机器学习在金融数学建模中的应用

中山大学数学学院

冯国灿

@ 第三届粤港澳金融数学建模

2022年10月22日

提纲

- 背景
- 机器学习与神经网络
- 深度学习及应用
- 金融数学建模
- 小结

0 背景

1. 金融领域

传统模式：（定性分析，基本面，基本指标）

新格局： 金融开放（2020 年 4 月 1 日，我国全面放开金融资本市场）

人才需求： 金融科技人才150万 （21世纪经济报道）

其中：金融量化 20万。

什么是金融数学建模，为什么需要量化分析

保尔·拉法格在《忆马克思》中谈到,马克思认为：“一种科学只有在成功地运用数学时,才算达到了真正完善的地步.”

大数据，人工智能， 机器学习 （案例：幻方量化）

“大湾区杯”
2022年粤港澳金融数学建模竞赛

竞赛时间：11月1日（周一）至11月8日（周一）（7天）。
指导单位：广东省科学技术协会
主办单位：
广东省工业与应用数学学会
粤港澳国家应用数学中心
万联证券

协办单位：泰迪智能有限公司
浙江核新同花顺网络信息股份有限公司
深圳点宽网络科技有限公司
广东聚智诚科技有限公司，

为什么要发起这项活动？

判别天

涨幅

≥1%，正类

其他，零类

≤-1%，负类

标记天

有效样本

0背景

- 机器学习，大数据，人工智能
- 深度学习， Deepface系列, Alpha Go : DeepMind
- 金融领域： 人

图像生成 x=arg max/rand p(x) / , p(x|w)
风格转换
captioning
画->诗

风格迁移

Real : a small zebra is grazing on some grass on a sunny day
Generated : a zebra walking across the grass in a field

1．机器学习与神经网络

机器学习（Machine Learning，ML）：

学习是人类具有的一种重要智能行为。

人：*学习方式 知识、经验、比较…，*

机器学习：根据数据（经验）发现或构造算法或程序实现数据的规则发现或分类。

类别：、

- 监督学习（*Supervised L.*）
- 无监督学习（*Unsupervised L.*）

  半监督学习*Semi-S.L.*
- 增强学习（Enforcement L.）

- 无监督学习 （聚类）

数据样本： $\{x_1, x_2, \ldots, x_n\}$，$x_n \in R^d$ *根据数据特点划分成两个（若干个）子集合，*

子集合内:

子集合间:

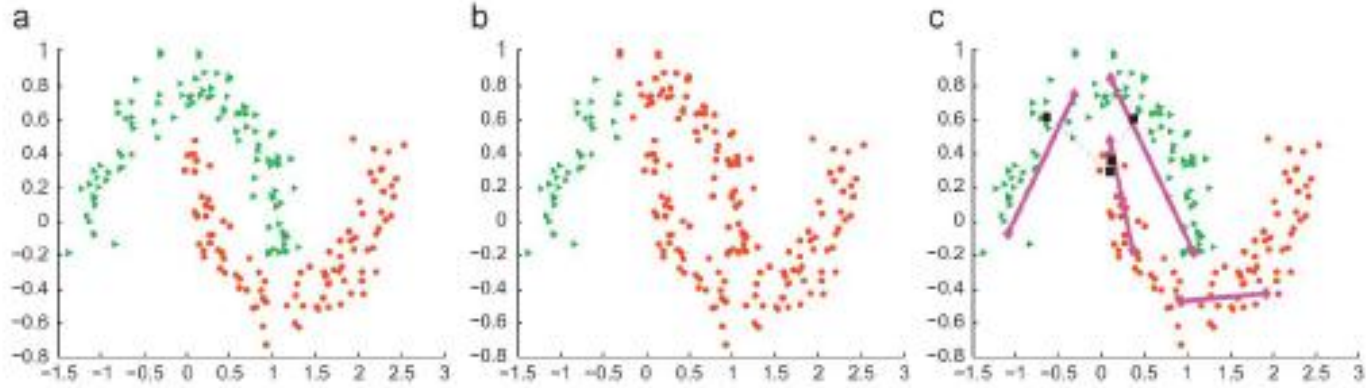


Fig. 5. Two "moons" and experimental results. (a) The original data set. (b) result given by Ncut. (c) result given by NSDR-Ncut. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- 监督学习 （分类, 回归）

给出：数据样本： $\{x_1, x_2, \ldots, x_n\}$ （$x_n \in R^d$）*和类标数据*

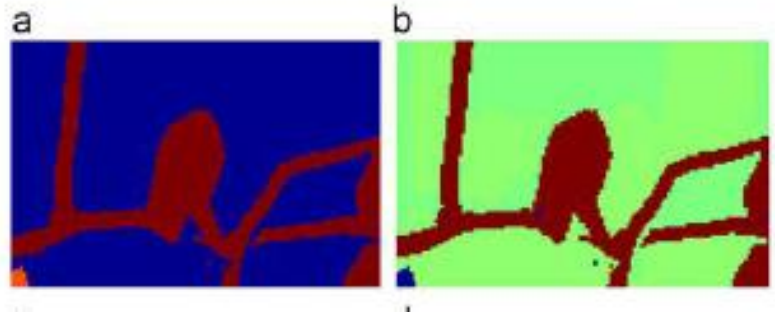
$\{y_1, y_2, \ldots, y_n\}$ （$y_n \in \{0, 1\}$）

推断： $x \to y$*的关系，* $y = f(x)$

- *半监督学习*

USL +SL

Fig. 14. Two-way segmentation result given by NSDR-Ncut with five pairs of 3×3 block-wise similarity constraints connected by magenta solid lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



其他学习：

协同，增强，
字典，迁移…

协同学习（collaborative learning)

原理（两类为例）：

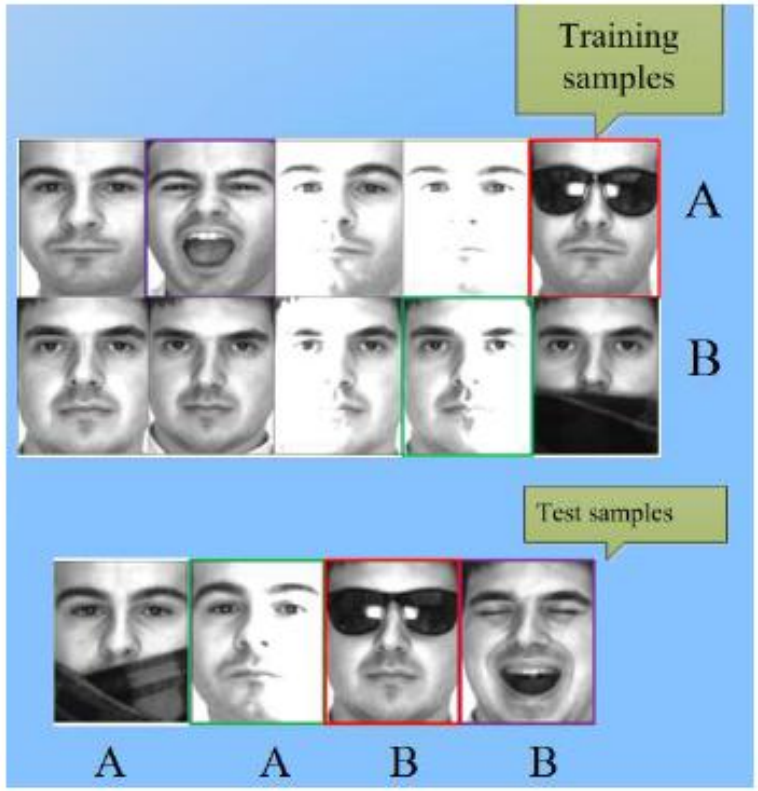
- *对象O_A 具有特征$a_1, a_2, \ldots, a_r$*
- *对象O_b 具有特征$b_1, b_2, \ldots, b_r$*

如果

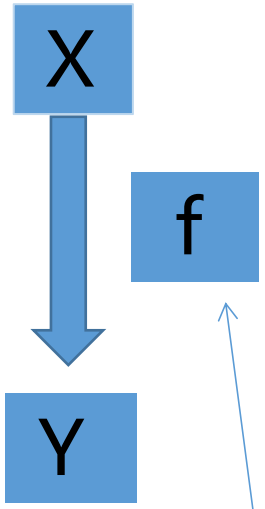
$sim(O_A, O_b)$ 高，

O_A 的特征$a_{j1} \ldots a_{jk}$缺失，

$b_{j1} \ldots b_{jk} \Rightarrow a_{j1} \ldots a_{jk}$



神经网络



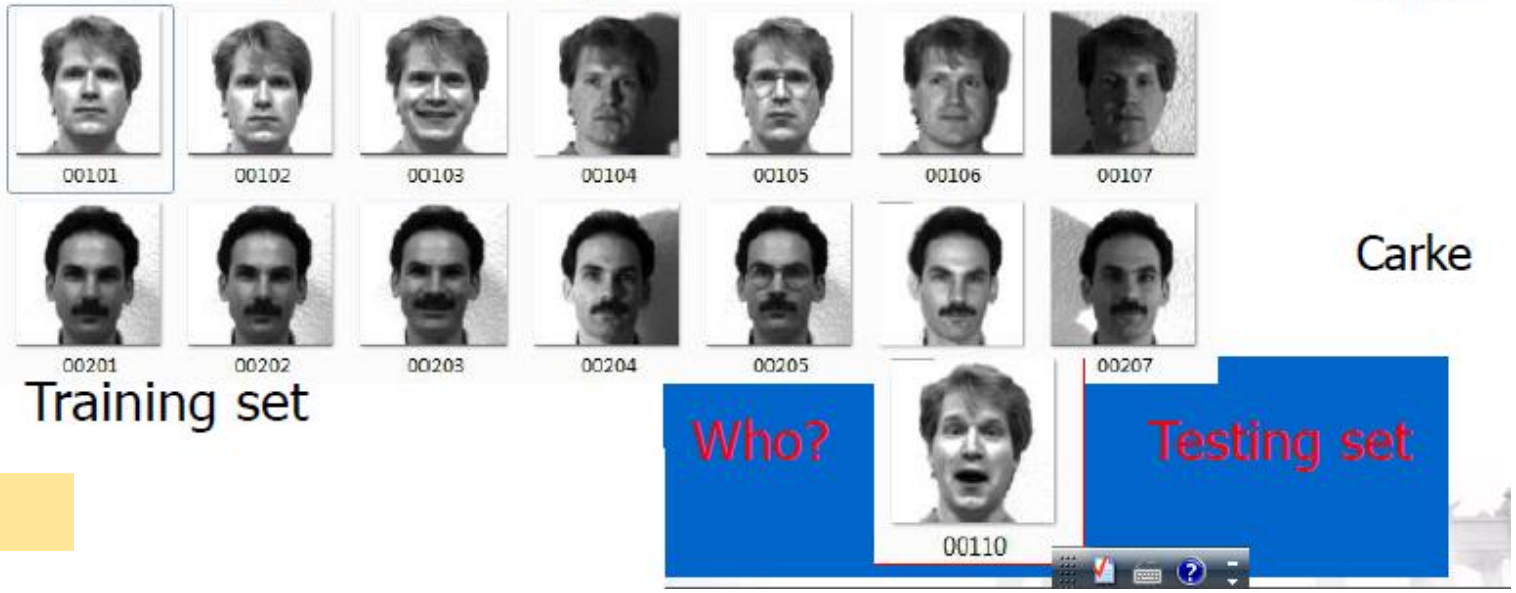
找一种结构实现　f　的功能

•监督学习 （分类, 回归）

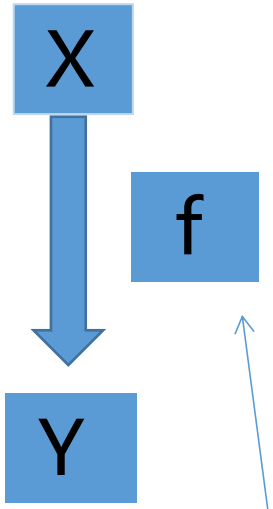
给出：数据样本：$\{x_1, x_2, \ldots, x_n\}$ （$x_n \in R^d$）和类标数据

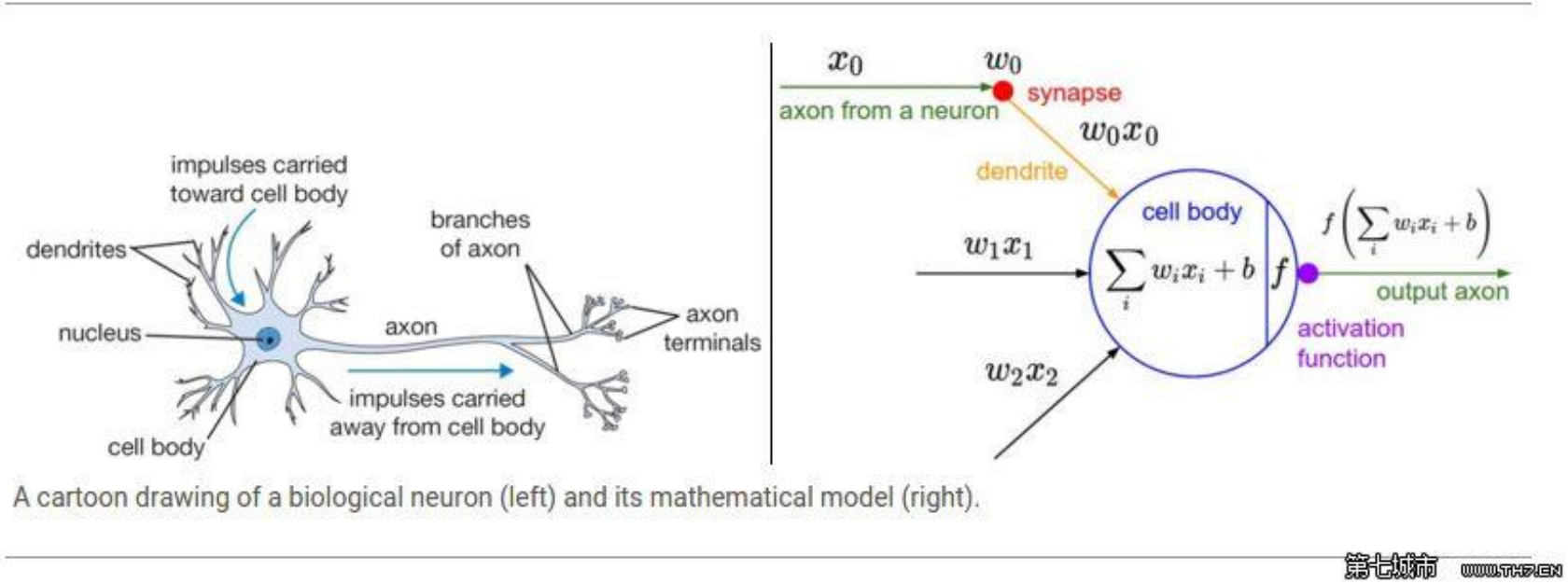
$\{y_1, y_2, \ldots, y_n\}$ （$y_n \in \{0, 1\}$）

