

# 视觉智能：进展与问题

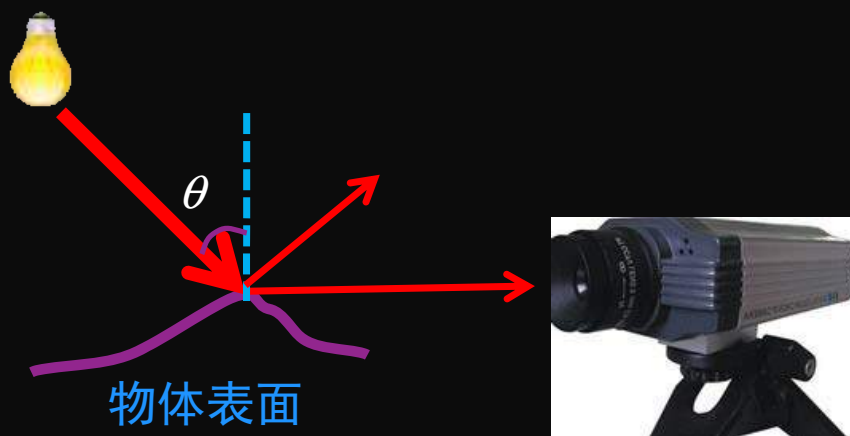
山世光 (sgshan@ict.ac.cn)

中国科学院计算技术研究所 研究员  
中科院智能信息处理重点实验室 常务副主任  
中科视拓（北京）科技有限公司 董事长/CTO

# 计算机视觉是什么？

## □ 相机或摄像机输出的是什么？

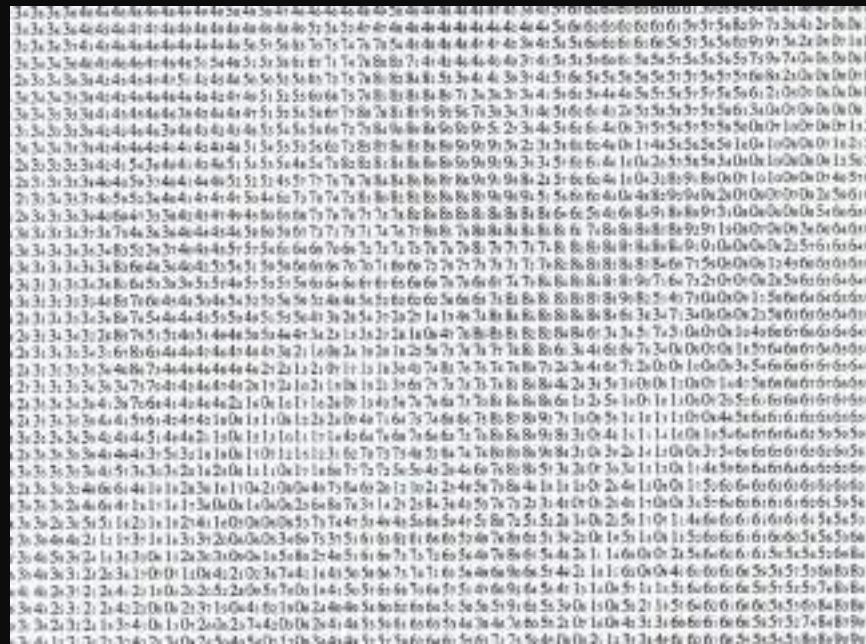
- 图像/视频：空间中某物体表面点反射或发射的不同波长的光强
- 物体表面不同材质对不同波长的光有不同的反射率（/吸收率）



R (红)	G (绿)	B 蓝
255	0	0
0	255	0
0	0	255
0	0	0
255	255	255
100	100	100
255	128	64

# 计算机视觉是什么？

解读w\*h\*3个0~255之间的数字中蕴藏的、人类可理解的内容  
(边界, 区域, 物体, 事件, 意义)



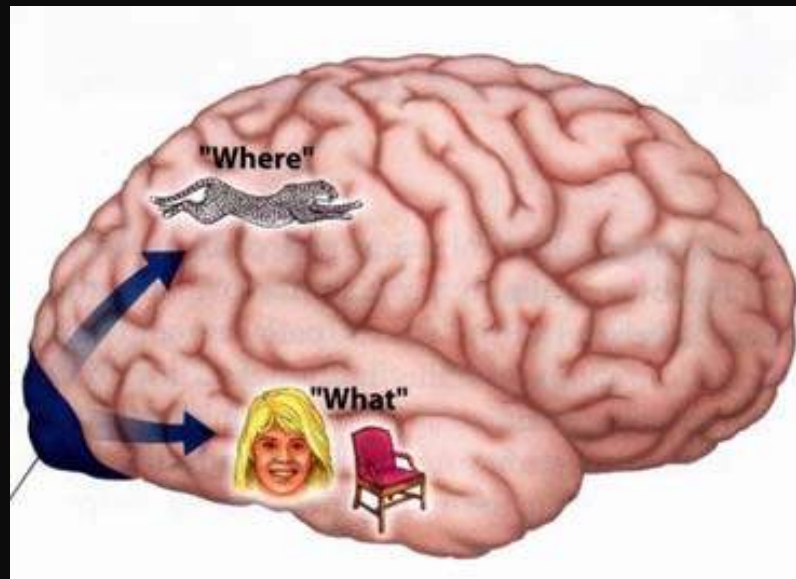
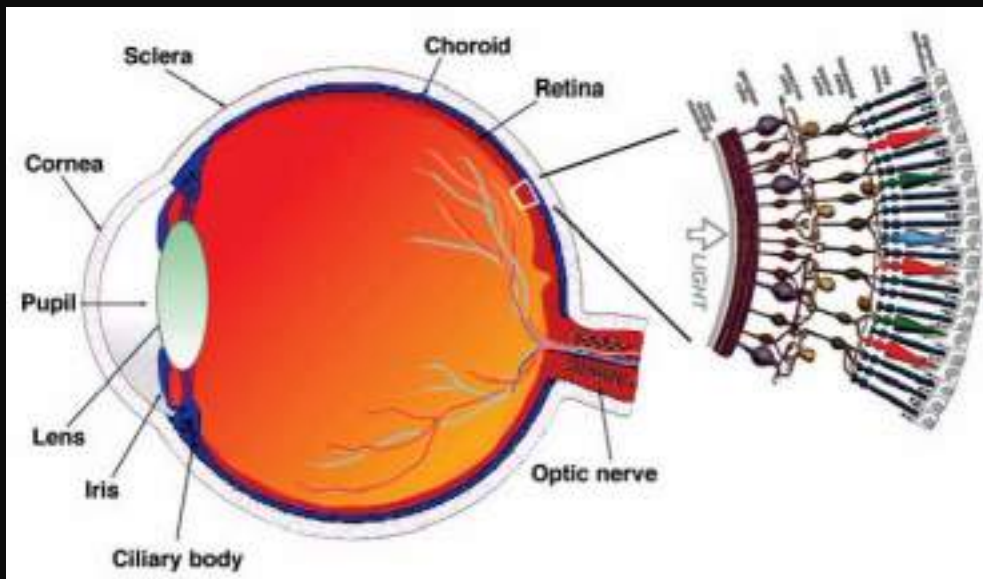
62	62	63	64	65	66	67	67	69	70	71	72	72	73	73	73	73	72	72	71	70	69	67	66	66	66	65	63	62	61	60	60	
61	62	63	64	66	66	67	68	68	69	70	71	71	72	72	73	73	72	72	71	71	70	69	68	66	66	65	65	63	62	61	60	60
61	62	63	64	66	66	68	68	69	70	71	72	73	73	73	72	72	71	71	69	68	67	66	66	65	65	64	63	62	61	61	61	
61	63	64	65	66	67	68	68	69	70	71	71	73	73	74	73	73	73	71	70	69	68	66	66	65	64	63	62	61	61	60	60	
61	63	64	65	67	68	69	70	70	71	71	72	55	53	69	72	72	71	71	70	69	68	67	66	65	64	63	62	60	60	60	60	
63	64	65	66	67	68	69	69	70	70	71	72	42	4	5	11	48	72	71	71	69	69	68	67	66	65	64	62	62	60	59	59	
63	65	66	66	68	68	69	70	71	71	72	18	4	4	7	8	66	71	70	69	68	68	67	66	65	64	63	61	59	59	58	58	
63	65	67	67	68	69	69	70	71	71	72	64	4	27	24	54	33	29	52	64	68	68	67	66	65	64	63	62	61	59	58	58	
64	65	66	66	68	69	70	71	71	71	24	24	12	17	24	45	60	37	43	35	52	66	68	67	66	65	64	63	61	60	59	58	57
65	66	67	67	68	69	71	49	6	6	6	5	34	38	12	47	34	17	29	54	43	63	67	66	65	64	63	62	60	59	58	57	
64	65	66	66	68	69	38	6	6	5	7	16	19	4	47	44	27	24	40	67	66	66	65	65	64	63	61	60	59	58	57		
63	64	65	65	67	68	6	6	5	5	6	8	9	20	27	51	78	41	44	66	65	65	65	65	64	63	62	60	59	58	57		
63	64	65	65	64	5	5	5	5	5	5	4	19	6	7	54	64	20	59	65	65	64	64	64	63	62	61	60	59	57	56	56	
63	64	64	65	14	5	6	5	5	4	5	4	18	7	5	4	19	10	11	05	64	64	64	63	61	66	62	61	60	59	58	56	
63	64	64	65	53	7	4	5	6	6	7	10	6	5	5	4	21	24	18	64	64	64	63	62	64	65	62	62	60	59	58	57	
64	64	64	65	50	4	4	4	5	11	6	6	4	6	35	16	29	66	64	64	63	61	72	67	63	62	61	59	58	57	56	56	
64	64	64	65	46	4	4	4	5	6	9	8	5	29	10	43	56	29	57	64	64	63	61	70	67	62	64	65	59	59	57	57	
64	64	64	65	66	27	5	4	4	5	6	6	6	18	66	20	57	60	46	38	75	70	62	61	70	67	62	61	60	59	58	58	
49	50	62	65	57	5	5	6	6	6	6	43	59	28	60	58	44	22	63	71	72	60	69	68	61	60	58	59	59	58	58		
42	52	57	52	26	5	5	5	5	5	5	70	50	43	61	62	64	39	42	64	60	62	56	63	65	65	67	61	53	53	53		
42	32	32	33	6	5	5	5	5	6	6	11	39	21	33	51	50	45	46	18	32	36	33	23	44	70	71	51	42	27	31		
50	50	51	30	5	5	5	6	6	6	6	42	69	28	34	42	30	43	37	26	29	40	26	29	26	35	42	35	33	18	19		
52	53	51	22	5	5	5	6	6	6	5	44	56	17	51	54	53	54	56	51	22	54	54	55	55	54	53	53	53	52	52		
54	54	53	8	5	5	5	6	5	6	13	52	42	21	51	54	51	49	49	50	22	41	45	42	41	40	41	44	43	42	42		
52	52	54	36	8	5	6	6	6	28	55	32	32	54	53	51	51	51	51	44	25	51	51	49	49	48	48	46	46	46	46		
54	54	52	53	9	7	5	6	6	6	40	54	29	52	51	53	56	55	52	51	38	52	52	50	49	46	46	45	45	46	47		
51	52	51	53	27	14	5	4	5	4	7	47	51	21	39	49	47	49	52	52	49	35	31	48	46	47	47	46	46	46	43		
48	50	51	53	25	14	7	8	4	4	17	46	40	18	43	47	46	49	52	54	53	53	54	18	50	49	46	47	47	47	45	45	
49	49	49	49	22	12	20	24	6	14	35	51	39	48	48	50	51	51	49	51	52	50	41	58	48	47	47	47	45	45	46		
51	49	50	50	22	13	19	38	13	12	42	50	46	73	50	50	49	48	49	48	49	48	49	45	46	44	44	44	42	45	47		
47	49	49	47	20	16	26	30	21	15	36	48	42	61	47	48	51	47	50	51	51	49	47	47	47	47	44	43	45	46	46		
48	50	48	52	19	13	33	36	18	18	36	49	51	54	47	47	49	46	46	49	49	49	49	47	44	53	44	44	46	46	45		



