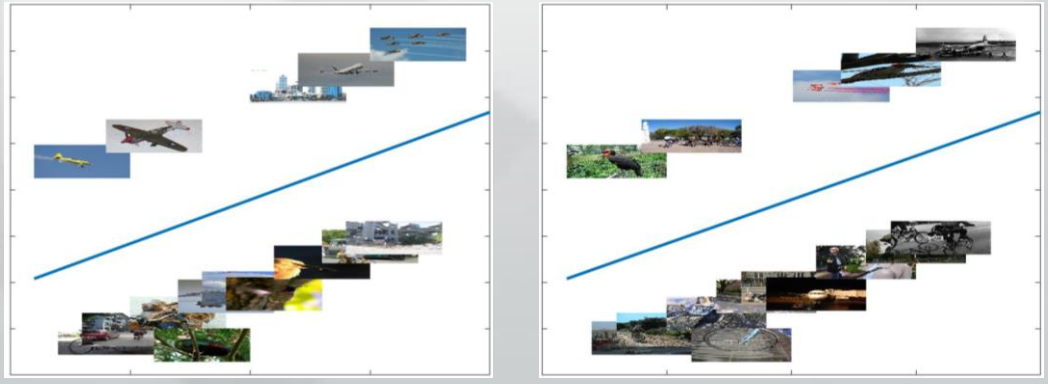
（a）例子图像　　（b）图像的直方图表示

图3-10 VOC2007 中的一幅例子图像及其根据词典计算得到的直方图表示

3.2 基于词袋的图像分类方法

3.2.2 基于词袋的图像分类

1. 对数据进行特征提取（使用VLFeat的denseSIFT）
2. 多特征融合
3. 利用K-Means算法构造词典
4. 图像量化
5. 基于SVM的图像分类
   - 支持向量机是一种线性分类模型，我们可以将其简单的表示成f=wx+b。
   - 支持向量机的原理是通过保留两类数据之间的最大间隔，实现最优的分离超平面。
   - 图3-11给了训练和测试时数据分布的示意图。



（a）训练数据分布　　（b）测试数据分布

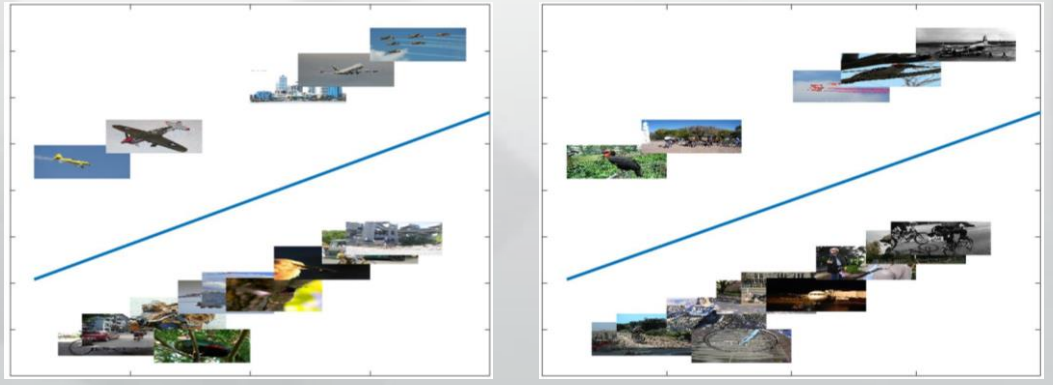
图3-11 训练数据和测试数据的分布

3.2 基于词袋的图像分类方法

词袋模型（Bag of Words）

3.2.2 基于词袋的图像分类

1. 对数据进行特征提取（使用VLFeat的denseSIFT）
2. 多特征融合
3. 利用K-Means算法构造词典
4. 图像量化
5. 基于SVM的图像分类
6. 识别准确率为42%



（a）训练数据分布　　（b）测试数据分布

图3-11 训练数据和测试数据的分布

3.3 基于Fisher向量的图像表示方法

机器学习领域有两类主流的解决方案：

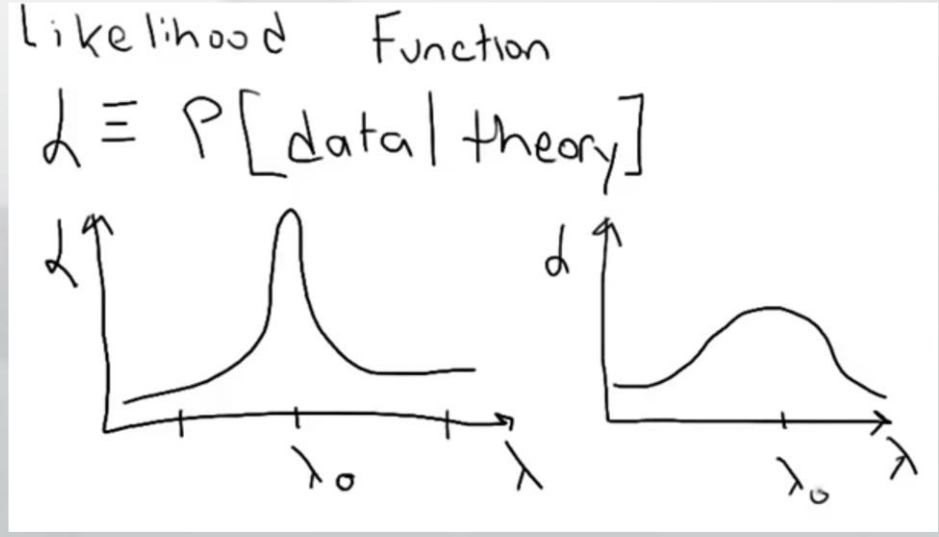
1. 通过对数据特征空间进行建模（一般是概率模型）并通过训练数据计算最优的模型参数，称为生成模型
2. 定义目标函数，通过优化算法计算模型的参数得到基于训练数据的最优模型

