GIFT: Graph-guIded Feature Transfer for Cold-Start Video Click-Through Rate Prediction

Sihao Hu*
husihao26@zju.edu.cn
Zhejiang University

Yi Cao*
dylan.cy@alibaba-inc.com
Alibaba Group

Yu Gong
gy910210@163.com
Alibaba Group

Zhao Li
lizhao.lz@alibaba-inc.com
Alibaba Group

Yazheng Yang
yazheng_yang@zju.edu.cn
Zhejiang University

Qingwen Liu
xiangsheng.lqw@alibaba-inc.com
Alibaba Group

Wengwu Ou
santong.oww@taobao.com
Alibaba Group

Shouling Ji†
sji@zju.edu.cn
Zhejiang University

ABSTRACT

Short video has witnessed rapid growth in China and shows a promising market for promoting the sales of products in e-commerce platforms like Taobao. To ensure the freshness of the content, the platform needs to release a large number of new videos every day, which makes the conventional click-through rate (CTR) prediction model suffer from the severe item cold-start problem.

In this paper, we propose **GIFT**, an efficient **G**raph-gu**I**ded **F**eature **T**ransfer system, to fully take advantages of the rich information of warmed-up videos that related to the cold-start video. More specifically, we conduct feature transfer from warmed-up videos to those cold-start ones by involving the physical and semantic linkages into a heterogeneous graph. The former linkages consist of those explicit relationships (e.g., sharing the same category, under the same authorship *etc.*), while the latter measure the proximity of multimodal representations of two videos. In practice, the style, content, and even the recommendation pattern are pretty similar among those physically or semantically related videos. Besides, in order to provide the robust id representations and historical statistics obtained from warmed-up neighbors that cold-start videos covet most, we elaborately design the transfer function to make aware of different transferred features from different types of nodes and edges along the metapath on the graph. Extensive experiments on a large real-world dataset show that our GIFT system outperforms SOTA methods significantly and brings a 6.82% lift on click-through rate (CTR) in the homepage of Taobao App.

1 INTRODUCTION

With the advent of video streaming apps such as Tiktok and Kwai, shooting short videos has become a much more popular way for everyone to share emotions, record lifestyle, spread news, and even advertise for goods. People can get the information they want from short videos of about 15 seconds anytime and anywhere, which significantly improves the efficiency of one's fragmented time use. In e-commerce platforms such as Taobao, a short video is produced by video authors to share a lifestyle or showcase an item, and will be displayed in the homepage's feeds stream to attract users' attention for their potential click and payment. Short videos in Taobao have witnessed rapid development in the past year: the number of short videos has grown five-fold to over ten million and the velocity of daily new video production has accelerated from tens of thousands to hundreds of thousands per day, which has greatly improved the richness and freshness of short videos, as well as the coverage for products.

Behind this, a video recommendation system plays a key role in accurately recommending appropriate short videos for a target user. Unfortunately, like most recommendation systems, it suffers from item cold-start problem[6] when faces a large number of new videos released every day. We analyze the reasons for the existence of the cold-start problem in recommendation systems: Mainstream methods such as collaborative filtering algorithms[11, 18, 27] require historical user-item interactions to calculate item co-occurrence relationship, which may lead to an unsatisfied result that new items without any user interactions can never be recommended. Deep learning methods[2, 26, 29] learn a meaningful id embedding for an entity with at least 5-10 occurrences[7] of that item are needed in the data, which is far beyond the average interaction times for the newly arrived videos. Furthermore, an unbalanced percentage of cold-start videos in the total can cause the model to overweight the id representation and statistical feature, which leads to poor performance on new videos.

Although considerable efforts have been made to cold-start problems in recommendation systems, most of them fall into the part of content-based methods[20, 22, 31], i.e., the performance lift of these models mainly comes from the introduction of the content information like category, text caption, image, and video representation. However, in the real industrial application, information mentioned above has long been incorporated into Taobao's feature systems to accurately portray the item, suggesting that content-based algorithms are no longer the silver bullet in our scenario (Taobao platform). Alternatively, the approach we seek is expected to capture some high-level information to represent a new video more precisely with the limited information.

To further step toward the above methods, we present **G**raph-gu**I**ded **F**eature **T**ransfer (GIFT) system, with a straightforward idea to construct linkages to guide feature transfer from warmed-up videos to cold-start ones. Precisely, the GIFT system consists of

*The first two authors are of equal contribution and in no particular order.
†Shouling Ji is the co-corresponding author.

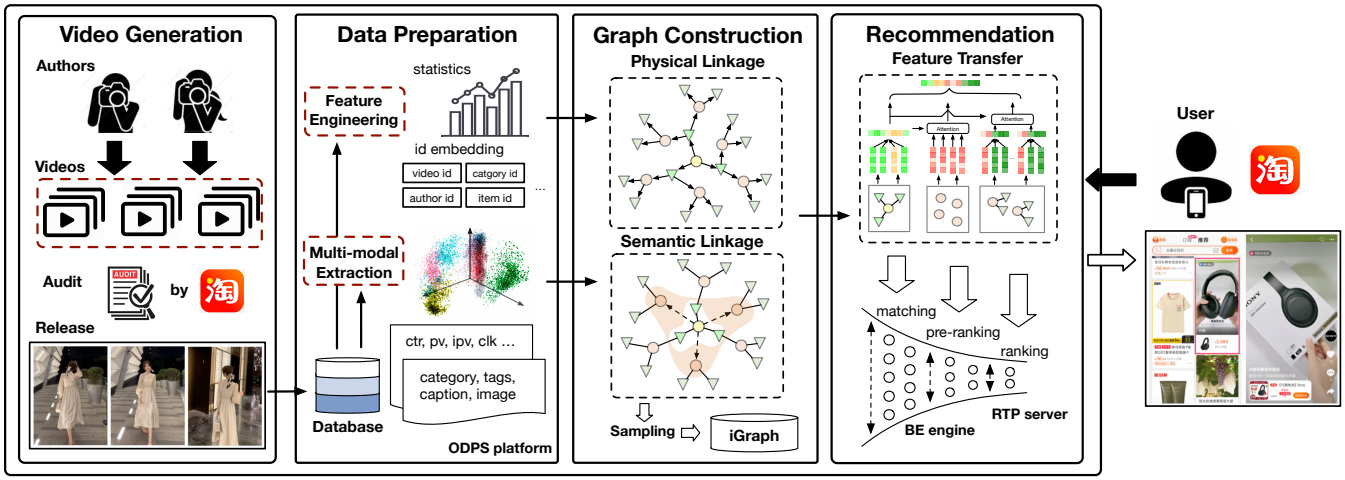


Figure 1: The workflow of our cold-start video recommendation system in Taobao App consists of video generation, data preparation, graph construction, and recommendation. The GIFT system works for graph construction and feature transfer process.

a graph construction module and a feature transfer model (GIFT network). Firstly, the graph construction module utilizes a heterogeneous graph to construct several types of linkages (physical and semantic linkage) between video and attribute nodes and several types of linkages. A physical linkage is built between a video node and an attribute node if the video has a certain attribute value. Based on this setting, the path "V-A-V" indicates two videos that share the same attribute value, e.g., same author or item. Empirical analysis shows that nodes linked by these relationships are not only similar in style and content, but also have similar performance (statistics) in e-commerce platforms, which sheds light for us to take the stable and robust feature (representation and statistics) of those warmed-up videos to compensate for that of neighbored new videos. By building the physical linkage, we are able to link more than 95% of cold videos to at least one warm video, However, two concerns are raised about this construction way: 1. We cannot assure that all cold videos can link to enough warm videos for effective transfer ($\geq$5); 2. The look-liked similarity does not necessarily guarantee the shortest vicinity in the semantic space. To address the problems, we use a pre-trained multi-modal model to jointly encode the title and cover image of the video into a semantic space, where a vector represents a video. By simply applying kNN search, we acquire k videos related to the target cold video without the requirement of a physical linkage.

