

Deep Interest with Hierarchical Attention Network for Click-Through Rate Prediction

Weinan Xu, Hengxu He*

Minshi Tan, Yunming Li, Jun Lang, Dongbai Guo

Alibaba Group

{stella.xu, minshi.tan, yunming.li, bill.lang, dongbai.guo}@lazada.com, hengxu.hhx@alibaba-inc.com

ABSTRACT

Deep Interest Network (DIN) is a state-of-the-art model which uses attention mechanism to capture user interests from historical behaviors. User interests intuitively follow a hierarchical pattern such that users generally show interests from a higher-level then to a lower-level abstraction. Modelling such **interest hierarchy** in an attention network can fundamentally improve the representation of user behaviors. We therefore propose an improvement over DIN to model arbitrary interest hierarchy: Deep Interest with Hierarchical Attention Network (DHAN). In this model, a multi-dimensional hierarchical structure is introduced on the first attention layer which attends to individual item, and the subsequent attention layers in the same dimension attend to higher-level hierarchy built on top of the lower corresponding layers. To enable modelling of multiple dimensional hierarchy, an expanding mechanism is introduced to capture one to many hierarchies. This design enables DHAN to attend different importance to different hierarchical abstractions thus can fully capture a user's interests at different dimensions (e.g. category, price or brand). To validate our model, a simplified DHAN is applied to Click-Through Rate (CTR) prediction and our experimental results on three public datasets with two levels of one-dimensional hierarchy only by category. It shows DHAN's superiority with significant AUC uplift from 12% to 21% over DIN. DHAN is also compared with another state-of-the-art model Deep Interest Evolution Network (DIEN), which models temporal interest. The simplified DHAN also gets slight AUC uplift from 1.0% to 1.7% over DIEN. A potential future work can be combination of DHAN and DIEN to model both temporal and hierarchical interests.

CCS CONCEPTS

• **Information systems** → **Users and interactive retrieval; Personalization.**

KEYWORDS

Click-Through Rate Prediction; Hierarchical Pattern; Hierarchical Attention Network; Recommendation

*This author deserves first author equal contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '20, July 25–30, 2020, Virtual Event, China

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8016-4/20/07...\$15.00

<https://doi.org/10.1145/3397271.3401310>

ACM Reference Format:

Weinan Xu, Hengxu He and Minshi Tan, Yunming Li, Jun Lang, Dongbai Guo. 2020. Deep Interest with Hierarchical Attention Network for Click-Through Rate Prediction. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, July 25–30, 2020, Virtual Event, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3397271.3401310>

1 INTRODUCTION

Click-Through Rate (CTR) prediction has become a core task in e-commerce since it is directly related to revenues. Furthermore, CTR also influences user satisfaction. Thus, CTR prediction modeling has drawn more attention from both academia and industry.

Many CTR prediction models such as Wide and Deep (WDL) [1] and Product Neural Network (PNN) [2] have used deep learning methods to extract item-level features and feature interactions. However, these models do not consider to capture user interests through historical behaviors. As the success of attention-based models, several state-of-the-art CTR prediction models [4, 5] have been proposed to capture user interests by extracting behavioral feature based on user historical behaviors. Deep Interest Network (DIN) [5], the first work to use attention-based model in recommendation systems, indicates the diversity of user interests, and uses attention mechanisms to activate historical behaviors with respect to the target item.

However, user interests intuitively follow a hierarchical pattern such that users generally show interests from a higher-level attributes (e.g. category, price or brand) and progressively to a lower-level attributes or items while DIN and other state-of-the-art models overlook it.

To tackle the above issue, a novel structure Deep Interest with Hierarchical Attention Network (DHAN) is proposed to model arbitrary interest hierarchy. In DHAN, a multi-dimension with arbitrary depth hierarchical structure is constructed such that horizontally, the structure captures interests in various dimensions (e.g. category, price or brand); vertically, for each dimension, the structure captures hierarchical interests. Specifically, the multi-dimensional structure is introduced on the first attention layer attending to individual item by different dimensions, and the hierarchical structure is introduced by subsequent attention layers in the same dimension to attend higher-level hierarchy built on top of the lower corresponding attention layers. In this way, DHAN can attend to attributes of different levels of various dimensions with different importance when constructing the overall interest representation.

Deep Interest Evolution Network (DIEN) [4] is another state-of-the-art model which considers dynamic changes of interests in time horizons. DHAN and DIEN focus on two different pathways

to model interests. We also conduct experiments to compare the effects of two different pathways towards CTR prediction.

