

Interpretable Click-Through Rate Prediction through Hierarchical Attention

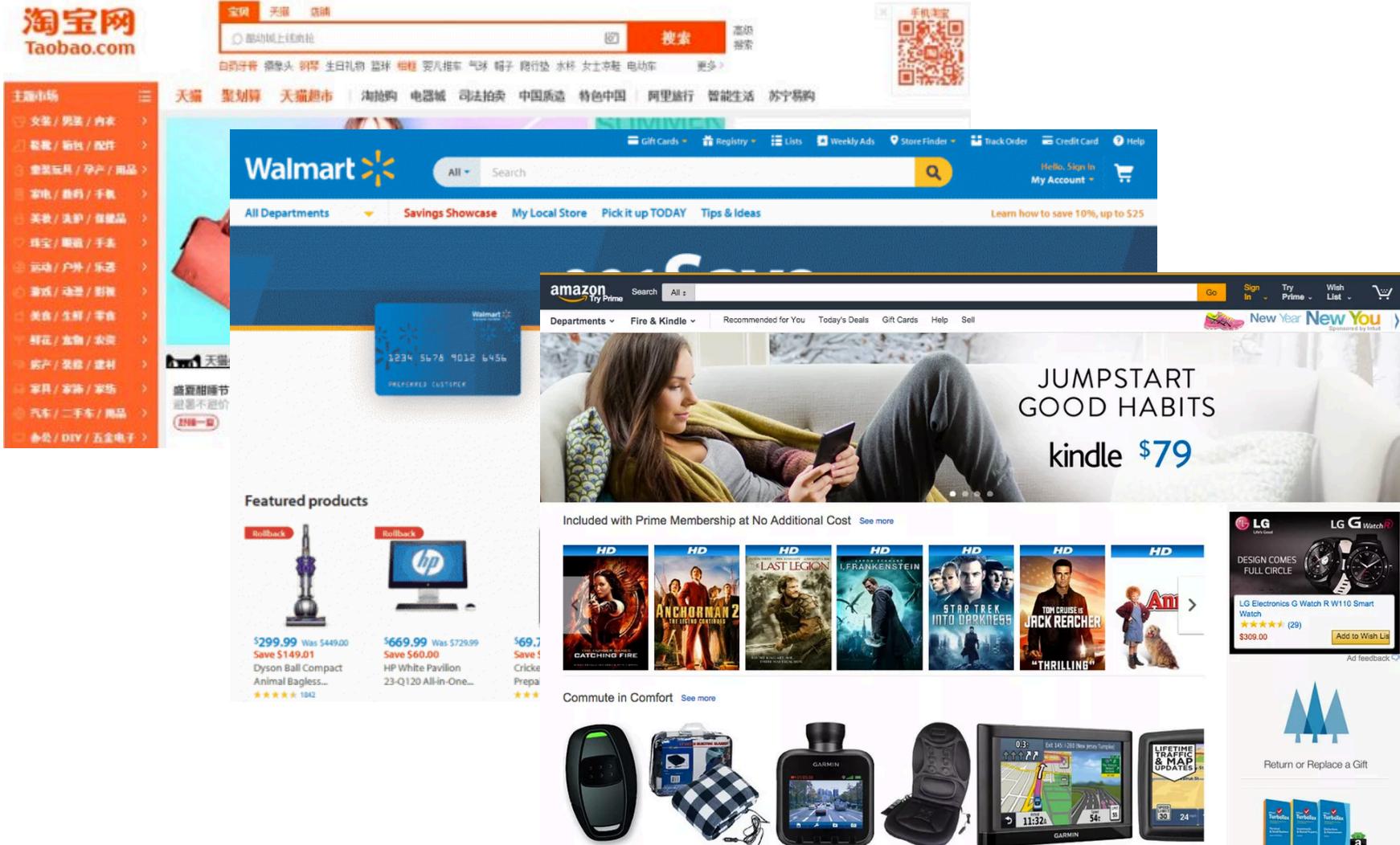
Zeyu Li, Wei Cheng, Yang Chen, Haicheng Chen, Wei Wang

UCLA, NEC Labs America, Google

WSDM 2020

Houston, Texas, USA

Recommender systems



Click-throughs

The Walmart homepage features a prominent '3-2-1 Save' banner at the top. Below it, there's a search bar and navigation links for 'All Departments', 'Savings Showcase', 'My Local Store', 'Pick it up TODAY', 'Tips & Ideas', and 'Learn how to save 10%, up to \$25'. The main content area displays several featured products, including a Dyson Ball Compact Animal Bagless vacuum, an HP Pavilion 22-q120 All-in-One computer, an HP Myoplex Original Ready-to-Drink gift set, and an HP Flyer Red 15.6" laptop.



