AiCon
全球人工智能与机器学习技术大会

深度学习在CTR预估的应用

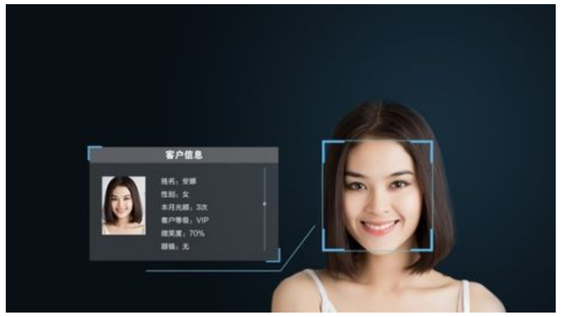
张俊林

新浪微博 AI Lab 资深算法专家

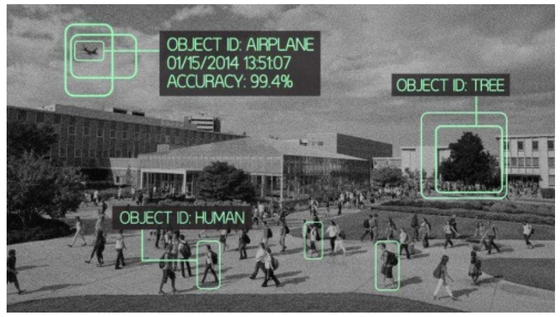
主办方 Geekbang. 极客邦科技 InfoQ

TABLE OF CONTENTES

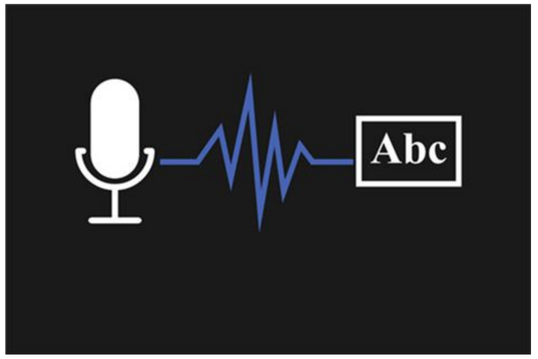
深度学习:各个领域的成功



人脸识别



物体识别



语音识别

机器翻译

风格转换

图片生成

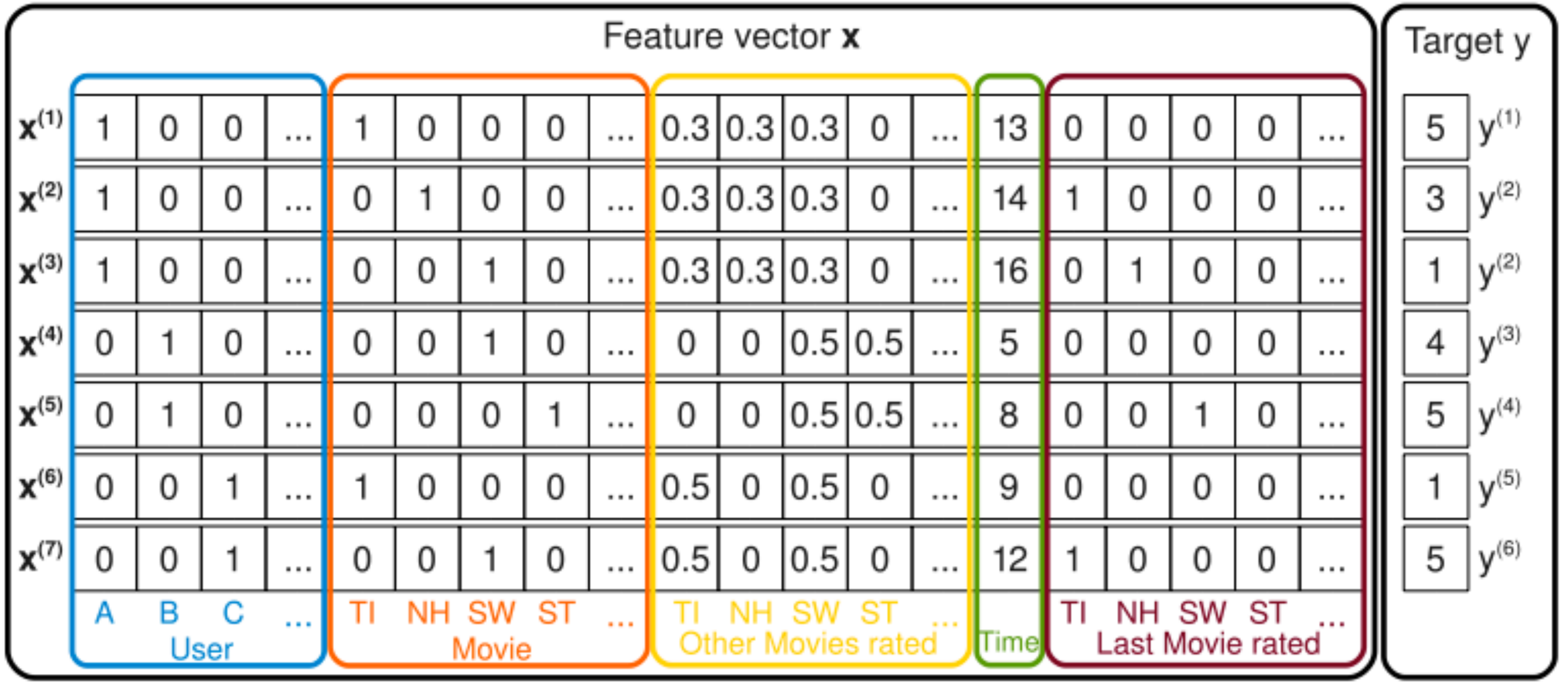
CTR任务的应用

计算广告

推荐系统

信息流排序

CTR任务例子



	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...	Target y
$x^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5 $y^{(1)}$
$x^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3 $y^{(2)}$
$x^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1 $y^{(2)}$
$x^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4 $y^{(3)}$
$x^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5 $y^{(4)}$
$x^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1 $y^{(5)}$
$x^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5 $y^{(6)}$

CTR任务的特点

- 大量离散特征
- 大量高维度稀疏特征
- 特征工程：特征组合对于效果非常关键

当CTR预估遇到深度学习

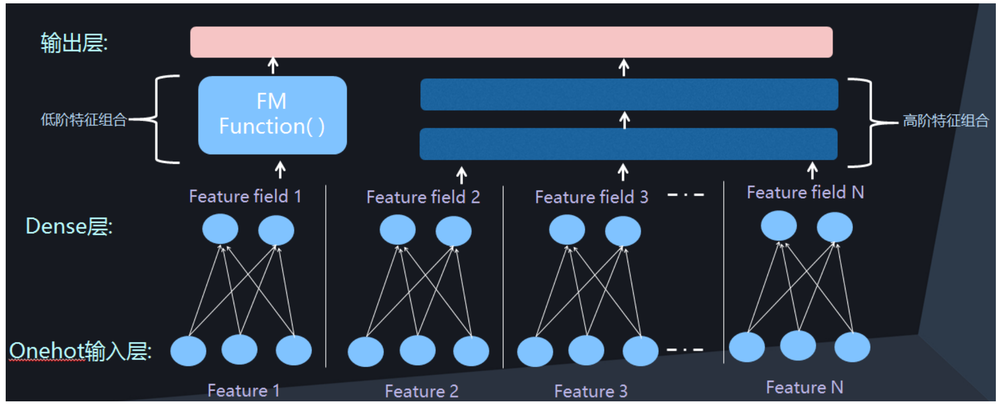
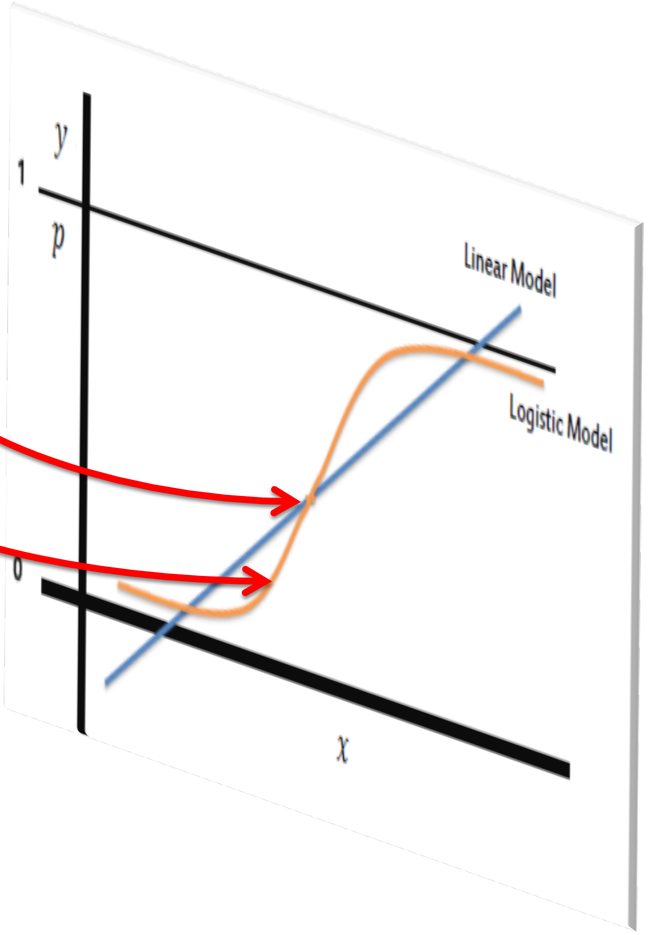


TABLE OF CONTENTES

线性模型：思路及问题

$$\text{Linear: } \hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i$$

$$LR: \hat{y}(x) = \frac{1}{1 + w_0 \exp(-w^T x)}$$



优势：

简单；可解释；易扩展；效率高；易并行

缺点：

难以捕获特征组合

线性模型改进：加入特征组合

改进版本 $\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j} x_i x_j$

两两特征组合

优势：

直接将两两组合特征引入模型

缺点：

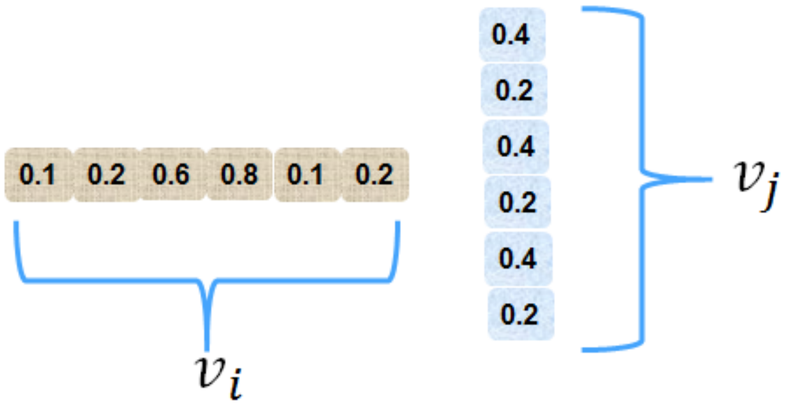
组合特征泛化能力弱

$w_{i,j} = 0 \quad if$ 在训练数据中$x_i x_j = 0$

FM模型

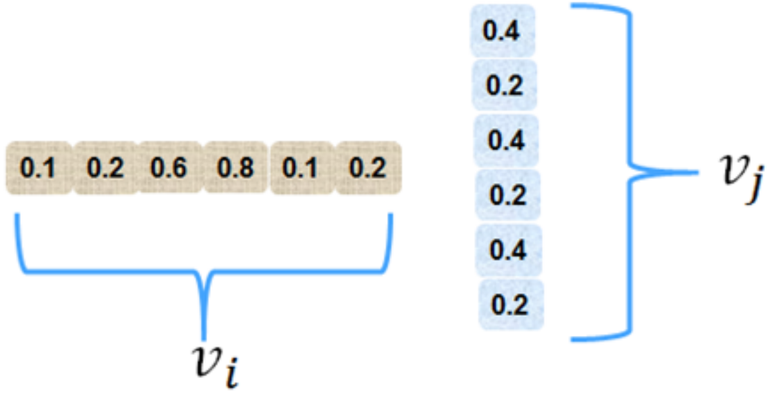
$$\mathrm{FM}: \hat{y}(x) := \underbrace{w_0 + \sum_{i=1}^{n} w_i x_i}_{\text{LR模型}} + \underbrace{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j}_{\text{Dense化两两特征}}$$

$$\langle v_i, v_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f} =$$



FM模型

$$\langle v_i, v_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f} =$$



$$w_{i,j} = \langle v_i, v_j \rangle \neq 0$$

even if 训练数据中$x_i x_j = 0$

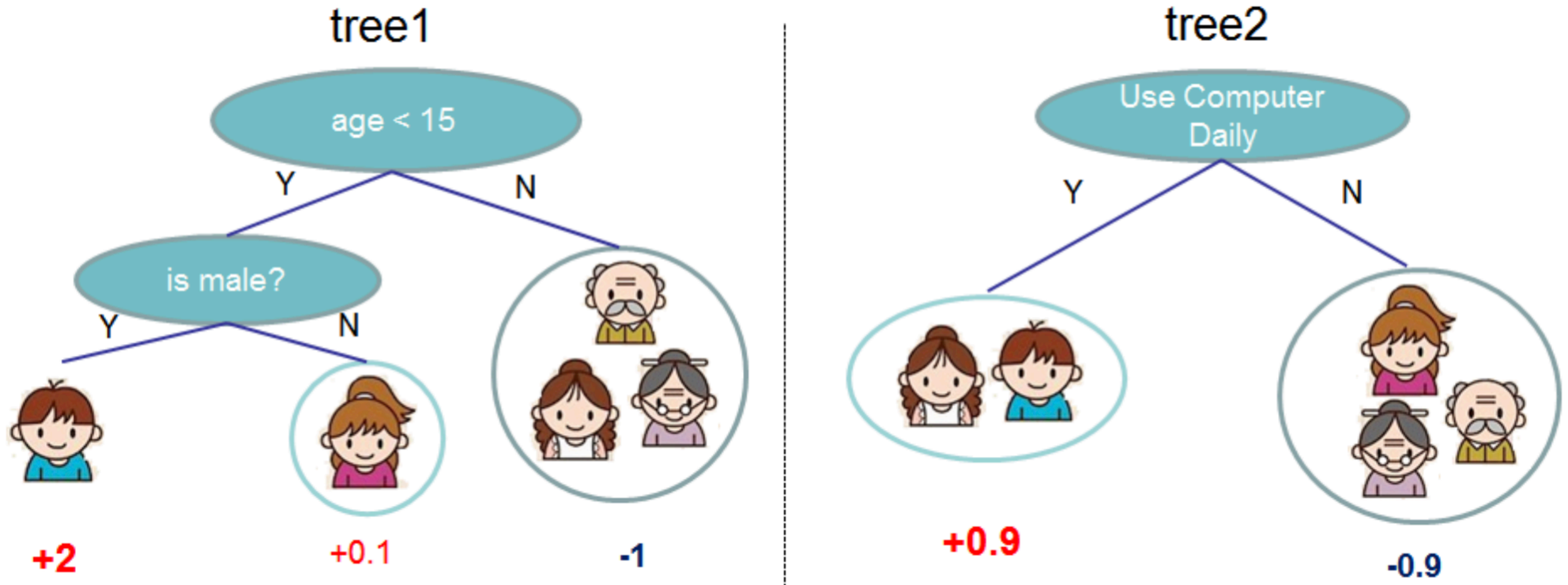
only if 在训练数据中存在k使得$x_i x_k \neq 0$

FM模型泛化能力强

GBDT模型

- Gradient Boosting Decision Tree:迭代的决策树算法，多棵决策树构成，所有子树的决策值累加得出预测值；
- 很多名称，还被称为：MART（Multiple Additive Regression Tree）／GBRT（Gradient Boosting Regression Tree） ／TreeLink等

GBDT模型(例：预测是否喜欢打游戏)

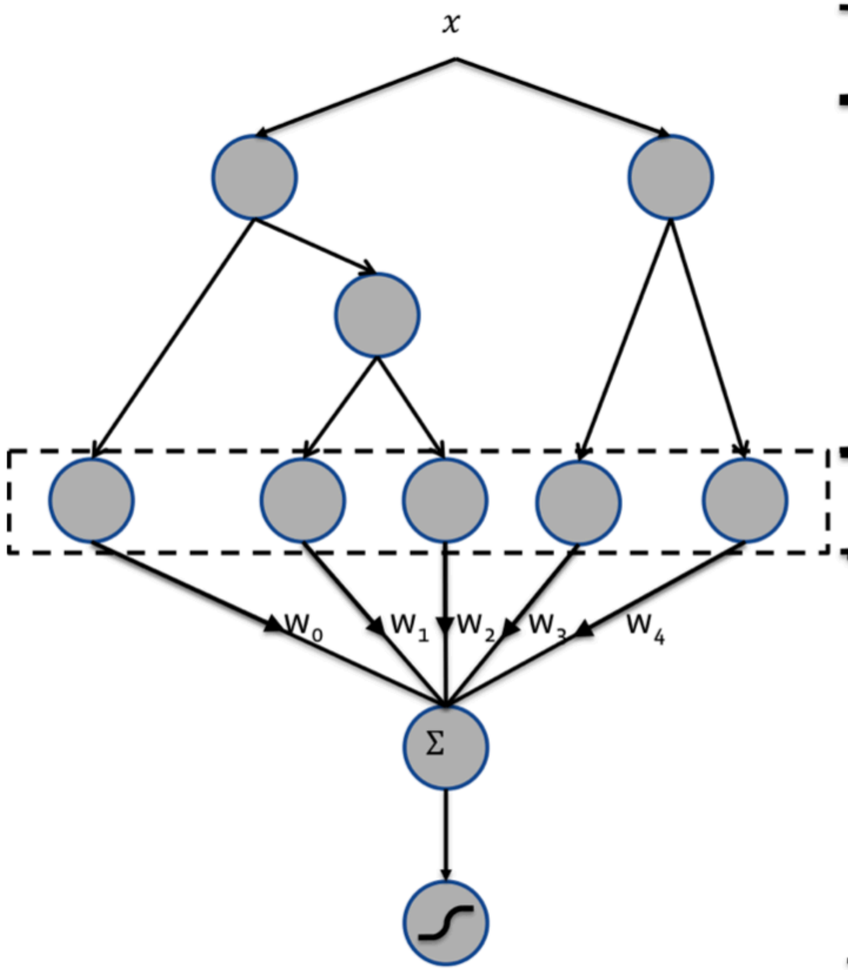


f( ) = 2 + 0.9= 2.9 f( )= -1 - 0.9 = -1.9

LR+GBDT模型

- Facebook首先提出（2014）
- 集成LR和GBDT各自的优势
  - GBDT发现有效的组合特征Feature Set
  - 将Feature Set引入LR模型中
- 目前广泛使用在各大互联网公司线上系统中

