M2GRL: A Multi-task Multi-view Graph Representation Learning Framework for Web-scale Recommender Systems

Menghan Wang  
Alibaba Group  
wangmengh@zju.edu.cn

Yujie Lin  
Alibaba Group  
wanrong.lyj@alibaba-inc.com

Guli Lin  
Alibaba Group  
guli.lingl@taobao.com

Keping Yang  
Alibaba Group  
shaoyao@taobao.com

Xiao-ming Wu  
Hong Kong Polytechnic University  
csxmwu@comp.polyu.edu.hk

ABSTRACT
Combining graph representation learning with multi-view data (side information) for recommendation is a trend in industry. Most existing methods can be categorized as multi-view representation fusion; they first build one graph and then integrate multi-view data into a single compact representation for each node in the graph. However, these methods are raising concerns in both engineering and algorithm aspects: 1) multi-view data are abundant and informative in industry and may exceed the capacity of one single vector, and 2) inductive bias may be introduced as multi-view data are often from different distributions. In this paper, we use a multi-view representation alignment approach to address this issue. Particularly, we propose a multi-task multi-view graph representation learning framework (M2GRL) to learn node representations from multi-view graphs for web-scale recommender systems. M2GRL constructs one graph for each single-view data, learns multiple separate representations from multiple graphs, and performs alignment to model cross-view relations. M2GRL chooses a multi-task learning paradigm to learn intra-view representations and cross-view relations jointly. Besides, M2GRL applies homoscedastic uncertainty to adaptively tune the loss weights of tasks during training. We deploy M2GRL at Taobao and train it on 57 billion examples. According to offline metrics and online A/B tests, M2GRL significantly outperforms other state-of-the-art algorithms. Further exploration on diversity recommendation in Taobao shows the effectiveness of utilizing multiple representations produced by M2GRL, which we argue is a promising direction for various industrial recommendation tasks of different focus.

CCS CONCEPTS
• Information systems → Recommender systems; • Computing methodologies → Multi-task learning; Learning latent representations.

KEYWORDS
Recommender system, Graph embedding, Multi-task, Multi-view

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:
Menghan Wang, Yujie Lin, Guli Lin, Keping Yang, and Xiao-ming Wu. 2018. M2GRL: A Multi-task Multi-view Graph Representation Learning Framework for Web-scale Recommender Systems. In KDD ’20: The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 22–27, 2020, San Diego, CA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3320713.3320734

1 INTRODUCTION
Recently, graph-based recommendation algorithms have significantly improved the prediction performance in academia via learning structural relations from graph data. But in industry there still remains many challenges to build a scalable graph-based recommendation algorithm and beat other industrial algorithms, one of which is how to incorporate graph representation learning with side information (e.g., item’s price, user’s profile). Side information (or multi-view data1) depicts different aspects of items (or users) and plays an important role in industrial recommender systems. In Taobao we have billions of items and each item has hundreds of features, and industrial experiences have shown that the huge volume of multi-view data could alleviate the sparsity problem and improve the recommendation performance [21, 24].

There are mainly two lines of research that explored how to utilize multi-view data in graph representation learning. One line of research is treating multi-view data (except rating data) as the attributes of items, which are then fed as input of graph-based algorithms. The other line of research is constructing a heterogeneous graph with multi-view data, and then applying graph representation learning techniques (e.g., metapath2vec [4]) to learn item embeddings. From the perspective of multi-view learning, these two kinds of research can be categorized into multi-view representation fusion. That is, data from multiple views are integrated into a single compact representation. In practice, these methods can effectively address the sparsity problem in recommendations [21].

However, multi-view representation fusion methods raise concerns in both engineering and algorithm aspects when deployed to web-scale recommendation tasks. First, one single vector of representation may lack the capacity to embed multi-view data. In

1 We use “multi-view data” to denote “side information” because we argue that “multi-view data” is a more general and reasonable term in the industrial scenario. In this paper, “multi-view data” also includes the user-item rating data.
We explicitly preserve the local structure of single-view data and which seeks to perform alignment between representations learned vectors especially, which could save a lot of engineering efforts. 2) view graph, the instance-level (item) graph, and the category-level decrease the risk of inductive biases.

