Deep Interest Evolution Network for Click-Through Rate Prediction

Guorui Zhou[*], Na Mou[†], Ying Fan, Qi Pi, Weijie Bian,
Chang Zhou, Xiaoqiang Zhu and **Kun Gai**
Alibaba Inc, Beijing, China
{guorui.xgr, mouna.mn, fanying.fy, piqi.pq, weijie.bwj, ericzhou.zc, xiaoqiang.zxq, jingshi.gk}@alibaba-inc.com

KDD2019

PanXiang

2018.11.20

CTR

- interest extractor layer to capture temporal interests from history behavior sequence
  - auxiliary loss to supervise interest extracting at each step
- interest evolving layer to capture interest evolving process that is relative to the target item
  - attention mechanism

GRU with attentional update gate (AUGRU)

GRU

model the dependency between behaviors

auxiliary loss

    uses the next behavior to supervise the learning of current hidden state

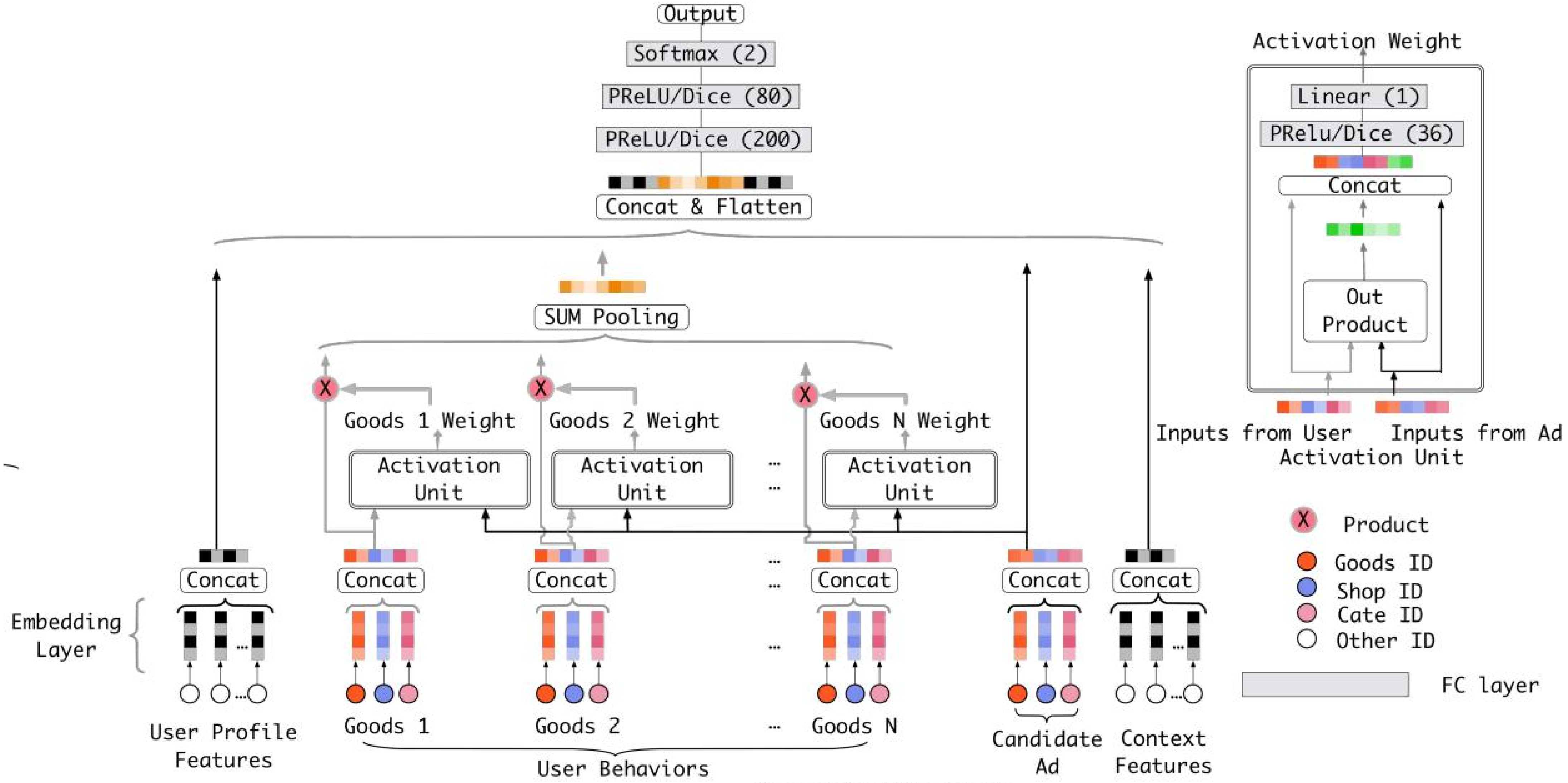
Attention

interest drifting phenomenon

    Each interest has its own evolution track

DIN

regard the behavior as the interest directly, while latent interest is hard tobe fully reflected by explicit behavior.

