Zero Shot on the Cold-Start Problem: Model-Agnostic Interest Learning for Recommender Systems

Philip J. Feng
NetEase Cloud Music, NetEase Inc.
Hangzhou, China
philipj.feng@outlook.com

Pingjun Pan∗
NetEase Cloud Music, NetEase Inc.
Hangzhou, China
panpingjun@corp.netease.com

Tingting Zhou
NetEase Cloud Music, NetEase Inc.
Hangzhou, China
hzzhoutingting15@corp.netease.com

Hongxiang Chen†
NetEase Cloud Music, NetEase Inc.
Hangzhou, China
hzchenhongxiang@corp.netease.com

Chuanjiang Luo
NetEase Cloud Music, NetEase Inc.
Hangzhou, China
luochuanjiang03@corp.netease.com

ABSTRACT
User behavior has been validated to be effective in revealing personalized preferences for commercial recommendations. However, few user-item interactions can be collected for new users, which results in a nullspace for their interests, i.e., the cold-start dilemma. In this paper, a two-tower framework, namely, the model-agnostic interest learning (MAIL) framework, is proposed to address the cold-start recommendation (CSR) problem for recommender systems. In MAIL, one unique tower is constructed to tackle the CSR from a zero-shot view, and the other tower focuses on the general ranking task. Specifically, the zero-shot tower first performs cross-modal reconstruction with dual autoencoders to obtain virtual behavior data from highly aligned hidden features for new users; and the ranking tower can then output recommendations for users based on the completed data by the zero-shot tower. Practically, the ranking tower in MAIL is model-agnostic and can be implemented with any embedding-based deep models. Based on the co-training of the two towers, the MAIL presents an end-to-end method for recommender systems that shows an incremental performance improvement. The proposed method has been successfully deployed on the live recommendation system of NetEase Cloud Music to achieve a click-through rate improvement of 13% ~ 15% for millions of users. Offline experiments on real-world datasets also show its superior performance in CSR. Our code is available†.

CCS CONCEPTS
• Information systems → Recommender systems; • Computing methodologies → Artificial intelligence.

†Pingjun Pan contributes to model implementation and online testing.
‡Hongxiang Chen is the corresponding author.
§https://github.com/LiangjunFeng/MAIL

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.

ACM Reference Format:
Philip J. Feng, Pingjun Pan, Tingting Zhou, Hongxiang Chen, and Chuanjiang Luo. 2021. Zero Shot on the Cold-Start Problem: Model-Agnostic Interest Learning for Recommender Systems. In Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM ’21), November 1–5, 2021, Virtual Event, QLD, Australia. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482312

1 INTRODUCTION
Recommender systems are playing important roles in various web and mobile applications, e.g., music platforms, which pick a fraction of items that a user might enjoy based on the user’s past behavior and current context [7, 25]. Despite the success of recent deep learning [30] and matrix factorization-based methods [17], a common challenge for recommender systems is the cold-start recommendation (CSR); that is, the recommendation model shows degraded performance in selecting satisfactory items for new users because of the lack of user-item interaction records [9, 15].

Different approaches have been proposed to handle the cold-start problem. A basic CSR method is to leverage personal information to generate recommendations for new users. This is reasonable because some basic attributes of users, e.g., age and occupation, can be collected when they register accounts, and people with similar attributes usually have similar interests and behaviors [2]. Some other auxiliary information is also used in CSR. For example, cross-domain knowledge is used by Hsu et al. [27] and Philip et al. [5] to enrich the feature representation of new users. Complementary heterogeneous information is leveraged to enrich user-item interactions by Lu et al. [13] using heterogeneous information networks. A stacked denosing autoencoder is used in a hybrid model to extract embedding knowledge for CSR by Wei et al. [22]. In [10, 20, 26], the popular meta-learning framework is applied to transfer knowledge from support sets and treat CSR as a few-shot learning task. However, the major limitation of these approaches is that new users’ behaviors and interests are still missing. The auxiliary information, such as semantic knowledge and pretrained embedding, cannot change the dilemma that the recommender system knows less about new users than old users.
Recently, zero-shot learning (ZSL) [6, 23], which has the same intention as CSR, has attracted ever-increasing attention in the computer vision field. Both ZSL and CSR involve the missing data problem of some objects in a specific scene. To provide a better understanding, we compare paradigms of ZSL and CSR in Figure 1. In the image recognition task, the training features of the unseen class are missing, and ZSL uses textual descriptions of objects to perform knowledge transfer from the seen class to the unseen class. In the cold-start recommendation task, behavior data of new users are missing, and CSR can utilize user attributes to generalize interests from old users to new users. Based on the zero-shot view, Li et al. [4] proposed the low-rank linear autoencoder (LLAE) for CSR. The LLAE addresses CSR with a linear autoencoder, where a new user’s behavior is learned from attributes with a low-rank constraint. Kumar et al. [24] also applied some classic zero-shot models, e.g., ESZSL [21] and ConSE [14], for unseen hashtag recommendation. However, the behavior data in CSR are different from the image features in ZSL. A fact that should be emphasized is that a user can interact with any items in the recommender system. The huge number of items in real-world applications make the models, state-of-the-art ranking performance. Although popular ranking models, e.g., ESMM [16] and DMR [18], suffer from the cold-start problem on new users, they usually perform well on old users. Incremental performance improvement can be expected by integrating the cold-start module with the ranking model together.