高斯混合模型是进行特征空间建模的概率模型之一

3.3 基于Fisher向量的图像表示方法

Fisher向量

- Fisher向量：$V_\Theta(X) = F_\Theta^{-\frac{1}{2}} \sum_{i=1}^{N} \nabla_\Theta \log p(\vec{x}|\Theta)$ $V_\Theta(X) \in \mathbb{R}^{D\cdot|\Theta|}$
- Fisher信息矩阵：$F_\Theta = \mathbb{E}_{\vec{x}}\left[U_\Theta(\vec{x})U_\Theta(\vec{x})^T\right]$ $U_\Theta(\vec{x}) = \nabla_\Theta \log p(\vec{x}|\Theta)$
- 高斯概率分布：$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- 高斯概率分布的log函数：$\log P(x) = \log\left[\frac{1}{\sqrt{2\pi}\sigma}\right] - \frac{(x-\mu)^2}{2\sigma^2}$



3.3 基于Fisher向量的图像表示方法

Fisher向量

3.3.1 高斯混合模型

- 高斯混合模型由几个高斯模型组成，假设高斯混合模型中包含k个高斯模型并且他们的序号用$k=\{1,\cdots,\mathrm{K}\}$表示，每个高斯模型都有一个中值μ，一个协方差Σ和一个混合概率π，所有高斯π相加为1，即：

$$\sum_{k=1}^{K}\pi_k=1.$$

- 高斯混合模型可以表示成：

$$P(\mathrm{X}=x)=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k).$$

3.3 基于Fisher向量的图像表示方法

Fisher向量

3.3.1 高斯混合模型

- 图3-12给出了一个高斯混合模型的例子，图中黑色的点是随机产生的5000个二维数据点，利用30个高斯模型对这些数据进行近似，对这30个高斯模型根据其中值和协方差（可以理解成图中蓝色椭圆的扁平程度）进行可视化。

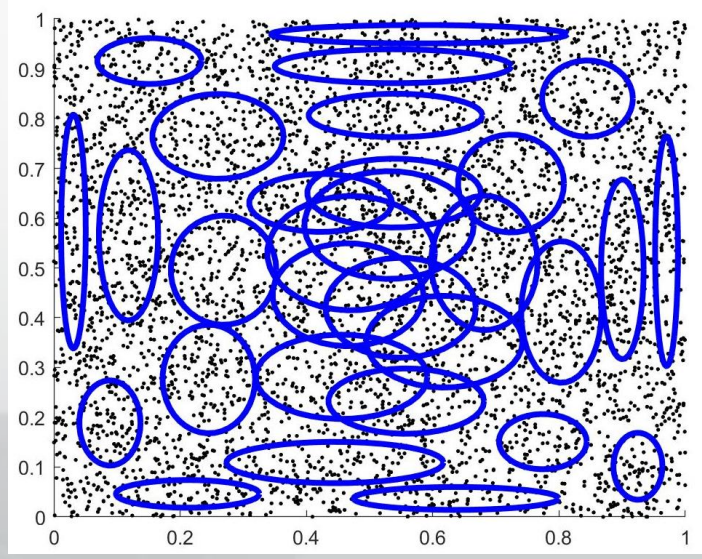


图3-12 利用k值为30的高斯混合模型近似5000个随机产生的点

3.3 基于Fisher向量的图像表示方法

Fisher向量

3.3.2 Fisher向量

- 计算30个拟合高斯的中值和协方差偏差：

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}},$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ik} [(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}})^2 - 1]$$

其中，系数q_{ik}是当前的数据点$\boldsymbol{x}_i$属于第k个高斯模型的概率

$$q_{ik} = \frac{\exp\left[-\frac{1}{2}(\vec{x}_i - \mu_k)^T {\Sigma_k}^{-1}(\vec{x}_i - \mu_k)\right]}{\sum_{t=1}^{K} \exp\left[-\frac{1}{2}(\vec{x}_i - \mu_t)^T {\Sigma_t}^{-1}(\vec{x}_i - \mu_t)\right]},$$

π_k是第k个高斯模型的先验，即一个数据点属于这个高斯模型的概率。整个高斯混合模型中所有高斯模型的先验之和为1，即：$\sum_{k=1}^{K} \pi_k = 1.$

3.3 基于Fisher向量的图像表示方法

3.3.2 Fisher向量

- Fisher向量的表示是对特征的所有维数都求数据中值偏差和协方差偏差，然后将所有结果表示成一个向量，即：

$$\mathrm{FV(I)}^T = [\cdots, \vec{u}_k, \cdots, \vec{v}_k, \cdots].$$

其中，$\vec{u}_k$是由u_{jk}的所有$j = 1,2,\cdots,\mathrm{D}$组成的向量，$j$表示的是特征的维数序号

- 例子：

1. 对所示的两幅猫的图像提取denseSIFT特征。每个SIFT特征的维数为128，从每幅图像中提取10000个特征点，因此特征集合的维数为128×20000。

