Published online: April 18, 2017

Article

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

Automated analysis of high-content microscopy data with deep learning

Oren Z Kraus[1,2,†], Ben T Grys[2,3,†], Jimmy Ba[1], Yolanda Chong[4], Brendan J Frey[1,2,5,6], Charles Boone[2,3,5,*] & Brenda J Andrews[2,3,5,**]

用深度学习自动分析高容量显微数据

陈泽宇

2017.11.29

摘要

- 传统机器学习对于多于一个的数据集就依赖于训练和调参，需要进行大量的分析

- 使用深度卷积神经网络(deep convolutional neural network)对酵母细胞图像进行分析，能够在蛋白的亚细胞定位上的自动分类上表现的更加准确和可靠

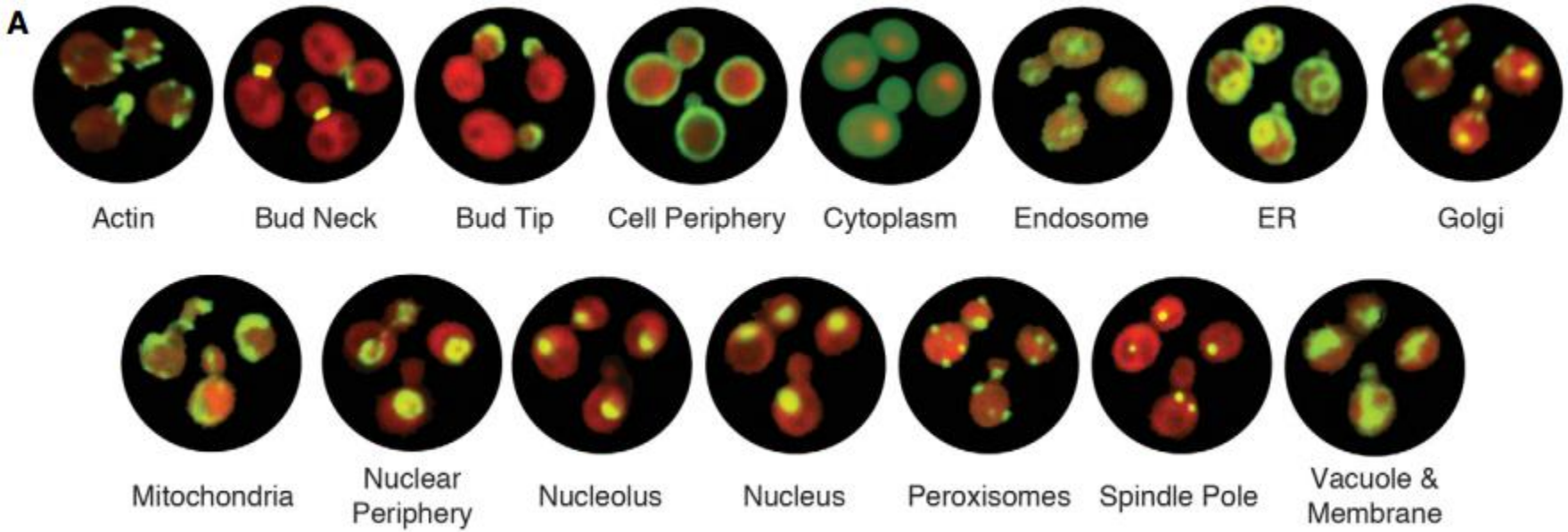
早前的工作

Using SVM

- 60 binary support vector machine(SVM)
- Trainging set containing >70000 cells
- >70% precision and recall
- 用于新的显微数据集还需要进行补充训练

- 通常一个细胞从图像中分割提取出来，包含了数百个像素强度的统计值，这些高维数据然后通过特征提取或者降维后去训练分类器。在数据集和数据集间这些特性是不可以通用的(每个数据集的特性都不进相同)，所以每次都需要去调参、重新分析。

Training and validating a deep neural network (DeepLoc) for classifying protein subcellular localization in budding yeast