The main contributions of this paper are summarized as follows:

- A novel structure DHAN is proposed to model arbitrary interest hierarchy with its scalability both on dimensions and hierarchical levels. The simplified structure with **two levels and one dimension** has achieved significant uplift on AUC compared with DIN over three public datasets.
- Compared with DIEN, the simplified DHAN has achieved slight uplift on AUC. The comparisons of modeling interest hierarchy over evolution indicate a potential future work to model both temporal and hierarchical interests.

2 DEEP INTEREST WITH HIERARCHICAL ATTENTION NETWORK

In this section, we introduce Deep Interest with Hierarchical Attention Network (DHAN) in detail. Firstly, we introduce the structure of DHAN. Secondly, structures of simplified DHAN referring to DIN and DIEN are presented.

2.1 The Structure of DHAN

In order to model interest hierarchy with multiple **dimensions** and arbitrary depth, Deep Interest with Hierarchical Attention Network (DHAN) is proposed. As shown in Fig. 1, multiple dimensions can be involved such as category, price, brand and style. The expanding mechanism is introduced above user behaviors and the target item in order to capture interests from one to many dimensions. For each dimension, the interest hierarchical structure is modeled like a tree from the lower-level attributes or items onto the higher-level attributes. Because of the hierarchical structure, the higher-level features are extracted based on the corresponding lower-level features by a series attention modules, enabling DHAN to attend different importance to different attribute levels. For each attention module, the activation unit shown in Fig. 1 is used for activating user interested items with respect to the given target item. On the top level, the overall interest representation incorporates user interests extracted from all levels and dimensions.

2.2 The Structure of Simplified DHAN

Considering the deployment of DHAN in recommendation scenarios, the simplified DHAN is used with one dimension and two attention levels.

In our experiments on public datasets, we use two categories of feature: *User Behaviors* and *Target Item*. The fields of *User Behaviors* are the list of user *reviewed item ids*, and the embedding of user behaviors is represented as $E = \{e_1, e_2, \dots, e_t, \dots, e_T\}$, where $e_t \in R^d$, T is the number of user's historical behaviors. *Target Item* is the item to recommend. The field of *Target Item* is its *item id*, and the embedding of target item is e_a where $e_a \in R^d$. The loss function is the negative log-likelihood function defined as:

$$L = -\frac{1}{N} \sum_{(x,y) \in S} (y \log p(x) + (1-y) \log(1-p(x))), \quad (1)$$

where S is the training set with data size N , and x is the last layer of feature extraction before passing into fully-connected layers.

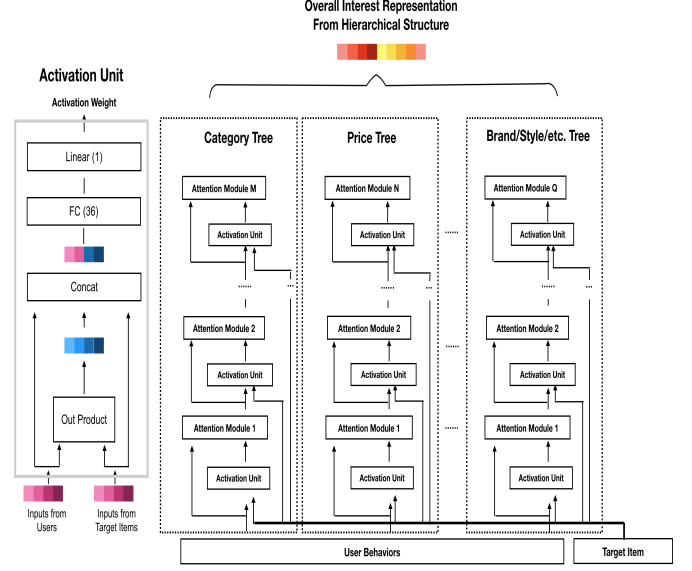


Figure 1: The main structure of DHAN. A multi-dimensional structure is introduced on the Attention Module 1. Each attention module is achieved based on the activation unit. For each tree, higher attention modules are built on top of the lower corresponding attention modules.

$y \in \{0, 1\}$ is the label of data, and $p(x)$ is the output after softmax, which is the probability of reviewing on sample x .