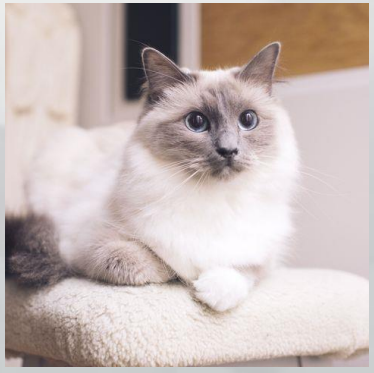


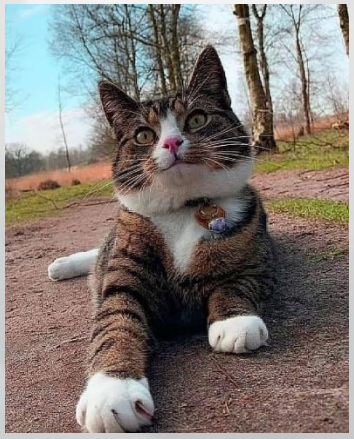
图3-13 高斯混合模型的两幅训练图像

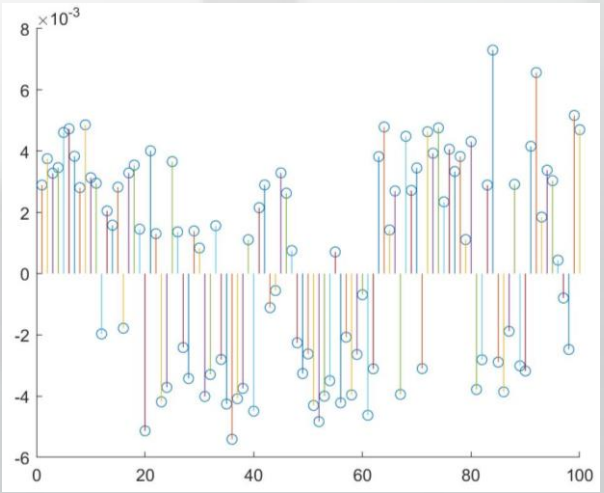
3.3 基于Fisher向量的图像表示方法

3.3.2 Fisher向量

- 例子：

1. 对所示的两幅猫的图像提取denseSIFT特征。每个SIFT特征的维数为128，从每幅图像中提取10000个特征点，因此特征集合的维数为128×20000。
2. 利用64个高斯组成的高斯混合模型去逼近特征分布。
3. 对于一幅新的图像，如图3-14（a）所示，从图像中提取denseSIFT特征得到一个128×10000维的特征矩阵，然后将提取出来的特征计算Fisher向量。
4. 得到最终的Fisher向量为16384×1维，显示结果在图3-14（b）中。（16384=128×64×2）





（a）Fisher向量表示的测试图像　（b）对测试图像的部分Fisher向量进行显示的结果

图3-14 Fisher向量的表示

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

- 基于深度学习的图像分类问题通过端到端（end-to-end）的特征提取和表示方法通过多层神经网络直接提取对目标有效的特征表示，并通过全连接层进行分类。
- 深度学习算法模型的基石是神经网络。比较经典的网络模型包括浅层的神经网络模型和深度学习的网络模型：
  1. 自动编码器（Auto encoder）
  2. 受限玻尔兹曼机（Restricted Bolzman Machine）
  3. 深度信念网络（Deep Belief Network）
  4. 卷积神经网络（Convolutional Neural Network）
  5. 循环神经网络（Recurrent Neural Network）
  6. 长短时记忆神经网络（Long Short Term Memory）

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

1. 自动编码器（Auto encoder）

- 主要思想是通过编码器学习从输入特征空间到一个新的特征空间的映射，并通过解码器学习从这个新的特征空间到原特征空间的映射。
- 映射和反映射的过程都是无监督的，因此称为自动编码器。
- 如果新的特征空间中特征的表示维度比原特征表示维度低，自动编码器可以进行数据降维。

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

1. 自动编码器（Auto encoder）

- **数据库**：MNIST 是计算机视觉算法中具有通用学术意义的基准验证数据库，一共包含10类阿拉伯数字，它包含 60000 个训练数据，10000 个测试数据，图像均为灰度图，通用的版本大小为 28×28。
- **网络结构**：输入数据的维数为28 × 28 = 784维，隐层神经元的个数设置为32，输出层的神经元个数与输入维数相同都是784维。

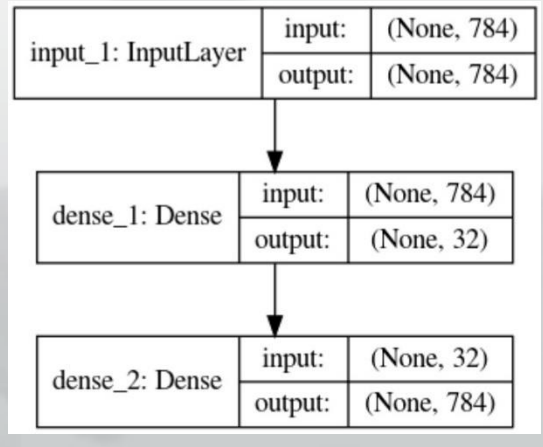


图3-15 Keras显示的自动编码器结构图

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

1. 自动编码器（Auto encoder）

- **结果：**我们利用这样一个网络对数据先降维然后重构，将重构后的数据与原数据进行对比，对比结果显示在图3-16中，我们显示了10幅测试图像，第一列图像是原图像，第二列图像是经过降维又重构的图像。