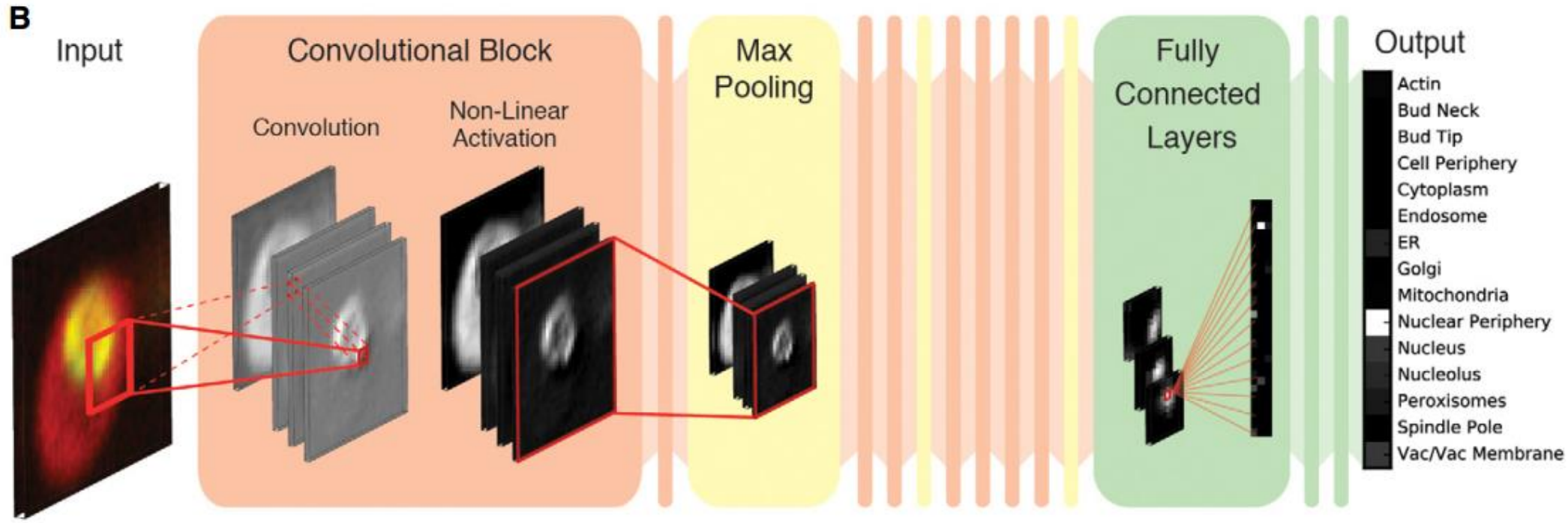


肌动蛋白、芽尖、细胞外围…

使用SVM分类器来鉴别这15个特征，和深层卷积神经网络作比较

Network Construction

Our network arranges 11 layers into eight convolutional blocks and three fully connected layers, consisting of over 10,000,000 trainable parameters in total.



机器如何看图

- 简单来说，每个图像都是一系列特定排序的图点（像素）。如果你改变像素的顺序或颜色，图像也随之改变。

25	2	1	44
223	7	6	60
196	8	2	148
249	1	3	40
60	7	1	154
59	1	7	213
214	7	3	163
89	182	219	13
74	146	113	72
89	18	244	85
1	4	8	97
3	4	2	121
2	1	2	131
7	6	8	47
3	5	5	126
7	6	8	121
5	3	1	237

Layers used to build ConvNets

- 卷积神经网络通常包含以下几种层
- **卷积层（Convolutional layer）**，卷积神经网路中每层卷积层由若干卷积单元组成，每个卷积单元的参数都是通过反向传播算法优化得到的。卷积运算的目的是提取输入的不同特征，第一层卷积层可能只能提取一些低级的特征如边缘、线条和角等层级，更多层的网络能从低级特征中迭代提取更复杂的特征。
- **线性整流层（Rectified Linear Units layer, ReLU layer）**，这一层神经的活性化函数（Activation function）使用线性整流（Rectified Linear Units, ReLU）$f(x)=max(0,x)f(x) = max(0, x)$。
- **池化层（Pooling layer）**，通常在卷积层之后会得到维度很大的特征，将特征切成几个区域，取其最大值或平均值，得到新的、维度较小的特征。
- **全连接层（ Fully-Connected layer）**，把所有局部特征结合变成全局特征，用来计算最后每一类的得分。

  Fully connected layers are then used for classification, in which elements in each layer are connected to all elements in the previous layer.

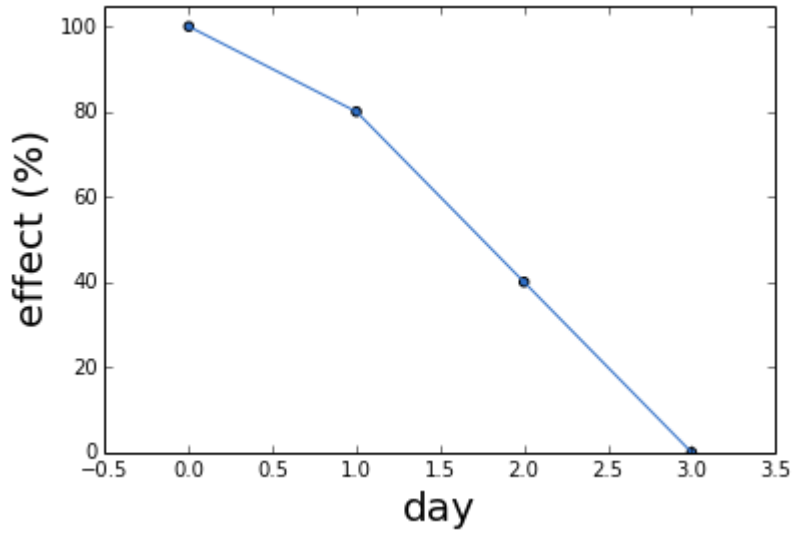
何谓卷积?