LR+GBDT模型



Input Features

Tree Splits → GBDT特征选择

Transformed Features

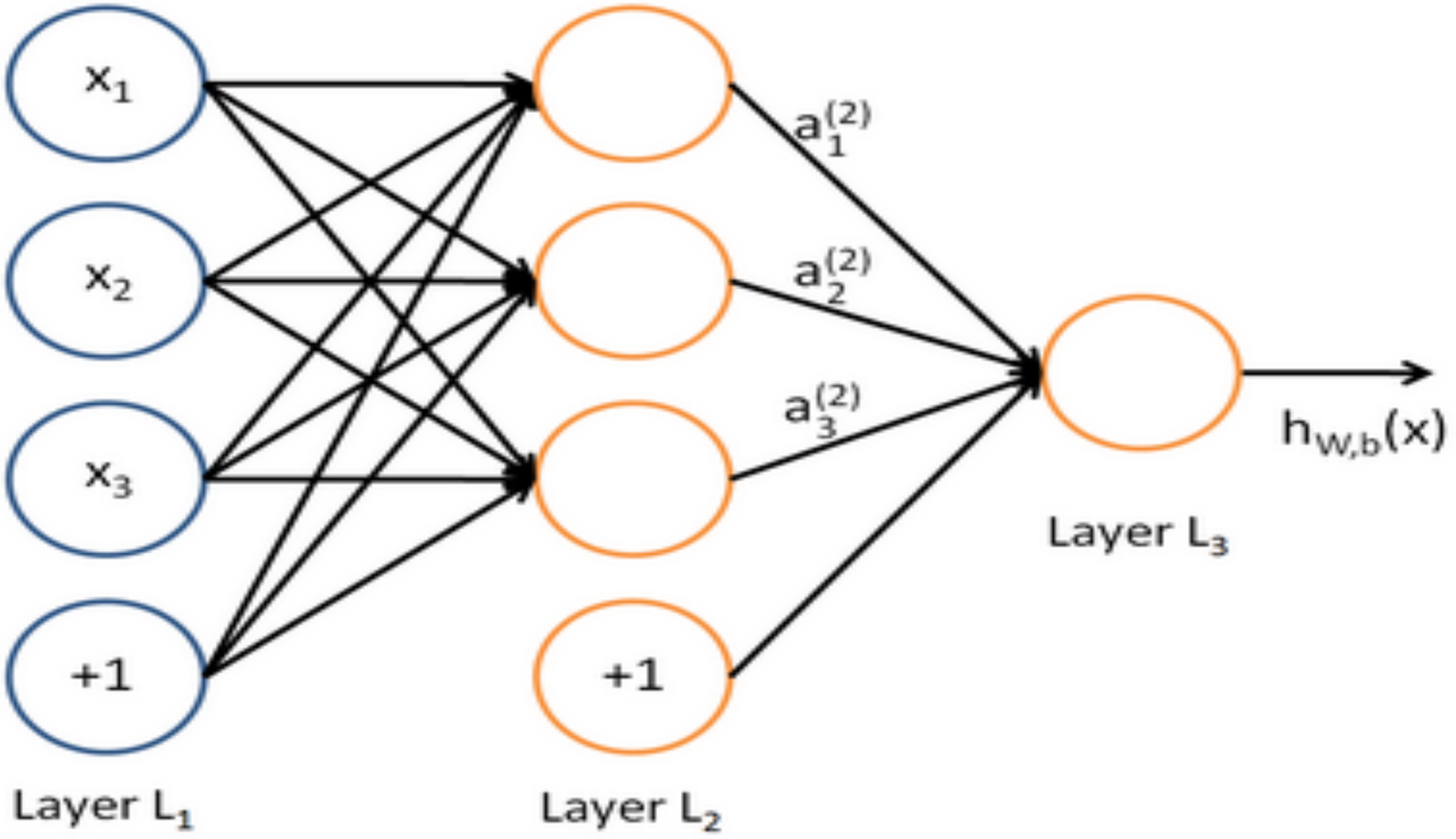
Linear Classifier → LR使用组合特征

GBDT+FM模型

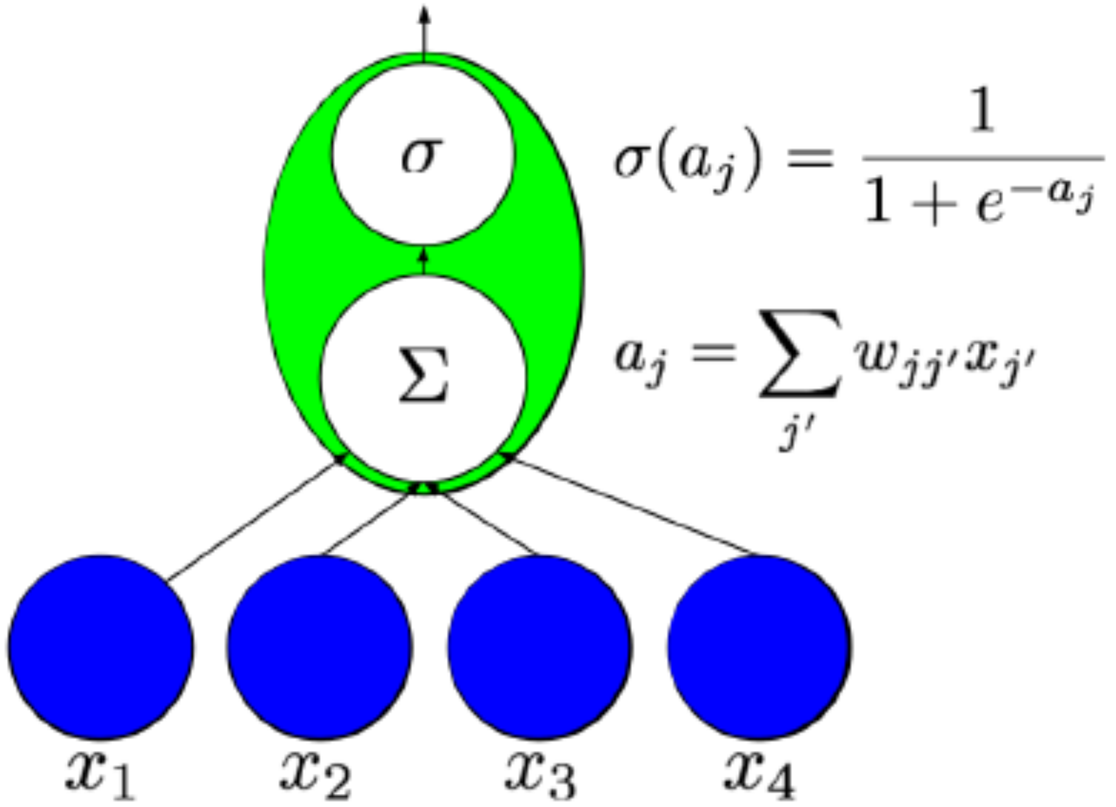
- 2014年提出（香港中文+百度）：GBFM模型
- 集成GBDT和FM各自的优势
  - GBDT发现最有效的组合特征Feature Set
  - 将Feature Set引入FM模型中
- GBFM：贪心策略选择组合特征

TABLE OF CONTENTES

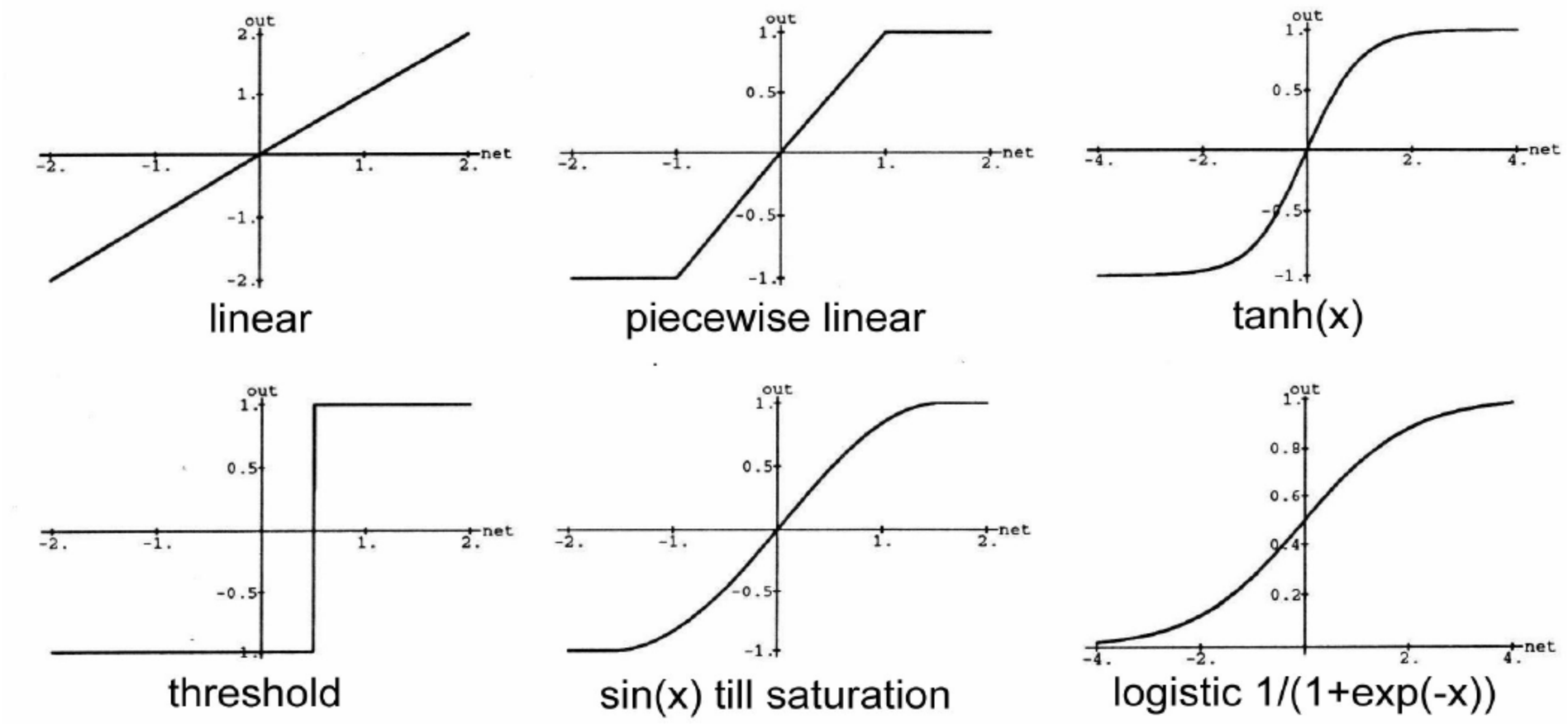
前向神经网络（MLP）



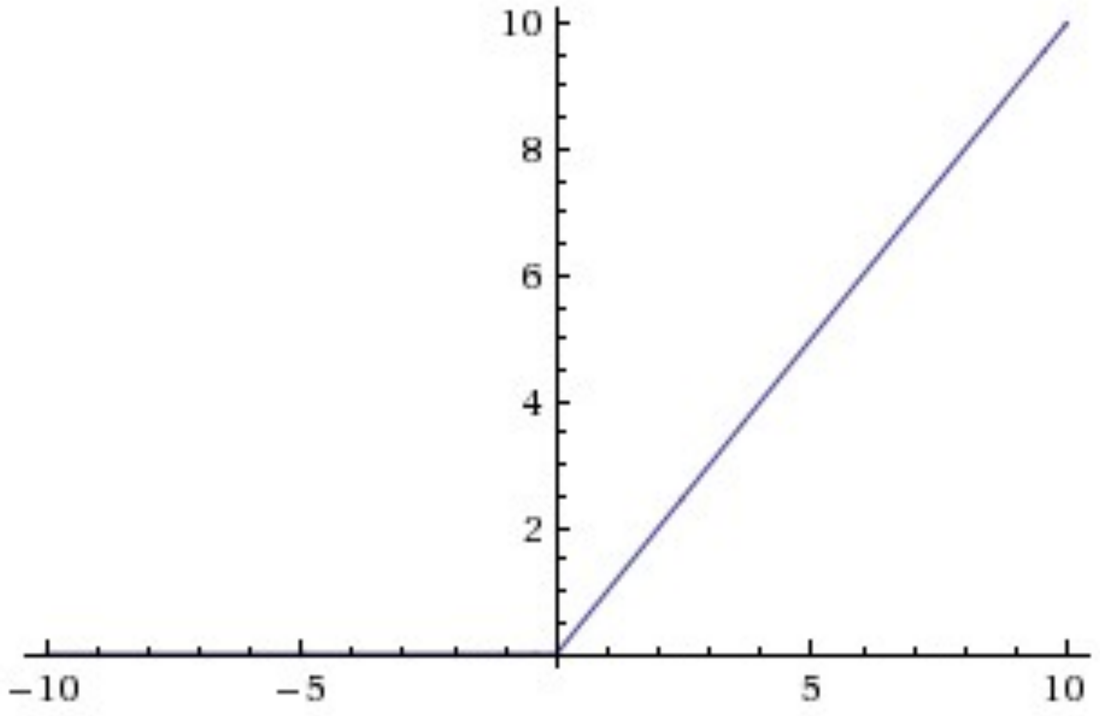
前向神经网络（MLP）：隐层节点



前向神经网络（MLP）：激活函数

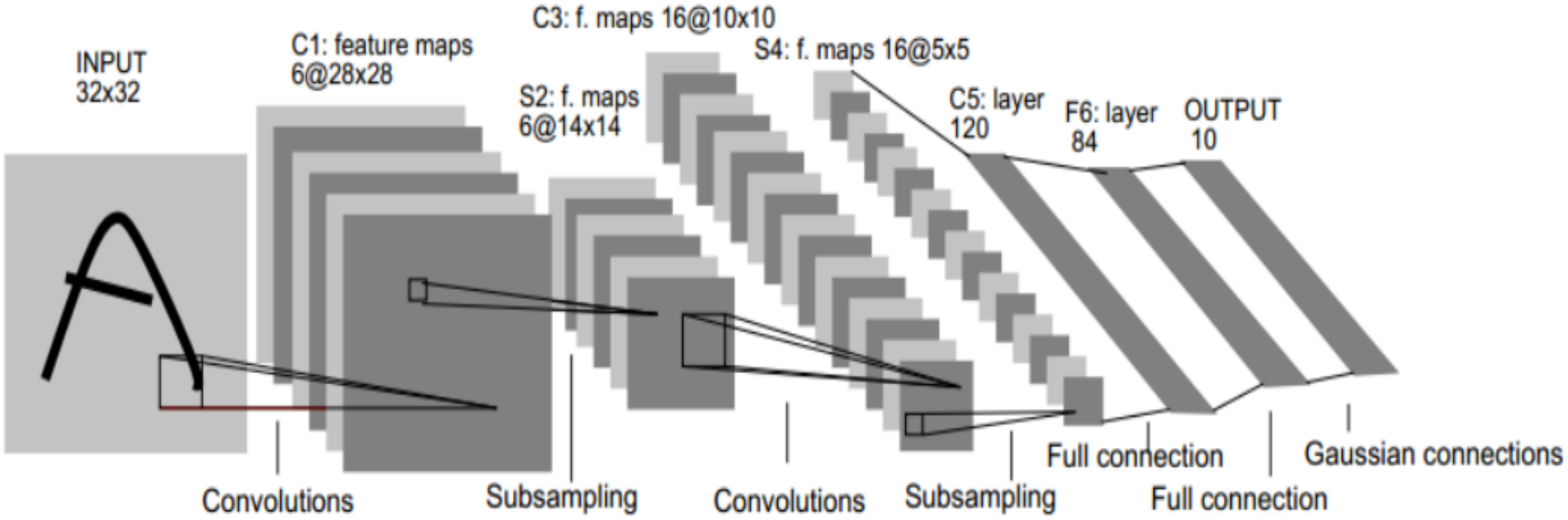


前向神经网络（MLP）：激活函数



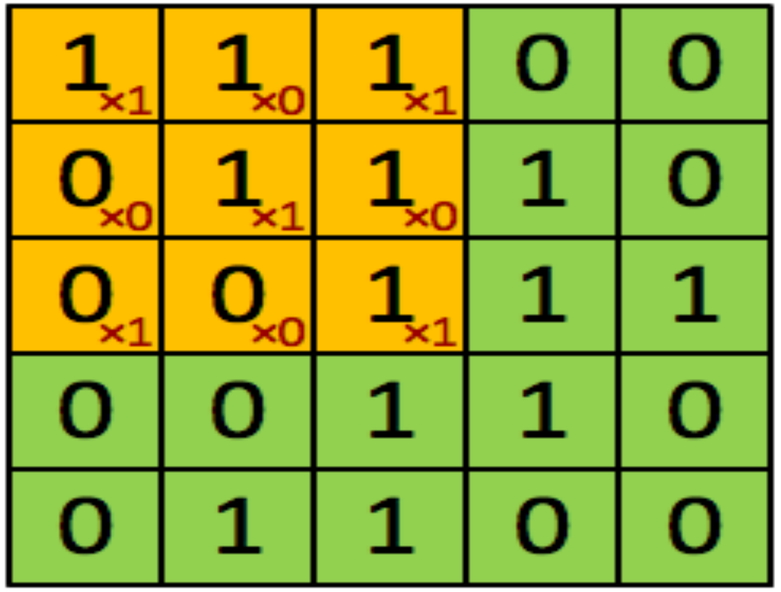
ReLU

CNN：LeNet-5整体结构

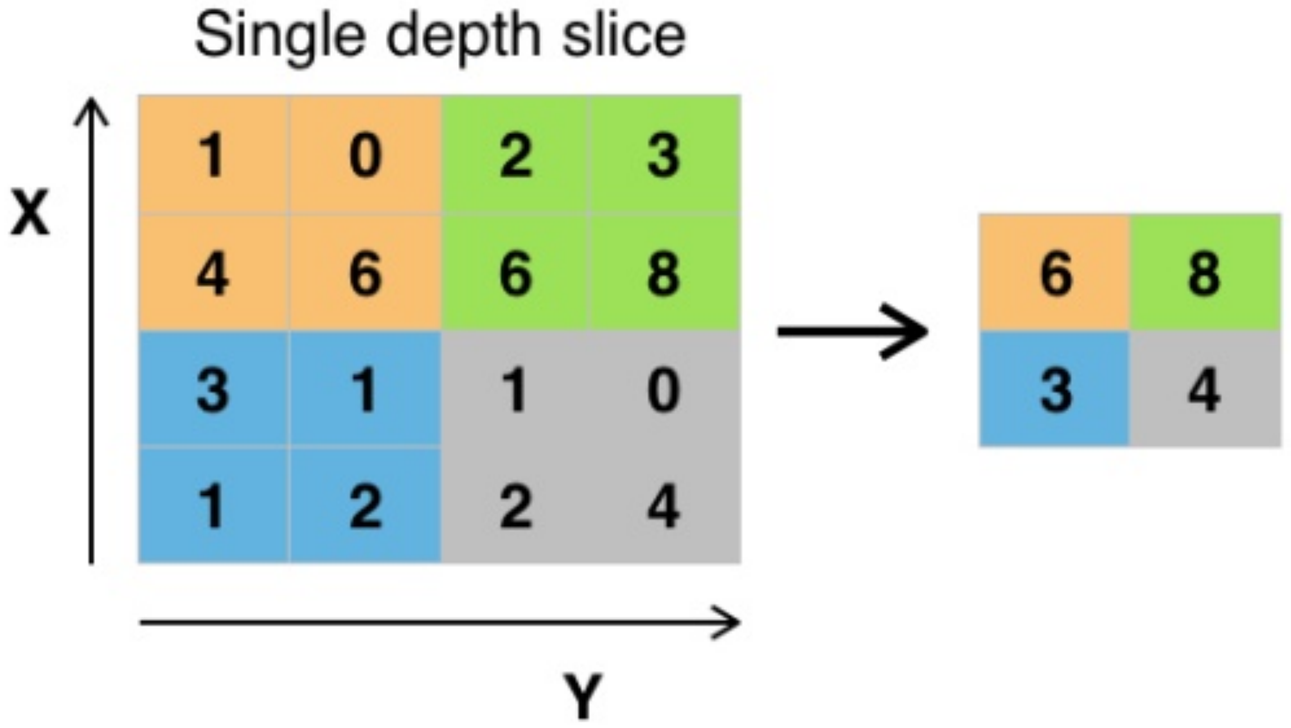


LeNet-5

CNN:卷积层

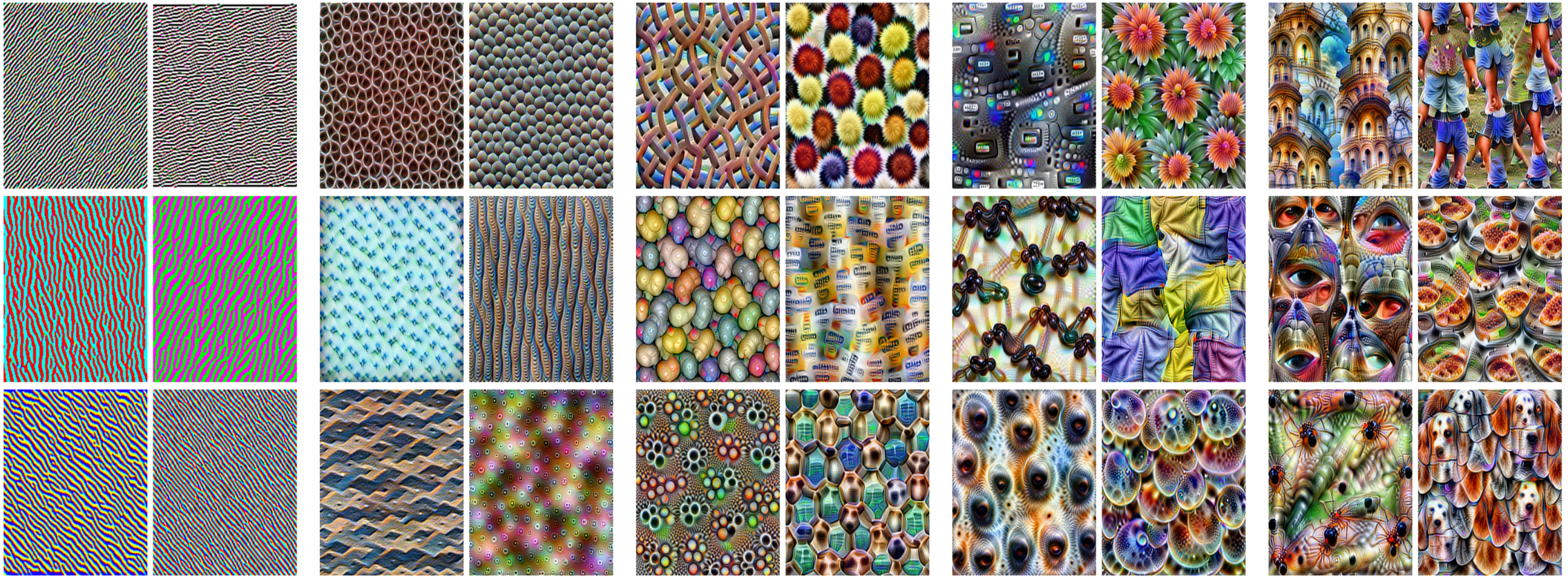

1 1 1 0 0
0 1 1 1 0
0 0 1 1 1
0 0 1 1 0
0 1 1 0 0
1 0 1
0 1 0
1 0 1
Image
4
Convolved
Feature

CNN:MaxPooling层



Example of Maxpool with a 2x2 filter and a stride of 2

CNN:学到了什么特征？



Edges (layer conv2d0)