The Amazon Black Friday Deals Week page is titled 'BLACK FRIDAY DEALS WEEK' and 'Deals on Amazon devices'. It features a grid of discounted products, each with a small image, price, and original price. The products include an Echo Spot Black (\$89.99), Fire TV Stick 4K w/Remote (\$34.99), Fire 7-inch Tablet w/Alexa (\$29.99), All New Echo Dot (3rd Gen) (\$24.00), Fire 7-inch Kids Tablet (\$69.99), All New Echo Plus (2nd Gen) (\$109.99), All New Echo Show (2nd Gen) (\$179.99), and Fire TV Cube w/Alexa (\$59.99).

The Taobao.com homepage features a large 'Fashion time!' advertisement with a woman holding a red handbag and a pair of sneakers. The left sidebar lists various product categories such as 文革 / 男装 / 内衣, 鞋靴 / 鞋类 / 配件, 儿童玩具 / 玩具 / 儿童用品, 家电 / 电子 / 手机, 美妆 / 化妆 / 保健品, 珠宝 / 钻石 / 手表, 运动 / 户外 / 乐器, 食品 / 动漫 / 影视, 美食 / 生鲜 / 食杂, 花卉 / 家居 / 园艺, 房产 / 基建 / 建材, 家具 / 家饰 / 家纺, 汽车 / 二手车 / 汽配, and 手机 / DIY / 五金电子. The right side of the page shows a QR code, a search bar, and various promotional banners for different products like a dress and a car.

Click-throughs



Two natural questions to ask:

1. How many advertisements will be clicked?
2. How many clicks will be purchased?

Click-throughs



Two natural questions to ask:

1. How many advertisements will be clicked?
2. How many clicks will be purchased?

Click-Through Rate Prediction (CTR)

- CTR:
 - Important role in recommendation system
 - Revenue of advertisements



Image: <https://www.lyfemarketing.com/blog/average-click-through-rate/>

Background

- CTR: *binary prediction*
- Pre-Deep Learning Model
 - FM: Factorization Machine
 - MF: Matrix Factorization
 - LR: Logistic Regression
- Deep learning based CTR model
 - DeepFM = FM module + Deep module
 - xDeepFM = CIN module + Deep module
 - CIN: Compressed Interest Network
 - and more ...

SOTA models with DNN

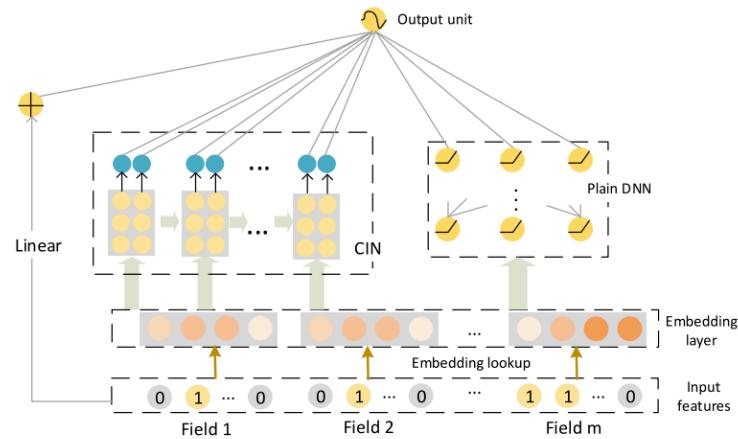


Figure 5: The architecture of xDeepFM.