# 来自生物视觉系统的启示

## □我们如何看见？

- 看：眼睛 == 摄像机
- 见：视皮层的what通路和where通路

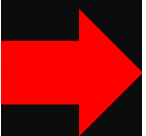
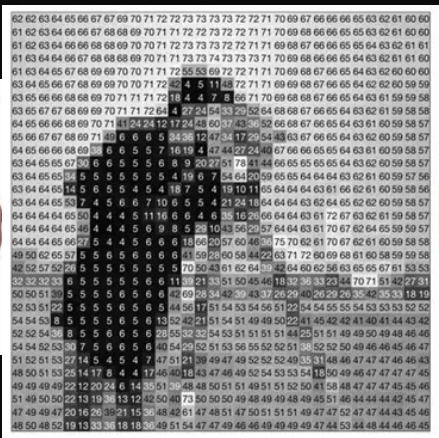
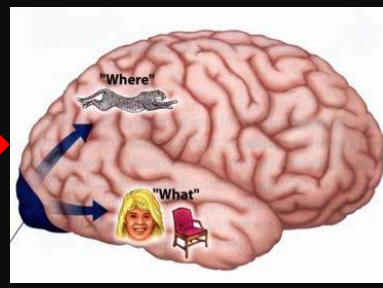
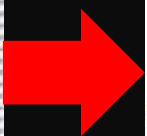




# 来自生物视觉系统的启示

## □视皮层：从宏观到微观

- 人类大脑共有860亿量级神经细胞，高度互联
- 大脑→脑区→神经细胞互联→单个神经细胞→生化反应
- 视皮层涉及多个脑区，几十亿神经细胞



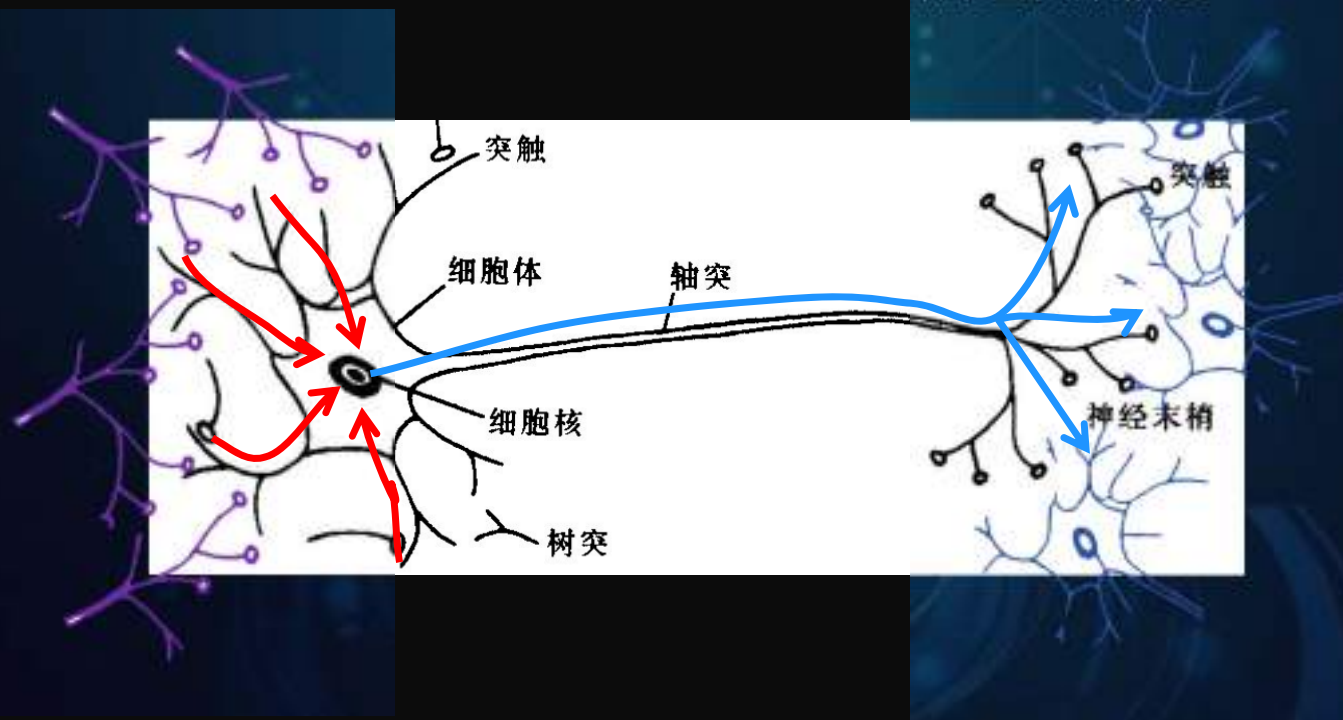
一张黑白照片，一位摄影师在用三角架上的照相机拍照，远处是...

# 来自生物视觉系统的启示

□神经细胞层面：高度互联，**每个神经元与数千神经元连接**

■感觉神经细胞的基本工作流程

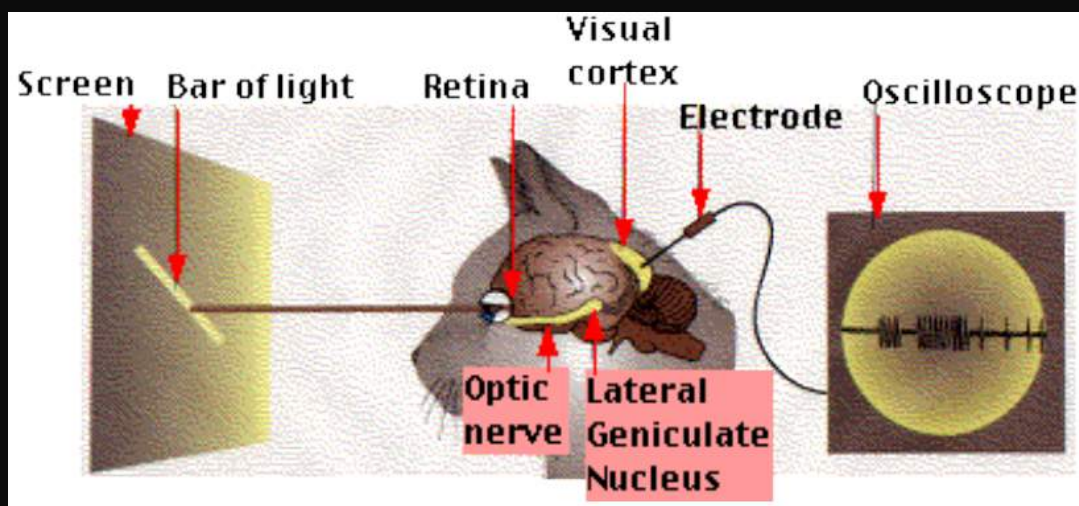
□树突收集信号，胞体汇集后做出决策，轴突向后传递决策信息



# 来自生物视觉系统的启示

## □每个神经元的功能

- 高度专业化的功能分工：对某个特定决策投赞成票或者弃权票
- 模式发现器：输入给自己的信息中，是否出现了自己感兴趣的模式

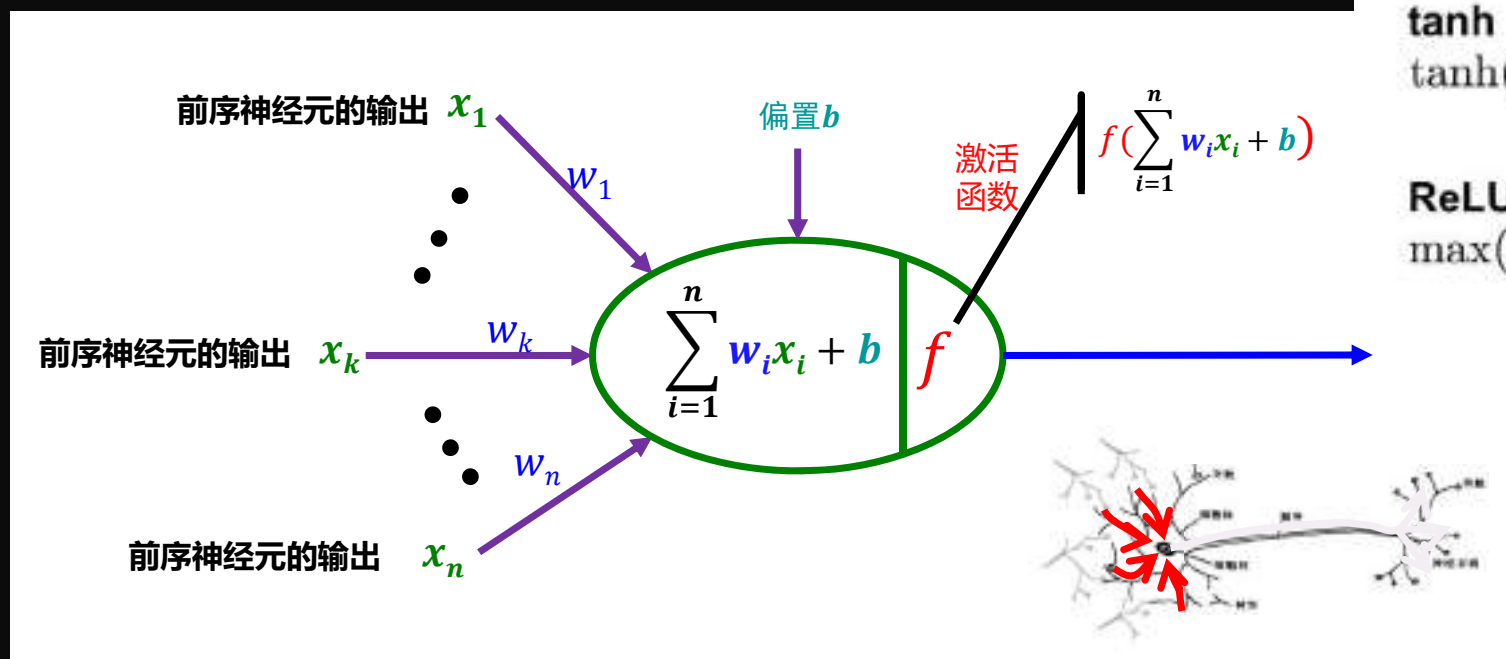


Hubel, D. H. & Wiesel, T. N. (1960s)