The main part of the simplified DHAN is the Hierarchical Attention Network (HAN) marked in Fig. 2. HAN is to model interest hierarchy with two levels which are item level and attribute level. The item-level representation is the item embedding E . According to DHAN, the representations of attribute levels and overall interests are generated based on two attention modules. In Fig. 2, the first attention module is applied such that item weights $W = \{w_1, w_2, \dots, w_t, \dots, w_T\}$ are calculated by attention mechanism through activation units. For HAN, item embedding and item weights are grouped based on their attributes to generate each attribute-level feature representation. Let the clustered item embedding be $I = \{I^1, I^2, \dots, I^l, \dots, I^L\}$ with L groups obtaining from E , and each I^l is $I^l = \{i_1^l, \dots, i_X^l\}$ where X is the length of user behaviors under group l . The corresponding weight of I^l is $w^l = \{w_1^l, \dots, w_X^l\}$ obtaining from W . To capture the user interests under each higher-level attribute l , each higher-level attribute feature is represented as c^l , which is a normalized weighted sum:

$$c^l = (w_1^l i_1^l + \dots + w_X^l i_X^l) / \sum_{k=1}^X w_k^l. \quad (2)$$

The second attention module is applied on top of the attribute level to extract overall interest representation under one dimension. The cluster weights of attributes are obtained by the second attention layer through the activation unit as $W_c = \{w_{c^1}, w_{c^2}, \dots, w_{c^l}, \dots, w_{c^L}\}$. The overall interest representation x is obtained by the attention mechanism as $x = \sum_{l=1}^L w_{c^l} c^l$.

After passing through the two attention modules, the output of HAN is the overall feature x . Compared with DIN [5] where x is

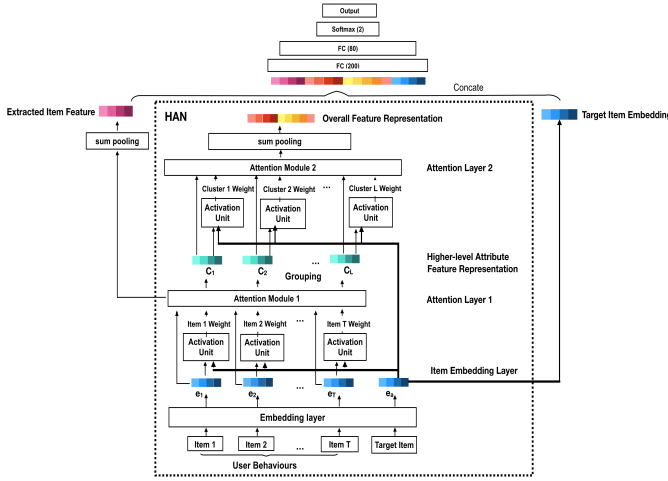


Figure 2: The structure of simplified DHAN, and the main structure HAN is marked.

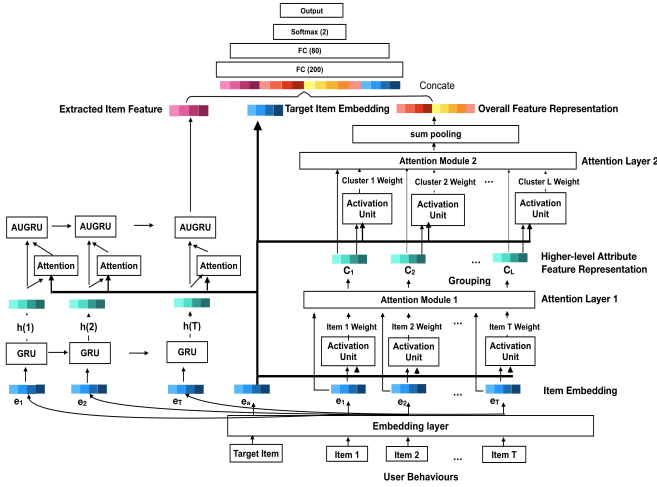


Figure 3: The structure of simplified DHAN which includes the extracted item feature from DIEN.

extracted based on only one item-related attention layer, x from HAN is extracted based on a hierarchical pattern by two attention layers. Thus, x from HAN can contain more information than that from DIN. However, if the number of higher-level attributes is small, x from HAN may have the **over-centralized issue** so that the capturing of interests is over concentrated to some higher-level attributes at early stage, causing the loss function trapped in local optima. To alleviate this issue, and to enhance the capability of feature expression before feeding into MLPs, x is concatenated with target item embedding e_a and **item-level feature representation** i_x . Accordingly, for the comparison of DHAN with DIEN, the item-level feature representation i_x uses the same i_x as DIEN with interest extraction layers by GRUs shown in Fig. 3. Thus, the overall interest representation is $X = [x, i_x, e_a]$, which is then fed into the following MLP for final CTR prediction.

3 EXPERIMENTS

In this section, we present our experiments in detail, including experimental setup, experiment results and analysis. We compare

Table 1: The statistics of datasets with DIN-generation.

