DEEP & CROSS NETWORK FOR AD CLICK PREDICTIONS

Ruoxi Wang
Stanford University
AdKDD2017, Halifax

Joint work with

Bin Fu (Google), Gang Fu (Google) and Mingliang Wang (Google, summer intern host)

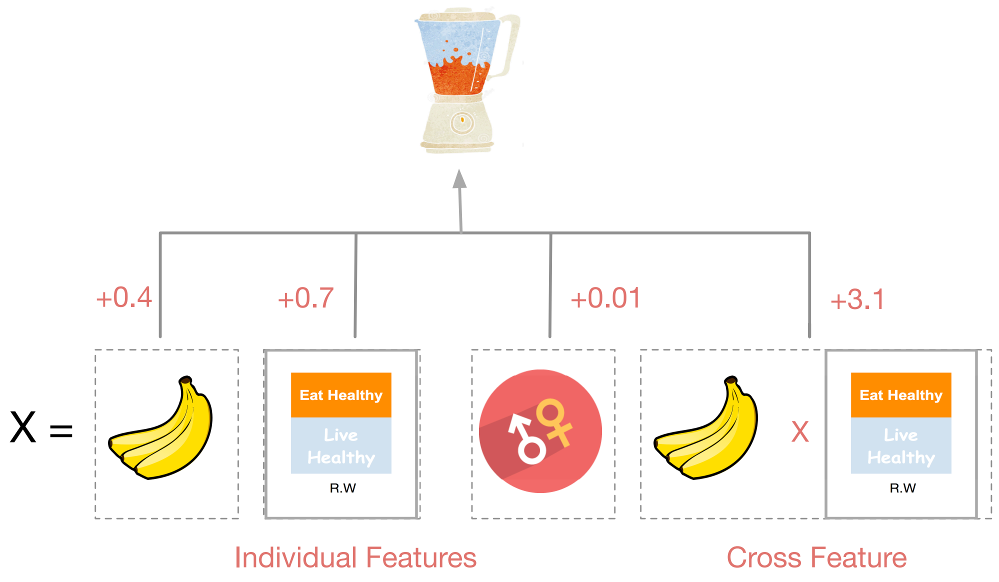
OUTLINE

- **Introduction**
- **Deep & Cross Network (DCN)**
- **Experimental Results**
- **Cross Network Analysis**

INTRODUCTION

PROBLEM AND CHALLENGE

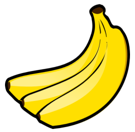
- **Goal**
  - Input $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, $\boldsymbol{x}_i$: data, y_i: label
  - Predict Ad click-through rate (CTR) accurately
- **Key**
  - Identify predictive feature crossings
  - Explore rare or unseen features
- **Challenge**
  - Large and sparse feature space
  - Manual feature engineering



RELATIONS TO EXISTING WORK

- Factorization Machines (FMs) [Rendle et al, 2010]
- Deep Crossing (DC) [Shan et al, 2016]
- Wide-and-Deep (W&D) [Cheng et al, 2016]

FM	DCN (our model)
☺ Handles sparse input	☺
☺ Generalizes well	☺
☹ 2nd-order interactions	☺ higher-order interactions

