中国科学院大学
University of Chinese Academy of Sciences

NSSC
中国科学院国家空间科学中心
National Space Science Center, CAS

基于深度卷积神经网络的星系形态分类研究

戴加明
导师：佟继周
中国科学院国家空间科学中心
中国科学院大学
daijiamingdl@gmail.com
2018.05.14

目录

- 研究背景与研究内容
- 星系形态分类标准
- 深度卷积神经网络CNN
- ResNet-26网络结构设计与实验结果分析
- 星系图像表征与结果分析
- 工作总结

1 研究背景与研究内容

星系物理研究与星系形态分类

- 现代天体物理学的三个主要研究方向：恒星物理、星系形成与演化、宇宙起源
- 星系形态是表示星系结构最直观的观测特征，它是不同运动状态的恒星轨道在天空中的投影
- 星系形态与星系的形成与演化有着密切的联系，是探究星系物理的重要参数，例：

✓早型星系：颜色偏红、星族年龄偏老年、椭圆星系

✓晚型星系：颜色偏蓝、星族成分较为年轻、有恒星盘以及旋臂结构的盘状星系

椭圆星系M60与旋涡星系NGC 4647

研究不同类型星系的物理特征，首先要做的是有效区分星系的不同形态

天文大数据时代的新挑战

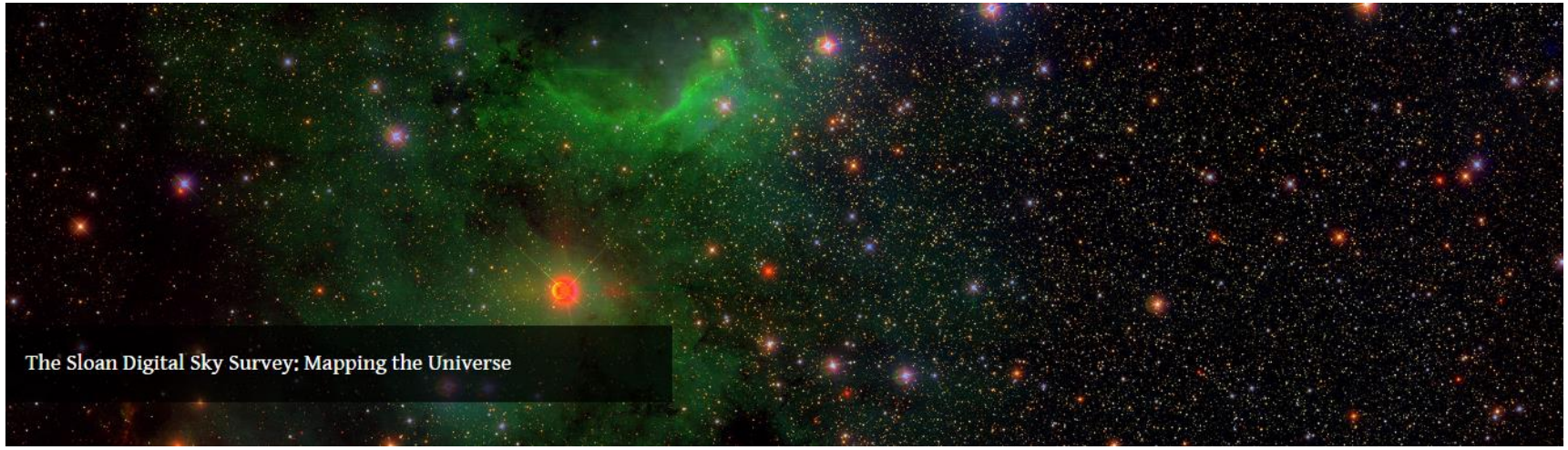
- 斯隆数字巡天（SDSS）三分之一天区观测，即获取12 亿张观测图像;
- 大口径全天巡视望远镜（LSST）计划每晚可产生15 TB 的原始观测数据；
- COSMOS 巡天，LAMOST巡天

星系观测数据呈现爆炸式增长

人工处理

自动化处理

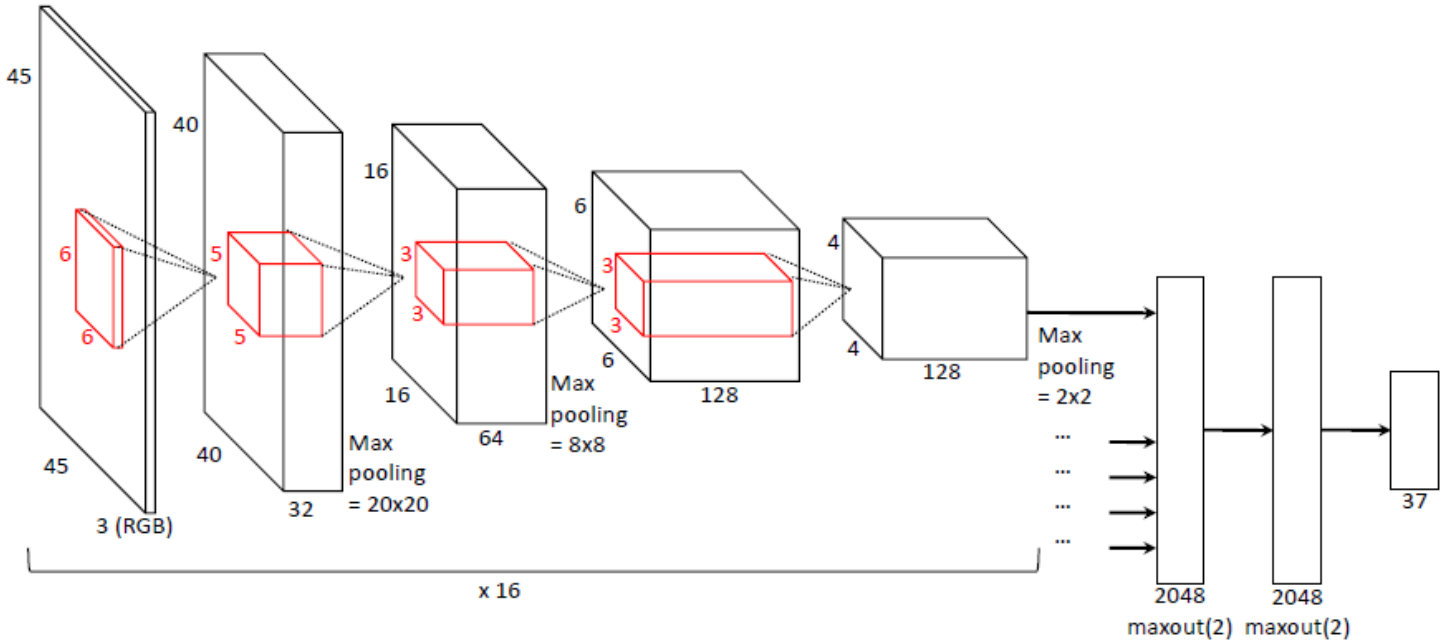
- 实时目标证认、特征提取、天体识别、随动观测优先级确认
- 海量、高维数据的实时处理和挖掘，新的科学问题的发现
- ……



自动化星系形态分类方法发展历程

模型提出者	年份	特征提取	分类器	
Naim	1995	bugle size, the number of arms	ANN	传统机器学习 ① 特征工程+分类器 ② 数据集相对较小 ③ 分类数目相对较少(2, 3, 5类)，而且随着分类数目的增多，分类准确率快速下降
Owens	1996	13 parameters	Decision tree, ANN	
De La Calleja	2004	PCA	ANN, LWR, Ensemble	
Banerji	2010	colors, shapes, concentration, texture	ANN	
Gauci	2010	photometric, spectra parameters	Decision tree, Fuzzy Logic	
Ferrari	2015	Concentration, Asymmetry, Smoothness, Gini coefficient, Moment, Entropy and Spirality	LDA	
Dieleman	2015	-	CNN	深度学习 自动提取特征
Gravet	2015.11	-	CNN	
Kim	2016.08	-	CNN(Star/galaxy)	
Aniyan	2017.05	-	CNN(Radio)	

Dieleman15 -自动化星系形态分类方法的里程碑

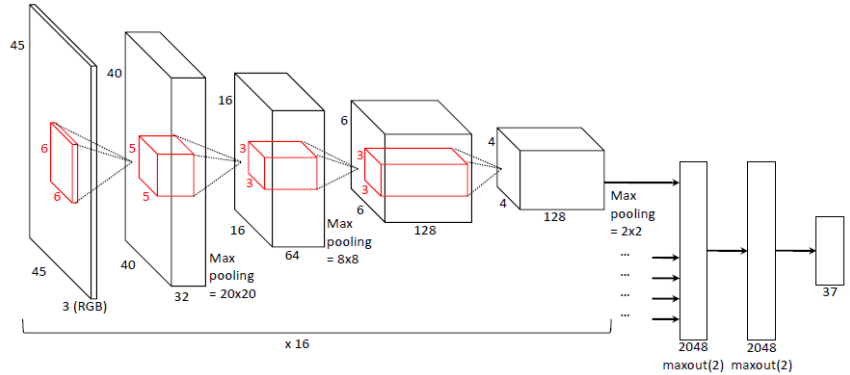


- 2013-2014 Galaxy Zoo-The Galaxy Challenge比赛冠军
- 2015年发表
- 7层（4个卷积层和3个全连接层）
- 4200万参数

Dieleman15

首次将深度学习技术应用于星系形态分类

Dieleman S, Willett K W, Dambre J. Rotation-invariant convolutional neural networks for galaxy morphology prediction[J]. Monthly notices of the royal astronomical society, 2015, 450(2): 1441-1459.