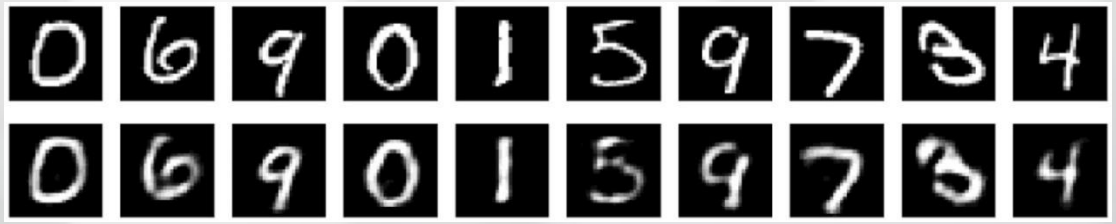


图3-16 MNIST数据库中的测试数据经过自动编码器重构之后的结果

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

2. 受限玻尔兹曼机（Restricted Bolzman Machine）

- 玻尔兹曼机是非确定性的生成式深度学习模型，包含两类结点：隐层节点和可见结点。
- 网络没有输出结点所以网络是非确定性的，是无监督学习网络的一种。
- 玻尔兹曼机的输入之间有连接，输入节点之间可以共享信息。
- 受限玻尔兹曼机是玻尔兹曼机的一种特殊类别，神经元间的连接只限制在可见结点和隐层结点之间

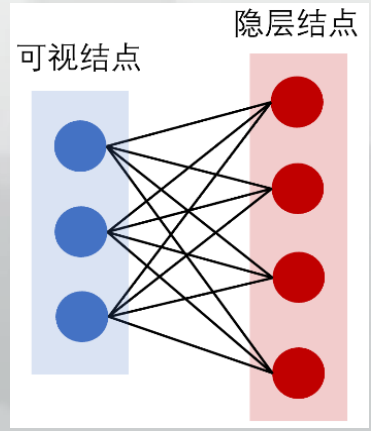


图3-17 受限玻尔兹曼机网络结构示意图

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

3. 深度信念网络（Deep Belief Network）

- 将多层受限玻尔兹曼机堆叠起来，利用梯度下降进行反向传播和微调，形成的网络即深度信念网络。
- 例子：一个由两个受限玻尔兹曼机组成的深度信念网络
  - ✓每层网络都是一个受限玻尔兹曼机
  - ✓前一个受限玻尔兹曼机的输出即为下一个受限玻尔兹曼机的输入

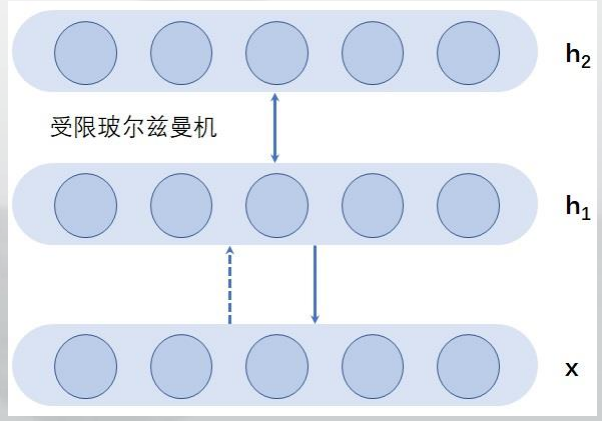


图3-18 由两层受限玻尔兹曼机构成的一个深度信念网络的结构图

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

4. 卷积神经网络卷积神经网络（Convolutional Neural Network）

- 卷积神经网络可以认为是应用最广泛、效果最显著的一种神经网络结构。
- 卷积神经网络的最基本操作单位是卷积和池化。
- 卷积神经网络同全接连网络相比大大减少了需要学习的网络模型参数

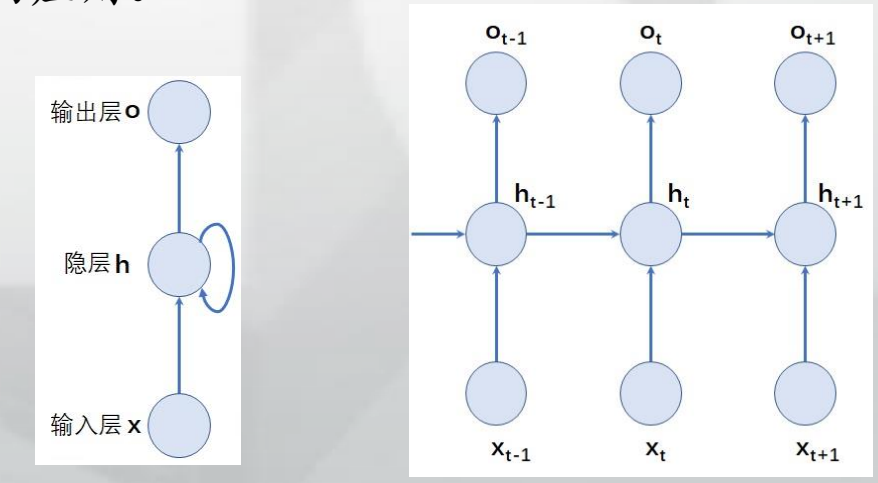
1. 卷积神经网络通过卷积操作避免了对当前神经元层的全连接操作，实现部分网络神经元的局部操作
2. 利用一个卷积核对图像进行处理时，共用卷积核的参数。