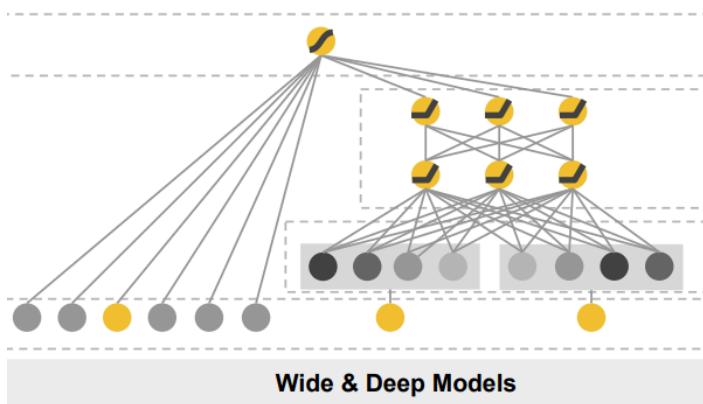


Figure 1: The Deep & Cross Network

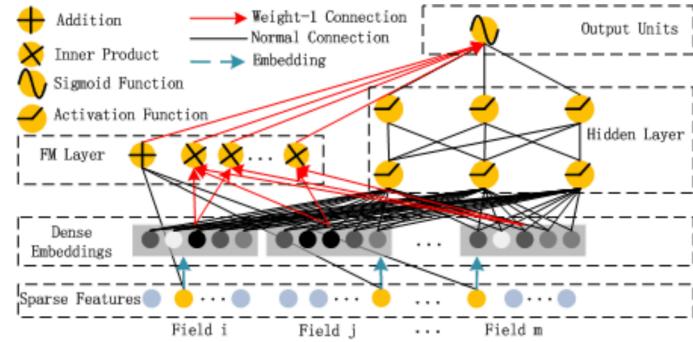
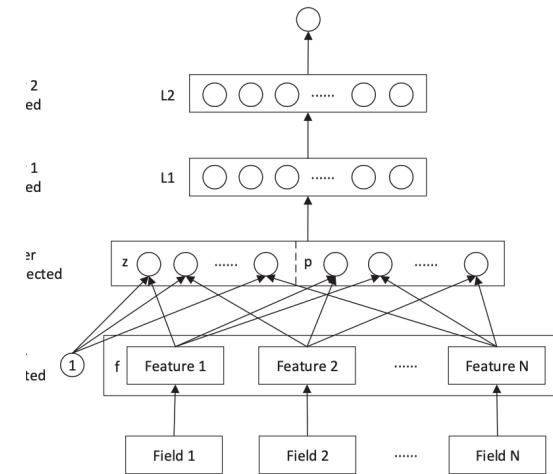
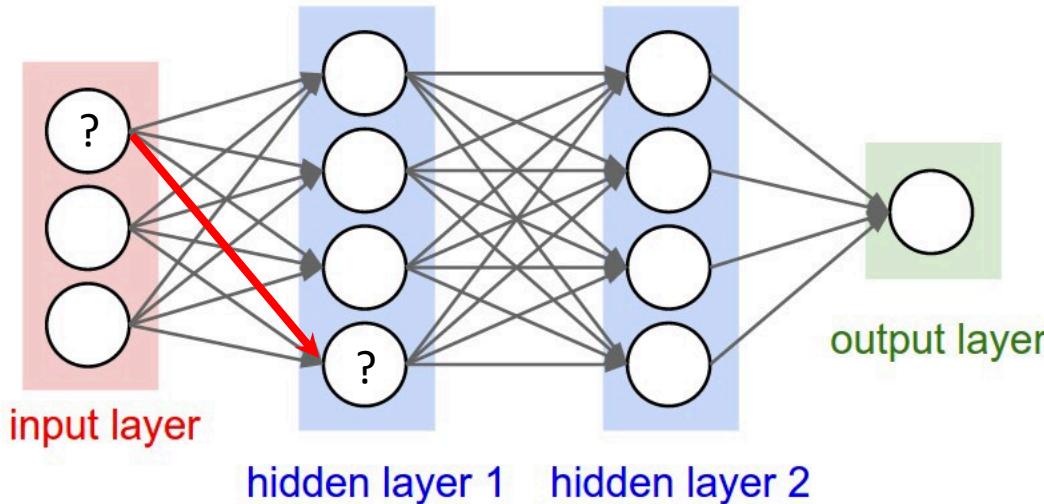


Figure 1: Wide & deep architecture of DeepFM. The wide and deep component share the same input raw feature vector, which enables DeepFM to learn low- and high-order feature interactions simultaneously from the input raw features.



Product-based Neural Network Architecture.