推断：x->y的关系， y=f(x)



神经网络

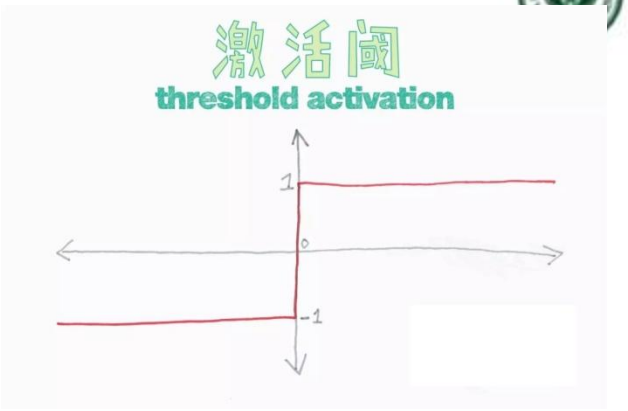




A cartoon drawing of a biological neuron (left) and its mathematical model (right).

找一种结构实现 f 的功能
Functionist ： y=f(x,θ)
Connectionist : 神经网

$$f(\sum_i w_i x_i + b)$$



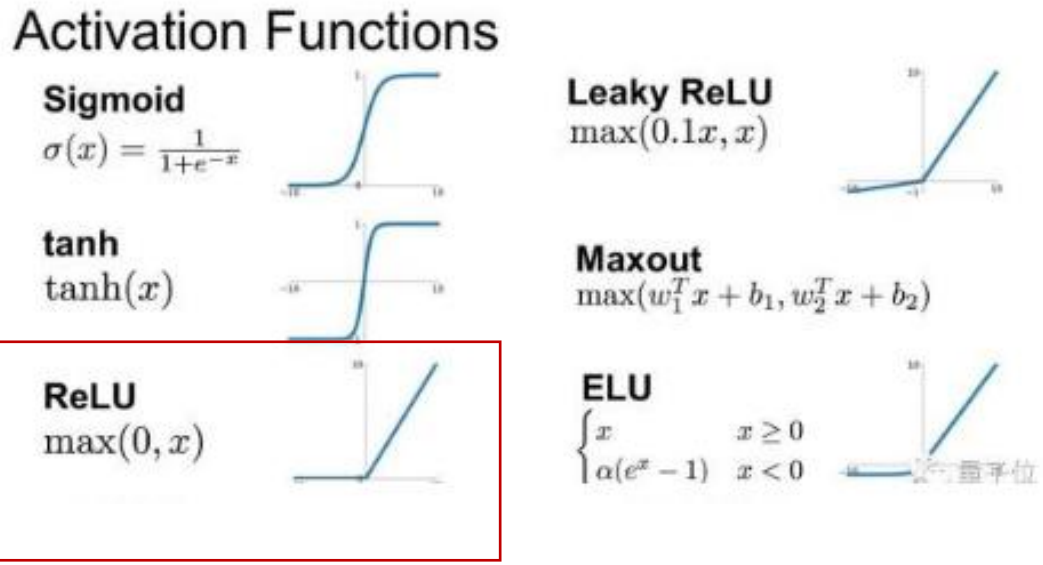
- 激活函数：

  0-1开关，

  Sigmoid

  relu(rectified linear function)

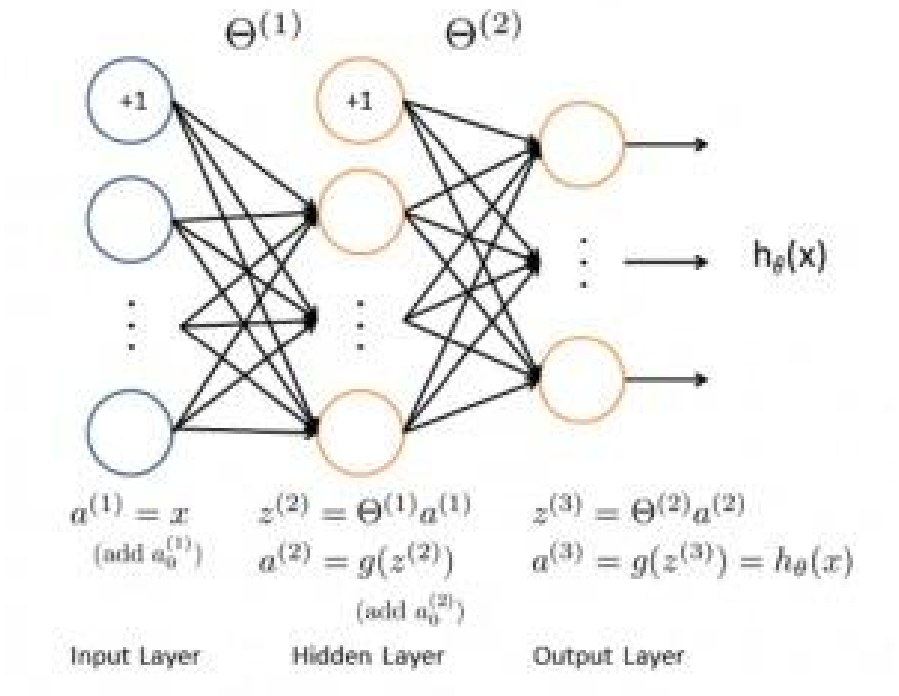
  f(x)=max(0,x)



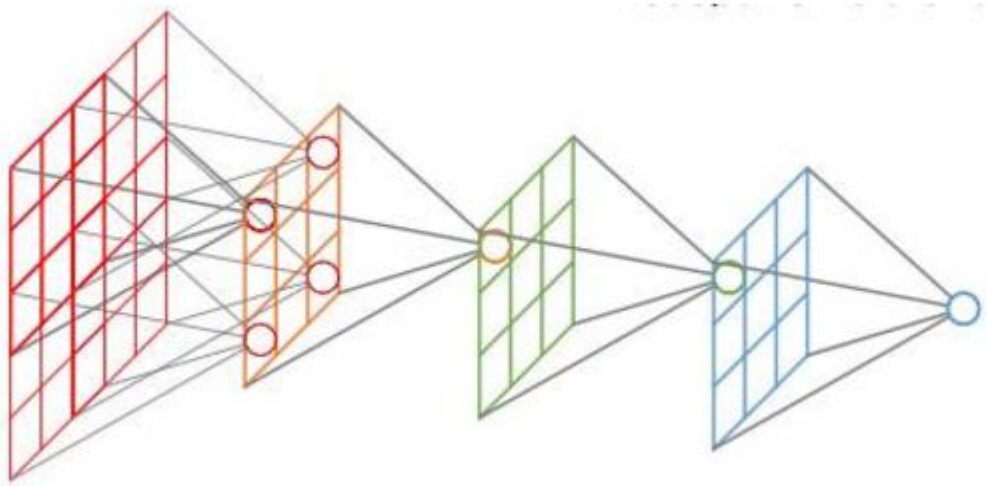
- 类别：感知机，ＢＰ，　Hopefiled NN, Boltzman机，ＲＢＭ，　…, CNN, RNN,

Convolutional NN(CNN) 卷积神经网

- BP



- CNN



局部连接

卷积的表达式为：

$$S(t) = \int x(t-a)w(a)da$$

离散形式是：

$$s(t) = \sum_a x(t-a)w(a)$$

这个式子如果用矩阵表示可以为：

$$s(t) = (X * W)(t)$$

其中星号表示卷积。

如果是二维的卷积，则表示式为：

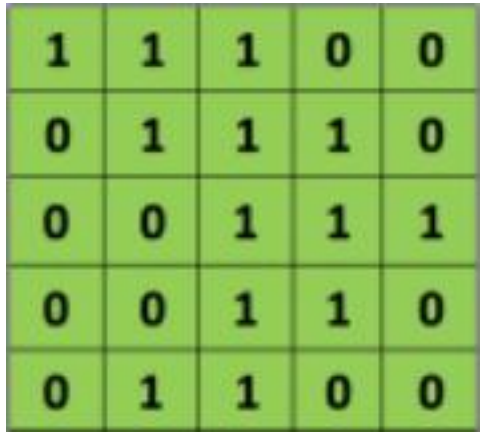
$$s(i,j) = (X * W)(i,j) = \sum_m \sum_n x(i-m, j-n)w(m,n)$$

在CNN中，虽然我们也是说卷积，但是我们的卷积公式和严格意义数学中的定义稍有不同,比如对于二维的卷积

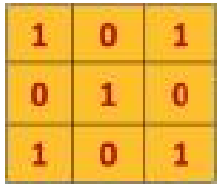
$$s(i,j) = (X * W)(i,j) = \sum_m \sum_n x(i+m, j+n)w(m,n)$$

提取与w相关的结构

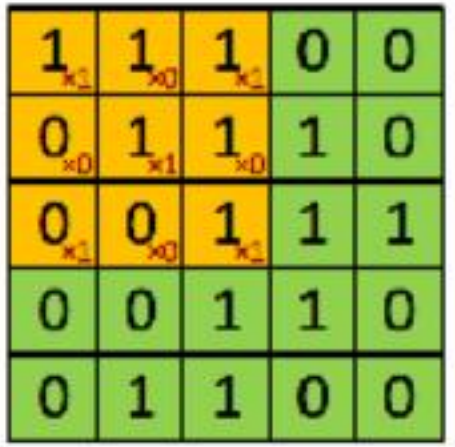
卷积运算过程



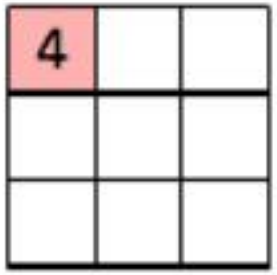
Binary image



filter

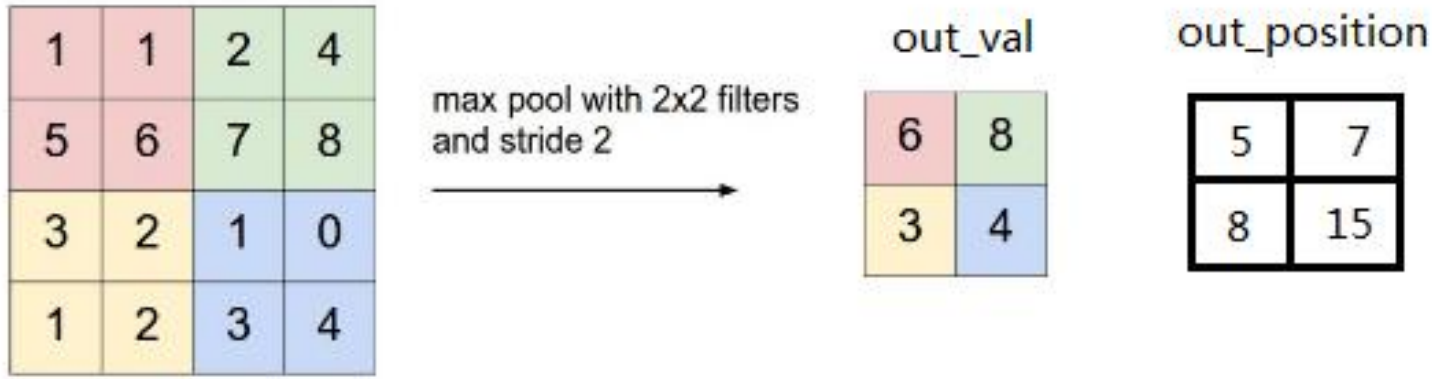


Image



Convolved Feature

Max pooling (池化)



max pool 前向传播

Relu(x)