Based on the above discussion, we propose a model-agnostic interest learning (MAIL) framework for recommender systems to address the cold-start problem in this paper. In MAIL, a new zero-shot tower and a general ranking tower are constructed together. From a cotraining view, the zero-shot tower formulates CSR as a ZSL task to provide virtual behavior data for new users, and the ranking tower can then capture new user’s interests based on the generated virtual data. Specifically, in the zero-shot tower, two spaces, i.e., user behavior space and user attribute space, are explicitly constructed and learned by dual autoencoders. In the common hidden space of autoencoders, cross-modal reconstruction is performed to overcome the modal-shift problem between user behavior and user attributes for old users. Based on the homogeneity of user’s attributes, new user’s behavior can be readily reconstructed from well-aligned hidden features. For the ranking tower, it is model-agnostic and can be implemented with any embedding-based ranking models. The ranking tower is assigned to share its embedding layer with the zero-shot tower to tackle the data sparsity problem for efficient model training. We show that integrating zero-shot learning together with popular ranking models makes recommender systems free from CSR so that performance improvement on state-of-the-art results is achieved. The contributions of this paper are fourfold:

1) A new framework, namely, the model-agnostic interest learning (MAIL) framework, is designed for recommender systems, which constructs a unique zero-shot tower to provide virtual behavior data for new users, enabling the ranking tower to capture their interests for cold-start recommendations.

2) Based on cross-modal reconstruction with dual autoencoders, the zero-shot tower in MAIL makes the hidden features of user attributes and user behavior highly aligned to generalize behavior data from old users to new users.

3) The ranking tower in MAIL can be implemented with any embedding-based deep models for practical application, and it also shares its embedding layer with the zero-shot tower to tackle the data sparsity problem for effective model training.

4) In addition to offline experiments, we deploy the proposed MAIL model on an online large-scale live recommendation system to verify the model effectiveness with millions of users in a real-world setting.

2 RELATED WORK

2.1 Zero-Shot Learning

In the computer vision field, intelligent models optimize millions of parameters using sufficient training samples to obtain recognition ability for target objects [1]. However, due to the long-tailed distribution of object categories, the collection of numerous samples for some rarely seen objects is generally challenging. To recognize unseen objects, zero-shot learning is explored by Lampert et al. [23] and Akata et al. [12] in recent years. From a macro perspective, existing ZSL methods can be grouped into three categories, including classic probability-based methods, compatibility-based methods, and generative models. Probability-based methods are represented by direct attribute-based prediction [8], as these methods learn a
number of probabilistic classifiers for attributes and combine the scores to make predictions. Incremental classifiers are also used by Feng et al. [33] to balance the learning between seen and unseen classes. Compatibility-based methods learn a compatibility function between the semantic embedding space and the vision feature space. For example, the ESZSL model proposed by Romera-Paredes et al. [21] applies square loss and implicit regularization between embedding spaces to rank the class possibilities in an embarrassingly simple way. A hashing model is trained by Ji et al. [29] to transfer knowledge and implement cross-model learning based on the semantic embedding space. Generative models train a conditional generator for visual features based on semantic features. Popular generative models include autoencoders and generative adversarial networks [3]. In comparison with the other two paradigms, generative models tackle the essential insufficient data problem for ZSL to achieve state-of-the-art results. The proposed zero-shot tower in MAIL is also constructed as a generative model to tackle the missing behavior problem for CSR.

2.2 Cold-Start Recommendation

While deep ranking models have achieved considerable success in recommender systems, difficulty usually arises in contending with new users with a nullspace for their interests, known as cold-start recommendation. Some solutions in solving this problem have been proposed in current publications, including the input of semantic knowledge, second-domain knowledge transfer, active learning and meta-learning. For example, Chou et al. [35] proposed a tensor factorization-based algorithm that exploits content features extracted from music audio to address CSR for the emerging application next-song recommendation. A unique characteristic of the algorithm is that it learns and updates the mapping between the audio feature space and the item latent space each time during the iterations of the factorization process, which makes the content features more effective. Li et al. [2] proposed an innovative cross-domain recommendation model based on partial least squares regression. The cross-domain model is able to purely use source-domain ratings to predict the ratings for cold-start users who never rated items in the target domains. Apart from utilizing auxiliary information from other scenes and datasets, some researchers have applied active learning approaches to deal with CSR. Zhu et al. [32] designed useful user selection criteria based on items’ attributes and users’ rating history, and combined the criteria in an optimization framework for selecting users. By exploiting the feedback ratings, accurate rating predictions for the other unselected users can then be generated. The popular meta-learning [28] approach is also applied to address CSR, which regards modeling each user’s preference as a learning task and minimizes the sum of the recommendation loss on training users to globally update the parameters of the meta-learner. By providing the initialization of parameters, the meta-learner guides the learning process for new users. The major difference between the proposed MAIL and the above works is that the MAIL addresses the essential missing behavior problem for new users by transferring knowledge from old users with a zero-shot tower. In addition to the basic user attributes, neither auxiliary information nor an additional dataset is added in the modeling for CSR, which is convenient when feature engineering and data collection are expensive.

3 PROPOSED APPROACH

3.1 Problem Formulation and Notations

3.1.1 Features in Recommender Systems. There are usually four categories of features in recommender systems: user attributes \( a \in \mathcal{A} \), user behavior \( v \in \mathcal{V} \), context \( c \in \mathcal{C} \), and target item \( t \in \mathcal{T} \). User attributes in \( \mathcal{A} \) contain the user ID, occupation, and so on; user behavior in \( \mathcal{V} \) is a sequential list of user interacted items with corresponding IDs. Context in \( \mathcal{C} \) contains the date, recall method, location, etc. Features of target item in \( \mathcal{T} \) are the item ID, category ID, and so on. The ranking model in a recommender system learns a mapping function \( f(x) \) to output the score \( y \) for an item, where \( x = \{a, v, c, t\} \) is the input. Generally, items with higher scores will be recommended to users with priority.