- Dieleman（2015）是基于深度卷积神经网络模型提出的自动星系形态分类方法，近几年来卷积神经网络模型在图像分类领域表现出的优异性能，是否能够应用到自动星系形态分类中进一步提升分类性能？

- 经神经网络训练后可以获得高维星系表征，这些表征除了用于星系形态分类，还有哪些用途？是否可以对高维星系表征数据进行降维和可视化，尝试挖掘其中的隐含的其他信息？这些隐含的信息对星系形态分类的后续研究是否有帮助？

研究内容

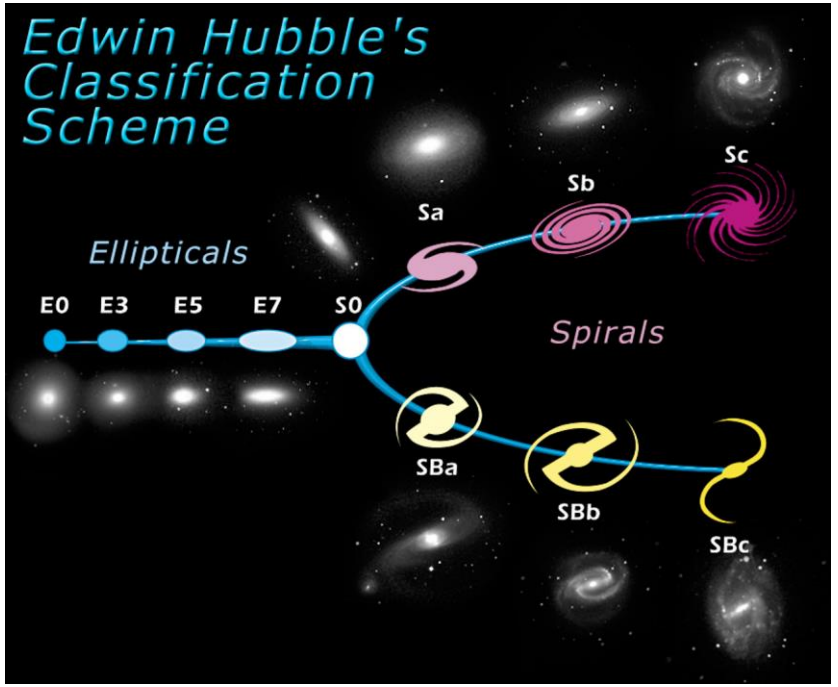
1. 调研星系形态分类标准，分析各类标准优缺点，选择一种形态分类方法作为星系形态分类模型的分类标准。
2. 调研以深度卷积神经网络模型为代表的深度学习技术的发展情况，分析模型算法与技术先进性、训练技巧以及优缺点等。
3. 基于深度残差网络，尝试改进残差单元，探索网络宽度与网络性能的关系，结合数据样本特点，设计一种改进的基于深度残差网络的星系形态分类模型，并进行实验验证，并与其他模型比较分析。
4. 开展高维星系表征数据可视化研究，运用高维数据可视化技术，针对神经网络训练得到的高维星系表征数据进行降维，可视化呈现星系图片本身潜在的全局结构和局部结构信息，寻找离群点，尝试分析星系形态高维抽象表征的内在联系与规律。

2 星系形态分类标准

——目视分类系统、模型化分类系统、非模型化分类系统

星系形态分类标准 — 目视分类系统

- 目视分类系统：直接凭借眼睛来对星系的形态进行分类
- 典型分类方法：哈勃序列（1926）、德沃库勒分类系统、叶凯士分类系统和范登伯分类系统等
- 优点：很好地用于近邻星系形态分类；
- 缺点：人为主观性强；耗时；不适用于高红移星系的分类



哈勃序列 [ESA/Hubble]

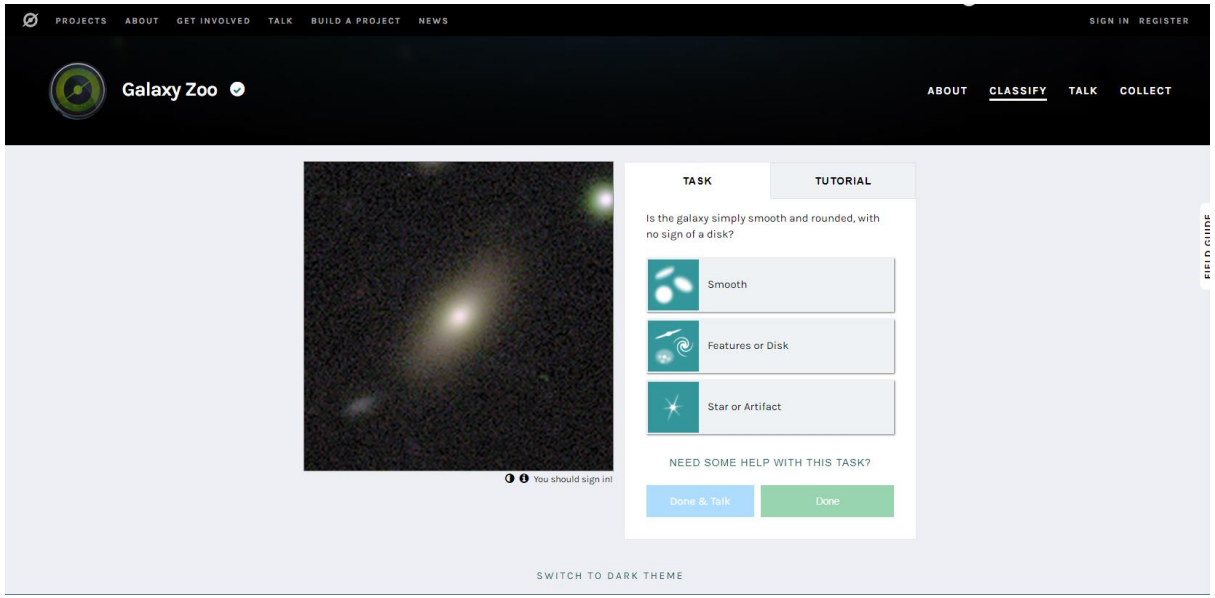
哈勃序列根据星系的光学波段形态特征进行分类，和许多星系的物理参数有关，至今仍被广泛使用。

星系形态分类标准 —模型化、非模型化分类系统

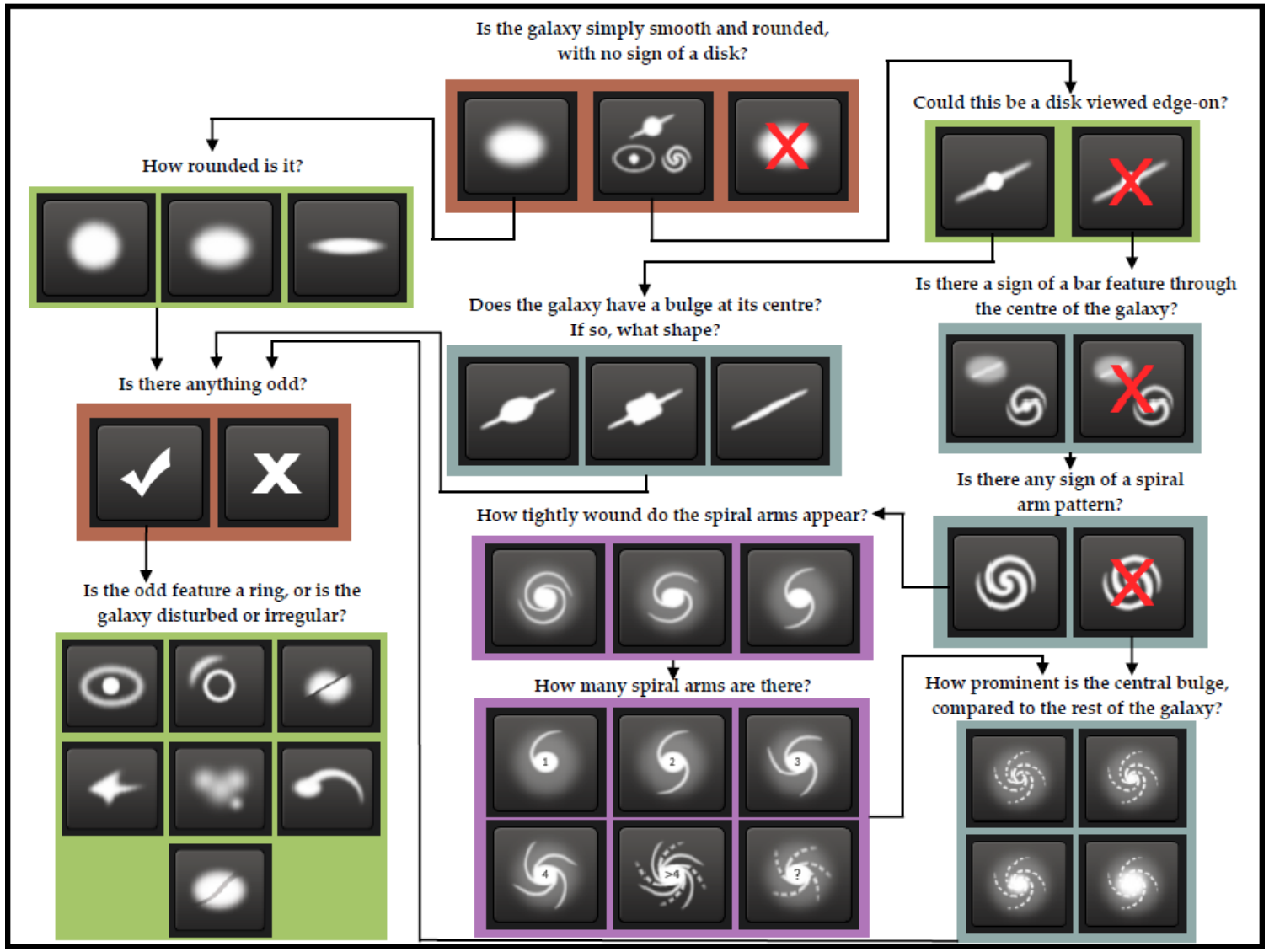
	模型化分类系统	非模型化分类系统
定义	使用星系的面亮度轮廓来对星系形态进行分类，星系的面亮度轮廓用不同的经验律($r^{1/4}$ 律或$r^{1/4}$ 律+ 指数律) 进行拟合	基于表示星系形态的结构参数来对星系形态进行分类，其参数包括聚集度指数(C) 、非对称指数（A）、簇聚指数（S）、基尼系数（G）和矩指数（$M20$）等
优点	可以用于大样本星系形态分类	适合大星系样本的星系形态的研究，具有丰富的物理内涵
缺点	• 依赖于假设的函数分布； • 没有考虑星系中存在棒、透镜或环等成份； • 对于高红移星系，限制于望远镜的观测分辨率，很难区分核球和盘	半自动化分类方法，需要天文学家进行综合判断，确定星系所属类别

星系动物园

- 什么是星系动物园（Galaxy Zoo，GZ）？
  - ✓ 在线众包项目，志愿者依据星系动物园决策树进行星系形态分类
  - ✓ 志愿者分类的图片随机出现，每张图片被40 到50 人分类，计算累计得分值，其结果与专业天文学家的分类结果高度一致。
- 星系动物园项目依次发布了第一代（GZ1）、第二代（GZ2）、第三代（Galaxy Zoo: Hubble）和第四代星系动物园（Galaxy Zoo: CANDELS），得到了大规模带标签的高质量星系图像数据集。



星系动物园2决策树



Task	Question	Responses	Next
01	*Is the galaxy simply smooth and rounded, with no sign of a disk?*	smooth	07
		features or disk	02
		star or artifact	end
02	*Could this be a disk viewed edge-on?*	yes	09
		no	03
03	*Is there a sign of a bar feature through the centre of the galaxy?*	yes	04
		no	04
04	*Is there any sign of a spiral arm pattern?*	yes	10
		no	05
05	*How prominent is the central bulge, compared with the rest of the galaxy?*	no bulge	06
		just noticeable	06
		obvious	06
		dominant	06
06	*Is there anything odd?*	yes	08
		no	end
07	*How rounded is it?*	completely round	06
		in between	06
		cigar-shaped	06
08	*Is the odd feature a ring, or is the galaxy disturbed or irregular?*	ring	end
		lens or arc	end
		disturbed	end
		irregular	end
		other	end
		merger	end
		dust lane	end
09	*Does the galaxy have a bulge at its centre? If so, what shape?*	rounded	06
		boxy	06
		no bulge	06
10	*How tightly wound do the spiral arms appear?*	tight	11
		medium	11
		loose	11
11	*How many spiral arms are there?*	1	05
		2	05
		3	05
		4	05
		more than four	05
		can't tell	05

