DEAR: Deep Reinforcement Learning for Online Advertising Impression in Recommender Systems

PaperID: 4386

Xiangyu Zhao[1], Changsheng Gu[2], Haoshenglun Zhang[2]
Xiwang Yang[2], Xiaobing Liu[2], Hui Liu[1], Jiliang Tang[1]

1: Data Science and Engineering Lab, Michigan State University
2: Bytedance

Recommender Systems

- Assisting users in their information-seeking tasks
  - Goal: suggesting items that best fit user's preferences

Music

Video

Ecommerce

News

Social Friends

Location based

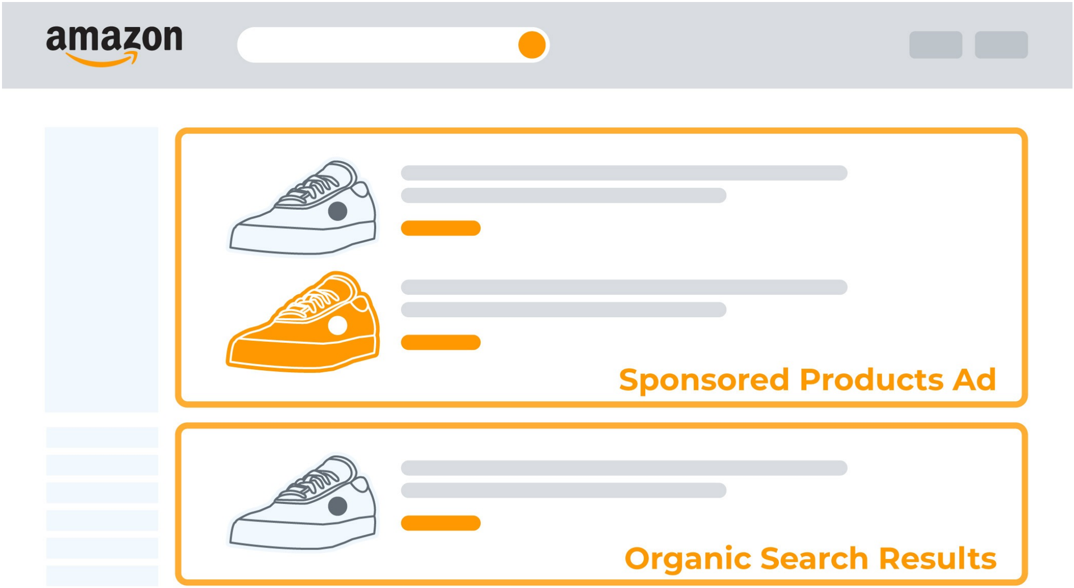
Online Ads

Online App

Content

Advertising in Recommender Systems

- Goal: maximizing the advertising revenue from advertisers
- Assigning the right ads at the right place to the right consumers



Online Advertising Challenges

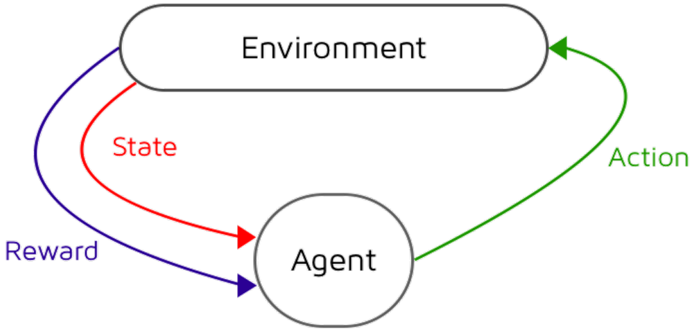
- Offline and static optimization

guaranteed delivery

real-time bidding

Online Advertising Challenges

- Reinforcement learning based online advertising



- Challenges:

VS

advertising revenue

user experience

An Example of Online Advertising Impression

- Three tasks
  - Interpolate an ad?
  - The optimal location?
  - The optimal ad?

- Goals of ad agent
  - Maximizing advertising revenue
  - Minimizing the negative influence of ads on user experience

Definition

- Markov Decision Process (MDP)
  - Advertising agent interacts with environment (users)

- State space S:
  - A state $s_t \in S$ is defined as a user's browsing history before time t and the information of current request at time t

$$s_t = concat(p_t^{rec}, p_t^{ad}, c_t, rec_t)$$

- Action space A:
  - The action $a_t \in A$ is to determine three internally related tasks: interpolate an ad? the optimal location? the optimal ad?