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

5. 循环神经网络（Recurrent Neural Network）

- 循环神经网络由若干个循环神经网络的单元构成。
- 这些单元由输入层、隐层和输出层构成。
- 在多个单元构成的循环神经网络结构中，隐层除了接收来自输入数据的信息之外，还接受来自历史隐层的信息。
- 循环神经网络可以处理时序信息
- 循环神经网络在机器翻译、语音识别、自然语言处理等领域都得到了非常广泛的应用。



（a）循环神经网络结构　　（b）展开后的循环神经网络结构

图3-19 循环神经网络结构

3.4 基于深度学习的图像分类

3.4.1 网络模型的主要类别

6. 长短时记忆神经网络（Long Short Term Memory）

- 循环神经网络中时间距离比较久远的信息会变得越来越少
- 长短时记忆神经网络在保留具有短时历史信息的隐层输入的同时，设计了一个能记录长时历史记忆的参数。

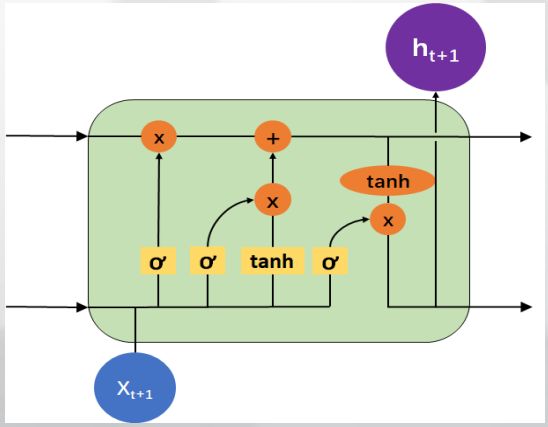


图3-20 LSTM单元结构设计示意图

3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

图像分类算法中经典的深度学习算法模型有：

1. LeNet-5
2. AlexNet
3. GoogleNet
4. VGGNet
5. ResNet

3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

1. LeNet-5

- LeNet-5网络包含5层网络

  1. 2个卷积池化层（卷积层加下采样层）和3个全连接层。
  2. LeNet5的具体的网络结构设计示意图以及每一层的网络参数设置，其中包括卷积特征数量、全连接层神经元个数等如图所示。