Deep neural network (DNN) module



- DNN
 - Widely used in CTR models
 - Unjustifiable element-wise computation within representations of input or hidden features
 - Unaffordable complexity for big feature dim or size

Image: <https://hackernoon.com/challenges-in-deep-learning-57bbf6e73bb>

Concerns of DNN

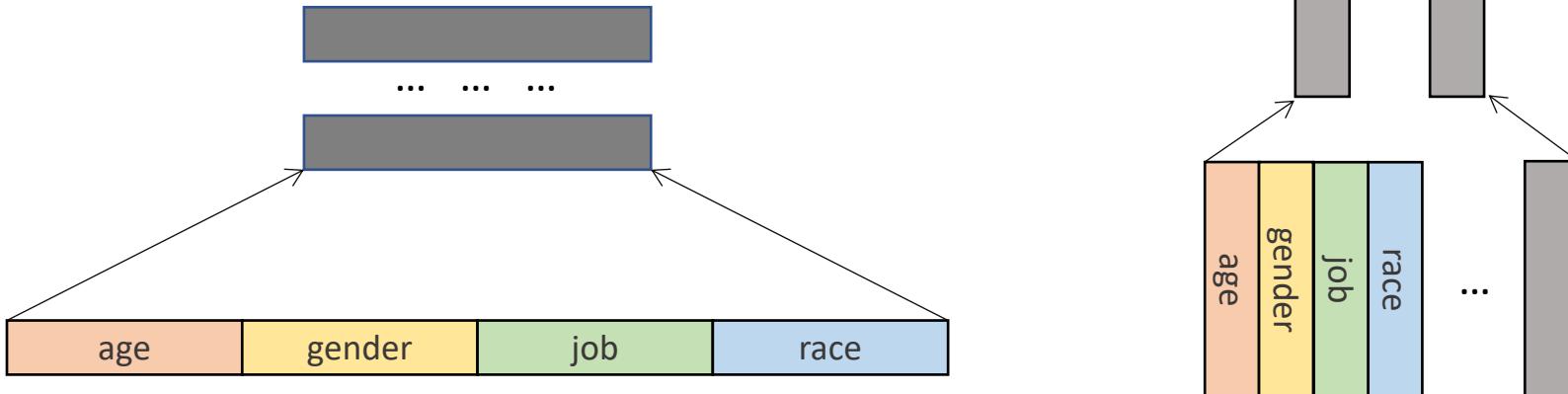
- Okay for online shopping with general purposes
 - Shopping on Amazon ...
- NOT okay for:
 - Medicine recommendation
 - Financial service recommendation
- Criteo:
 - 4 billions clicks in 24 hrs



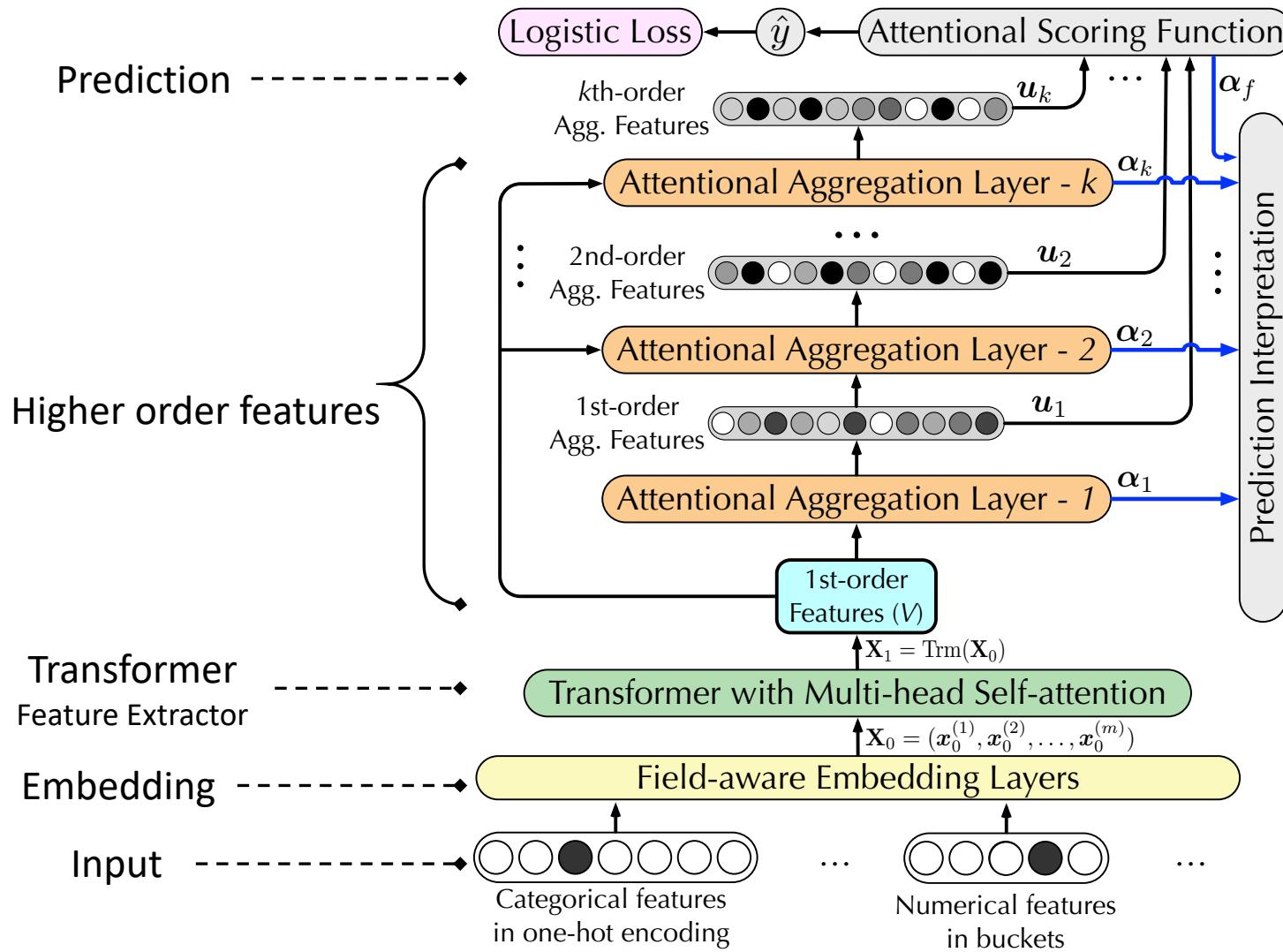
Images: <https://www.fajarmag.com/an-increasing-trend-of-online-shopping/>;
https://www.tes.com/lessons/arP_sMT1GxDHQ/medicine-by-the-minute;
<https://www.wealthandfinance-news.com/awards/>