公式:$f(x)*g(x)=\int_{-\infty}^{\infty}f(\tau)g(x-\tau)d\tau$

1. $f(x)*g(x)f(x)*g(x)$ 表示 $f(x)f(x)$ 和 $g(x)g(x)$ 的卷积，注意此处自变量为 x；
2.它是对 $(-\infty,\infty)(-\infty,\infty)$ 区间上对 τ求积分；
3.积分对象为两个函数的乘积：$f(\tau)$ 和 $g(x-\tau)$ 。
4.等式右边只有 $g(x-\tau)g(x-\tau)$ 提到了 x，其他部分都在关注 τ

例子

- 试想小明有一段时间每天都要去输液，输的药会在身体里残留直至失效，药效随着时间是不断衰落的。 这里为简便起见，假设药效 4 天就失效，而且药效持续函数是离散的。如下图所示：



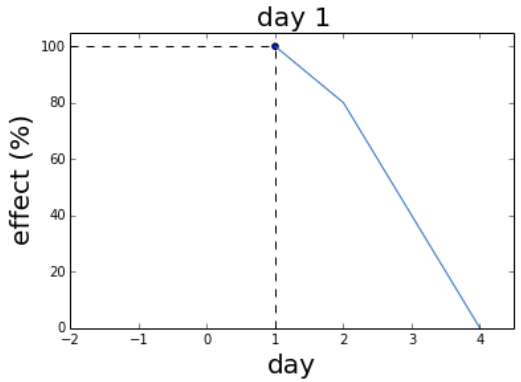
记天数为 t，每天输液的药量为m(t)，药效函数为eff(t)，小明身上残留的药效为 rest(t)
其中药效函数：

$$\text{eff}(t) = \begin{cases} 100\% & t=0 \\ 80\% & t=1 \\ 40\% & t=2 \\ 0\% & t>=3 \end{cases}$$

图中，横坐标为天数，纵坐标为药效。输液当天（day=0）药效为 100%，第二天减弱为 80%，第三天减弱为 40%，第四天减弱为 0。

下面观察一下小明从第一天起，连续三天输液后身上所留下的药效（假设每天药量固定为10）。

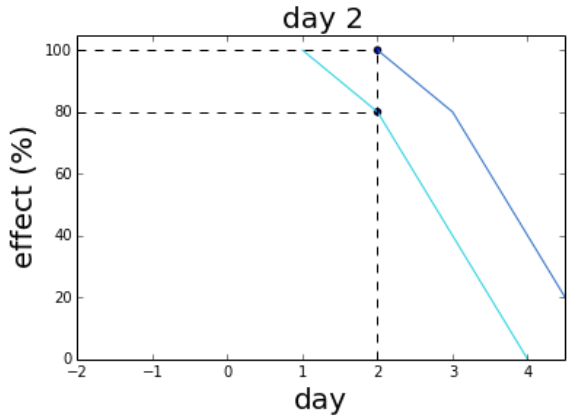
第一天，小明去医院输完液后，药效为 10（rest(t)=m(t)·eff(0)）。



