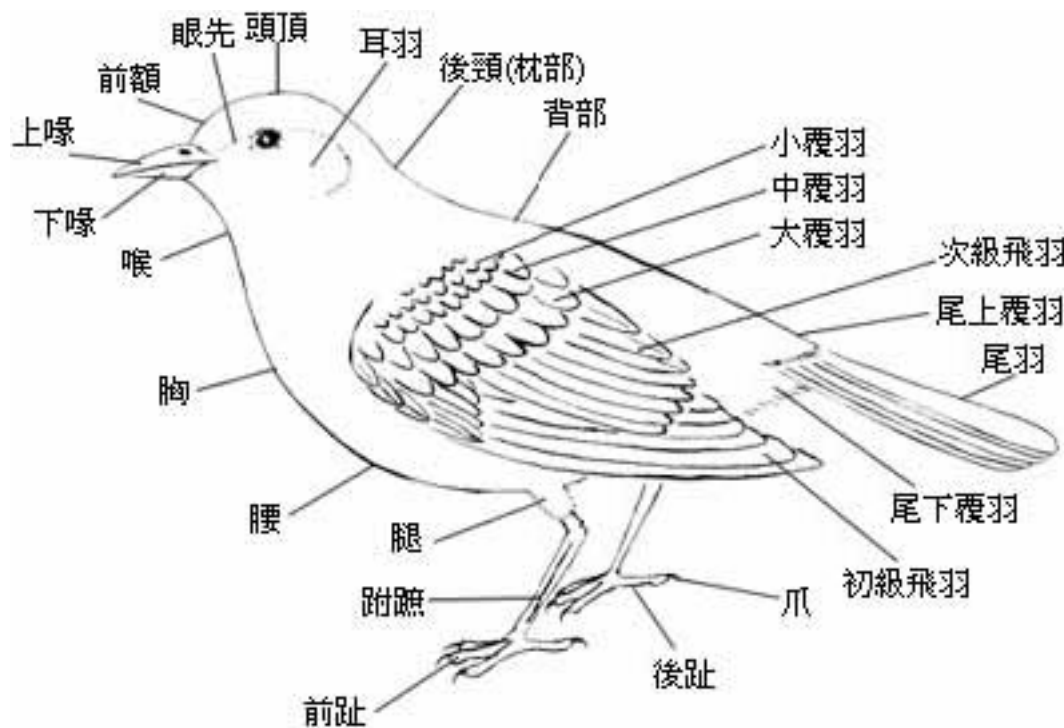


深度学习

稀疏自编码器

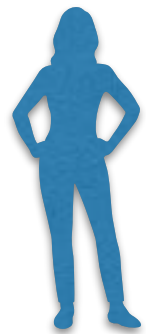
如何识别一只鸟



寻找特征

特征组合

宏观特征的量化



鸟的眼睛可以看做一个特征，那么如何描述这个特征？

特征不具有稳定的形态，难以准确描述。



房子的面子可以看做是房子的一个特征，那么如何描述这个特征？

特征具有稳定的形态，可以准确描述。

宏观特征的量化

宏观特征是对象或事件在宏观上具有的特性或属性。宏观特征可能是静态的，也可能是动态的。

静态的宏观特征具有表达无歧义性且可量化的特性。

动态的宏观特征具有抽象性、多态性。难以量化。

小练习

以下场景中包含的特征哪些是动态宏观特征？

A. 森林中地上的树叶



B. 一个杯子中的水量

C. 电视屏幕上某一点的像素值

D. 照片中的人脸



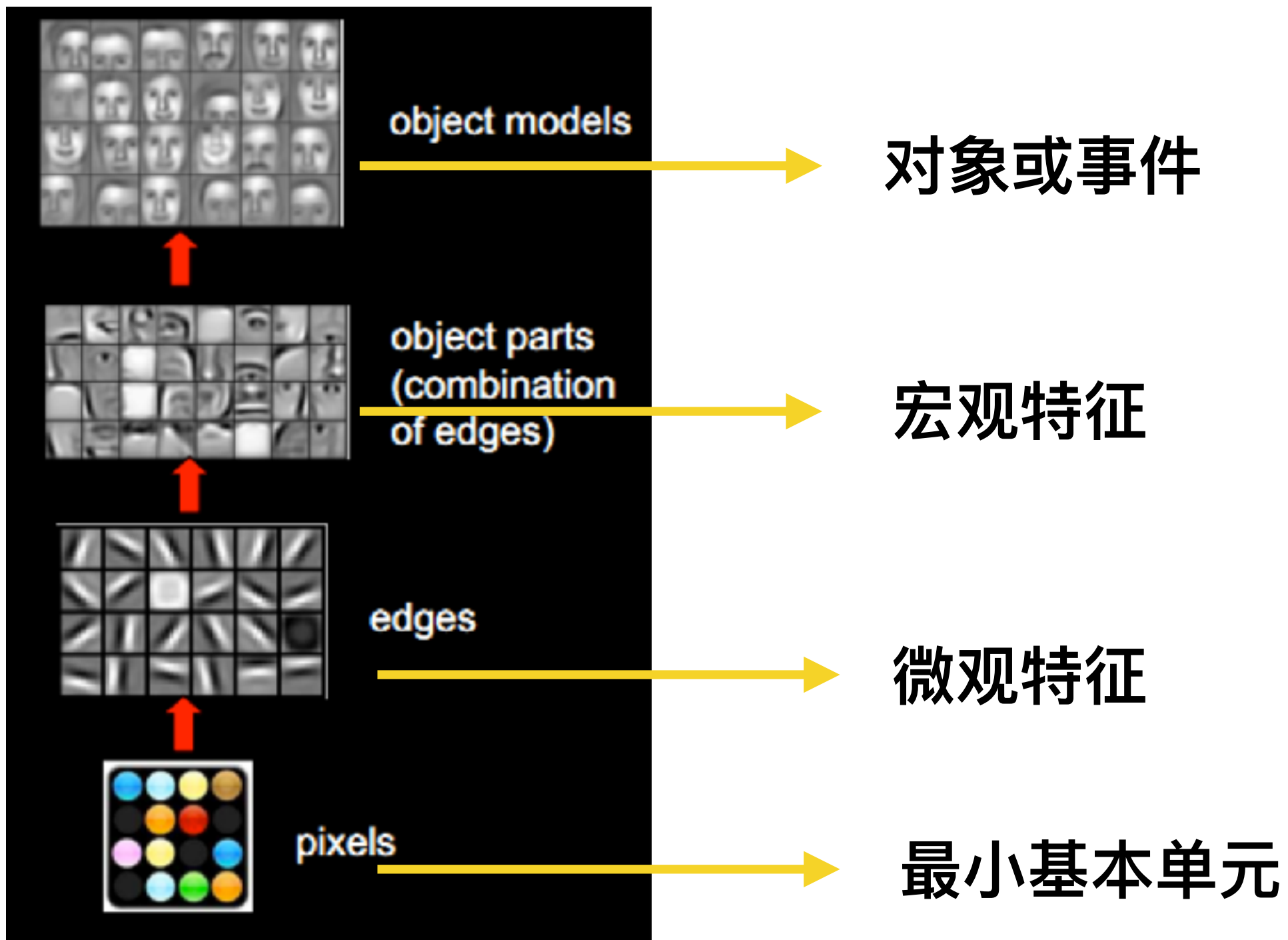
动态宏观特征的表达难题



20世纪70年代基于文法规则的语义识别就被大部分研究机构抛弃了。原因是文法规则太过于庞大，其规模甚至超过了语句的规模。而识别率却很低。