Cnn： 将具有滤波器特征的选择并保留

Convolutional networks convolutional (LeCun, 1989),

- 机器学习与神经网络
- 深度学习
- 应用案例
- 问题思考

2 深度学习

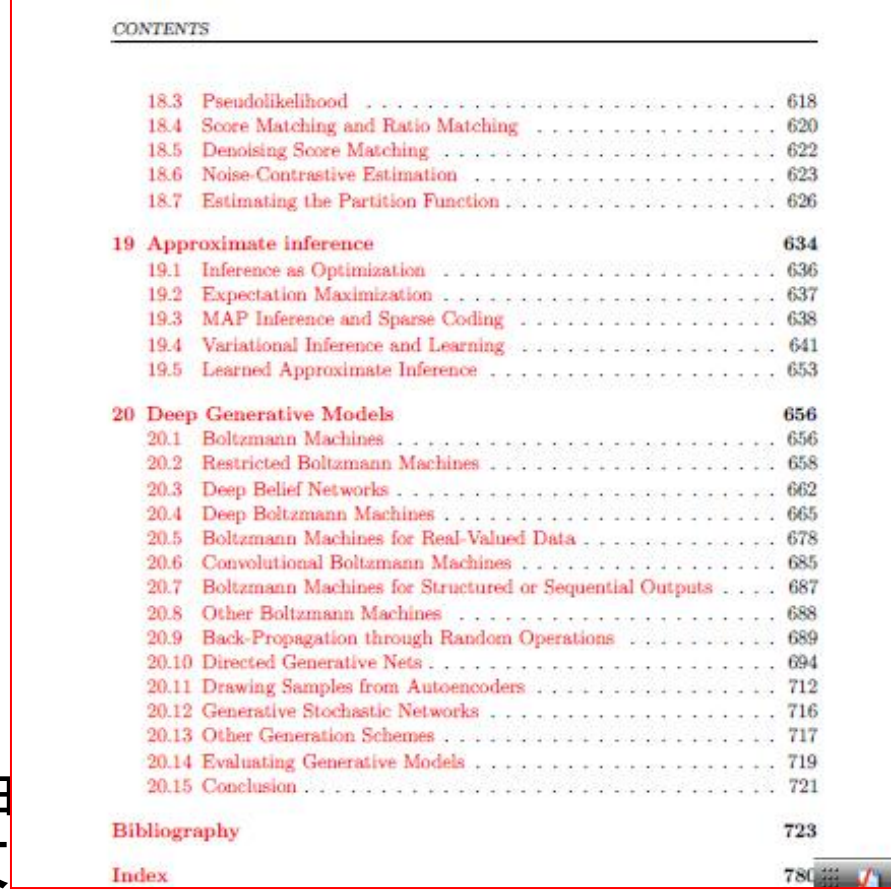
- Deep Learning （书）

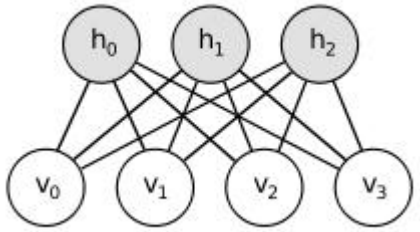
  Ian **Goodfellow**， Yoshua Bengio， Aaron Courville

[1] Geoffery E**. Hinton,** Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006 Jul 28;313(5786):504-7.

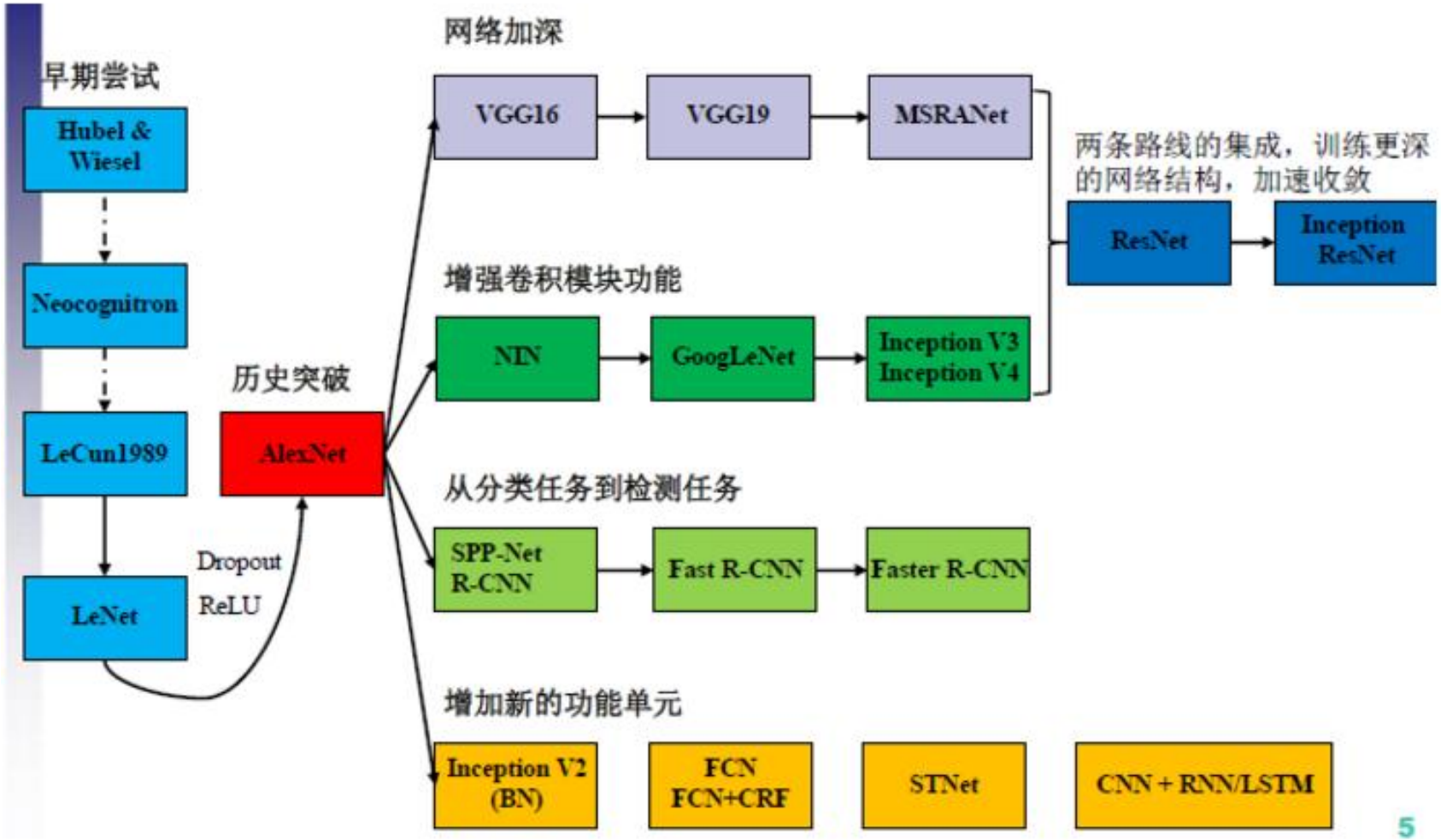
Alex在2012年提出的Alexnet网络结构模型引爆了神经网络的应用热潮，并赢得了2012届图像识别大赛的冠军

- Deep Restrict Boltzman M(DRBM)
- DCNN ,Alexlet, DesenseNet, ResNet, U-Net… ,GAN
  对抗生成网络（GAN）， Wasserstein-GAN
- D recursive NN





RBM 深度学习


早期尝试
Hubel & Wiesel
Neocognitron
LeCun1989
LeNet
Dropout
ReLU
历史突破
AlexNet
网络加深
VGG16
VGG19
MSRANet
增强卷积模块功能
NIN
GoogLeNet
Inception V3
Inception V4
两条路线的集成，训练更深的网络结构，加速收敛
ResNet
Inception ResNet
从分类任务到检测任务
SPP-Net
R-CNN
Fast R-CNN
Faster R-CNN
增加新的功能单元
Inception V2 (BN)
FCN
FCN+CRF
STNet
CNN + RNN/LSTM