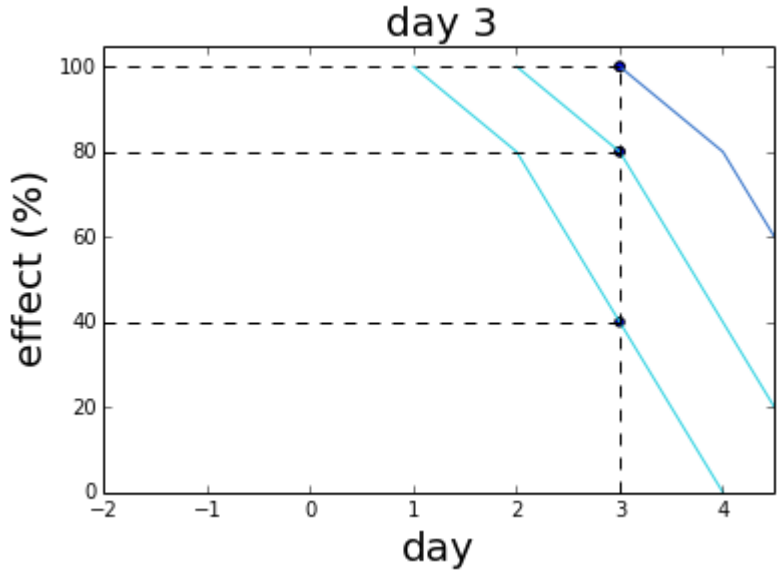
第二天，小明去医院准备输液

输液前，他身上带着前一天的药效，此时已经衰减为 10*80%=8，即 m(t−1)·eff(1)。

输液后，他身上携带的药效为：8 + 10 = 18（rest(t)=m(t−1)·eff(1)+m(t)·eff(0)）

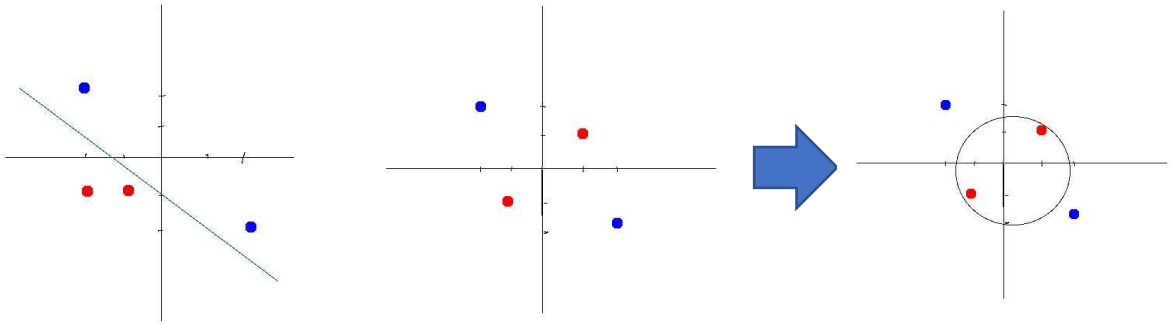


第三天，小明去医院准备输液
输液前，他身上带着前两天的药效，第一天的此时已衰减为 10· 40%=4（ m(t-2)·eff(2)），第二天的此时衰减为 10· 80%=8（ m(t-1)·eff(1)）。
输液后，他身上携带的药效为：4 + 8 + 10 = 22
rest(t)=m(t-2)·eff(2)+m(t-1)·eff(1)+m(t)·eff(0)



小明第 t 天身上残留的药效 $\mathrm{rest}(t) = \sum_{i=1}^{n} \mathrm{m}(t-i)\,\mathrm{eff}(i)$ ，
其中 n为药效有效的最大天数。

何谓激活？



线性可分　　线性不可分

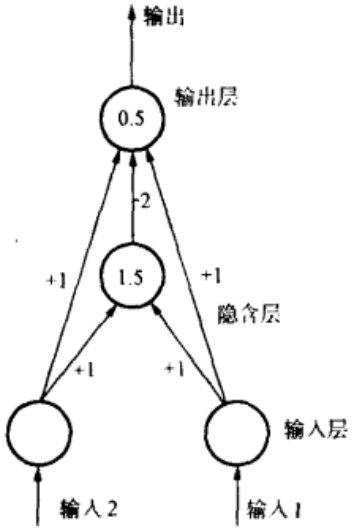
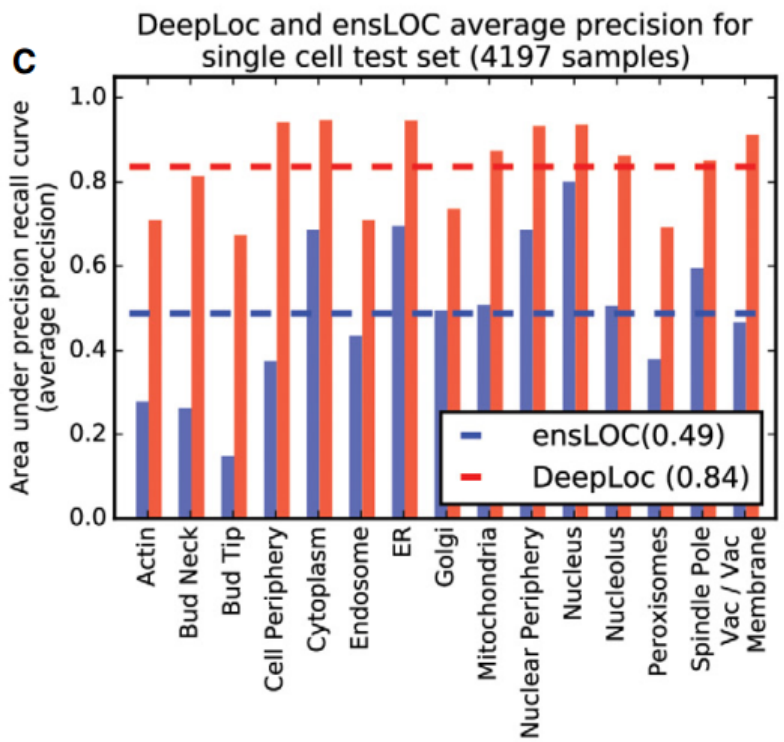