动态宏观特征难以量化与直接使用

微观特征



微观特征

思考： 图中的一个像素值算不算图像的特征？

像素值不能算作一张包含信息的图片的特征，因为单个像素值不具有区分性。一般来讲特征应该具有以下特点：富有信息量、有区分性。

思考： 像素值组合成的特定像素块是否为特征？

如果这个像素块是包含信息的，则可以看做是一个特征。

微观特征

思考：微观特征是动态的吗？如果是动态的会发生什么？

微观特征具有稳定性，即其是静态特征。

首先不存在动态微观特征，微观特征均是静态的。如果其具有动态特征的表现，那么要么它不是特征，要么这个特征是继续可微的。

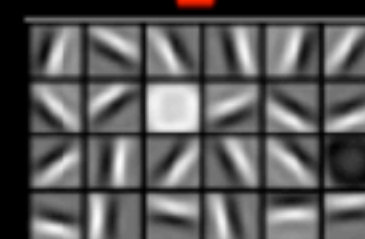
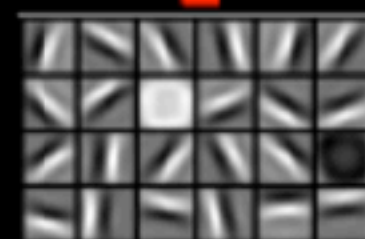
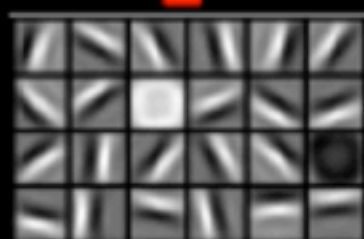
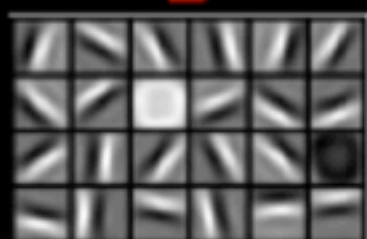
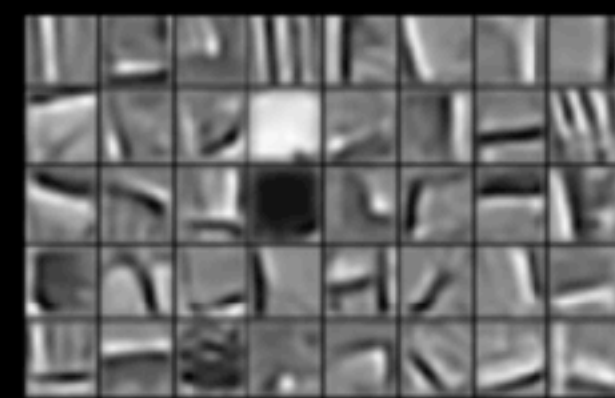
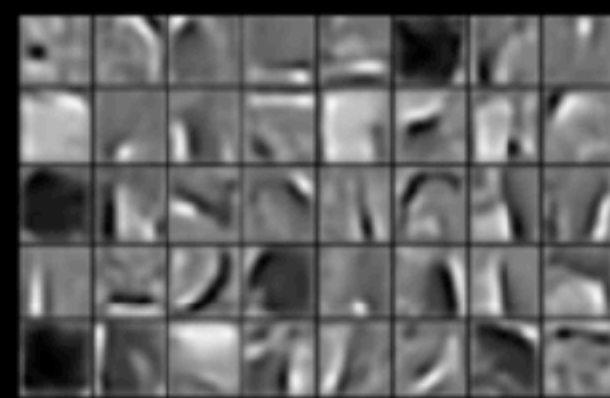
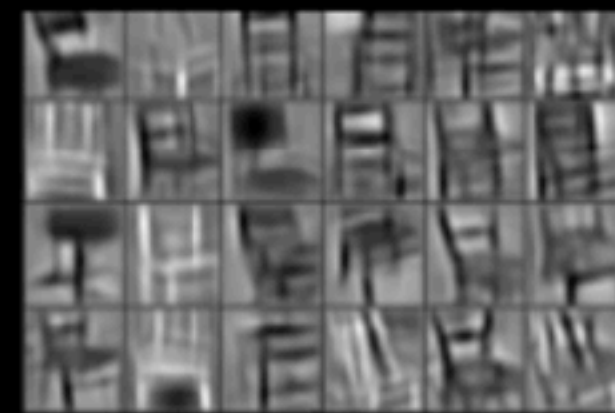
Features learned from different object classes.

Faces

Cars

Elephants

Chairs



[Lee, Grosse, Ranganath & Ng, 2009]

从不同对象学习到了相同的微观特征

Author: WangQi
Email: wangqikaixin@gmail.com

神经网络中输入与特征的关系

神经网络的输入包括两类：样本对象与样本特征。

输入样本特征的例子：

预测房价中输入房屋面积。

预测肿瘤性质中输入肿瘤大小。

预测上市公司股票走势输入财报。

输入样本对象的例子：

识别图片中的物体时输入图片

语音识别中输入语音

语义分析中输入文本

当样本特征不可量化和或者样本特征无法代表原始数据的全部信息是，考虑将样本输入神经网络。

自然界信息的特点

- **信息是冗余的。**例如人像照片：肤色信息是冗余的，并不需要那么多像素就可以表达出肤色。
- **信息是包含噪声的。**例如人像照片：除非在特定的环境中拍摄，否则照片中肯定包含噪声数据。
- **特征（指微观特征）是稀疏的。**例如人像照片：仅需要十余种微观特征的不同组合即可表达出照片内容。