$x =$ 

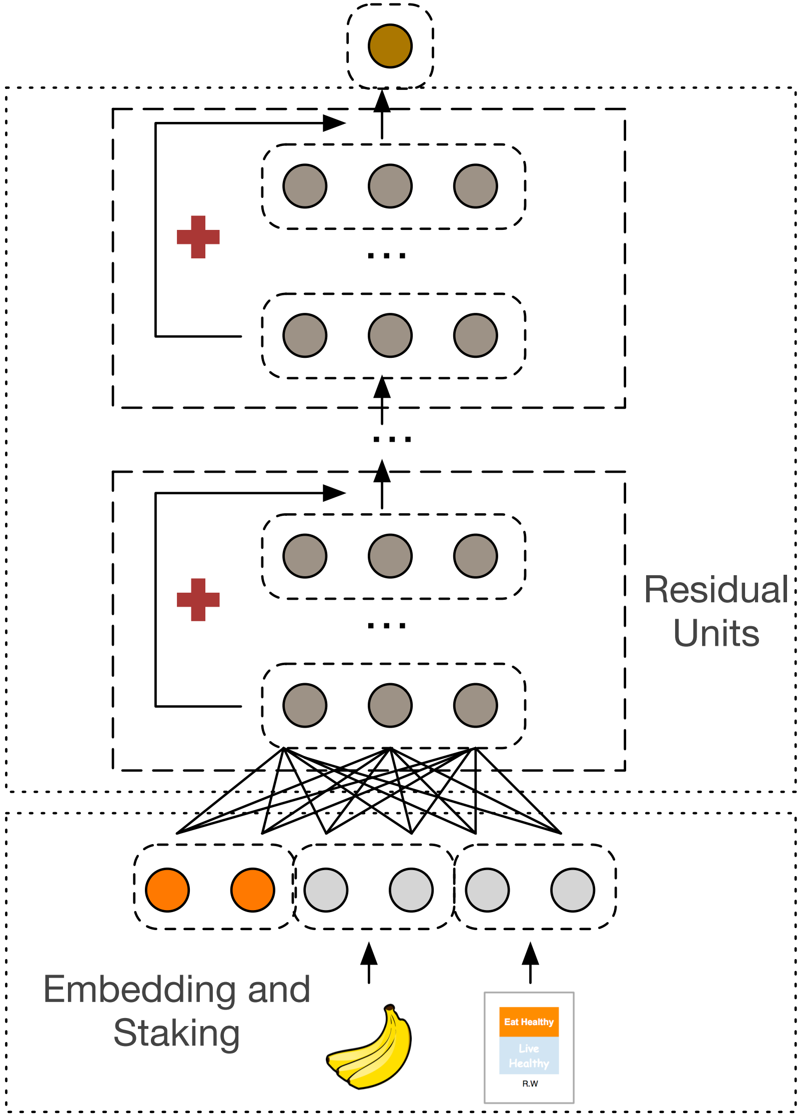
$$\mathbf{v} = \begin{bmatrix} 0 \\ 6 \\ 1 \\ 8 \end{bmatrix}$$

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

RELATIONS TO EXISTING WORK

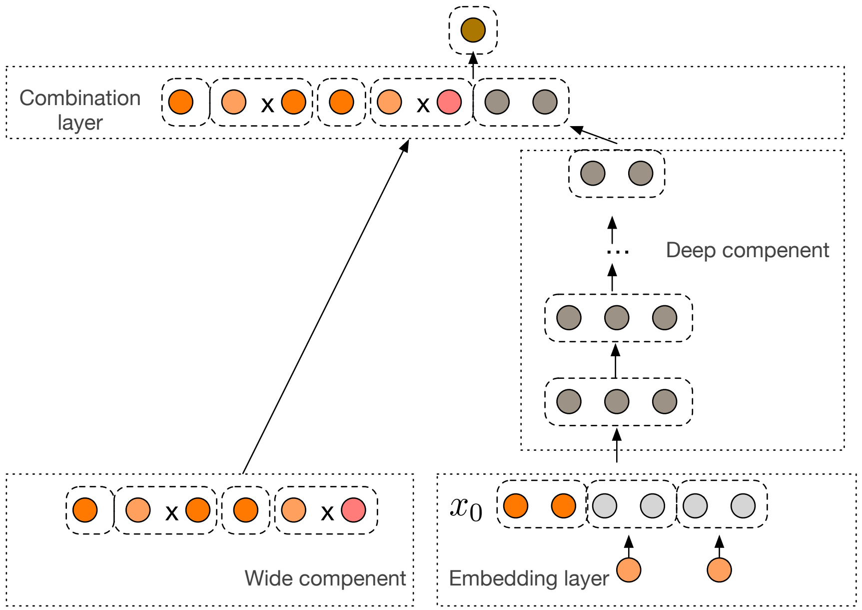
- Factorization Machines (FMs) [Rendle et al, 2010]
- Deep Crossing (DC) [Shan et al, 2016]
- Wide-and-Deep (W&D) [Cheng et al, 2016]