星系动物园2数据集使用方法

方法1:直接使用图像标签中的概率值

方法2:在方法1的基础上进一步选取干净样本

干净样本的选取规则：满足设定的阈值

如选取旋涡星系：$f_{features/disk} \geq 0.430$,
$f_{edge-on,no} \geq 0.715, f_{spiral,yes} \geq 0.619$

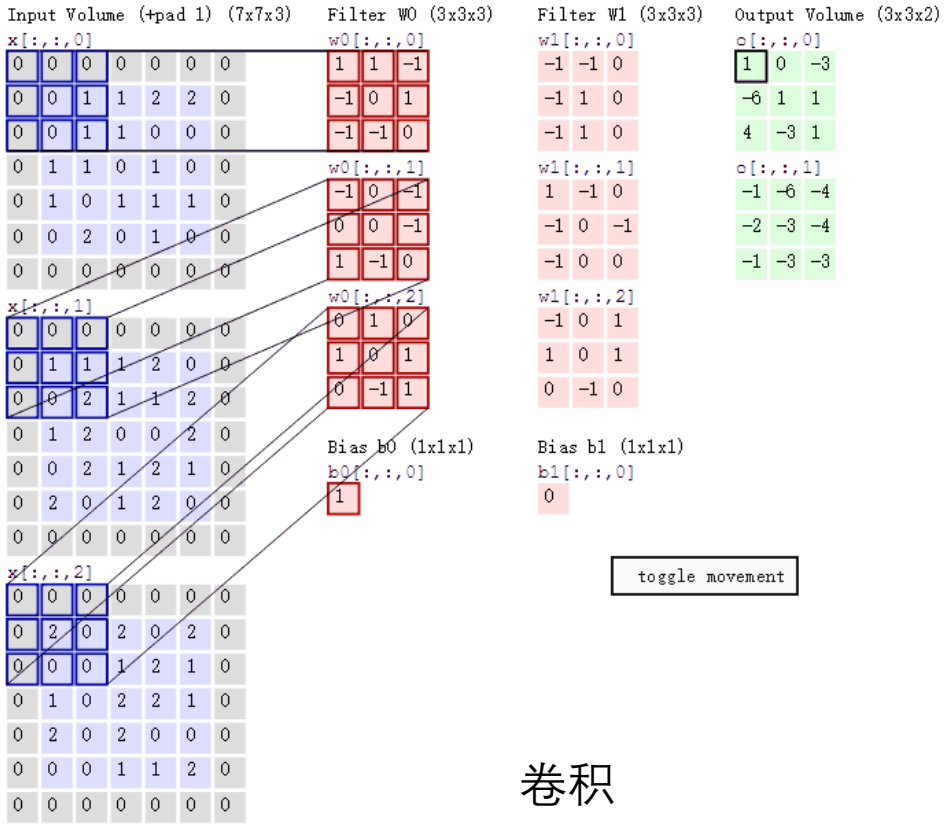
干净样本阈值选取规则

Task	Previous task	Vote fraction
01	-	-
02	01	0.430
03	01,02	0.715
04	01,02	0.715
05	01,02	0.715
06	-	-
07	01	0.469
08	06	0.420
09	01,02	0.602
10	01,02,04	0.619
11	01,02,04	0.619

3 深度卷积神经网络CNN

深度卷积神经网络 CNN

- 是专门用来处理具有类似网格结构数据（例如图像数据）的神经网络
- 可直接使用图像的原始像素作为输入，自动地提取有效特征，避免了复杂的特征工程，同时具有对缩放、平移、旋转等畸变不变性，具有很强的泛化性。
- 构成：卷积层、池化层、全连接层

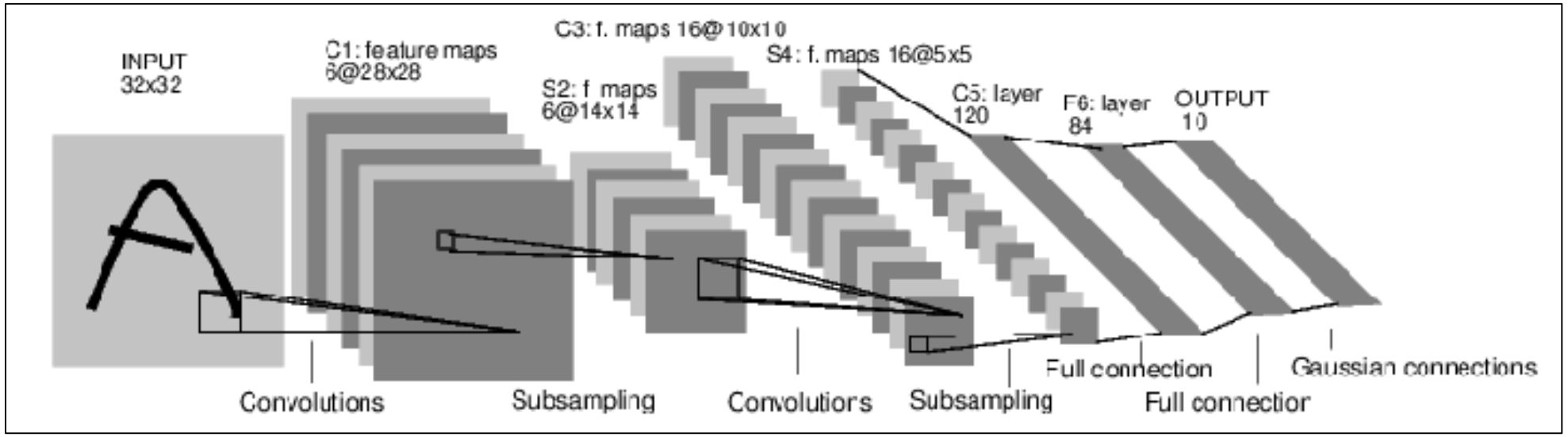


卷积

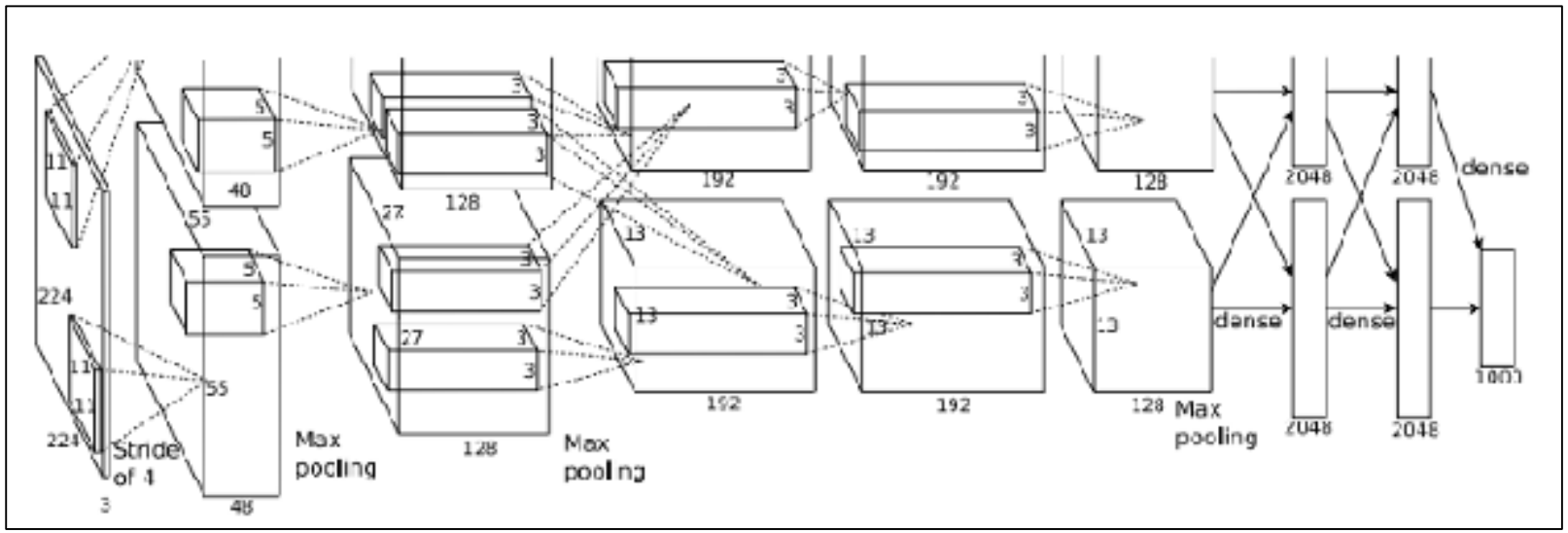
- 经典的CNN 结构： LeNet5、 AlexNet 、VGG 、Google Inception、ResNet