深度学习-DCNN（卷积神经网）

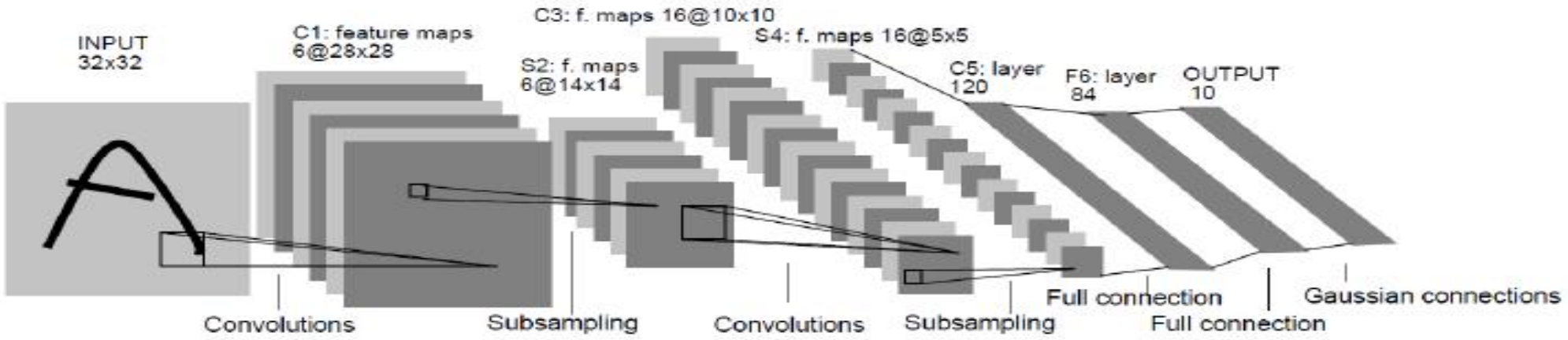


图 4. 卷积神经网络结构。

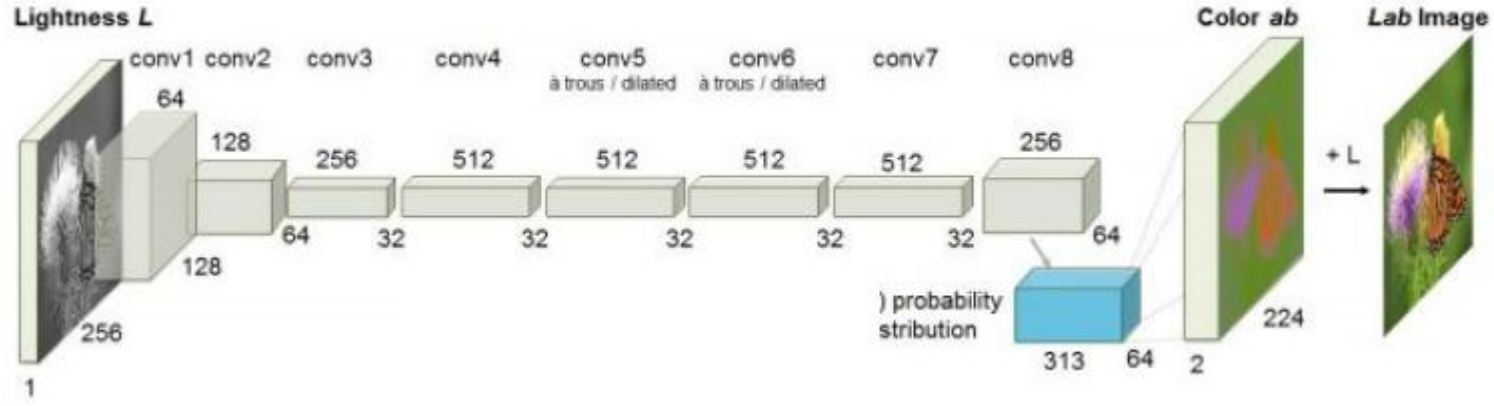
应用

Deepface++,

Imagenet Large scale visual patte
(ILSVPC) 大赛

Alpha go, zero,

简单的例子　灰度图像的彩色化



输入

彩色化

原图像

灰度图像的彩色化

Rechard Zhang 2017

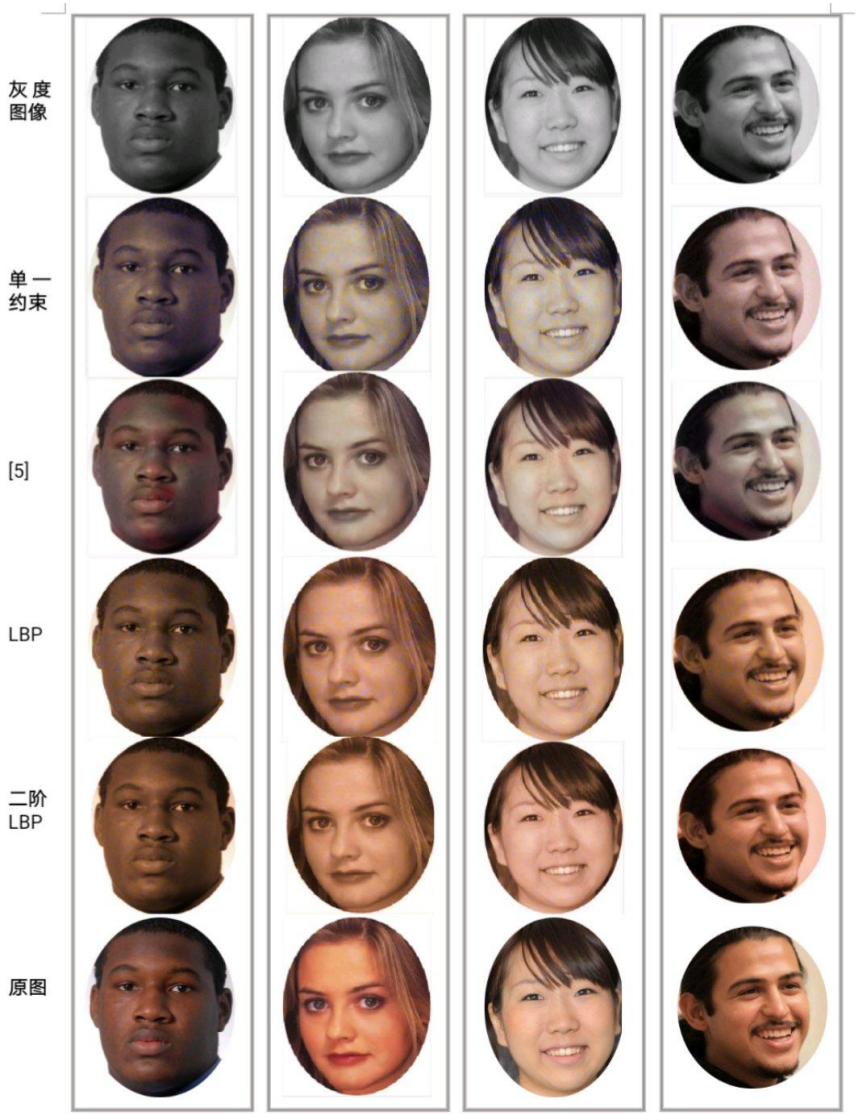
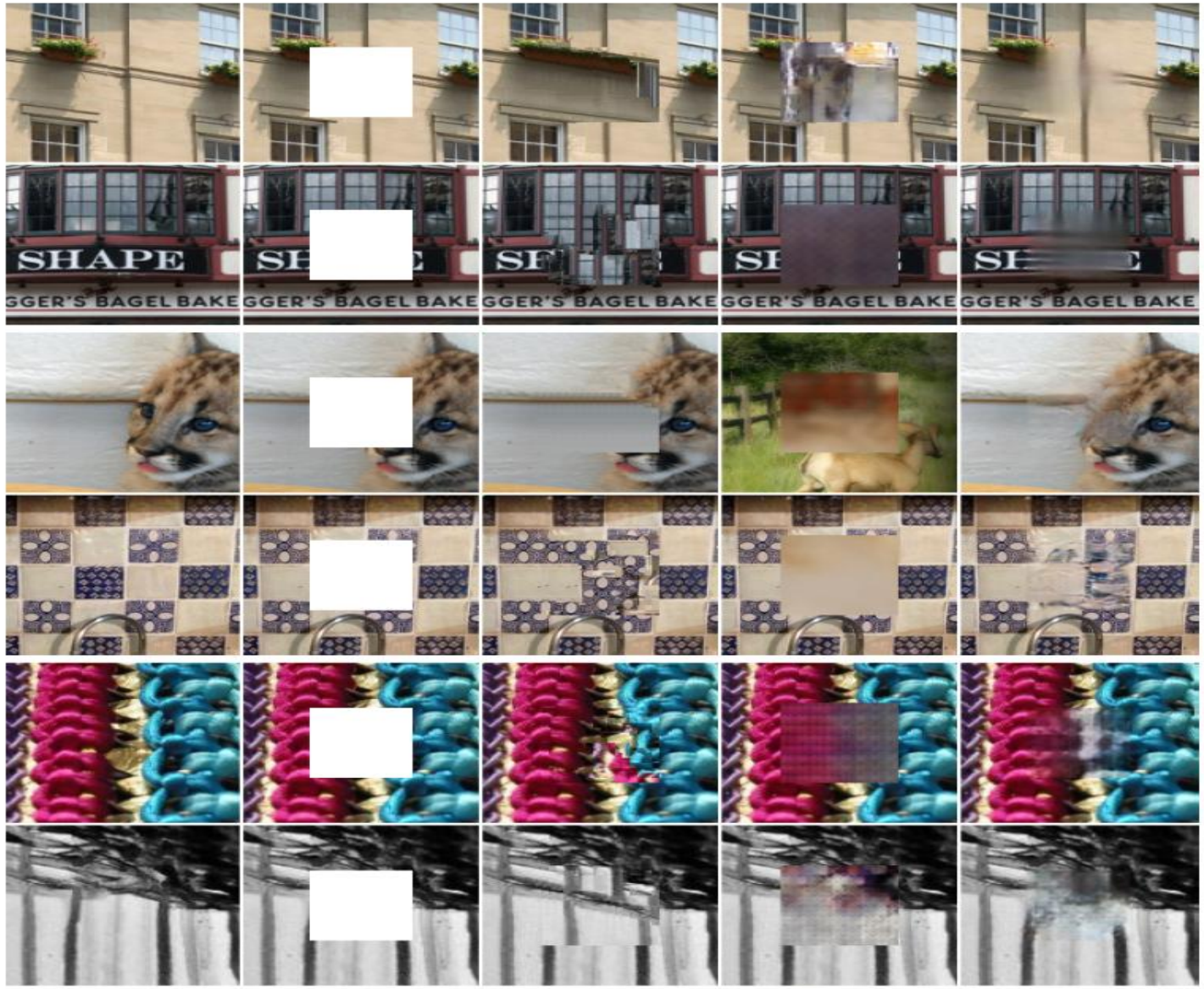


图 27. 本文将实验结果作用于一个人脸库中，注意本实验考虑到了不同人像，人种的多样性。从不同种族挑选了典型的人像进行比较，（包括亚裔，非裔，拉丁裔，美裔）实验顺序如前文一样，从上往下，从灰度到[5]到一阶二阶在到原图。



Original image | input context | Criminisi | Semantic inpainting | Our method

Captioning (字幕)

Retr.
1. Top view of the lights of a city at night, with a well-illuminated square in front of a church in the foreground;
2. People on the stairs in front of an illuminated cathedral with two towers at night;

Gen.
A square with burning street lamps and a street in the foreground;

Retr.
1. Tourists are sitting at a long table with beer bottles on it in a rather dark restaurant and are raising their bierglaeser;
2. Tourists are sitting at a long table with a white table-cloth in a somewhat dark restaurant;

Gen.
Tourists are sitting at a long table with a white table cloth and are eating;

Retr.
1. A dry landscape with light brown grass and green shrubs and trees in the foreground and large reddish-brown rocks and a blue sky in the background;
2. A few bushes at the bottom and a clear sky in the background;

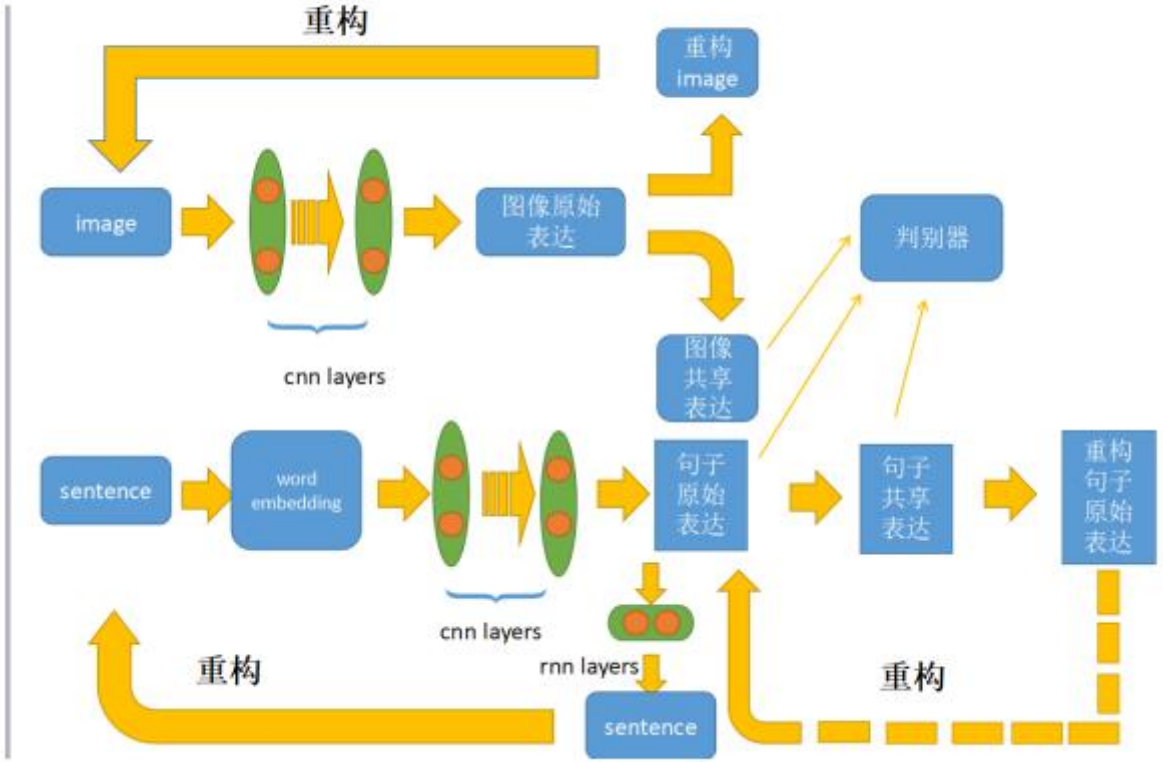
Gen.
A dry landscape with green trees and bushes and light brown grass in the foreground and reddish-brown round rock domes and a blue sky in the background;

Retr.
1. Group picture of nine tourists and one local on a grey rock with a lake in the background;
2. Five people are standing and four are squatting on a brown rock in the foreground;

Gen.
A blue sky in the background;

Figure 1: Examples of the generated and two top-ranked retrieved sentences given the query image from IAPR TC-12 dataset. The sentences can well describe the content of the images. We show a failure case in the fourth image, where the model mistakenly treats the lake as the sky and misses all the people. More examples from the MS COCO dataset can be found on the project page: `www.stat.ucla.edu/~junhua.mao/m-RNN.html`.

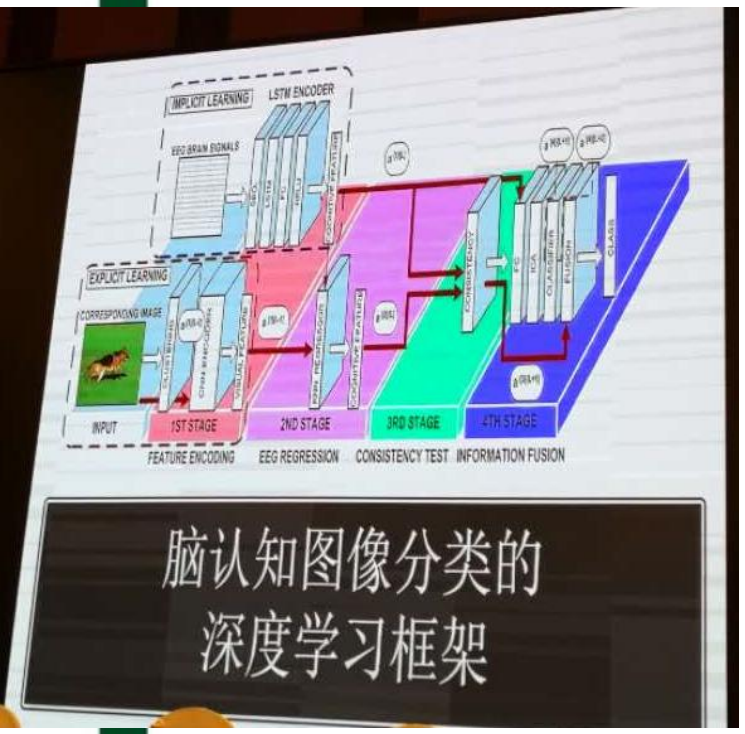
Image to text:

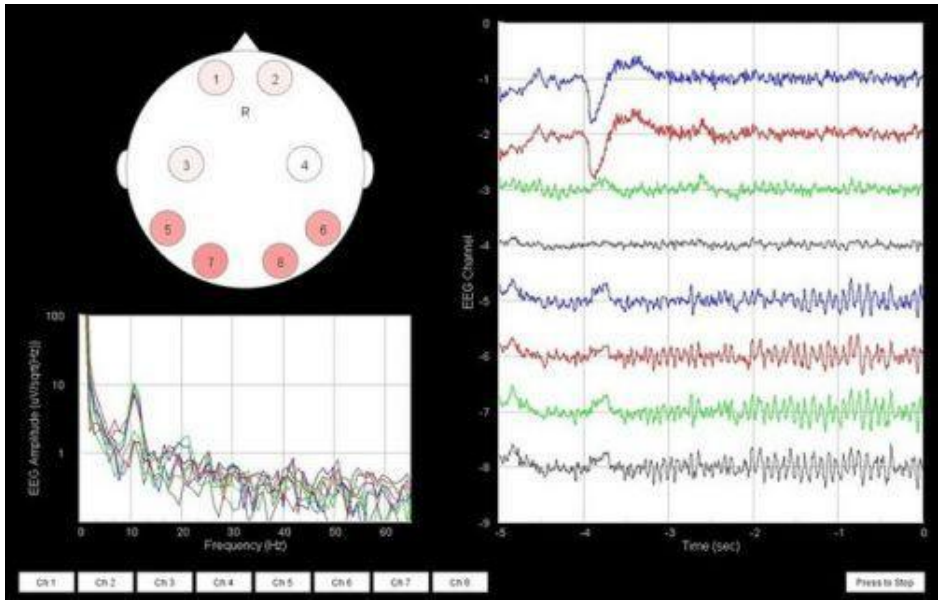


Real: a small zebra is grazing on some grass on a sunny day

Generated : a zebra walking across the grass in a field

读出你的心中所想



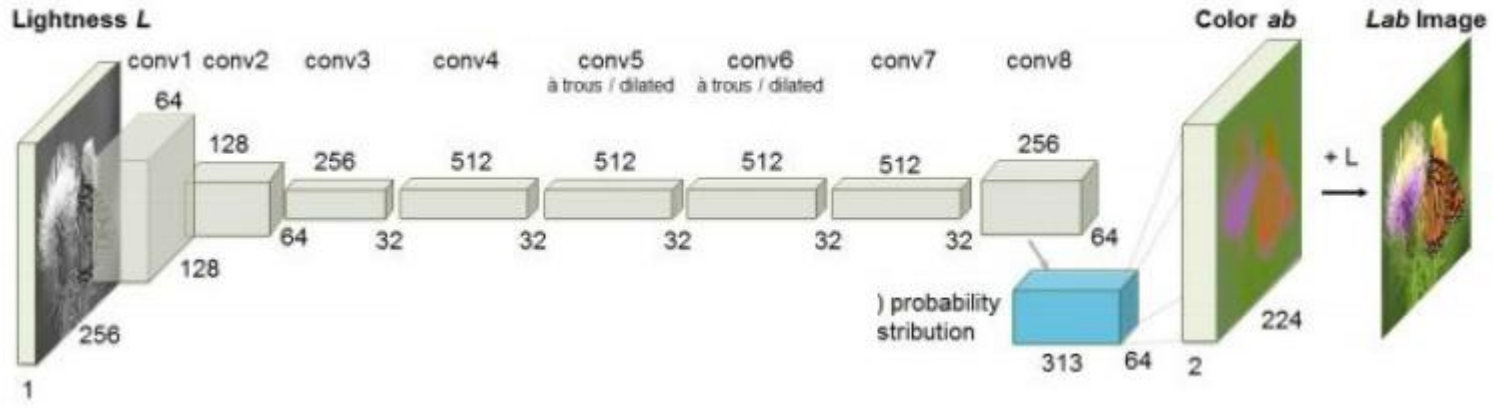


物体的类别：　鱼，猫，花　．．．

Spampinata et al 2017 EEG的图像分类 82.9% 的识别率 （40类）

f(脑电波)→所思
f(场景)→描述(诗，散文，小说，绘画…)
f(症状) →疾病
f(历史数据、因子) → 趋势
⋮

F



优势：海量样本的可以学习用分析数学形式难以描述的结果

不足： 外延性?