So we do not need to configure size of representation methods: 1) We use multiple representation vectors sentation fusion from multiple different views. Following this methodology, we can influence the final representations of the three nodes. The idea of relations of nodes of different graphs should be modeled to make from three graphs, i.e., the green, blue, and yellow graph. Then, in Figure 1 we can learn three independent sets of representations cross-view relations between different graphs. More concretely, in industry there are billions of items and every item may serve multiple interests and needs. But due to engineering concerns, the size of representation vector is often set to 128 or 256 even for algorithms based on rating data (single-view), which is already a trade-off between efficiency and accuracy. Fusing multi-view data into one small vector may further sacrifice the accuracy. Second, multi-view data may come from different distributions; inductive bias may be introduced if they are not properly handled. For example, as shown in Figure 1, we can construct three single-view graphs from users’ behavior sequences (self-loops are ignored), i.e., the shop-view graph, the instance-level (item) graph, and the category-level graph. These three graphs have different structures and contain unique information, which is ignored by multi-view representation fusion methods. A more proper utilization of graph representation learning with multi-view data is needed in industry.

Instead, we argue that a more plausible method is to first learn a separate graph representation for each view of data, and then model cross-view relations between different graphs. More concretely, in Figure 1 we can learn three independent sets of representations from three graphs, i.e., the green, blue, and yellow graph. Then, relations of nodes of different graphs should be modeled to make the learned representations more reasonable and interpretable. For example, the ID 1 shoe J1 belongs to category C2 and is sold in shop S1, so the I1 – C2 and J1 – S1 cross-view relations should influence the final representations of the three nodes. The idea of modeling cross-view relations is initially inspired by multi-view representation alignment [15], another branch in multi-view learning, which seeks to perform alignment between representations learned from multiple different views. Following this methodology, we can circumvent the two above-mentioned concerns of multi-view representation fusion methods: 1) We use multiple representation vectors to represent multi-view data and the procedure can be distributedly performed. So we do not need to configure size of representation vectors especially, which could save a lot of engineering efforts. 2) We explicitly preserve the local structure of single-view data and model cross-view relations, which is more reasonable and should decrease the risk of inductive biases.

Particularly, in this paper, we propose a multi-task multi-view graph representation learning framework (M2GRL) to learn node representations for web-scale recommender systems. M2GRL constructs one graph for each single-view data and learns multiple separate representations from multiple graphs, and then performs alignment to model cross-view relations. M2GRL is composed of two types of tasks: intra-view task and inter-view task. The intra-view task learns the representations of nodes within a single view, and the inter-view task models the cross-view relations of two different graphs. As M2GRL tends to have many intra-view and inter-view tasks, we apply homoscedastic uncertainty [13] to adaptively tune the loss weights of tasks during training. The main contributions of this paper are:

- We propose a novel framework M2GRL for graph representation learning with multi-view data. To our best knowledge, it is the first work to apply multi-view representation alignment in web-scale recommender systems.
- Our M2GRL is scalable, flexible and extensible. M2GRL supports unlimited number of views of data and can be distribut edly deployed to handle billion-scale data samples. Also, existing graph representation learning algorithms can be easily incorporated into M2GRL. Besides, the multiple representations learned by M2GRL provide embeddings of items from different aspects, which can benefit downstream recommendation tasks with different focus.
- Through extensive offline experiments and online A/B tests, we show that M2GRL achieves state-of-the-art performance compared to other industrial recommendation algorithms. Further, a use case of diversity recommendation in Taobao demonstrates the effectiveness and benefit of the use of multiple representations learned by M2GRL.

The rest of the paper is organized as follows. In Section 2, we introduce the related works, including graph representation learning for recommendation and recommendation with multi-view data. Our proposed M2GRL and its implementation details are presented in Section 3. We then show offline and online experimental results in Section 4. In Section 5, we describe a use case of diversity recommendation in Taobao with multiple representations produced by M2GRL. Finally, we conclude our work in Section 6.

2 RELATED WORK

In this section, we discuss the following two lines of research work that are closely relevant to this paper.

2.1 Graph Representation Learning for Recommendation

Graph representation learning is a hot topic in research community as it can learn structure informations from graph data. In the past few years, many algorithms have been proposed to learn graph representations. These methods could be categorized into three broad categories: 1) Factorization methods (e.g., LINE [1], NetSMF [19]) aim to approximately factorize the adjacency matrix and preserve structural proximities. 2) Random walk based techniques(e.g., DeepWalk [18], Node2Vec [6], metapath2vec [4]) use random walks on graphs to obtain node representations; they can be easily deployed in a distributed manner and thus are widely
used in industrial applications. 3) Graph convolutional networks (GCNs) [2, 8, 9] perform (approximate) spectral convolutions to aggregate information from node’s neighborhood in graphs. GCNs have shown superiority against factorization methods and random walk based method in small datasets. But GCNs suffer from the efficiency and over-smoothing problems, which prevent their further expansion in industry. In the recommendation area, all the three kinds of methods have been explored to improve the performance of recommender systems. Especially, in industry, Wang et al. [21] propose a graph embedding algorithm that first uses random walks to generate samples and then apply word2vec [17] to learn node representations for recommendation at Taobao. Ying et al. [24] develop a GCN algorithm that combines random walks and graph convolutions to generate embeddings of nodes for recommendation at Pinterest. In this paper, we propose a general framework that can cope with all the three kinds of methods.