DC (and DNN-based model)	DCN (our model)
☺ Complex interactions	☺
☺ $\forall\ (smooth)\ f, \forall \epsilon, \| \hat{f} - f \| < \epsilon$	☺
☹ Implicit crossing: linear + ReLu (or Sigmoid)	☺ Explicit & bounded crossing: e.g, $x_1x_2,\ x_1x_3x_4$



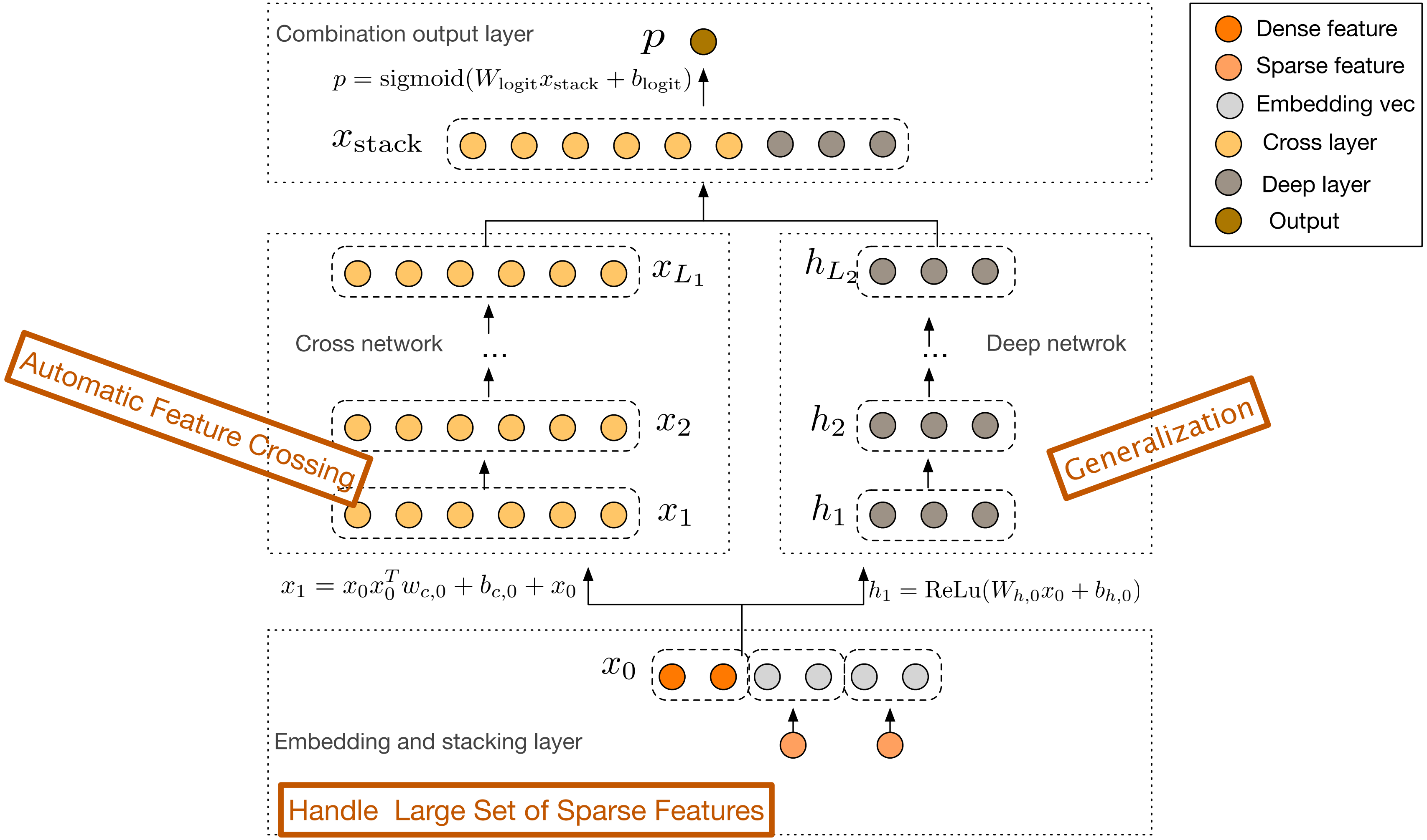
RELATIONS TO EXISTING WORK

- Factorization Machines (FMs) [Rendle et al, 2010]
- Deep Crossing (DC) [Shan et al, 2016]
- Wide-and-Deep (W&D) [Cheng et al, 2016]