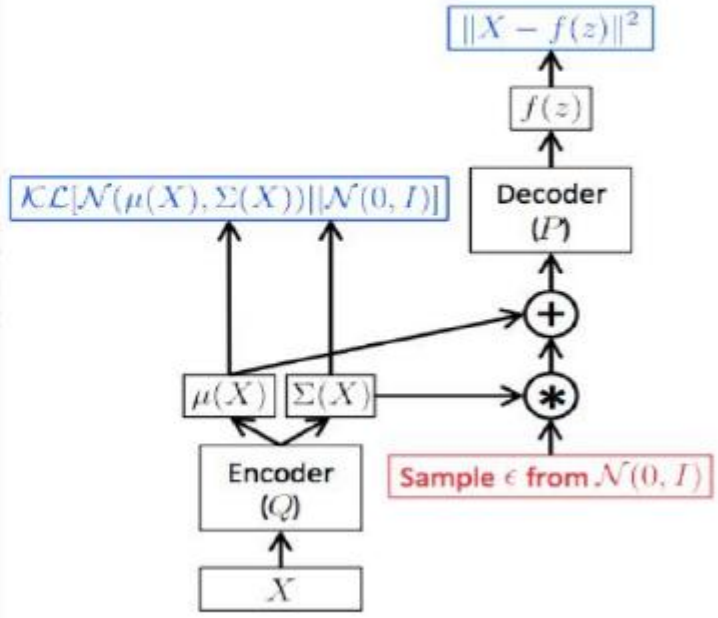
- 机器学习与神经网络
- 深度学习
- 应用案例
- 问题思考

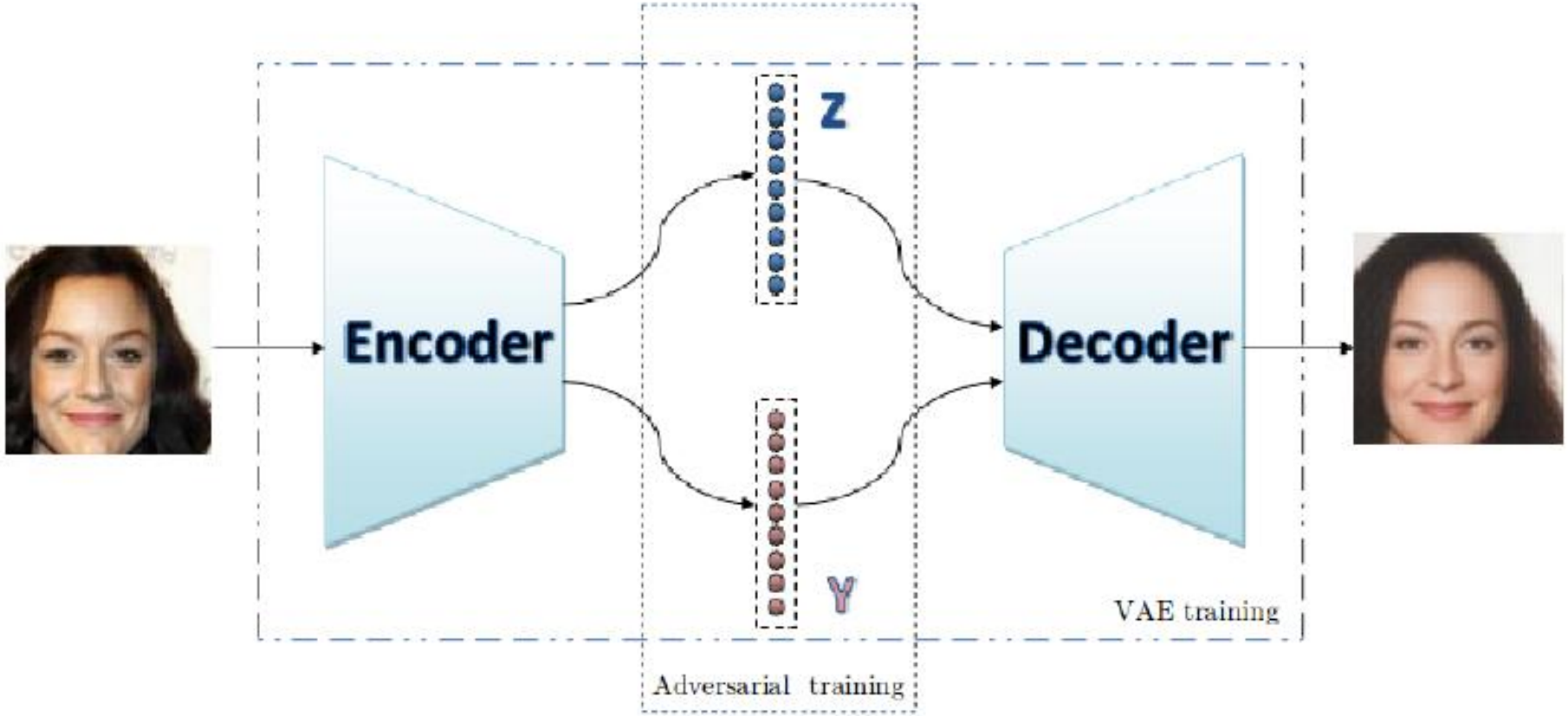
Face editing（人脸像编辑）

- Disentangled representation learning by latent space adversarial variational autoencoders (Defang Li)

Variational Autoencoders (VAE) [1]

- Based on Maximum likelihood (ML) learning
- The encoder $Enc(x) = q_\phi(z|x)$.
- The decoder $Dec(z) = p_\theta(x|z)$.
- $p(z) = N(0, I)$
- Reparametrization trick




z
Encoder
Decoder
y
VAE training
Adversarial training

目标函数

VAE training:

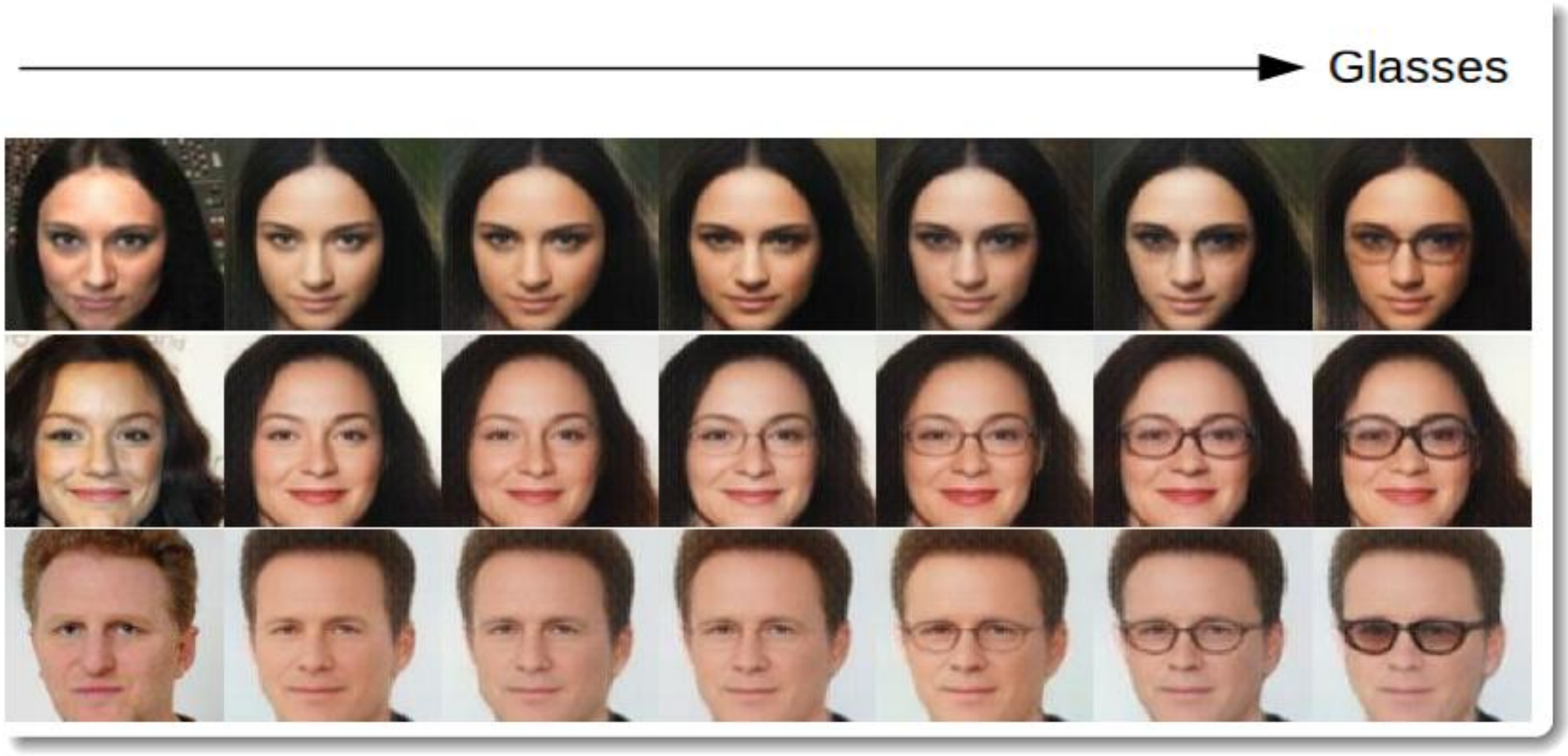
$$(1)\ \max_{\theta,\phi,\psi} -\sum_{i=1}^{L}\alpha_i KL(q_\psi(y_i|x)\parallel p(y_i)) - \beta KL(q_\phi(z|x)\parallel p(z)) + \gamma E_{q_\psi(y|x),q_\phi(z|x)} log\ p(x|y,z)$$

Adversarial training:

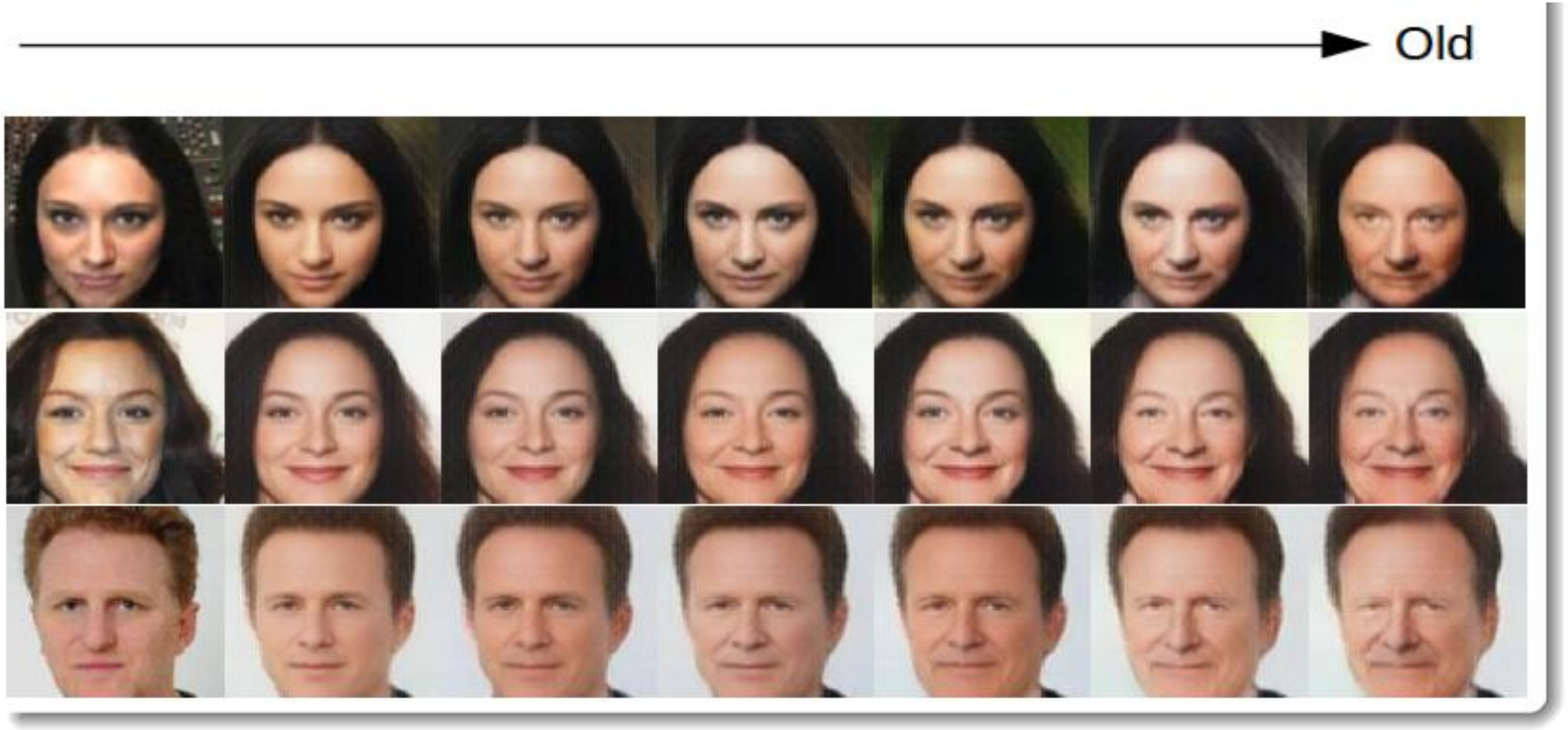
$$(2)\ \min_{\phi,\psi}\sum_{i=1}^{L} KL(q_\psi(y_i|x)\parallel p(y_i)) + KL(q_\phi(z|x)\parallel p(z)) + max(0, m - E_{\tilde{y}\sim p(y),\tilde{z}\sim p(z)}(KL(q_\phi(z|p_\theta(x|\tilde{y},\tilde{z}))\parallel p(z))))$$