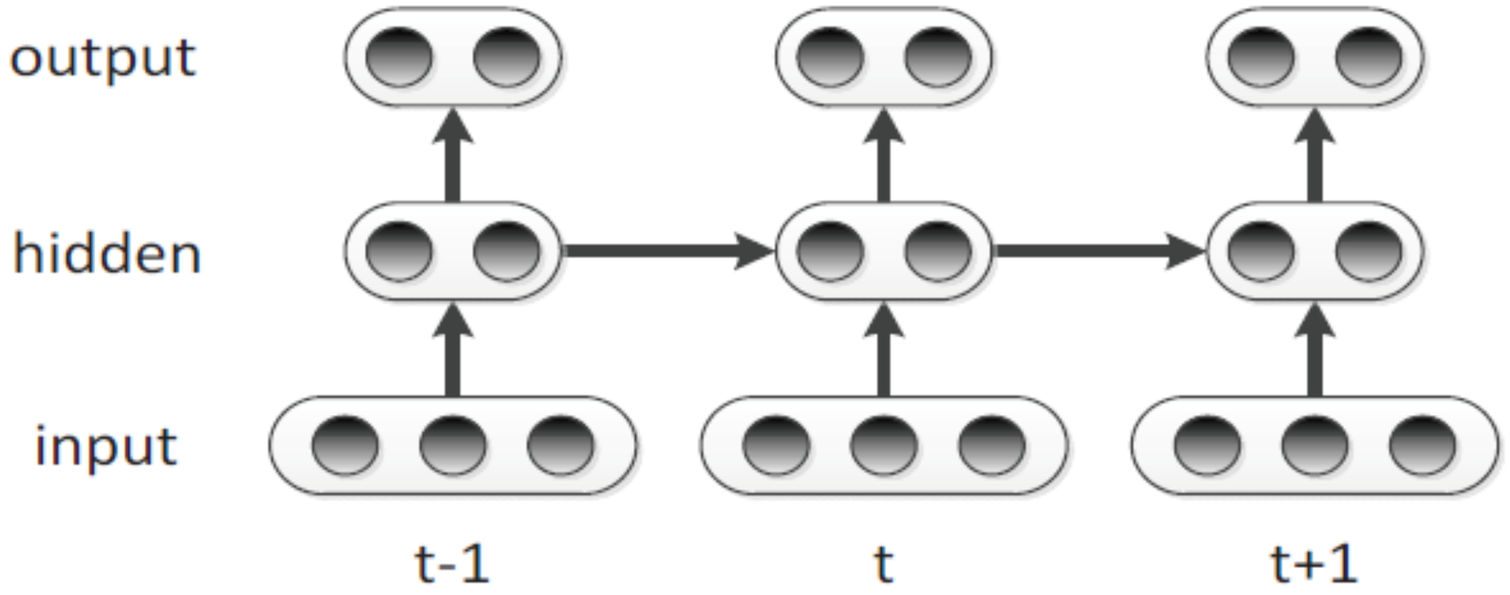
Textures (layer mixed3a)

Patterns (layer mixed4a)

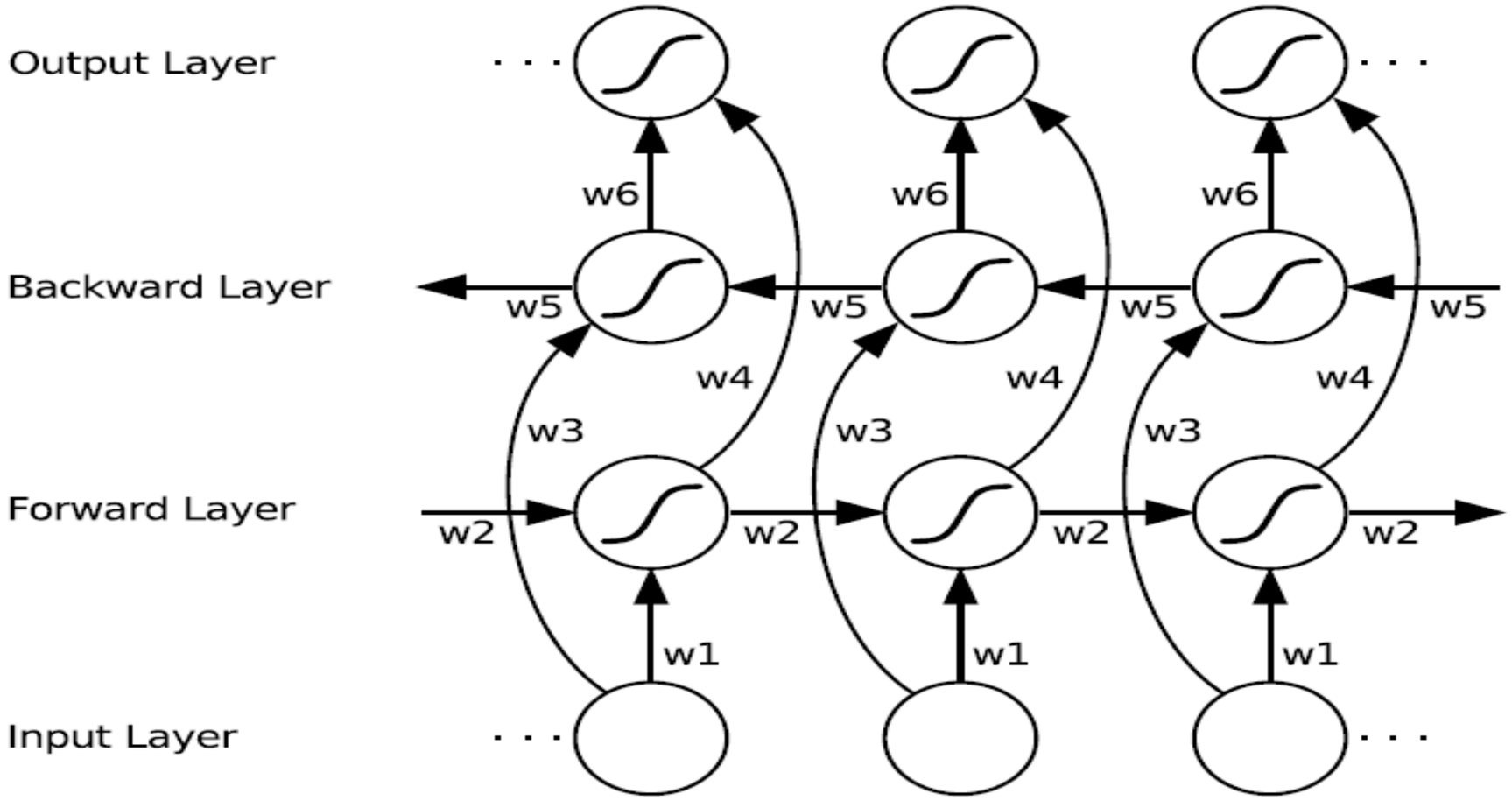
Parts (layers mixed4b & mixed4c)

Objects (layers mixed4d & mixed4e)

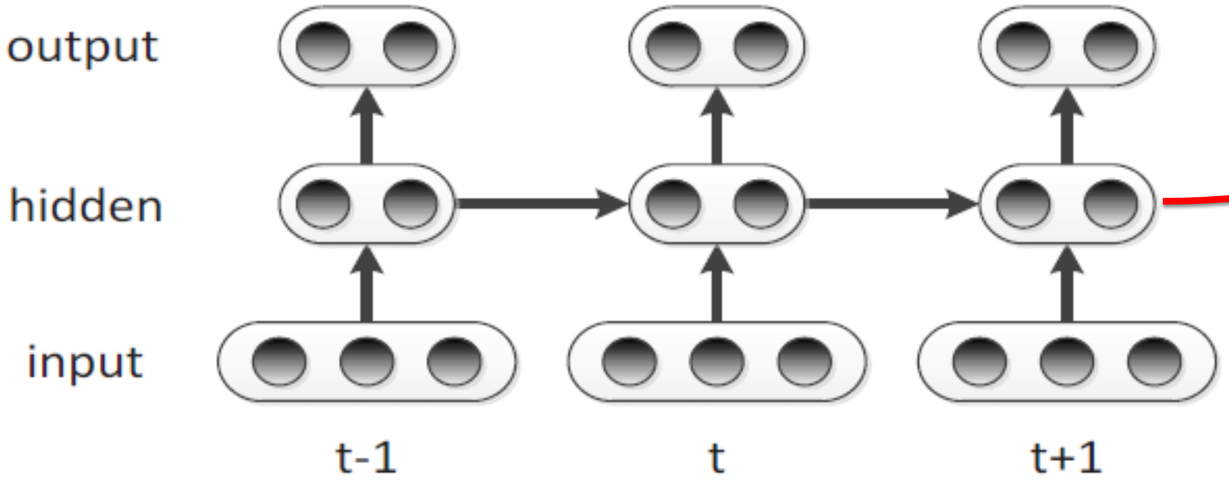
RNN

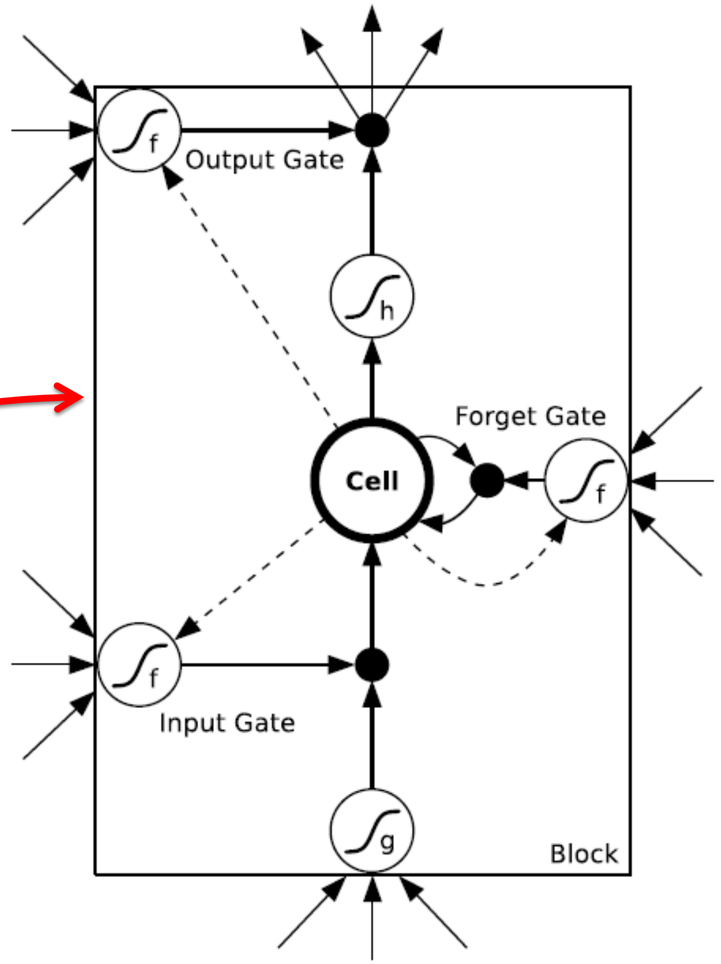


双向RNN


Output Layer
Backward Layer
Forward Layer
Input Layer
w1
w2
w3
w4
w5
w6

LSTM


output
hidden
input
t-1
t
t+1


Output Gate
Forget Gate
Cell
Input Gate
Block

双向深度LSTM

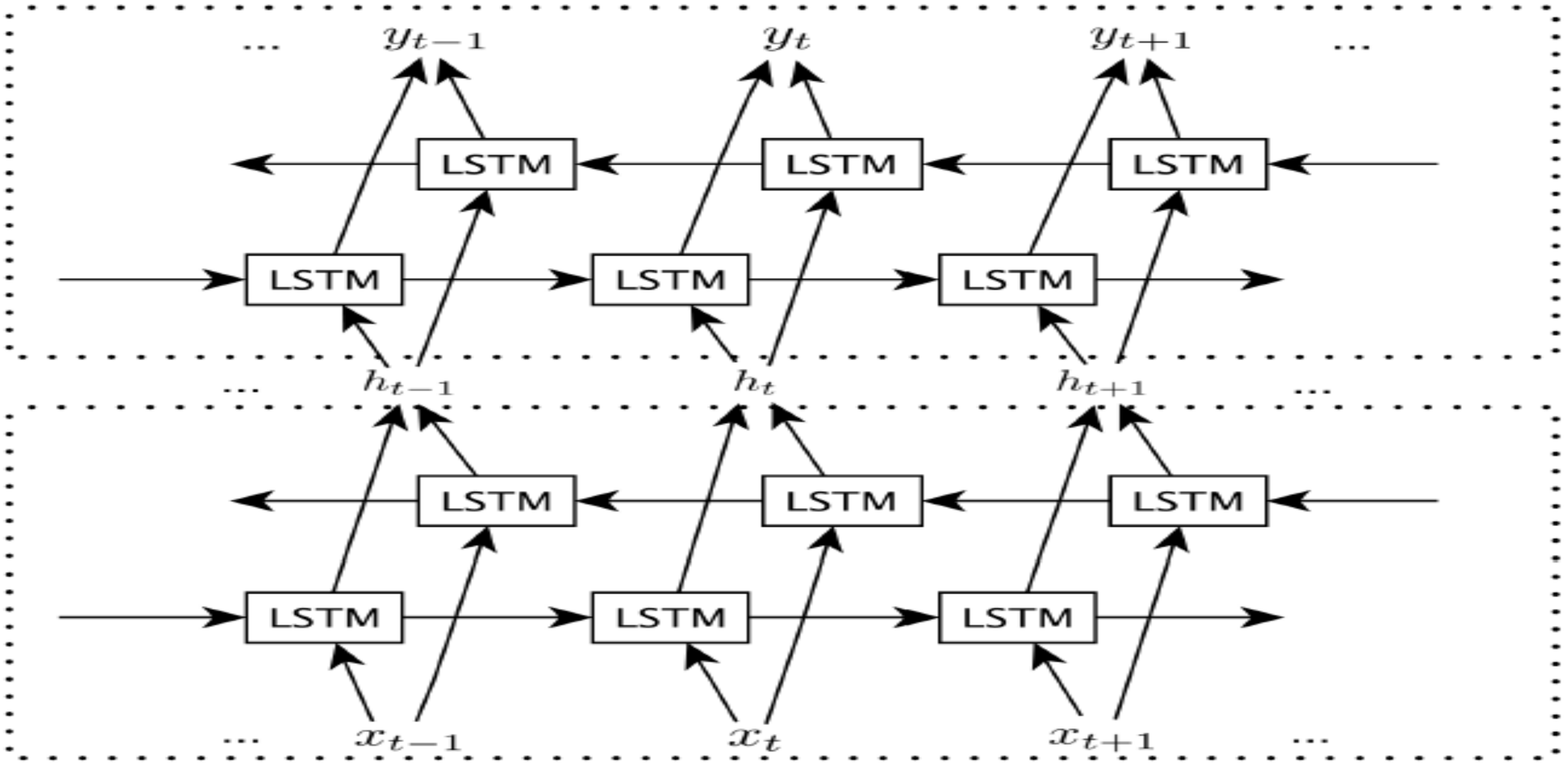


TABLE OF CONTENTES

深度学习CTR模型要解决的几个关键问题

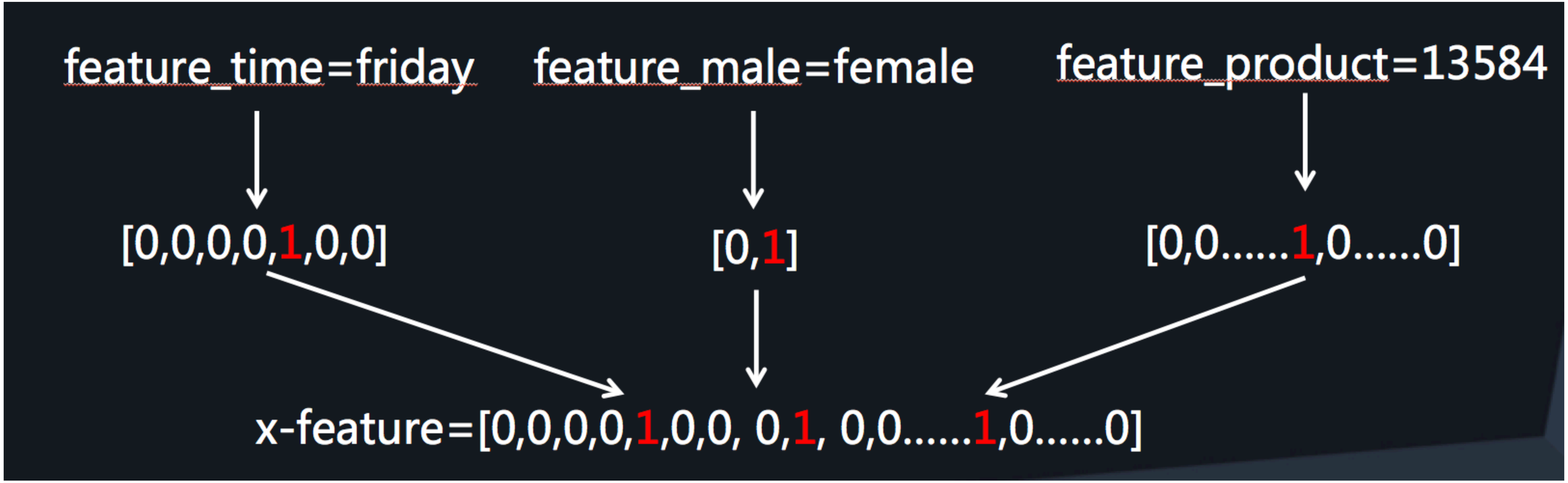
- CTR任务特点：大量离散特征的表示问题
- CTR任务特点：如何快速处理大量高维度稀疏特征？（OneHot 2 Dense）
- 特征工程：如何从手工到自动？（深度学习的优势）
- 特征工程：如何捕获和表达两两组合特征？（FM机制神经网络化）
- 特征工程：如何捕获和表达多组组合特征？（利用Deep网络）