图片来源：http://cs231n.github.io/assets/conv-demo/index.html

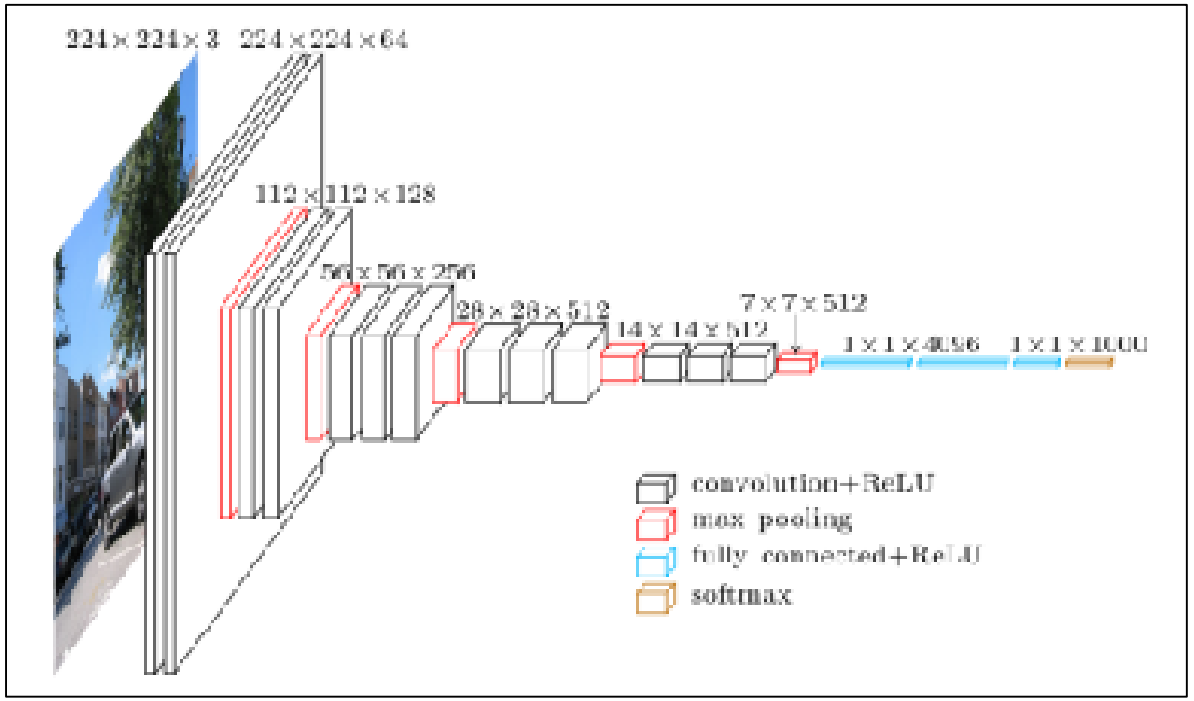
LeNet5



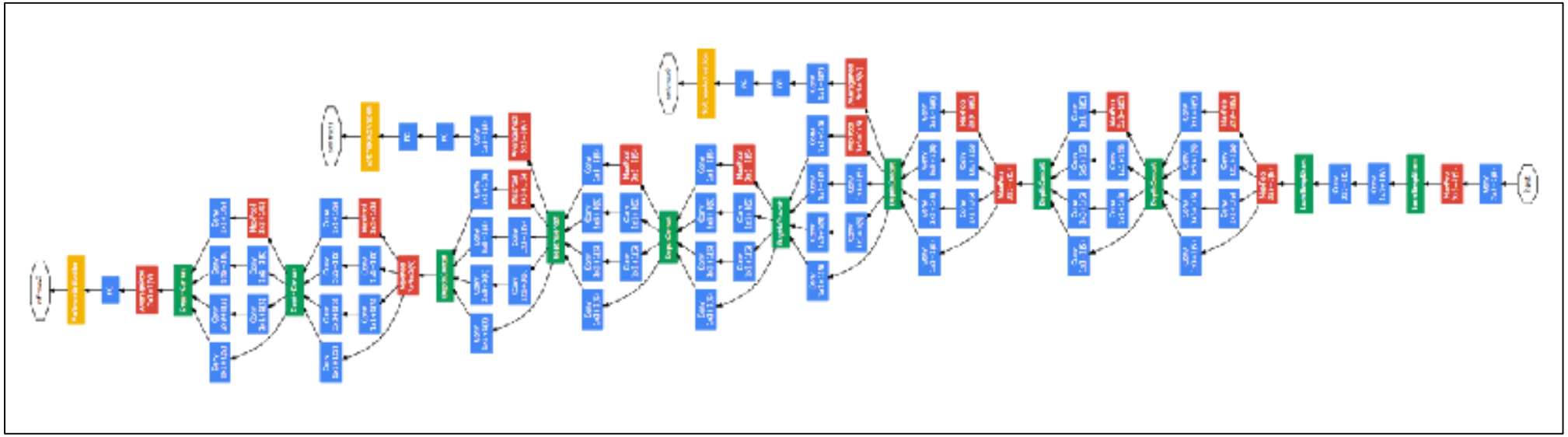
AlexNet



VGG



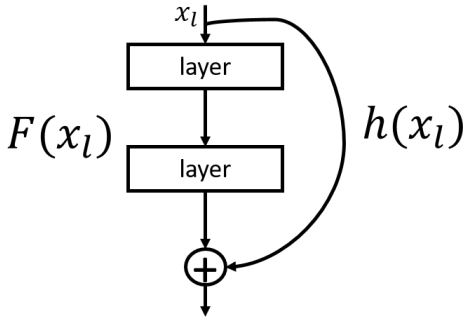
Google Inception



深度残差网络ResNet

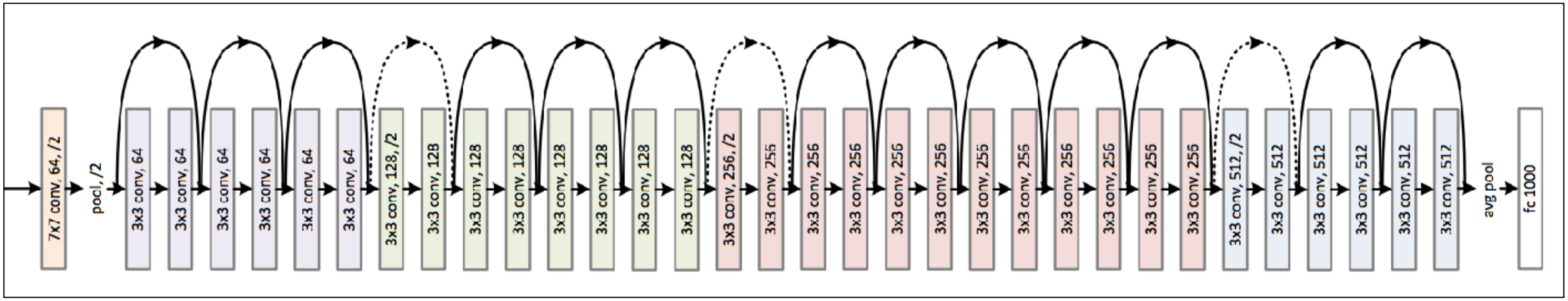
残差单元:假设某段神经网络的输入是x，期望输出是H(x)，如果直接把输入x传到输出作为初始结果，那么需要学习的目标就是F(x)=H(x)-x，即残差

短连接



残差单元

$$x_{l+1} = f\big(h(x_l) + F(x_l)\big)$$
$$h = identity$$
$$f = identity$$
$$x_{l+1} = x_l + F(x_l)$$



4 ResNet-26网络结构设计与实验结果分析

网络结构设计

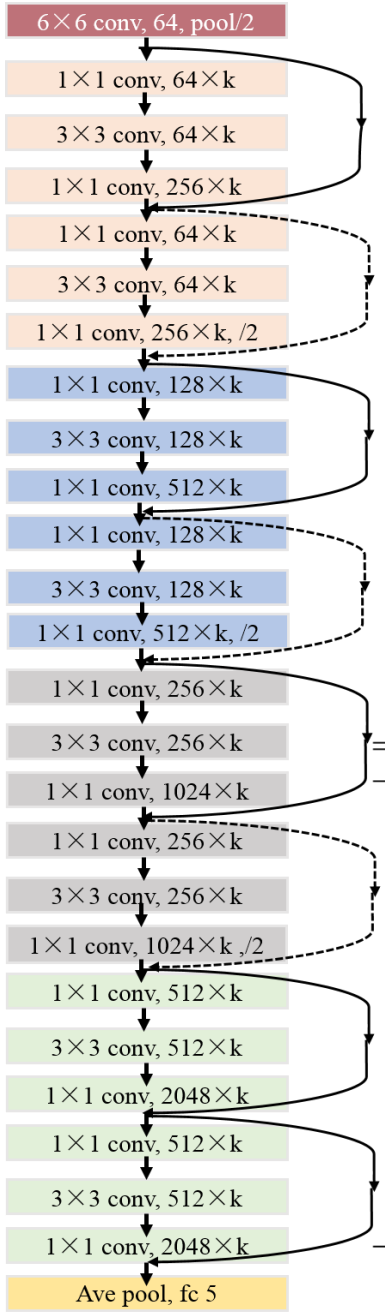
改进的残差单元：

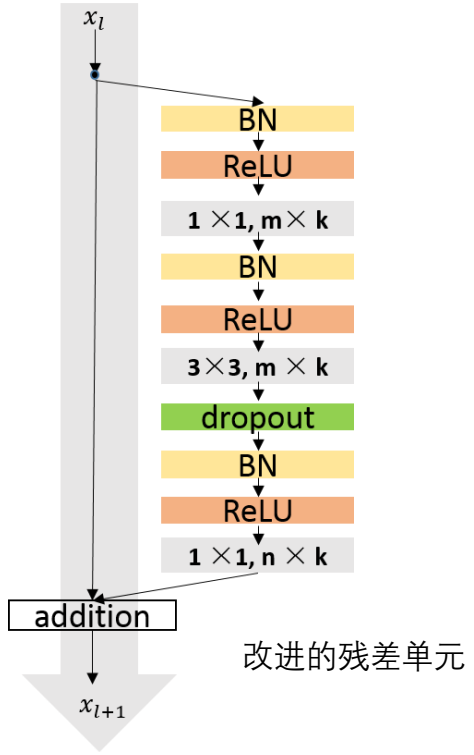
- 3 个卷积层：1 x 1，3 x 3，1 x 1卷积；
- 前2个卷积层的通道数一样，第3 个卷积层的通道数一般为前2个卷积层通道数的4 倍；
- 使用预激活“pre-activation” 方式，即“BN-ReLU-Conv ”；
- 在3 x 3 卷积之后加入dropout 。

网络结构：

- 第一个卷积层卷积核，尺寸为6x6，通道数为64，步长为1，后为2 x 2，步长为2 的最大池化层；
- 最大池化层之后为4个卷积组：Conv2，Conv3，Conv4 和Conv5，其中每个卷积组包含N 个残差单元；
- 最后一层是平均池化层，大小为4 x 4，池化层的输出为1 x 1 x 4096，最后作用于一个5 个神经元的全连接层*softmax*。

在设计的残差网络模型中，改进了残差单元，减少了网络的层数，加宽了网络的宽度（通道数目），同时融合了Dieleman 模型的优点，网络的总层数为$2N$ +2层





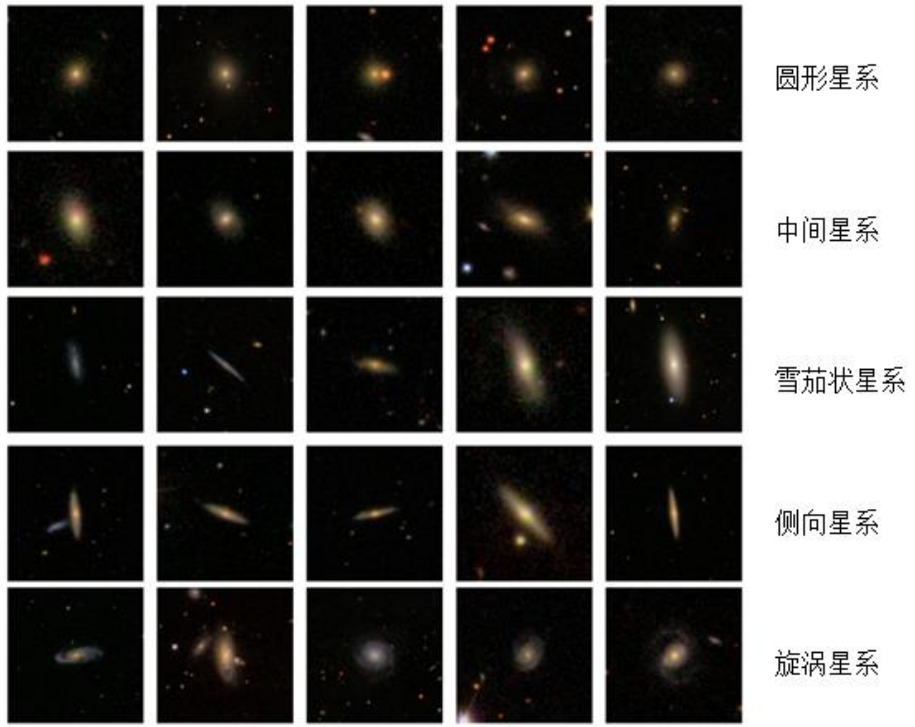
改进的残差单元

Layer name	Output size	Depth
Conv 1	64 × 64	6 × 6, 64
Max-pooling	32 × 32	2 × 2, stride 2
Conv 2	16 × 16	$\begin{bmatrix} 1\times1, 64\times k \\ 3\times3, 64\times k \\ 1\times1, 256\times k \end{bmatrix} \times N$
Conv 3	8 × 8	$\begin{bmatrix} 1\times1, 128\times k \\ 3\times3, 128\times k \\ 1\times1, 512\times k \end{bmatrix} \times N$
Conv 4	4 × 4	$\begin{bmatrix} 1\times1, 256\times k \\ 3\times3, 256\times k \\ 1\times1, 1024\times k \end{bmatrix} \times N$
Conv 5	4 × 4	$\begin{bmatrix} 1\times1, 512\times k \\ 3\times3, 512\times k \\ 1\times1, 2048\times k \end{bmatrix} \times N$
Avg-pooling	1 × 1	4 × 4, 5 – d, $softmax$