Guided by these two types of linkages, we are able transfer features from neighbored nodes, not only the id representations of warmed-up video nodes but also historical statistics from the attribute nodes like page view (PV), click-through rate (CTR), which are exactly the most lacking information for a cold video. Secondly, we present the feature transfer mechanism (GIFT network) to transfer information precisely. We elaborately design the transfer function for different hops or types of nodes that are visited along a metapath from the target video node, i.e., a sequence of nodes linked by pre-defined types of edges in order (a structural path at the meta-level). Concretely, we apply the attention mechanism[23] to extract the most informative feature from the video nodes and attribute nodes separately to ensure the awareness of different types of transferred features. Moreover, two classes of linkages and different types of physical linkage are also considered in the transfer process as illustrated in Section 3.2. It is noted that our approach uses a jumping-connection schema to directly transfer the information of nodes on the metapath, which sets GIFT network apart from standard GNNs[10, 13]. Finally, to acquire an effective model, especially robust id representation of warm videos, we pre-trained our model on the whole set of videos firstly and then fine-tune it on the logs of cold-start videos for domain adaptation. By utilizing the above-mentioned instance- and model-based transfer strategies, GIFT achieves **1** AP gain of AUC average over the base model, which is a huge boost for the click-through rate prediction task.

Additionally, the whole GIFT system has been deployed on Taobao, the biggest e-commerce platform in China, serving for the cold-start video recommendation in the section of guess you like in homepage of Taobao App. The gain of **6.82**% on CTR further illustrates the effectiveness of our system.

Contributions. To summarize, the contributions are as follows:

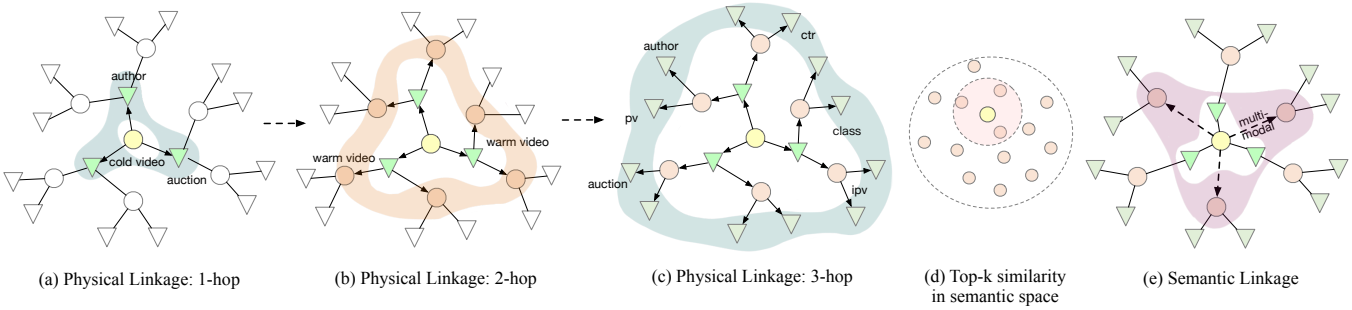
- We present a novel heterogeneous graph construction method to build physical and semantic linkages between the warmed-up videos and cold-start ones.
- We propose the GIFT network to transfer the information from the warmed-up videos to cold-start ones, and validate its effectiveness on a large real-world dataset. The source code will be released on GitHub[1].
- We implement and deploy the whole GIFT system on the homepage of Taobao App. Online AB test shows it achieves significant performance improvements for cold-start video recommendation.

2 PRELIMINARY

2.1 Item Cold-start Problem

In this paper, we focus on addressing the item-side cold-start problem for CTR prediction task, which refers to the problem that new items has no or rare prior events, like rating or click logs, make

[1]https://github.com/Bayi-Hu/GIFT-Graph-guided-Feature-Transfer-Network



(a) Physical Linkage: 1-hop (b) Physical Linkage: 2-hop (c) Physical Linkage: 3-hop (d) Top-k similarity in semantic space (e) Semantic Linkage

Figure 2: Construction rules for two types of linkages. (a)-(c): physical linkage construction; (d): toy example for top-k similarity in the semantic representation space;(e): semantic linkage construction based on top-k similarity.

them miss the opportunity to be recommended and remain "cold" all the time.

In our application, the term "*item*" refers to video, and a *new* or *cold-start video* is defined as a short video released less than 3 days (inclusive) in Taobao, and a *warmed-up video* is defined as a video released more than 3 days. It should be noted that our GIFT system can be easily extended to other item cold-start recommendation applications, e.g., the product, advertisement CTR prediction tasks.

2.2 Workflow of Cold-start Video Recommendation System

Figure 1 gives a brief illustration of the workflow of cold-start video recommendation system in Taobao, which consists of four processes: video generation, data preparation, graph construction, and recommendation. Our GIFT system works for the latter two parts. Specifically, in video generation process, after produced by authors and audited by the platform, new videos can be added up to the video pool and recommended to users. Then followed is the data preparation process, which calculates the statistics of videos periodically, as well as other id information and the multi-modal representations extracted based on videos' caption/title and cover image. These content information are especially useful for cold-start recommendation since the statistics and id representation of a new video are usually missing and unrobust. Third, our GIFT system constructs a heterogeneous graph by building physical and semantic linkages between new and warmed-up videos. The computation graph for each cold video is pre-sampled and stored in Alibaba's iGraph database to support real-time inference. Finally, in recommendation process, we use GIFT network to enhance the robustness of video feature to precisely predict the CTR for each item given the user. A typical industrial recommendation procedure in Alibaba follows a multi-stage paradigm to trade off the accuracy and computation overhead (for more details about our system implementation, please refer to Appendix-A). In each stage, the model predicts CTR for each given item and then selects top ranked ones for the following stages. In theory, GIFT network works for all the stages. In practice, we implement and deploy it in the "ranking" stage, powered by Alibaba's RTP server.

3 GRAPH CONSTRUCTION

Graph has been a very popular structure to represent the ubiquitous relationships between entities in e-commerce[8, 13]. In GIFT system,

(a) Target Video (b) Same Item (c) Same Author (d) Semantic Similar

Figure 3: Videos connected by physical and semantic linkages: (a): the target cold video; (b): A warmed-up video share the same item with (a); (c): A warmed-up video share the same author with (a); (d): A warmed-up video connected by the semantic linkage to the (a).

we construct a heterogeneous graph consisting of two types of nodes and multiple types of edges to guide the subsequent feature transfer process. Concretely, two types of link construction methods are introduced to ensure the effectiveness of subsequent feature transfer and a high coverage rate for the cold-start videos. It is noted that a video node uniquely identifies a video and an attribute node represents a specific attribute value like category id or historical IPV.

3.1 Physical Linkage

The overall construction process of physical linkage can be divided into three sub-steps: one-hop, two-hop, and three-hop linkage construction as illustrated in Figure 2(a)-(c), where the central round yellow node represents the target cold-start video.

The **one-hop linkages** are naturally set between a video and some attribute nodes(denoted as green triangle nodes) as in Figure 2(a) and denote that a video has certain attribute values, e.g., a

video has an author id of its producer and belongs to a specific class of style, along with statistics like average stay time that calculated based on historical user visited logs. If we directly make use of these information without expanding the computation graph outward, it is a typical implementation of the content-based methods.