# 深度学习源起——单神经元计算模型

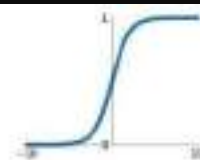
## 单神经元计算模型

- 加权求和（卷积） + 非线性激活函数



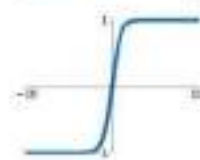
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



tanh

$$\tanh(x)$$



ReLU

$$\max(0, x)$$

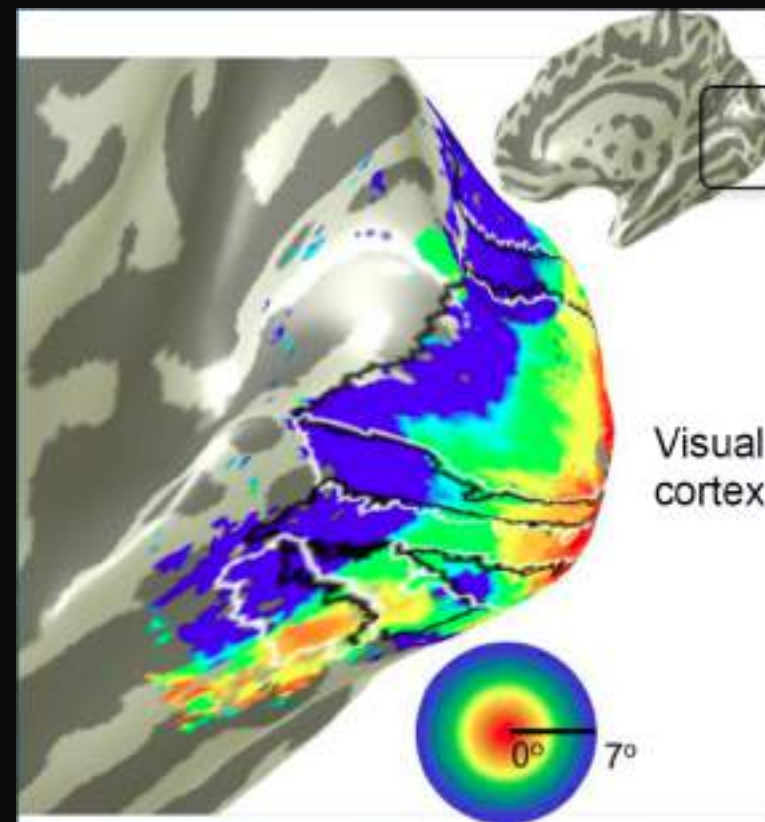
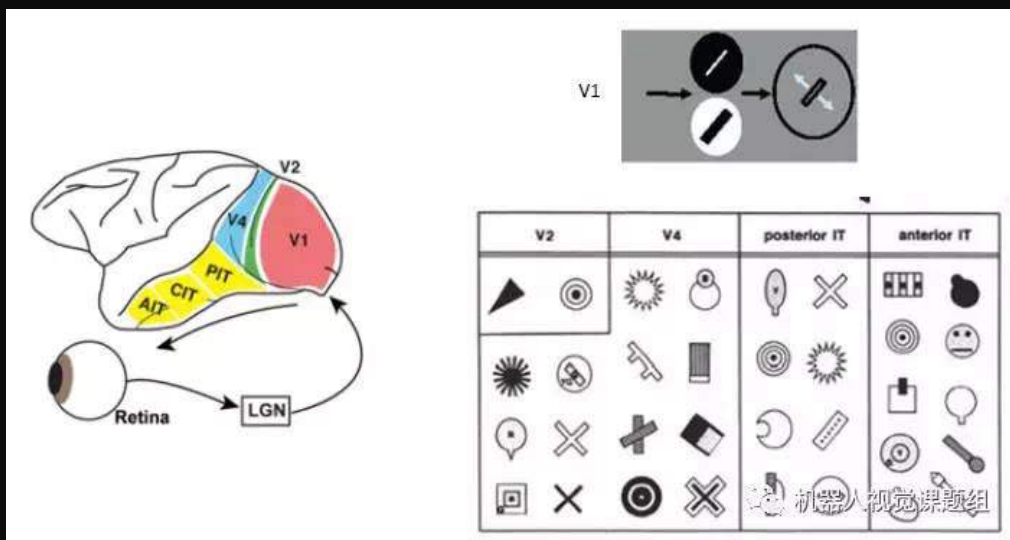




# 深度学习源起——层级感受野假设

## □ 层级感受野

- 一个神经细胞看的更远（视野更大）、能处理更复杂的任务，是因为他站在了其他神经元的肩膀上！
- 类比：社会结构

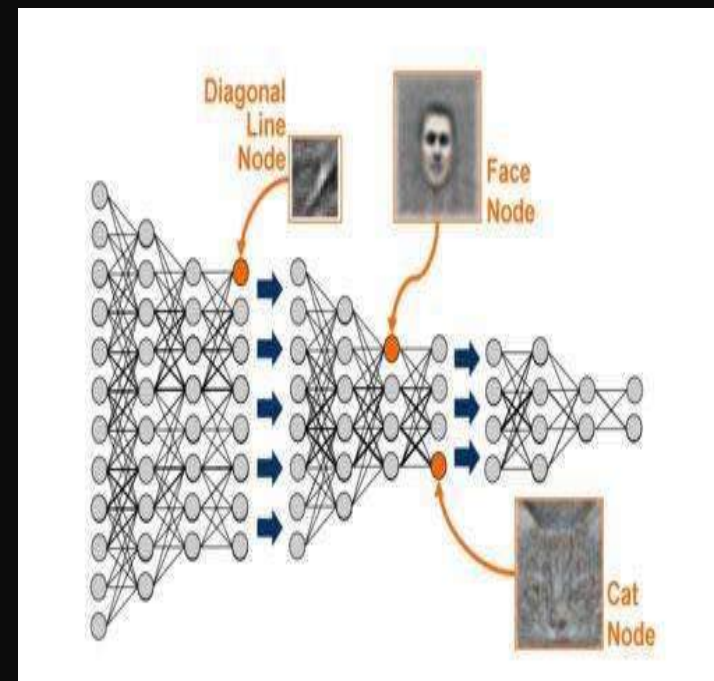
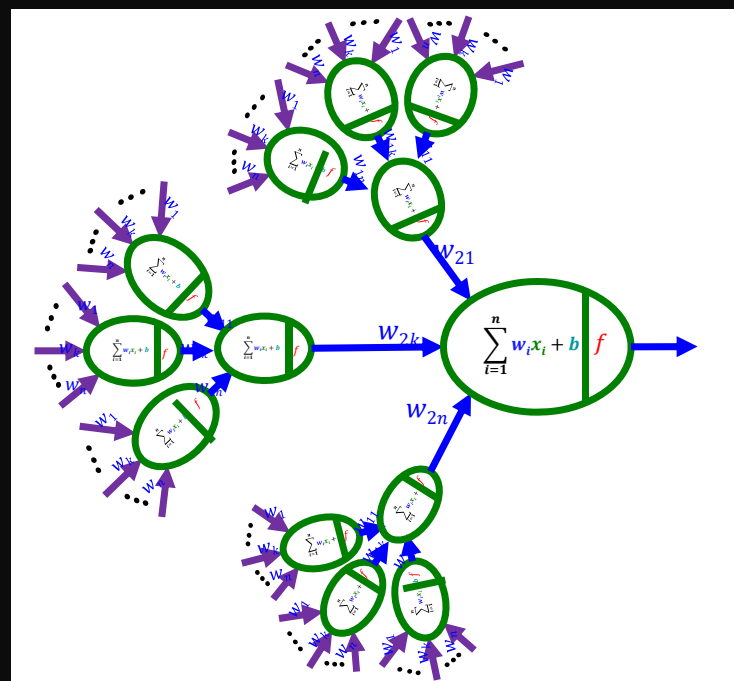
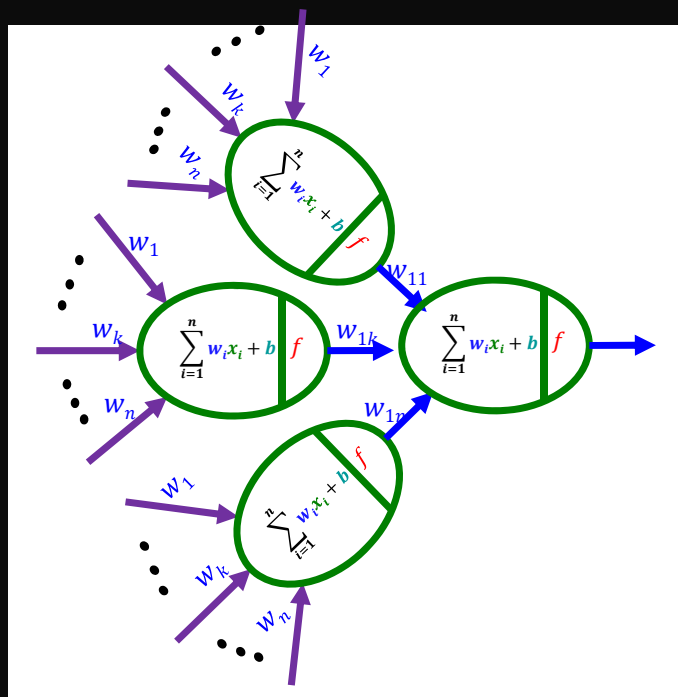


Retinotopy images courtesy of Jesse Gomez in the Stanford Vision & Perception Neuroscience Lab.