W&D	DCN (our model)
☺ Memorization + Generalization	☺
☹ No efficient method to select cross features	☺ Automatic + efficient

DEEP & CROSS NETWORK (DCN)



DCN: ARCHITECTURE & ADVANTAGES

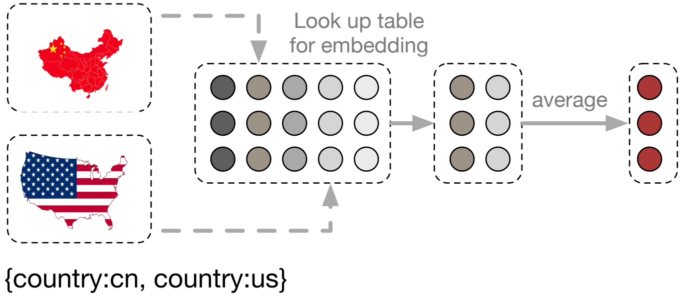
- Joint training
- No need of manual feature engineering

DCN: EMBEDDING AND STACKING

- Inputs are mostly categorical features
  (e.g. "country=usa")
- One-hot vector encoding
  (e.g. "[0,1,0]")
- Leads to excessively high-dimensional feature spaces
- Input of our model:
  output from embedding

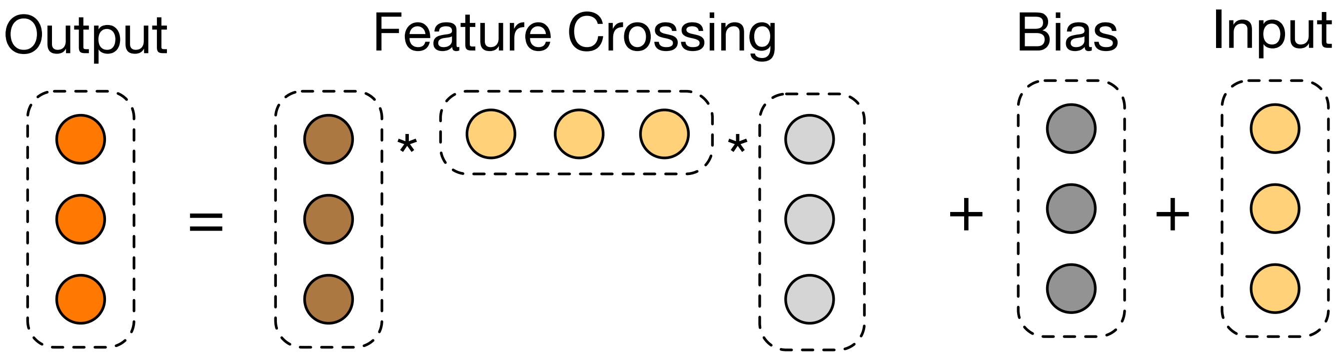
Low dimensional embedding

$$\mathbf{x}_{\text{embed},i} = W_{\text{embed},i}\mathbf{x}_i$$



Stacking

$$\mathbf{x}_0 = \left[\mathbf{x}^T_{\text{embed},1}, \ldots, \mathbf{x}^T_{\text{embed},k}, \mathbf{x}^T_{\text{dense}}\right]$$



$$y = \underbrace{x_0 * x' * w}_{=\begin{bmatrix} x_1\tilde{x}_1 & x_1\tilde{x}_2 & \dots & x_1\tilde{x}_d \\ x_2\tilde{x}_1 & x_2\tilde{x}_2 & \dots & x_2\tilde{x}_d \\ \vdots & \vdots & \ddots & \vdots \\ x_d\tilde{x}_1 & x_d\tilde{x}_2 & \dots & x_d\tilde{x}_d \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}} + b + x$$