$$(3)\ \min_{\theta} E_{\tilde{y}\sim p(y),\tilde{z}\sim p(z)}(KL(q_\phi(z|p_\theta(x|\tilde{y},\tilde{z}))\parallel p(z))) + \sum_{i=1}^{L} E_{\tilde{y}\sim p(y),\tilde{z}\sim p(z)}(KL(q_\psi(y_i|p_\theta(x|\tilde{y},\tilde{z}))\parallel p(y_i)))$$

实验结果



戴眼镜

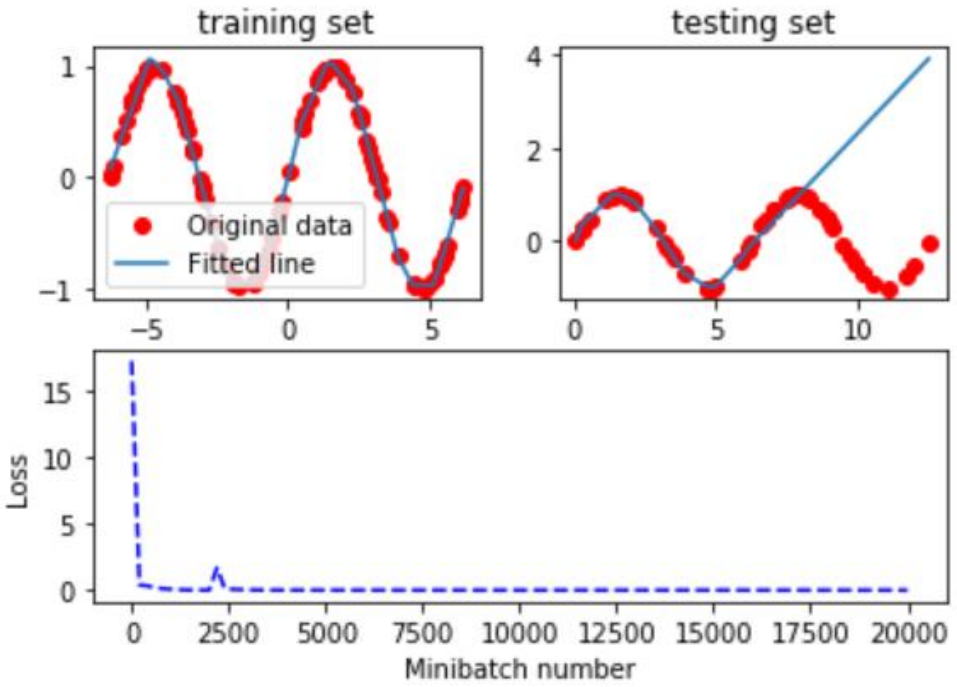


变老

- 机器学习与神经网络
- 深度学习
- 应用案例
- 问题思考

神经网络学习： y=sin(x),

采用的BP(back propagation)神经网络，有两个隐藏层,激活函数是relu。 100 个采样点.
用训练数据[-2 π, 2 π] , 得到： y=f(x,θ) ,其中 θ 为网络的参数.
预测 [0 π, 4π] 测试数据是从给定区间上随机采样得到的。



目前，解决大数据问题有一种倾向：

神经网络，特别是用深度学习。

但是深度学习自身的基础？还有些困境

挑战 1

1. 训练样本大(data hungry) BPL(B M Lake,2015.12 science)
   样本增加技术，预处理（特征提取）

   医学数据?

挑战2

1. 有效的结构、网络节点，层，参数的意义

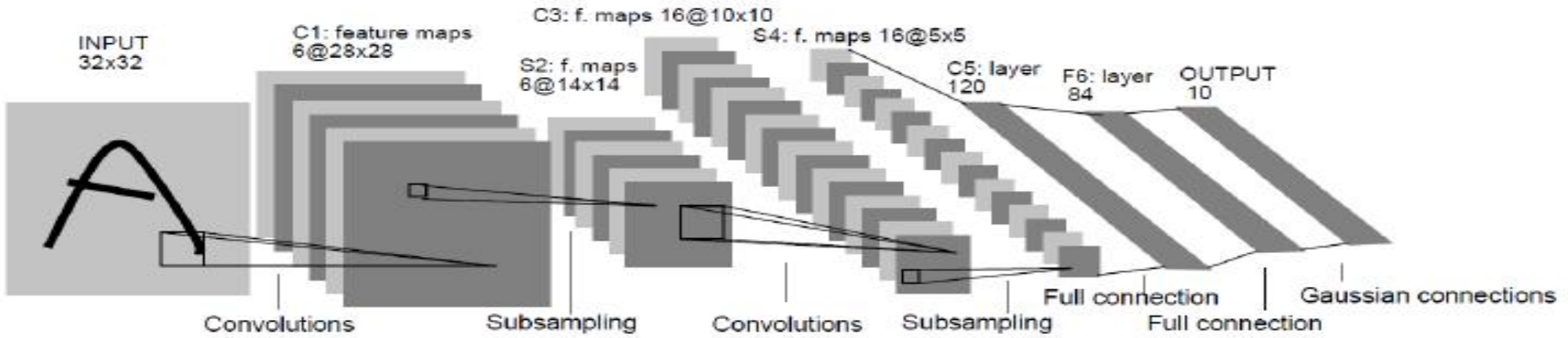


图 4. 卷积神经网络结构。

利用深层网络结构（多达152层）[5]在数据集ImageNet上的图片识别准确率达到了95.06%，已经超越了人眼识别的准确率（94.9%）

挑战3：误差域

对抗扰动

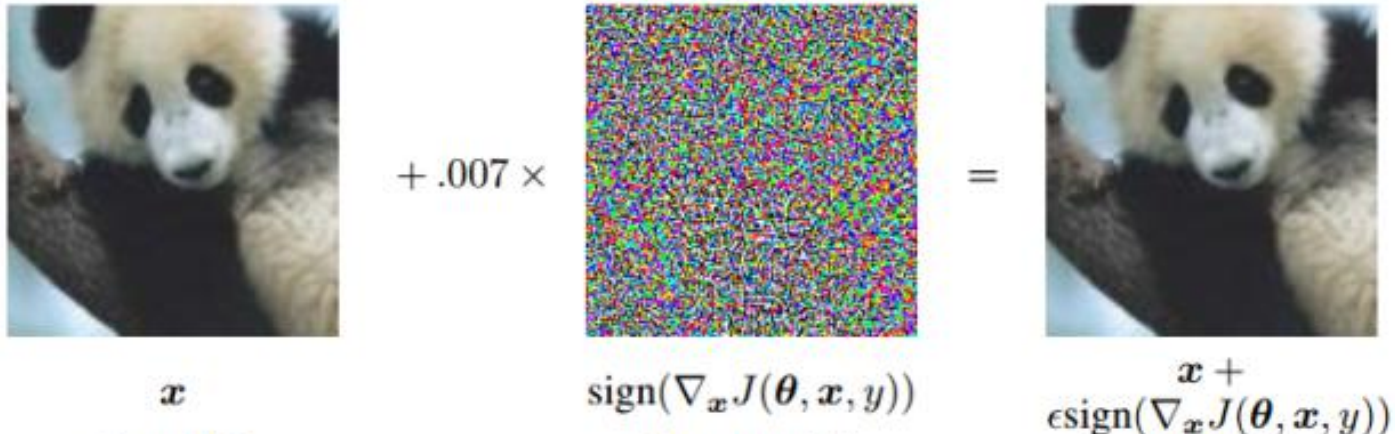
人几乎无法识别的扰动，
深度学习算法却分错。

(a) (b)

图 5. 受对抗扰动的识别。

建立揭示问题机理的数学模型是解决问题的关键

原图　　随机噪音　　攻击样本



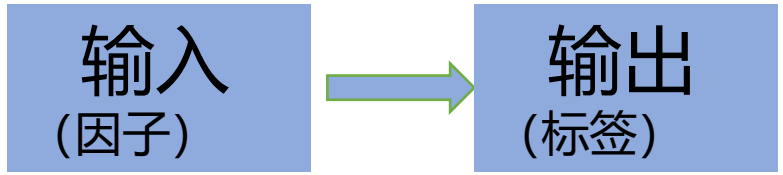
噪音样本训练，　收敛域的估计

7 如何用机器学习探讨金融建模

- 机理模型的探讨
- 经验模型（机器学习-统计学习）

如何运用：

1. 清晰的目的： 解决金融领域某类问题（问题驱动：高频，中线，选股，择时，对冲...）从历史数据中，找规律，建立高收益，低风险的收益。
2. 建模要有逻辑。 （合理，影响的因子）
3. 建立样本集。
4. 建立学模型（结构），学习系统的参数
5. 验证策略的可行性（收益，风险，边界等）
6. 试盘。



注意事项： 风险意识。逻辑，学习适用的场景等

金融数学建模

特点

1. 金融背景，逻辑
2. 数学模型
3. 评价（收益，风险)
4. 回测
5. 实测
6. 分析评估

因子名称	因子类型	IC 均值
股息率	估值	3.34%
营业利润/营业总收入	盈利能力	2.05%
有形资产/带息债务	偿还能力	2.03%
净利润/营业总收入	盈利能力	1.98%
销售净利率	盈利能力	1.96%
净资产收益率 ROE	盈利能力	1.94%
	……	
市销率	估值	-0.70%
市销率 TTM	估值	-0.75%
销售成本率	盈利能力	-1.11%
市盈率 TTM	估值	-1.34%
市盈率	估值	-1.40%
营业总成本/营业总收入	盈利能力	-1.71%
市净率	估值	-2.53%

表 4 因子 IC 显著性统计表

(3)因子策略效果检验

对于 IC 显著性大的因子还需要进一步考察其选股能力的实际表现，即观察因子值高的股票是否能够保持盈利。为此建立了双因子选股模型，分别权重为 50%，并对不同的双因子组合从 2011 年至 2020 年进行回测分析，得到如下图表（其中-1 代表反方向）：

模型组合	最大回撤	夏普比
股息率、市净率(-1)	35.17%	0.79
股息率、营业利润/营业总收入	32.06%	0.47
股息率、营业总成本/营业总收入(-1)	37.06%	0.59
市净率(-1)、营业利润/营业总收入	33.4%	0.31
市净率(-1)、营业总成本/营业总收入(-1)	37.26%	0.24
营业利润/营业总收入、营业总成本/营业总收入(-1)	37.45%	0.14

2020 年“大湾区杯”粤港澳金融数学建模竞赛题目

A 题　大湾区指数增强策略

指数增强策略采用量化增强模型，追求高于标的指数回报水平的投资，同时力求进行有效的风险控制、降低交易成本、优化投资组合。指数增强策略不会对跟踪标的成份股进行完全复制，而是会对部分看好的股票增加权重，不看好的股票则减少权重，甚至完全去掉。通过对交易成本模型的不断监测，尽可能让交易成本降到最小。综合来看，就是既做到超额收益，又控制主动风险。

某证券公司选取三十支大湾区金股(表 1)构建了大湾区指数(399999)，指数行情如图 1 所示，附录一为成份股的行情数据，附录二为大湾区指数行情数据。

表 1　大湾区指数成份股

公司名称	股票代码	公司名称	股票代码	公司名称	股票代码
分众传媒	002027	华侨城 A	000069	塔牌集团	002233
亿纬锂能	300014	金地集团	600383	粤水电	002060
立讯精密	002475	保利地产	600048	顺丰控股	002352
风华高科	000636	招商积余	001914	中顺洁柔	002511
国星光电	002449	中国平安	601318	美盈森	002303