# 深度学习源起——多神经元互联模型

□ 多个单神经元互联形成多层神经网络

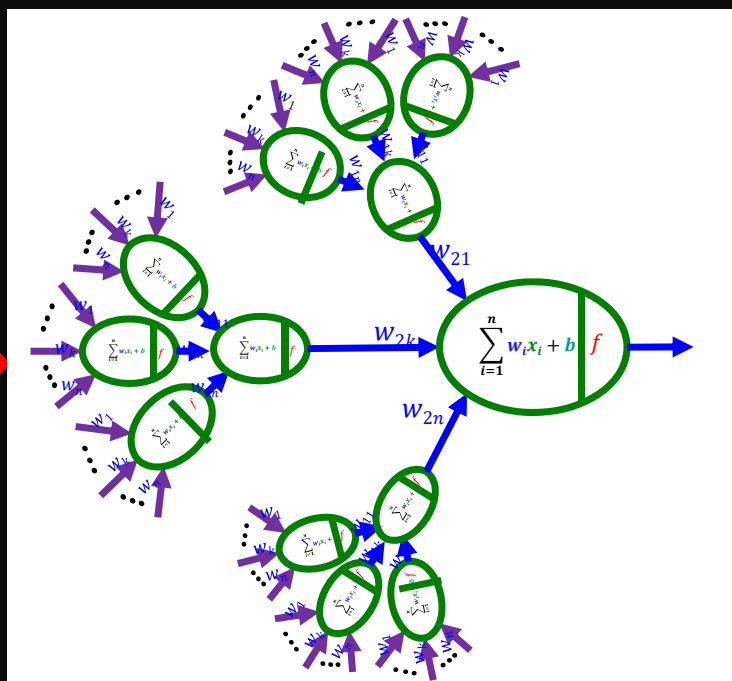
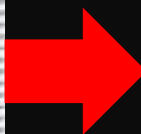
■ 从输入到输出，中间多个隐含层（**隐层级多即所谓深度学习之深**）



# 深度学习源起——多神经元互联模型

## □ 多个单神经元互联形成多层神经网络

- 从输入到输出，中间多个隐含层（**隐层级多即所谓深度学习之深**）



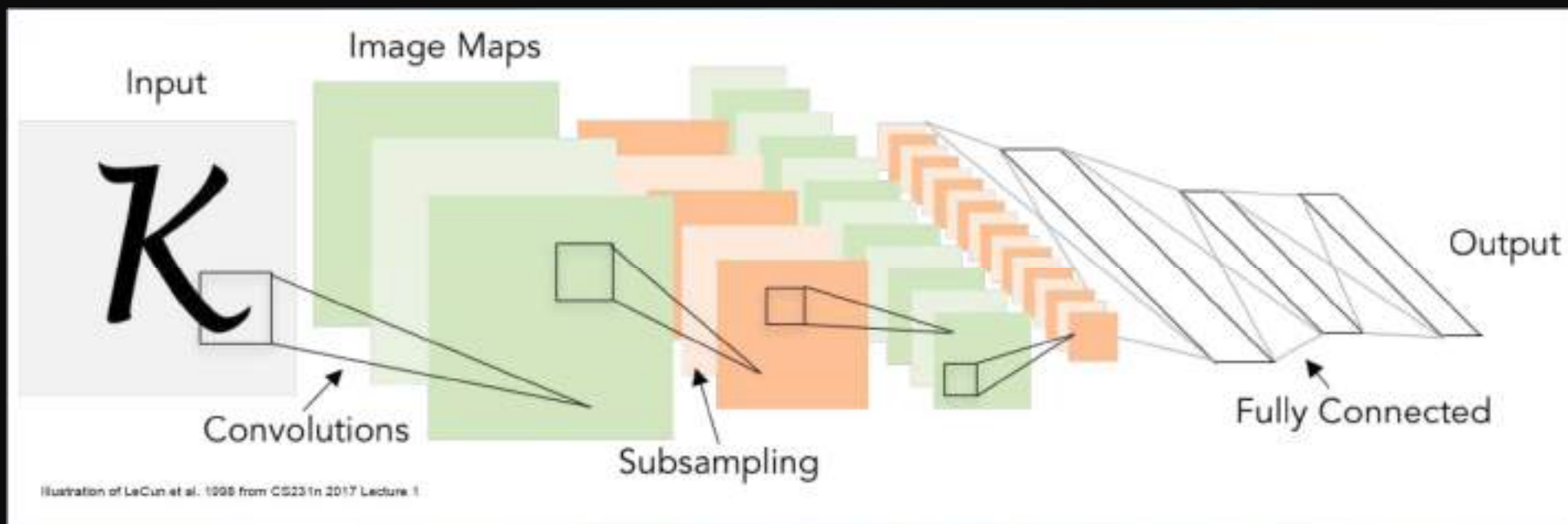
一张黑白照片，一位摄影师在用三角架上的照相机拍照，远处是...



# 卷积神经网络CNN

## □ 一种模拟层级感受野的前馈神经网络

- 若干卷积层（非线性激活后），交叉若干Pooling(下采样)层
- 若干全连接层，最后一层输出为类别标签（或目标回归值）



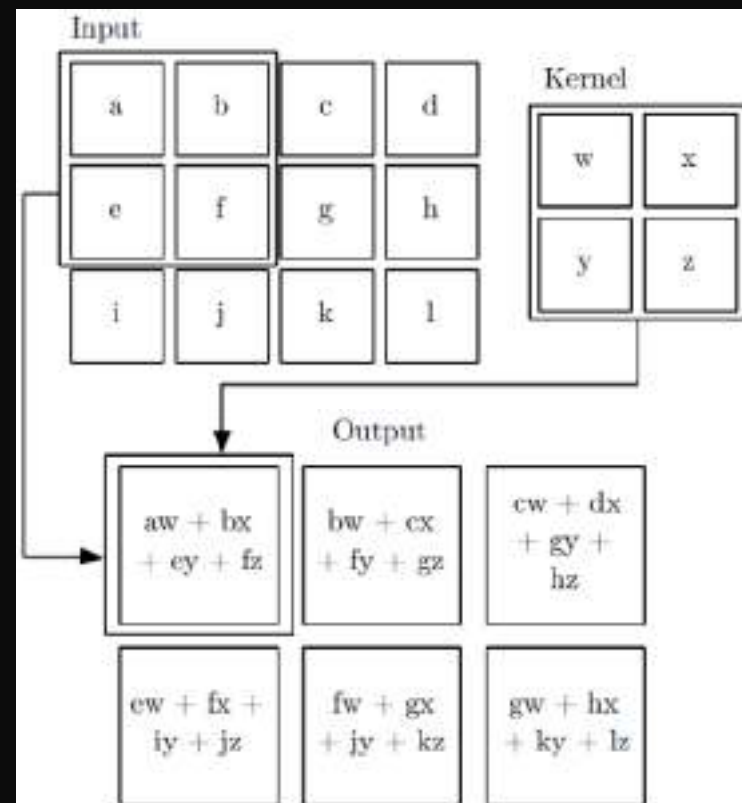
# 卷积神经网络CNN

## □卷积操作

- 图像与核(权重)的加权求和

## □作用

- 本质上是一种滤波器，或局部特征提取器

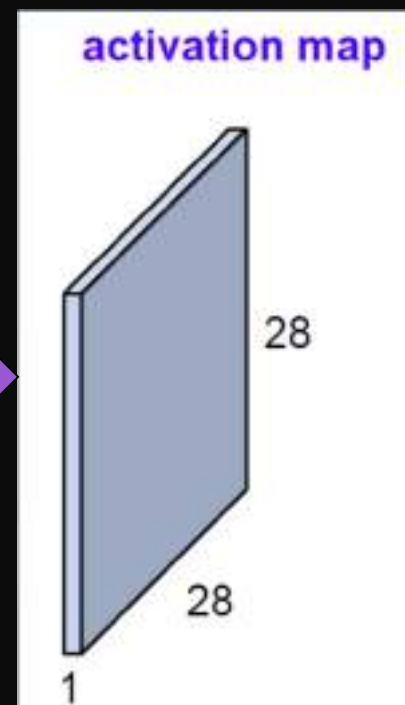
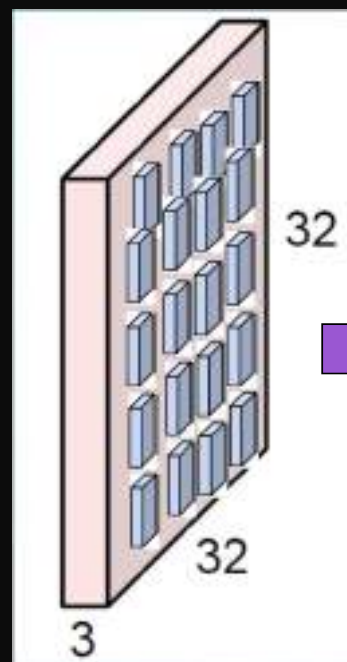
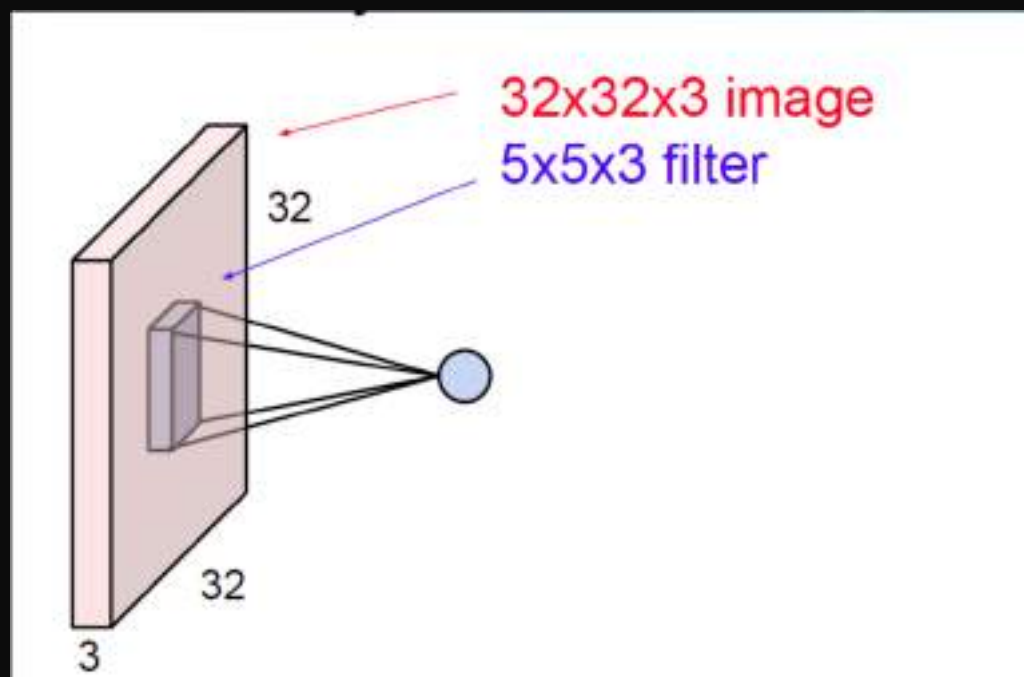