Note that most of the mentioned features are categorical, which are represented by one-hot vectors with high dimension. In deep learning models, one-hot features are usually transformed into low-dimensional dense features by an embedding layer. Hence, the notations, e.g., \( a \) and \( v \), actually denote the embedding vectors. For example, user attributes are \( a = [a_1, \ldots, a_{n_a}] \in \mathbb{R}^{n_a \times d} \), where \( n_a \) is the number of attributes and \( d \) is the embedding dimension. For the user behavior sequence \( v = [v_1, \ldots, v_{n_v}] \in \mathbb{R}^{n_v \times d} \), each item \( v_i \) is obtained by performing element-wise multiplication between the embedding of the item ID and the embedding of the user’s action. Moreover, some dense features can be directly used without the embedding transformation.

3.1.2 Zero-Shot Learning for CSR. In cold-start recommendation, we have two scopes: the old user scope \( \mathcal{S} \) and the new user scope \( \mathcal{O} \). For old users, four categories of features are ready for ranking models, which are denoted as \( x_o = \{a_o, v_o, c_o, t_o\} \). For new users, the behavior \( v_o \) is missing, and features of new users are denoted as \( x_n = \{a_n, c_n, t_n\} \). In this paper, we train a zero-shot tower \( \hat{v}_n = g(a_n) \) based on old users, which can be applied to new users to obtain virtual behavior data \( \hat{v}_n = g(a_n) \) to tackle the cold-start problem. The notations \( \hat{v}_n = g(a_n) \) and \( \hat{v}_o = g(a_o) \) denote that only old users’ behavior data \( v_o \) are available for the zero-shot tower while new users’ behavior data are not available. To summarize, the whole model in this paper can be formulated as \( y_n = f(x_n) = f(a_n, c_n, t_n, \hat{v}_n) \) for old users and \( \hat{y}_o = f(x_o) = f(a_o, c_o, t_o, \hat{v}_o) \) for new users, where \( \hat{v}_o = g(a_o) \).

3.2 Model-Agnostic Zero-Shot Interest Learning

3.2.1 Overall Idea. The overall idea of the proposed MAIL is illustrated in Figure 2. In addition to a general ranking tower, a new zero-shot tower can be observed. The ranking tower outputs recommendations based on the completed data by the zero-shot tower, and the zero-shot tower enjoys the dense embedding by the ranking tower. The co-training of the two towers make the recommender systems free from CSR so that an incremental performance improvement is achieved.

3.2.2 Zero-Shot Tower for Cold-Start Problem. Here, we show how to construct the zero-shot tower for CSR in detail.

First, it is clarified that the zero-shot tower is trained on old users and makes inference on new users to transfer behavior data from old users to new users based on their attributes. Hence, two kinds
of embedding features, i.e., user attributes $a_i$ and user behavior $v_i$, are used for model training. The embedding features are trained by the ranking tower. Through learning from the dense embedding vectors, the zero-shot tower obtains more effective training and extraction processes.

To weight different features, the self-attention technique is then applied on $a_i$ and $v_i$ to construct attribute features $h_a^i$ and behavior features $h_v^i$, respectively. Taking $h_a^i$ as an example, the attention mechanism is performed as follows:

$$h_a^i = \sum_{i=1}^{n_a} \alpha_i a_{si},$$

(1)

where $n_a$ is the number of user attributes and $a_{si} \in \mathbb{R}^d$ is the embedding vector of the $i$-th user attribute in $a_i$. The attention weight $\alpha_i$ is obtained by

$$\alpha_i = \frac{exp(e_i)}{\sum_{j=1}^{n_a} e_j},$$

$$e_i = z^T \tanh(wa_{si} + b),$$

(2)

where $z \in \mathbb{R}^{d_z}, w \in \mathbb{R}^{d_z \times d_a}$, and $b \in \mathbb{R}^{d_z}$ are trainable parameters, and $d_z$ is the given attention dimension.

Based on $h_a^i$ and $h_v^i$, cross-modal reconstruction is performed to align user attributes and behavior for a more effective attribute-based behavior generation. In the attribute space, the cross-modal reconstruction loss is:

$$\mathcal{L}_a = ||D1(p_a) - a_i||_2^2 + ||D2(p_a) - v_i||_2^2,$$

(3)

and in the behavior space, the cross-modal reconstruction loss is:

$$\mathcal{L}_v = ||D1(q_a) - a_i||_2^2 + ||D2(q_a) - v_i||_2^2,$$

(4)

where $p_a = E1(h_a^i) \in \mathbb{R}^{d_h}, q_a = E2(h_v^i) \in \mathbb{R}^{d_h}$, and $d_h$ is the given hidden dimension. $E1$ and $E2$ are encoders for user attributes and user behavior, and $D1$ and $D2$ are corresponding decoders, which are shown in Figure 2.

The cross-modal reconstruction loss assigns the hidden features $p_a$ and $q_a$ to reconstruct $a_i$ and $v_i$, respectively, with the common decoders, which actually makes $p_a$ and $q_a$ have the same prediction ability for user attributes and user behavior. Take the exposition a step further, we denote $p_a \in \mathbb{R}^{n \times d_h}$ as a batch of $p_a$, $q_a \in \mathbb{R}^{n \times d_h}$ as a batch of $q_a$, and $n$ as the batch size. The maximum mean discrepancy (MMD) constraint is then applied to $p_a$ and $q_a$ to minimize their distribution distance and guide the convergence of cross-modal
reconstruction, which can be formulated as:

$$L_d = D_{\mathcal{H}}(P_s, Q_s) = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(p_{si}) - \frac{1}{n} \sum_{j=1}^{n} \phi(q_{sj}) \right\|^2_2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} k(p_{si}, p_{sj}) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} k(q_{si}, q_{sj}) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(p_{si}, q_{sj}),$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS), $\phi : P_s, Q_s \rightarrow \mathcal{H}$, $k$ is a Gaussian kernel function:

$$k(p_s, q_s) = \exp(-\|p_s - q_s\|^2 / 2\sigma^2),$$

and $\sigma = 1$ is the bandwidth in this paper.