改进的深度残差网络结构

数据集选取

- 数据集选自Galaxy Zoo - The Galaxy Challenge，星系图片来自SDSS DR7，共61578张，大小为424×424×3像素，标签采用GZ2的分类标准
- SDSS对星系的观测包括5 个光学波段（u、g、r、i和z），一般取前3 个波段（u、g和r）合成相应的RGB 星系图像
- 依据干净样本的阈值选取规则，选取5类星系：圆形星系、中间星系、雪茄状星系、侧向星系和旋涡星系，总计28790张图片



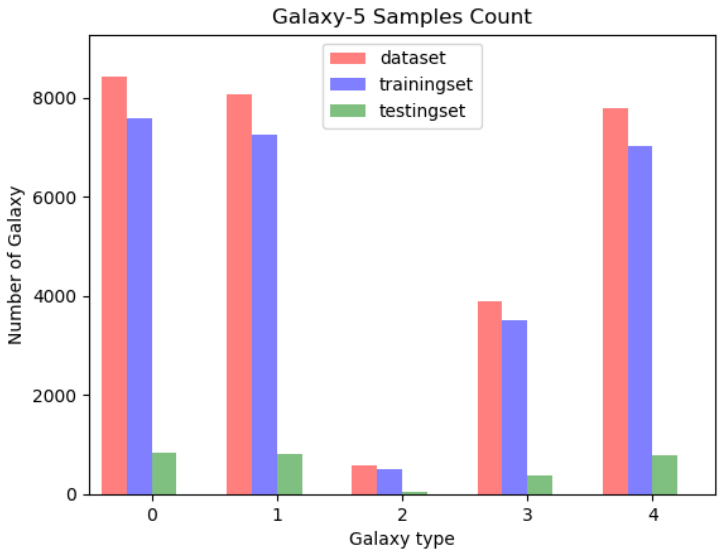
GZ2 中随机抽取的5类星系图片

数据集选取

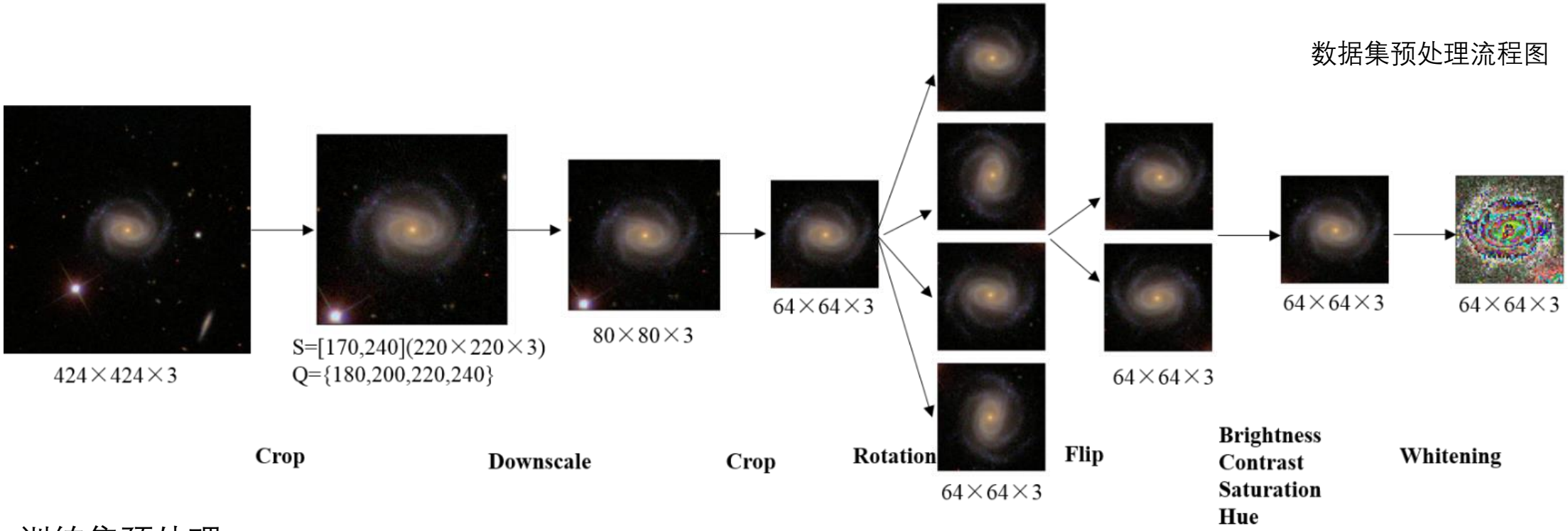
- 将28790 张干净样本按9：1 的比例划分训练集和测试集

	Completely round smooth 0		In-between smooth 1		Cigar-shaped smooth 2		Edge-on 3		Spiral 4		Total
	N_1	Proportion	N_2	Proportion	N_3	Proportion	N_4	Proportion	N_5	Proportion	
Training set	7591	29%	7262	28%	520	2%	3513	14%	7025	27%	25911
Testing set	843	29%	807	28%	58	2%	390	14%	781	27%	2879
Data set	8434	29%	8069	28%	578	2%	3903	14%	7806	27%	28790

- 训练集和测试集中星系图片数目是同分布的



数据预处理



训练集预处理

- 中间裁剪（S 训练尺度抖动）
- 下采样
- 随机裁剪
- 随机旋转
- 水平翻转
- 光学畸变
- 图像白化

数据增强，
可使数据量增加15万倍

测试集预处理

- 中间裁剪（Q 测试尺度抖动）
- 下采样
- 中间裁剪
- 图像白化

模型确定实验-超参数选择

超参数的设置是决定模型最终分类性能的关键

1. 第一个卷积层卷积核大小与网络性能的关系
2. 加宽因子k和每组残差单元数N与网络性能的关系

1

Size of Conv 1	Accuracy%）
3×3	92.1181
6×6	95.2083 √
7×7	93.7153

2

k,N	Number of layers	Accuracy%）
$k=1, N=1$	14	93.4028
$k=1, N=2$	26	94.7569
$k=2, N=1$	14	93.0556
$k=2, N=2$	26	95.2083 √

结论：当第一层卷积核的尺寸为6 X 6，加宽因子k为2，每组残差单元数为 2 以及 Dropout 为 0.8 时，模型具有最优的分类性能，称为ResNet-26，共计2600万参数。

3. Dropout 大小与网络性能的关系
4. 星系图像类型与网络性能的关系

3

Dropout	Accuracy（%）
0.5	92.8819
0.7	94.2917
0.8	95.2083 √

4

	Test(Q)	Accuracy(%)	F1	AUC
Gray	240	93.4722	0.9342	0.9786
RGB √	220	**95.2083**	**0.9515**	**0.9823**

ResNet-26分类结果

精确率，召回率，F1值

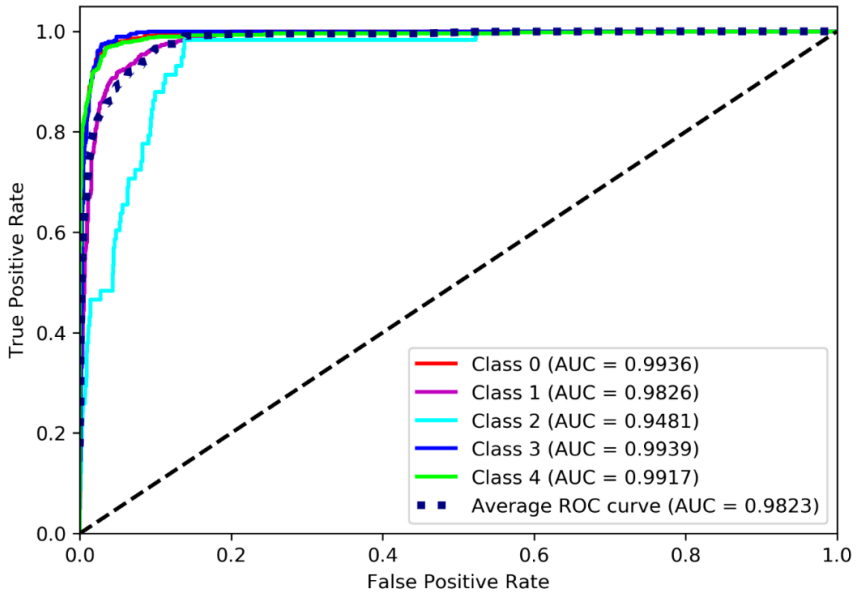
Class	Precision	Recall	F1
0	**0.9611**	0.9634	0.9622
1	0.9561	0.9431	0.9495
2	0.7234	0.5862	0.6476
3	0.9412	0.9485	0.9448
4	0.9573	**0.9782**	**0.9677**
Average	0.9512	0.9521	0.9515

混淆矩阵

	0	1	2	3	4
0	**815**	21	0	0	10
1	29	**762**	0	0	17
2	0	4	**34**	18	2
3	0	3	12	**368**	5
4	4	7	1	5	**763**

注：列为真实标签，行为预测标签，0、1、2、3、4分别表示圆形星系、中间星系、雪茄状星系、侧向星系和旋涡星系。

5个星系类的ROC曲线



结论：5个类别星系的分类准确率分别为:圆形星系，96.6785%；中间星系，94.4238%；雪茄状星系，58.6207%；侧向星系，94.3590%和旋涡星系，97.6953%；

模型在圆形星系、中间星系、侧向星系和旋涡星系的分类识别过程中均表现优秀，F1值都在0.94以上；其中圆形星系、侧向星系和旋涡星系的AUC值都超过0.99。模型针对旋涡星系的分类性能最好，其分类准确率达到了97.6953%；针对雪茄状星系的分类性能表现一般，训练集中的雪茄状星系图片数量相对较少是造成分类性能一般的一个原因。

不同网络模型分类性能对比

不同模型的测试平均准确率

Model	Image side		Accuracy(%)
	Train(S)	Test(Q)	
Dieleman(Dieleman et al. 2015)	[170,240]	180,200,220,240	93.8800
AlexNet(Krizhevsky et al. 2012)	[170,240]	180,200,220,240	91.8230
VGG(Simonyan & Zisserman 2014)	[170,240]	180,200,220,240	93.1336
Inception(Szegedy et al. 2016)	[170,240]	180,200,220,240	94.2014
ResNet-50(He et al. 2016a)	[170,240]	180,200,220,240	94.0972
ResNet-26	[170,240]	180,200,220,240	**94.6875**