2.2 Recommendation with Multi-view Learning

In multi-view representation learning, there are two major kinds of methods: 1) multi-view representation fusion, which tries to fuse multi-view data into a single compact representation. 2) multi-view representation alignment, which aims to capture the relationships among multiple different views through feature alignment. In recommendation literature, multi-view data except rating data are collectively called side information, such as temporal information, item’s description, and users’ social network. Currently multi-view representation fusion methods are the main stream in recommendation area [10, 16, 22, 25, 26]. However, there are also some multi-view representation alignment methods. For example, Elkahky et al. [5] use a deep learning approach to map users and items to a latent space where the similarity between users and their preferred items is maximized. Jiang et al. [12] introduce a deep cross-modal retrieval method, which considers learning a multi-modal embedding from the perspective of optimizing a pairwise ranking problem while enhancing both local alignment and global alignment. However, these works are not designed for graph data, whereas the focus of this paper is to align embeddings of nodes across different graphs.

3 M2GRL FRAMEWORK

In this section, we first introduce the problem setup and how we construct graphs from users’ behavior history. Then we present the overall structure of our M2GRL framework and explain its components in detail.

3.1 Problem Setup

Our task is to learn high-quality representations of items in E-commerce scenarios. These representations can then be used for recommendation by either a nearest-neighbor lookup or many downstream ranking algorithms. As an item contains sets of features that reflect different aspects of the item, we generate multiple representations for items based on their features. Concretely, we use three-view data in this paper: 1) The instance-view data records the user-item rating data. 2) The category-view data records the category information of items. Category information is a high-level abstract of items. 3) The shop-view data records the shop information of items. Taobao is a consumer-to-consumer (C2C) platform with millions of shop selling goods on it. Many shops may sell a same item but with different prices and services. Note that M2GRL is scalable to more views of data. For clarity and convenience, below we will describe our method with the above three-view data.

Graph construction. We first construct three graphs (i.e., item (instance) graph, category graph, and shop graph) for M2GRL from users’ sequential behaviors. Take the item graph for example, we assume that two items are connected if they occur consecutively in one user’s behavior history. The construction of category and shop graph are similar to that of item graph. Figure 1 shows an example of graph construction on three-view data. As we remove consecutive duplicates, the three graphs are of different structures.

Node sequence sampling. A common method to generate node sequences for training samples is random walk and its variants. In practice, we generate training samples via extracting sessions from users’ behavior history. The main phases are as follows:

- Data Clean. We noticed that users may click one item and quickly go back to previous pages, which indicates that users are probably not interested in that item. We remove items whose duration of the stay after a click are less than two seconds. This operation is also applied in graph construction.
- Session split and merge. We extract timestamps from log data that record when user open and close the Taobao App, and use these timestamps to split user behaviors into sessions. Sometimes a session will be hours-long (e.g., running App in the background), so we split a session into two sessions if there is a one-hour idle period. As for session merge, we merge two consecutive sessions if the time span is less than 30 minutes.

The category and shop data is bundled with item data, so once an item session is determined we can easily get the corresponding category and shop sessions.

3.2 Overall Structure

We present the graphical structure of M2GRL in Figure 2. Particularly, M2GRL consists two types of tasks: the intra-view task learns the representations of item within a single-view graph, and the inter-view task models the relations of nodes between two different views. We choose five tasks where three tasks are intra-view tasks and the others are inter-view tasks. All the representations generated by the intra-view and inter-view tasks are collectively called multiple representations.

3.3 Intra-view Representation Learning

We treat the intra-view task as a representation learning problem on a homogeneous graph. The intra-view tasks are conditionally independent given the inter-view tasks, so we can apply state-of-the-art graph representation learning methods without much modification. Here we choose the skip-gram model with negative sampling (SGNS), a simple but effective model, to learn node embeddings. SGNS is scalable and can be easily deployed in a distributed manner; many works have shown its effectiveness in extreme large graphs.