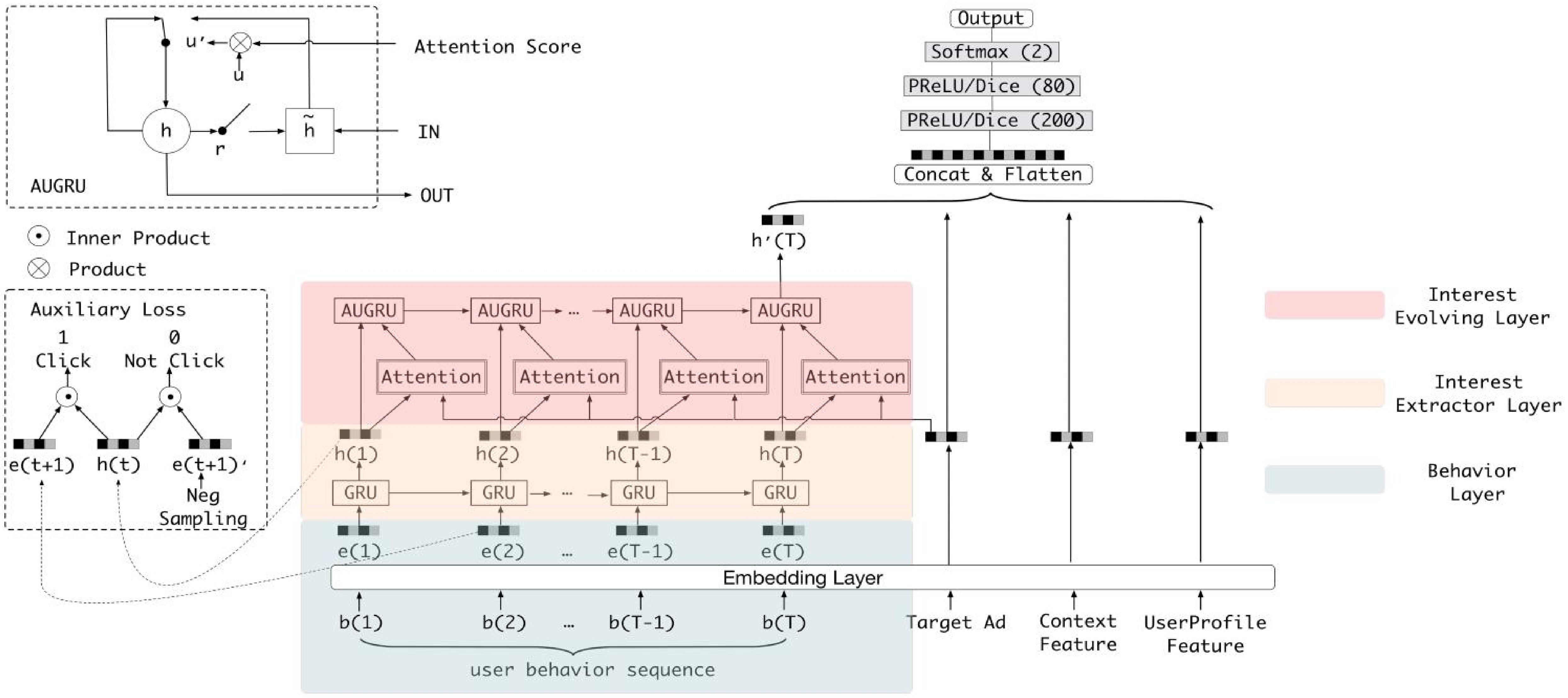


Deep Interest Network

$$\boldsymbol{v}_U(A) = f(\boldsymbol{v}_A, \boldsymbol{e}_1, \boldsymbol{e}_2, .., \boldsymbol{e}_H) = \sum_{j=1}^{H} a(\boldsymbol{e}_j, \boldsymbol{v}_A)\boldsymbol{e}_j = \sum_{j=1}^{H} \boldsymbol{w}_j \boldsymbol{e}_j, \tag{3}$$

V(A)=f(Va,e1,e2,eh)

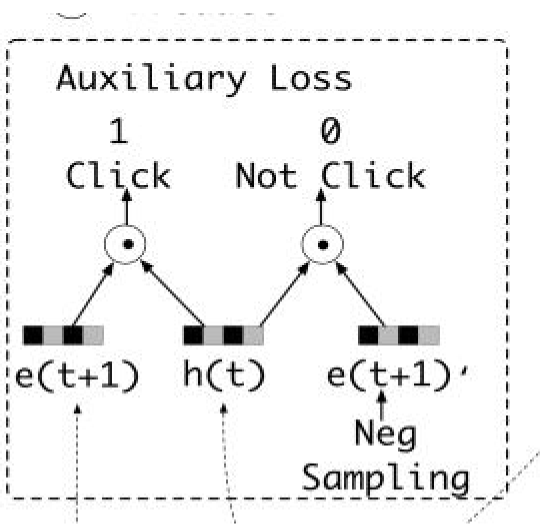
where $\{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_H\}$ is the list of embedding vectors of behaviors of user U with length of H, $\boldsymbol{v}_A$ is the embedding vector of ad A. In this way, $\boldsymbol{v}_U(A)$ varies over different ads. $a(\cdot)$ is a feed-forward



AUGRU
u'
u
h
r
h̃
Attention Score
IN
OUT
Inner Product
Product
Auxiliary Loss
1
Click
0
Not Click
e(t+1)
h(t)
e(t+1)'
Neg
Sampling
Output
Softmax (2)
PReLU/Dice (80)
PReLU/Dice (200)
Concat & Flatten
h'(T)
AUGRU
Attention
h(1)
h(2)
h(T-1)
h(T)
GRU
e(1)
e(2)
e(T-1)
e(T)
Embedding Layer
b(1)
b(2)
b(T-1)
b(T)
user behavior sequence
Target Ad
Context
Feature
UserProfile
Feature
Interest
Evolving Layer
Interest
Extractor Layer
Behavior
Layer

Interest Extractor Layer

$$\sigma(\mathbf{x_1}, \mathbf{x_2}) = \frac{1}{1+\exp(-[\mathbf{x_1}, \mathbf{x_2}])}$$



h(t) interest

$$L_{aux} = -\frac{1}{N}\Big(\sum_{i=1}^{N}\sum_{t} \log \sigma(\mathbf{h}_t^i, \mathbf{e}_b^i[t+1]) + \log(1 - \sigma(\mathbf{h}_t^i, \hat{\mathbf{e}}_b^i[t+1]))\Big),$$

$$L_{target} = -\frac{1}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}}^{N} (y \log p(\mathbf{x}) + (1-y)\log(1-p(\mathbf{x}))), \quad (1)$$