The **two-hop linkages** expand from certain types of attribute nodes to video nodes (pink round nodes) as in Figure 2(b). Empirically, videos introducing the same item may have very similar content and thus more likely to be clicked by the same group of users. With the assumption that videos nodes sharing the same attribution may follow the same recommendation pattern. The edge between video node and attribution node represents how tight the connection is. In our implementation, the video nodes connected by two-hop linkages are constraint as the warmed-up videos, which have been released for more than 3 days so that they are able to accumulate enough user behavior logs and their statistics are reliable. By building the two-hop linkage we are able to link more than 95% of cold videos to at least one warm video.

If the CTR model solely takes the computation graph up to this point as input, it can be regarded as a representation transfer model, which can transfer relatively robust representations from the two-hop linked video nodes. As presented in Figure 2(c), we take a step outward by setting **three-hop linkage** from the outermost warmed-up video nodes to neighbored attribute nodes, just the same rule as the one-hop linkage. Three-hop linkage can link to attribute nodes that represent informative value of the warmed-up videos like historical statistics like page view (PV), average stay time, click-through rate (CTR), etc., which are exactly what cold-start video lacks and needs to be made up for.

3.2 Semantic Linkage

The above method solely focuses on the linkages that physically exist, and though analysis does show the similar style and content shared by video nodes linked physically, there are still some concerns about this type of construction way: 1. We cannot assure that all cold videos can link to the warm video; 2. The look-liked similarity does not necessarily guarantee the shortest vicinity in the semantic space.

In this section, we construct the **semantic linkage** to solve these two problems further. By taking two factors into account, i.e., the title and cover image of the video, we use a pre-trained multimodal model lxmert[19] to encode these two types of information into a semantic space, where a vector represents a video as presented in Figure 2(d). Consequently, kNN search can be easily applied into this space to acquire k videos that relate to the target video without any requirement of physical relationship. Moreover, the vicinity calculated through metrics like cosine distance ensures the top-k similarity in the multimodal semantic space. By taking the semantic-neighbored videos as the source node, we further links the attribute nodes one-step outward, which represents the historical statistics of PV, CTR and so on, just the same rule as the three-hop physical linkage in Section 3.1.

We test the similarity of linked videos with some randomly selected cases in Figure 3, which suggests a relatively high correlation between the target video and its neighbored videos both in terms of style and content.

4 CTR PREDICTION MODEL

The existing CTR prediction models[2, 28, 29] can be divided into user-side and item-side categories. This paper mainly focuses on the item-side model designing since it is directly related to addressing the item cold-start problem. It is worth noting that the proposed GIFT network does **not couple with** any user-side model. Yet to facilitate the description, we select DIN[29] as our base model.

4.1 Base Model: DIN

DIN has been successfully applied in Taobao. It mainly proposes the target attention mechanism to adaptively learn the representation of user interests from historical behaviors w.r.t. the target item, as shown in the left of Figure 4. It consists of several parts:

Embedding Layer. The input of DIN in our scenario contains several categorical id features, e.g., video id, item id, etc. These id features cannot be directly input into the model because of their extremely high dimensionality. For instance, the number of item ids is about billions. Therefore, the widely-used embedding technique is adopted to embed the original sparse features into low-dimensional dense vectors, which significantly eases the model computing process. For user and items (both the target item and items in user behavior sequence), id embeddings are concatenated with its other continuous features together to generate the overall embedding as depicted in the left of Figure 4.

Target Attention Mechanism. The original DIN proposes a local activation unit to calculate the representation vector of user interests by considering the relevance of historical behaviors w.r.t. the candidate item, which means the representation vector varies over different candidate items. This local activation unit can be re-implemented under the framework of attention[23] by considering the representation of target item $\boldsymbol{h}_t$ as the query and $\boldsymbol{H}_u$ as key and value, where $\boldsymbol{H}_u = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_H\}$ is the set of embedding vectors of items in the user behaviors with length of H. Specifically, we formalize it as follows:

$$\boldsymbol{h}_u = \text{Attention}\left(\boldsymbol{h}_t \boldsymbol{W}^Q, \boldsymbol{H}_u \boldsymbol{W}^K, \boldsymbol{H}_u \boldsymbol{W}^V\right) \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{2}$$

where the projections matrices $\boldsymbol{W}^Q \in \mathbb{R}^{d\times d}$, $\boldsymbol{W}^K \in \mathbb{R}^{d\times d}$, $\boldsymbol{W}^V \in \mathbb{R}^{d\times d}$ are learnable parameters and d is the dimension of hidden space. $\boldsymbol{H}_u$ can be regarded as a matrix $\in \mathbb{R}^{d\times H}$, and $\boldsymbol{h}_t \in \mathbb{R}^d$ is the hidden representation of target item that passed through an multi-layer network. The temperature $\sqrt{d}$ is introduced to produce a softer attention distribution for avoiding extremely small gradients. This mechanism can be called as **target attention**.

Loss. The objective function used in DIN and GIFT is both the negative log-likelihood function defined as:

$$L = -\frac{1}{N}\sum_{(x,y)\in\mathcal{D}} (y \log f(x) + (1-y)\log(1-f(x))) \tag{3}$$

where D is the training set, with x as the input of the model and $y \in \{0, 1\}$ as the ground-truth label, $f(x)$ is the output after the softmax layer that represents the predicted probability whether sample x is clicked.

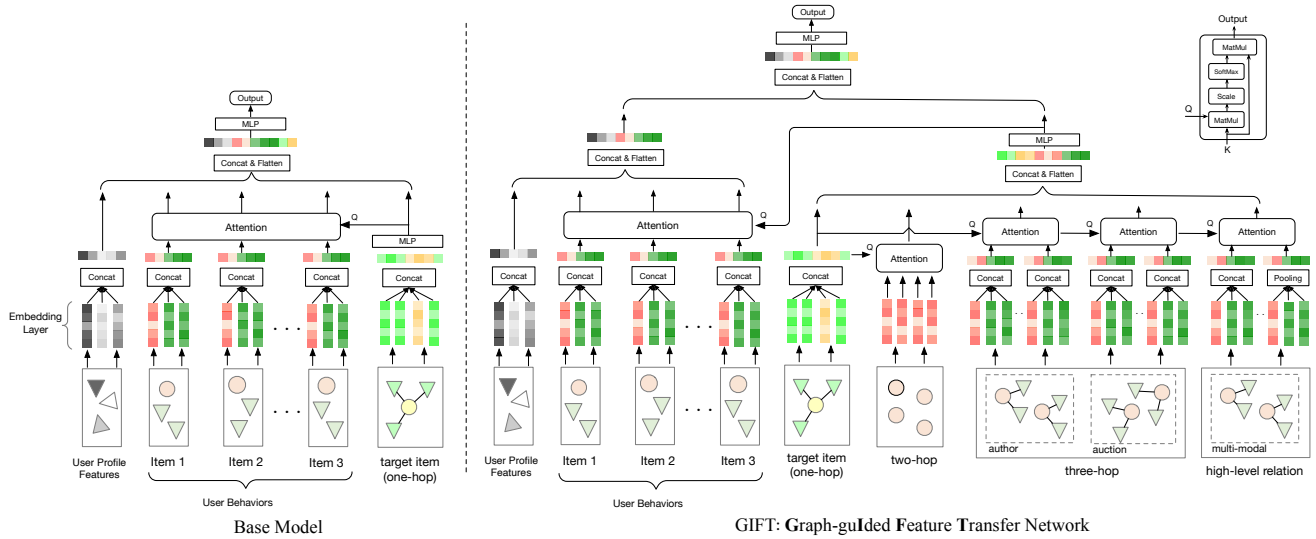


Figure 4: Metapath-guided Graph Transfer Neural Network: Left is the user-side model and right is the item-side model.