We first generate node sequences with the node sequence sampling method discussed in Sec. 3.1. Then we apply the Skip-Gram
algorithm \cite{17} to learn node embeddings. The training objective of the Skip-gram model is to find node representations that are useful for predicting the surrounding nodes in a node sequence. More formally, take the item graph for example, given a sequence of training items \( i_1, i_2, i_3, \ldots, i_T \), the objective of the Skip-gram model is to maximize the average log probability

\[
L_{\text{intra}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(i_{t+j} | i_t) 
\]

where \( c \) is the window size of the context nodes in the sequences. In practice, we find \( c = 9 \) is a good tradeoff between accuracy and training efficiency. The basic Skip-gram formulation defines \( p(i_{t+j} | i_t) \) using the softmax function:

\[
p(i_0 | i_t) = \frac{\exp(p_{i_0}^T v_{i_t})}{\sum_{w=1}^{W} \exp(p_w^T v_{w})} \]

where \( v_{w} \) and \( p_{i_0} \) are the “input” and “output” vector representations of \( w \), and \( W \) is the number of items in the item vocabulary.

We use the negative sampling method to approximately maximize the log probability of the softmax function. We first apply negative sampling to generate \( k \) negative samples for each positive example, and then Eq. (2) can be transformed into

\[
\log \sigma(p_{i_0}^T v_{i_t}) + \sum_{t=1}^{k} \log \sigma(-v_{i_0}^T v_{i_j}) \]

where \( P_n(i) \) is the noise distribution for negative sampling, and \( \sigma \) is the sigmoid function \( \sigma(x) = \frac{1}{1 + e^{-x}} \). The goal of Eq. (3) is to distinguish the representation of the target item and negative samples. The final loss function is Eq. (1) with log \( p(i_{t+j} | i_t) \) replaced by Eq. (3).

### 3.4 Inter-view Alignment

The goal of inter-view tasks is to model the cross-view relations (e.g., instance - category (I-C) and instance - shop (I-S)). If one item \( i \) has attribute \( x \), we define there is a cross-view relation \( (i, x) \) between the item and \( x \), and vice versa. Instead of directly adding constraints to different embedding spaces, we propose an inter-view alignment technique to transform information across views and learn associations of entities in a relational embedding space. Note that one inter-view task is responsible for one type of cross-view relations (e.g., I-C relations or I-S relations). Specifically, taking the I-C alignment task for example, we first map two different embeddings into one relation space via an alignment transformation matrix \( W_{ic} \). Then, a loss of a inter-view task is formulated as Eq. 4,

\[
L_{\text{inter-}c} = \sum_{(i,c) \in S(i,c)} \sigma(W_{ic} \cdot e_i) \sigma(W_{ic} \cdot e_c) - \sum_{(i,c \neq S(i,c))} \sigma(W_{ic} \cdot e_i) \sigma(W_{ic} \cdot e_c),
\]

where \( \sigma \) is the sigmoid activation function, and \( c' \) are chosen by negative sampling. The inter-view tasks also generate representations; \( \sigma(W_{ic} \cdot e_i) \) and \( \sigma(W_{ic} \cdot e_c) \) can be regarded as two representations from different aspects in the relational embedding space.

### 3.5 Learning Task Weights with Homoscedastic Uncertainty

M2GRL is concerned about learning a global optimization with respect to multiple related tasks. A naive but popular approach to compute the final loss function is to simply perform a weighted linear sum of the loss of each individual task:

\[
L_{\text{total}} = \sum_i w_i \cdot L_{\text{intra}_i} + \sum_j w_j \cdot L_{\text{inter}_j},
\]

where \( \{w_i\} \) and \( \{w_j\} \) are hyper-parameters that balance the importance of losses. However, manually tuning these hyperparameters are expensive and intractable in web-scale recommendation scenarios. An alternative approach \cite{23} in practice is to optimize each task iteratively, which, may stick at a local optimum and fail to perform well in some tasks.

Instead, we use the idea of homoscedastic uncertainty \cite{13} to weigh the loss of each task automatically during model training. Follow the work of \cite{14}, we adjust each task’s relative weight in the final loss function by deriving a multi-task loss function based on maximizing the Gaussian likelihood with task-dependant
We use the approximated form of Eq. (10) for each task with high similarity scores in the map are chosen as candidates. With uncertainty as follows:

\[ L_{t,\text{total}}(x; \theta, \sigma_t) = \sum_{t} L_t(x_t; \theta_t, \sigma_t). \]  

(6)

where \( L_t \) is the classification loss function for task \( t \), \( x_t \) and \( \theta_t \) are the data and model parameter for task \( t \) (either intra-view task or inter-view task), and \( \sigma_t \) is the corresponding task uncertainty.