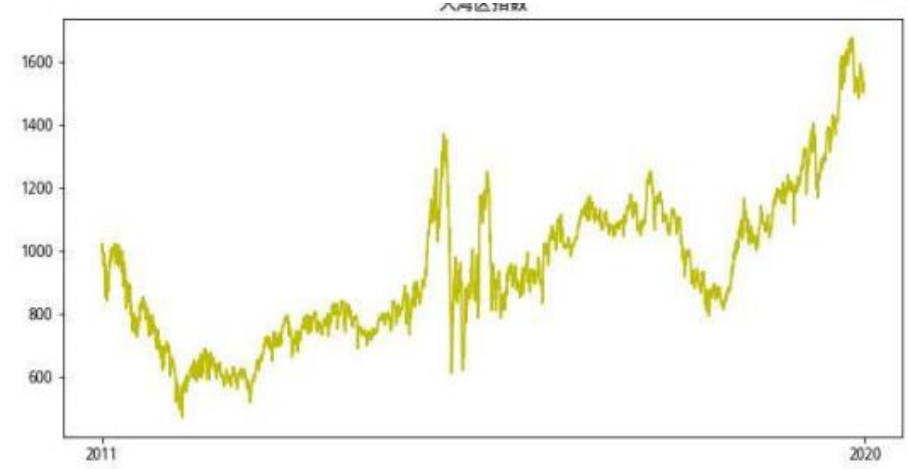


图 1　大湾区指数行情实况(2011-2020)

目前该公司采用的大湾区指数增强策略为：在 30 支成份股中选取上一周最强势的 10 支股票，作为下一周投资标的，每周星期一换仓，以当天的收盘价为准考虑收益，每支股票投资额固定为本金的 10%，手续费万分之 2.5。

1）请计算该投资策略从 2011-2020 年的收益率曲线，并且参照大湾区指数评价其投资成效。

2）如果调仓时间不变，但单支股票的投资额可以灵活变化，请找出该投资策略最优的收益率曲线。

3）调仓时间不变，单支股票的投资额为本金的 10%，根据市场交易数据建模，设计你自己的选股策略。

优秀论文https://www.tipdm.org/wq2jszx/1821.jhtml

2020年“大湾区杯”粤港澳金融数学建模竞赛题目

A题-大湾区指数增强策略（点击下载）

B题-知识产权证券化（点击下载）

2020年“大湾区杯”粤港澳金融数学建模竞赛优秀论文（选）

1. 13095A（点击下载）
2. 13353A（点击下载）
3. 13445A（点击下载）
4. 12634B（点击下载）

Markowitz 模型解决的是不同资产间投资权重配置的问题：给定某个收益水平，求出风险最小的投资组合；或给定某个风险水平，求出收益最高的投资组合。其基本思想是：如果将资产的收益率视为一个数学上的随机变量，那么资产的期望收益是该随机变量的数学期望（均值），而投资风险可以用该随机变量的方差来表示。

设N为资产的个数，w_i为第 i 个资产的权重，E_i为第 i 个资产的期望收益率，σ_i为第 i 个资产收益率的标准差，σ_{ij}为第 i 个资产收益率与第 j 个资产收益率的协方差。那么资产组合的收益率和方差满足如下公式：

$$\sigma^2 = \sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{i \neq j}^{N} w_i w_j \sigma_{ij} \tag{3}$$

$$E_p = \sum_{i=1}^{N} w_i E_i \tag{4}$$

那么，根据 Markowitz 模型：给定资产组合的目标收益率，可以通过最小化方差来求出资产配置 w；给定资产组合的目标方差，可以通过最大化收益来求出资产配置 w。模型的数学表示如下：

$$\min \quad \sigma^2 \tag{5}$$

$$\text{s. t.} \quad \sum_{i=1}^{N} w_i = 1 \tag{6}$$

$$E_p = R_p$$

$$w_i \geq 0 \qquad i = 1 \sim N \tag{7}$$

或者

$$\max \quad E_p \tag{8}$$

$$\text{s. t.} \quad \sum_{i=1}^{N} w_i = 1 \tag{9}$$

$$\sigma^2 = \sigma_p^2$$

$$w_i \geq 0 \qquad i = 1 \sim N \tag{10}$$

其中，R_p为目标收益率，σ_p^2为目标方差。

$$U = R_p - \frac{\delta}{2}\sigma^2 \tag{11}$$

其中，δ为风险厌恶系数，用来权衡风险与收益之间的关系，其具体大小取决于投资者自身的风险厌恶程度。经资料查阅，业界与学界一般将这一风险厌恶系数设为 2.5，经济学家称之为标准风险价格。

那么，以上两个模型等价于如下模型：

$$\max \quad U = R_p - \frac{\delta}{2}\sigma^2 \tag{12}$$

$$\text{s. t.} \quad \sum_{i=1}^{N} w_i = 1 \tag{13}$$

$$w_i \geq 0 \qquad i = 1 \sim N \tag{14}$$

基于机器学习的金融量化建模

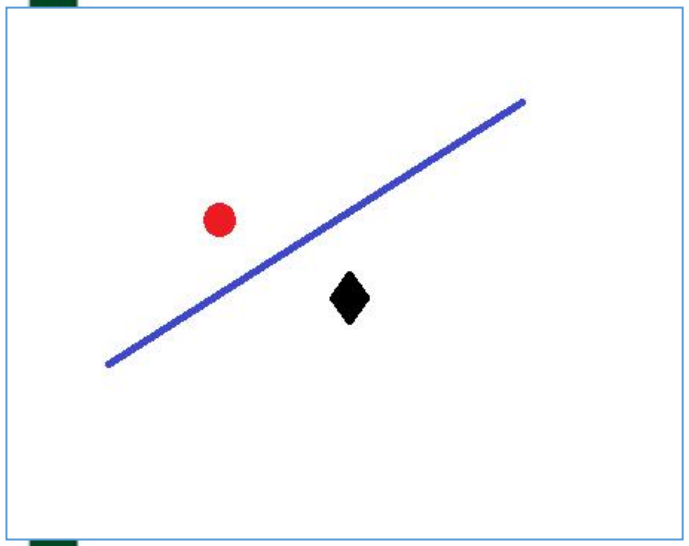
分析确定问题的目标 （如：价格预测，趋势预测 ...

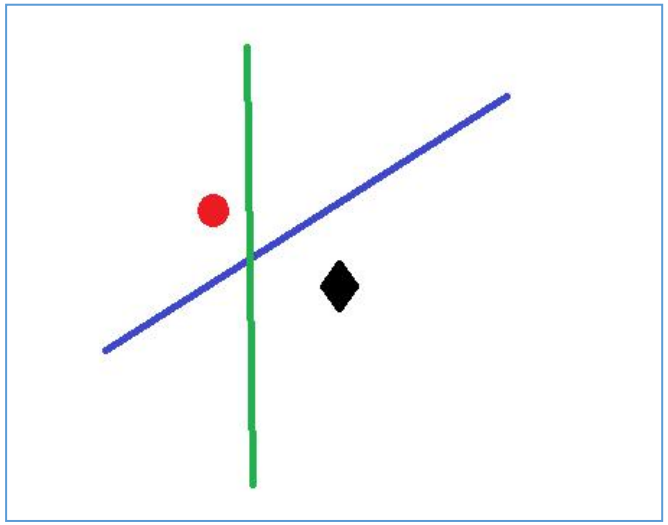
1. 数据备制 （选取合理的样本， 均衡性，...)
   - （Xi, Yi) ,标签的有序性， 类别标签

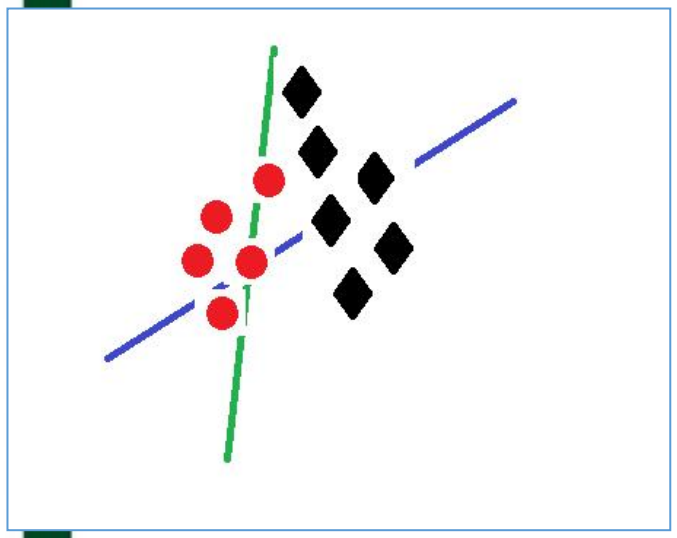
   考虑的问题：
   - x->y （x y是否具有关联, 数据中是否包含y 的因子）
   - 样本的充分性（样本量），是否需要增强，清洗
   - 样本的均衡性

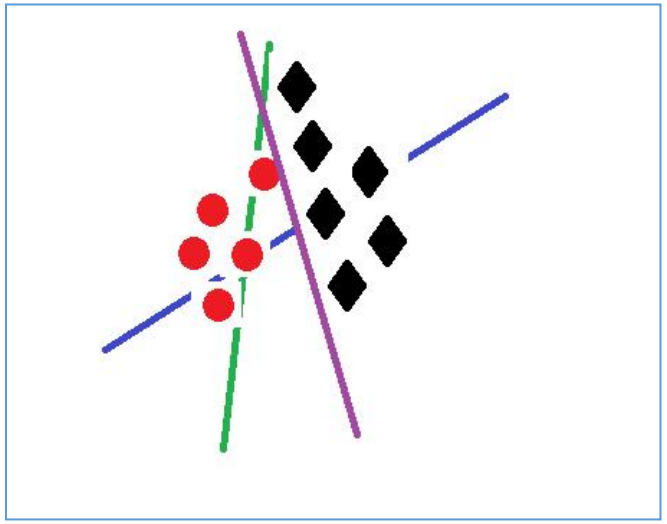
2. 网络结构 选择 （CNN , RNN, UNet,RNet
3. 实现
4. 评价
   - 建模的**合理性**
   - 有效性： 专业：（F1， 召回，准去率），： 领域：客观评价：收益 （收益率），风险（回撤） ， 鲁棒性（时候的条件）
   - 实测

数据的位数与样本数



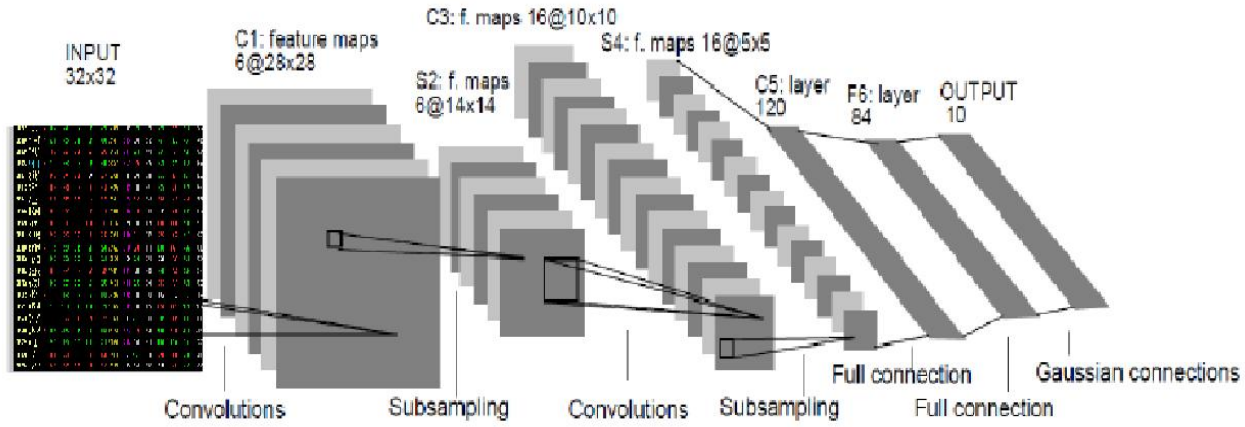






A 题

- （xi, yi)
- xi 上周的数据， yi 本周 最好 (何为最好?) k (k=10?) 只股票
- 构建神经网络
  （DCNN， ResNet,
  LSTM, Unet,...)
- 训练，分析测试
- 评价



判别天

涨幅

≥1%，正类

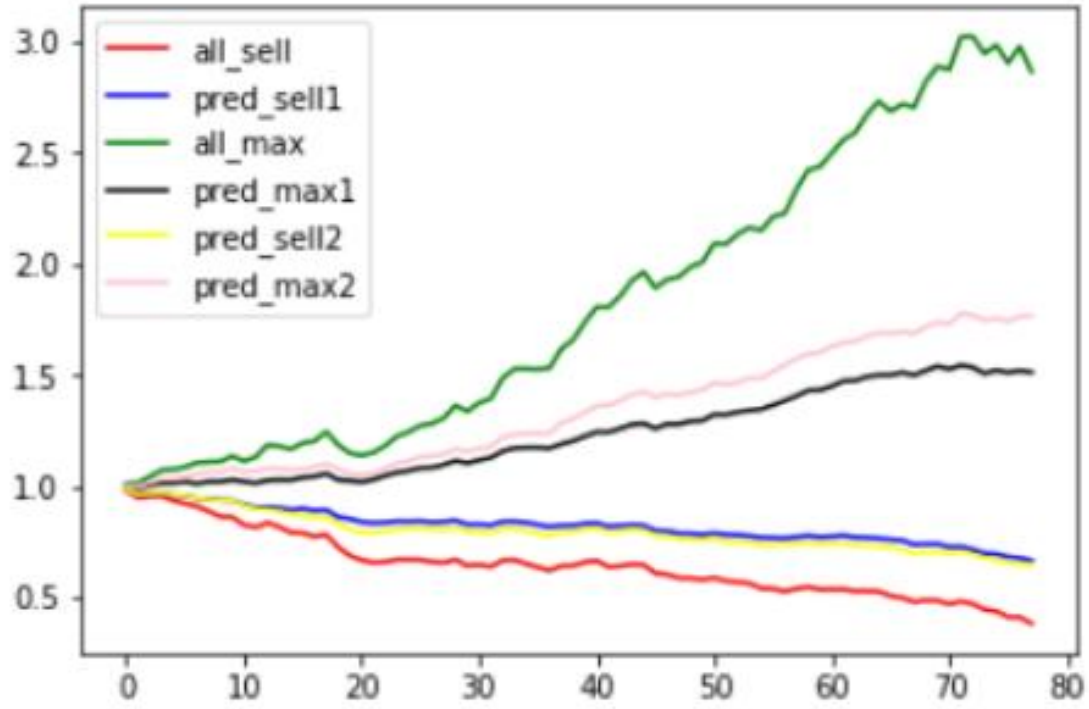
其他，零类

≤-1%，负类

标记天

有效样本

盈利曲线图：


all_sell
pred_sell1
all_max
pred_max1
pred_sell2
pred_max2
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
60
70
80