DCN: CROSS NETWORK

$$\boldsymbol{x_{l+1}} = \boldsymbol{x_0 x_l^T w_l} + \boldsymbol{b_l} + \mathbf{x_l}$$

- Generate all d^2 cross pairs
- $d^2 \rightarrow d$ by an immediate embedding
- Optimization selects informative crossings
- Residual network

EXPERIMENTAL RESULTS

CRITEO DISPLAY ADS DATA

- 13 integer features and 26 categorical features
- 11 GB user logs from a period of 7 days (~ 41 million records)
- Improvement of 0.001 in logloss is considered as practically significant

Best test logloss from different models

Model	DCN	DC	DNN	FM	LR
Logloss	**0.4419**	0.4425	0.4428	0.4464	0.4474

✓ **Outperforms DNN with 60% less memory!**

- DC: deep crossing (the same embedding (stacking) layer as DCN)
- DNN: deep neural network (the DCN model with no cross network)
- FM: factorization machine based model (proprietary details)
- LR: logistic regression (all single features + carefully selected cross features)

COMPARISON: DCN & DNN (CRITEO)

#parameters needed to achieve a desired logloss

Logloss	0.4430	0.4460	0.4470	0.4480
DCN	**7.9E+05**	**7.3E+04**	**3.7E+04**	**3.7E+04**
DNN	3.2E+06	1.5E+05	1.5E+05	7.8E+04

✓ ~ an order of magnitude more memory efficient!

Best logloss achieved with various memory budgets

#params	5.0E+04	1.0E+05	4.0E+05	1.1E+06	2.5E+06
DCN	**0.4465**	**0.4453**	**0.4432**	**0.4426**	**0.4423**
DNN	0.4480	0.4471	0.4439	0.4433	0.4431

✓ consistently outperforms!

✓ captured meaningful feature interactions!

NON-CTR (DENSE) DATASETS

Forest datatype

(581012 samples and 54 features)

Model	DCN	DNN	DC
Accuracy	**0.9740**	0.9737	0.9737

✓ **Performs well on non-CTR data!**

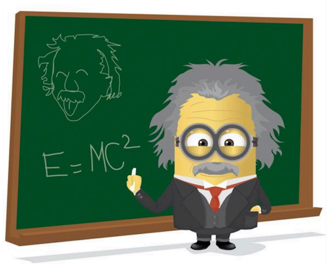
Higgs

(11M samples and 28 features)

Model	DCN	DNN
Logloss	**0.4494**	0.4506

✓ **DCN outperforms with 50% of the memory used in DNN!**

CROSS NETWORK ANALYSIS



DCN: CROSS NETWORK ANALYSIS

- Consider an l- layer cross network
- Our effective hypothesis functions live in the space of degree $l + 1$ polynomials
- We use only O(d) parameters to characterize them

DCN: CROSS NETWORK ANALYSIS

- $P_n(x) = \left\{\sum_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d} \mid 0 \le |\boldsymbol{\alpha}| \le n, \boldsymbol{\alpha} \in N^d\right\}$; $O(d^n)$ parameters

DCN: CROSS NETWORK ANALYSIS

- $P_n(x) = \{\sum_{\alpha} w_{\alpha} x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d} \mid 0 \le |\boldsymbol{\alpha}| \le n, \boldsymbol{\alpha} \in N^d\}$; $O(d^n)$ parameters
- $\boldsymbol{x_{i+1}} = \boldsymbol{x_0 x_i^T w_i} + \mathbf{x}_{\boldsymbol{i}}$; Input: $\boldsymbol{x}_0 = [x_1, x_2, \ldots, x_d]^T$; Output: $g_l(\boldsymbol{x}_0) = \boldsymbol{x}_l^T w_l$
- Explicitly applies feature crossing at each layer, and reproduces:

$$\left\{ \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}(\mathbf{w}_0, \ldots, \mathbf{w}_l) x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d} \,\middle|\, 0 \le |\boldsymbol{\alpha}| \le l+1, \boldsymbol{\alpha} \in \mathbb{N}^d \right\}$$