注：结果取自每一个测试尺度运行10次取最大值，然后再求平均值。高亮标识最好结果。

不同模型的测试准确率，精确率，召回率，F1值和AUC值

Model	Accuracy(%)	Precision	Recall	F1	AUC
Dieleman(Dieleman et al. 2015)	94.6528	0.9455	0.9465	0.9456	0.9793
AlexNet(Krizhevsky et al. 2012)	92.2569	0.9207	0.9226	0.9215	0.9809
VGG(Simonyan & Zisserman 2014)	93.6458	0.9348	0.9365	0.9353	0.9846
Inception(Szegedy et al. 2016)	94.5139	0.9447	0.9451	0.9448	**0.9852**
ResNet-50(He et al. 2016a)	94.6875	0.9458	0.9469	0.9461	0.9823
ResNet-26	**95.2083**	**0.9512**	**0.9521**	**0.9515**	0.9823

注：结果基于每一个测试尺度运行10次取最大值，高亮标识最好结果。

在相同的GPU 服务器上，使用相同的训练集和测试集，测试不同模型的准确率、精确率、召回率、F1 值和AUC 值。

结论：通过与Dieleman 模型、AlexNet、VGG、Inception 和ResNet-50 的对比实验，ResNet-26 取得了较优的分类结果。

5 星系图像表征与结果分析

星系图像表征的降维方法t-SNE

- T 分布随机近邻嵌入t-SNE (t-Distribution Stochastic Neighbor Embedding)

是一种非线性降维方法，通常应用于高维数据的可视化，即将高维数据降维至二维或三维空间，具有保留数据局部特征以及揭示全局特征的优点。

假设数据集 X 包含 N 个样本 $X=\{x_1,x_2,\cdots,x_N\}$，每个样本为 D 维向量。t-SNE 的目标是计算投影 $Y=\{y_1,y_2,\cdots,y_N\}$。通常 $y_i\in\mathbb{R}^d$ 对应于 $x_i\in\mathbb{R}^D$。典型地，$d=2$ 且 $D\gg d$.

首先，计算条件概率 $p_{j|i}$。$p_{j|i}$ 表示数据点 x_i 和 x_j 的相似性，当 x_i 和 x_j 邻近时，$p_{j|i}$ 值越大。其定义为

$$p_{j|i}=\frac{exp(-||x_i-x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}exp(-||x_i-x_k||^2/2\sigma_i^2)}. \tag{5.1}$$

其中 σ_i 是是以数据点 x_i 为均值的高斯分布标准差。

然后，定义高维空间中的联合分布 p_{ij}，使其为对称的条件概率：

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2N}. \tag{5.2}$$

下一步，在低维空间中，使用更重长尾分布的 t 分布（自由度为 1）定义联合分布 q_{ij}：

$$q_{ij}=\frac{(1+||y_i-y_j||^2)^{-1}}{\sum_{k\neq l}(1+||y_k-y_l||^2)^{-1}}. \tag{5.3}$$

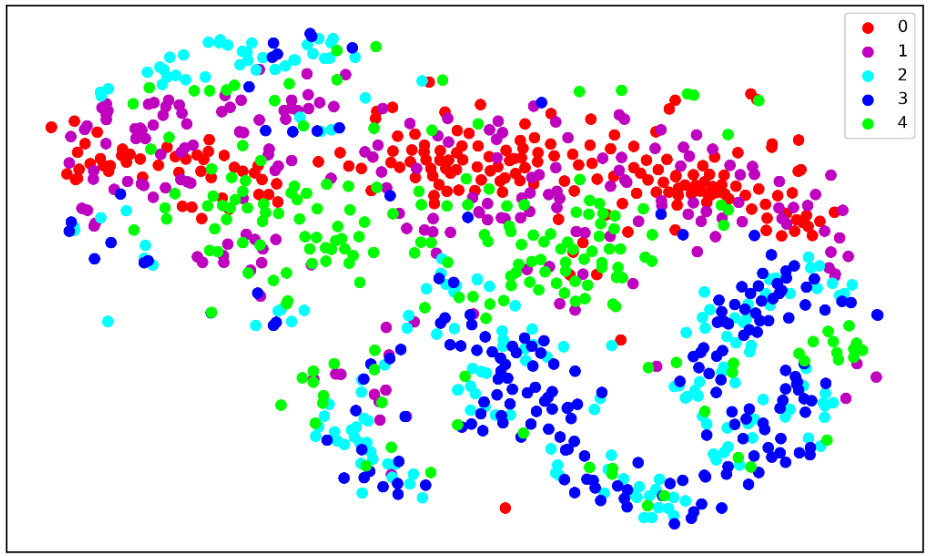
低维空间中数据点 y_i 由最小化分布 Q 和 P 的 Kullback-Leibler 散度（KL 散度）求得，即 t-SNE 的目标函数 C 为

$$C=KL(P|Q)=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}}. \tag{5.4}$$

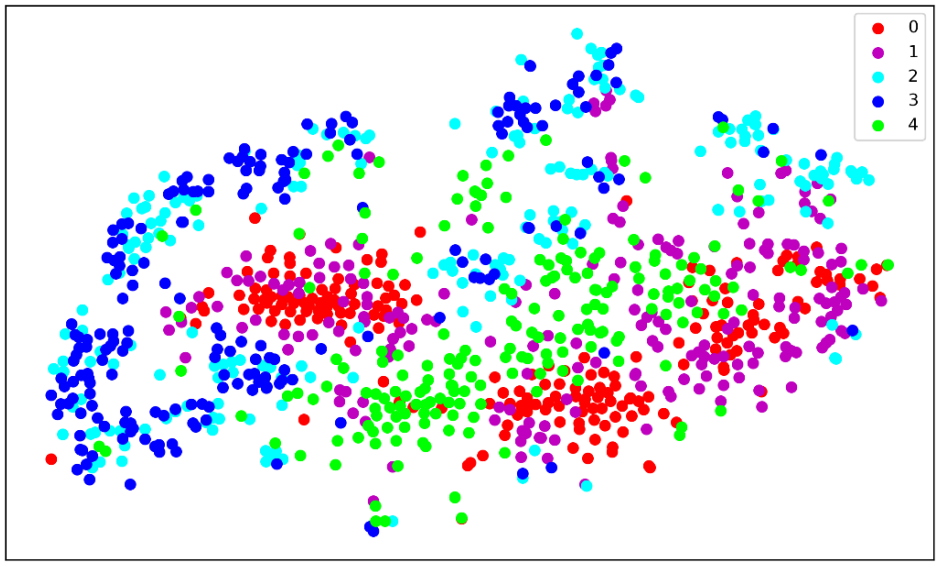
P 和 Q 的 KL 散度使用如下梯度公式求解：

$$\frac{\delta C}{\delta y_i}=4\sum_j(p_{ij}-q_{ij})(y_i-y_j)(1+||y_i-y_j||^2)^{-1}. \tag{5.5}$$