# 卷积神经网络

□卷积层：局部连接，加权求和

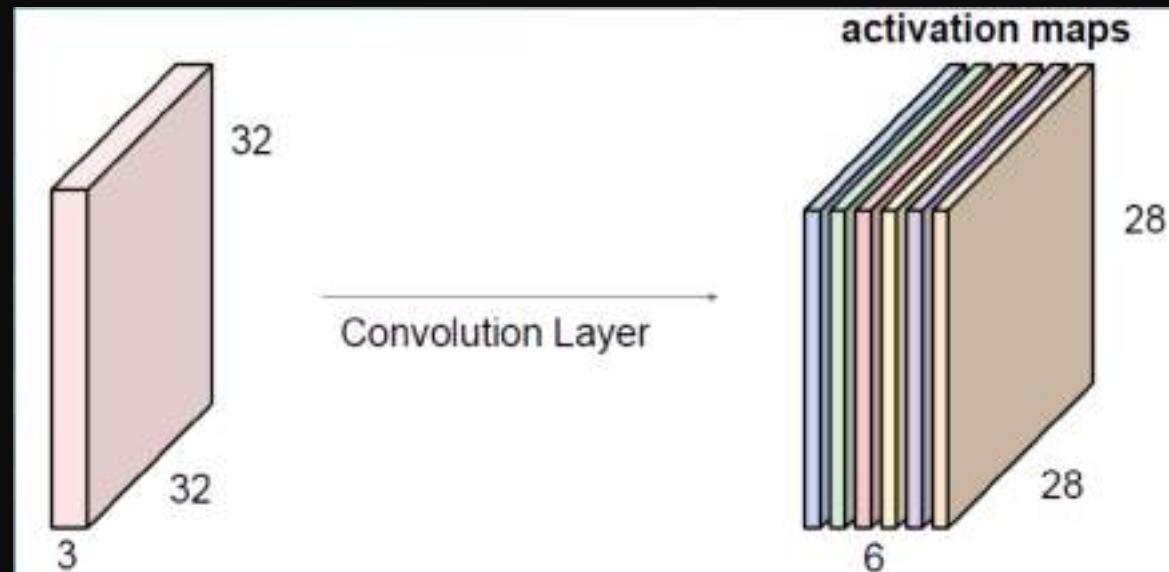
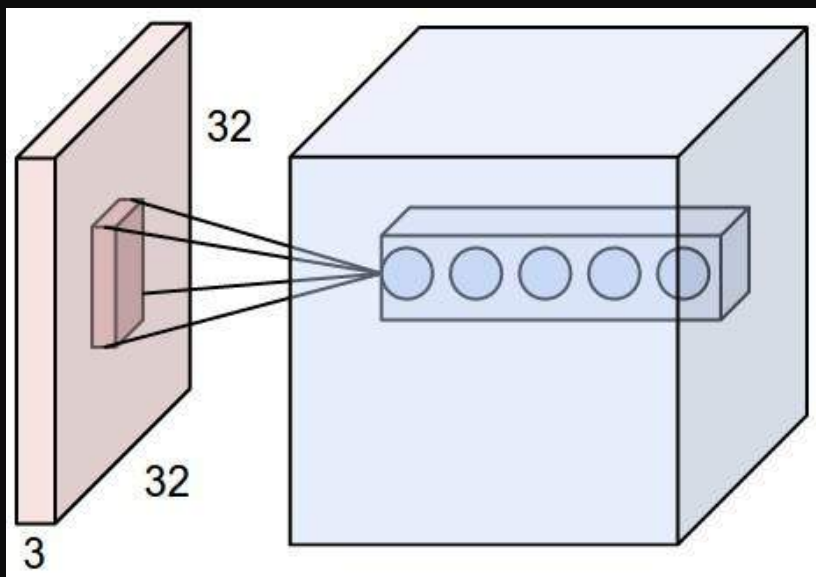
□功能：特征提取器，越来越复杂的特征，逐级抽象



# 卷积神经网络

□卷积层：局部连接，加权求和

□功能：特征提取器，越来越复杂的特征，逐级抽象

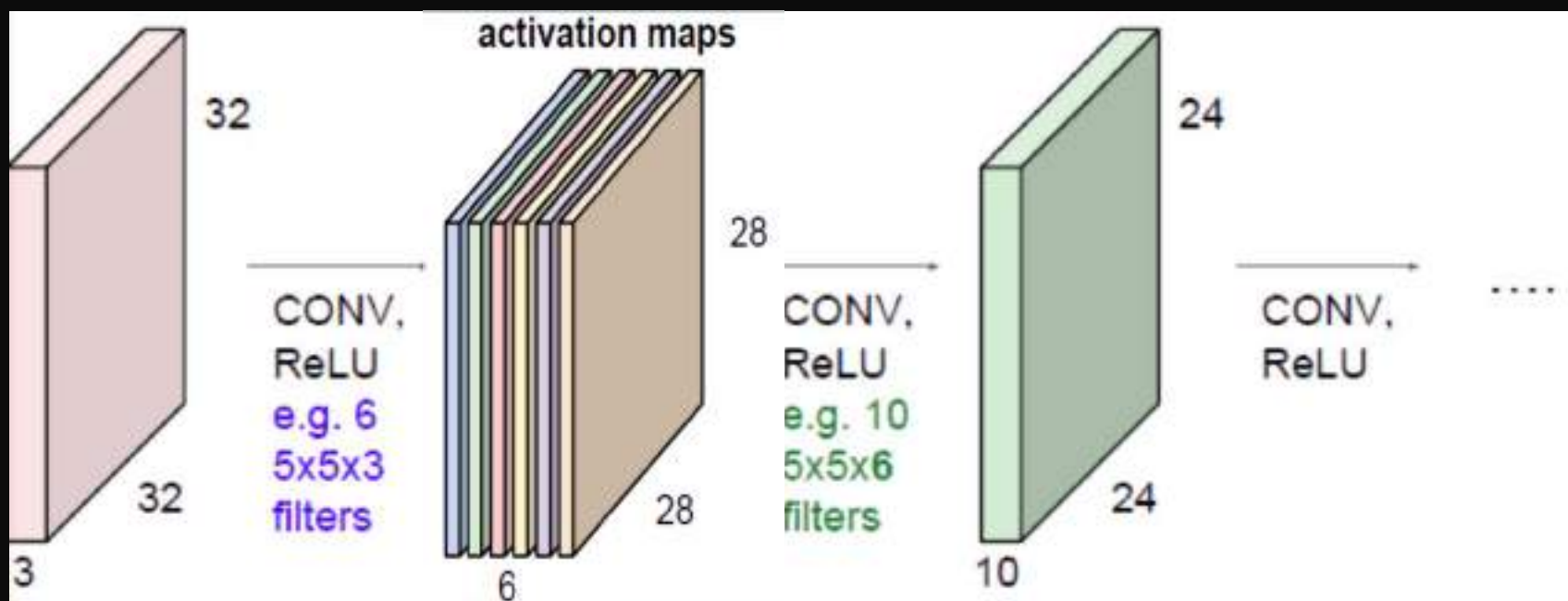


注：此页图片来自斯坦福大学李飞飞教授胶片

# 卷积神经网络CNN

□卷积层：局部连接，加权求和

□功能：特征提取器，越来越复杂的特征，逐级抽象

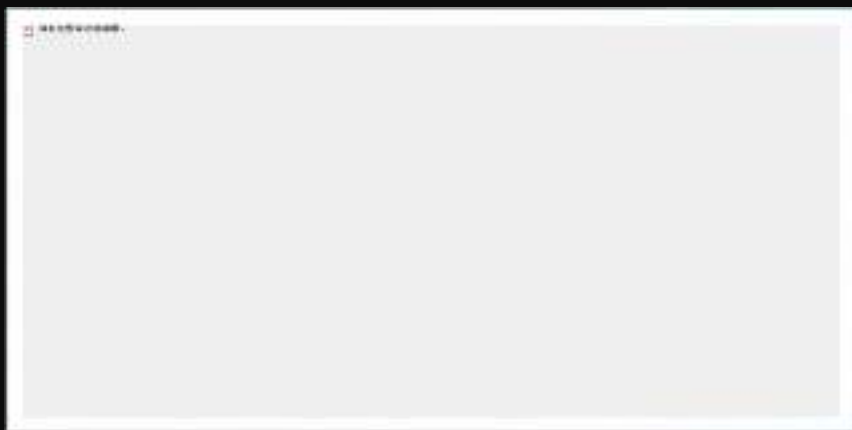


注：此页图片来自斯坦福大学李飞飞教授胶片

# 卷积神经网络CNN

## □与普通滤波器的不同

- LoG, DoG, Gabor: 权重系数人工设计
- CNN卷积: 权重系数学习而来



# 卷积神经网络CNN

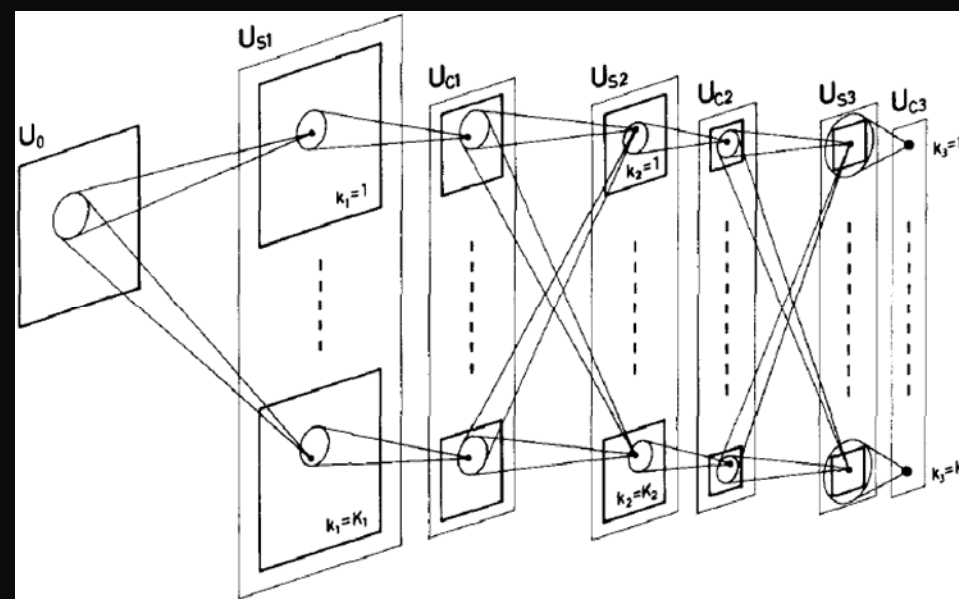
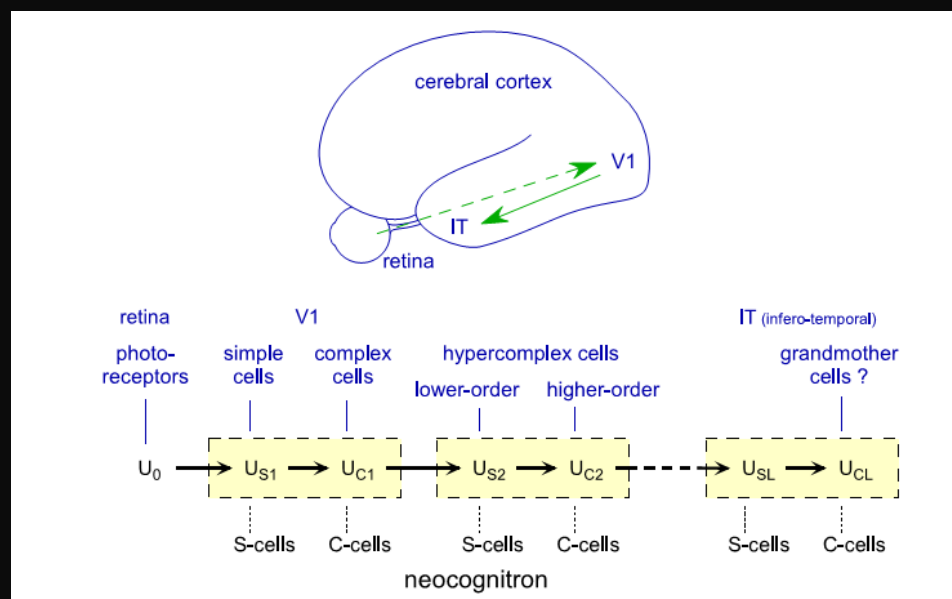
- K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, vol. 36, pp. 193–202, 1980
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989
- Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998