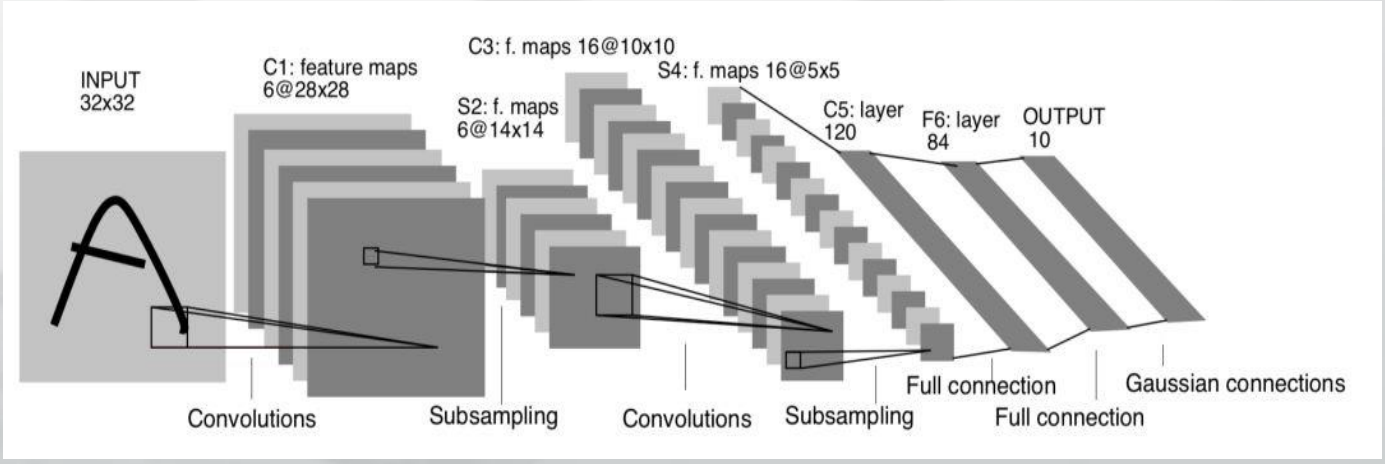
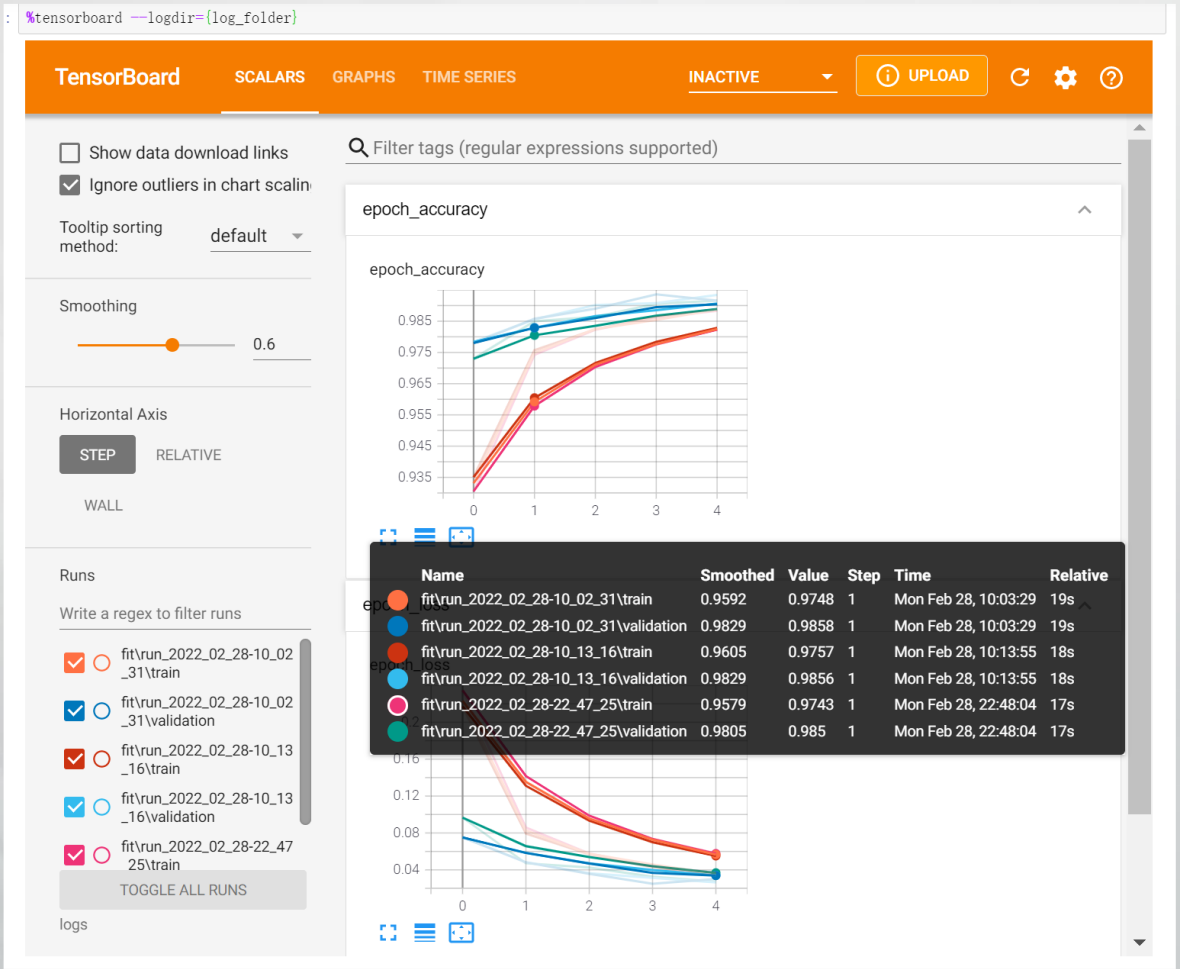