Dataset	User	Goods	Categories	Samples
Six-Category	259541	138454	6	463624
Kindle Shop	65599	40241	48	1221635
Electronics	188757	33000	51	1286668

our model simplified DHAN with the state-of-the-art models in public datasets. For simplicity, we use DHAN to notate simplified DHAN in this section. DIN is set as our baseline model in order to study the effect of including the extracted overall feature from the hierarchical structure compared with considering only the item-level extracted feature. Additionally, the model comparison between DHAN and DIEN is also conducted with the same public dataset but a different sample generation method because DIEN considers user interests evolution in time horizons. Besides, Wide and Deep (WDL) model and PNN model are also involved for comparisons because they are widely used in Embedding&MLP. Each experiment is repeated 5 times, and mean and standard deviation are reported.

RelaImpr[3] metric is used to measure the relative improvement over models, which is defined as:

$$RelaImpr = \left(\frac{AUC(measured\ model) - 0.5}{AUC(base\ model) - 0.5} - 1 \right) \times 100\%. \quad (3)$$

3.1 Comparison with DIN

Experimental Setup

Amazon Dataset is used as our benchmark, which contains product review information and metadata from Amazon. Our task is to predict $(k+1)$ -th reviewed item by using the first k reviewed items. For each user, we construct positive sample by taking user’s historical reviews, and generate a negative sample from items not reviewed by this user. Training and testing data sampled from three different datasets: Six-Category, Kindle Shop, and Electronics are used for the performance evaluation of DHAN. The Six-Category contains six subsets: Video Games, Movies and TV, Industrial and Scientific, Electronics, CDs and Vinyl, Apps for Android, and we pick their category names as our category tags and only take 10% of reviews. For the other two sampled datasets, we only keep categories that contain more than 300 items to make the datasets more concentrated, and evaluate the model performance on digital and physical goods separately.

Experiment Results and Analysis

Table 2 describes the results of DHAN compared with DIN and Embedding&MLP on public datasets. Here, DIN is treated as the state-of-the-art baseline. Although DIN beats basic Embedding&MLP (WDL and PNN) with its capability of finding user interested items by attention mechanisms, DHAN beats DIN since firstly, DHAN captures the overall feature representation of user interests with the hierarchical pattern, and secondly, DHAN combines the overall features with item-level features to achieve effective interest expression. The model with hierarchical attention network helps DHAN to beat DIN with remarkable AUC improvement, i.e. 12.58% RelaImpr on Amazon Six-Category Dataset, 14.79% on Kindle Shop Dataset and 21.03% on Electronics Dataset.

The effect of the higher-level activation is visualized in Fig. 4, which illustrates interest intensity of higher-level attributes (category in the case) with respect to the target item. Obviously, the higher-level activation attends to high relevance to the target item

Table 2: Results on Amazon datasets, with DIN as baseline.

Model	Amazon (Six-Category)		Amazon (Kindle Shop)		Amazon (Electronics)	
	mean \pm std	Relaimpr	mean \pm std	Relaimpr	mean \pm std	Relaimpr
WDL	0.6949 \pm 0.00000	-11.49%	0.8928 \pm 0.00032	-0.10%	0.8420 \pm 0.00115	-5.14%
PNN	0.6999 \pm 0.00046	-9.23%	0.8363 \pm 0.00069	-14.49%	0.8498 \pm 0.00066	-2.96%
DIN	0.7202 \pm 0.00007	0.00%	0.8932 \pm 0.00090	0.00%	0.8605 \pm 0.00054	0.00%
DHAN	0.7479 \pm 0.00010	12.58%	0.9514 \pm 0.00291	14.79%	0.9363 \pm 0.00111	21.03%

with the attention weight as 40% on handphone cases, which support the effect of the hierarchical structure in DHAN. Fig. 5 shows the comparison of DHAN and DIN. After 100000 global steps, DHAN shows its significant uplift on both AUC and training loss.

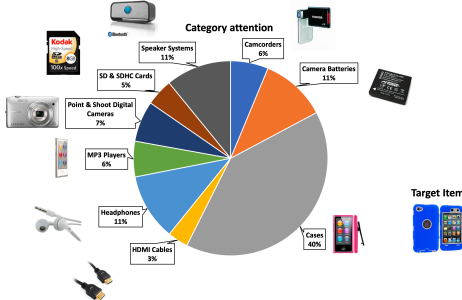


Figure 4: Visualization of higher-level activation in DHAN. This user has the highest interests on Cases category, which is the most relevant category to the target item.