Our idea -- InterHAt

- Interpretability
 - Attention mechanism
 - Avoid flat concatenation of features
 - Avoid DNN and dim-wise computation
- Efficiency
 - Shrunk problem size

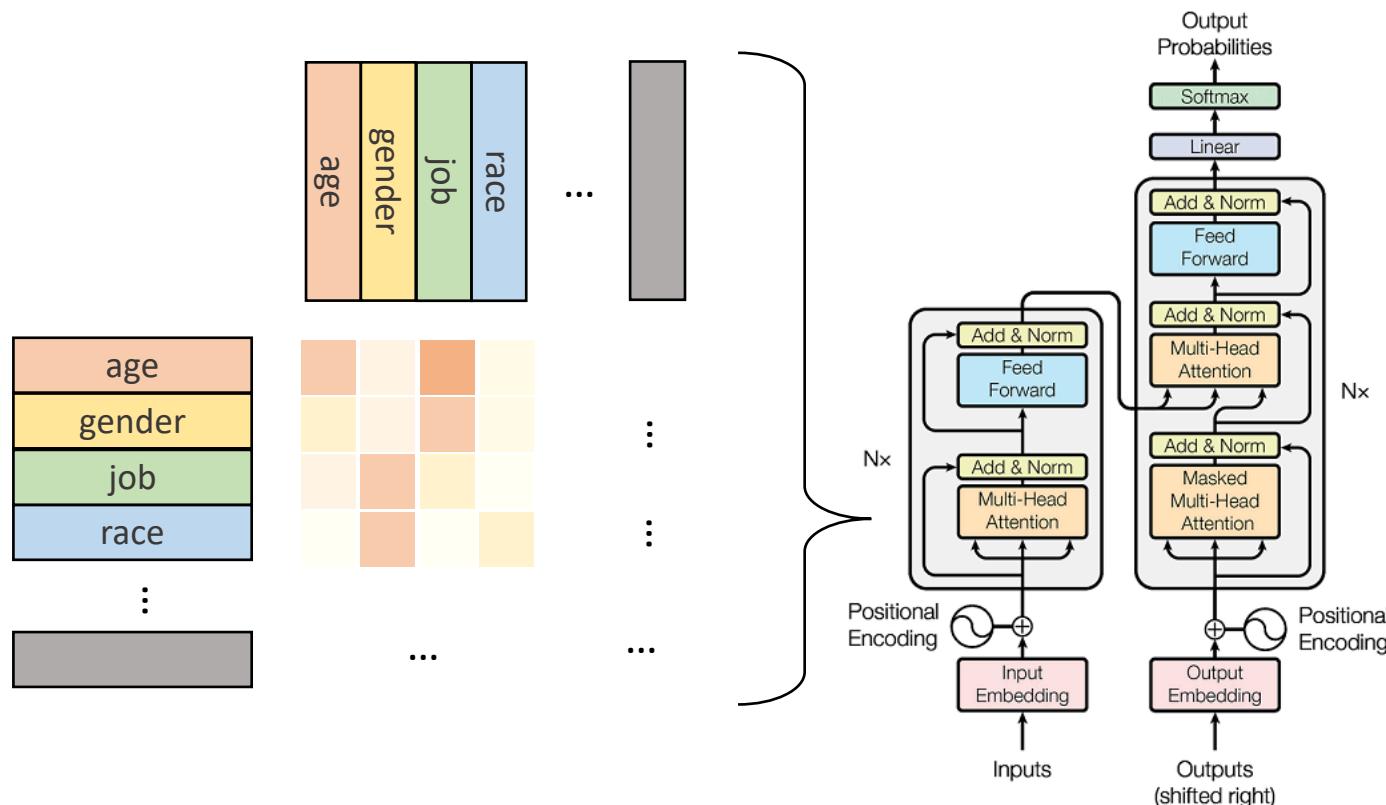


Interpretable CTR via Hierarchical Attention



Polysemy

○ Self attention in Transformer



Right Figure: Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Hierarchical Attention

- Input: i-th order features:
- Generate aggregated feature:

$$\alpha_i^{(j)} = \frac{\exp(\mathbf{c}_i^T \text{ReLU}(\mathbf{W}_i \mathbf{x}_i^{(j)}))}{\sum_{j' \in F} \exp(\mathbf{c}_i^T \text{ReLU}(\mathbf{W}_i \mathbf{x}_i^{(j')}))},$$

$$\mathbf{u}_i = \text{AttentionalAgg}(\mathbf{X}_i) = \sum_{j=1}^m \alpha_i^{(j)} \mathbf{x}_i^{(j)},$$

- Output (i+1)-t order features:

$$\mathbf{x}_{i+1}^{(j)} = \mathbf{u}_i \circ \mathbf{x}_1^{(j)} + \mathbf{x}_i^{(j)}, \quad j \in \{1, \dots, m\},$$

Evaluation

- Datasets
 - Performance evaluation
 - Critio, Avazu, Frappe
 - Interpretability study
 - MovieLens-1m dataset (reviews as clicks)
- Baselines
 - FM, Wide&Deep, DCN, PNN, DeepFM, xDeepFM
- Metrics
 - Area Under ROC Curve (AUC)
 - Cross Entropy (LogLoss)

| Dataset | Criteo | Avazu | Frappe |
|-------------------------|---------|--------|--------|
| #. of features (C + N) | 22 + 14 | 21 + 0 | 7 + 0 |
| #. of total records | 13.8M | 12.1M | 288K |
| #. of distinct features | 605.7K | 23.8K | 5,382 |