CTR任务中的特征类型

- 连续特征
  - 收入，身高，体重.....
  - 适合DNN处理
- 离散特征
  - 职业，性别，毕业学校.....
  - 不适合DNN处理

离散特征如何让DNN可以处理？

- 直观思路：离散特征使用Onehot表达

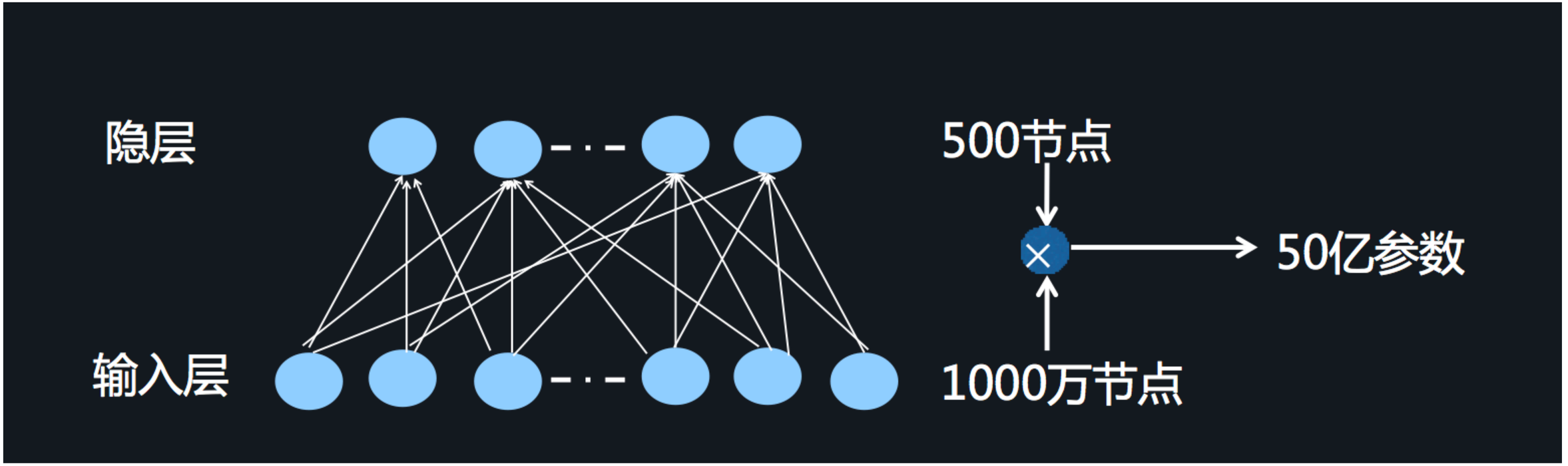


离散特征如何让DNN可以处理？

- 直观思路：离散特征使用Onehot表达

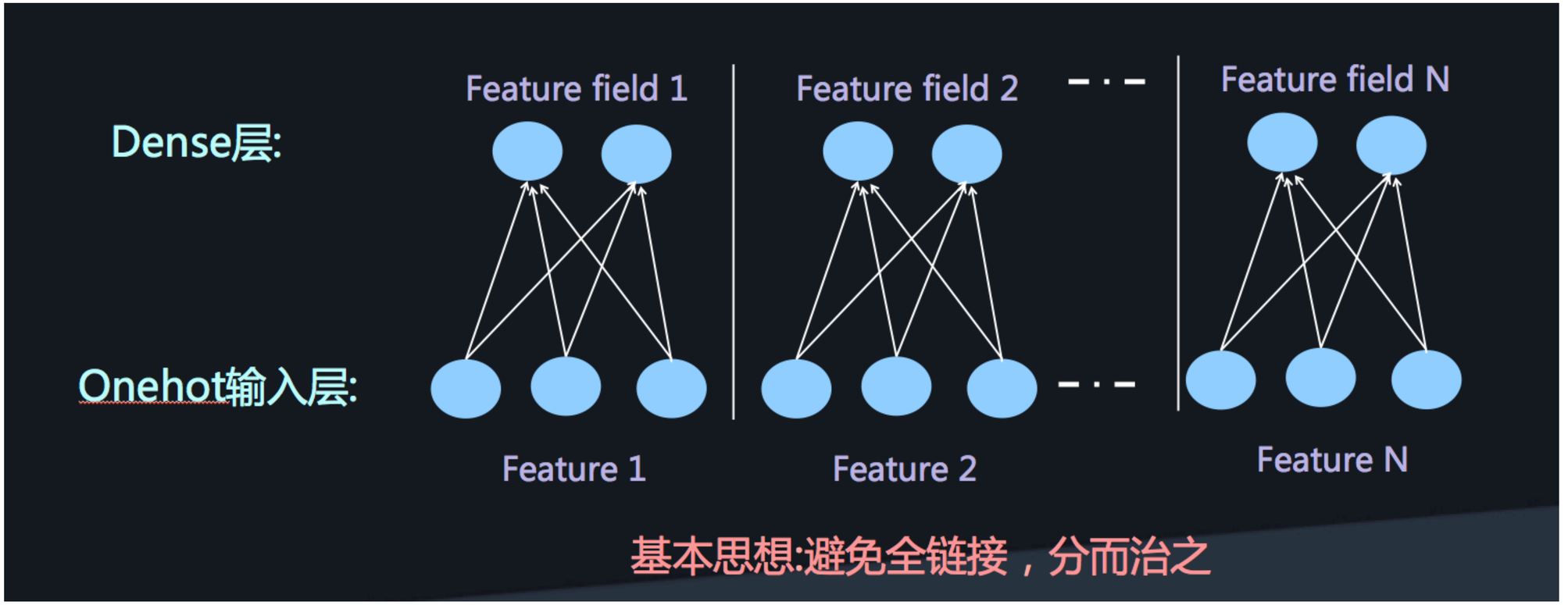
离散特征如何让DNN可以处理？

- Onehot作为DNN输入的问题：CTR预估任务里不可行

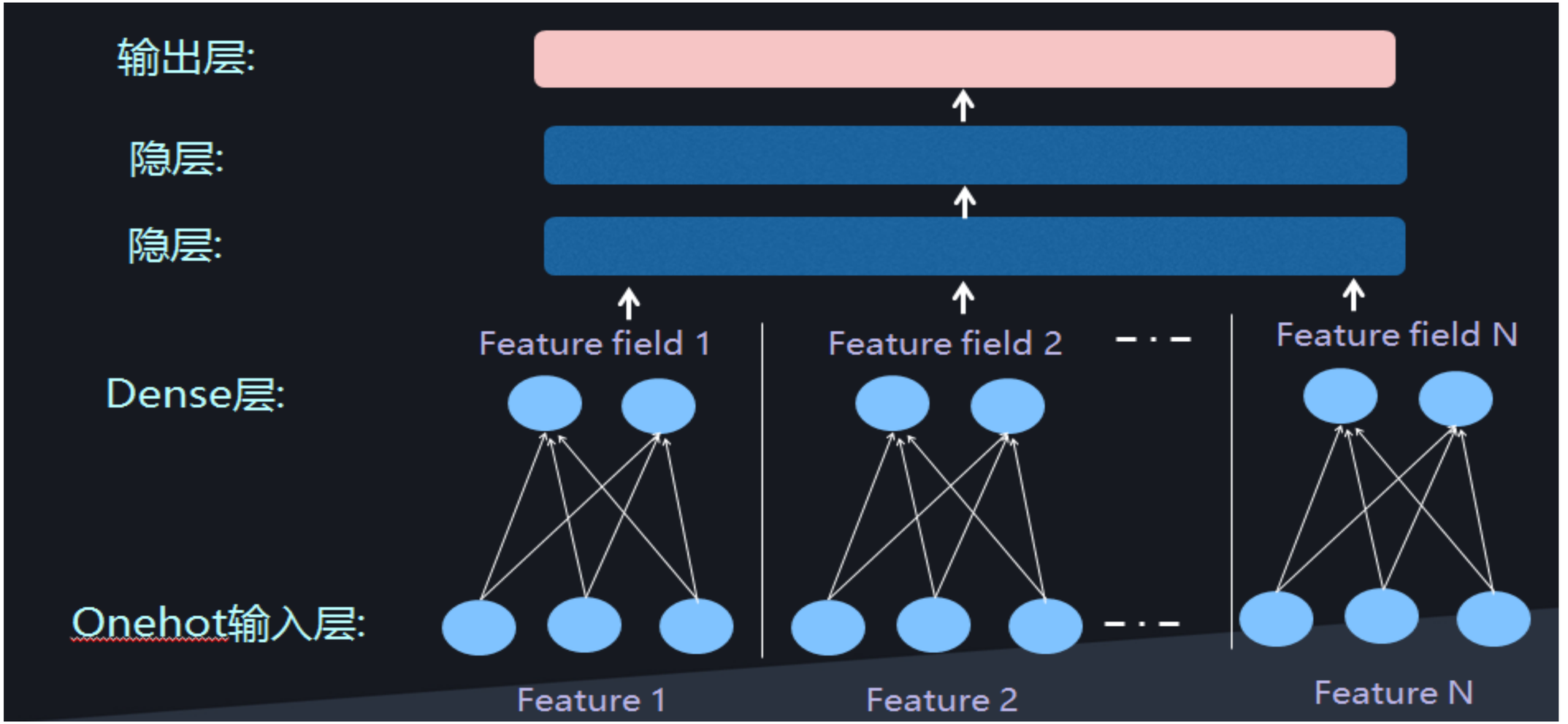


离散特征如何让DNN可以处理？

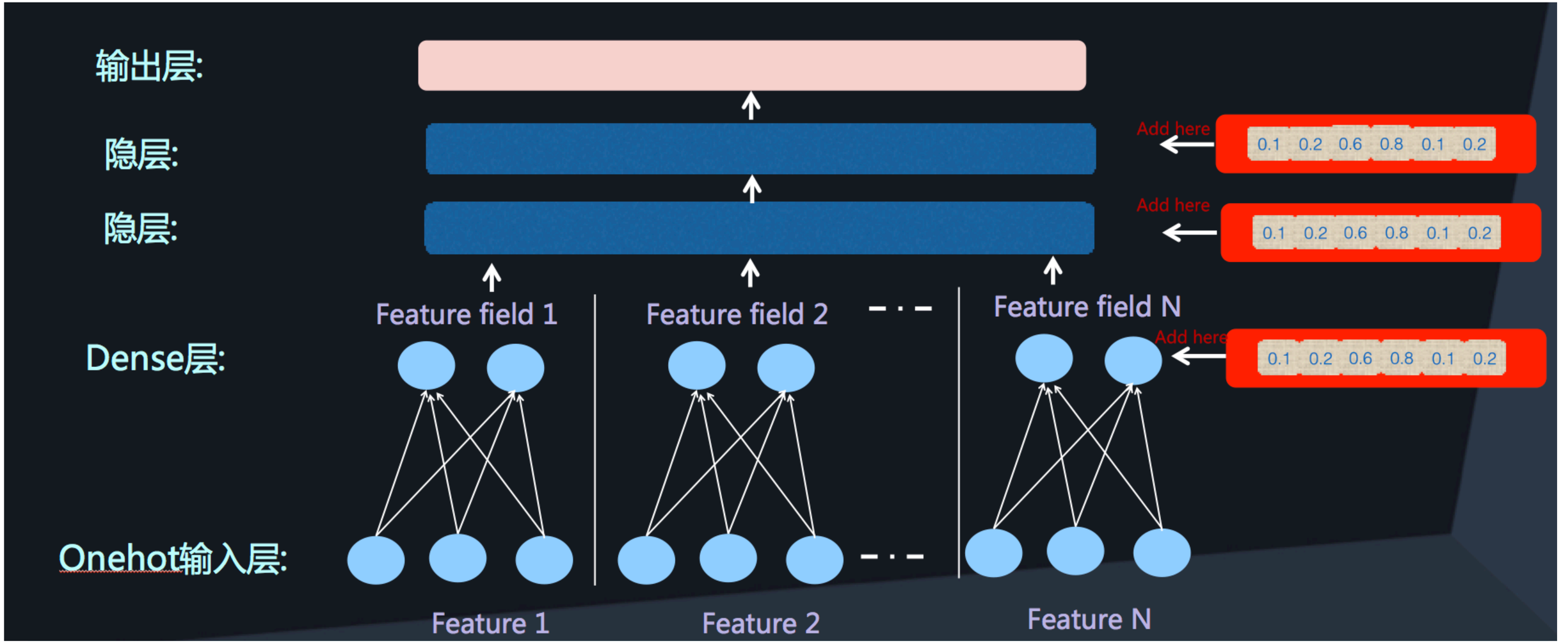
- 解决思路：从OneHot到Dense Vector



形成DNN结构

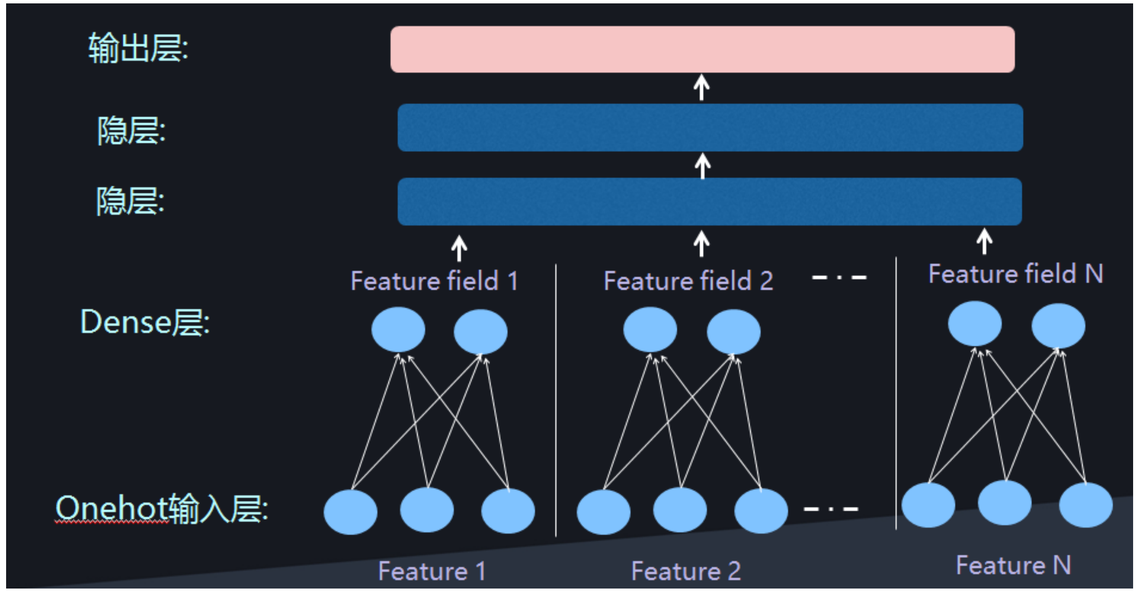

输出层:
隐层:
隐层:
Feature field 1
Feature field 2
Feature field N
Dense层:
Onehot输入层:
Feature 1
Feature 2
Feature N

再加入连续特征


输出层:
隐层:
隐层:
Add here
0.1 0.2 0.6 0.8 0.1 0.2
Add here
0.1 0.2 0.6 0.8 0.1 0.2
Add here
0.1 0.2 0.6 0.8 0.1 0.2
Feature field 1
Feature field 2
Feature field N
Dense层:
Onehot输入层:
Feature 1
Feature 2
Feature N

这是通用的深层模型结构

这就是FNN模型：Factorisation-Machine Supported Neural Networks

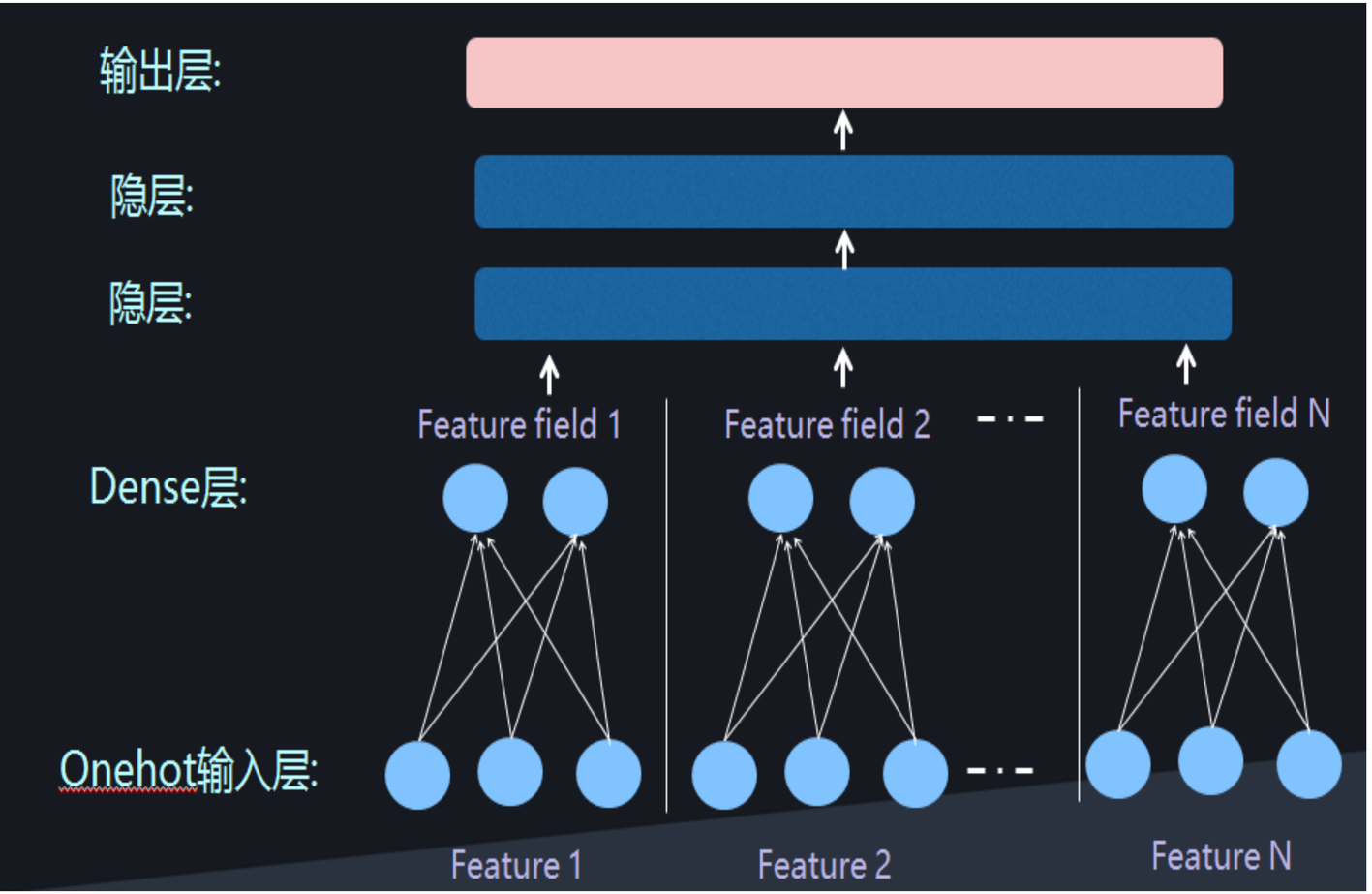


Wide&Deep模型的Deep部分：相同的结构

很多其它改进模型的Deep部分：相同的结构

几乎所有DL+CTR模型的输入部分:这种Onehot2Dense映射

DNN输入问题解决了，但是.....



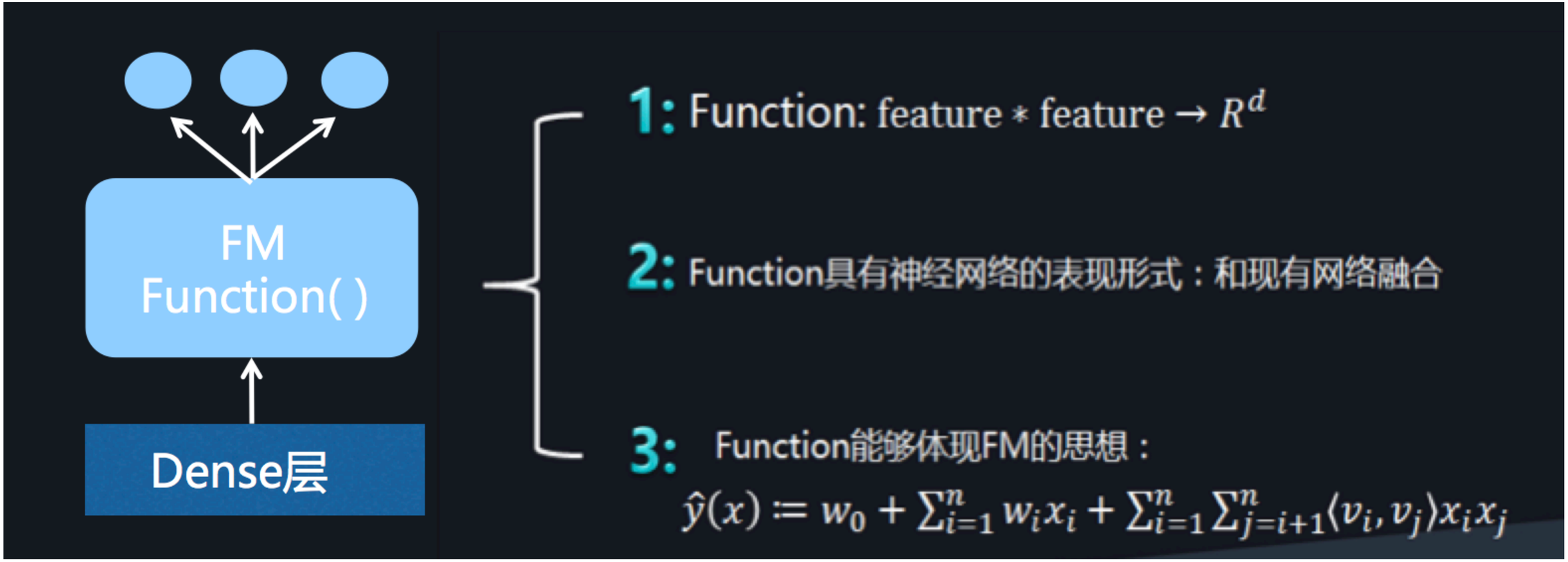
低阶和高阶特征组合隐含地体现在隐层

我们希望：把低阶特征组合单独建模

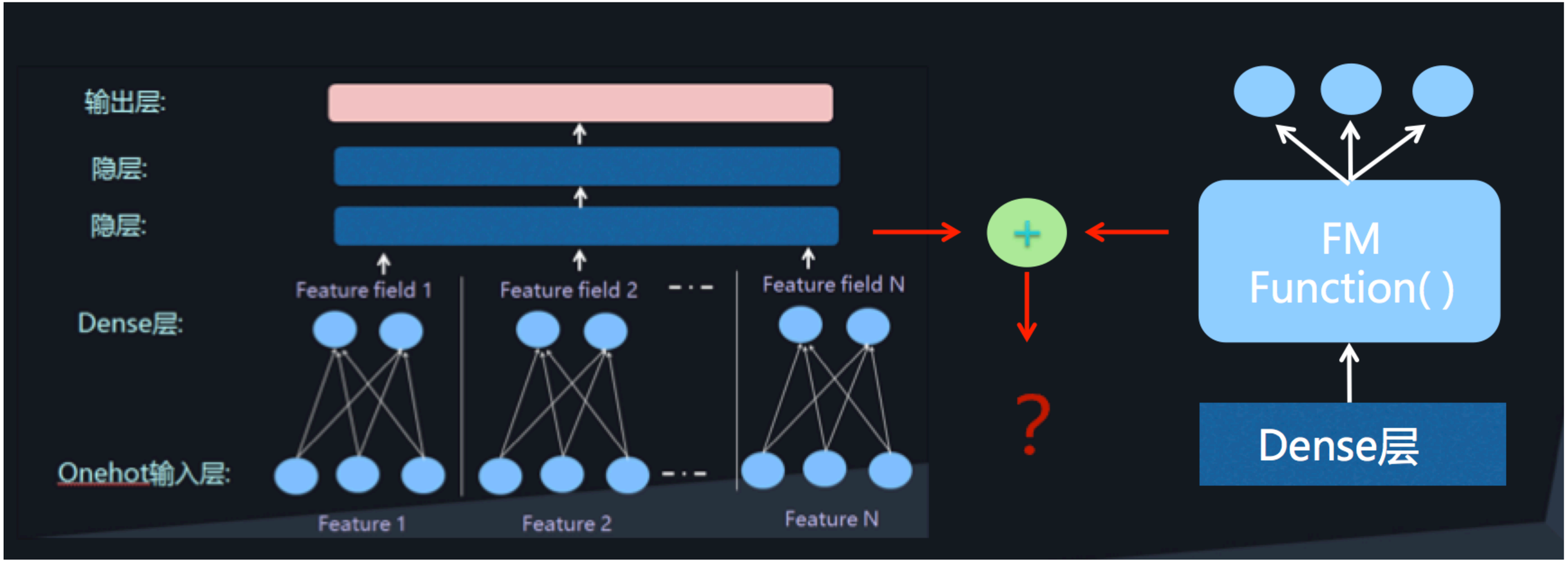
?