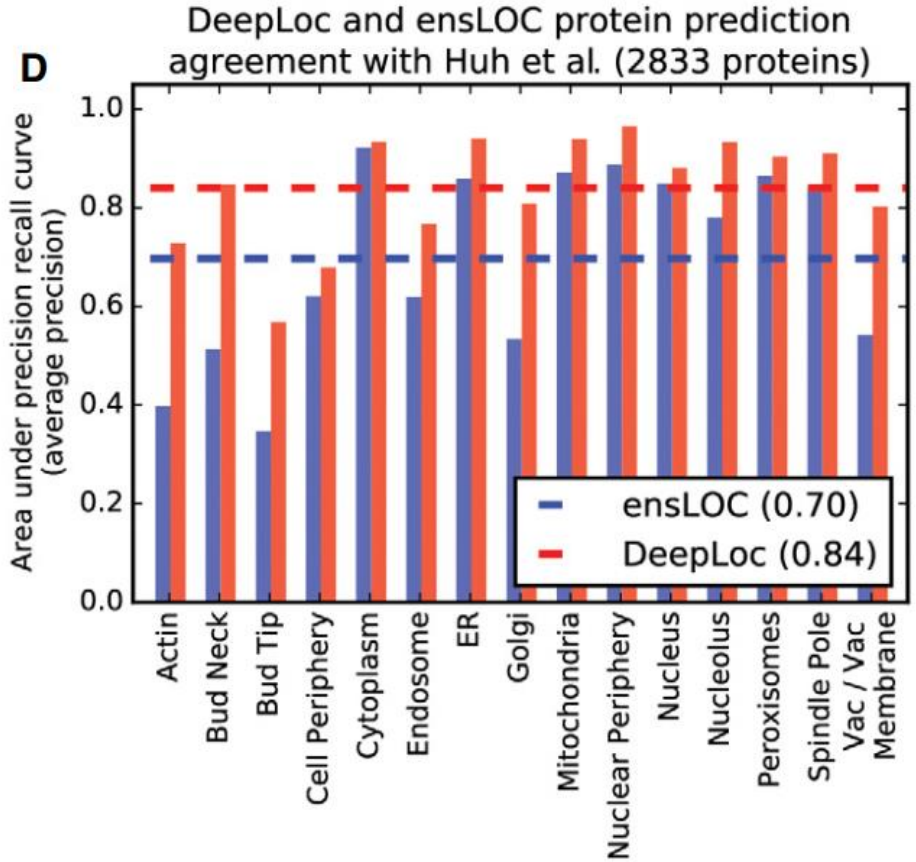
图 3.3　用于解决异或问题的多层人工神经网络

另外一种方法是引入非线性函数
我们来看异或问题(xor problem)

输入 1	输入 2	隐含层	输出单元
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

我们可以设计一种神经网络，通过激活函数来使得这组数据线性可分。
激活函数我们选择阀值函数（threshold function），也就是大于某个值输出1（被激活了），小于等于则输出0（没有激活）。这个函数是非线性函数。

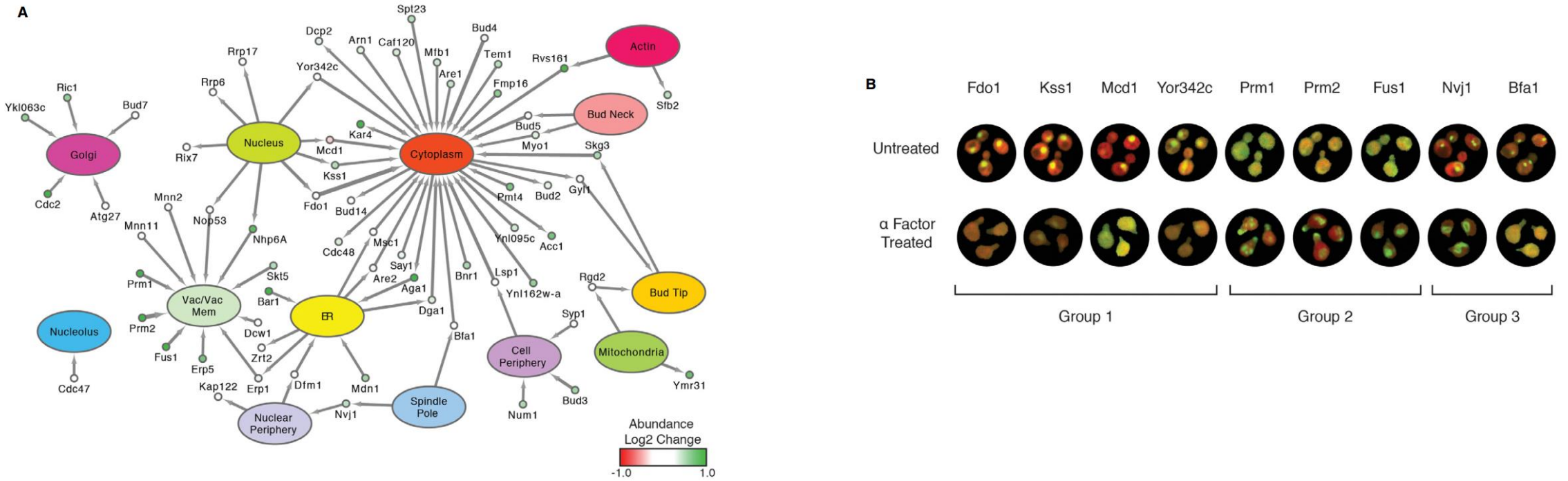

C
DeepLoc and ensLOC average precision for single cell test set (4197 samples)
Area under precision recall curve (average precision)
1.0
0.8
0.6
0.4
0.2
0.0
ensLOC(0.49)
DeepLoc (0.84)
Actin
Bud Neck
Bud Tip
Cell Periphery
Cytoplasm
Endosome
ER
Golgi
Mitochondria
Nuclear Periphery
Nucleus
Nucleolus
Peroxisomes
Spindle Pole
Vac / Vac Membrane