DCN: CROSS NETWORK ANALYSIS

- $P_n(x) = \{\sum_\alpha w_\alpha x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \mid 0 \le |\boldsymbol{\alpha}| \le n, \boldsymbol{\alpha} \in N^d\}$; $O(d^n)$ parameters
- $\boldsymbol{x}_{i+1} = \boldsymbol{x}_0 \boldsymbol{x}_i^T \boldsymbol{w}_i + \mathbf{x}_i$; Input: $\boldsymbol{x}_0 = [x_1, x_2, \dots, x_d]^T$; Output: $g_l(\boldsymbol{x}_0) = \boldsymbol{x}_l^T w_l$
- Explicitly applies feature crossing at each layer, and reproduces:

$$\left\{ \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}(\mathbf{w}_0, \dots, \mathbf{w}_l) \boxed{x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}} \,\middle|\, 0 \le |\boldsymbol{\alpha}| \le l+1, \boldsymbol{\alpha} \in \mathbb{N}^d \right\}$$

✓ *cross term of degree* $|\boldsymbol{\alpha}| = \sum_i \alpha_i$

DCN: CROSS NETWORK ANALYSIS

- $P_n(x) = \{\sum_{\alpha} w_{\alpha} x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \mid 0 \le |\boldsymbol{\alpha}| \le n, \boldsymbol{\alpha} \in N^d\}$; $O(d^n)$ parameters
- $\boldsymbol{x}_{i+1} = \boldsymbol{x}_0 \boldsymbol{x}_i^T \boldsymbol{w}_i + \mathbf{x}_i$; Input: $\boldsymbol{x}_0 = [x_1, x_2, \dots, x_d]^T$; Output: $g_l(\boldsymbol{x}_0) = \boldsymbol{x}_l^T w_l$
- Explicitly applies feature crossing at each layer, and reproduces:

$$\left\{ \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}(\mathbf{w}_0, \dots, \mathbf{w}_l) \boxed{x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}} \;\middle|\; \boxed{0 \le |\boldsymbol{\alpha}| \le l+1}, \boldsymbol{\alpha} \in \mathbb{N}^d \right\}$$

✓ *cross term of degree $|\boldsymbol{\alpha}| = \sum_i \alpha_i$*

✓ *all cross terms of degree $0 \sim l+1$*

DCN: CROSS NETWORK ANALYSIS

- $P_n(x) = \{\sum_{\alpha} w_{\alpha} x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \mid 0 \le |\boldsymbol{\alpha}| \le n, \boldsymbol{\alpha} \in N^d\}$; $O(d^n)$ parameters
- $\boldsymbol{x}_{i+1} = \boldsymbol{x}_0 \boldsymbol{x}_i^T \boldsymbol{w}_i + \mathbf{x}_i$; Input: $\boldsymbol{x}_0 = [x_1, x_2, \dots, x_d]^T$; Output: $g_l(\boldsymbol{x}_0) = \boldsymbol{x}_l^T w_l$
- Explicitly applies feature crossing at each layer, and reproduces:

$$\left\{ \sum_{\boldsymbol{\alpha}} \boxed{c_{\boldsymbol{\alpha}}(\mathbf{w}_0, \dots, \mathbf{w}_l)} \boxed{x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}} \;\middle|\; \boxed{0 \le |\boldsymbol{\alpha}| \le l+1}, \boldsymbol{\alpha} \in \mathbb{N}^d \right\}$$

✓ *cross term of degree $|\boldsymbol{\alpha}| = \sum_i \alpha_i$*

✓ *O(d) parameters*

✓ *all cross terms of degree $0 \sim l + 1$*

DCN: CROSS NETWORK ANALYSIS

- $P_n(x) = \{\sum_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \mid 0 \le |\boldsymbol{\alpha}| \le n, \boldsymbol{\alpha} \in N^d\}$; $O(d^n)$ parameters
- $\boldsymbol{x}_{i+1} = \boldsymbol{x}_0 \boldsymbol{x}_i^T \boldsymbol{w}_i + \mathbf{x}_i$; Input: $\boldsymbol{x}_0 = [x_1, x_2, \dots, x_d]^T$; Output: $g_l(\boldsymbol{x}_0) = \boldsymbol{x}_l^T w_l$
- Explicitly applies feature crossing at each layer, and reproduces:

$$\left\{ \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}(\mathbf{w}_0, \dots, \mathbf{w}_l) x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \;\middle|\; 0 \le |\boldsymbol{\alpha}| \le l+1, \boldsymbol{\alpha} \in \mathbb{N}^d \right\}$$