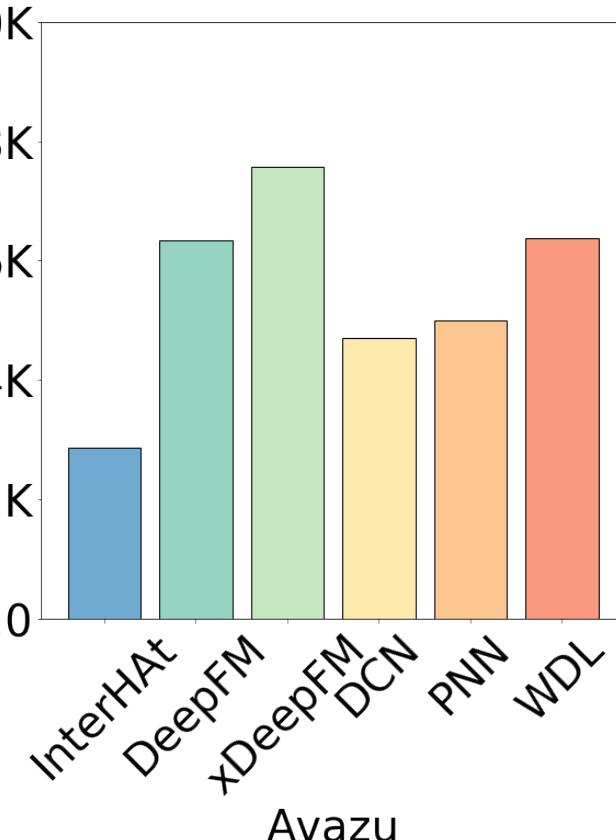
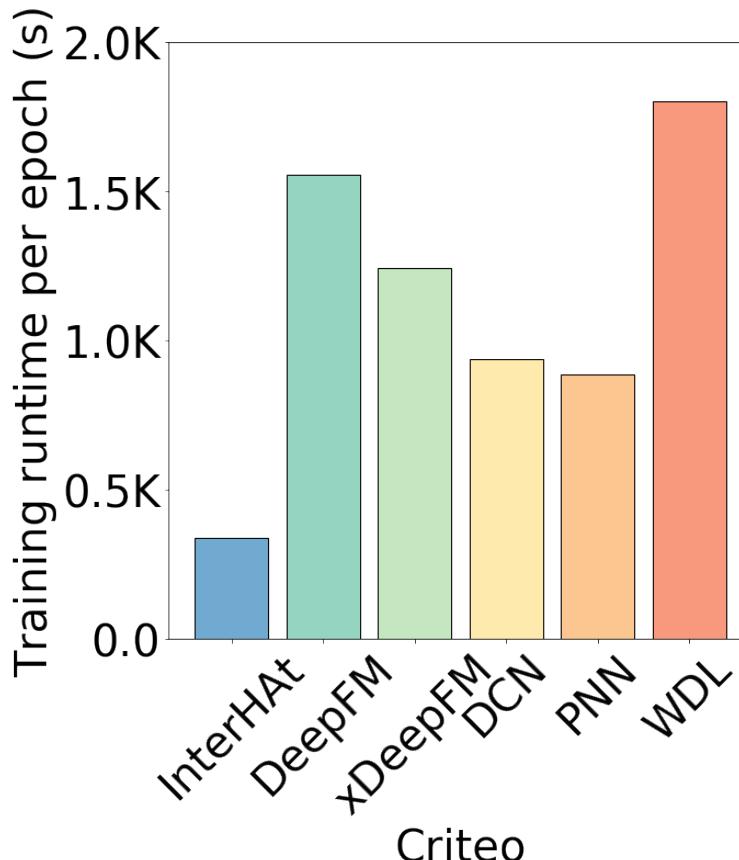
Performance

- Comparable with SOTA models
- Perform better on categorical features
 - SOTA models have close performance
 - Need better ways for encoding numeric features