D
DeepLoc and ensLOC protein prediction agreement with Huh et al. (2833 proteins)
Area under precision recall curve (average precision)
1.0
0.8
0.6
0.4
0.2
0.0
ensLOC (0.70)
DeepLoc (0.84)
Actin
Bud Neck
Bud Tip
Cell Periphery
Cytoplasm
Endosome
ER
Golgi
Mitochondria
Nuclear Periphery
Nucleus
Nucleolus
Peroxisomes
Spindle Pole
Vac / Vac Membrane

Using DeepLoc to identify protein dynamics in response to mating pheromone

- 信息素（英语：pheromone，音译作费洛蒙），也称做外激素，指的是由一个个体分泌到体外，被同物种的其他个体通过嗅觉器官（如副嗅球、犁鼻器）察觉，使后者表现出某种行为，情绪，心理或生理机制改变的物质。它具有通讯功能。几乎所有的动物都证明有信息素的存在。1959年发表雌蚕蛾会分泌性信息素，是科学界首次证明了性信息素是存在的[1]。
- 将MATa 细胞暴露在mating pheromone下，细胞将停留在G1期而且极性生长
- 使用暴露在mating pheromone中40、80、120分钟的MATa细胞的GFP-ORF图片进行分析
- DeepLoc能够在数小时内对单个细胞的不同的蛋白进行合理的分类，而且不需要额外的训练
- SVM需要数周时间进行训练而且需要进行参数优化
- 鉴定到了由于α-因子扰动造成位置变化的270个蛋白，其中100个蛋白参与到了结合和有性生殖中。还参与到了细胞融合、交配时的核融合、以及极性生长。
- 这表明DeepLoc有能力去鉴别那些早已证明参与到mating response过程中的蛋白，可用来进行实验验证，找出那些富有生物学意义的结果。

- 提取像素强度作为蛋白丰富度的度量标准
- 发现82个蛋白的丰富度改变了2倍以上，75个蛋白的丰富度增加，7个蛋白的丰富度减少。
- 比较了蛋白丰富度和基因表达量改变直接的关系，二者直接是正相关的
- 和前者的工作相互补、相吻合

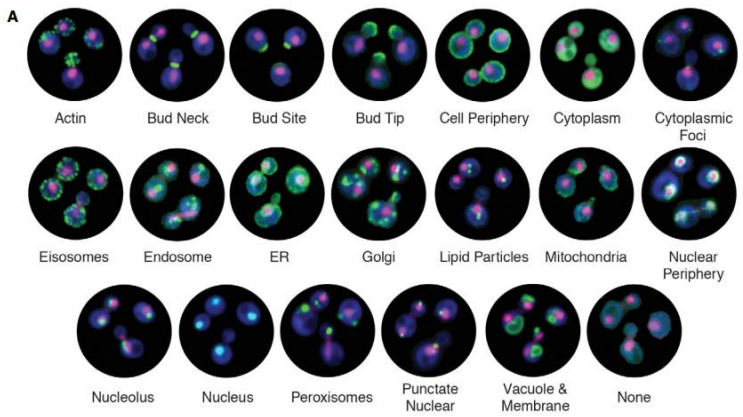


Protein dynamics in response to mating pheromone

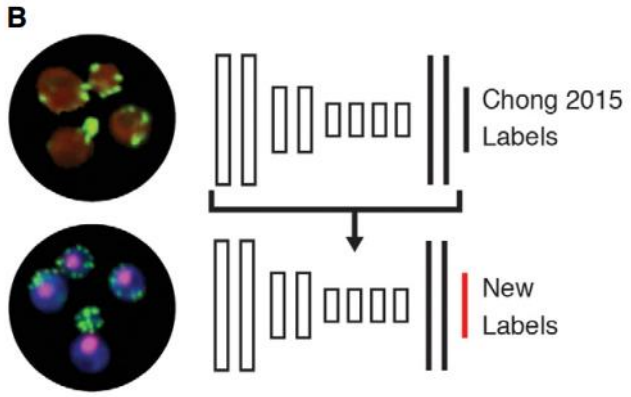
Assessing the transferability of DeepLoc to new and different microscopy datasets

- we used transfer learning to classify image sets that significantly diverge from the images used to train DeepLoc.
- Transfer learning 顾名思义就是就是把已学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。

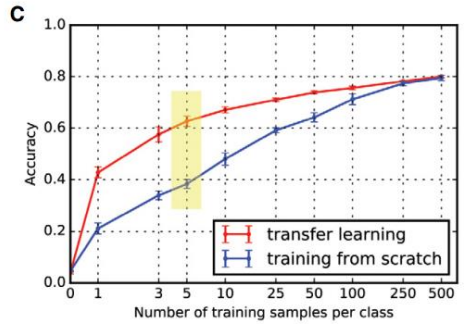
- A new HTP confocal microscope and strains contained different red fluorescent markers