To summarize, the training loss of the zero-shot tower is:

$$L_{zst} = L_a + L_o + L_d.$$  

(7)

The first two items in $L_{zst}$ require both $p_s$ and $q_s$ to have the reconstruction ability for $a_s$ and $\tilde{v}_s$, while the third item in $L_{zst}$ makes $p_s$ and $q_s$ enjoy the same hidden space for a better convergence of cross-modal reconstruction. With the triple-aligned hidden features, the virtual behavior data can be readily generated for new users from their attributes as follows:

$$\tilde{v}_o = D2(p_o) = D2(E1(h_o^s)),$$

(8)

where $h_o^s$ is the attention weighted features of $a_o$ and can be obtained by Eqs. (1) and (2).

3.2.3 Ranking Tower for Recommendation. Here, we show how to construct the ranking tower for recommendation in detail.

Generally, the ranking tower is model-agnostic and can be implemented by any embedding-based models for estimating the click-through rate (CTR), post-click conversion rate (CVR), or post-view click-through&conversion rate (CTCVR) [16]. Here, the basic deep model shown in Figure 2 is introduced to provide a better understanding.

The ranking tower treats new users and old users in different ways, which means there should be a flag to indicate whether a user is a new user, which can be obtained by

$$\text{flag} = \begin{cases} 
\text{true} & \sigma = \emptyset \\
\text{false} & \text{otherwise}, 
\end{cases}$$

where $\text{flag} = \text{true}$ for new users and $\text{flag} = \text{false}$ for old users.

Based on the embedding lookup table and flag, the input $x$ of the ranking tower can be denoted as:

$$x = \begin{cases} 
\{a_o, c_o, t_o, \tilde{v}_o\} & \text{flag} = \text{true} \\
\{a_s, c_s, t_s, \tilde{v}_s\} & \text{flag} = \text{false}, 
\end{cases}$$

(10)

where $\tilde{v}_o$ obtained from Eq. (8) is the virtual behavior data for new users.

A basic sum pooling layer is then added to obtain the concatenated hidden features $z = [z^a, z^o, z^f, z^t] \in \mathbb{R}^{4d}$ from $x$, where $d$ is the embedding dimension. Taking $z^a$ as an example, the sum pooling is performed as follows:

$$z^a = \sum_{i=1}^{n_a} a_i,$$

(11)

where $a_i$ is the $i$-th attribute embedding vector in $a$.

The final output $\hat{y}$ is obtained by a multilayer perceptron, which is formulated as:

$$h_{z1} = \text{LeakyReLU}(\text{BN}(\text{Dense}(\text{Dropout}(z, 0.5))))$$

$$h_{z2} = \text{LeakyReLU}(\text{BN}(\text{Dense}(\text{Dropout}(h_{z1}, 0.5))))$$

(12)

and the training loss of the ranking tower is:

$$L_{rt} = - \frac{1}{n} \sum_{i=1}^{n} [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)],$$

(13)

where $n$ is the batch size.

In summary, the overall training loss of MAIL is:

$$L = L_{zst} + L_{rt}.$$  

(14)

where two towers are cotrained to tackle the CSR for recommender systems. It is worth mentioning that, in practical applications, not only will there be new users in the test data but there will also be new users in the training data. It is difficult for a ranking and CSR two-stage model to address the missing data problem in the ranking model’s training phase because the CSR model based on embedding features has to be trained after the ranking model. In contrast, the proposed MAIL is designed as an end-to-end model to obtain virtual behavior data from the zero-shot tower for the ranking tower whenever the need arises and is hence more practical for applications.

![Figure 3: Illustration of the autoencoder used in MAIL, where the encoder and decoder are designed with residual structures for efficient feature extraction.](image)

3.2.4 Implementation Tricks and the Algorithm of MAIL. Here, implementation tricks and the training algorithm of MAIL are summarized.

1) Cotraining with two optimizers: The two towers in MAIL are trained by two optimizers, respectively. One optimizer $G_A$ is used for $L_{zst}$, and the other $G_B$ is for $L_{rt}$. It is noted that the embedding
look-up table is completely trained by $G_B$ together with the ranking tower and is not affected by $G_A$. Because the aim of the zero-shot tower is to reconstruct embedding vectors instead of changing them, applying the gradients of the zero-shot tower on the embedding look-up table would degrade model performance.

2) Autoencoders with residual structure: Two autoencoders are used in the zero-shot tower to perform the cross-modal reconstruction. Considering the sparse information in recommender systems, a residual structure is designed for the autoencoder to perform the reconstruction task better. Taking the left autoencoder in Figure 2 as an example, the residual structure is shown in Figure 3. The forward-propagation of the residual encoder can be formulated as follows:

$$z_s^1 = \text{LeakyReLU}(\text{Dense}(\text{Dropout}(h_s^0, 0.5)))$$
$$z_s^2 = \text{LeakyReLU}(\text{Dense}(\text{Dropout}(\text{Concat}(z_s^1, h_s^0), 0.5)))$$
$$p_s = \text{LeakyReLU}(\text{Dense}(\text{Concat}(z_s^2, h_s^0)))$$

and the forward-propagation of the residual decoder can be formulated as:

$$z_s^3 = \text{LeakyReLU}(\text{Dense}(\text{Dropout}(p_s, 0.5)))$$
$$z_s^4 = \text{LeakyReLU}(\text{Dense}(\text{Dropout}(\text{Concat}(z_s^3, p_s), 0.5)))$$
$$a_s^c = \text{Dense}(\text{Concat}(z_s^4, p_s))$$

The principle of the residual structure is similar to that of the wide&deep model [34], where both the wide module’s memory ability and the deep module’s generalization ability are retained by feature concatenation.