We represent the likelihood of the model for each task as a scaled version of the model output \( f_t(x) \) with the uncertainty \( \sigma_t \) squashed through a softmax function:

\[ P(C = c|x_t, \theta_t, \sigma_t) = \frac{\exp\left(\frac{1}{\sigma_t} f_c(x_t)\right)}{\sum_{c'=1}^{C} \exp\left(\frac{1}{\sigma_t} f_{c'}(x_t)\right)}. \]  

(7)

Using the negative log likelihood, we express the classification loss with uncertainty as follows:

\[ L_t(x_t, \theta_t, \sigma_t) = \sum_{c=1}^{C} -C_t \log P(C_c = 1|x_t, \theta_t, \sigma_t) \]

\[ = -C_t \log(\exp(\frac{1}{\sigma_t} f_c(x_t))) + \log \sum_{c'=1}^{C} \exp\left(\frac{1}{\sigma_t} f_{c'}(x_t)\right). \]  

(8)

Applying the same assumption in [14]:

\[ \exp\left(\frac{1}{\sigma_t} f_c(x_t)\right) \approx \left(\sum_{c'} \exp(f_{c'}(x_t))\right)^{-\sigma_t} \]  

(9)

allows to simplify Eq. 8 to:

\[ L_t(x_t, \theta_t, \sigma_t) \approx \frac{1}{\sigma_t} \sum_{c=1}^{C} -C_t \log P(C_c = 1|x_t, \theta_t) + \log(\sigma_t^2). \]  

(10)

We use the approximated form of Eq. (10) for each task \( t \) in Eq. (6) and get the final multi-task loss. The \( \sigma_t \) can be interpreted as the relative weight of the loss of task \( t \). When \( \sigma_t \) increases, the corresponding weight decreases. Additionally, \( \log(\sigma_t^2) \) serves as a regularizer to avoid overfitting. For numerical stability, we trained the network to predict \( \log(\sigma_t^2) \) instead of \( \sigma_t^2 \). All network parameters and the uncertainty tasks weights are optimized with stochastic gradient descent (SGD).

### 4 SYSTEM DEPLOYMENT

In this section, we introduce the deployment of M2GRL in Taobao’s recommendation platform as shown in Figure 3. In the candidate generation stage, the system constructs multiple single-view graphs after extracting multi-view data from the raw log. Then the system runs M2GRL in a distributed manner to produce the multiple representations. In Taobao we have many recommendation modules (downstream tasks) and we use different strategies (e.g., task 1 and task 2 in Fig. 3) to generate \( I2I \) (item2item) similarity maps based on the goal of each module. Given an item as a request, items with high similarity scores in the map are chosen as candidates. Note that the size of candidates is much smaller (usually thousands) compared with the size of the corpus (hundreds of millions). Then in the ranking stage, rank models request \( I2I \) similarity maps for candidates and use online models to score candidates. Items with top scores are finally displayed to users as recommendations.

Training one version of M2GRL takes a dozen hours, so the model is deployed offline in a daily-run mode. We currently support two downstream tasks. One is general recommendation that generates the similarity maps using inner product search on the instance-view representations of M2GRL. The other is diversity recommendation that first uses the multi-view metric model introduced in Sec. 6.1 to produce new representations, and then generate the similarity maps via inner product search. More downstream tasks with different focus are under development.

### 5 EXPERIMENTS

In this section, we conduct comprehensive experiments to verify the effectiveness of our proposed M2GRL framework.

#### 5.1 Offline Evaluation

##### 5.1.1 Datasets and Experimental Settings

We choose two datasets for evaluating recommendation performance. One is Taobao, which is one of the most widely-used public dataset for recommendations. For Taobao, we construct three single-view graphs (i.e., shop, instance, category) according to Section 3.1. For the Movielens, we treat tag of a movie as the movie’s category information and construct single-view graphs similarly. Note that Movielens does not have shop information. So we only use two graphs in M2GRL for Movielens. The statistics of the two datasets are shown in Table 1.