星系图像数据集的t-SNE 可视化及分析



training-1000 原始样本映射

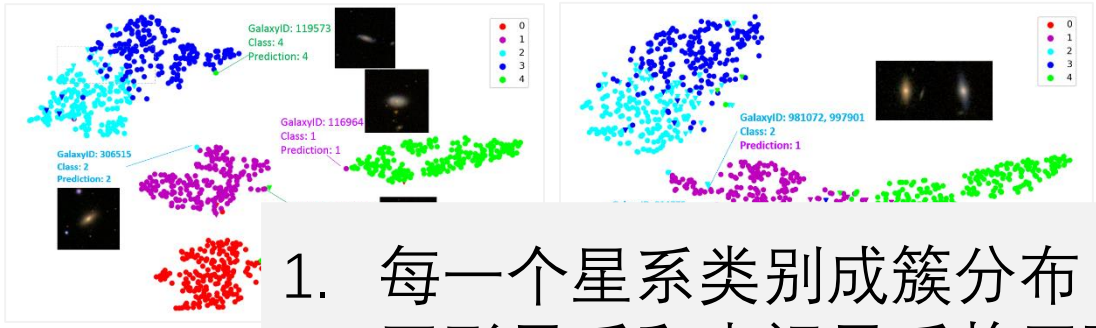


testing-1000 原始样本映射

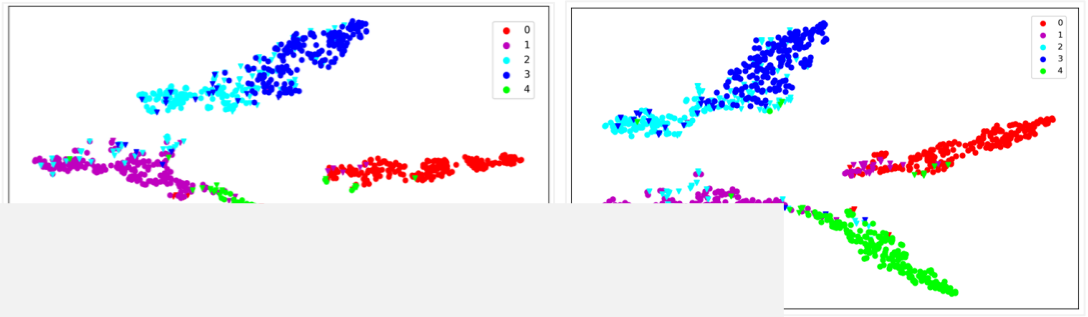
424*424*3=539,328维

星系形态分类模型的t-SNE可视化及分析

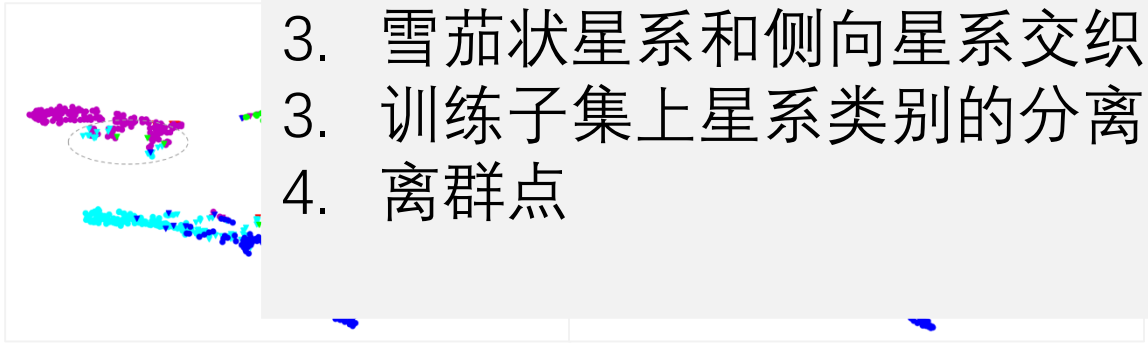
Dieleman 模型最后一层全连接层映射

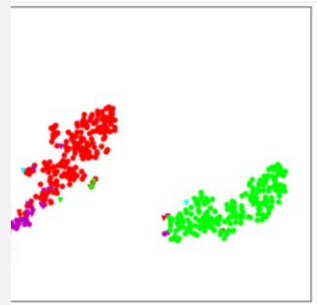


AlexNet 最后一层全连接层映射



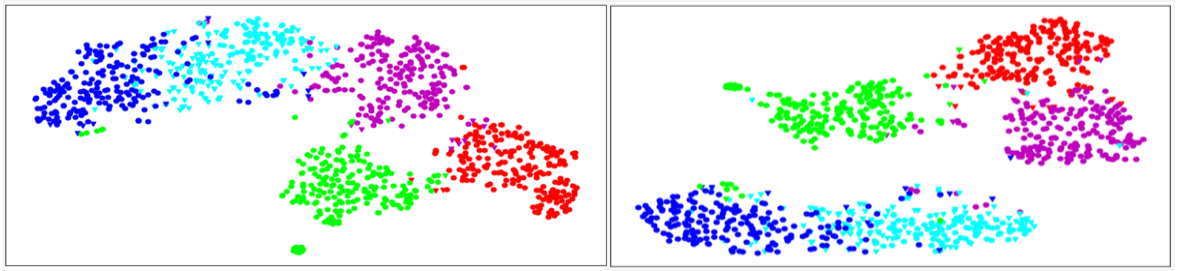
VGG最后一层



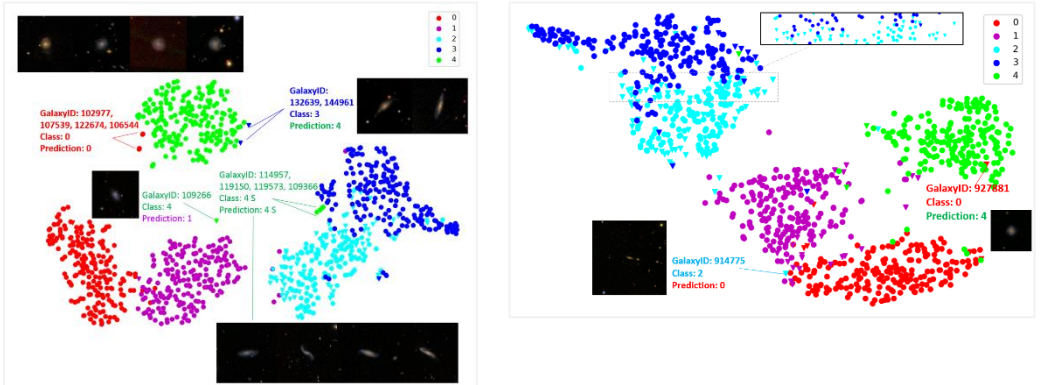


1. 每一个星系类别成簇分布；
2. 圆形星系和中间星系趋于聚合；
3. 雪茄状星系和侧向星系交织缠绕在一起；
3. 训练子集上星系类别的分离程度比测试子集效果好；
4. 离群点

ResNet-50 最后一层平均池化层映射

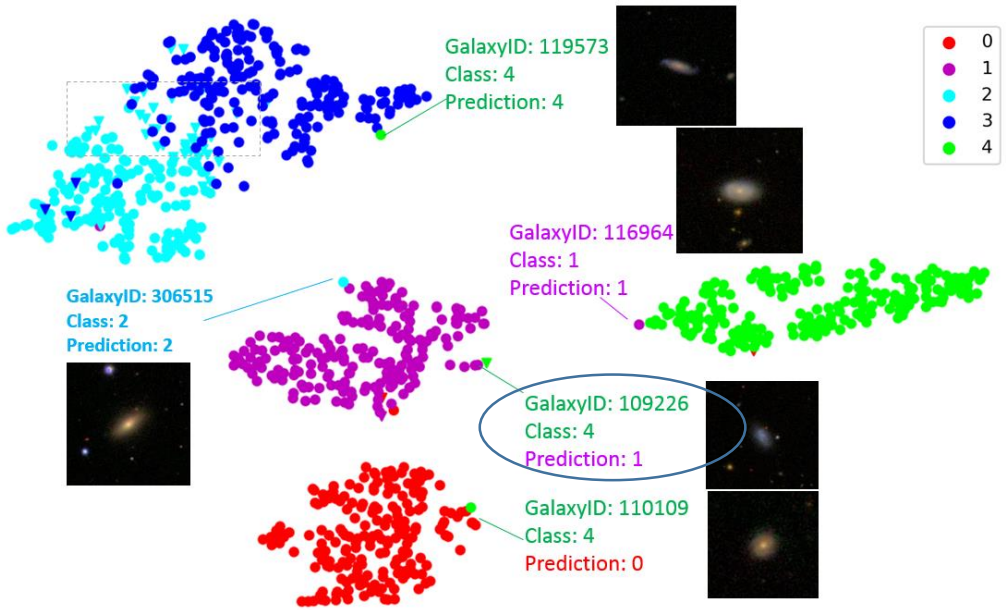


ResNet-26 最后一层平均池化层映射

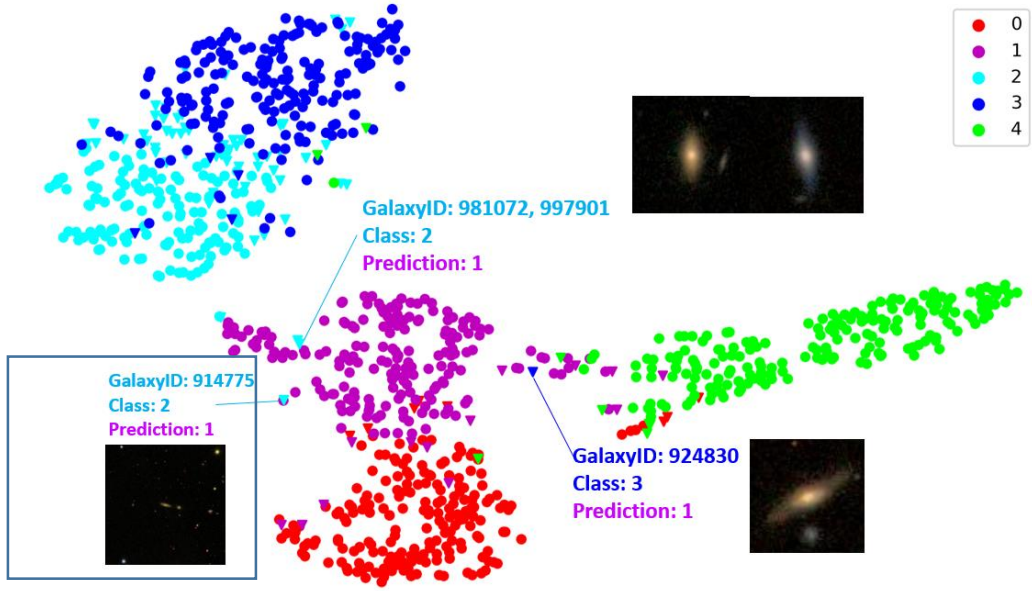


离群点分析

基于 training-1000 的 Dieleman 模型最后一层全连接层映射



基于 testing-1000 的 Dieleman 模型最后一层全连接层映射



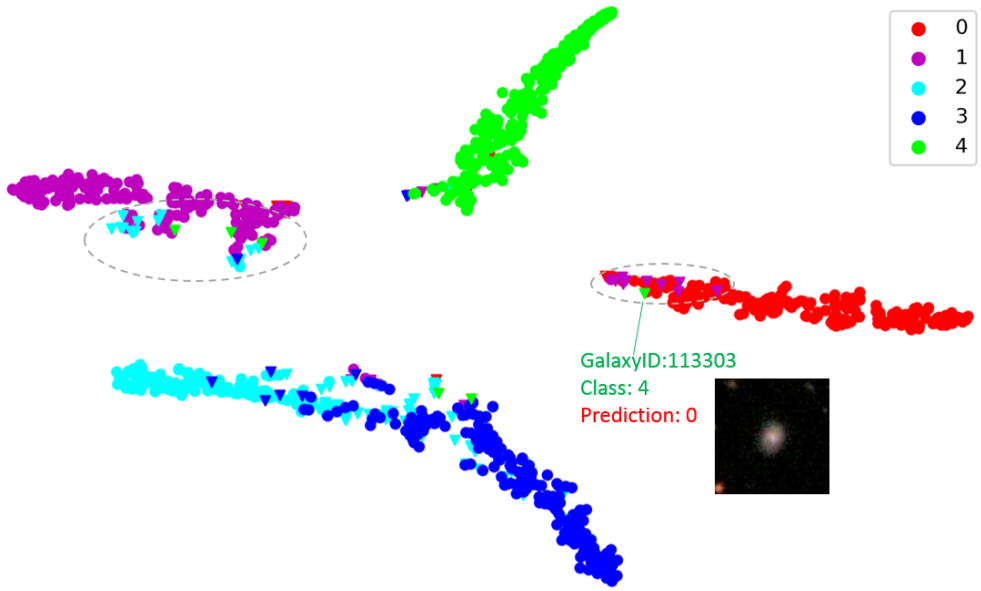
左图：蓝色簇下面的绿黄色的点，它属于旋涡星系，也被分给了旋涡星系，但是却出现在蓝色簇中，这说明它们的结构相似。经检查，这个数据点在数据集中代号（GalaxyID）为119573，看起来和侧向星系很相似。

另一个在红色簇（圆形星系）附近的绿黄色数据点，它属于旋涡星系，却被分给了圆形星系。经检查，这个数据点为代号110109的旋涡星系，其实是圆形星系，也就是说，它的标签是错误的。

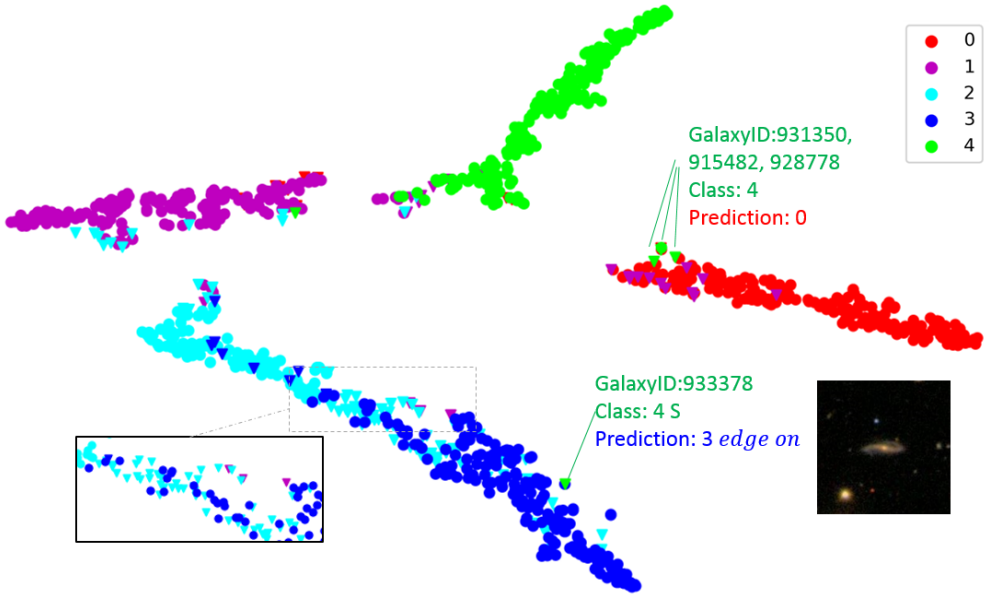
右图：一些离群点，如代号为981072 和997901 的雪茄状星系被预测给了中间星系，代号为924830 的侧向星系被预测给了中间星系。检查之后，发现这些离群点的图片都和中间星系很像，即它们不是典型的原属标签星系。