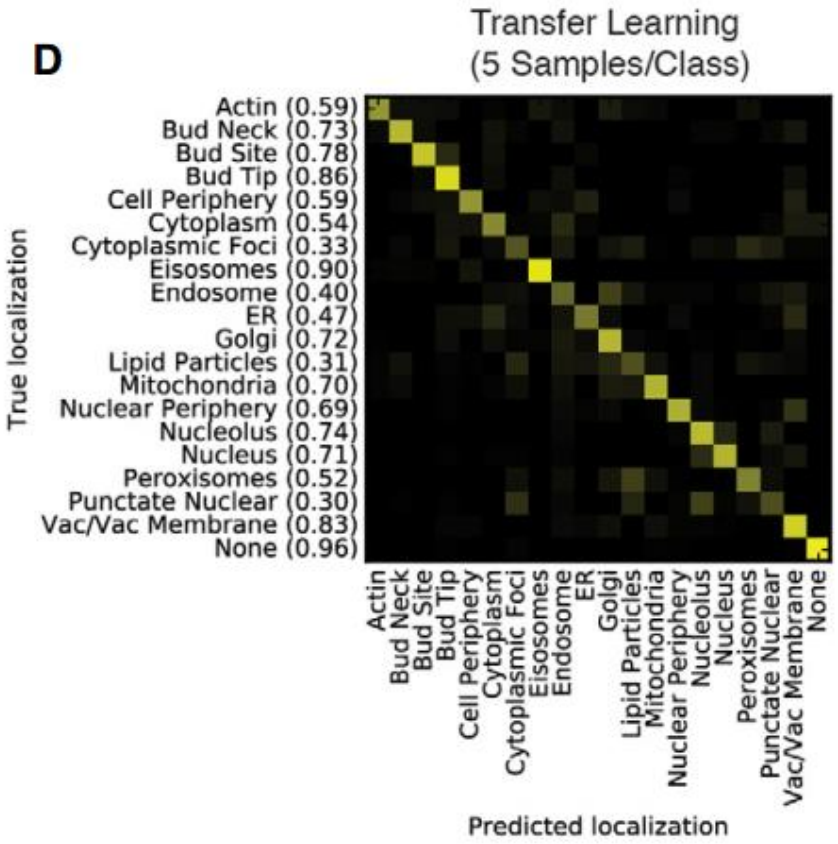


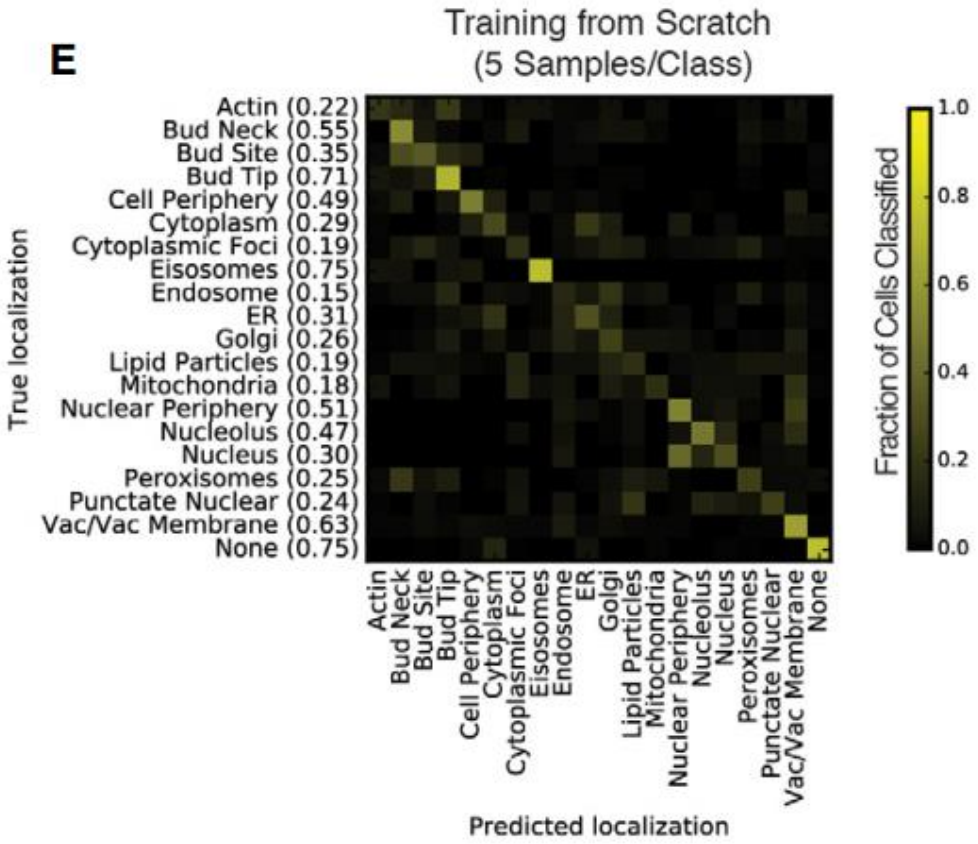
Cytoplasmic foci, eisosomes, and lipid particles



contrasted the performance of this network with one trained from scratch using the same amount of training input