图3-22 LeNet5网络架构示意图

3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

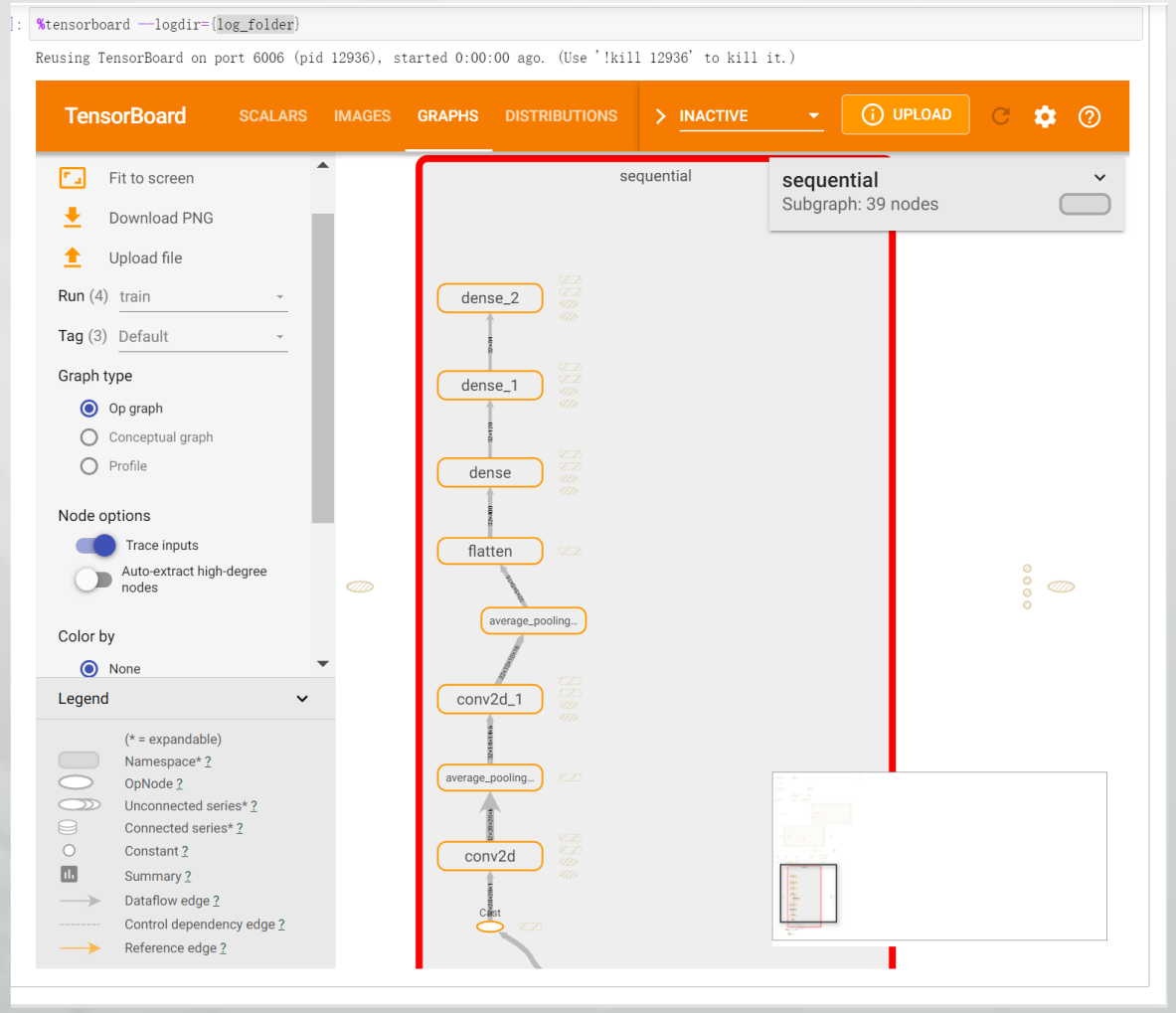
- LetNet-5



3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

- LetNet-5



3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

2. AlexNet

- AlexNet是第一个真正意义上的深度网络，与LeNet5的5层相比，它的层数增加了3层
- 输入也从28变成了224
- AlexNet包含3个卷积池化层（卷积加池化层）、2个卷积层和3个全连接层
- 网络结构中每卷积层的卷积核维数和个数以及全连接层中的神经元个数设置如图所示。