离群点分析

基于training-1000的VGG最后一层全连接层映射



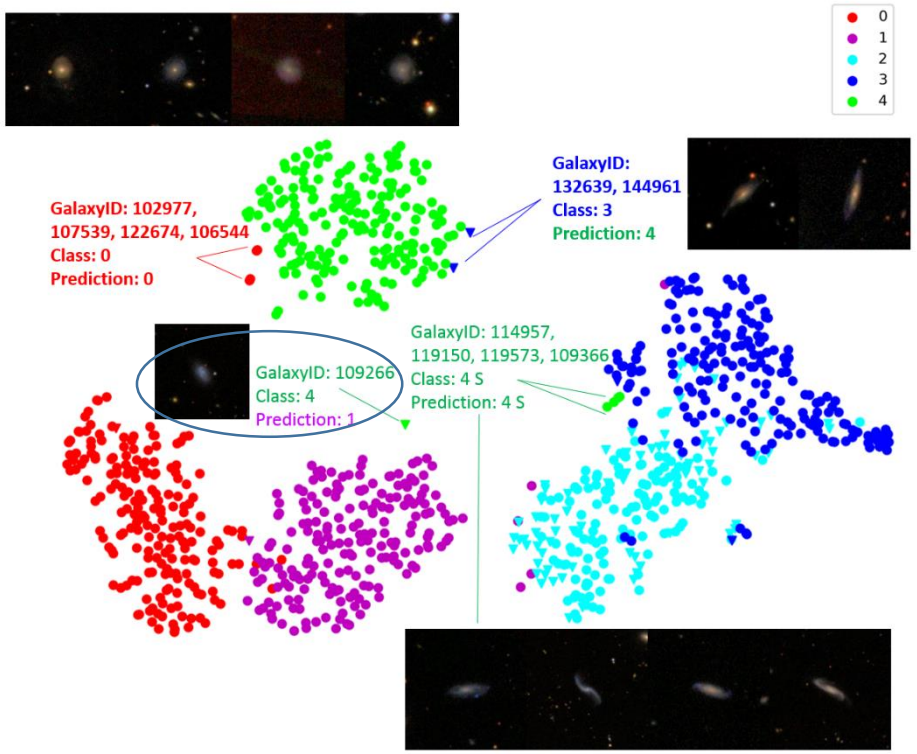
基于testing-1000的VGG最后一层全连接层映射



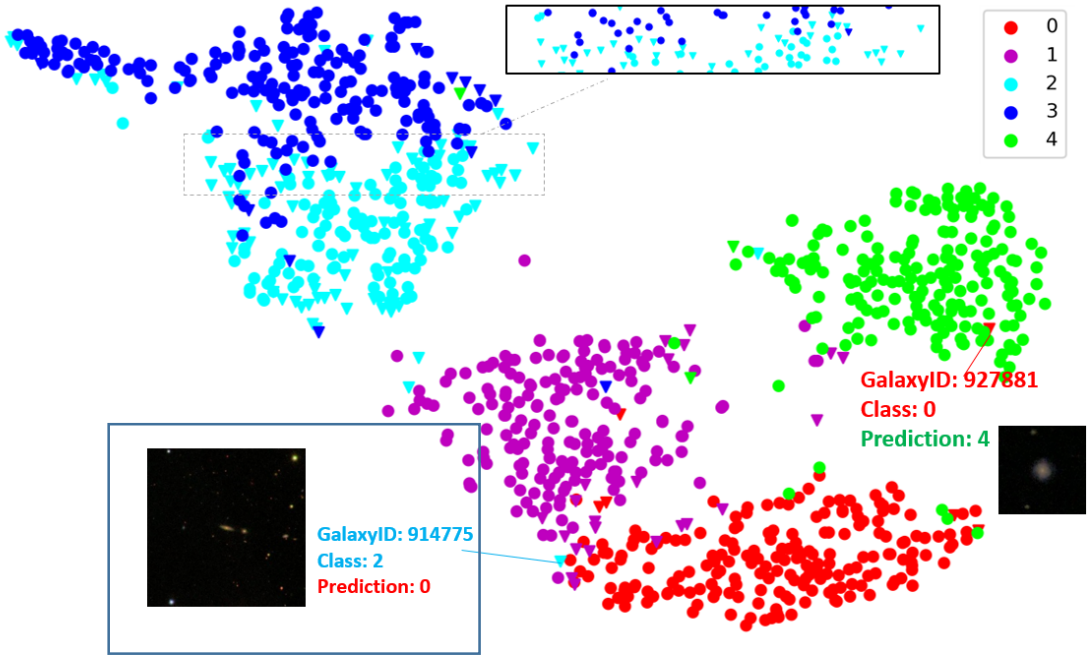
虚椭圆框中：分别有9 张中间星系被错分给圆形星系，16 张雪茄状星系被错分给中间星系。因为圆形星系、中间星系和雪茄状星系都属于平滑星系，它们之间并没有严格的界限去区分它们。可以解释它们是被错分的，也可以解释它们的标签是错误的，但却被模型分类正确。

离群点分析

基于training-1000的 ResNet-26 最后一层平均池化层映射



基于testing-1000的 ResNet-26 最后一层平均池化层映射



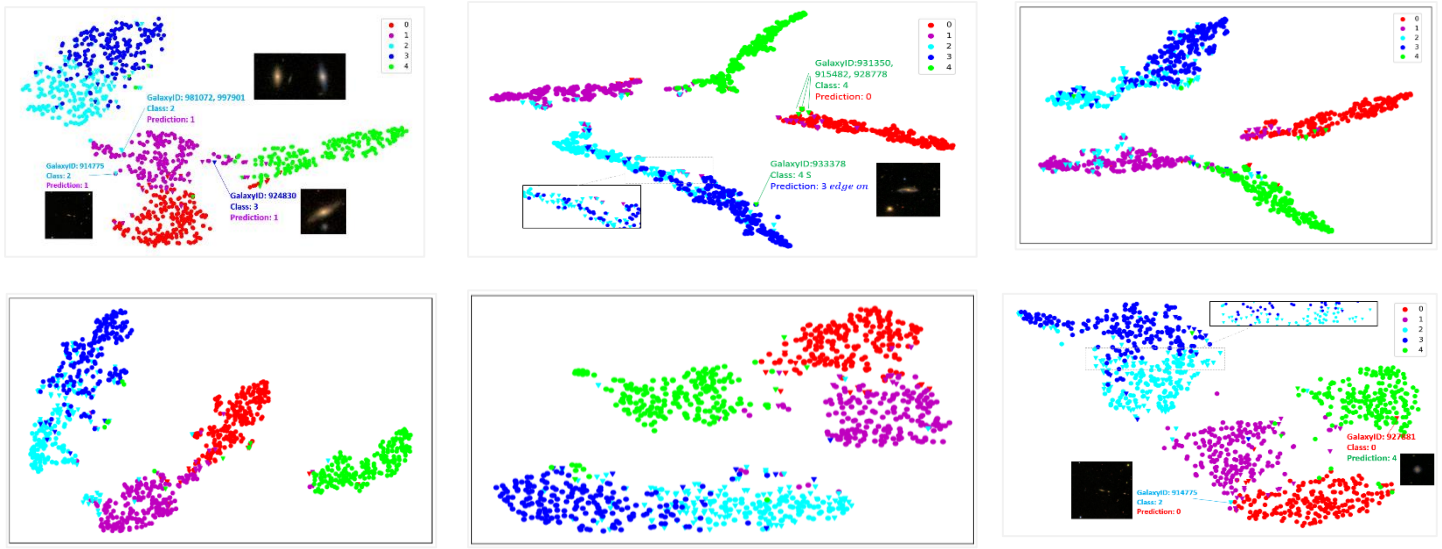
左图：红色离群点（GalaxyID: 102977, 107539, 122674, 106544），它们属于圆形星系，也被预测给了圆形星系，但是它们却出现在绿黄色簇所代表的旋涡星系中，检查之后，发现其中有三张真的圆形星系（GalaxyID: 102977, 107539, 106544）却跟旋涡星系有着相似的结构，另一张代号为122674 的图片其实是旋涡星系，却被标记成圆形星系。

另外4 张旋涡星系（GalaxyID: 114957, 119150, 119573,109366）出现在侧向星系簇和雪茄状星系簇附近，检查之后发现，它们都是瘦长的旋涡星系，看起来非常像侧向星系和雪茄状星系。

离群点分析

- 109266 在training-1000 子集中被标记为旋涡星系，在Dieleman和ResNet-26中全都被错分给中间星系。检查之后发现，它是一张非常模糊的旋涡星系以至于很难识别它。
- 914775 在testing-1000 子集中被标记为雪茄状星系，被Dieleman模型错分给了中间星系，被ResNet-26错分给了圆形星系。检查之后发现，它是一张非常小的图片，星系处于图片中央非常小的位置，很难识别它属于圆形星系、中间星系还是雪茄状星系。

星系交织和趋于聚合现象



雪茄状星系

侧向星系

- 雪茄状星系和侧向星系交织缠绕，从图像上看，雪茄状星系和侧向星系形态相似；
- 圆形星系和中间星系趋于聚合；

✓ 是否可以有助于优化GZ2决策树？

✓ 交织和趋于聚合是否和星系形态分类模型相关？

✓ ……

6 工作总结

工作总结

- 将深度学习领域最新的研究成果引入星系形态分类领域，提出并设计了基于改进的深度残差网络的星系形态分类框架ResNet-26，通过实验验证了模型优良的分类性能；同时实现其他 5 种流行的CNN模型，进行对比实验，证明了ResNet-26较优的分类性能和泛化能力；
- 将高维数据可视化技术（t-SNE）引入星系形态分类的后续研究中，可视化从神经网络中所学习到的高维星系表征，得到一些有价值的发现。

后续工作

- 数据集中各类星系样本数量分布不均衡，针对数据样本较少的雪茄状星系测试准确率不高，后续可考虑从大规模数据集中获取星系类别更全面、样本数量更充分、数据质量更优的星系图片作为训练样本，进一步训练模型；
- 未来除了可尝试运用Inception Module 进一步改进网络结构外，还可以将非模型化分类系统中的结构参数加入到模型中，构建专家系统与神经网络相结合的混合模型，以提升模型的分类性能；
- 在星系图像表征可视化方面，仅利用t-SNE 降维方法，进行了星系高维表征可视化并对可视化后的映射进行了初步分析，未来可与天文学家进行更深层次的信息挖掘工作。

Resources

- Dataset and Code
  - Data: Galaxy Zoo - The Galaxy Challenge (https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge)
  - My models in Tensorflow: https://github.com/Adaydl/

Q & A

Thank you !