Definition

- Reward R:
  - Income of ad r_t^{ad}
  - Influence of an ad on the user experience r_t^{ex}

$$r_t(s_t, a_t) = r_t^{ad} + \alpha \cdot r_t^{ex} \qquad r_t^{ex} = \begin{cases} 1 & continue \\ -1 & leave \end{cases}$$

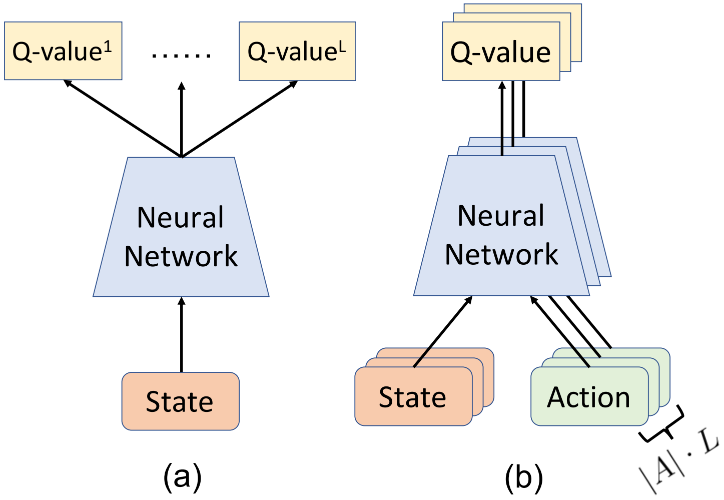
- Transition probability P:
  - The state transition from s_t to s_{t+1} after taking the action a_t

$$p(s_{t+1}|s_t, a_t, ..., s_1, a_1) = p(s_{t+1}|s_t, a_t)$$

- Discount factor γ:
  - Discount factor $\gamma \in [0,1]$ is introduced to measure the present value of future reward

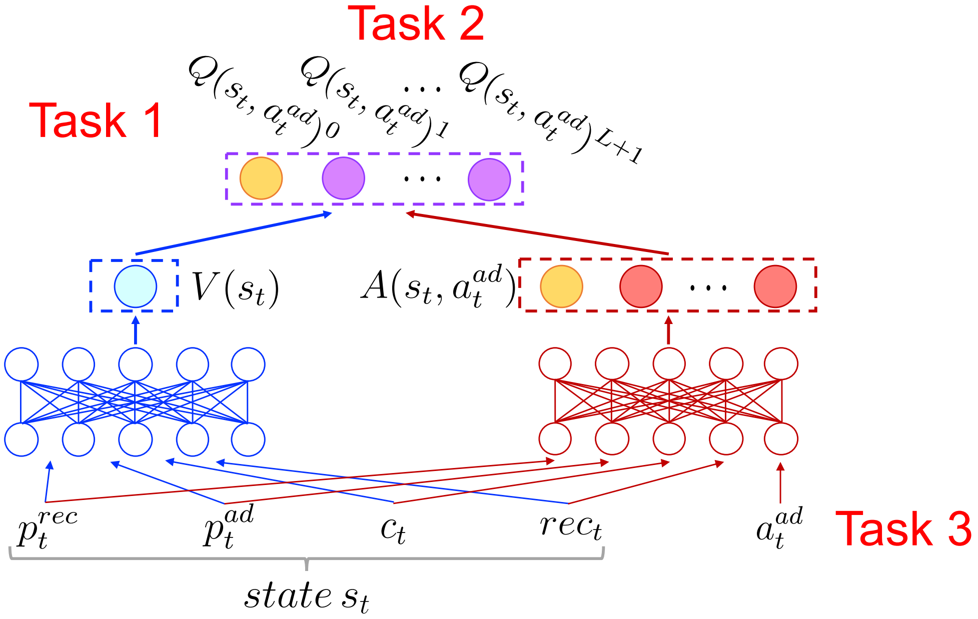
Classic DQN Architectures

- Assumptions
  - There are $|A|$ candidate ads for each request
  - The length of the rec-list is L