# 卷积神经网络CNN

## □ Neocognitron, Fukushima 1980

- 层级感受野：简单细胞 → 复杂细胞 → 超复杂细胞...
- 中间层**无监督训练**（自组织），分类层有监督训练



# 卷积神经网络CNN

## □ LeCun et al. 1989

- 简化了Neocognitron的结构

- 训练方法

- 监督训练

- BP算法

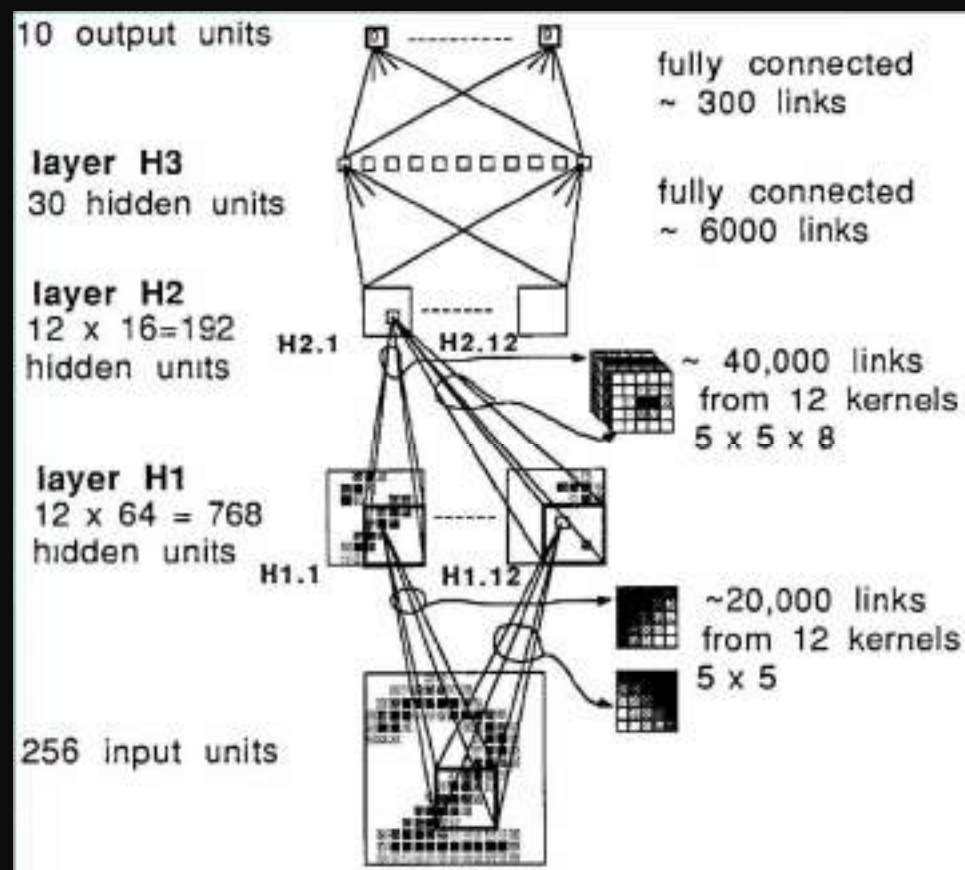
- Sigmoid Loss

- 随机梯度下降SGD

- 正切函数收敛更快

- 应用

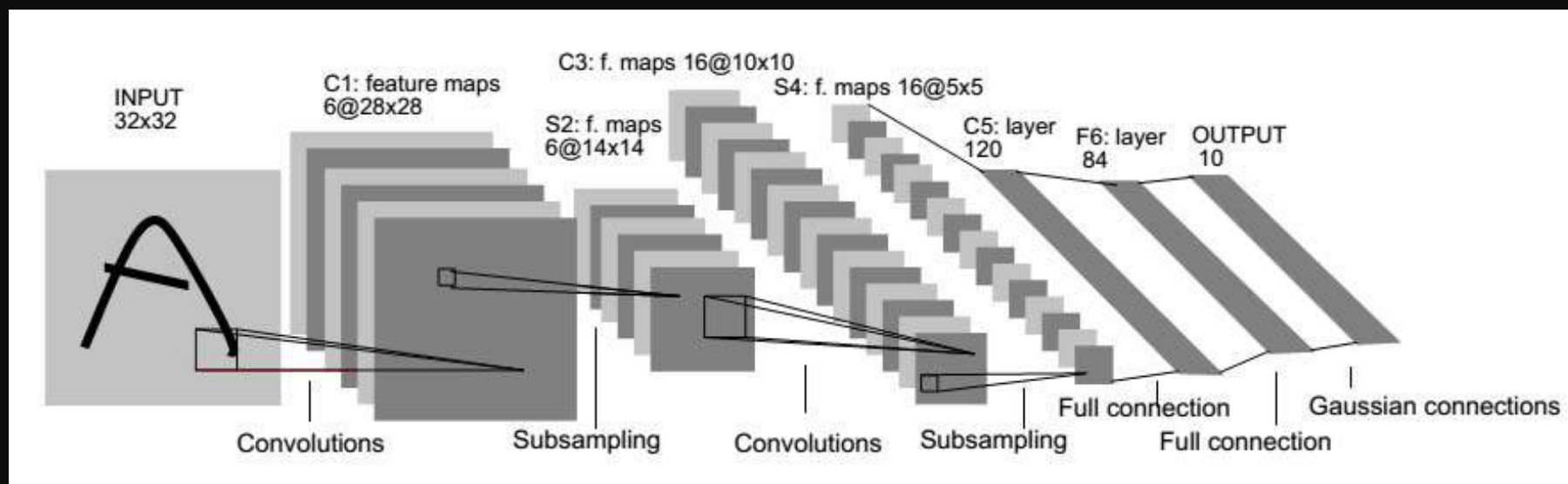
- 邮编识别



# 卷积神经网络CNN

## □ CNN: LeCun et al 1998

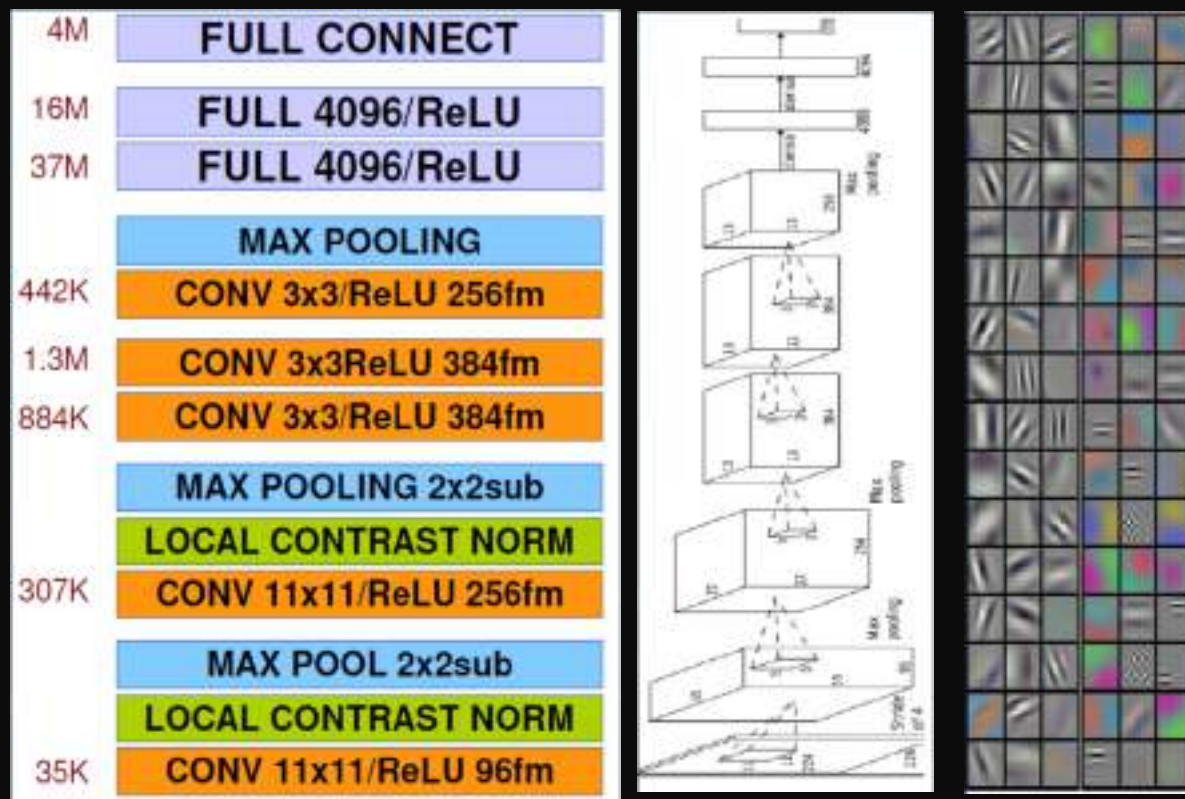
- 2个卷积层+3个全连接层
- Pooling(/subsampling)



# 卷积神经网络CNN

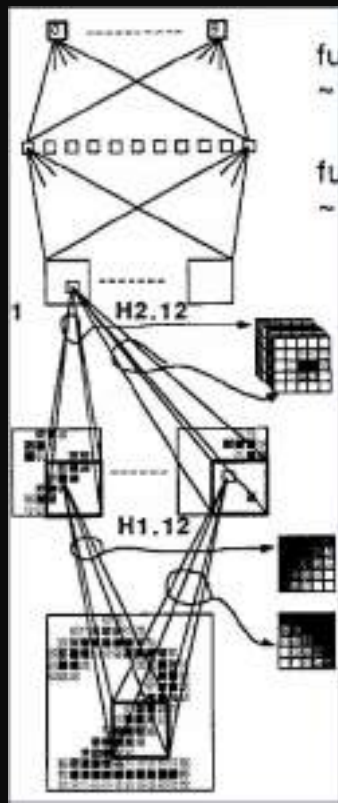
## □ AlexNet

- 650K神经元
- 6000万参数要学习
- 训练
  - BP on GPU
- 各种优化技巧
  - ReLU
  - Dropout
  - 数据增广
  - ...