3) Training algorithm of MAIL: The algorithm of MAIL is summarized in Algorithm 1.

3.3 Feasibility Analysis of Attribute-based Knowledge Transfer

In MAIL, the zero-shot tower generalizes behavior data from old users to new users based on their attributes. A basic assumption behind the knowledge transfer is that people with similar attributes have similar behaviors. However, there seems to be a contradiction in that the user attribute space is much smaller than the user behavior space in recommender systems. For better clarification, we denote the number of attributes as $n_a$ and assume that each attribute has $k_a$ different values. Similarly, the length of a user’s behavior sequence is denoted as $n_b$, and the number of items that can be interacted with users is denoted as $k_b$. The condition $n_a k_a >= n_b k_b$ seems to be required for effective learning from attributes to behavior, while both $n_a$ and $k_a$ may be larger than $n_b$ and $k_b$, respectively, in a real recommender system.

Here, we highlight that the core of user behavior is user interest instead of items. Compared with learning what items users have interacted with, it is better to learn what style of items users like for recommender systems. This is one of the important reasons why the dense embedding vector $v$ is implemented as the learning target of the zero-shot tower. Based on the dense vector, a user’s interest $\phi(r)$ can be readily represented by a set of items as

$$\phi(r) = \{ r \mid (r, v)[2] < \epsilon \},$$

where $\epsilon$ is an error bound representing the interest range. As long as the generated $\hat{a} \in \phi(r)$, it is feasible for the ranking tower to capture user interest. Hence, the object of the user attribute space is actually the user interest space instead of the user behavior space. Meanwhile, the kinds of user interests $n_a$ are much smaller than $n_b k_b$, which makes the attribute-based knowledge transfer available for CSR.

| Table 1: Statistics of the Two Datasets |
|-------------------------------|-----------------|-----------------|
| Dataset                  | Public | Industrial |
| # All users             | 1,141,729 | 613,221 |
| # New users             | 456,691 | 282,081 |
| # Items                 | 846,811 | 52,831 |
| # User attributes       | 9      | 21     |
| # Context features      | 4      | 75     |
| # Avg. length of behavior sequence | 63 | 79 |
| # Item features         | 6      | 105    |

4 EXPERIMENTS

Here, the MAIL is validated on both public and industrial datasets. First, we introduce experimental settings and compared methods in detail. Next, the recommendation performance and an ablation study are given to show the state-of-the-art results. Visualized features are also presented to provide an intuitive understanding.
### 4.1 Experimental Settings and Compared methods

#### 4.1.1 Datasets. Two large-scale real-world dataset are used here.

1) **Public Dataset** The Alimama Dataset\(^1\) contains ad displays and click logs randomly sampled from Taobao over 8 days. It includes 1.14 million users and 0.84 million items with 6 million user-item interaction logs. The logs from the first seven days are used as the training set, and logs from the last day are used as the testing set. Note that the dataset is provided without new users. Hence, we randomly select forty percent of users in the training and testing set and treat them as new users by removing their behaviors.

2) **Industrial Dataset** The industrial dataset from NetEase Cloud Music contains live recommendation and user behavior logs randomly sampled from a company (abiding by the double-blind principle) over 8 days. It includes 613 thousand users and 52 thousand items with 2 million user-item interaction logs. In the industrial dataset, approximately half of the users are new users. The statistics of the two datasets are summarized in Table I.

#### 4.1.2 Compared Methods. Seven models are compared.

1) **Emblr** Logistic regression is a classic linear model that can be regarded as a shallow neural network. Here, we implement logistic regression with an embedding layer, i.e., Emblr, to content with sparse IDs. Emblr is the baseline of LLAE.

2) **LLAE (2019)** LLAE [4] is a zero-shot learning-based cold-start method, that generates behavior data for logistic regression with a linear low-rank autoencoder. As discussed in [4], the weight of low-rank loss is decided by cross-validation and set to 30.

3) **BaseDNN** BaseDNN is the ranking tower shown in Figure 2, which is used here as a baseline of MetaEmb and MAIL-Base.

4) **MetaEmb (2019)** MetaEmb [28] is a meta-learning-based cold-start method, that learns to generate desirable initial embeddings for new IDs. The base model is implemented as the BaseDNN.

5) **MAIL-Base (Ours)** MAIL-Base is the proposed two-tower model shown in Figure 2, which uses the zero-shot tower to tackle the cold-start problem for BaseDNN.

6) **DMR (2020)** DMR [18] is a state-of-the-art model for recommender systems, that consists of an item-to-item network and a user-to-item network to combine the thought of collaborative filtering for the ranking task. DMR is the baseline of MAIL-DMR.

#### 4.1.3 Metrics. The area under the ROC curve (AUC) and the group area under the ROC curve (GAUC) are used as metrics. The model performance on both new users and old users are reported. In addition, for each result, the relative improvement (RI) between the baseline and the improved method is provided for an intuitive comparison. RI is obtained as follows:

\[
RI(a, b) = \frac{b - a}{a} \times 100\%.
\]

where \(a\) is the result of the baseline and \(b\) is the result of the improved method.