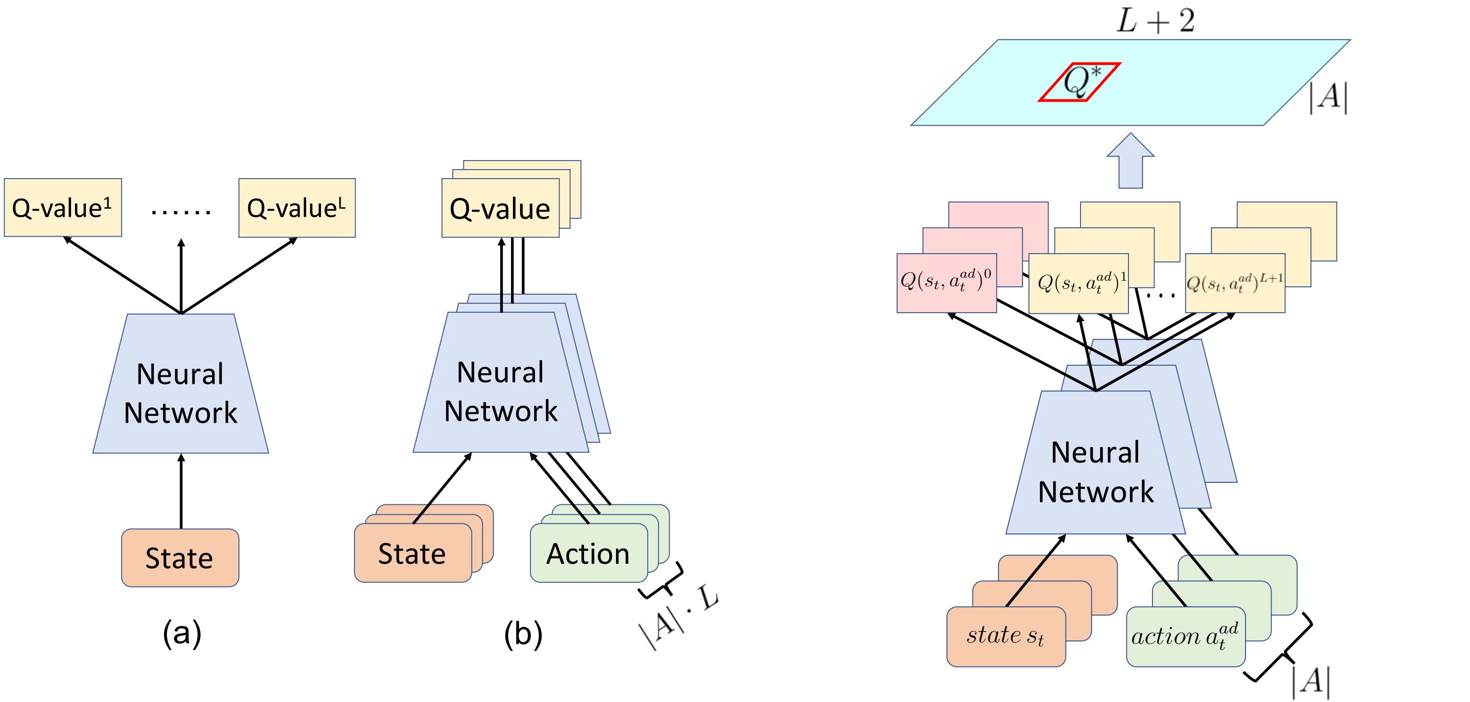
Novel DQN Architecture

- Three tasks
  - Task 1: Interpolate an ad?
  - Task 2: The optimal location?
  - Task 3: The optimal ad?



Comparison

- The first individual DQN architecture that can simultaneously evaluate the Q-values of multiple levels' related actions



Experimental Settings

- Dataset from the short video app Douyin

Table 1: Statistics of the Douyin video dataset.

session	**user**	**normal video**	**ad video**
1,000,000	188,409	17,820,066	10,806,778
session time	**session length**	**session ad revenue**	**rec-list with ad**
17.980 min	55.032 videos	0.667	55.23%

- Metric: accumulated reward of a recommendation session

Overall Performance Comparison

- Baselines
  - Wide & Deep
  - DeepFM
  - GRU4REC
  - Hierarchical DQN

Table 2: Overall performance comparison.

method	reward	improvement	p−*value*
W&D	9.12	20.17%	0.000
DFM	9.23	18.75%	0.000
GRU	9.87	11.05%	0.000
HDQN	10.27	6.712%	0.002
DEAR	**10.96**	-	-

Component Study

- DEAR-1: supervised training
- DEAR-2: no RNN
- DEAR-3: classical DQN (b)
- DEAR-4: no $Q(s, a)=V(s)+A(s, a)$
- DEAR-5 : random ad
- DEAR-6: random location

Table 3: Component study results.

variant	reward	improvement	p−$value$
DEAR-1	9.936	10.32%	0.000
DEAR-2	10.02	9.056%	0.000
DEAR-3	10.39	5.495%	0.001
DEAR-4	10.57	3.689%	0.006
DEAR-5	9.735	12.58%	0.000
DEAR-6	9.963	10.01%	0.000
DEAR	**10.96**	-	-

Conclusion

- A deep RL framework DEAR with a novel DQN architecture for online advertising in recommender systems

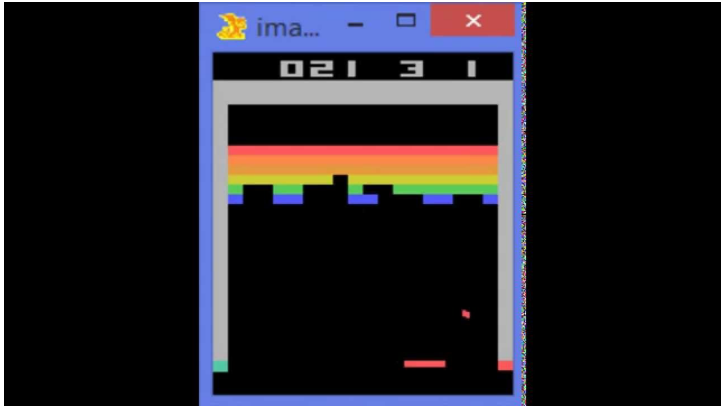
- Determine three internally related actions at the same time
  - Interpolate an ad?
  - The optimal location?
  - The optimal ad?

- Simultaneously maximize the revenue of ads and minimize the negative influence of ads on user experience

Future Work

- Jointly optimizes advertising and recommending strategies

- More applications such as video games



Thanks

zhaoxi35@msu.edu