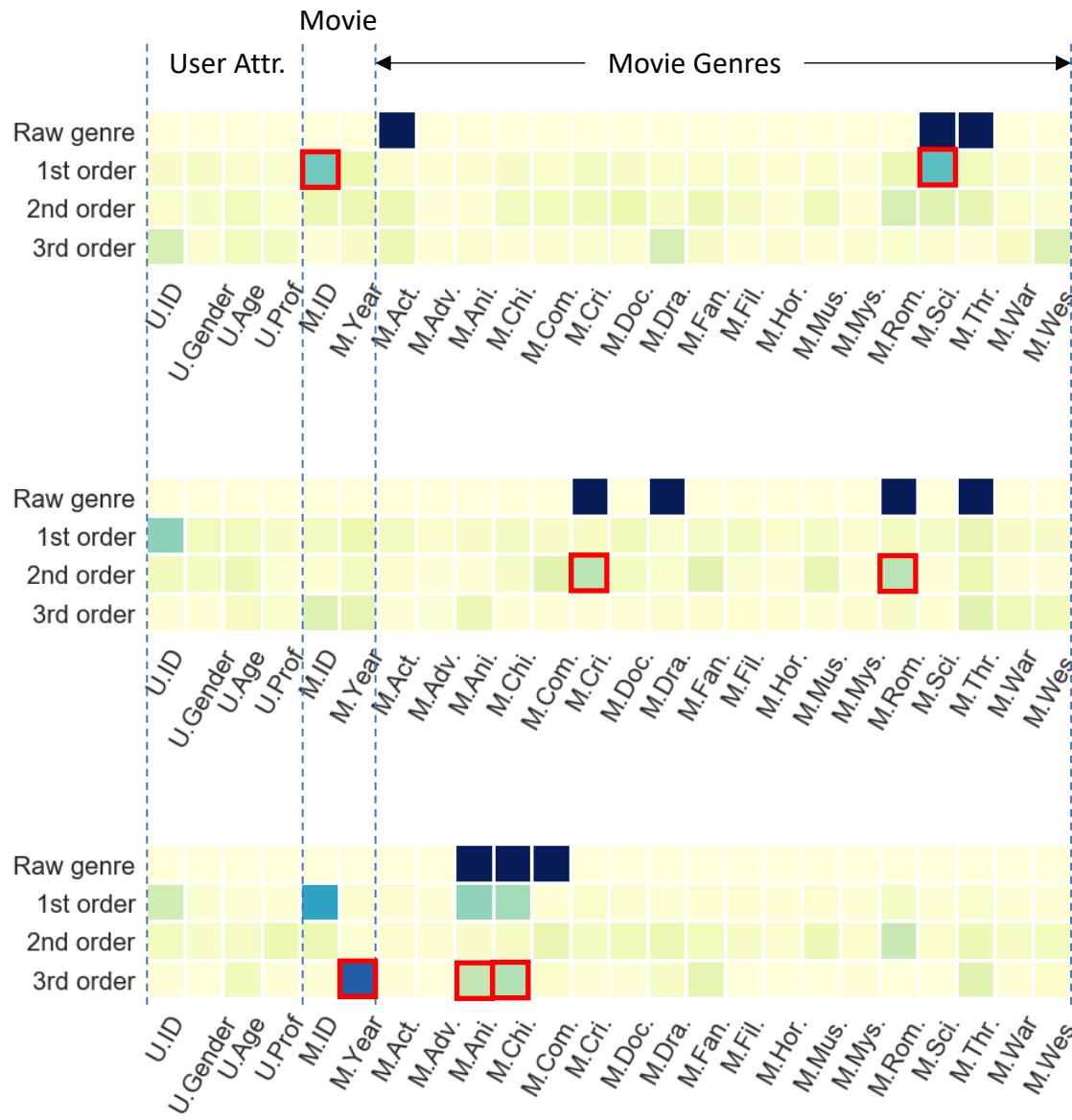
| Dataset | Criteo | | Avazu | | Frappe | |
|------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | Metrics | Logloss | AUC | Logloss | AUC | Logloss |
| FM | 0.4814 | 0.7525 | 0.3951 | 0.7508 | 0.4480 | 0.8625 |
| Wide&Deep | 0.4577 | 0.7845 | 0.3920 | 0.7564 | 0.2571 | 0.9500 |
| DCN | 0.4590 | 0.7826 | 0.3921 | 0.7564 | 0.2335 | 0.9616 |
| PNN | 0.4547 | 0.7887 | 0.3916 | 0.7569 | 0.2177 | 0.9642 |
| DeepFM | 0.4560 | 0.7866 | 0.3920 | 0.7561 | 0.2410 | 0.9520 |
| xDeepFM | 0.4563 | 0.7874 | 0.3917 | 0.7569 | 0.2043 | 0.9694 |
| InterHAt-S | 0.4608 | 0.7820 | 0.3919 | 0.7577 | 0.2151 | 0.9616 |