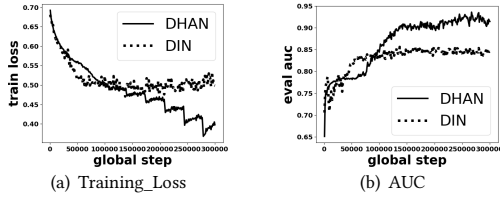


Figure 5: Training loss and AUC of DHAN and DIN on the Amazon Electronics Dataset.

3.2 Comparison with DIEN

Experimental Setup

Similar to 3.1, we still use Electronics, Kindle Shop, and Six-Category datasets from Amazon to construct our training and testing samples. To keep in accordance with the DIEN sample generation method which dedicates to extracting sequential features, each users' sequential review history is taken as one record if the user has at least 5 reviewed items. Due to the different criteria of selecting samples from each dataset, the statistics in Table 3 is different from those in Table 2. A negative sample is generated by selecting items not reviewed by each user. Since the sequential review history is taken based on each user in DIEN, when sampling in Six-Category dataset, we take 10% users' review history, instead of 10% of entire review history directly.

Experiment Results and Analysis

Table 4 shows the results on public datasets where DIEN is the state-of-the-art baseline. Among all the attention-based models, DIEN performs better than DIN since DIEN takes interests evolution in time horizon into account. However, DHAN finally beats DIEN by 1.09% on Amazon Six-Category, 1.75% on Kindle Shop and 1.50% on Electronics. The improvement of DHAN is slight because DIEN proposes a interest extractor layer to capture temporal interest from

Table 3: The statistics of datasets with DIEN-generation.

Dataset	User	Goods	Categories	Samples
Six-Category	49673	139531	6	99346
Kindle Shop	46776	40221	48	93552
Electronics	73958	33175	51	147916

Table 4: Results on Amazon datasets, with DIEN as baseline.

Model	Amazon (Six-Category)		Amazon (Kindle Shop)		Amazon (Electronics)	
	mean \pm std	Relaimpr	mean \pm std	Relaimpr	mean \pm std	Relaimpr
WDL	0.9011 \pm 0.00357	-3.41%	0.7936 \pm 0.00292	-18.49%	0.7023 \pm 0.00034	-9.34%
PNN	0.9127 \pm 0.00030	-0.63%	0.8215 \pm 0.00068	-10.74%	0.7123 \pm 0.00154	-4.82%
DIN	0.9075 \pm 0.00000	-1.88%	0.8321 \pm 0.00011	-7.81%	0.7138 \pm 0.00000	-4.17%
DIEN	0.9153 \pm 0.00000	0.00%	0.8602 \pm 0.00000	0.00%	0.7231 \pm 0.00000	0.00%
DHAN	0.9198 \pm 0.00167	1.09%	0.8665 \pm 0.00079	1.75%	0.7264 \pm 0.00148	1.50%

historical behaviors, but DHAN does not capture this feature. In addition, training and testing sample generation method to adapt to the sequential behaviors in DIEN reduces the AUC gap between DHAN and DIN. This will be addressed in future work.

4 CONCLUSION

In this paper, a novel model Deep Interest with Hierarchical Attention Network (DHAN) is proposed to model user interest hierarchy for multiple dimensions and arbitrary depth. An expanding mechanism in the item level is introduced to capture **one to many** dimensions. For each dimension, the higher-level features are extracted based on the corresponding lower-level features by a series of attention modules. Considering the deployment in recommendation scenarios, a simplified DHAN has been applied to three public datasets, and has achieved significant uplift of AUC around 12% to 21% over DIN and slight uplift of AUC around 1.0% to 1.7% over DIEN. The outperformance of DHAN supports the effectiveness of modeling interest hierarchy leads to more accurate CTR prediction. The differences in uplifts are because DIEN models interest evolution. Thus, a potential future work is to model user interests considering both interest hierarchy and evolution. Furthermore, as more dimensions are involved in item features, another future work is to replace the predefined dimensions and automatically capture user interests towards those dimensions.

REFERENCES

- [1] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ipsir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
- [2] Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. 2016. Product-based neural networks for user response prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 1149–1154.
- [3] Ling Yan, Wu-jun Li, Gui-Rong Xue, and Dingyi Han. 2014. Coupled group lasso for web-scale ctr prediction in display advertising. In *Proceedings of the 31th International Conference on Machine Learning*. 802–810.
- [4] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5941–5948.
- [5] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.

A OPEN SOURCE CODES

Our open-source codes with parameters are available on GitHub: <https://github.com/stellaxu/DHAN>.