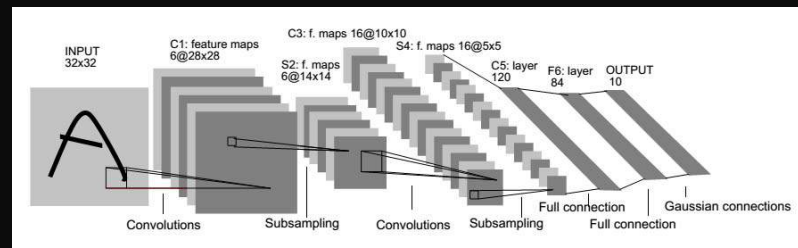


A. Krizhevsky, L. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, 2012.

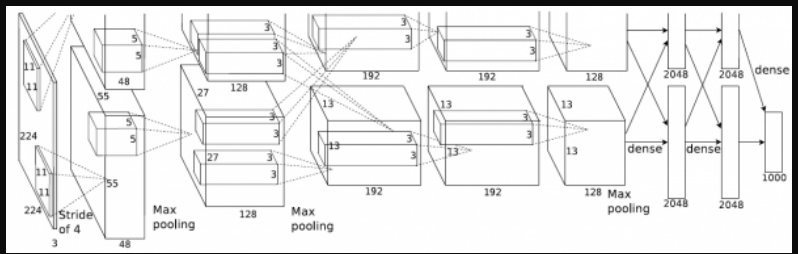
# Renaissance神经网络 (以CNN为例)



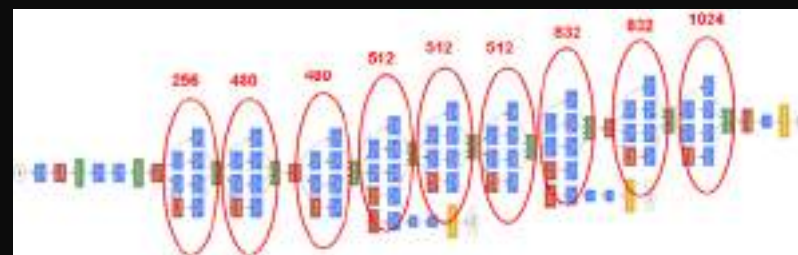
1989



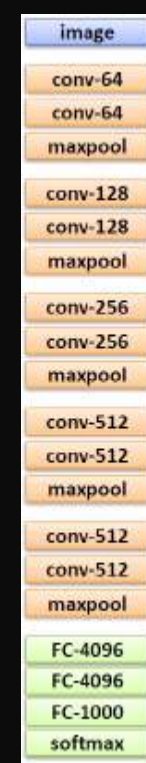
1998



2012



2015



2014

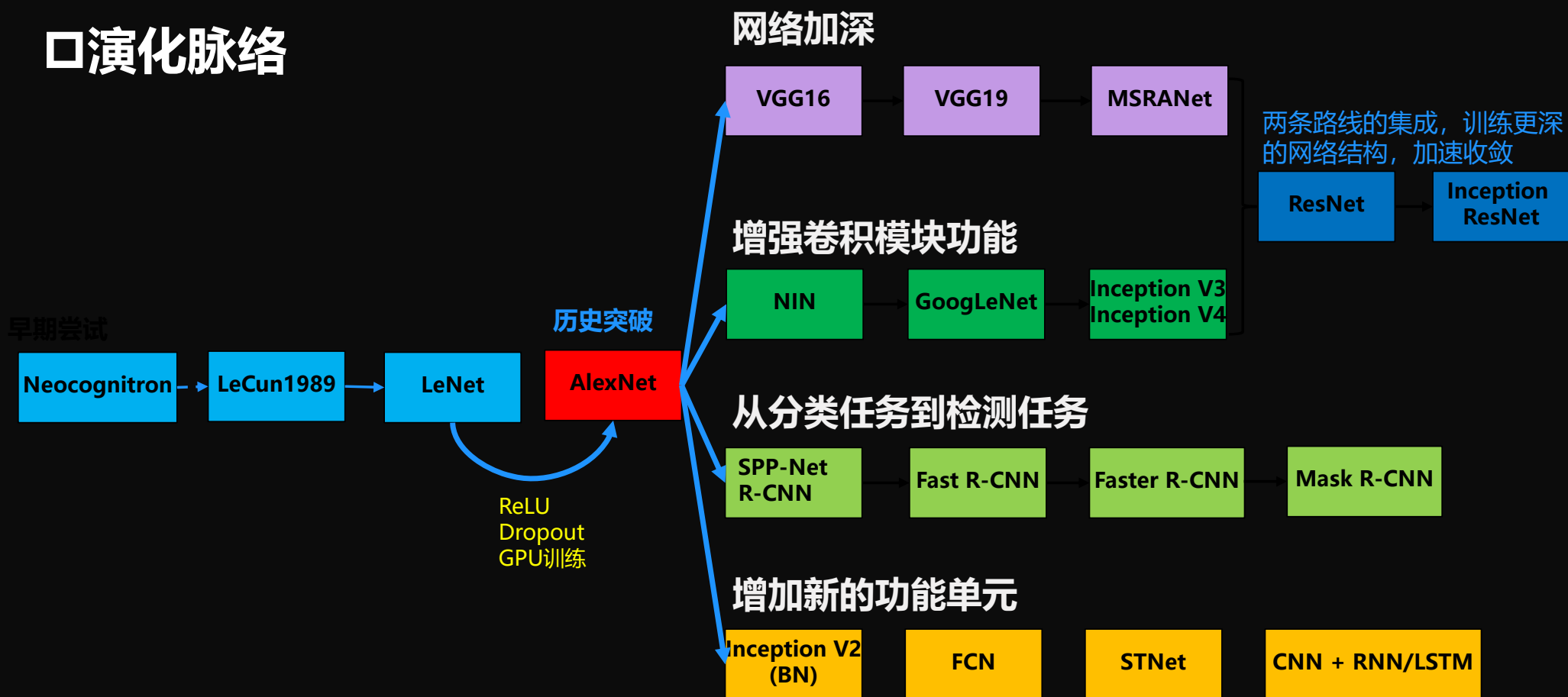


2015



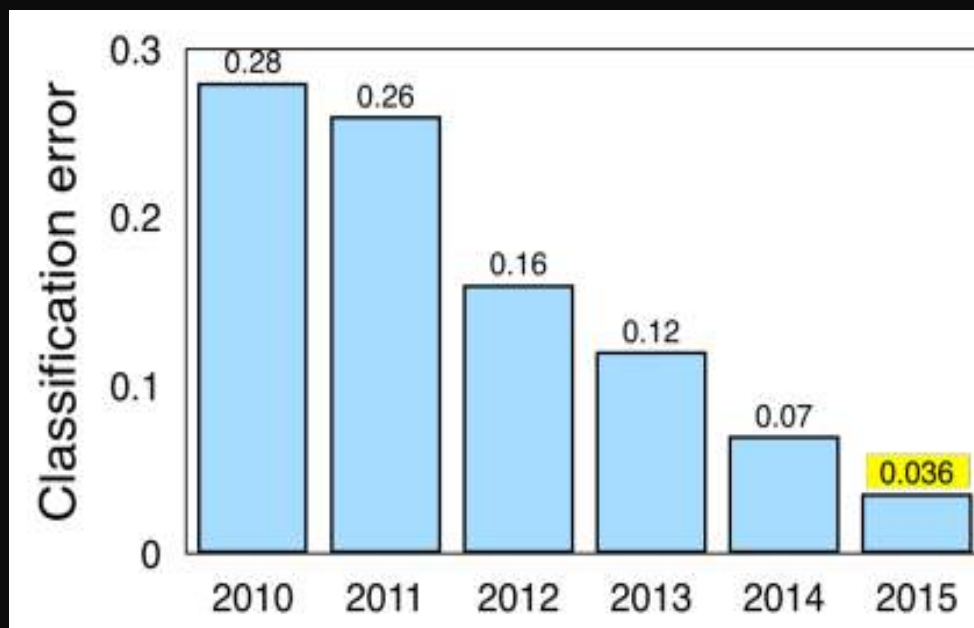
# Renaissance神经网络 (以CNN为例)

## 演化脉络



# 深度学习为计算机视觉带来的进步

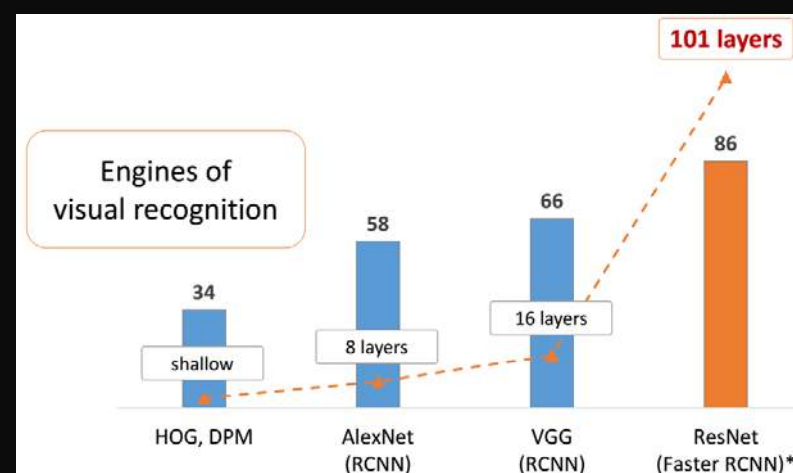
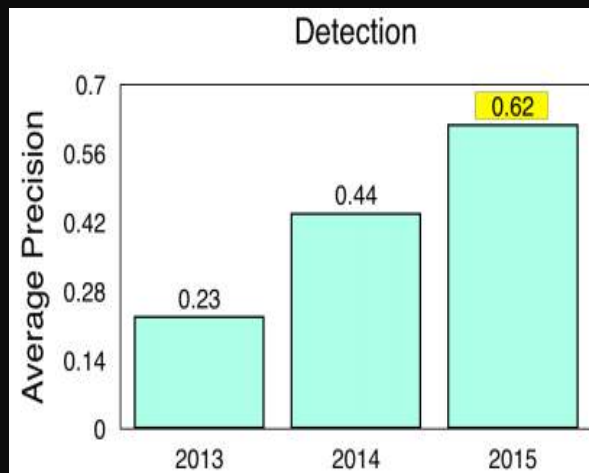
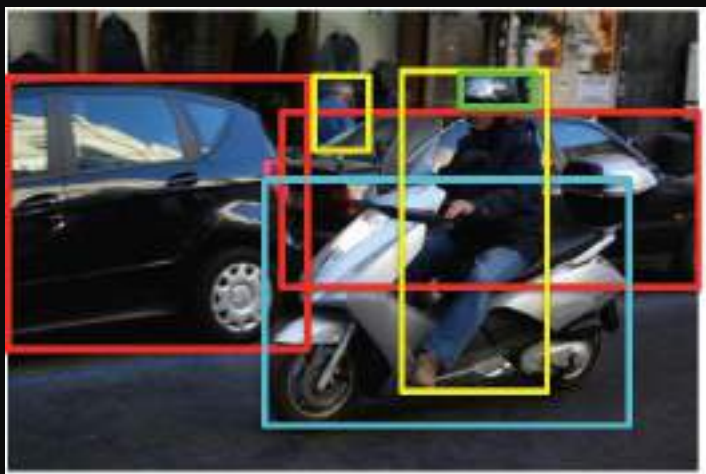
□ 图像分类上的跨越式进步【1000类Top5错误率: 26%→2.3%】