---

2https://grouplens.org/datasets/movielens/
Table 1: Datasets Statistics.

| Dataset    | #Item   | #Category | #Shop   | #Item-edges | #Cate-edges | #Shop-edges | #Item-Cate links | # Item-Shop links |
|------------|---------|-----------|---------|-------------|-------------|-------------|------------------|-------------------|
| Taobao     | $1.0 \times 10^7$ | $1.2 \times 10^4$ | $1.3 \times 10^6$ | $1.8 \times 10^{10}$ | $3.5 \times 10^5$ | $4.4 \times 10^5$ | $1.5 \times 10^8$ | $1.4 \times 10^8$ |
| Movielens  | 27,278  | 1,128     | -       | 10,463,449  | 1,712,087   | -           | 9,110,188        | -                 |

Table 2: Comparisons of different models on offline datasets.

| Models   | Taobao | Movielens |
|----------|--------|-----------|
|          | HitRate@50 | Recall@50 | Precision@50 | F1@50 | HitRate@50 | Recall@50 | Precision@50 | F1@50 |
| LINE     | 9.55%   | 1.96%     | 0.22%     | 0.40% | 24.89%  | 11.08%    | 0.62%     | 1.17%  |
| DeepWalk | 31.66%  | 12.11%    | 1.13%     | 2.06% | 34.72%  | 17.43%    | 0.97%     | 1.83%  |
| Node2Vec | 31.79%  | 12.20%    | 1.14%     | 2.08% | 36.43%  | 18.14%    | 1.01%     | 1.91%  |
| EGESaxy  | 32.42%  | 12.45%    | 1.17%     | 2.13% | 42.46%  | 22.01%    | 1.23%     | 2.33%  |
| GraphSage| 7.16%   | 1.98%     | 0.19%     | 0.35% | 2.23%   | 0.76%     | 0.05%     | 0.09%  |
| GRU4Rec  | 27.32%  | 11.45%    | 1.08%     | 1.97% | 30.81%  | 13.39%    | 0.76%     | 1.44%  |
| YouTube-DNN | 29.97% | 11.97%    | 1.11%     | 2.03% | 38.27%  | 18.73%    | 1.06%     | 2.01%  |
| M2GRL    | 33.24%  | 12.65%    | 1.19%     | 2.17% | 42.62%  | 22.16%    | 1.24%     | 2.35%  |

- **Deepwalk** [18]. This approach learns low-dimensional feature representations for nodes in graphs by simulating uniform random walks.
- **LINE** [20]. This method learns graph embedding via a customized loss function to preserve the first-order or second-order proximities in graphs separately.
- **Node2Vec** [6]. This algorithm treats feature learning in graphs as a search-based optimization problem. It takes a biased random walk that is a tradeoff between DFS and BFS.
- **EGESaxy** [21]. This method incorporates side information into graph embedding via a weighted average layer to aggregate the embeddings of side information. It is a typical multi-view representation fusion method.
- **Graphsage** [8]. This method learns node embeddings via aggregating information recursively from neighborhoods of the nodes in a graph. It is a variant version of graph convolutional networks.
- **GRU4Rec** [11]. A GRU-based RNN model for session-based recommendations, which outperforms traditional methods significantly.
- **YouTube-DNN** [3]. A deep neural network based recommendation approach proposed by YouTube, which is popularly used for industrial recommendation systems.

Hyperparameter tuning is conducted by grid search, and each method is tested with best hyperparameters for a fair comparison. For M2GRL, we generate recommendations via inner product search on the instance-view(item) representations.

### Offline metrics

To evaluate the offline performances of different methods, we use **HitRate@K**, **Precision@K**, **Recall@K** and **F1@K** metrics.

- **HitRate@K** represents the proportion of test cases ($n_{hit}$) which has the correctly recommended items in a top K position in a ranking list, defined as
  \[
  \text{HitRate@K} = \frac{n_{hit}}{N}
  \]
  where $N$ denotes the number of test data.