### 4.2 Results on Public and Industrial Datasets

In the experiments, we set the learning rate to 0.001, embedding size to 32, batch size to 1024, and dropout rate to 0.5. The maximum length of the behavior sequence is set to 100 for all users. If the length of a user’s behavior sequence is less than 100, we concatenate the behavior sequence with zero padding. In MAIL, dual autoencoders are used. The dimension of hidden layers in encoders is 1024-512-512, and the dimension of hidden layers in the decoders is 512-1024-\(d_t\), where \(d_t\) denotes the dimension of target vectors. The dimension of the hidden layers in the multilayer perceptron of the ranking tower is 512-256. The Adam optimizer [31] is used with \(\beta_1 = 0.9\) and \(\beta_2 = 0.999\) in this paper.

The results are summarized in Table II. LLAE performs better than the basic Emblr method. An AUC improvement of 0.6% ~ 0.9% is obtained on new users, which is significant for large-scale real-world datasets. However, LLAE is limited to linear transformations and based on linear logistic regression. In comparison with the other five deep models, the weak representation ability of LLAE can be observed, which demonstrates the effectiveness of nonlinear transformation and a deep structure.

For MetaEmb and MAIL-Base, the BaseDNN model is used as a baseline and an AUC improvement of 0.8% ~ 1.3% is obtained on new users. The proposed MAIL-Base generates behavior data

### Table 2: Results on Public and Industrial Datasets

| Method     | Public Dataset |          |          | Industrial Dataset |          |          |
|------------|----------------|----------|----------|--------------------|----------|----------|
|            | AUC   | RI | GAUC | RI | AUC   | GAUC | RI | AUC   | GAUC | RI |
| EmbLR      | 0.5800 | 0.00% | 0.5607 | 0.00% | 0.6252 | 0.00% | 0.5892 | 0.00% | 0.6682 | 0.00% | 0.6365 | 0.00% | 0.7728 | 0.00% | 0.6419 | 0.00% |
| LLAE       | 0.5854 | 0.93% | 0.5646 | 0.69% | 0.6253 | 0.01% | 0.5892 | 0.00% | 0.6713 | 0.46% | 0.6384 | 0.29% | 0.7737 | 0.11% | 0.6422 | 0.05% |
| BaseDNN    | 0.5862 | 0.00% | 0.5658 | 0.00% | 0.6271 | 0.00% | 0.5897 | 0.00% | 0.6771 | 0.00% | 0.6380 | 0.00% | 0.7782 | 0.00% | 0.6459 | 0.00% |
| MetaEmb    | 0.5911 | 0.84% | 0.5682 | 0.42% | 0.6273 | 0.03% | 0.5900 | 0.05% | 0.6834 | 0.93% | 0.6404 | 0.37% | 0.7788 | 0.07% | 0.6466 | 0.10% |
| MAIL-Base  | 0.5934 | 1.22% | 0.5710 | 0.91% | 0.6271 | 0.00% | 0.5898 | 0.02% | 0.6849 | 1.15% | 0.6413 | 0.51% | 0.7788 | 0.07% | 0.6460 | 0.01% |
| DMR        | 0.5879 | 0.00% | 0.5668 | 0.00% | 0.6302 | 0.00% | 0.5911 | 0.00% | 0.6823 | 0.00% | 0.6420 | 0.00% | 0.7833 | 0.00% | 0.6551 | 0.00% |
| MAIL-DMR   | 0.5958 | 1.34% | 0.5733 | 1.14% | 0.6302 | 0.00% | 0.5915 | 0.06% | 0.6895 | 1.05% | 0.6465 | 0.70% | 0.7838 | 0.06% | 0.6558 | 0.10% |

Red font and blue font are the highest and the second highest results in each column. EmbLR is the baseline of MetaEmb and MAIL-Base; DMR is the baseline of MAIL-DMR.

---

\(^1\)https://tianchi.aliyun.com/dataset/dataDetail?dataId=56
based on user attributes to address CSR, while the MetaEmb gives a better initial embedding for the ranking model. In comparison, the generative strategy of MAIL is more effective than the initialization strategy of MetaEmb. Moreover, by comparing MAIL-Base and MAIL-DMR, it can be observed that MAIL can use better models in the ranking tower to achieve an incremental performance improvement. For the public dataset, MAIL improves the AUC on new users from 0.5934 to 0.5958 by replacing BaseDNN with DMR. For the industrial dataset, the AUC on new users is improved from 0.6849 to 0.6895. The model-agnostic property is very attractive for practical applications since there are various ranking models for different scenarios in the real world.

Furthermore, although cold-start methods, i.e., LLAE, MetaEmb, and MAIL, focus on the new users, we find that they show a slight AUC improvement on old users. For the public dataset, an AUC improvement of 0% ~ 0.06% on old users can be observed. For the industrial dataset, an AUC improvement of 0% ~ 0.1% on old users is obtained. This may be explained by noting that cold-start methods generate virtual behavior data or make a better initial embedding, which alleviates the sparse data problem for ranking models that enables performance improvement. The results of the GAUC metric are similar to those of AUC.

Table 3: Results of Ablation Study on Public and Industrial Datasets

| Metrics          | New Users | Old Users |
|------------------|-----------|-----------|
|                  | AUC       | RI        | AUC       | RI         |
| **Public Dataset** |           |           |           |            |
| BaseDNN          | 0.5862    | -1.21%    | 0.6271    | 0.00%      |
| MAIL-None        | 0.5834    | -1.68%    | 0.6263    | -0.12%     |
| MAIL-Single      | 0.5844    | -1.51%    | 0.6267    | -0.06%     |
| MAIL-Dual        | 0.5897    | -0.62%    | 0.6272    | 0.01%      |
| MAIL-Base        | 0.5934    | 0.00%     | 0.6271    | 0.00%      |
| **Industrial Dataset** |         |           |           |            |
| BaseDNN          | 0.6771    | -1.13%    | 0.7788    | 0.00%      |
| MAIL-None        | 0.6755    | -1.37%    | 0.7782    | -0.07%     |
| MAIL-Single      | 0.6764    | -1.24%    | 0.7784    | -0.05%     |
| MAIL-Dual        | 0.6802    | -0.68%    | 0.7788    | 0.00%      |
| MAIL-Base        | 0.6849    | 0.00%     | 0.7788    | 0.00%      |