# 深度学习为计算机视觉带来的进步

## □ 物体检测上的跨越(ImageNet竞赛)

- 图像中200类物体检测mAP: 23% → **66%**(2016年) → **73%**(2017)
- 视频中 30类物体检测mAP: 68%(2015) → **81%**(2016年) → **82%**(2017)



Pascal VOC2007 检测性能演化

# 深度学习为计算机视觉带来的进步

## □ 视频结构化技术（中科视拓Seeta视觉引擎）

### ■ 监控场景行人车辆检测、跟踪与属性估计





# 深度学习为计算机视觉带来的进步

## □ 无人机视觉技术（中科视拓Seeta视觉引擎）

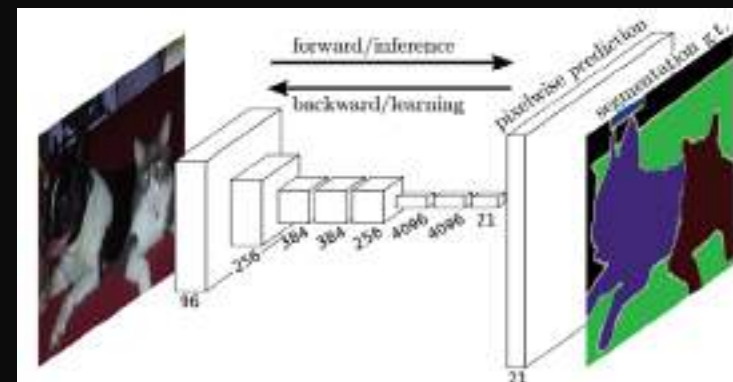
- 地面目标检测与跟踪技术：检测率90%以上



# 深度学习为计算机视觉带来的进步

## 语义分割任务

- VOC2012数据集
- 50%(2013) → **75% (2015)** → **86.9%(2017)**



方法	VOC2012
NUS-UDS CVPR13	50.0
FCN-8s CVPR15	62.2
DeepLab ICLR15	71.6
FCN + CRF-RNN ICCV15	74.7
CMT-FCN-ResNet-CRF 2016	80.0
DeepLabv3-JFT 2017	86.9

VOC2012分割测试Mean IU Accuracy比较



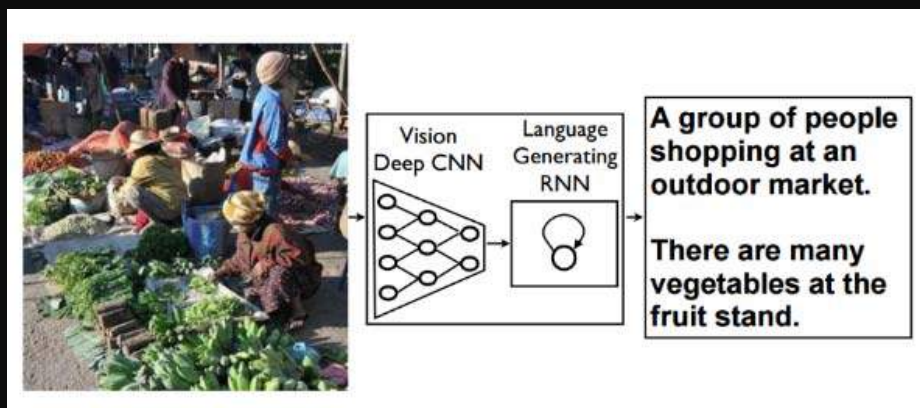
# 深度学习为计算机视觉带来的进步

## □ Image Captioning (看图说话)

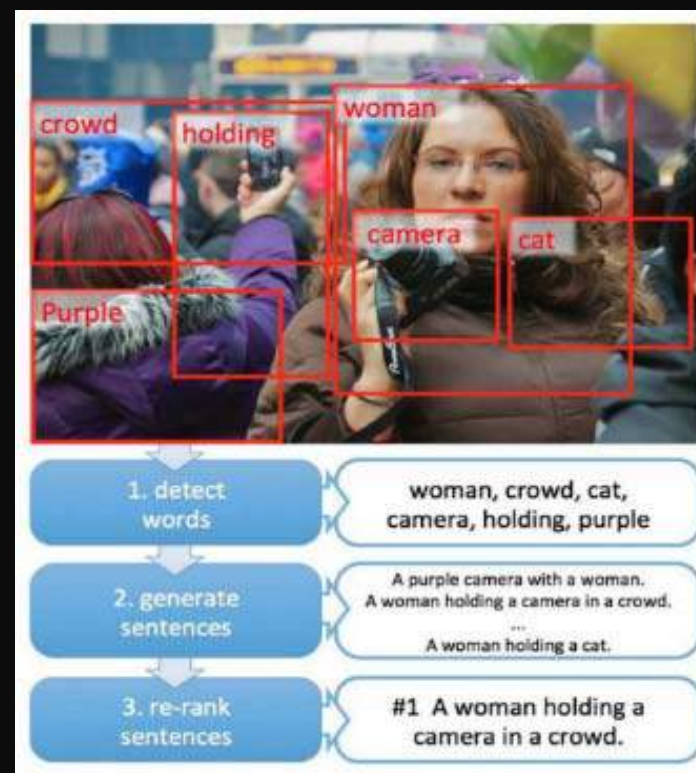
■ MSCOCO2015上 (Google系统)

□ 27.3% 优于人, 31.7% 通过图灵测试

## □ 视觉与语言的联姻



Show and Tell: A Neural Image Caption Generator (a work from Google)



From Captions to Visual Concepts and Back (a work from Microsoft)

# 深度学习为计算机视觉带来的进步

## □VQA: Visual Question Answering

### □主要方法

- 图像用CNN提取视觉特征
- 语言(问句) 用LSTM形成特征
- 二者互动, 以分类或预测方式得到答案 (非常类似于机器翻译)

### □基本进展

- 取得了有限的进步, 依赖于大量问答句子进行训练 (推理能力有限)



# 深度学习为计算机视觉带来的进步

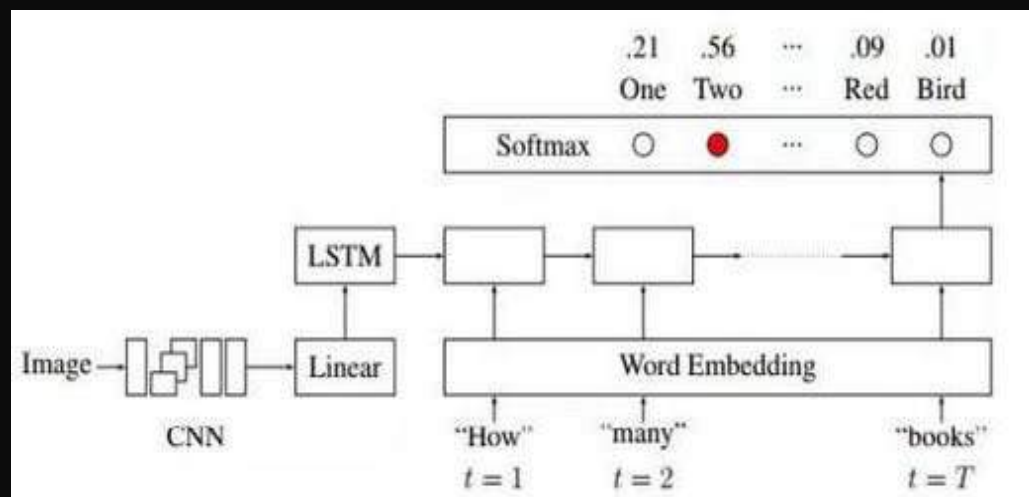
## □VQA: Visual Question Answering

### □主要方法

- 图像用CNN提取视觉特征
- 语言(问句) 用LSTM形成特征
- 二者互动, 以分类或预测方式得到答案 (非常类似于机器翻译)

### □基本进展

- 取得了有限的进步, 依赖于大量问答句子进行训练 (推理能力有限)

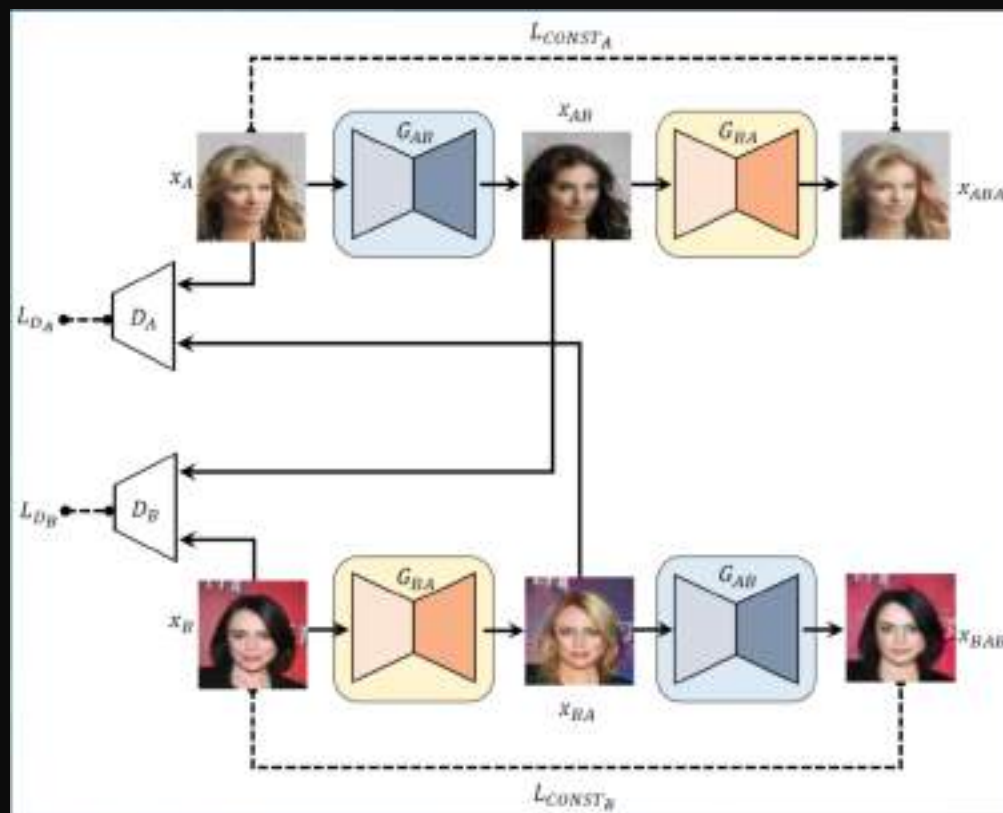


# 深度学习为计算机视觉带来的进步

## □ 图像合成及风格转换

- 生成对抗网络GAN
- 生成目标风格的逼真图像
- 风格的学习

## □ 其意义在于 “举一反三”



# 深度学习为计算机视觉带来的进步

## □ 图像合成及风格转换

- 生成对抗网络GAN
- 生成目标风格的逼真图像
- 风格的学习

□ 其意义在于 “**举一反三**”

