Deep Learning over Multi-field Categorical Data

本文提出两种模型旨在解决现有模型需要手动生成交叉特征、难以充分学习到高阶交叉特征的问题。

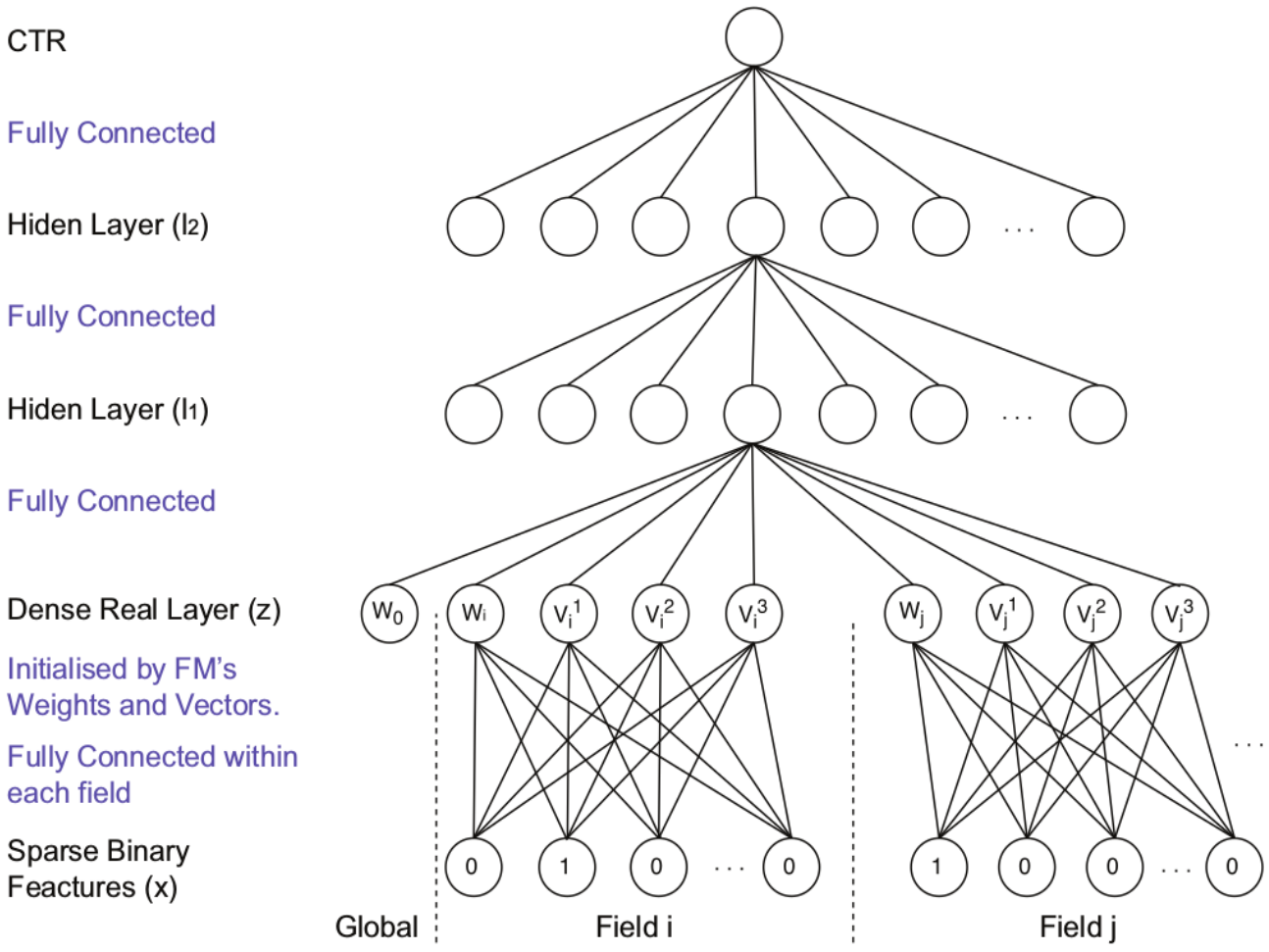
本文拟解决问题： **CTR**预估过程中高维离散属性带来的计算复杂问题

feature works:

- 使用 momentum 方法训练 DNN
- 本文FNN采用的 partial 连接方式可以尝试扩展到更高层，因为这种方式计算更简单，模型更鲁棒，更像人脑

Factorisation Machine supported Neural Network (FNN)

FNN 结构



Output Layer:

$$\hat{y} = sigmoid(W_3 l_2 + b_3)$$

Hidden Layer:

$$l_i = tanh(W_i l_{i-1} + b_i)$$

$$l_0 = (w_0, z_1, z_2, \ldots, z_n)$$

Dense Real Layer: 将高维稀疏向量转换为低维连续向量；

对于 $filed_i$, 若设定embedding的个数为k，W_0^i 的形状是 $(k+1, end_i - start_i)$, 则有:

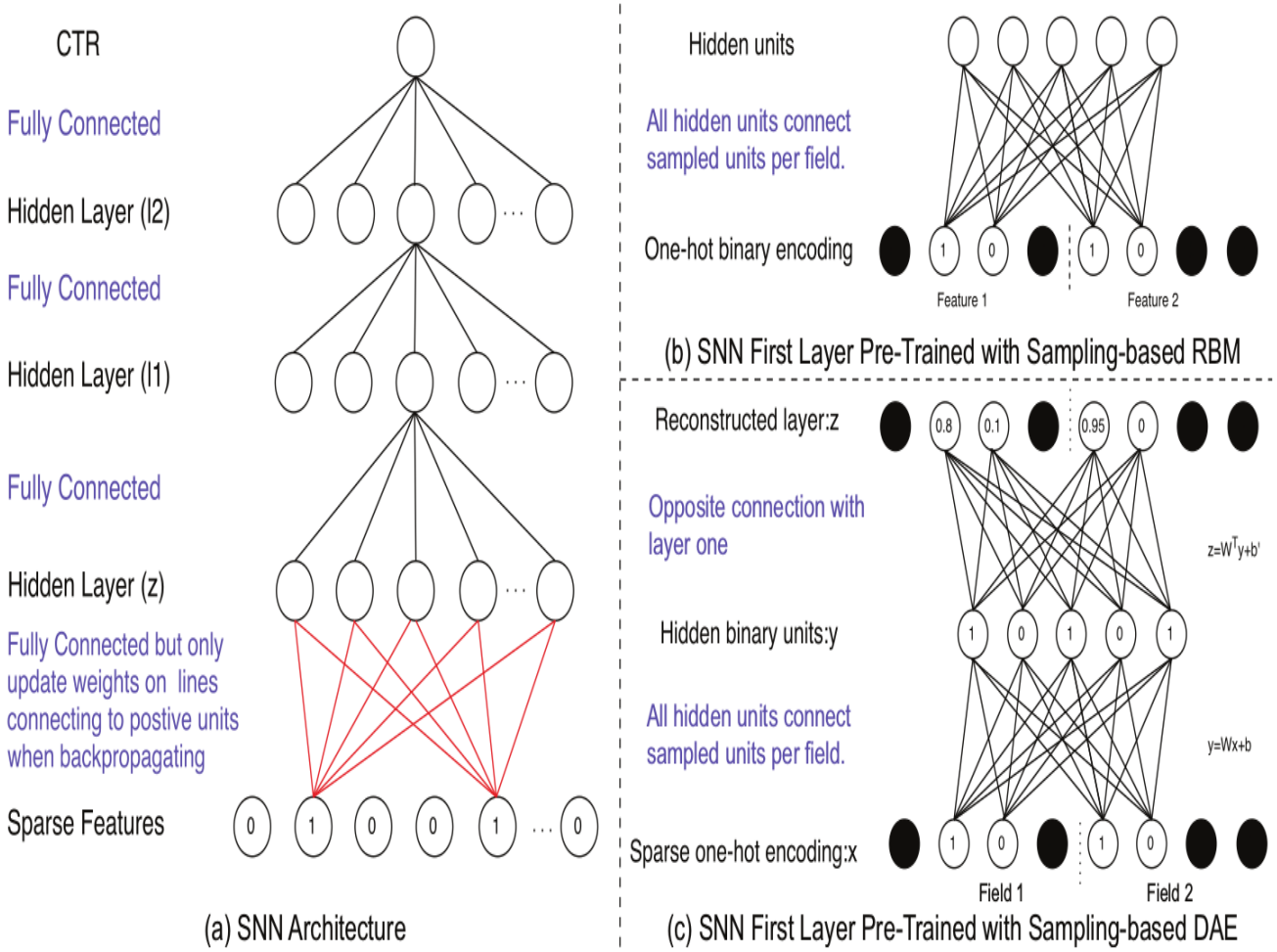
$$z_i = W_0^i \cdot x[start_i : end_i] = (w_i, v_i^1, v_i^2, \ldots, v_i^k)$$

FNN 训练流程

- 使用SGD训练FM模型，用FM的embedding初始化Dense Real Layer的权重;
- 采用基于 contrastive divergence 方法的 layer-wise RBM pre-traing 初始化Hidden Layer的参数;
- fine-tuning with back propagation

Sampling-based Neural Network (SNN)

SNN的结构



SNN与FNN的区别有两点：

- SNN的第一层是全连接的，而FNN的第一层是基于Field进行b部分连接的;
- SNN的第一层权重初始化采用 Sampling-based RBM/DAE， 而FNN的第一层权重初始化需要进行FM的预训练;

此外，本文 SNN 所谓 Sampling based 是指，选中 one hot 的向量中非零的一项， 再选中 m 个值为零的项参与训练，如此一来可以减小训练复杂度;

实验验证

采用 IPinYou 的数据，共计 19.50M 样例, 14.79K 正例，onehot后共有 937.67K 的 binary features.

与LR和FM进行对比，参数设置如下：

- 抑制 overfitting 措施
  - early stopping
  - 对比了 L2 regularisation 和 dropout, 其中 dropout 是最优的
- learning rate
  比较了[1, 0.1, 0.01, 0.001, 0.0001]
- Sample number
  比较了[1, 2, 4], m=2是最好的
- active function
  比较了[linear, sigmoid, tanh], tanh是最好的
- hidden layer structure
  - architecture: 比较了[diamond, constant, increasing, decreading], 其中 diamod(200,300,100) 是最好的;