把低阶特征组合单独建模

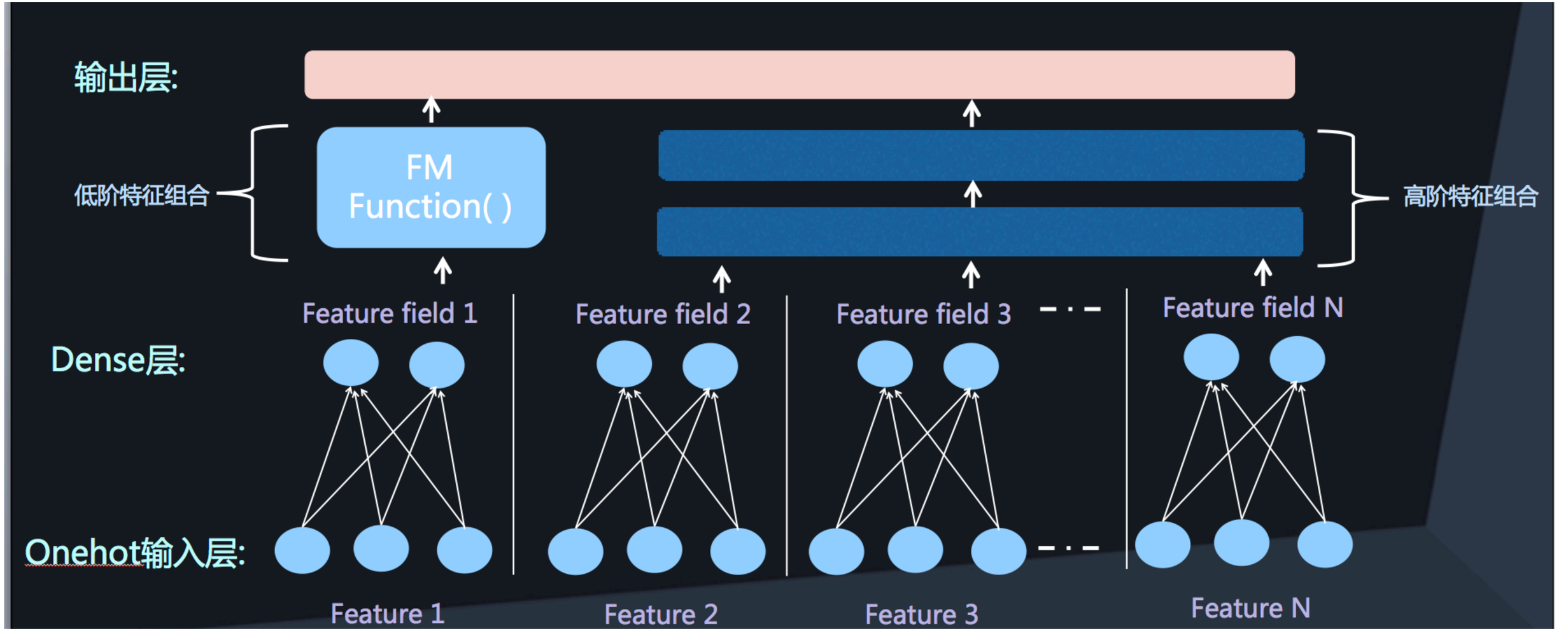
- 首先需要：定义一个神经网络版的低阶特征组合模型



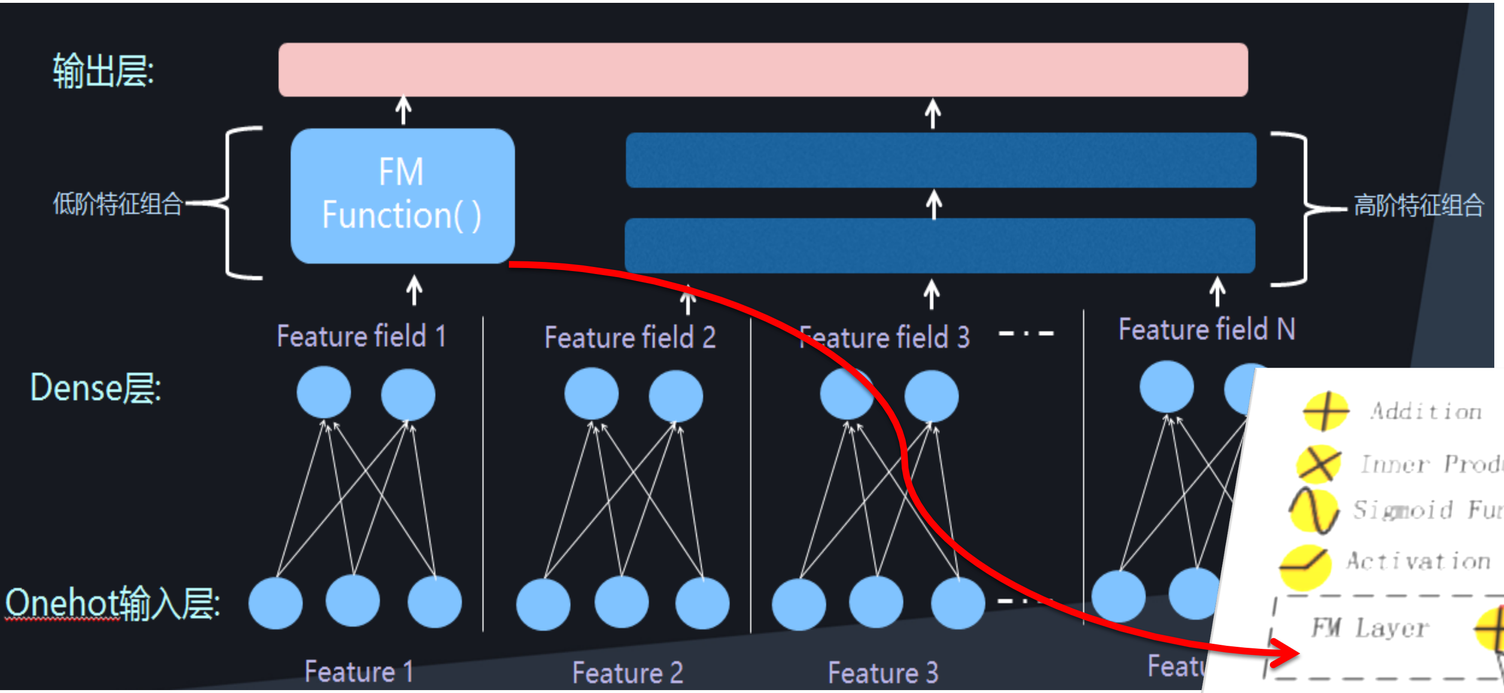
把低阶特征组合模型插入网络结构中

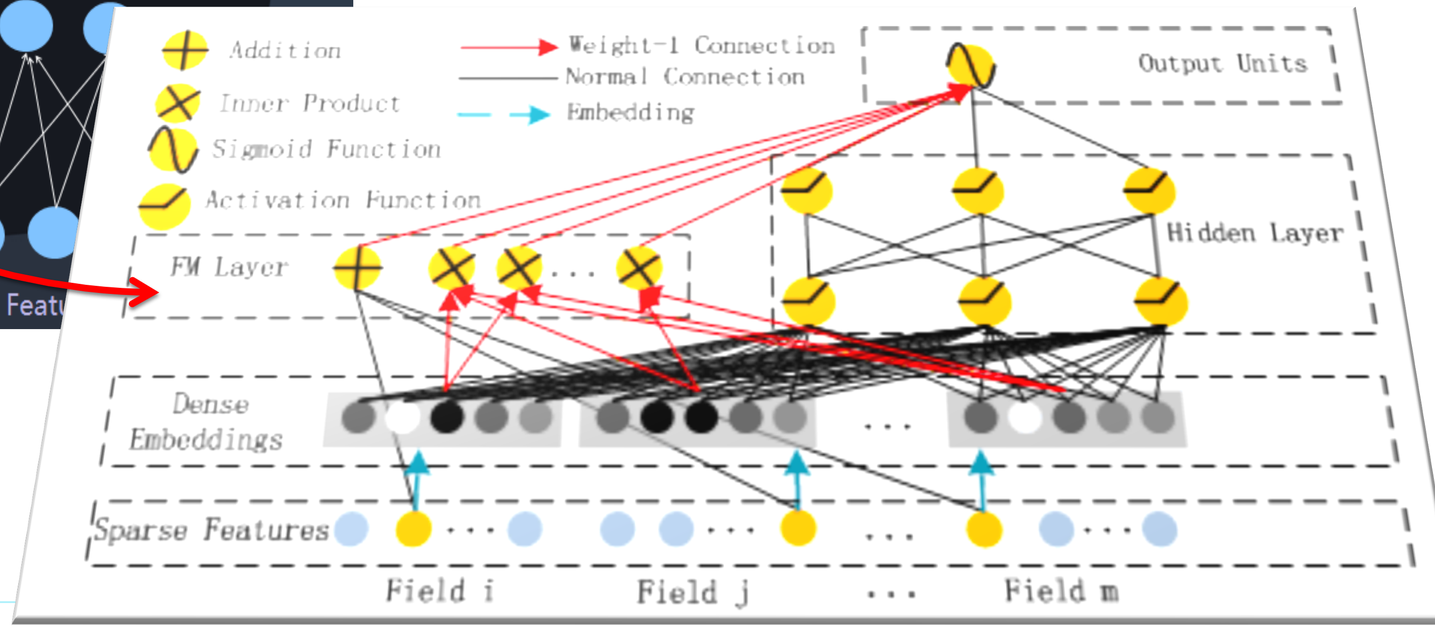


典型网络融合结构之一：并行结构



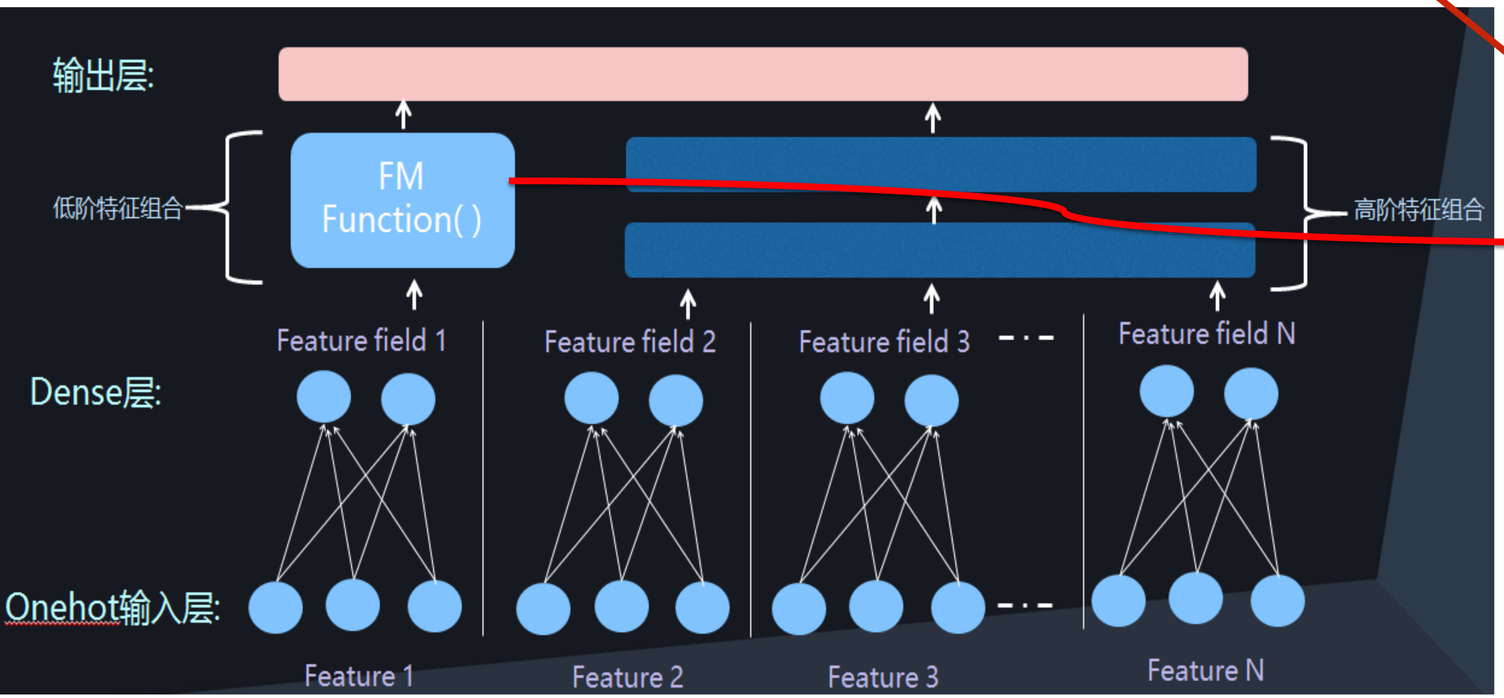
并行结构实例：DeepFM模型

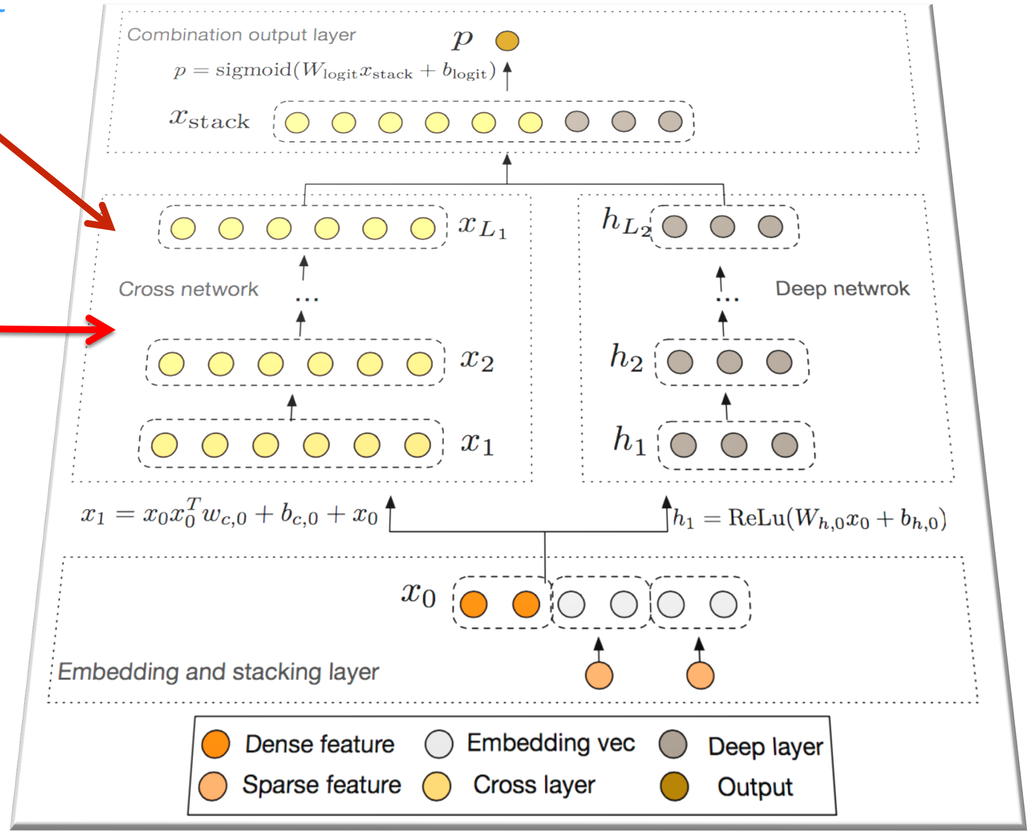




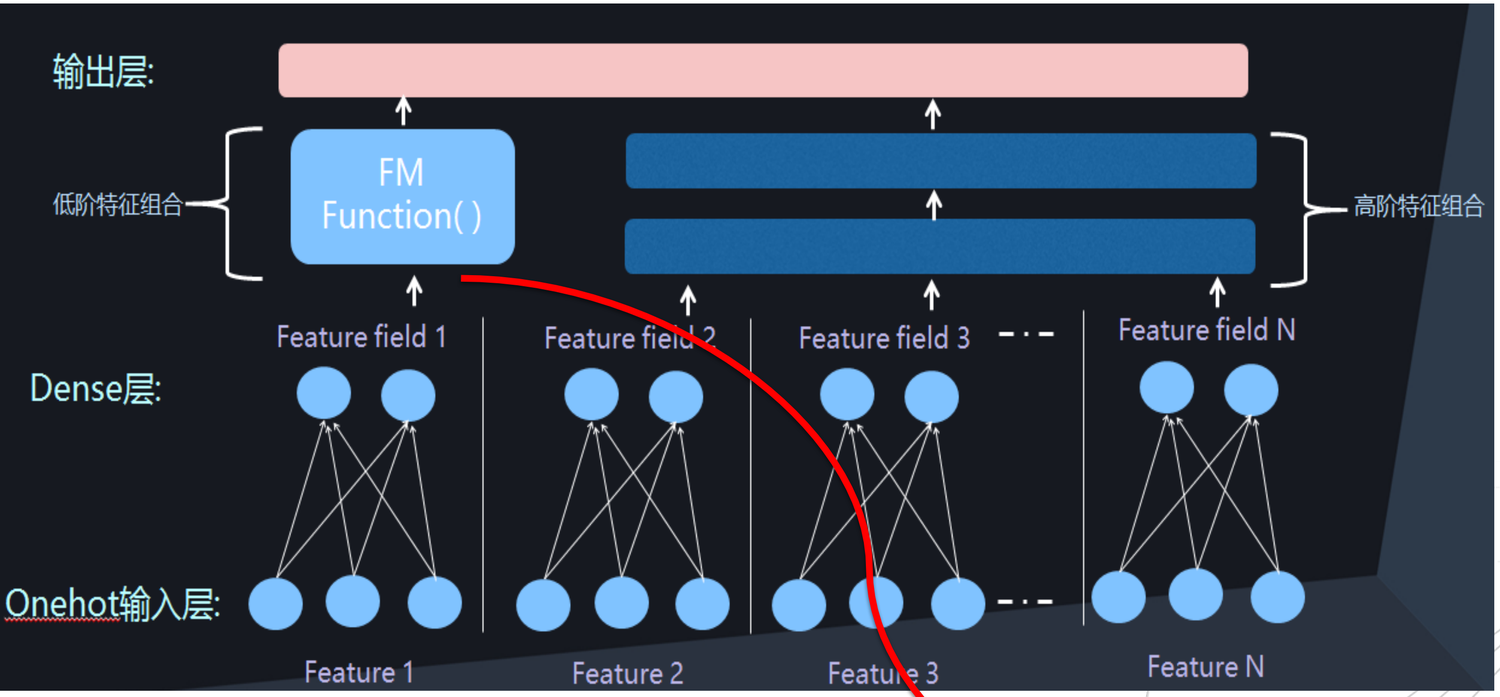
并行结构实例：Deep&Cross模型

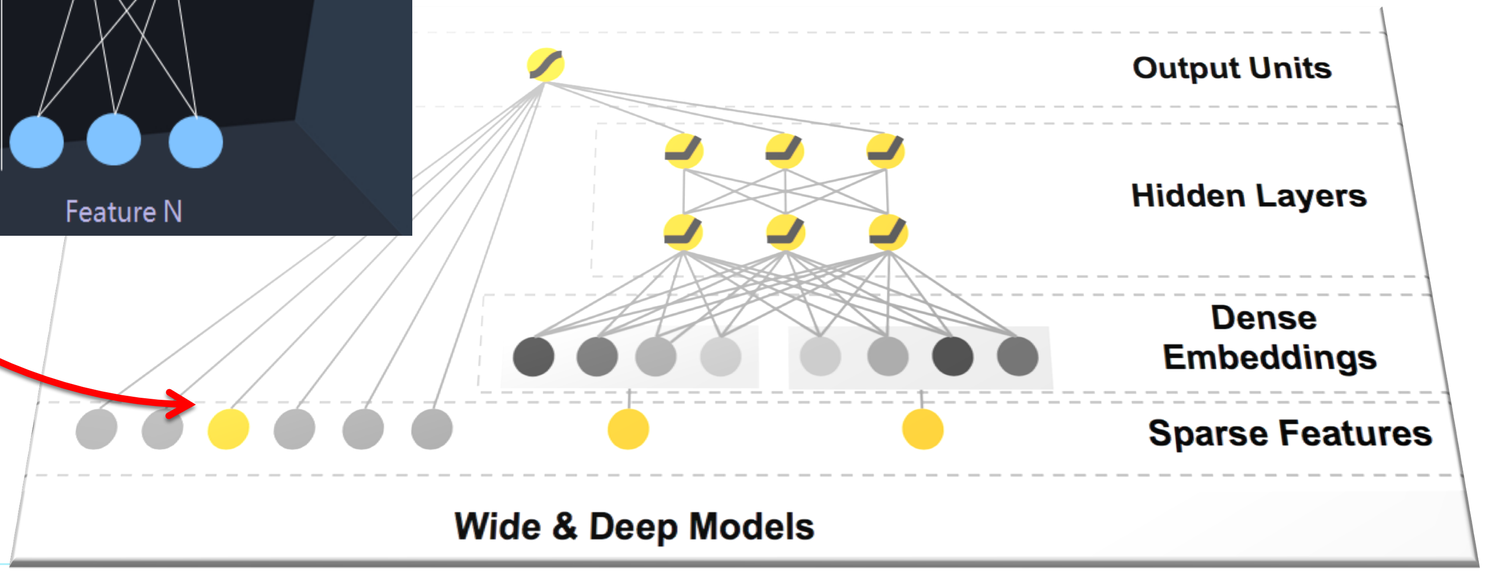
FM Function=Cross Network=ResNet





并行结构实例：Wide&Deep模型



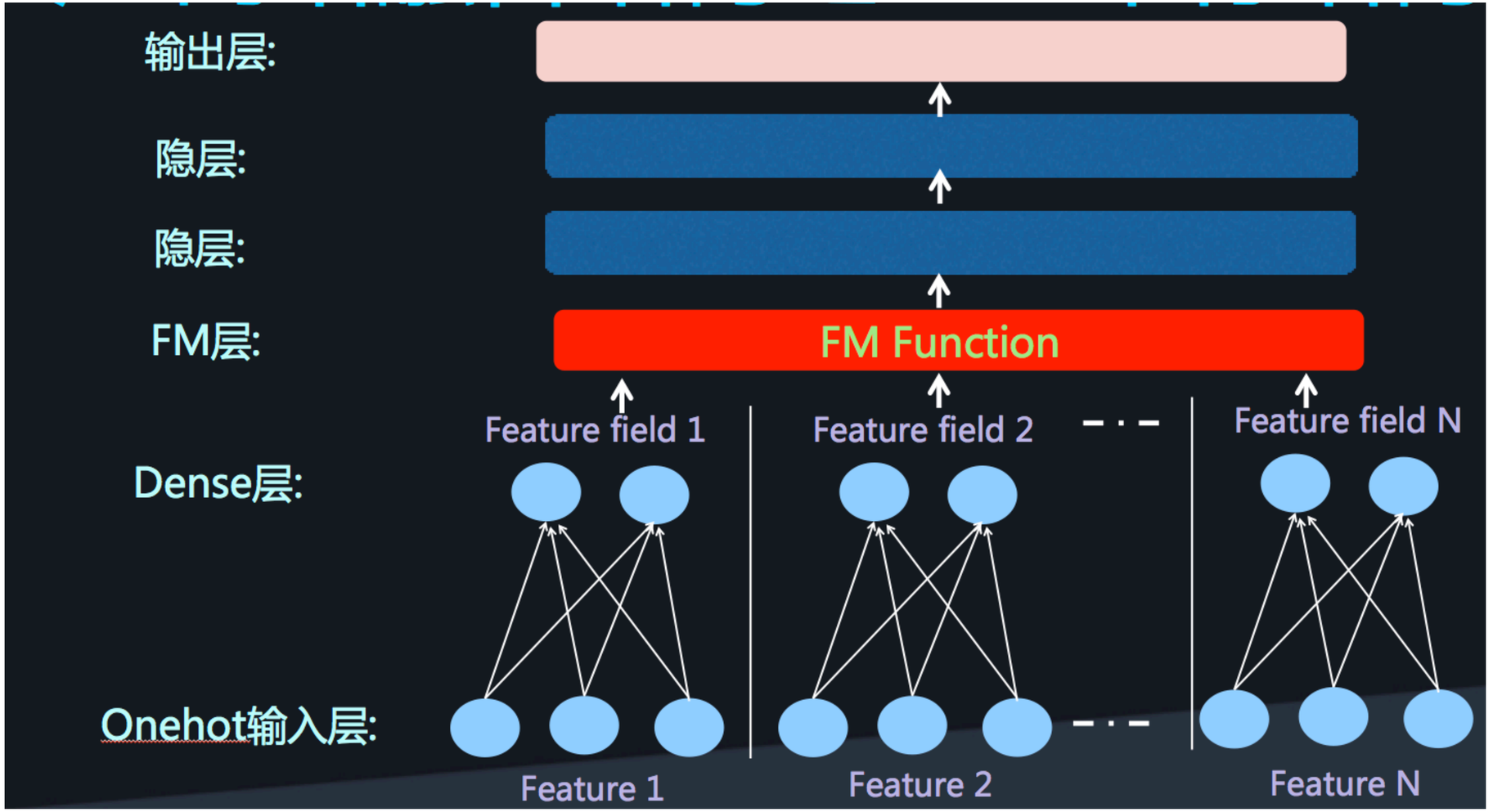


Wide网络LR模型，无FM组合

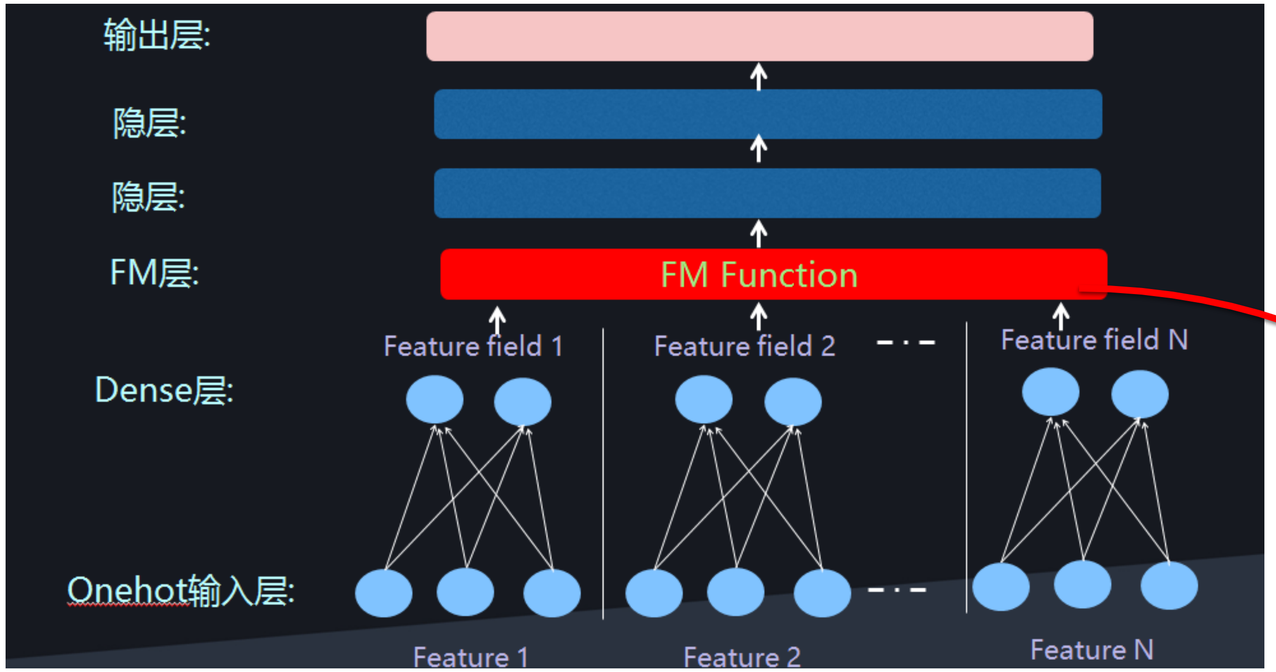
两个不同的输入结构