4.2 GIFT: Graph-guided Feature Transfer

The overall structure of the proposed GIFT network is demonstrated in Figure 4, which shares a similar user-side structure as DIN. Given the constructed computation graph for a candidate video, we need to clearly distinguish the hop of a node and the type of edges it connected with to apply different feature transfer strategies since different types of nodes contain different types of features and nodes connected by different linkages follow some certain patterns that need to be modeled separately. Therefore, we extract these heterogeneous relation paths to guide the feature transfer process with the help of metapath.

Metapath[3], defined as a relation sequence connecting two objects, is proposed to capture the specific structural relation between objects. Figure 2(e) shows two types of metapath we used in the feature transfer procedure: "$\rho_1 = v_t \xrightarrow{p} a \xrightarrow{p} v \xrightarrow{p} a$" indicates a target video node v_t shares the attribute nodes with warmed-up video nodes who connect with more abundant attribute nodes like historical statistics; "$\rho_2 = v_t \xrightarrow{s} v \xrightarrow{p} a$" indicates the video node v_t connects to the warmed-up video nodes through the semantic linkage, who also link to attribute nodes through the physical linkage. p and s represent the physical and semantic linkage, respectively.

Metapath has been widely used in graph mining community to constrain a meta structure in a heterogeneous graph that represents a specific semantic pattern. We further introduce the concept of metapath-guided Neighbors to represent sets of nodes visited along the given metapath ρ since we focus more on nodes to transfer the feature than the metapath itself. ·

Definition 1. ***Metapath-guided Neighbors**[3]. Given a node o and a metapath ρ (start from o) in the graph, the metapath-guided neighbors is defined as the set of all visited objects when the object o walks along the given metapath. In addition, we denote the i-th step neighbors of object o as $\mathcal{N}_\rho^{(i)}(o)$. Taking Figure 2(e) as an example, given the metapath "$\rho_1 = v_t \xrightarrow{p} a \xrightarrow{p} v \xrightarrow{p} a$", we can get metapath-guided neighbors as $\mathcal{N}_{\rho_1}^{(1)}(v_t) = \{a_1, a_2\}, \mathcal{N}_{\rho_1}^{(2)}(v_t) = \{v_1, v_2, v_3\}$. $\mathcal{N}_{\rho_1}^{(0)}(v_t)$ is v_t itself.*

It is worth noting that a video node represents the video id of itself and the attribute node represents a certain attribute value like category id. Following this setting, we use set of nodes to describe features used in our model, e.g., $\{v_t\} \cup \mathcal{N}_{\rho_1}^{(1)}(v_t)$ represents the id and attributes of video v_t.

Metapath-guided Feature Transfer. With the help of metapath-guided neighbor, it is clear to describe the following feature transfer process. To begin with, we obtain a representation vector $\boldsymbol{h}_t$ by embedding layer to represent the original information of target video v_t, which contains both id embedding and dense features like historical statistics. For a cold-start item, this information is not robust because of very limited data. Firstly, we transfer the id representation of the warmed-up video nodes to $\boldsymbol{h}_t$. The id representation of a warmed-up video is calculated based on sufficient user behavior logs, thus containing abundant information to transfer. The neighbored video nodes come from $\mathcal{N}_{\rho_1}^{(2)}(v_t)$ and $\mathcal{N}_{\rho_2}^{(1)}(v_t)$ and the generation of id representation $\mathbb{E}^{(2)}$ can be formalized as:

$$\mathbb{E}^{(2)} = \left\{ \text{Embed}(v_i) \,|v_i \in \mathcal{N}_{\rho_1}^{(2)}(v_t) \cup \mathcal{N}_{\rho_2}^{(1)}(v_t) \right\} \tag{4}$$

where $\mathbb{E}^{(2)}$ is the set of id representations of node $v_i \in \mathcal{N}_{\rho_1}^{(2)}(v_t) \cup \mathcal{N}_{\rho_2}^{(1)}(v_t)$. $\mathbb{E}^{(2)}$ can be seen as a matrix $\in \mathbb{R}$ and Embed$(*)$ denotes the embedding layer operation. Then, instead of mean or sum pooling $\mathbb{E}^{(2)}$ into a vector, we adopt the idea of target attention as follows:

$$\boldsymbol{h}^{(2)} = \text{Attention}\left(\boldsymbol{h}_t \boldsymbol{W}_2^Q, \mathbb{E}^{(2)} \boldsymbol{W}_2^K, \mathbb{E}^{(2)} \boldsymbol{W}_2^V\right) \tag{5}$$

This adaptive pooling strategy obtains more importance based on the dot-product of $\boldsymbol{h}_t$ and $\boldsymbol{h}_i, v_i \in \mathcal{V}_{(2)}$, which forces the model to pay more attention to nodes that have more similar representations to $\boldsymbol{h}_t$, thus plays a role to automatically filter those most informative id representations for v_t.

Since $\boldsymbol{h}^{(2)}$ does not contain any statistical information of the warm videos that are deficient for a cold-start video, we take a further step to get this information involved. To transfer statistical feature to the target cold item, similarly, we generate the set of

Table 1: Statistics of the three datasets.

Dataset		# Users	# Items	# Samples	Edge Type	# Edges	Path Type	# Paths
Taobao	full	4.98×10^7	2.2×10^7	5.78×10^8	V-A	2.2×10^7	V-A-V	1.9×10^8
	cold	3.0×10^7	4.8×10^5	1.38×10^8	V-I	2.1×10^7	V-I-V	5.7×10^7
	test	2.1×10^6	1.2×10^5	9.4×10^6	V-V	2.8×10^8	V-V	2.8×10^8
MovieLens	full	6, 038	3, 671	943, 022	M-A	15, 398	M-A-M	54, 164
	cold	5, 884	729	154, 562	M-D	4, 210	M-D-M	13, 448
	test	5, 252	674	38, 640				
DBook	full	10, 534	20, 934	631, 349	B-A	20,934	B-A-B	210, 400
	cold	7, 890	3, 207	72, 130				
	test	5, 409	2, 861	18, 032				

video representation $\mathbb{E}^{(3)}$ as follows:

$$\mathbb{E}_r^{(3)} = \begin{cases} \left\{ \text{Pool}\left(\text{Embed}\left(\{v\} \cup \mathcal{N}_{\rho_1,r}^{(1)}(v)\right)\right) | v \in \mathcal{N}_{\rho_1,r}^{(2)}(v_t) \right\}, r \in \mathcal{R}_p \\ \left\{ \text{Pool}\left(\text{Embed}\left(\{v\} \cup \mathcal{N}_{\rho_2}^{(1)}(v)\right)\right) | v \in \mathcal{N}_{\rho_2}^{(1)}(v_t) \right\}, r \in \mathcal{R}_s \end{cases} \tag{6}$$

where $\mathcal{N}_{\rho_1,r}^{(2)}(v_t), r \in \mathcal{R}_p$ refers to the set of video nodes that connected with edge type r on metapath ρ_1, $\mathcal{R}_p$ is the set of physical edge types and p is a certain type of physical linkage, e.g., item and author; $\mathcal{N}_{\rho_2}^{(2)}(v_t)$ refers to the set of video nodes connected by the semantic edges on metapath ρ_2 and in this case $r \in \mathcal{R}_s = \{S_{mm}\}$. Then we give our three-hop transfer function as follows:

$$\boldsymbol{h}_r^{(3)} = \text{Attention}\left(\boldsymbol{h}_t \boldsymbol{W}_3^Q, \mathbb{E}_r^{(3)} \boldsymbol{W}_3^K, \mathbb{E}_r^{(3)} \boldsymbol{W}_3^V\right) \tag{7}$$

$$\boldsymbol{h}^{(3)} = \boldsymbol{h}_{r_1}^{(3)} \oplus \boldsymbol{h}_{r_2}^{(3)} \oplus \ldots \oplus \boldsymbol{h}_{r_i}^{(3)}, r_i \in \mathcal{R}_p \cup \mathcal{R}_s \tag{8}$$

It is noted that attentions are computed separately on $\mathbb{E}_r^{(3)}$ according to different types of linkages the node connected with. This strategy bases on the observation that patterns of a set of videos vary from the types they relate with. For example, videos showing the same item may have more similar content about the product but be shot in a very different style compared to videos made by the same author. The model will be unable to be aware of the difference caused by this factor if all the video nodes are mixed for feature transfer. Finally, we concatenate $\{\boldsymbol{h}_r^{(3)} | r \in \mathcal{R}\}$ to a vector to acquire the whole $\boldsymbol{h}^{(3)}$ in Eq. 8, where $\oplus$ represents the concatenate operation.