大脑神经网络的稀疏激活性

谣言

经常的可以看到：人类只开发出了人脑中只5%的功能。

正解

人脑中一般最多只有1%-4%的神经元处于激活状态。



神经元具有稀疏激活性

神经元的抑制与激活

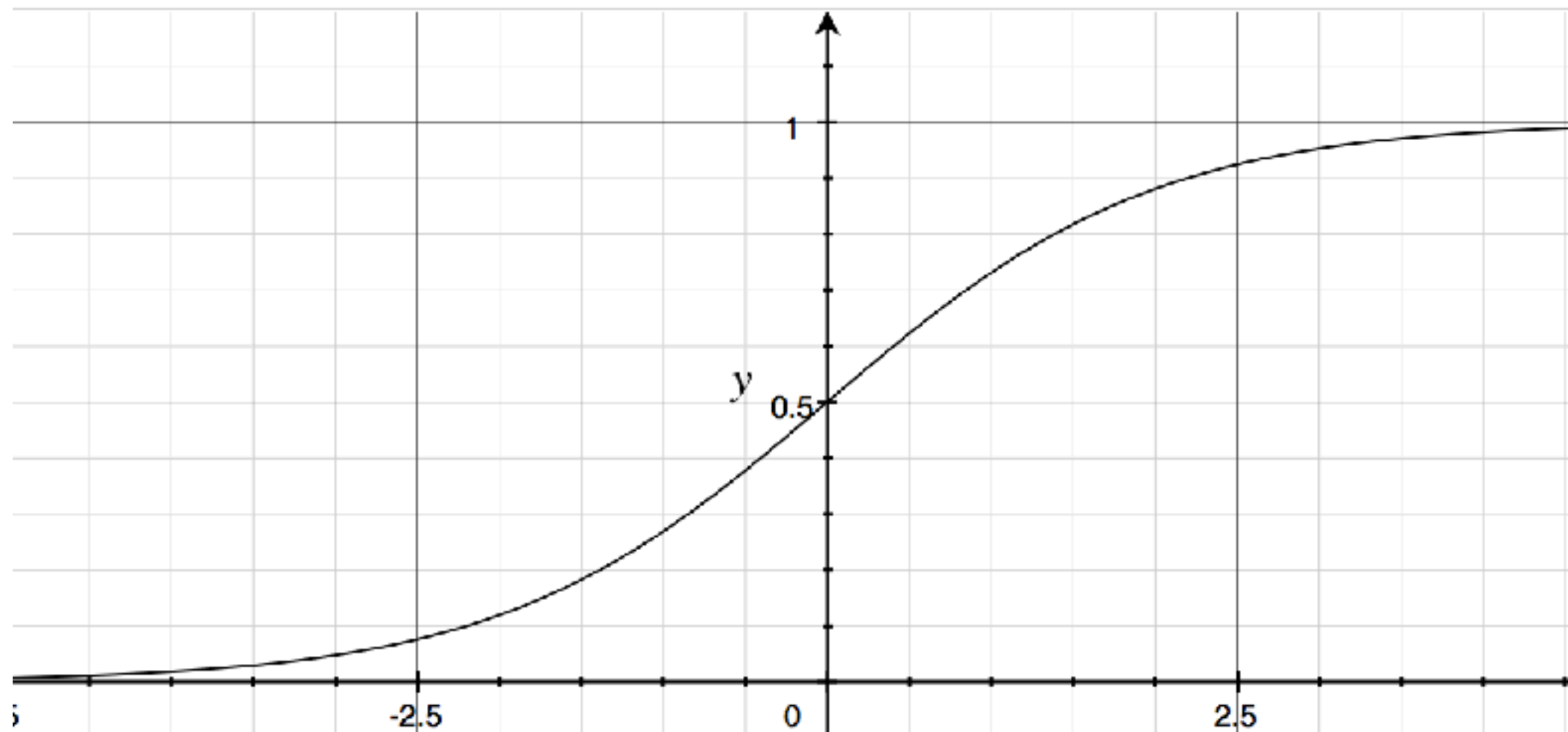
生物神经元中，突触可以分为抑制性与兴奋性两种。分别控制下一个神经的抑制与激活。

抑制：降低或中断信号传递。

激活：将信号传入下一个（或一批）神经元。

神经元的抑制与激活

sigmoid神经元把输出接近于0定义为抑制，接近于1定义为激活。



神经元的抑制与激活

ReLU把神经元输出为（或接近）0定义为抑制，反之为激活。

tanh神经元把输出接近-1定义为抑制，接近于1定义为激活。

不同种类的神经元，神经元的激活与抑制定义也不尽相同。

稀疏自编码器

目的

降低神经元的平均激活度。

激活函数

以sigmoid激活为例。

方法

稀疏性约束

意义

将数据变为稀疏的特征表达。

使用场景

特征表达、压缩

隐藏层神经元的激活数量

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^m (a_j^{(2)}(x^{(i)}))$$