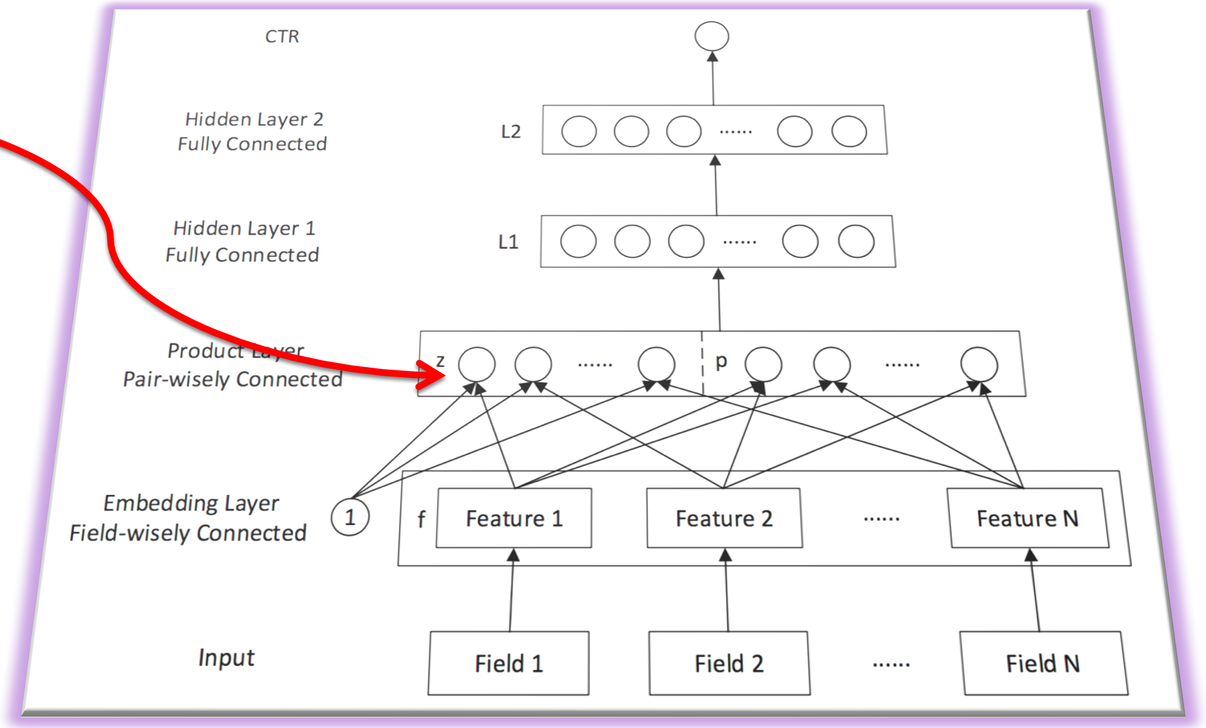
Wide部分手工特征工程+Cross 特征

典型网络融合结构之二：串行结构

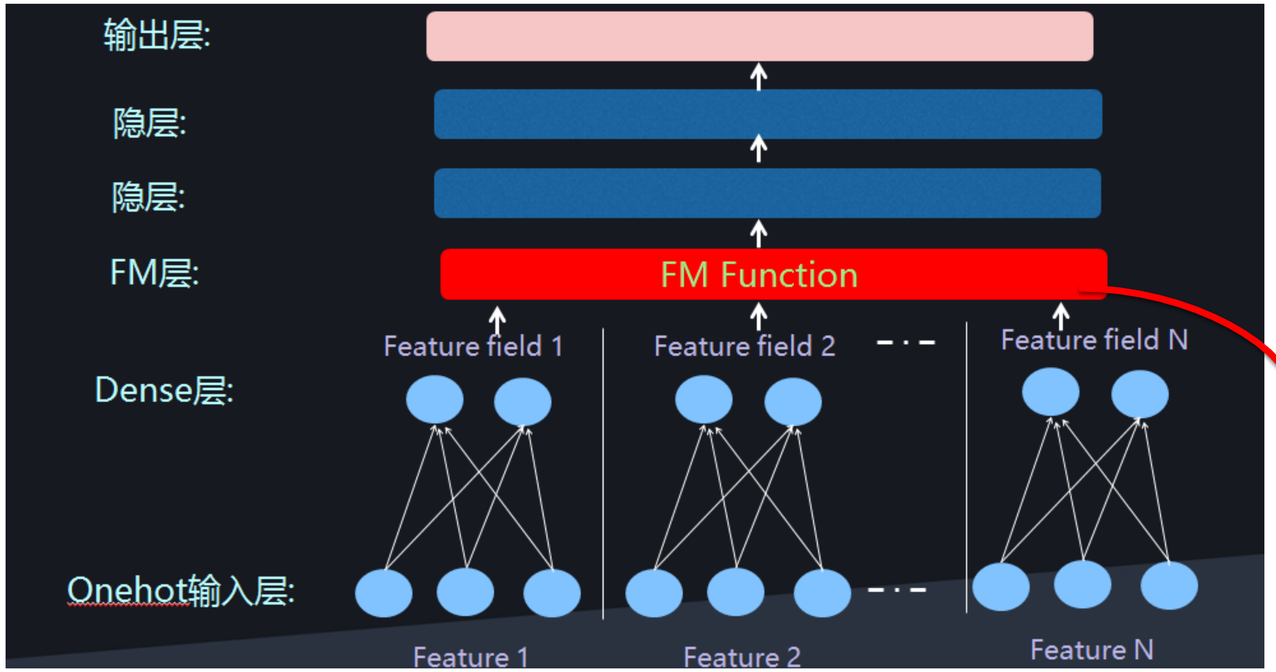


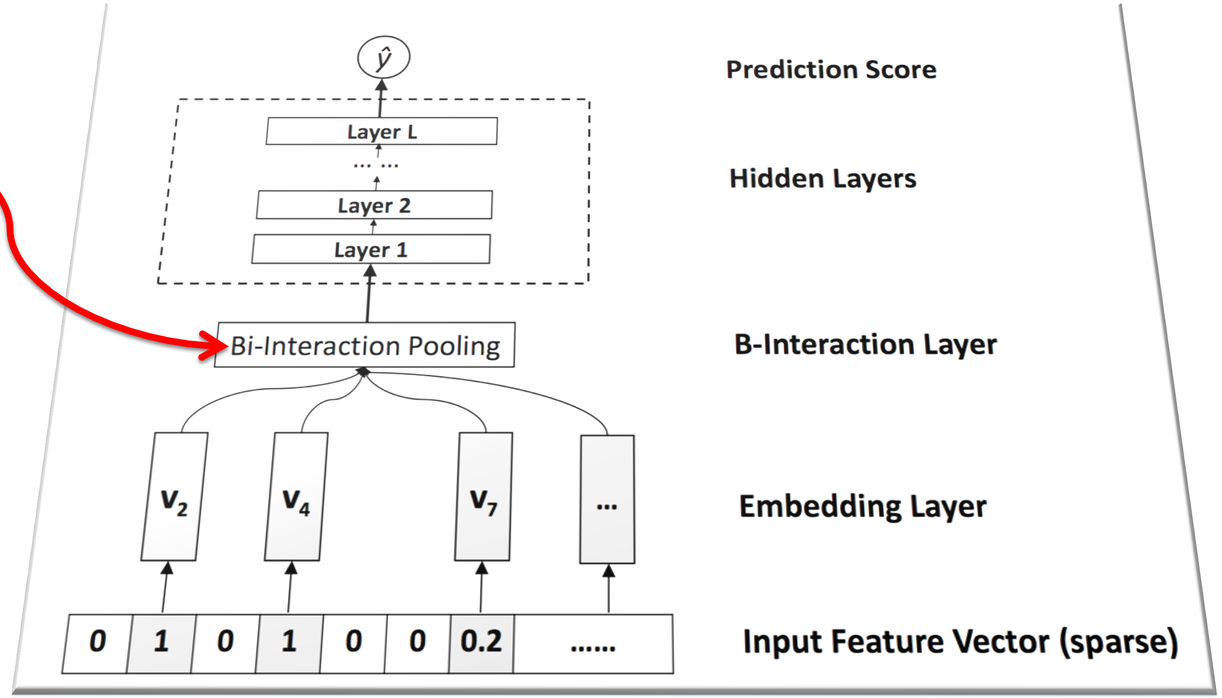
串行结构实例：PNN模型



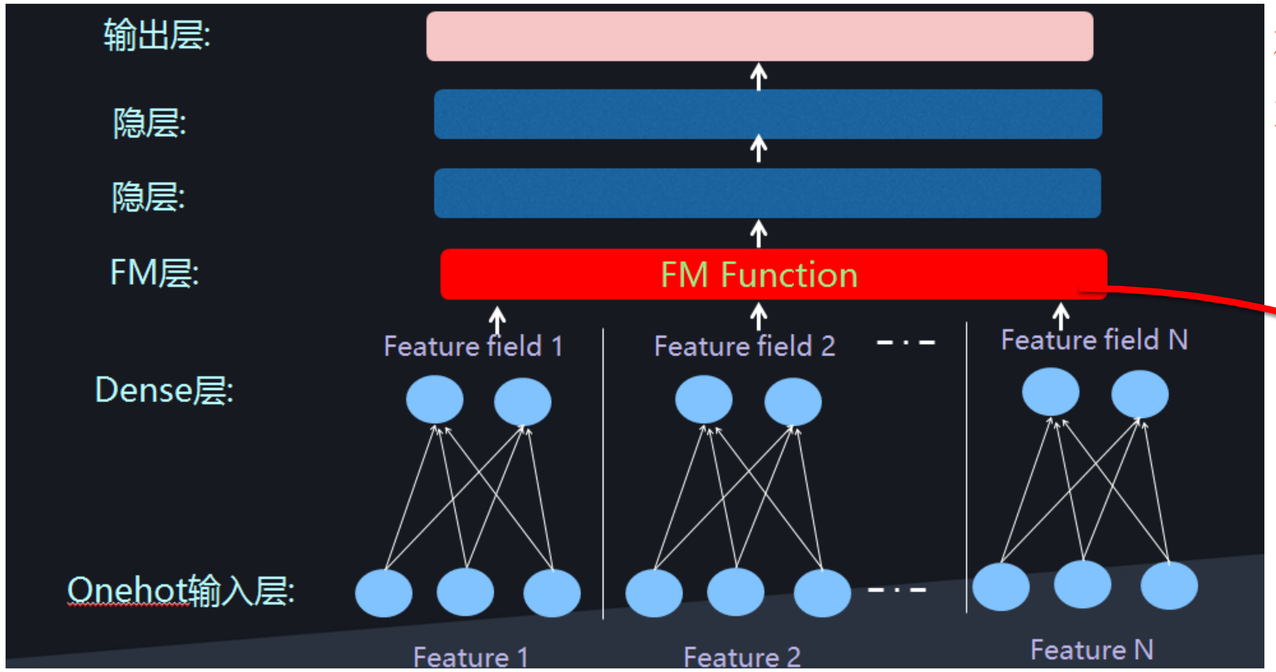


串行结构实例：NFM模型

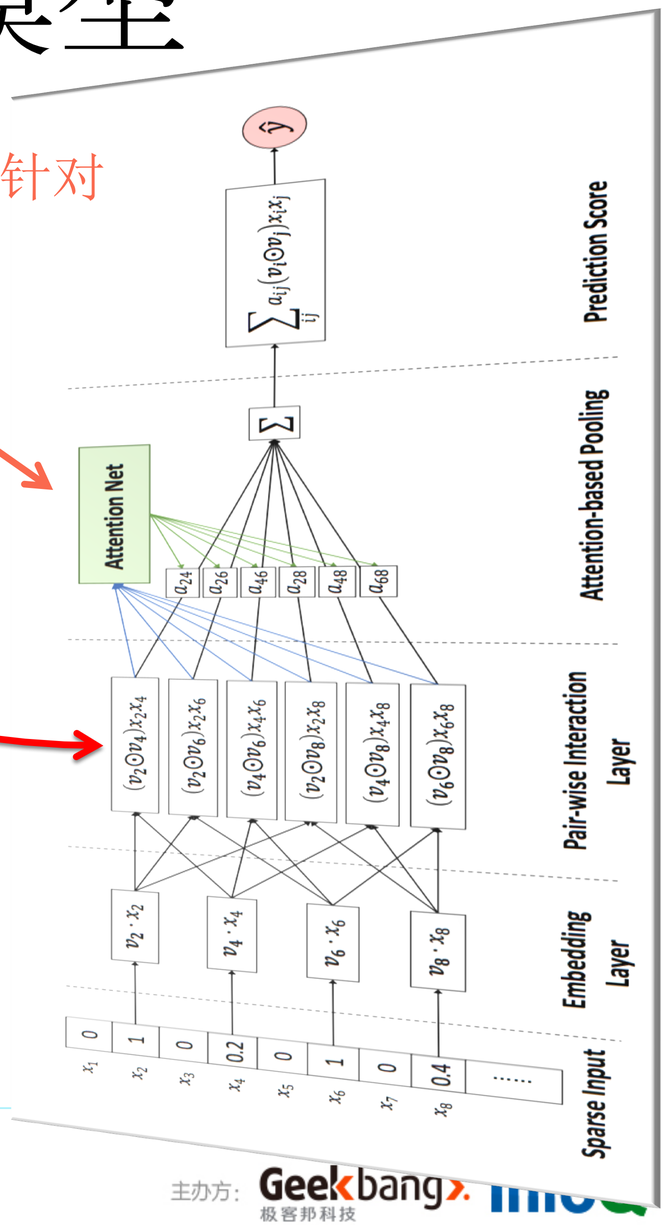




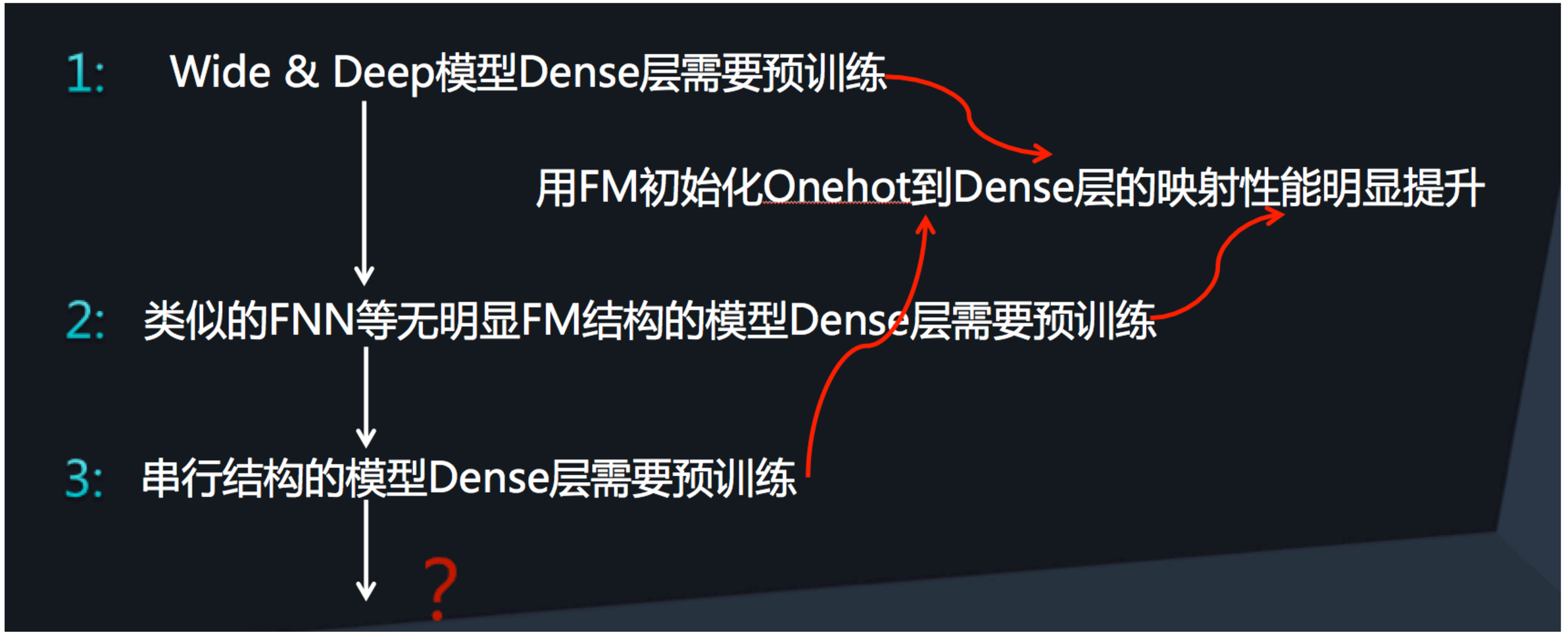
串行结构实例：AFM模型



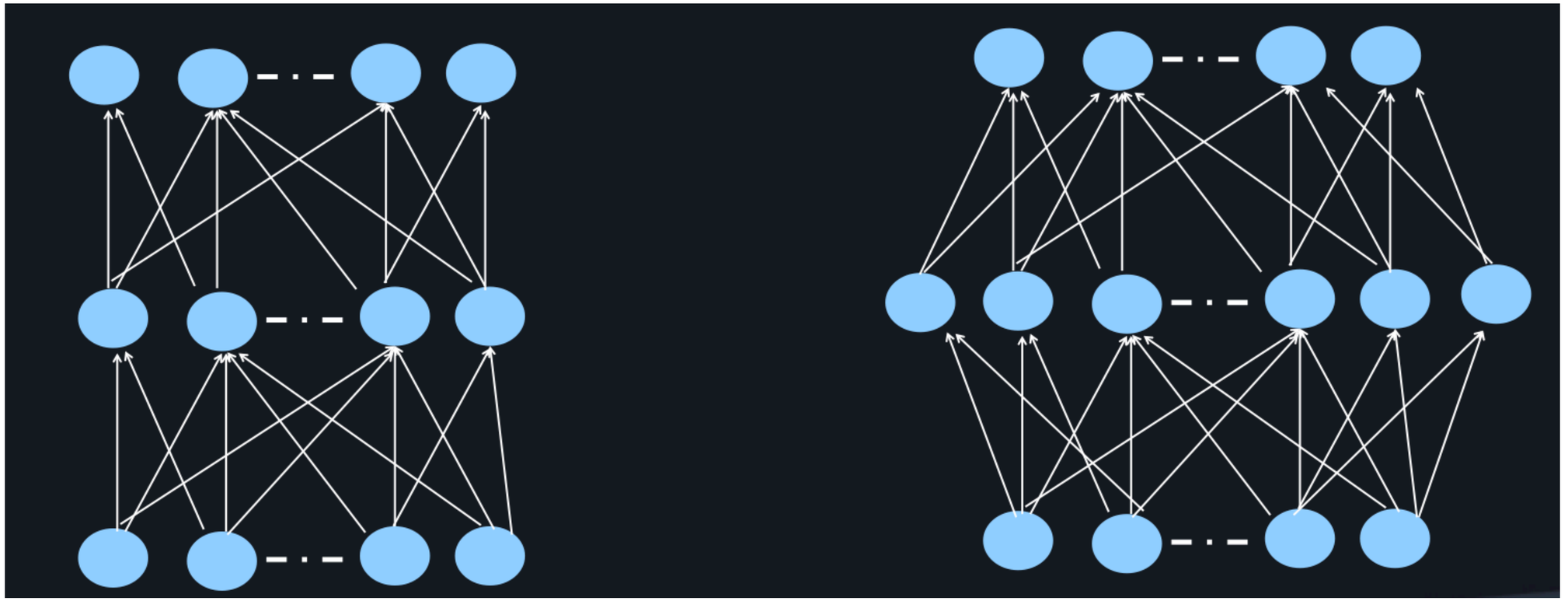
相比NFM模型只是多了针对组合特征的Attention



模型训练与优化：Dense层的预训练

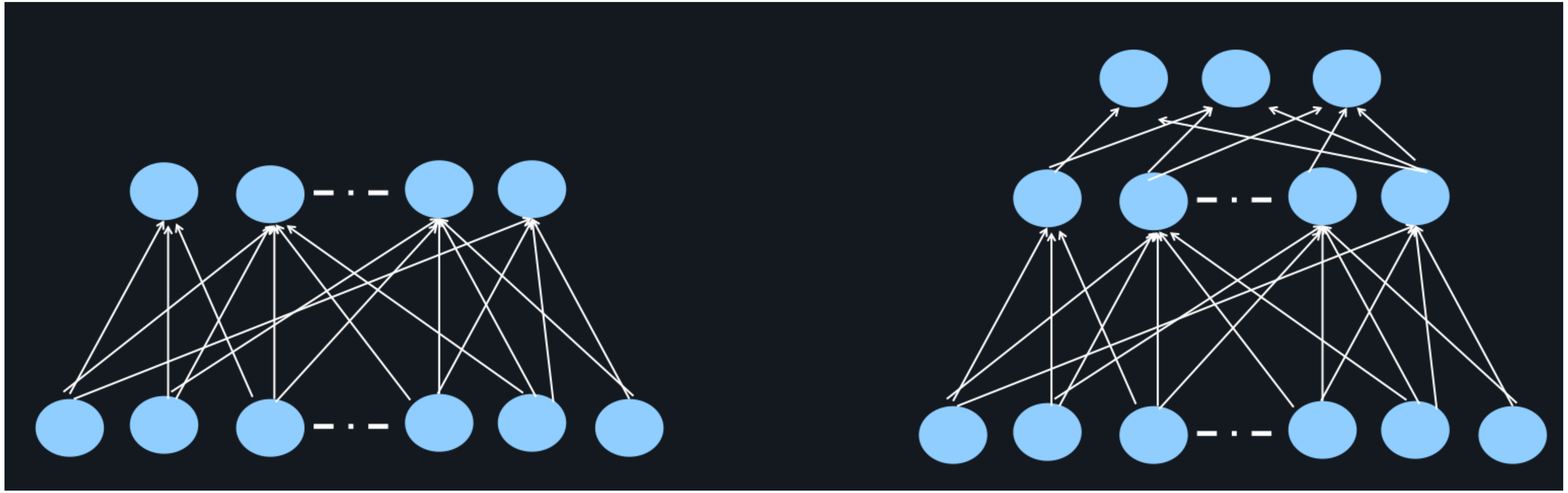


模型训练与优化：Deep网络隐层网络结构



平行结构或者菱形结构效果较好

模型训练与优化：Deep网络隐层的层深



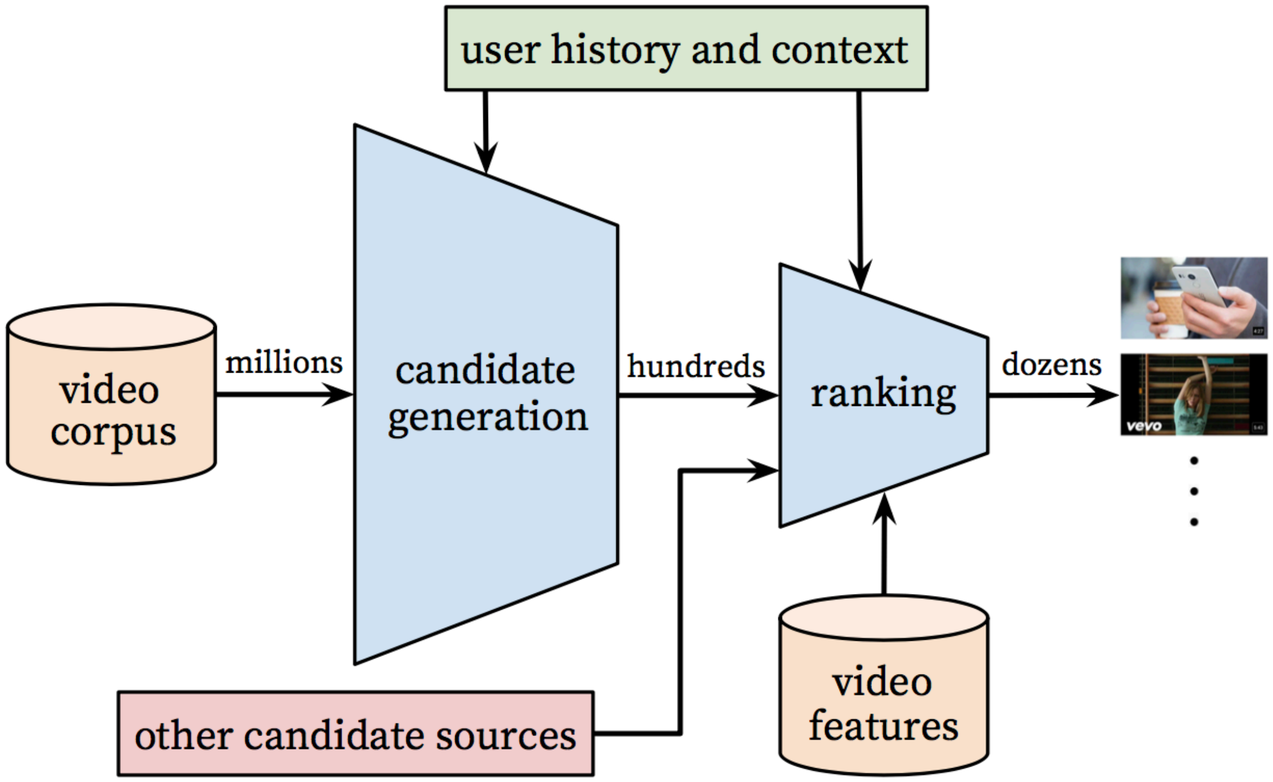