where $\mathbf{x} = [\mathbf{x}_p, \mathbf{x}_a, \mathbf{x}_c, \mathbf{x}_b] \in \mathcal{D}$, $\mathcal{D}$ is the training set of size N. $y \in \{0, 1\}$ represents whether the user clicks target

$$L = L_{target} + \alpha * L_{aux},$$

Interest Evolving Layer

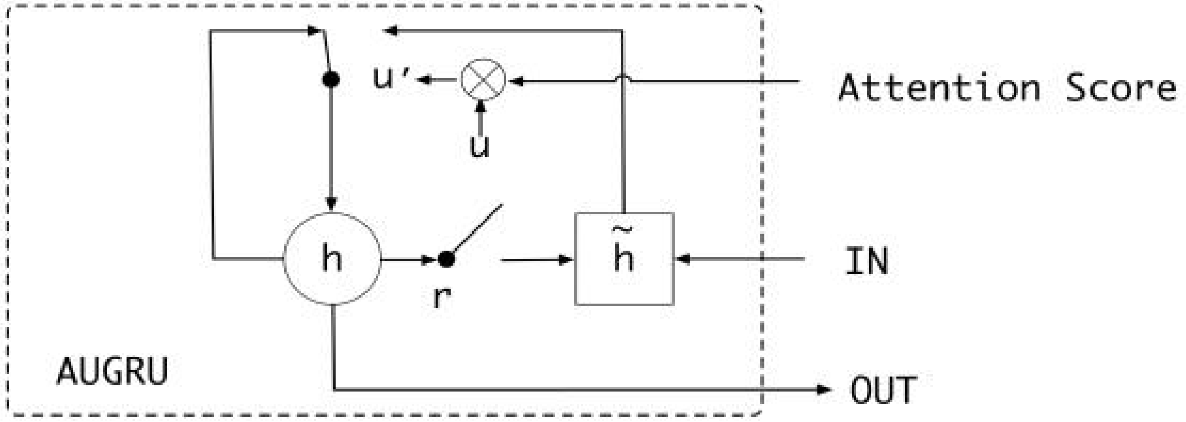
interest drift

may interest inkinds of books during a period of time, and need clothesin another time.

interest individual

We only concerns the evolving process that is relative to target item

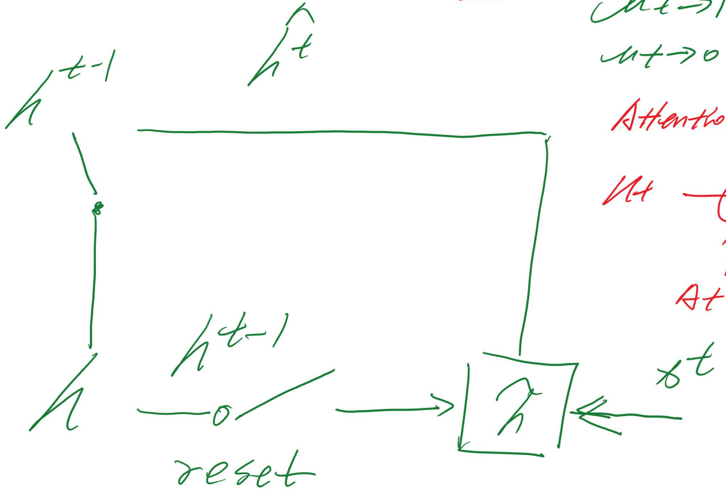
Interest Evolving Layer



$h_t = (1 - u_t) h^{t-1} + u_t \hat{h}^t$.