上式中m表示样本数量，i表示样本索引，j表示神经元索引，a表示激活度。所以：

$a_j^{(2)}(x^{(i)})$ 表示隐藏层（第二层）第j个神经元在输入样本 $x^{(i)}$ 的情况下的激活度

通过运行前向传播算法，得到隐藏层的输出，就可以求得激活度。

KL散度

稀疏性约束方法：KL散度

KL散度可以用来衡量两个概率分布的相似性，对于离散型的随机变量，其概率分布P和Q的KL散度公式如下：

$$D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$$

由于KL散度是两个概率分布P和Q差别的非对称性的度量，所以在此处，我们需要规定P表示真实分布即期望的激活度，Q表示理论分布即实际的激活度。

通过KL散度可以计算神经元的激活度与期望的激活度的距离

隐藏层神经元的激活度

如果我们用：

$\hat{\rho}_j$ 表示实际激活度，用 ρ 表示期望的激活度，则它们之间的 KL 散度可以写成：

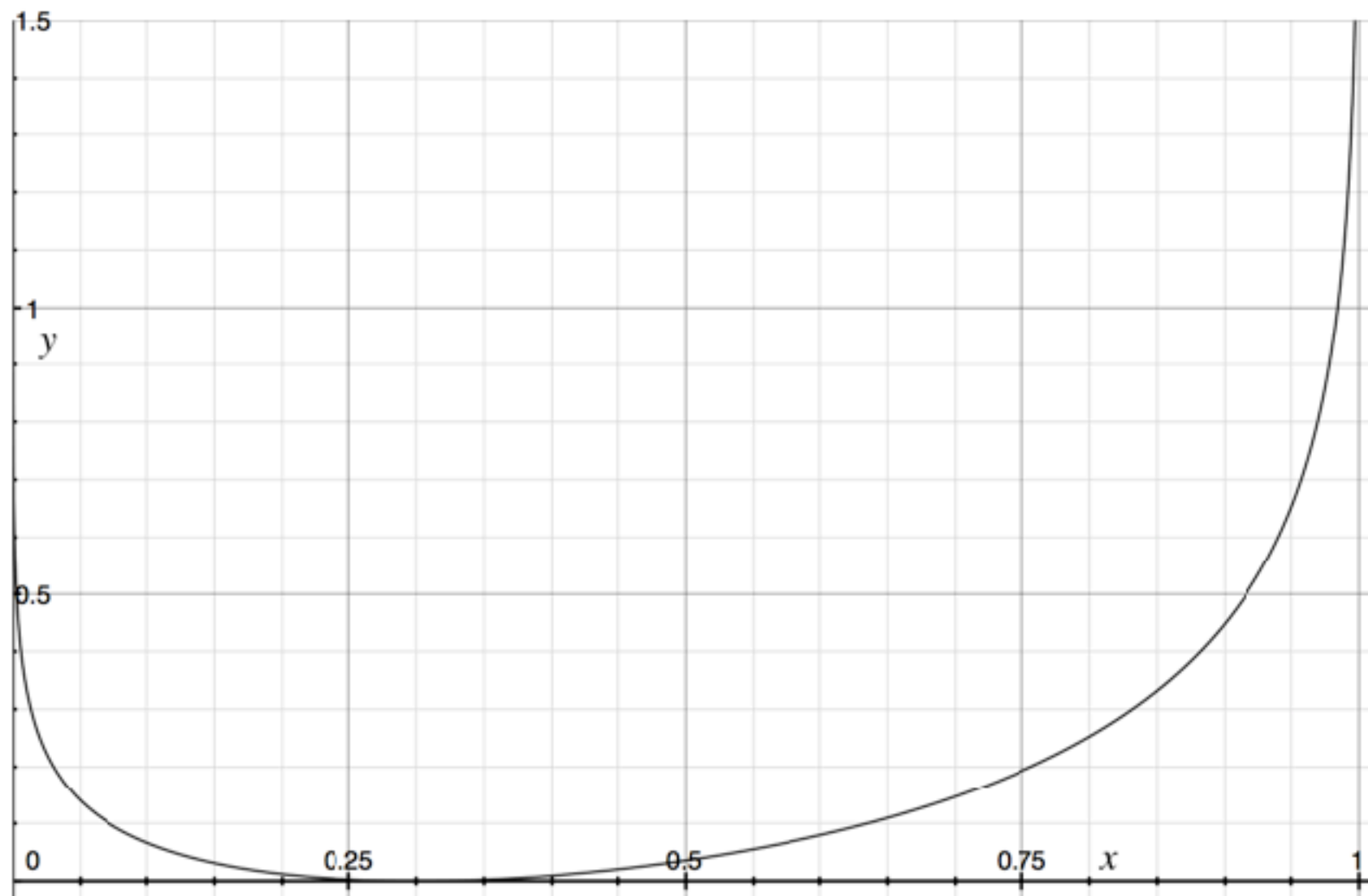
$$KL(\rho||\hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j}$$