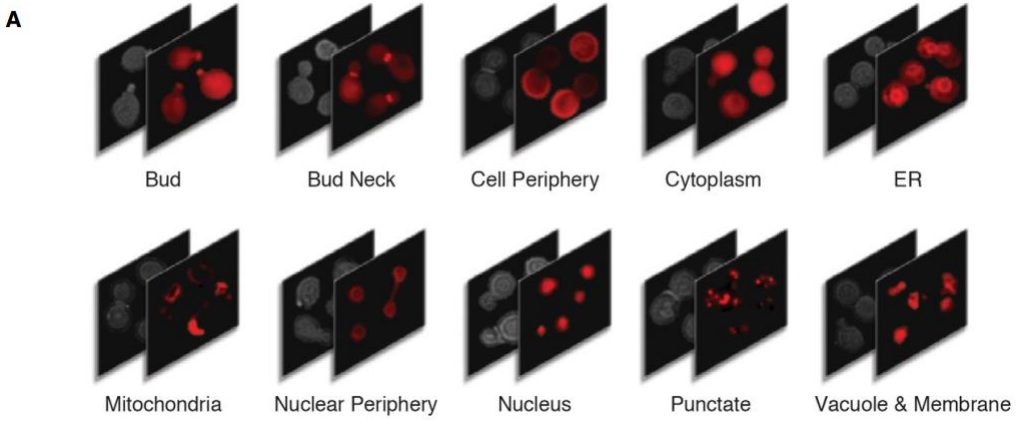


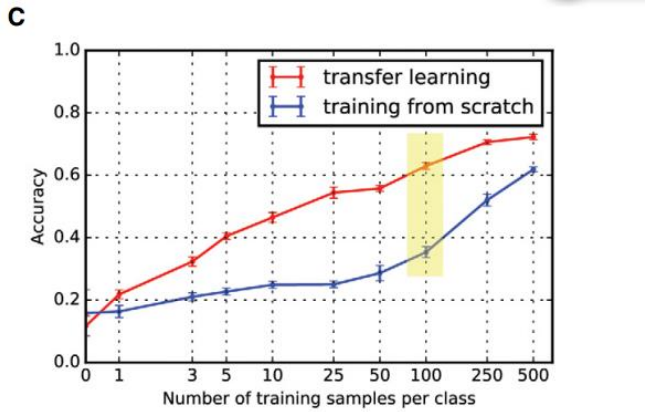


with several localization categories achieving accuracies above 80%

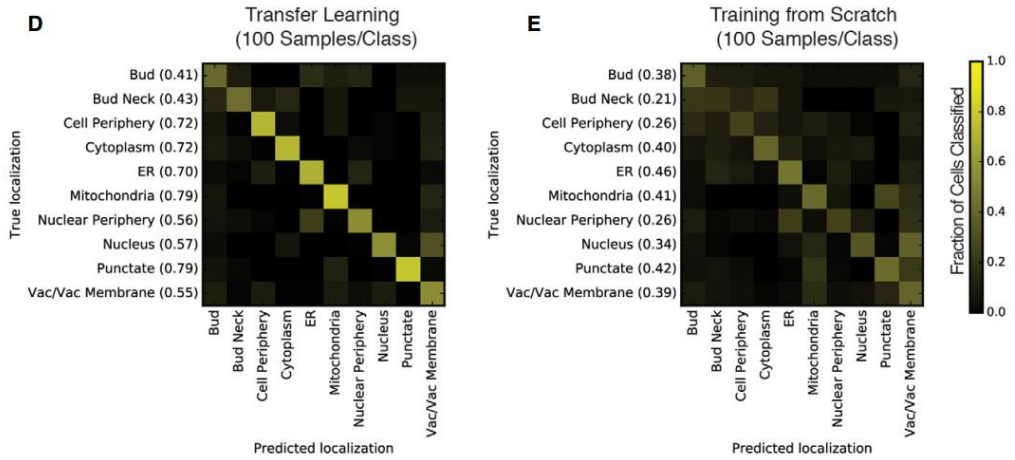
this is a 63.4% improvement in performance using transfer learning over training from scratch

- Next, we used our transfer learning protocol to classify images generated by the Schuldiner laboratory using a different microscope and fluorescent markers .





- to classify protein localizations in this dataset with an average accuracy of 63.0% after training with only 100 samples per class



ranged from 79% for the mitochondrial and “punctate” compartments to 41% for the bud compartment

- Despite these classification errors, the performance of DeepLoc is a significant achievement given that these images have previously only been classified by manual inspection, and that the imaging protocols were highly divergent from those that are optimized for automated analysis.

- （1）用一句话概括此研究的主要结论或创新点。

使用深度卷积神经网络(deep convolutional neural network)对酵母细胞图像进行分析，能够在蛋白的亚细胞定位上的自动分类上表现的更加准确和可靠，而且可以适用于不同条件(不同的高分辨率显微镜、不同的mraker)的图像。

（2）此研究对你有何启发。

很有意思！机器学习、深度学习、迁移学习，随着时间的发展，最后将是什么样子呢？

尝试着找一些具有生物学意义的分类任务/识别任务，模仿着开展研究

（3）此研究还存在哪些问题可以改进。

准确率还可以继续提高？