两层或三层

TABLE OF CONTENTES

Google：Youtube视频推荐

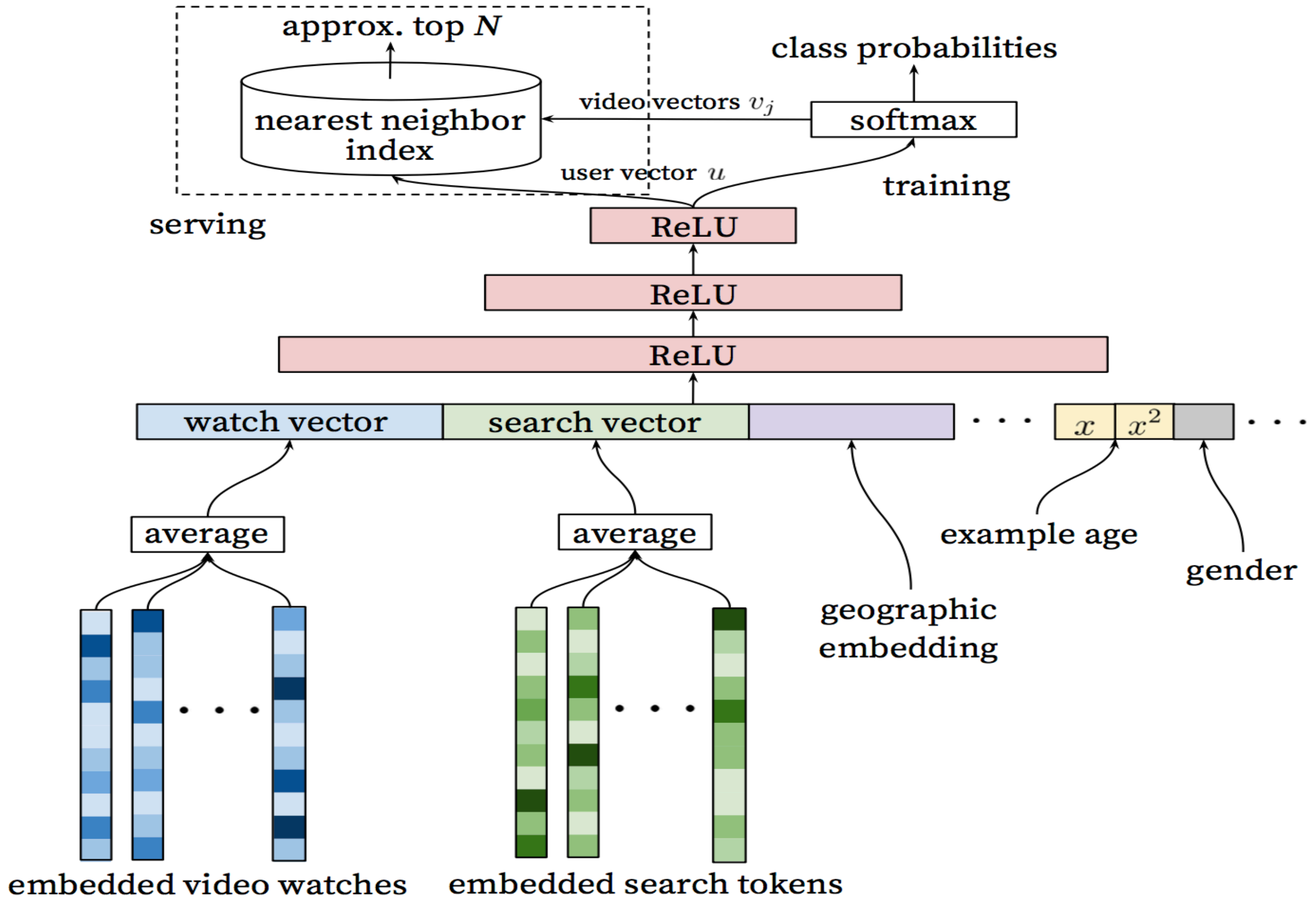
- Google 2016年工作
  - 论文：Deep Neural Networks for YouTube Recommendations
  - 已上线，效果优于传统模型
- 主要关注
  - 视频推荐
  - 深度神经网络的应用

Google：Youtube视频推荐



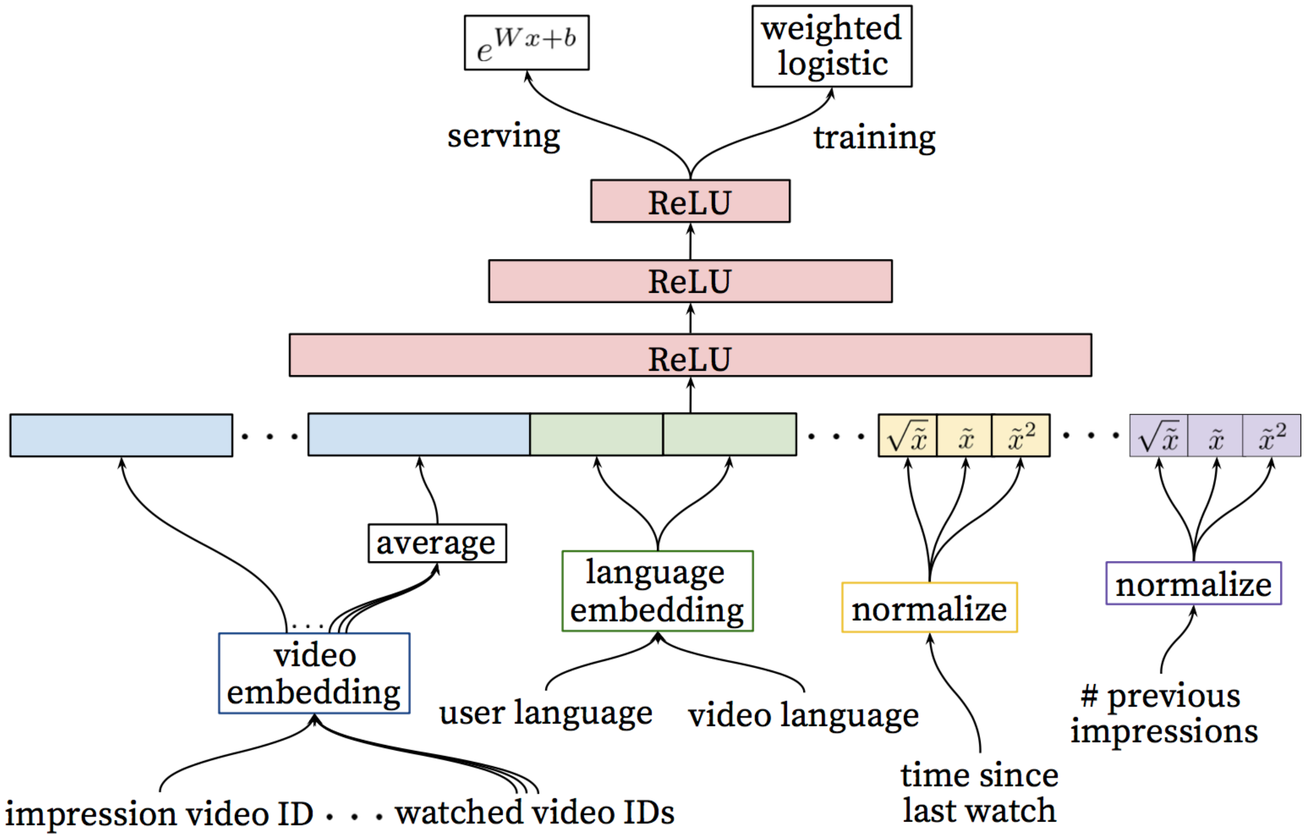
推荐系统架构

Google：Youtube视频推荐



训练阶段

Google：Youtube视频推荐



在线服务阶段

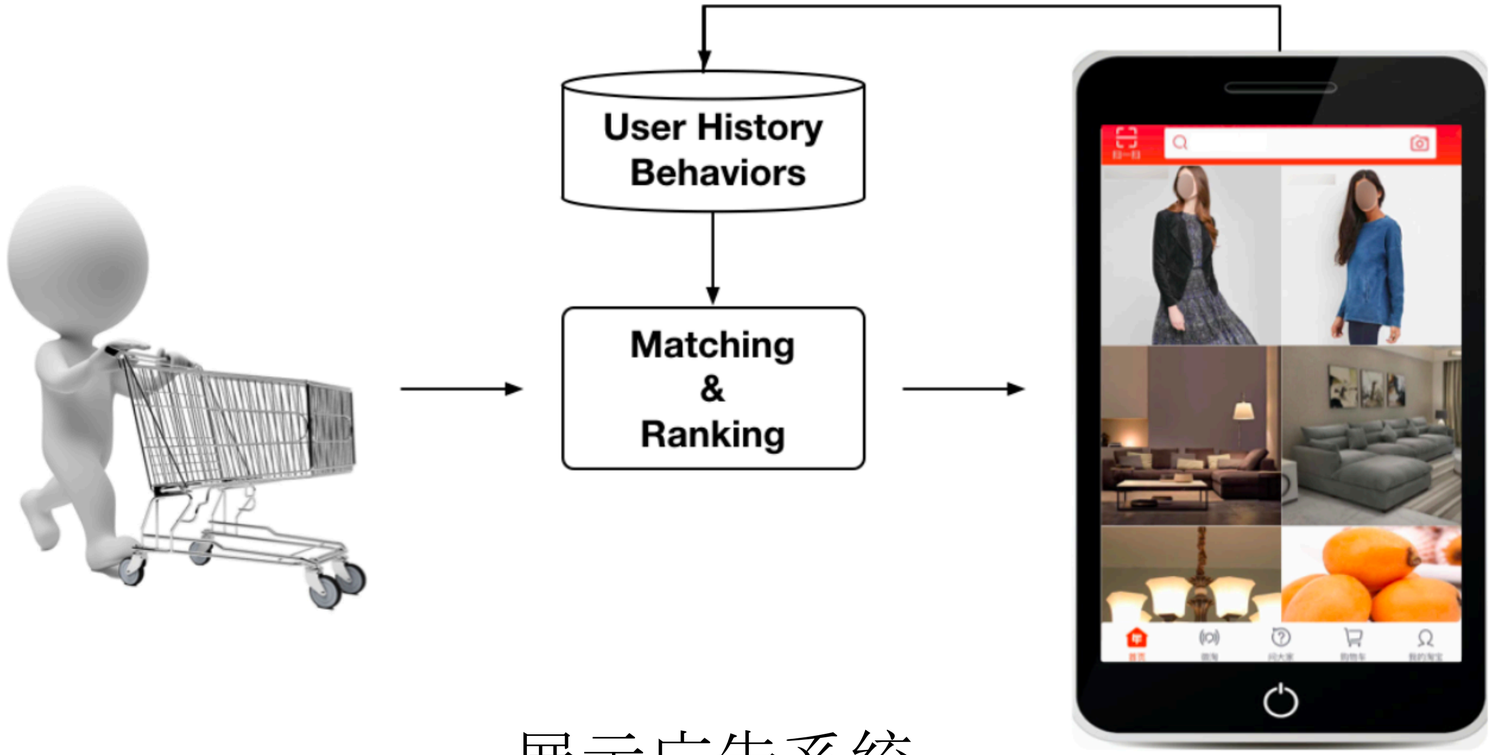
Google：Youtube视频推荐

Hidden layers	weighted, per-user loss
None	41.6%
256 ReLU	36.9%
512 ReLU	36.7%
1024 ReLU	35.8%
512 ReLU → 256 ReLU	35.2%
1024 ReLU → 512 ReLU	34.7%
1024 ReLU → 512 ReLU → 256 ReLU	34.6%