- **Recall@K** represents the ability of coverage in the user’s ground truth. It’s calculation is
  \[
  \text{Recall@K}(u) = \frac{|P_u \cap G_u|}{|G_u|}
  \]
  where $G_u$ is the best baseline. It beats Node2Vec, DeepWalk, and LINE (the same line of algorithms), mainly due to the utilization of side information. Side information can alleviate the sparsity problem and thus improve the performance of recommendations. However, in EGESaxy embeddings of side information are directly averaged with item embedding, ignoring the heterogeneities of items and side information. Our method M2GRL takes a multi-view representation alignment method to address this issue and outperforms EGESaxy. 2) Node2Vec and DeepWalk outperform LINE because Line only preserve information of a two-hop neighborhood, while Node2vec and DeepWalk can learn information from a variable-length range of neighborhood. Moreover, Node2vec takes weighted random walks and is superior to DeepWalk on both the Movielens and Taobao datasets. 3) We notice that Graphsage performs very poorly. One main reason is that the item graph is built on users’ sequential behaviors and it is dense and noisy, which may interfere the effectiveness of Graphsage because the training phase of Graphsage aims to learn aggregator parameters rather than representations of each node in a graph. In addition, the shallow structure of Graphsage (two or three layers) also limits the performance of learned representations. 4) GRU4Rec and YouTube-DNN
are sequential recommender systems, and they are empirically inferior to random-walk-based methods (e.g., DeepWalk and M2GRL). Meanwhile, deploying sequential recommendation algorithms to online environment needs extra engineering efforts and computation overhead: the system has to call users’ sequential behavior to predict and rank items for every recommendation request. In contrast, M2GRL uses a lookup table. From the above analysis, we can conclude that M2GRL is scalable and competitive.

5.2 Online A/B Test

We report the online A/B test experiments of M2GRL and other deployed industrial algorithms (i.e., EGES and Item-CF) at Taobao. We choose top popular 150 million items and use 57 billion samples for training. M2GRL is implemented on PAI-tensorflow (a machine learning platform in Alibaba), and is trained in a distributed manner with 25 server nodes and 100 worker nodes. The total amount of memory used in training is 2000GB. The chosen metric is Click-Through-Rate (CTR) on the homepage of Mobile Taobao App. We walk, and Item-CF consistently in terms of CTR, which demonstrates the effectiveness of utilization of multi-view graph embedding. Compared with the best baseline EGES, M2GRL achieves a 5.76% improvement on average, which is a significant boost in recommendation performance considering that we have hundreds of millions of active users in Taobao everyday. Further, EGES outperforms DeepWalk consistently, which reveals the effectiveness of the side information. The results is similar to what we conclude in the offline experiments.

5.3 Analysis of Loss Weighting

To evaluate the influence of the adaptive loss weighting mechanism (denoted by M2GRL_adaptation), we experiment with two variants of M2GRL: M2GRL_uniform uses uniform weights and M2GRL_static uses manually assigned weights. We perform a grid search and find the optimal weights are approximately proportional to the distribution of the train data. Concretely, we set 0.05 as the loss weight for the instance-view task, and 1 for other tasks. From Table 3, we can find that M2GRL_adaptation outperforms other two versions on two datasets, showing the superiority of the adaptive loss weighting mechanism. Note that M2GRL_static performs better than M2GRL_uniform, but finding proper task weights is too expensive. It’s almost impossible to search task weights manually for a daily updated industrial algorithm due to computation cost concerns.

Besides, in Figure 5 we display the uncertainty variable \(\sigma^2\) (loss weight is \(1/\sigma^2\)) and losses of two tasks of M2GRL: the I-C inter-view task and the I intra-view task. Note that the horizontal axis is the number of training steps and the loss weight of each task is \(1/\sigma^2\).

| Dataset   | M2GRL_uniform | M2GRL_static | M2GRL_adaptation |
|-----------|---------------|--------------|-----------------|
| Movielens | 42.24%        | 42.50%       | 42.62%          |
| Taobao    | 33.08%        | 33.11%       | 33.24%          |
threshold. First we can see that $\sigma^2$ in Figure 5a decreases as the training step increases, and $\sigma^2$ in Figure 5b is below threshold so the loss weight is set to 200. Compared with the I task, the loss weight of the I-C task increases relatively. Then in Figure 5c and 5d the loss of both tasks decrease smoothly as the training steps increase. Note that the loss of I-C task is much smaller and thus is prone to be interfered by the I task. The adaptive weight mechanism helps to prevent lost fluctuations and guarantee convergence of each task.

5.4 Visualization
In this section, we present some representations of real-world cases in Taobao to further illustrate the effectiveness of M2GRL. We visualize the instance embeddings of items via the Embedding Projector, a visualization tool provided by Tensorflow. We apply principal component analysis (PCA) and draw the results in Figure 6. In Figure 6a, we choose five cloth-related categories (i.e., T-shirt, coat, suit, windcheater, and cheongsam), and display 1000 randomly chosen items for each category. Here one color indicates one category of items. We can find that items of one category are clustered in the feature space and distances between different categories are different. There are two interesting observations: the T-shirt clustering is far away from other clusterings, and the clusterings of windcheaters and coats have many overlaps. The far distance can be explained that T-shirts have short sleeve and sometimes treated as undergarment, while items of other four categories are outer garment. The reasons of overlaps are that windcheaters and coats are similar in function and windcheaters are annotated as coats in some shops in Taobao. Although users may click these cloth-related items iteratively in a short time, M2GRL can well capture their category relations, which demonstrates the effectiveness of the learned embeddings of M2GRL with multi-view data. In Figure 6b, we also display the embeddings of 12,000 categories in the Taobao dataset. We can find that these categories are distributed unevenly. Some categories tend to form clusterings and others are rather discrete, which further reveal M2GRL’s ability to model category relations.

