Addressing Cold-start Problem in Click-Through Rate Prediction via Supervised Diffusion Modeling

Anonymous submission

Algorithm 1: Supervised diffusion model framework.

Input: f_θ: A pretrained backbone model.
Input: $\phi_{\text{ID}}^{\text{new}}$: Pretrained ID embeddings by the backbone.
Input: $\mathcal{D}$: A recommendation dataset.
Randomly initialize U-Net.
while not converge **do**
  Sample a batch sample $\mathcal{B}$ from $\mathcal{D}$.
  Extract cold ID embeddings $\mathbf{e}$ of items in $\mathcal{B}$.
  Update parameters of U-Net by optimizing $\mathcal{L}$.
end while
while item id embedding is not replaced **do**
  Get the item id and related side information.
  Generate warm up embedding $\mathbf{w}$ by Equation (10).
  Replace the item id embeddings by $\mathbf{w}$.
end while

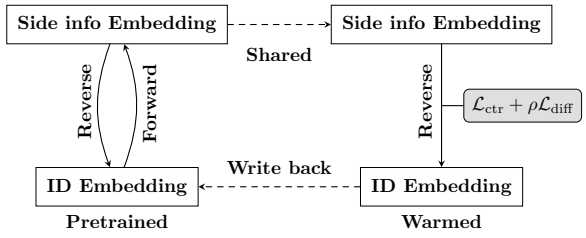

Figure 1: An illustration of the process for generating the warm-up embeddings.

Method

We outline the main steps of our methodology in Algorithm 1. First, a backbone model is pre-trained to provide the initial item embeddings. These pre-trained embeddings are then converted into $\mathbf{z}_0$ in the diffusion process. During the training of the diffusion model, the parameters of the backbone model are frozen. After completing the training of the diffusion model, we write the warmed-up embeddings back into the original embedding space, as shown in Figure 1.

Datasets

We report the statistics of the datasets we used in our experiments in Table 1.

MovieLens-1M [1]: It is one of the most well-known datasets for evaluating recommendation algorithms. This dataset comprises one million instances of movie ratings across thousands of movies and users. The movie features include movie ID, title, year of release, and genres, while the user features encompass age, gender, and occupation. We transfer ratings into binary (The ratings less than 4 are turned to 0 and the others are turned into 1).

Taobao Display Ad Click [2]: It contains 114000 randomly selected users from the Taobao website, covering 8 days of ad display and click log data (26 million records). The ad features include category ID, campaign ID, brand ID, advertiser ID, and price. User features consist of micro group ID, cms_group_id, gender, age, consumption level, shopping depth, occupation, and city level. The dataset includes a label of 1 for click behavior and 0 for non-click behavior.

CIKM 2019 EComm AI [3]: It is an E-commerce dataset comprising 62 million instances. Each item is characterized by four categorical features: item ID, category ID, shop ID, and brand ID. User features encompass user ID, gender, age, and purchasing power. Each instance is tagged with a behavioral label('pv', 'buy', 'cart', 'fav'). We transform the label into a binary format (1 for a purchase action, 0 otherwise).

More Experiments

Generalization Experiments. Beside applying DeepFM, Wide & Deep, and DCN as backbone models, we also conduct experiments on PNN. PNN (Qu et al. 2016) is a CTR prediction model that employs a product layer to explore the interactions among inter-field categories. We present the results in Figure 2. The results show that our method outperforms others in most cases.

Ablation Study. We also conduct an ablation analysis on the diffusion model concerning the dropout mechanism across three datasets using DCN as the backbone model. The results are shown in Table 2. We can observe similar results: on the MovieLens-1M dataset, the inclusion of

[1] http://www.grouplens.org/datasets/movielens/
[2] https://tianchi.aliyun.com/dataset/dataDetail?dataId=56
[3] https://tianchi.aliyun.com/competition/entrance/231721

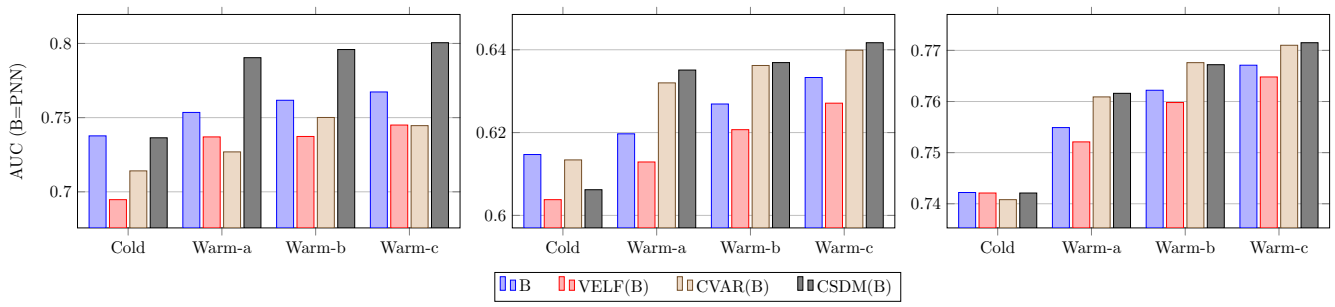

Figure 2: AUC scores evaluated across various stages using PNN as backbone model, conducted over three datasets with 10 runs per model.