当实际激活度接近期望激活度时，KL散度接近于极小值。

当实际激活度远离期望激活度时，KL散度区域无穷大。

隐藏层神经元的激活度

举例来说，在下图中，我们设定期望激活度为0.3（实际中通常设置**0.05**，但这样会使图像变得不易观察，故以0.3为例），横坐标表示实际激活度，纵坐标表示KL散度。



稀疏性自学习

自编码器中，通过设置均方误差代价函数，并执行梯度下降算法，ANN便可以迭代更新参数，达到特征提取的目的。同样的，我们也希望自编码器能够自己学会稀疏特征的表达，这时候我们需要改进代价函数，将隐藏层的激活度——KL散度加入代价函数即可。

代价函数

稀疏性约束的代价函数就是：

$$J_{KL} = \sum_{j=1}^{N^{(2)}} KL(\rho || \hat{\rho}_j)$$

所以总体的代价函数可以表示为：

$$J_{sparse}(w, b) = J(w, b) + \beta \sum_{j=1}^{N^{(2)}} KL(\rho || \hat{\rho}_j)$$

其中 $J(W, b)$ 代表原有的代价函数， β 控制稀疏性惩罚因子的权重。 $\hat{\rho}_j$ 则来自于前向传播。

这里需要注意的是在反向传播算法中，我们也需要做一些修改，我们只需要在求隐藏层的残差时再加入一项残差导数即可（输出层的残差不变）：

$$\delta_j^{(2)} = \sum_{k=1}^{N^{(3)}} \delta_k^{(3)} \cdot w_{kj}^{(3)} \cdot \sigma'(z_j^{(2)}) + \beta \left(-\frac{\rho}{\hat{\rho}_j} + \frac{1 - \rho}{1 - \hat{\rho}_j} \right) \cdot \sigma'(z_j^{(2)})$$

如何理解平均激活度低？

假如有100个神经元，平均激活度是5%代表着什么？

理解1：一次正向传播过程中，大约有5个神经元在工作。

理解2：一百次正向传播过程中，某个神经元大约工作了5次。

两种理解方式均正确，但“理解2”表达更准确。“理解1”没有指明单个神经元的激活度。我们计算神经元激活数量的公式表达了我们希望每一个神经元均具有较低的平均激活度，而非总体的激活度低。

稀疏激活性与稠密激活性对比

举例：MNIST手写数字图片使用AE与SAE隐藏层的激活情况。

输入层与输出层均为784个神经元。AE的隐藏层设置为50个神经元，SAE的隐藏层设置为200个神经元。当训练完成后：

输入一个标记为“5”的样本

输入一个标记为“7”样本

AE的50个神经元全部激活 → 激活单元相同 ← AE的50个神经元全部激活
SAE大约有10个神经元激活 → 激活单元不相同 ← SAE大约有10个神经元激活

结论：稀疏激活性学到的特征与其它特征之间的区分度更大，特征与特征之间影响更小。

稀疏自编码器的应用

1. 稀疏特征表达。
2. 数据压缩。
3. 用于组成其它其它自编码器。

小结

- 宏观特征包含动态特征与静态特征两种。
- 动态宏观特征难以量化。
- 微观特征均可以量化，但往往难以直观表示，微观特征一般由神经网络学得。
- 神经网络的输入包括样本对象与样本特征两种。
- 通常稀疏激活性的神经网络在特征表达上更有优势。
- KL散度可以衡量神经网络的隐藏层的激活度与期望激活度的距离。
- 给自编码器的代价函数加入激活度代价就成为了稀疏自编码器的激活函数。
- 稀疏自编码器的隐藏层神经元数量通常要多一些。

THANKS