✓ $\boldsymbol{\alpha} \ne \boldsymbol{\beta} \Rightarrow c_{\boldsymbol{\alpha}} \ne c_{\boldsymbol{\beta}}$

✓ cross term of degree $|\boldsymbol{\alpha}| = \sum_i \alpha_i$

✓ O(d) parameters

✓ all cross terms of degree $0 \sim l+1$

$$e.g., c_{\boldsymbol{\alpha}} = \sum_{i,j,k \in P_{\boldsymbol{\alpha}}} w_0^{(i)} w_1^{(j)} w_3^{(k)} + w_0^{(i)} w_2^{(j)} w_3^{(k)} + w_1^{(i)} w_2^{(j)} w_3^{(k)} \quad (l = 3)$$

RECAP

Proposed the DCN that

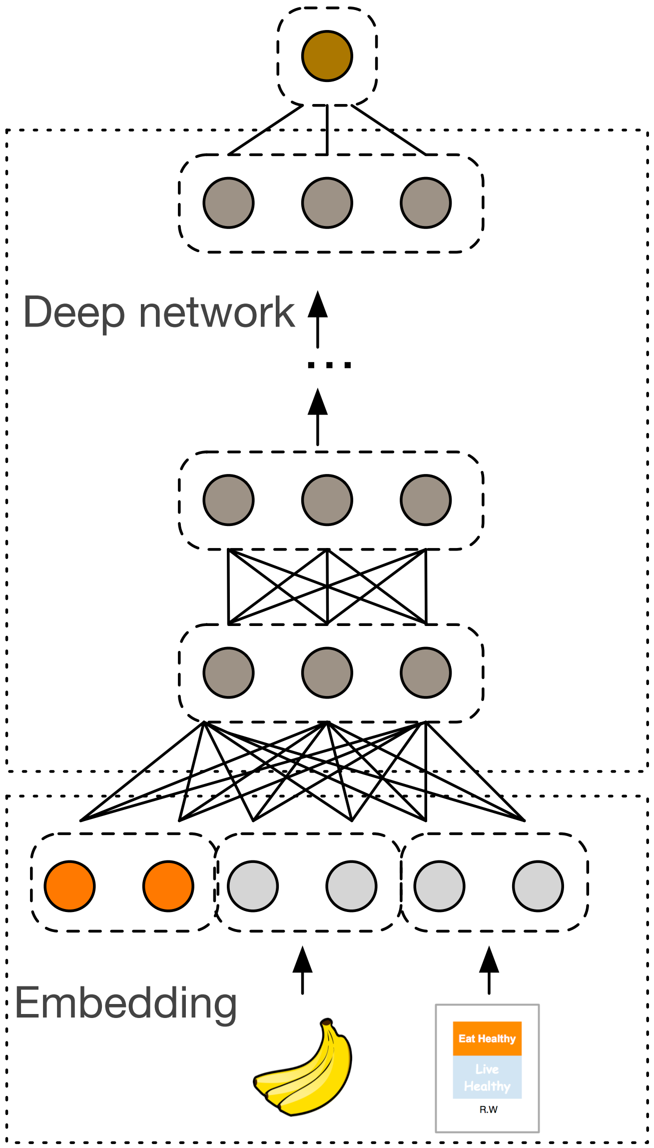
- handles a large set of sparse and dense features
- learns explicit cross features of bounded degree jointly with traditional deep representations
- delivers state-of-the-art performance on Criteo CTR dataset, in terms of both model accuracy and memory usage