Dataset	MovieLens-1M	Taobao AD	CIKM 2019
#user	6040	1141729	1050000
#item	3706	864811	3934201
#instance	1000209	25029435	62428486

Table 1: Statistics of datasets used in our experiments.

Dropout	Dataset	Cold	Warm-a	Warm-b	Warm-c
w	ML-1M	**0.7362**	0.7971	0.8020	**0.8059**
w/o	ML-1M	0.7360	0.7971	0.8020	0.8057
w	TaobaoAD	0.6097	0.6352	**0.6387**	**0.6436**
w/o	TaobaoAD	**0.6103**	**0.6361**	0.6382	0.6428
w	CIKM	0.7425	**0.7601**	**0.7657**	**0.7706**
w/o	CIKM	0.7425	0.7541	0.7629	0.7681

Table 2: An ablation test on the dropout function of diffusion models: "w" indicates that dropout is enabled, whereas "w/o" signifies that dropout is disabled. ML-1M stands for MovieLens-1M. DCN is used as the backbone model.

dropout provides minimal improvement. In contrast, on the CIKM 2019 dataset, incorporating dropout into the diffusion process positively impacts and enhances the model's performance.

Proofs

Definition of $q_\sigma(\mathbf{z}_{1:T}|\mathbf{z_0}, \mathbf{h})$

We rewrite the definition of $q_\sigma(\mathbf{z}_{1:T}|\mathbf{z_0}, \mathbf{h})$ in the paper:

$$q_\sigma(\mathbf{z}_{1:T}|\mathbf{z}_0, \mathbf{h}) := q_\sigma(\mathbf{z}_T|\mathbf{z}_0, \mathbf{h}) \prod_{t=2}^{T} q_\sigma\left(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{h}, \mathbf{z}_0\right) \tag{1}$$

Definition of $q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{h})$

As shown in our main paper, we have defined:

$$q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{h}) = \mathcal{N}(\mathbf{z}_{t-1}|\kappa_t \mathbf{z}_t + \lambda_t \mathbf{z}_0 + \nu_t \mathbf{h}, \sigma_t^2 \mathbf{I}) \tag{2}$$

where,

$$\kappa_t = \sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}$$

$$\lambda_t = \sqrt{\alpha_{t-1}} - \sqrt{\alpha_t}\sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}$$

$$\nu_t = \sqrt{c_{t-1}} - \sqrt{c_t}\sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}$$

Lemma 1. *Given the definitions of $q_\sigma(\mathbf{z}_{1:T}|\mathbf{z}_0, \mathbf{h})$ in Equation (1) and $q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{h})$ in Equation (2), we have:*

$$q_\sigma(\mathbf{z}_t|\mathbf{z}_0, \mathbf{h}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{c_t}\mathbf{h}, (1-\alpha_t)\mathbf{I}) \tag{3}$$

Proof. Following (Song, Meng, and Ermon 2020), we prove the statement by induction. First, for $t = T$, we already have:

$$q_\sigma(\mathbf{z}_t|\mathbf{z}_0, \mathbf{h}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{c_t}\mathbf{h}, (1-\alpha_t)\mathbf{I}) \tag{4}$$

To prove the Equation (3) holds for $t < T$, we have

$$q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_0, \mathbf{h}) := \int_{\mathbf{z}_t} q_\sigma(\mathbf{z}_t|\mathbf{z}_0, \mathbf{h}) q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{h}) \tag{5}$$

Since $q_\sigma(\mathbf{z}_t|\mathbf{z}_0, \mathbf{h})$ and $q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{h})$ are both Gaussian, from (Bishop 2006) (Equation 2.115), we know that $q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_0, \mathbf{h})$ is also Gaussian. We denote it as $\mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$. where

$$\begin{aligned}\mu_{t-1} &= \kappa_t(\sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{c_t}\mathbf{h}) + \lambda_t \mathbf{z}_0 + \nu_t \mathbf{h} \\ &= \sqrt{\alpha_{t-1}}\mathbf{z}_0 + \sqrt{c_{t-1}}\mathbf{h} \end{aligned} \tag{6}$$

$$\Sigma_{t-1} = \sigma_t^2 \mathbf{I} + \frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}(1-\alpha_t)\mathbf{I} \tag{7}$$

$$= (1-\alpha_{t-1})\mathbf{I} \tag{8}$$

Therefore,

$$q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_0, \mathbf{h}) = \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{z}_0 + \sqrt{c_{t-1}}\mathbf{h}, (1-\alpha_{t-1})\mathbf{I}) \tag{9}$$

By applying induction, we establish that Equation (3) holds for $t \leq T$. □

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*.

Qu, Y.; Cai, H.; Ren, K.; Zhang, W.; Yu, Y.; Wen, Y.; and Wang, J. 2016. Product-Based Neural Networks for User Response Prediction. In *ICDM*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.