$u_t \to 1$ update

$u_t \to 0$ remember



Attention:

$u_t \otimes \to u_t'$

At score { Value / vec

$$\mathbf{u}_t = \sigma(W^u \mathbf{i}_t + U^u \mathbf{h}_{t-1} + \mathbf{b}^u), \tag{2}$$

$$\mathbf{r}_t = \sigma(W^r \mathbf{i}_t + U^r \mathbf{h}_{t-1} + \mathbf{b}^r), \tag{3}$$

$$\tilde{\mathbf{h}}_t = \tanh(W^h \mathbf{i}_t + \mathbf{r}_t \circ U^h \mathbf{h}_{t-1} + \mathbf{b}^h), \tag{4}$$

$$\mathbf{h}_t = (\mathbf{1} - \mathbf{u}_t) \circ \mathbf{h}_{t-1} + \mathbf{u}_t \circ \tilde{\mathbf{h}}_t, \tag{5}$$

Interest Evolving Layer

$$a_t = \frac{\exp(\mathbf{h}_t W \mathbf{e}_a)}{\sum_{j=1}^{T} \exp(\mathbf{h}_j W \mathbf{e}_a)},$$

where $\mathbf{e}_a$ is the concat of embedding vectors from fields in category ad, $W \in \mathbb{R}^{n_H \times n_A}$, n_H is the dimension of hidden state and n_A is the dimension of advertisement's embedding vector. Attention score can reflect the relationship between advertisement $\mathbf{e}_a$ and input $\mathbf{h}_t$, and strong relativeness leads to a large attention score.

$$\mathbf{i}'_t = \mathbf{h}_t$$

AIGRU $\mathbf{i}'_t = \mathbf{h}_t * a_t$

AGRU $\mathbf{h}'_t = (1 - a_t) * \mathbf{h}'_{t-1} + a_t * \tilde{\mathbf{h}}'_t,$

AUGRU

$$\tilde{\mathbf{u}}'_t = a_t * \mathbf{u}'_t,$$

$$\mathbf{h}'_t = (1 - \tilde{\mathbf{u}}'_t) \circ \mathbf{h}'_{t-1} + \tilde{\mathbf{u}}'_t \circ \tilde{\mathbf{h}}'_t,$$

EXPERMENTS

Table 2: Results (AUC) on public datasets

Model	Electronics (mean± std)	Books (mean ± std)
BaseModel (Zhou et al. 2018c)	0.7435 ± 0.00128	0.7686 ± 0.00253
Wide&Deep (Cheng et al. 2016)	0.7456 ± 0.00127	0.7735 ± 0.00051
PNN (Qu et al. 2016)	0.7543 ± 0.00101	0.7799 ± 0.00181
DIN (Zhou et al. 2018c)	0.7603 ± 0.00028	0.7880 ± 0.00216
Two layer GRU Attention	0.7605 ± 0.00059	0.7890 ± 0.00268
DIEN	$\mathbf{0.7792 \pm 0.00243}$	$\mathbf{0.8453 \pm 0.00476}$

Table 3: Results (AUC) on industrial dataset

Model	AUC
BaseModel (Zhou et al. 2018c)	0.6350
Wide&Deep (Cheng et al. 2016)	0.6362
PNN (Qu et al. 2016)	0.6353
DIN (Zhou et al. 2018c)	0.6428
Two layer GRU Attention	0.6457
BaseModel + GRU + AUGRU	0.6493
DIEN	**0.6541**

Table 4: Effect of AUGRU and auxiliary loss (AUC)

Model	Electronics (mean ± std)	Books (mean ± std)
BaseModel	0.7435 ± 0.00128	0.7686 ± 0.00253
Two layer GRU attention	0.7605 ± 0.00059	0.7890 ± 0.00268
BaseModel + GRU + AIGRU	0.7606 ± 0.00061	0.7892 ± 0.00222
BaseModel + GRU + AGRU	0.7628 ± 0.00015	0.7890 ± 0.00268
BaseModel + GRU + AUGRU	0.7640 ± 0.00073	0.7911 ± 0.00150
DIEN	$\mathbf{0.7792 \pm 0.00243}$	$\mathbf{0.8453 \pm 0.00476}$