4.3 Ablation Study

There are three items in the training loss $L_{zd}$ of the zero-shot tower, including two cross-modal reconstruction items and an MMD item. Here, an ablation study is performed to explore the effectiveness of the three items. Five models are designed for comparison as follows:

1) **BaseDNN** BaseDNN is the ranking tower shown in Figure 2.
2) **MAIL-None** MAIL-None has no zero-shot training loss in the zero-shot tower, but only has a linear mapping from user attributes to user behaviors. The ranking tower of MAIL-None is BaseDNN.
3) **MAIL-Single** The zero-shot tower of MAIL-Single is trained by Eq. (3). The ranking tower of MAIL-Single is BaseDNN.
4) **MAIL-Dual** The zero-shot tower of MAIL-Dual is trained by Eqs. (3) and (4). The ranking tower of MAIL-Dual is BaseDNN.
5) **MAIL-Base** The zero-shot tower of MAIL-Base is trained by Eq. (7). The ranking tower of MAIL-Base is BaseDNN.

Table 2 gives the results of MAIL with different zero-shot towers on both public and industrial datasets. As the table indicates, MAIL-None and MAIL-Single show worse results than BaseDNN. For the public dataset, the AUC on new users of BaseDNN is 0.5862, while that of MAIL-None and MAIL-Single is 0.5834 and 0.5844, respectively. For the industrial dataset, the AUC on new users of BaseDNN is 0.6771, while that of MAIL-None and MAIL-Single is 0.6755 and 0.6764, respectively. The two models generate virtual data depending on user attributes and do not regularize the autoencoders by setting user behavior as input, which results in overfitting and performance degradation. In comparison, the results of MAIL-Dual and MAIL-Base are much better, the stronger regularization by dual cross-modal reconstruction and maximum mean discrepancy reduces the modal-shift problem between user behavior and user attributes to achieve satisfactory virtual data for the ranking tower.

Also, it is noticed that the unreliable data generated by MAIL-None and MAIL-Single slightly affect the model performance on old users. For the public dataset, an AUC decrease of 0.06% ~ 0.12% is obtained. For the industrial dataset, an AUC decrease of 0.05% ~ 0.07% is obtained. In comparison, MAIL-Dual and MAIL-Base are not observed to affect the ranking performance on old users.

Figure 4: The training loss of MAIL on the public and industrial datasets.

In addition, we show the training loss of MAIL-Base in Figure 4, including the ranking tower loss $L_{zr}$, the behavior-based cross-modal reconstruction loss $L_{zr}$, the attribute-based cross-modal reconstruction loss $L_{zd}$, and the MMD loss $L_{zd}$. Generally, the training loss decreases rapidly in one hundred batches for both of the public and industrial datasets, which verifies the learnability from user attributes to user behaviors.
4.4 Visualization Study

In this subsection, we perform a visualization study for various hidden features in the zero-shot tower using the t-SNE [36].

4.4.1 Visualization for the Public Dataset. For the public dataset, we provide a visualization comparison between the generated features $\hat{v}$ and the real features $v$ to validate the effectiveness of the zero-shot tower. Note that the new users in the public dataset are made by removing their behavior data from the dataset; hence, we can perform the comparison on both the old users ($\hat{v}_s$ and $v_s$) and new users ($\hat{v}_o$ and $v_o$). The visualization is given in Figure 5. As shown, most of the yellow points are overlapped with purple points and the generated distribution is covered by the real distribution, which demonstrates the effectiveness of the zero-shot learning loss for the modal-shift problem.

4.4.2 Visualization for the Industrial Dataset. For the industrial dataset, we provide a visualization comparison between the new user’s features and old user’s features to validate the assumption that users enjoy a common attribute space to allow the behavior transfer from old users to new users. Specifically, $h^a_s$ is compared with $h^o_o$, and $\hat{v}_s$ is compared with $\hat{v}_o$. The visualization is given in Figure 6. It is observed that the new users’ attribute distribution $h^o_s$ is almost the same as the old user’s attribute distribution $h^o_o$, which validates the assumption that all users enjoy a common attribute space. Moreover, based on the common feature space, the generated $\hat{v}_s$ and $\hat{v}_o$ are also similar to each other, which demonstrates the feasibility of attribute-based behavior transfer from old users to new users.

4.5 Results from Online A/B Testing

We conducted online A/B testing to validate the model performance for real applications in our recommender system. The MAIL-DMR and DMR were separately deployed for comparison, where three million users were assigned to each model. In the half-month of A/B testing, the proposed MAIL-DMR improves CTR by 13% ~ 15% and CTCVR by 3% ~ 4% relatively compared with DMR, which is the last version of the recommender model in our system. Note that users would perform some conversion actions, e.g., purchase goods and collect videos, only when they are genuinely interested in the recommended items. Hence, we highlight the contribution of MAIL-DMR to the improvement of CTCVR in real applications, which is more difficult than the improvement of CTR.

5 CONCLUSION

In this paper, the MAIL model is proposed to address the cold-start problem for recommender systems in a two-tower framework. Inspired by the thought of zero-shot learning, a new zero-shot tower is designed and cotrained with a model-agnostic ranking tower in MAIL. By performing the cross-modal reconstruction and the maximum mean discrepancy-based modal alignment, the zero-shot tower generalizes behavior data from old users to new users based on their attributes. Using the virtual behavior data, the ranking tower can then capture new user’s interests better for recommendation. An attractive property of the proposed method for practical application is that the ranking tower in MAIL can be implemented with any embedding-based ranking models, which
allows MAIL to achieve an incremental performance improvement based on previous works. To verify the effectiveness of MAIL, ranking experiments, an ablation study, and a visualization study are performed on two large-scale real-world datasets. Furthermore, we have deployed the proposed MAIL on an online large-scale live recommendation system of a commercial application to show the significant improvements of CTR and CTCVR.