案例：基于嵌套LSTM的股票交易策略研究

LSTM 模型

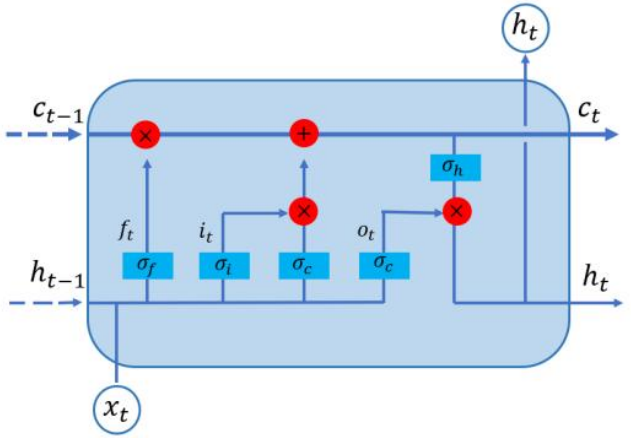


图 2-2　LSTM 网络结构示意图

（图中x_t代表 t时刻的输入向量，h_t代表 t时刻的隐层状态值，c_t代表 t时刻的记忆单元，i_t，f_t，o_t分别为输入门、遗忘门和输出门，σ为各项激活函数）

在我们的输入序列中，并非每个时刻的数据都能对模型的预测提供帮助，这时就需要通过输入门来对输入数据做一个筛选工作。而保存在记忆细胞中的历史信息也不一定都与当前的预测相关，因此需要通过遗忘门来选择是否保留或者遗忘。同时，当前记忆是否有必要输出也将由输出门来决定。这三个门的更新将由当前时刻的输入x_t，以及上一时刻的隐层输出h_{t-1}，以及激活函数共同决定：

$$input\ gate:\ i_t = \sigma_i(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \tag{2.18}$$

$$forget\ gate:\ f_t = \sigma_f(x_t W_{xf} + h_{t-1} W_{hf} + b_f) \tag{2.19}$$

$$output\ gate:\ o_t = \sigma_o(x_t W_{xo} + h_{t-1} W_{ho} + b_o) \tag{2.20}$$

Stacked LSTM 模型

络（Stacked LSTMs，简称 SLSTM）。在堆叠式 LSTM 中，上一层 LSTM 的输出将作为下一层的输入。

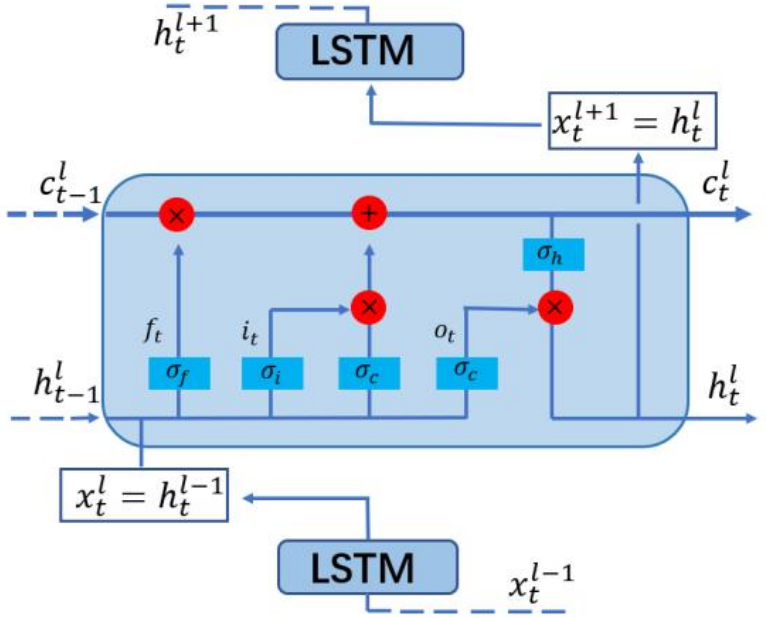


图 2-3 堆叠 LSTM 结构示意图

结构选择

与全连接层；以下将详细介绍该模型。

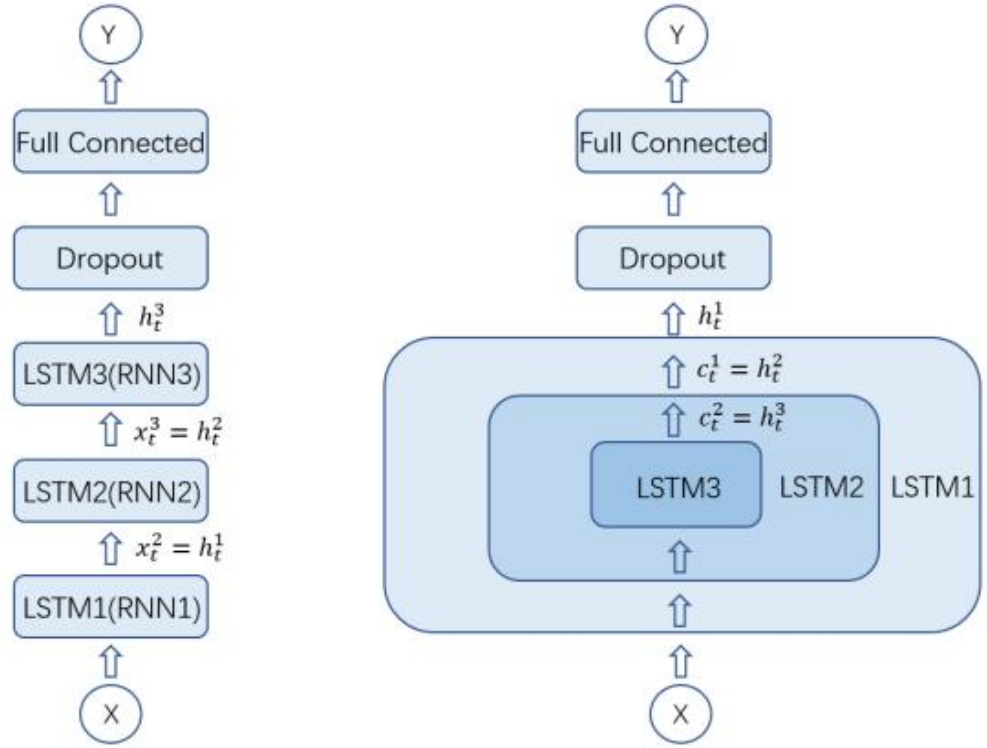


图 3-1 基于 NLSTM 建立的股票预测模型（右）与对比模型（左）

1)股票选取

表 3-1 实验股票名称与代码

名称	西部建设	雅化集团	特力 A	安妮股份	捷顺科技
代码	sz002302	sz002497	sz000025	sz002235	sz002609
名称	恒通科技	上海新阳	华力创通	太空智造	正业科技
代码	sz300374	sz300236	sz300045	sz300344	sz300410
名称	天奇股份	金轮股份	长青股份	远望谷	
代码	sz002009	sz002722	sz002391	sz002161	

2)数据选取

表 3 2 sz002497 交易日 29/12/2017 开盘第一分钟的原始分笔交易数据

日期	时间	开盘	价格	成交量	买单量	卖单量
12/29/2017	9:30:00	13.5	13.5	54	97	582
12/29/2017	9:30:03	13.5	13.5	914	828	1002
12/29/2017	9:30:06	13.5	13.47	67	866	1492
12/29/2017	9:30:09	13.5	13.5	54	864	1485
12/29/2017	9:30:15	13.5	13.47	18	1085	1233
…	…	…	…	…	…	…
12/29/2017	9:30:39	13.5	13.44	21	1164	1295
12/29/2017	9:30:42	13.5	13.49	240	1391	1194
12/29/2017	9:30:48	13.5	13.49	7	957	1224
12/29/2017	9:30:54	13.5	13.49	81	1642	1434
12/29/2017	9:30:57	13.5	13.49	127	1320	1434

3)根据问题给标签(x_i->y_i)

预处理：清洗，标准化,均衡化降采样，增强，…

数据预处理

如表 3-3 所示，我们的模型输入是一段长度为 100 的序列数据，共有八个特征，即输入维度为 100*8。在数据输入模型之前，我们还需要对其进行归一化处理。在此，我们将根据各指标的不同特点，分成两类来处理。

表 3-3 sz002497 在 2017 年 12 月 26 日的部分交易数据

价格	成交量	买单量	卖单量	开盘价	最大值	最小值	均价
12.69	94	996	283	12.76	12.94	12.53	12.7766
12.75	19	1016	637	12.76	12.94	12.53	12.7766
12.78	70	1146	766	12.76	12.94	12.53	12.7766
12.71	156	953	805	12.76	12.94	12.53	12.7758
12.71	3	1058	1097	12.76	12.94	12.53	12.7757
…	…	…	…	…	…	…	…
12.73	12	877	821	12.76	12.94	12.53	12.7246
12.75	10	868	880	12.76	12.94	12.53	12.7246
12.74	22	879	646	12.76	12.94	12.53	12.7246
12.74	8	874	641	12.76	12.94	12.53	12.7246
12.72	614	658	673	12.76	12.94	12.53	12.7246

$$price_t = 10 * \left(\frac{price_t}{close} - 1\right) \tag{3.3}$$

另外三个与数量相关的特征（分别是成交量、买单量与卖单量）将用最大值来做归一化，这里用$volume$来代表与数量相关的特征，其归一化方式为：

$$\text{volume}_t = \frac{volume_t}{\max(volume)} \tag{3.4}$$

标准化:比如价格的处理，成交量的处理。

结果及评价：精准率，召回率，F1

表 4-2 RNN 模型对 002497 的预测结果（2017 年 7 月）

预测值 / 真实值	买点	不交易	卖点	总计	召回率
买点	452	1976	17	2445	18.49%
不交易	2222	78500	4995	85717	91.58%
卖点	0	490	2071	2561	80.87%
总计	2674	80966	7083	90723	买卖点
精准率	16.90%	96.95%	29.24%	准确率	准确率
F1	17.64%	94.34%	42.92%	89.31%	50.40%

RNN

表 4-3 SLSTM 模型对 002497 的预测结果（2017 年 7 月）

预测值 / 真实值	买入点	不交易	卖出点	总计	召回率
买入点	1883	544	18	2445	77.01%
不交易	10424	60009	15289	85717	70.00%
卖出点	0	282	2299	2561	89.77%
总计	12307	60810	17606	90723	买卖点
精准率	15.30%	98.67%	13.06%	准确率	准确率
F1	25.51%	81.97%	22.73%	70.75%	83.54%

SLSTM

表 4-4 NLSTM 模型对 002497 的预测结果（2017 年 7 月）

预测值 / 真实值	买入点	不交易	卖出点	总计	召回率
买入点	1951	489	5	2445	79.80%
不交易	4856	71324	9537	85717	83.21%
卖出点	0	279	2282	2561	89.11%
总计	6807	72092	11824	90723	买卖点
精准率	28.66%	98.93%	19.30%	准确率	准确率
F1	42.17%	90.42%	31.73%	83.28%	84.56%

NLSTM

领域：收益，风险

模型的总体评价

表 4-5 002497（2017 年 7 月）各模型预测总结

交易点类别	评价指标	RNN	堆叠 LSTM	嵌套 LSTM
买点	精准率	16.90%	15.30%	**28.66%**
	召回率	77.01%	77.01%	**79.80%**
	F1	17.64%	25. 51%	**42.17%**
不交易	精准率	96.95%	98.67%	**98.93%**
	召回率	**91.58%**	70.00%	83.21%
	F1	**94.34%**	81.97%	90.42%
卖点	精准率	**29.24%**	13.06%	19.30%
	召回率	80.87%	**89.77%**	89.11%
	F1	**42.92%**	22.73%	31.73%
准确率		**89.31%**	70.75%	83.28%
买卖点准确率		50.40%	83.54%	**84.56%**

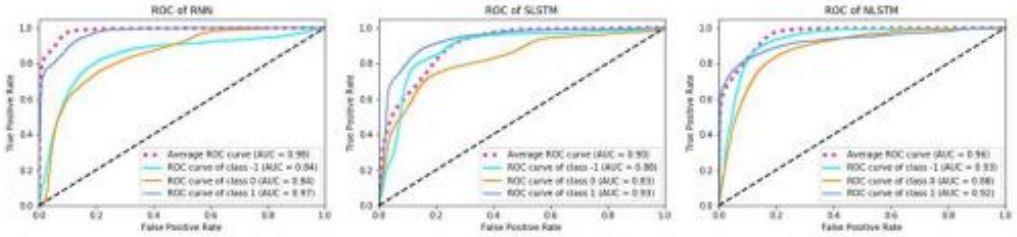


图 4-1 002497 的 ROC 与 AUC 图（2017 年 7 月）

小结

- 机器学习
- 金融量化分析要点
- 实践

中山大學
SUN YAT-SEN UNIVERSITY

谢谢！