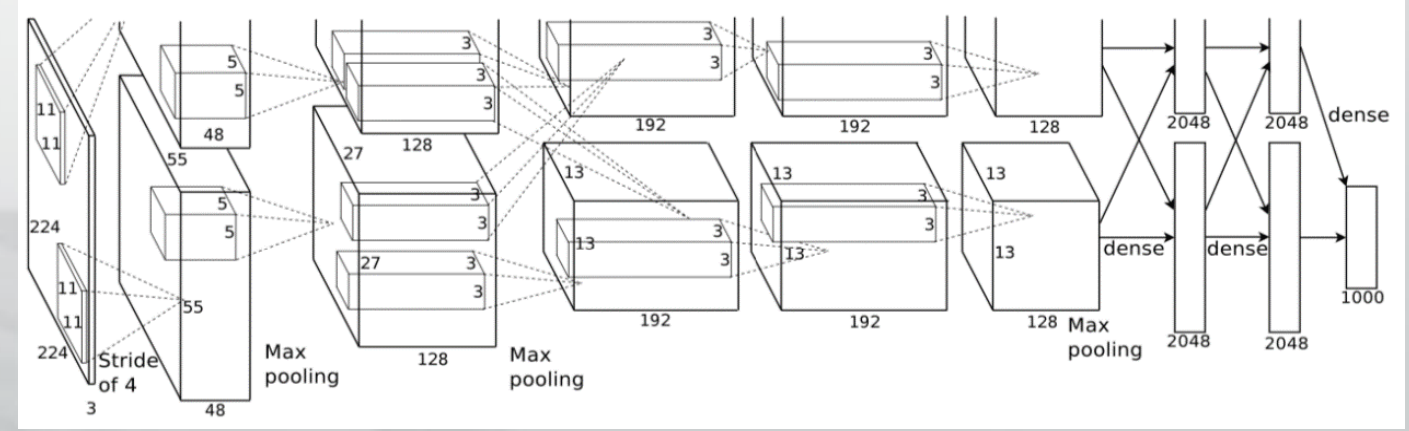


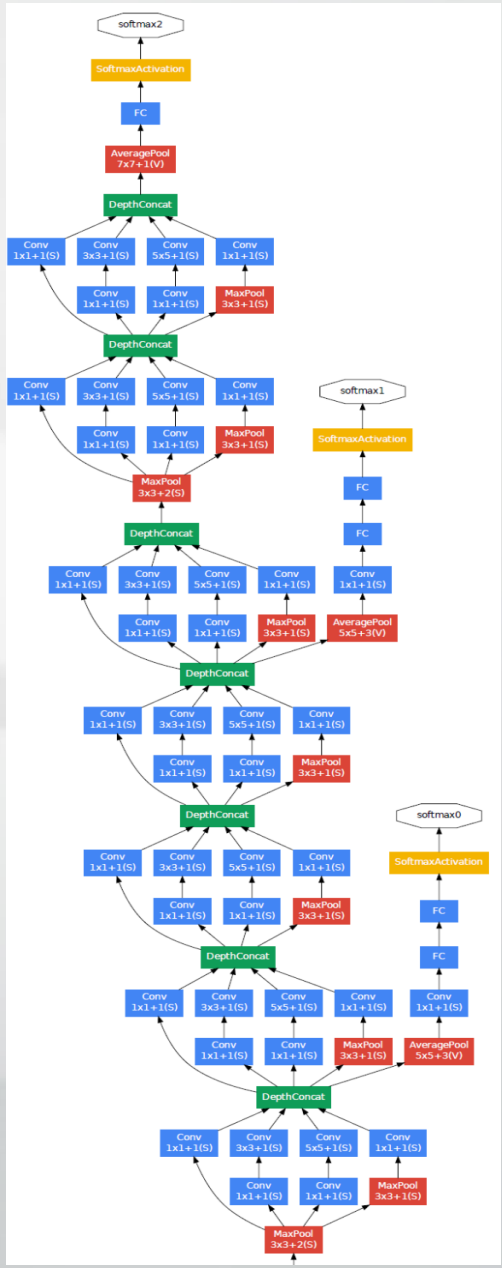
图3-23 AlexNet网络结构示意图

3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

3. GoogLeNet/Inceptions

- Inception模块显著地减少网络中参数的数量
- 没有使用卷积神经网络中的全连接层
- 使用了平均池化



(a)GoogLeNet第二部分　　(b)GoogLeNet第一部分

图3-24 GoogLeNet网络结构图

3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

4. VGGNet

- VGGNet的名字源于牛津大学Visual Geometry Group的缩写，是该组提出了一组深度学习网络结构的总称
- VGG16包含了16个卷积/全连接层，所有的卷积使用的基本都是3×3的卷积，所有的池化使用的都是2×2的池化。
- 图3-25给出了所有VGG网络的设置，表中conv后面的数字表示的分别是感受野的大小和特征通道的个数，即conv（感受野大小）-（特征通道的个数）。

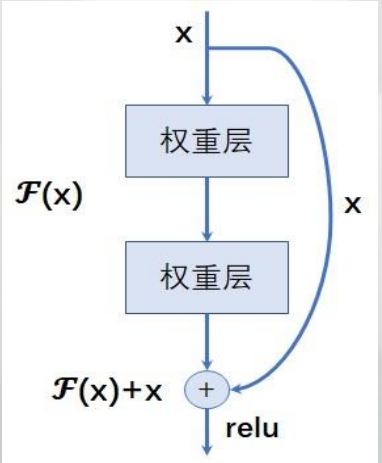
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

5. ResNet

- ResNet提出了一种新型的跳跃链接
- ResNet的这个单元可以将处理前的原始数据进行保留并与处理后的数据$\mathcal{F}$(x)相加，从而保留原始数据的特征。
- ResNet网络中大量使用了批量归一化（batch normalization）处理。



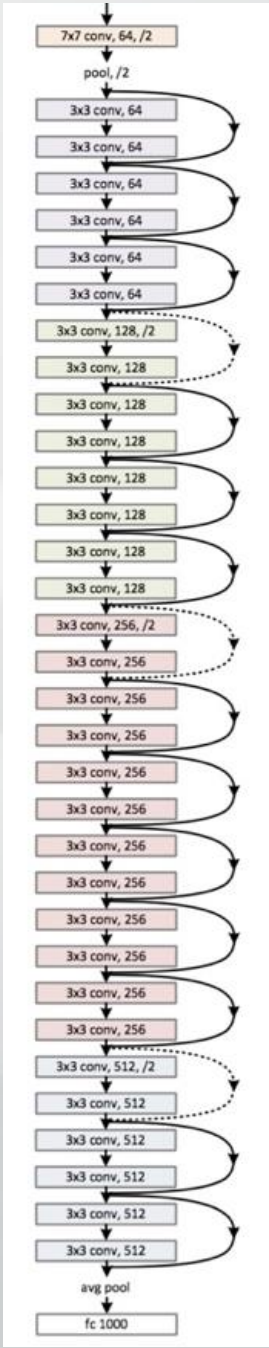
（a）ResNet中跳跃链接的设计

3.4 基于深度学习的图像分类

3.4.2 经典的图像分类模型

5. ResNet

- 图中给出了ResNet的整个网络架构示意图，ResNet包含34层网络结构
- 同VGGNet类似，它基本上都采用了3×3的卷积核，
- 只有在最开始的图像输入的第一层卷积层采用了7×7的卷积核。



（b）ResNet整个网络结构

3.4 基于深度学习的图像分类

3.4.3 深度学习算法与传统算法的比较

- 相似的地方，例如词袋模型中通过低层特征对图像进行编码的过程与深度学习算法中卷积层作用类似，深度学习算法中的卷积层实际上也是提取低层特征然后编码的过程
- 不同之处在于：
  - 词袋模型中特征提取和表示过程都是预先定义好的，不能根据具体的问题或者设定的目标对这个过程进行监督或者优化
  - 深度学习模型中端到端的学习方式能够更好的结合具体的问题，对低层特征特征的提取和描述进行指导