REFERENCES

[1] Yann Lecun, Yoshua Bengio, Geoffrey Hinton. 2015. Deep learning. Nature, 521, 7553 (2015), 436-444.
[2] Jiajun Li, Ke Lu, Zi Huang, and Heng Tao Shen. 2017. Two birds one stone: on both cold-start and long-tail recommendation. In Proceedings of the 25th ACM international conference on Multimedia (MM), 32 (2017), 898-906. ACM.
[3] Martin Arjovsky, Soumith Chintala, and Leon Bottou. 2017. Wasserstein GAN. In Proc. ICLR, 2017, 1-32.
[4] Jiangli Li, Mengmeng Jiang, Ke Lu, Lei Zhu, Yang Yang, and Zi Huang. 2019. From zero-shot learning to cold-start recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, 33 (2019), 4189-4196.
[5] Xinhuang Wang, Zhaohui Peng, Zhenzhang Wang, S Yu Philip, Wenhui Fu, Xiaokang Xu, and Xiaoguang Hong. 2019. CDLM: cross-domain recommendation for cold-start users via latent feature mapping. Knowledge and Information Systems, 2019, 1-24.
[6] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiaoang Hong, Yang Yang, and Zi Huang. 2018. Recent advances in zero-shot recognition: toward data-efficient understanding of visual content. IEEE Signal Processing Magazine, 35, 1 (2018), 112-125.
[7] Wang-Cheong Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In Proceedings of International Conference on Data Mining (ICDM) 2018, 197-206.
[8] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In Proc. IEEE Conf. CVPR 2009, 951-958.
[9] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoziy Tang. 2017. Collaborative filtering and deep learning based recommendation system for cold start items. Expert Systems with Applications, 69 (2017), 29-39.
[10] Hooyop Lee, Jinhae Im, Seongwon Jang, Hyunsook Cho, and Seheung Chung. 2019. MeLU: meta-learned user preference estimator for cold-start recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019, 1073-1082.
[11] Fan Wang, Wei Yi Zhong, Xiaoang Xu, Wu Jiarui, Zhi Ji, and Zhi Ji. 2020. Privacy-aware cold-start recommendation based on collaborative filtering and enhanced trust. In IEEE 7th International Conference on Big Data Science and Advanced Analytics 2020, 655-663.
[12] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2013. Label-embedding for attribute-based classification. In Proc. IEEE Conf. CVPR, 2013, 819-826.
[13] Yuanlu Fu, Yuan Fang, and Chuan Shi. 2020. Meta-learning on Heterogeneous Information Networks for Cold-start Recommendation. In Proceedings of ACM SIGKDD International Conference on Knowledge & Data Mining. ACM 2020, 1563-1574.
[14] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrew Frome, Greg S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. arXiv pre-print https://arxiv.org/abs/1312.5650, 2013.
[15] Jovian Lin, Karunari Sugiyama, Min-Yen Kan, and Tat-Seng Chua. 2018. Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In ACM SIGIR 2013, 283-292.
[16] Xia Ma, Liqun Zhao, Guan Huang, Zhi Wang, Zelin Hu, and Xiaoliang Hu. 2018. Ensemble multi-task recommendation: an effective approach for estimating post-click conversion rate. In 41st International ACM SIGIR Conference 2018, 1137-1140.
[17] Dheeraj Bolde, Sheetal Girase, and Debajyoti Mukhopadhyay. 2015. Matrix factorization model in collaborative filtering algorithms: a survey. Procedia Computer Science 49 (2015), 136-146.
[18] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2013. Label-embedding for attribute-based classification. In Proc. IEEE Conf. CVPR, 2013, 819-826.
[19] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoziy Tang. 2017. Collaborative filtering and deep learning based recommendation system for cold start items. Expert Systems with Applications, 69 (2017), 29-39.
[20] Hooyop Lee, Jinhae Im, Seongwon Jang, Hyunsook Cho, and Seheung Chung. 2019. MeLU: meta-learned user preference estimator for cold-start recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019, 1073-1082.
[21] Fan Wang, Wei Yi Zhong, Xiaoang Xu, Wu Jiarui, Zhi Ji, and Zhi Ji. 2020. Privacy-aware cold-start recommendation based on collaborative filtering and enhanced trust. In IEEE 7th International Conference on Big Data Science and Advanced Analytics 2020, 655-663.
[22] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2013. Label-embedding for attribute-based classification. In Proc. IEEE Conf. CVPR, 2013, 819-826.
[23] Yuanlu Fu, Yuan Fang, and Chuan Shi. 2020. Meta-learning on Heterogeneous Information Networks for Cold-start Recommendation. In Proceedings of ACM SIGKDD International Conference on Knowledge & Data Mining. ACM 2020, 1563-1574.
[24] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrew Frome, Greg S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. arXiv pre-print https://arxiv.org/abs/1312.5650, 2013.
[25] Jovian Lin, Karunari Sugiyama, Min-Yen Kan, and Tat-Seng Chua. 2018. Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In ACM SIGIR 2013, 283-292.
[26] Xia Ma, Liqun Zhao, Guan Huang, Zhi Wang, Zelin Hu, and Xiaoliang Hu. 2018. Ensemble multi-task recommendation: an effective approach for estimating post-click conversion rate. In 41st International ACM SIGIR Conference 2018, 1137-1140.