| InterHAt | 0.4577 | 0.7845 | 0.3910 | 0.7582 | 0.2026 | 0.9696 |

Efficiency

- InterHAt trains faster than other baselines



Interpretability



Conclusion

- InterHAt:
 - Efficiency and interpretability issues of CTR task
 - Efficiency:
 - Avoiding **deep** fully connect neural networks
 - Interpretability:
 - Attention mechanism
 - Interpretability v.s. Explanability
 - Nice performances on both aspects!
 - Try it out:
 - <https://github.com/zyl193/InterHAt>

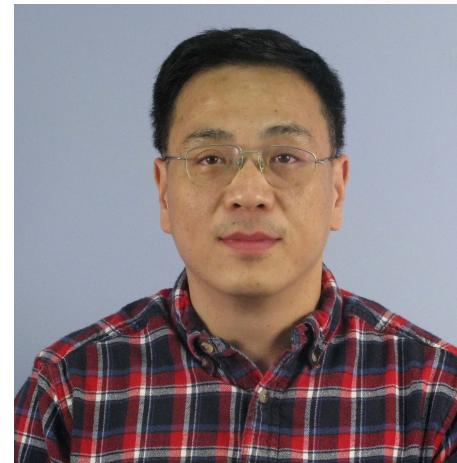
Questions?



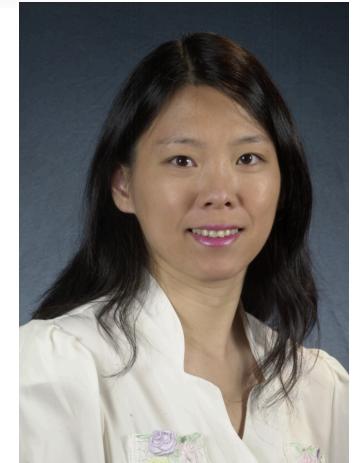
Wei Cheng



Yang Chen



Haifeng Chen



Wei Wang