Given v_t's original representation $\boldsymbol{h}_t$ and transfer embedding $\boldsymbol{h}^{(2)}$, $\boldsymbol{h}^{(3)}$ that encodes different types of information from the warmed-up videos, we concatenate and pass them through a MLP-layer to obtain the final representation $\boldsymbol{h}'_t$, and take it as the query vector for the user-side target attention calculation same as DIN. The remaining settings of the network structure are just the same as the base model. It is noted that due to the robustness of $\boldsymbol{h}'_t$, the efficiency of target attention is enhanced either.

4.3 Training Strategy

Pre-training. GIFT network serves for the short video newly launched on the Taobao, yet in order to acquire an effective model, we must pre-train the model on the whole set of videos, since it is needed to firstly generate robust representations for the warmed-up videos that frequently appeared both in the users' video sequence and the neighborhood of the target new video. Moreover, because of the unbalanced distribution of the new and old video, recommend logs collected from the new video (around 1/10 daily) are insufficient to train a model with a good generalization performance.

Fine-tuning. However, domain bias between new and old videos not only exists in terms of data volume but also in the recommendation pattern: model trained on the logs of old videos more rely on the id embedding or historical statistics compared to the model trained on the new videos, which should assign more weights on the content information to well generalize to the cold-start videos. Therefore, we fine-tune our pre-trained model on the logs of cold-start videos collected over the past 1 month to obtain sufficient training logs. By using this adaptation strategy, our model boosts the offline performance by 0.86% in terms of AUC (Table 2).

5 EXPERIMENTS

In this section, we perform a series of offline and online experiments to answer the following research questions one by one:

- **Q1**: How does GIFT perform on CTR prediction task for cold-start items compared to state-of-the-art baselines?
- **Q2**: How do different parts of the model/feature/graph contribute to the final performance lift of GIFT? Is it easy to generalize to other methods?
- **Q3**: How is the effectiveness and efficiency of the GIFT system in Taobao's cold-start video recommendation application?

5.1 Datasets and Competitors

Taobao Dataset. Taobao dataset includes hundreds of millions of user interaction logs with short videos in the homepage of Tabao App. The dataset consists of two parts: D_{full} and D_{cold}, where D_{full} is uniformly sampled from 15-day user interaction logs of all videos, and D_{cold} is a subset of D_{full} that only includes user interaction logs on new videos that launched less than seven days. Both D_{full} and D_{cold} are used for the training phase. We further sample the testing set D_{test} from the following one-day logs after D_{full} and D_{cold} are collected, which also only includes logs of cold videos. It is worth noting that the collection of this dataset strictly simulates the online environment.

MovieLens Dataset. MovieLens (1M) dataset is a public dataset collected from IMDb, which contains 6,040 users, 3881 movies, 21 categories, and nearly 1,000,000 samples. To make it suitable for CTR prediction task, we follow the setting of DIN[29] by labeling the samples with rating of 4 and 5 to be positive and the rest to be negative. Following the setting of MetaHIN[15], we regard movies released after 1998 as new items. For this dataset, we construct "Movie-Author" and "Movie-Director" relationship as the physical linkage.

Table 2: Comparison of classification performance on three dataset

Type	Model	Taobao		MovieLens		DBook	
		AUC	RelaImpr	AUC	RelaImpr	AUC	RelaImpr
Handcrafted Features	LR	0.7218	-13.63%	0.7693	-3.54%	0.7531	-9.25%
	SVM	0.7339	-8.92%	0.7722	-2.51%	0.7673	-4.16%
	GBDT	0.7377	-7.44%	0.7734	-2.08%	0.7687	-3.66%
DNNs	DNN	0.7423	-5.65%	0.7753	-1.40%	0.7739	-1.75%
	Wide&Deep	0.7465	-4.01%	0.7757	-1.25%	0.7754	-1.25%
	DeepFM	0.7508	-2.33%	0.7759	-1.18%	0.7760	-1.04%
	DIN	0.7568	0.00%	0.7792	0.00%	0.7789	0.00%
Cold-Start Methods	DropOutNet	0.7573	0.19%	0.7791	-0.04%	0.7796	0.25%
	ACCM	0.7550	-0.70%	0.7798	0.21%	0.7810	0.75%
Ours	GIFT	**0.7670**	**3.97%**	**0.7827**	**1.25%**	**0.7861**	**2.58%**
	GIFT[1]	**0.7693**	**4.87%**	-	-	-	-

DBook Dataset. DBook dataset is collected from the Douban website in China. We label the samples with rating of 5 to be positive and the rest to be negative. The definition of the new book follows the setting of MetaHIN[15]. For this dataset, we construct "Book-Author" relationship as the physical linkage.

MovieLens and DBook datasets are also divided into D_{full} and D_{cold}, and testing sets are randomly separated from D_{cold}. For all the methods, we train them on the D_{full} and test on D_{test}. This is based on an observation that the performance of all methods trained on D_{full} is significantly better than that of D_{cold}, because of the larger sample size of D_{full}. D_{cold} is only used for fine-tuning strategy as described in Section 4.3.

Competitors. We compare proposed GIFT network with three types of baselines: the first type is conventional machine-learning methods based on handcrafted features, like Logistic Regression (LR), Support vector machine (SVM), Gradient Boosting Decision Tree (GBDT). The second type is DNN-based methods, including:

- DNN[2]: a video recommendation approach proposed by Youtube and has been widely adopted in industrial application.
- Wide&Deep[1]: a method proposed by Google that adopted a wide model to tackle manually designed cross-product features and a deep model to capture high-order feature interactions.
- DIN[29]: our base model as described in Section 4, which uses target attention to learn the representation of user interests from historical behaviors w.r.t. the target item.

As for the third type of baselines, we implement two cold-start approaches:

- DropOutNet [24]: a method that applies dropout technique to make DNNs generalize to the missing input that is common for the cold-start items. DIN is implemented as the base model of DropOutNet.
- ACCM[17]: a hybrid model that attentively integrates id embedding and content information into one vector to adapt to both the warm and cold items. For the fairness of comparison, we re-implement its base model as DIN.

We do not put any pure content-based models into the competitors because the existing feature systems in Taobao has included abundant content information long before.