阿里巴巴：Deep Interest Network

- 阿里妈妈2017年工作
  - 论文：Deep Interest Network for Click-Through Rate Prediction
  - 已上线，效果明显优于传统模型
- 主要关注
  - 兴趣的多样性
  - 局部激活（Local Activation）

阿里巴巴：Deep Interest Network



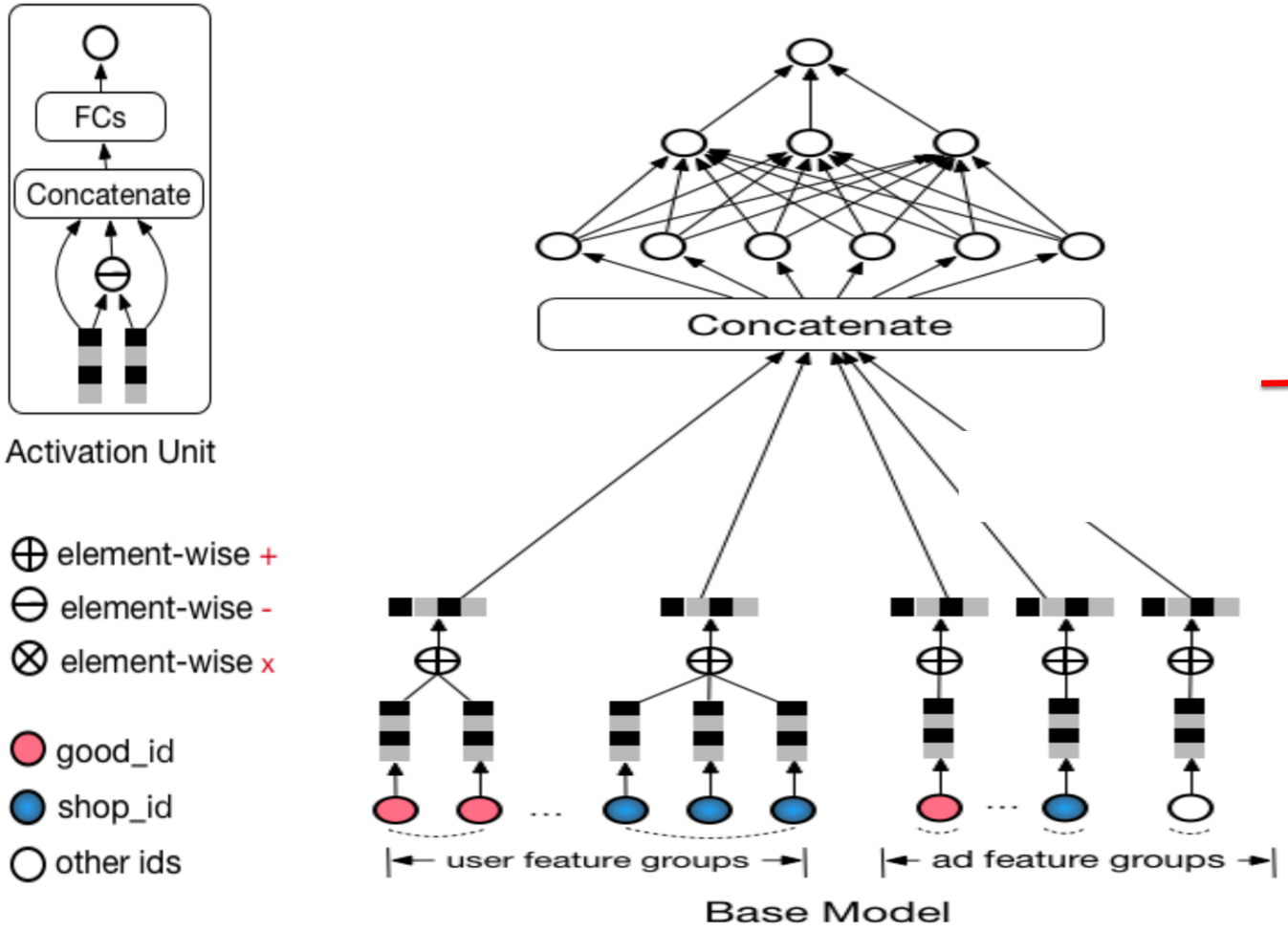
展示广告系统

阿里巴巴：Deep Interest Network

Feature Category	Feature Name	Dimemsion	Type	#Nonzero Ids/Sample
User Profile Features	gender	2	one-hot	1
	age_level	~ 10	one-hot	1
	...	...	...	...
User Behavior Features	visited good_ids	$\sim 10^9$	multi-hot	$\sim 10^3$
	visited shop_ids	$\sim 10^7$	multi-hot	$\sim 10^3$
	visited cate_ids	$\sim 10^4$	multi-hot	$\sim 10^2$
	...	...	...	...
Ad Features	good_id	$\sim 10^7$	one-hot	1
	shop_id	$\sim 10^5$	one-hot	1
	cate_id	$\sim 10^4$	one-hot	1
	...	...	...	...
Scene Features	pid	~ 10	one-hot	1
	time	~ 10	one-hot	1
	...	...	...	...

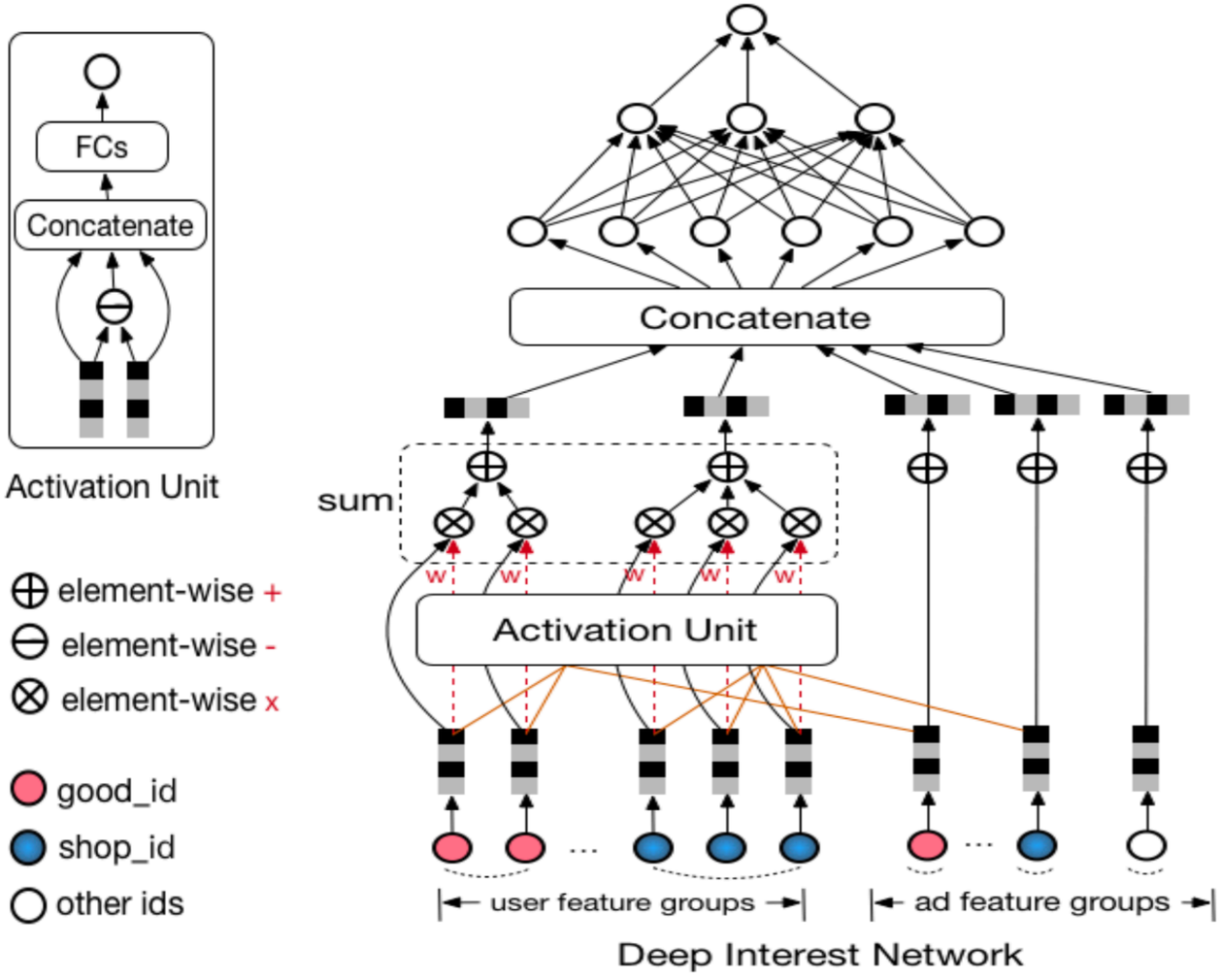
使用特征

阿里巴巴：Deep Interest Network



基准模型

阿里巴巴：Deep Interest Network



DIN模型

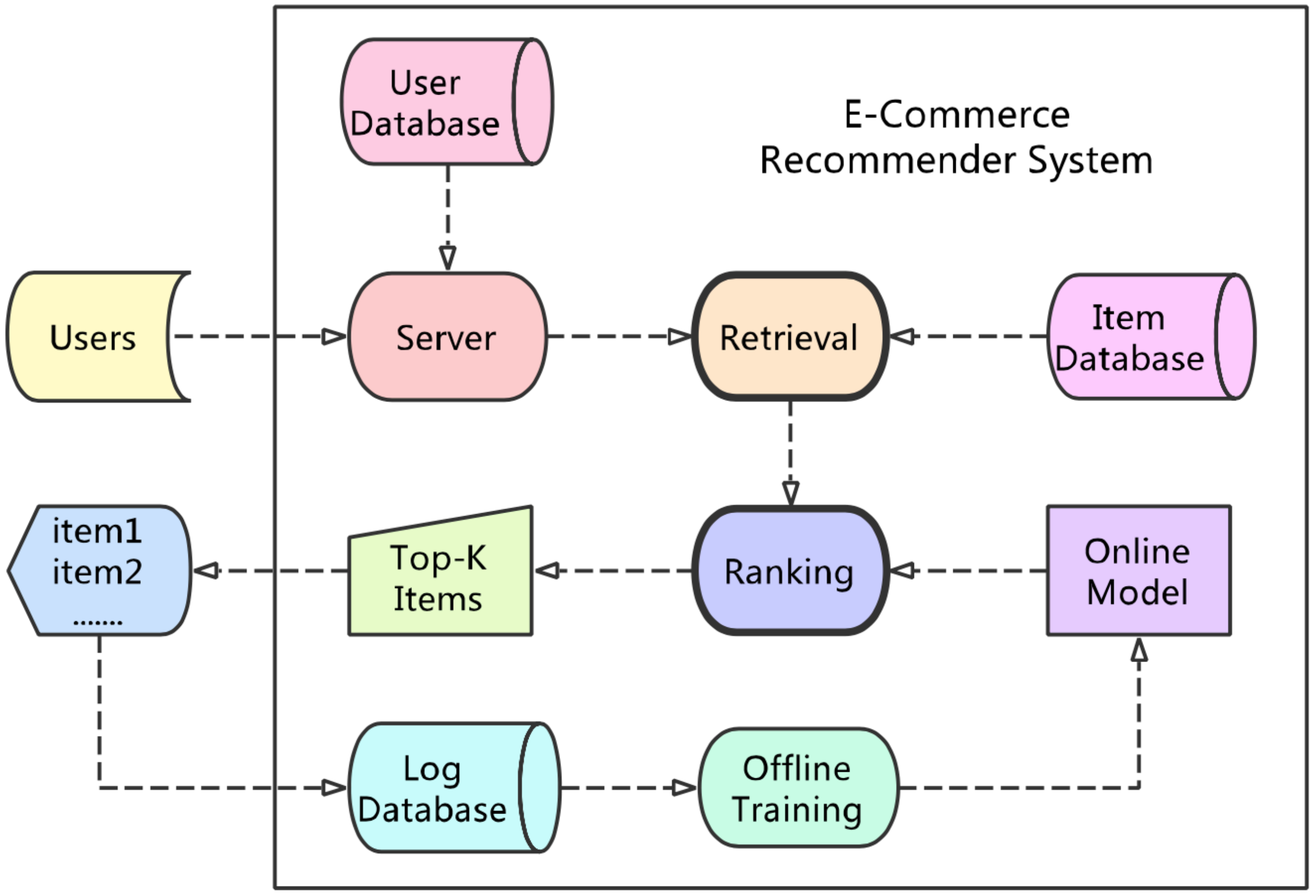
阿里巴巴：Deep Interest Network

Local Activation

京东商城：Telepath

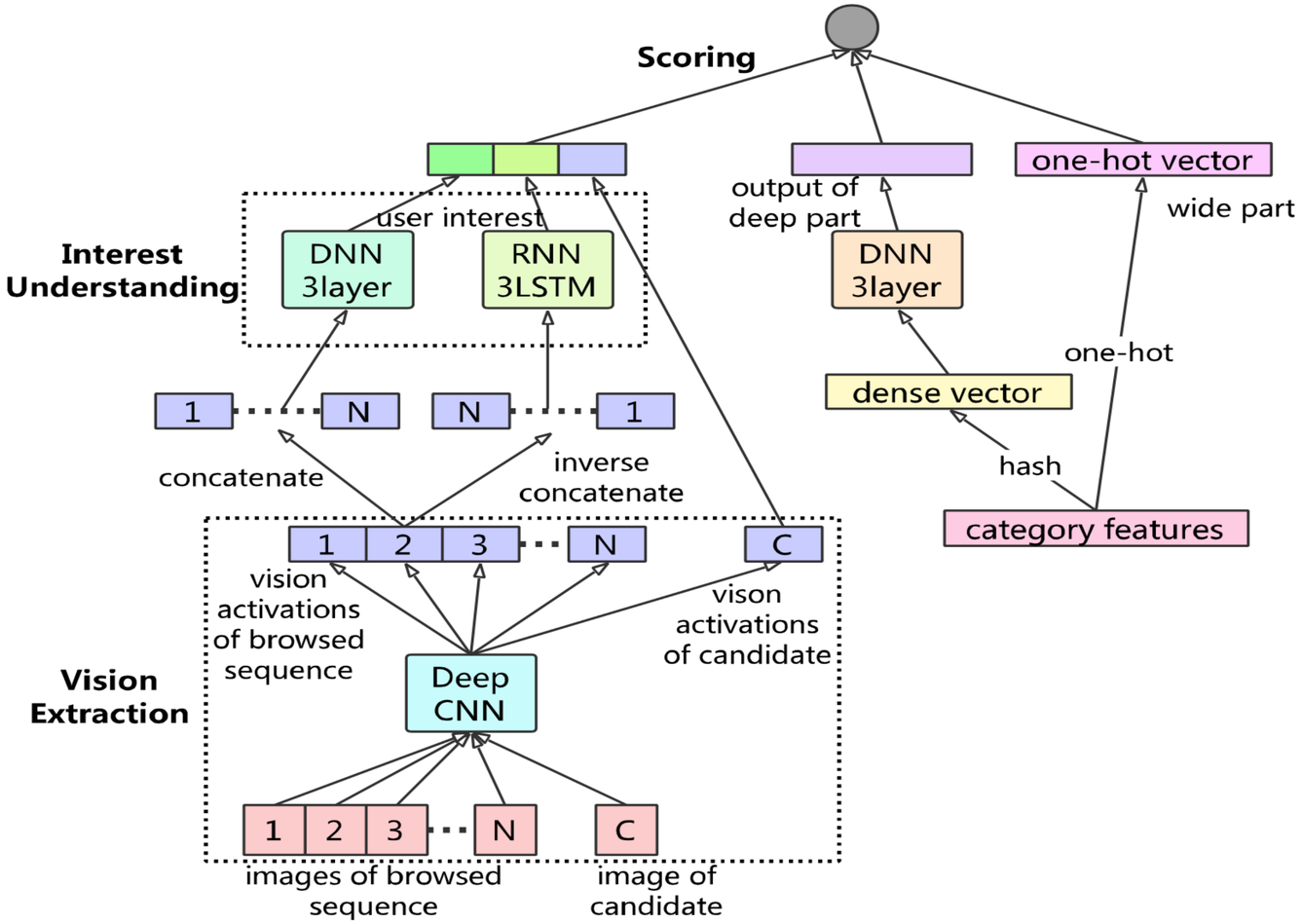
- 京东2017年工作
  - 论文：Telepath: Understanding Users from a Human Vision Perspective in Large-Scale Recommender Systems
  - 已上线，效果明显优于传统模型
- 主要关注
  - 用于推荐和广告
  - 融合图片和行为等各种信息

京东商城：Telepath



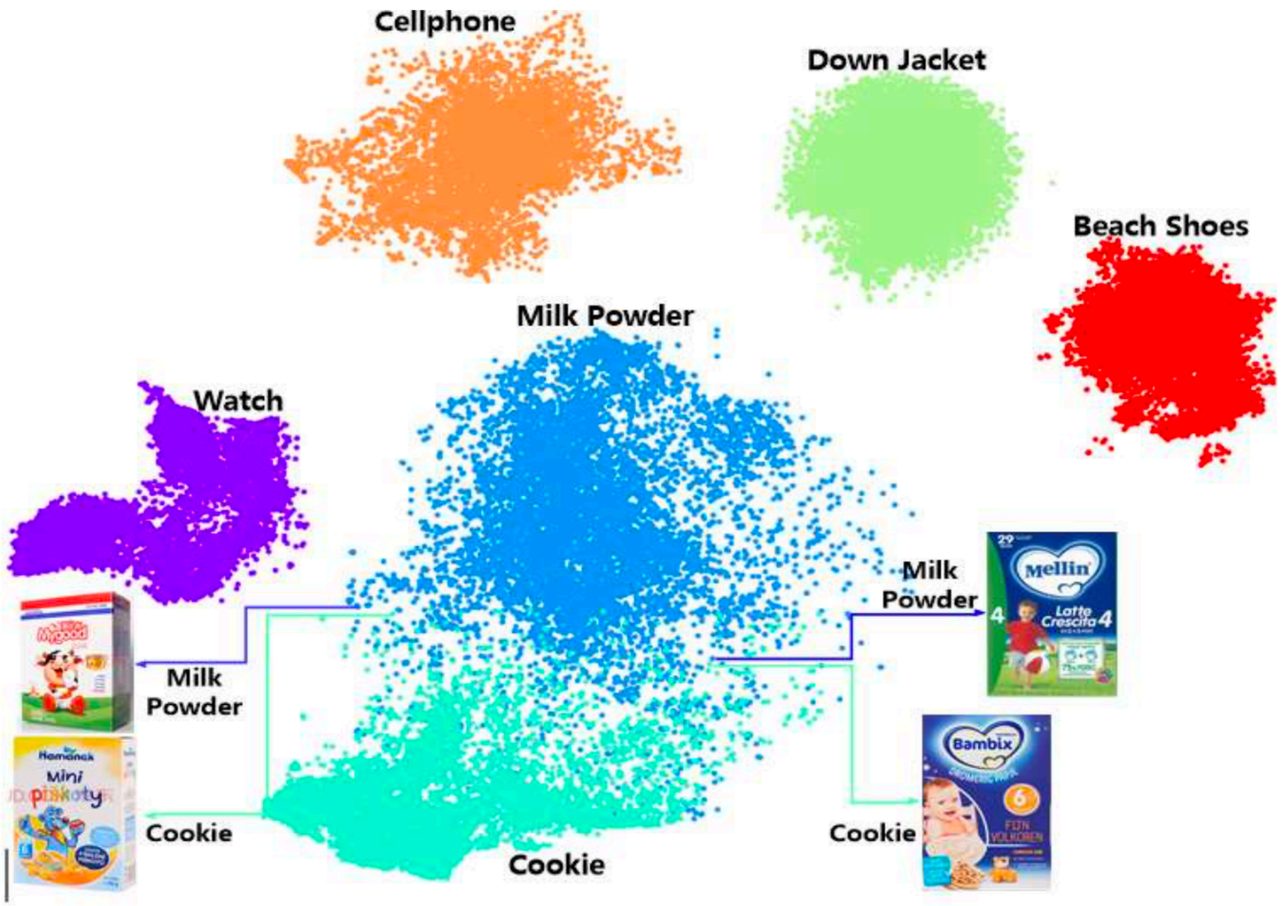
推荐系统架构

京东商城：Telepath



Telepath网络结构

京东商城：Telepath



视觉激活item可视化

京东商城：Telepath

Date	Day1	Day2	Day3	Day4
CTR	+5.15%	+8.07%	+10.5%	+6.15%
GMV	+ 126.48%	+9.1%	+18.4%	-19.24%
ROI	+129.53%	+14.35%	+14.2%	-17.44%
Date	**Day5**	**Day6**	**Day7**	**Average**
CTR	+4.63%	+2.11%	+9.48%	**+6.58%**
GMV	+8.53%	+143.09%	+ 145.74%	**+61.72%**
ROI	+9.17%	+161.36%	+147.79%	**+65.57%**

上线对比效果

Thanks!