DEEP & CROSS NETWORK FOR AD CLICK PREDICTIONS

Ruoxi Wang, Bin Fu, Gang Fu, Mingliang Wang

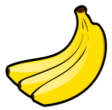
RELATIONS TO EXISTING WORK

- Factorization Machines (FMs) [Rendle et al, 2010]
- Deep Neural Networks (DNNs)
- Deep Crossing (DC) [Shan et al, 2016]
- Wide-and-Deep Model (W&D) [Cheng et al, 2016]



RELATED WORK

- Factorization Machines (FMs) [Rendle et al, 2010]
- Field-aware Factorization Machines (FFMs) [Juan et al, 2016]
- Deep Crossing (DC) [Shan et al, 2016]
- Wide-and-Deep Model (W&D) [Cheng et al, 2016]

$x =$  $\longrightarrow \quad \mathbf{v}_{f_1} = \begin{bmatrix} 0 \\ 8 \\ 1 \\ 6 \end{bmatrix} \quad \mathbf{v}_{f_2} = \begin{bmatrix} 0 \\ 3 \\ 1 \\ 4 \end{bmatrix}$

$$\langle \mathbf{v}_{i,f_1}, \mathbf{v}_{j,f_2} \rangle x_i x_j$$

FORMULA FOR MONOMIAL COEFFICIENT

$$c_{\boldsymbol{\alpha}} = M_{\boldsymbol{\alpha}} \sum_{\mathbf{i} \in B_{\boldsymbol{\alpha}}} \sum_{\mathbf{j} \in P_{\boldsymbol{\alpha}}} \prod_{k=1}^{|\boldsymbol{\alpha}|} w_{i_k}^{(j_k)}$$

- M_{α} is a constant independent of $\boldsymbol{w}_i$'s
- $B_{\alpha} = \left\{ y \in \{0, 1, \dots, l\}^{|\alpha|} \;\middle|\; y_i < y_j \wedge y_{|\alpha|} = l \right\}$
- P_{α} is the set of all the permutations of the indices $(\underbrace{1, \dots, 1}_{\alpha_1 \text{ times}}, \dots, \underbrace{d, \dots, d}_{\alpha_d \text{ times}})$

EFFICIENT PROJECTION

$$\mathbf{x}_p^T = \begin{bmatrix} x_1\tilde{x}_1 \ldots x_1\tilde{x}_d & \ldots & x_d\tilde{x}_1 \ldots x_d\tilde{x}_d \end{bmatrix} \begin{bmatrix} | & & & \\ \mathbf{w} & 0 & \ldots & 0 \\ | & | & & \\ 0 & \mathbf{w} & \ldots & 0 \\ & | & & \\ \vdots & \vdots & \ddots & \vdots \\ & & & | \\ 0 & 0 & \ldots & \mathbf{w} \\ & & & | \end{bmatrix}$$

HYPER-PARAMS TUNING RANGE

CRITEO

- # Hidden layers: 2 ~ 5; Hidden layer size: 32 ~1024
- # Cross layers: 1 ~ 6 (DCN)
- # Residual units: 1 ~ 5; Input dimension and cross dimension: 100 ~ 1026 (DC)
- Initial learning rate: 0.0001 - 0.001

Non-CTR

- # Deep layers: 1 ~ 10; Layer size: 50 ~ 300
- # Cross layers: 4 ~ 10
- # Residual units: 1 ~ 5; Input dimension and cross dimension: 50 ~ 300 (DC)

HYPER-PARAMS FOR BEST MODELS

CROTEO

- DCN: 2 deep layers of size 1024 + 6 cross layers
- DNN: 5 deep layers of size 1024
- DC: 5 residual units with input dimension 424 + cross dimension 537
- LR: 42 cross features

FOREST

- DCN : 8 cross layers of size 54 + 6 deep layers of size 292
- DNN: 7 deep layers of size 292
- DC: 4 residual units with input dimension 271 + cross dimension 287

HIGGS

- DCN: 4 cross layers of size 28 + 4 deep layers of size 209
- DNN: 10 deep layers of size 196

RESULTS WITH STD (CRITEO)

- DCN: $0.4422 \pm 9 \times 10-5$
- DNN: $0.4430 \pm 3.7 \times 10-4$
- DC: $0.4430 \pm 4.3 \times 10-4$