5.2 Performance Comparison (Q1)

In the field of CTR prediction, AUC value is widely used as a metric of goodness of order by ranking all the items with predicted CTR[28, 29]. We further adopt the idea of RelaImpr[29] to measure the relative improvement over models. Since for a random guesser, the value of AUC is 0.5, the RelaImpr is defined as below:

$$\text{RelaImpr} = \left(\frac{\text{AUC(measured model)} - 0.5}{\text{AUC(base model)} - 0.5} - 1 \right) \times 100\% \quad (9)$$

Table 2 shows the results of ten methods on three testing datasets. Obviously, all the DNNs beat non-deep learning models significantly, which demonstrates the power of deep learning in recommendation field. For DNN-based methods, Wide&Deep with elaborately designed "wide" structure outperforms DNN, and DeepFM performs better than Wide&Deep further; DIN stands out significantly among all the DNN-based methods, especially on Taobao Dataset with rich user behaviors. We also observe that two cold-start methods do not work well on Taobao dataset: DropOutNet outperforms the base model with only 0.19% of RelaImpr, ACCM performs even worse than DIN with -0.70% of RelaImpr. Instead, GIFT achieves superior performance compared with all the competitors on all the datasets. The performance boost mainly comes from the graph construction and feature transfer mechanism, which can be observed by comparing GIFT and its base model DIN (**0.01** absolute AUC gain and **3.97**% of RelaImpr on Taobao dataset), and shows it does bring the information that cold-start videos covet most. In e-commercial recommendation systems with hundreds of millions of traffic, 0.01 absolute AUC gain is significant enough and worthy of model deployment.

It should be noted that MovieLens and DBook datasets do not contain any statistical feature for GIFT to transfer, nor do they have multimodal features to construct the semantic linkages to guide feature transfer, thus limiting the effectiveness of GIFT (1.25% and 2.58% of RelaImpr). Moreover, the definitions of new items in these two datasets are not time-sensitive, e.g., a new movie or a new book may have been on the shelves for years, which is quite different from our application. Here we just want to prove that GIFT also works for other type of item-side cold-start recommendation.

We also compare GIFT to the GIFT[1] that fine-tunes on the D_{cold} to reduce the domain bias, and an **0.86**% RelaImpr is observed on Taobao dataset. We do not report GIFT[1] on the other two datasets

Table 3: Transferred Video Feature in Taobao Dataset.

Granularity	Feature Name	Dim	Type
Representation Feature	video_id	32	one-hot
	item_id	32	one-hot
	author_id	16	one-hot
	category_id	16	one-hot
	title_token_ids	16	multi-hot
	...	...	...
Statistical Features	ctr_15d	1	numeric
	pv_cnt_15d	1	numeric
	ipv_cnt_15d	1	numeric
	clk_cnt_15d	1	numeric
	...	..	..

because this fine-tuning strategy does not work well for them. This is because the size of these two fine-tuning datasets is too small to make the model over-fit on them. In Taobao's real-world application, we have nearly 30% of the total sample available for fine-tuning.

5.3 Ablation Study (Q2)

We then study how the different elements of the feature/model/-graph contribute to performance lift. Since the public datasets differ significantly from our scenarios, we only conduct ablation studies on the Taobao dataset.

Transferred Feature-level Ablation. It is easy to wonder what kind of transferred features play the most crucial role in feature transfer. The transferred features can be divided into two categories as demonstrated in Table 3. The representation features of warmed-up video are dense vectors generated by the embedding layer and compensate for the inadequate training of representation feature of new videos. Statistical features represent the historical performance of the video in Taobao, which are completely absent from new videos. In the experiment, we remove these two types of features in turn and evaluate the performance of ablated transferred features.

The results are presented in Table 4, where we use (-) to denote the removed part. Specifically, **R**(-) means removing the representation feature from the transferred feature, and **S**(-) denotes removing the statistical features. From the table, we can observe that AUC drops 2.63% of RelaImpr, which means the performance boost of GIFT owes much to the id representation. We conjure the reason is that the representation features are not only important for new videos, but also essential for attention calculation, i.e., it is difficult to calculate the accurate attention scores just based on the fixed numeric statistical features. We also observe that S(-) drops 1.27% of RelaImpr, which suggests that historical statistics of the warmed-up video are also a very important supplement to the original feature of new videos.

Model-level Ablation. We remove transferred features generated by different phases of model in turn to evaluate the performance of ablated models. Results are presented in Table 4, where $\boldsymbol{h}^{(2)}$(-) means removing the transferred embedding $\boldsymbol{h}^{(2)}$ but keeping $\boldsymbol{h}^{(3)}$ and $\boldsymbol{h}^{(3)}$(-) vice versa. From the table, we can see that the performance drops when either $\boldsymbol{h}^{(2)}$ or $\boldsymbol{h}^{(3)}$ is removed (-1.69% and -2.85% of RelaImpr). Removing both of them impacts the performance more significantly, suggesting that the two transferred

Table 4: Ablation study on feature-, graph- and model-level

Category	Operator	AUC	RelaImpr
-	Base Model	0.7568	-3.82%
Feature	**R**(-)	0.7600	-2.62%
	S(-)	0.7636	-1.27%
Graph	PL(-)	0.7603	-2.51%
	PL(author)(-)	0.7631	-1.46%
	PL(item)(-)	0.7620	-1.87%
	SL(-)	0.7622	-1.80%
Model	$\boldsymbol{h}^{(2)}$(-)	0.7635	-1.31%
	$\boldsymbol{h}^{(3)}$(-)	0.7594	-2.85%
-	GIFT	0.7670	0.00%

embeddings work well together to generate more expressive video representations. Another observation is that the $\boldsymbol{h}^{(3)}$ contribute more than $\boldsymbol{h}^{(2)}$ to the final performance lift. This is reasonable because $\boldsymbol{h}^{(2)}$ only contains the video_id representation of neighbored videos, but $\boldsymbol{h}^{(3)}$ includes not only other id information but also the statistical feature the cold video lacks.

Graph-level Ablation. We study the contribution of each type of linkages by masking a specific type of linkages in turn. The results are presented in Table 4, where PL(author)(-) means removing the physical linkages built by the same author relation and PL(item)(-) means removing edges built by the same item relation. PL(-) means removing both of these two types of physical linkages, and SL(-) means removing the type of semantic linkages. As we can see, the performance of PL(author)(-) exceeds PL(item)(-) with 0.42% of RelaImpr, which suggests PL(item) is a stronger relationship compared to PL(author) for feature transfer. We also observe that semantic linkages contribute a lot to the performance lift (the AUC drop reaches -1.80% of RelaImpr), which suggests that physical and semantic linkages can compensate for each other to conduct more feature transfer.

5.4 Applying GIFT to Other Models (Q2)

GIFT is a model paradigm and can generally apply to various models. To prove this, we conduct an experiment to examine whether it can bring improvements to other models. Specifically, we equip it for DNN-based models like DNN, Wide&Deep, DeepFM, DIN on Taobao dataset. For DNN, Wide&Deep, DeepFM, we directly apply our methods to improve them. For DIN, it becomes our standard GIFT network. The results of AUC on Taobao dataset are shown in Figure 5. First, after applying with GIFT, all the baselines achieve better performance. This shows that item-side graph feature transfer can be easily applied to improve their performance on the cold-start CTR prediction task. Second, the performance lift of DIN (3.97% of RelaImpr) exceeds other baselines. This is because GIFT can enhance the robustness of video embedding, which improve the effectiveness of target attention by calculating more accurate attention scores between the target video and videos user have clicked.

5.5 Study of Online A/B Test (Q3)

To verify the effectiveness of the proposed method in a real-world scenario, we implement and deploy GIFT for the cold-start video

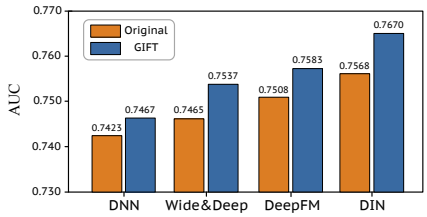


Figure 5: Performance (AUC) comparison of different models enhanced by our GIFT method on Taobao dataset.

recommendation on the homepage of Taobao App. The pipeline of our recommendation system consists of three stages: matching, pre-ranking, and ranking, equipped with Swing[27] and MIND[12] in matching, DNN[2] in pre-ranking and DIN[29] in ranking. For our recommendation system, GIFT is deployed in the ranking stage to compare with the base model DIN, and all the other conditions remain the same. For more details about our system implementation, please refer to Appendix-A. We conducted the strict online A/B test on the live application spanning from Sep. 21, 2020, to Sep. 27, 2020 and involves hundreds of millions of users per day. Online evaluation shows that GIFT has achieved **6.82%** lift on CTR metric (from 4.180% to 4.465%) with the 20% of overall traffic on the test bucket and 80% on the baseline bucket, which can bring huge economic benefits to the e-commerce platform.

5.6 Study of Online Response Time (Q3)

We evaluate GIFT's efficiency in the real-world application by comparing the time cost between the baseline system and the system equipped with GIFT. The response time (in milliseconds) with thousands of queries per second during Sep. 23, 2020 is presented in Figure 6, where the gray and green lines represent the response time of the baseline system and our system. Empirical evidence shows that the response time of our system is only about 4 ms more than that of the baseline system on average, which is due to the extra graph data reading and computational overhead boost of inference, but still far below the maximum response time of industrial recommendation systems (300 ms). Since the neighbored graph information for each cold-start video is sampled offline and pre-stored in iGraph server, and the queries of graph data and features are asynchronous, the lift in response time brought by graph reading is negligible.

5.7 Scalability Analysis (Q3)

We investigate the scalability of GIFT that has been deployed on multiple workers for optimization. Figure 7(a) shows the training time of 1 million steps w.r.t. the number of workers on Taobao dataset. The figure shows that GIFT is quite scalable on the distributed platform, as the training time decreases significantly when adding up the number of workers. Finally, when the number of workers reaches 400, the training speed of GIFT is very close to that of DIN, which indicates GIFT is scalable enough to be adopted in practice. We also investigate the scalability of GIFT w.r.t. the maximum number of neighbors for each cold-start video. As presented in Figure 7(b), the training time cost (green line) for 1 million steps has a linear growth at the beginning of the maximum number of neighbors scales up and then reaches a limit when it is larger than 50,

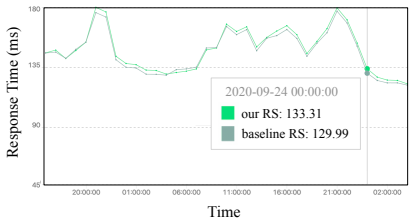


Figure 6: Comparison of Online Response Time

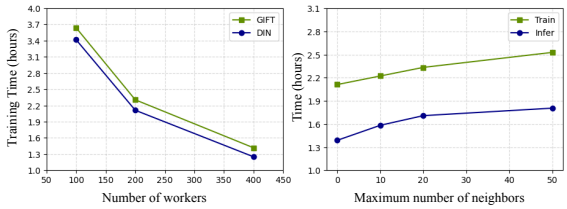


Figure 7: Study of Scalability. (a) presents the training overhead w.r.t. the number of workers; (b) presents the scalability test w.r.t the maximum number of neighbors

this is because most cold-start videos cannot be linked to more than 50 videos as the frequency of node degree follows the power-law distribution. Similar is the trend of time overhead of inference on the Taobao dataset (blue line). The empirical evidence guarantees that GIFT does not require a relatively large number of neighbors to achieve good results, thus restricting the time overhead as well.

6 CONCLUSION

In this paper, we focus on the task of CTR prediction for cold-start video recommendation in Taobao. We present an efficient graph-guided feature transfer system, GIFT, which establishes physical and semantic relations between warmed-up videos and cold-start ones. We also elaborately design the graph-guided transfer function to transfer different types of information from the warmed-up videos to the target cold-start video. When evaluated on real-world datasets, GIFT network achieves significantly better results than other competitors. In addition, we deploy GIFT for the cold-start video recommendation in Taobao, the largest e-commerce platform in China. Online evaluation further verifies its effectiveness in practical scenarios. Our study is expected to shed light on the item cold-start problem in the recommendation field.

7 RELATED WORK

Substantial efforts have been made toward the item cold-start recommendation, and existing research methods mainly focus on the following two directions:

- **Content-based methods.** content-based methods refer to the methods that exploit content information, such as item attributes, to address the cold-start dilemma. Content-based Filtering [20, 22] is proposed with the assumption that if a user likes a product, it is very likely that she/he will prefer other attribute-similar items. By building this attribute relation, the content-based filtering is able to make cold-start item recommendation without requiring any behavior logs for new items. Along this line, different kinds of side information and algorithm are explored

under different scenarios[5, 31]. Another line of works is the hybrid models[9, 20] that take both cares of behavior logs and the content information. Take e-commerce recommendation[25] for example, innate attributes of item like seller, category, brand, style etc are utilized as the initial feature for the item representation learning, makes their embeddings more robust and effective, despite the missing of user's behaviors for the cold-start items.

- **Transfer learning methods**. Transfer learning is widely used in the cold-start recommendation systems since it targets on applying knowledge extracted from the warmed-up items (source domain) to the cold-start items (target domain). Based on the transfer object, it can be roughly divided as instance- and model-based transfer learning. Instance-based transfer learning[14, 16, 32] aims to extract useful information from the source domain data and convert them to the target domain with weighted tricks. Methods like feature mapping[16], data sharing[32] or building explicit relationships like graph structure[14] between two domains can be used to enhance the information of the cold-start items. It is noted that the graph-based transfer method elaborated in this paper falls into this category; Model-based transfer learning[4, 21, 30] aims to transfer model parameters trained on the abundant samples of source domain to the target domain meanwhile to prevent the negative transfer caused by the distribution gap. Training techniques like pre-training and fine-tuning[30], confusion regularization[21], adversarial training[4] are applied to make the model easily adapt to the target domain.

REFERENCES

[1] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016.

[2] P. Covington, J. Adams, and E. Sargin. Deep neural networks for youtube recommendations. In *RecSys*, 2016.

[3] S. Fan, J. Zhu, X. Han, C. Shi, et al. Metapath-guided heterogeneous graph neural network for intent recommendation. In *SIGKDD*, 2019.

[4] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*. PMLR, 2015.

[5] Z. Gantner, L. Drumond, C. Freudenthaler, S. Rendle, et al. Learning attribute-to-feature mappings for cold-start recommendations. In *ICDM*, 2010.

[6] J. Gope and S. K. Jain. A survey on solving cold start problem in recommender systems. In *ICCCA*, 2017.

[7] M. Grbovic and H. Cheng. Real-time personalization using embeddings for search ranking at airbnb. In *SIGKDD*, 2018.

[8] S. Hu, X. Zhang, J. Zhou, S. Ji, et al. Turbo: Fraud detection in deposit-free leasing service via real-time behavior network mining. In *ICDE*, 2021.

[9] B. M. Kim, Q. Li, C. S. Park, et al. A new approach for combining content-based and collaborative filters. *Journal of Intelligent Information Systems*, 2006.

[10] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*, 2016.

[11] Y. Koren, R. Bell, and C. Volinsky. tion techniques for recommender systems. *Computer*, 2009.

[12] C. Li, Z. Liu, M. Wu, Y. Xu, et al. Multi-interest network with dynamic routing for recommendation at tmall. In *CIKM*, 2019.

[13] Z. Li, X. Shen, Y. Jiao, X. Pan, et al. Hierarchical bipartite graph neural networks: Towards large-scale e-commerce applications. In *ICDE*, 2020.

[14] S. Liu, I. Ounis, C. Macdonald, and Z. Meng. A heterogeneous graph neural model for cold-start recommendation. In *SIGIR*, 2020.

[15] Y. Lu, Y. Fang, and C. Shi. Meta-learning on heterogeneous information networks for cold-start recommendation. In *SIGKDD*, 2020.

[16] S. J. Pan, J. T. Kwok, Q. Yang, et al. Transfer learning via dimensionality reduction. In *AAAI*, 2008.

[17] S. Shi, M. Zhang, Y. Liu, and S. Ma. Attention-based adaptive model to unify warm and cold starts recommendation. In *CIKM*, 2018.

[18] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009.

[19] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv:1908.07490*, 2019.

[20] P. B. Thorat, R. Goudar, and S. Barve. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 2015.

[21] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[22] R. Van Meteren and M. Van Someren. Using content-based filtering for recommendation. In *MLnet/ECML2000 workshop*, 2000.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al. Attention is all you need. *arXiv:1706.03762*, 2017.

[24] M. Volkovs, G. W. Yu, and T. Poutanen. Dropoutnet: Addressing cold start in recommender systems. In *NIPS*, 2017.

[25] J. Wang, P. Huang, H. Zhao, Z. Zhang, et al. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *SIGKDD*, 2018.

[26] S. Xin, Z. Li, P. Zou, C. Long, et al. Atnn: Adversarial two-tower neural network for new item's popularity prediction in e-commerce. In *ICDE*, 2021.

[27] X. Yang, Y. Zhu, Y. Zhang, X. Wang, et al. Large scale product graph construction for recommendation in e-commerce. *arXiv:2010.05525*, 2020.

[28] G. Zhou, N. Mou, Y. Fan, Q. Pi, et al. Deep interest evolution network for click-through rate prediction. In *AAAI*, 2019.

[29] G. Zhou, X. Zhu, C. Song, Y. Fan, et al. Deep interest network for click-through rate prediction. In *SIGKDD*, 2018.

[30] K. Zhou, H. Wang, , et al. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *CIKM*, 2020.

[31] Y. Zhu, J. Lin, S. He, B. Wang, et al. Addressing the item cold-start problem by attribute-driven active learning. *TKDE*, 2019.

[32] F. Zhuang, P. Luo, H. Xiong, Q. He, et al. Exploiting associations between word clusters and document classes for cross-domain text categorization. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2011.

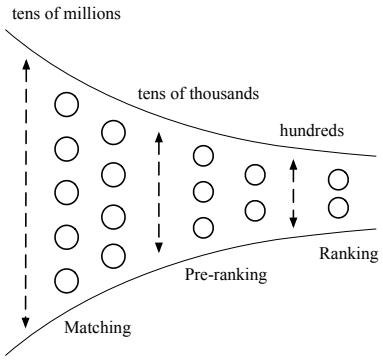


Figure 8: The cascade architecture for industrial recommendation system.

A SYSTEM IMPLEMENTATION

In this section, we introduce the details about the implementation and deployment of the GIFT system to promote this paper's reproducibility.

First, it should be noted that in industrial e-commerce scenarios, an eligible recommendation system should be able to select tens of items out of billions of items for each user in less than 300 milliseconds to meet the users' needs. As it is impossible to predict CTR for each pair between the user and all the items, the recommendation systems in Taobao always follow a multi-stage cascade architecture to trade off the efficiency and accuracy, i.e., to use some simple models to select a candidate item set from a large size of item pool and then use more complex models to filter this candidate set further. The whole process can be divided into three stages: matching, pre-ranking, and ranking.

In the matching stage, a candidate set of items (about tens of thousands) that are similar to items users have interacted with will be selected based on item-to-item scores, pre-calculated by Swing[27]. We also implement MIND[12] to conduct vector retrieval in our pipeline; Pre-ranking can be seen as a simplified version of the ranking phase, considering the computation cost challenge of online serving with a larger size of the candidate set to be calculated (around tens of thousands). We implement the DNN[2] model for the pre-ranking stage in our pipeline; Ranking stage plays a vital role in the whole cascade architecture. Since it tackles only hundreds of items, the model can take into account much more fine-grained features like cross-feature and adopt more intricate model architecture, like the target attention of DIN. In our implementation, the GIFT system works for the ranking stage and can be easily generalized to the matching and pre-ranking stages because it only involves calculation on the target item side. (interactive calculation between user- and item-side like target attention can not be supported by matching and pre-ranking stage because of the restriction on response time).

Figure 9 gives a brief illustration of the GIFT system that has been deployed in the homepage of Taobao App, which consists of four key components: iGraph server, ABFS server, real-time prediction server and model management module. iGraph is an in-memory graph database developed by Alibaba Group, providing storage, real-time query, and update services for large-scale graph data; ABFS is a uniform feature service built by Alibaba Group, which

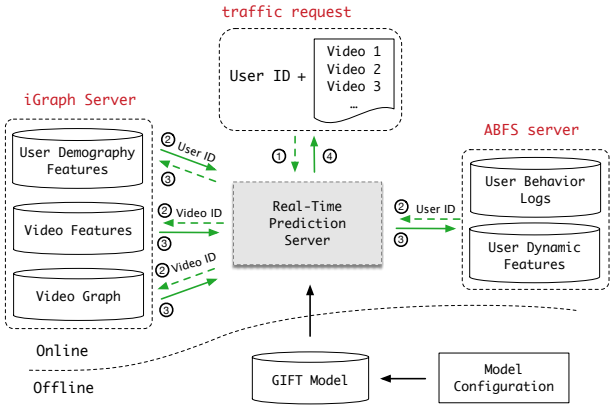


Figure 9: The overall framework of GIFT system deployed in Taobao.

can provide real-time user behavior logs with very low latency (tens of milliseconds) and generate dynamic user features like the length of stay time on a video. Specifically, when a client makes a prediction request with the user id and the candidate videos selected by matching and pre-ranking models, the real-time prediction server queries the user demography features and video features stored in the iGraph server and the neighbored video nodes of the target video. Simultaneously, the real-time prediction server asks the ABFS server to return the clicked video sequence and the dynamic user feature. Taking above as the input, the prediction server makes a real-time prediction and returns to the client the predicted CTR for each candidate item. The execution order is marked by the number.

In our implementation, the user demography feature and video feature are relatively stable and thus can be updated in a low frequency like a daily basis. To keep the new videos fresh, we construct the video graph at a higher frequency of updates, e.g., hourly basis. It should be noted that the neighbored video nodes are sampled offline and pre-stored as a sequence for each cold-start video in iGraph server. To this end, there is no need to consider the overhead of graph sampling in online inference. As the CTR model in the ranking stage, GIFT is trained incrementally in a daily basis by the model management module. It is also very helpful to adopt online learning to improve the performance of new video recommendation in practice.

B EXPERIMENT DETAILS

We implement the DNN part of Wide&Deep, DeepFM, DNN, DIN and GIFT just the same architecture, i.e., a three-layer MLP with 512, 256 and 128 hidden units for Taobao dataset, and 128, 64, 32 for MovieLens and DBook datasets. For all attention layers in above models, we set the number of hidden units to 128 for Taobao dataset and 32 for other two datasets. Adagrad optimizer is adopted in all the methods, the learning rate of 1e-4 is set. For other baselines, the grid search strategy is applied to find the optimal hyper-parameters. For a certain type of physical linkage, we sample top-20 neighbors according to the ascending order of time interval between them and the target item; For semantic linkage, we sample top-20 neighbors according to the descending order of cosine similarity. The default value of 20 is chosen based on the observation that there are only marginal improvements when we continually increase this value.