6 BEYOND RECOMMENDATION ACCURACY
A high recommendation accuracy is not a silver bullet to a successful recommender system; there are many other factors need to be considered to evaluate broader aspects of user satisfaction, such as diversity, utility, and user coverage. We argue that utilizing multiple representations is a promising direction for various recommendation tasks with different focus. Below we describe a use case of diversity recommendation in Taobao with multiple represenations.

6.1 Diversity Recommendation with Multiple Representations - A Downstream Use Case
In Taobao, recommendation diversity is an important metric with respect to user’s long-term experience. Customers will get bored if the system keeps recommending similar items of one category (e.g., skirts with different colors). Diversity recommendation in Taobao refers to recommending items whose categories did not appear in users’ recent behavior logs (usually two weeks). In traditional recommendation algorithms, items with the same category tend to have a high similarity score, which contradicts the goal of diversity recommendation. We argue that the multiple representations learned by M2GRL can be utilized for diversity recommendation in a principled way. Heuristically, diversity recommendation is about finding two items that are close in instance-view embedding space but far away in category-view embedding space. Thus, we propose a simple multi-view metric learning model. We use two representations, one is $e_i$ from instance-view embedding space and the other is $e_{ic} = \sigma(W_{ic} \cdot e_i)$ from instance-category relational embedding space. For a item pair $(a, b)$, we use two metric matrices ($i.e., M_i$ and $M_{ic}$) to measure their distance:

$$d_i(a, b) = (e^a_i - e^b_i)^T M_i(e^a_i - e^b_i)$$

$$d_{ic}(a, b) = (e^a_{ic} - e^b_{ic})^T M_{ic}(e^a_{ic} - e^b_{ic})$$

$$d(a, b) = d_i(a, b) + d_{ic}(a, b)$$

Then we use contrastive loss [7] to learn the metric matrices, which is

$$L_{metric} = \frac{1}{2N} \sum_{n=1}^{N} y_{n}d^2 + (1 - y)\max(margin - d, 0)^2,$$
We propose M2GRL, a multi-task multi-view graph representation learning framework for web-scale recommender systems. M2GRL constructs one graph for each single-view data, learns multiple separate representations from multiple graphs, and performs alignment to model cross-view relations. M2GRL supports unlimited number of views of data and can be distributedly deployed to handle billion-scale data samples. We also design an adaptive weight tuning mechanism to make M2GRL more scalable. We deployed M2GRL at Taobao, and the offline experiments and online A/B test both shown the superiority of M2GRL over other competitive benchmarks. Besides, we explore a multi-view metric model for diversity recommendation with two representations learned by M2GRL, which shows a promising result. As there are many recommendation tasks with different focus in industry, we argue that useful representations generated by M2GRL can be further utilized in the future to tackle these, e.g., tag recommendation and recommendation with explanations.

REFERENCES

[1] Amr Ahmed, Nino Shervashidze, Shrvan Narayanamurthy, Vanja Josifovski, and Alexander J Smola. 2013. Distributed large-scale natural graph factorization. In WWW. ACM, 37–48.

[2] Joao Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013).

[3] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM conference on recommender systems. ACM, 191–198.

[4] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In SIGKDD. ACM, 135–144.

[5] Ali Mandyoub Elkahlby, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In WWW. WWW, 278–288.

[6] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In SIGKDD. ACM, 855–864.

[7] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In CVPR. Vol. 2. IEEE, 1735–1742.

[8] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems. 1024–1034.

[9] William I. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. arXiv preprint arXiv:1709.05584 (2017).

[10] Xiaogang He, Lizi Liao, Hanwang Zhang, Luqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In WWW International World Wide Web Conferences Steering Committee, 173–182.

[11] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1510.06939 (2015).

[12] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. 2016. LINE: Large-scale information network embedding. In Proceedings of the 23rd ACM international conference on Multimedia. ACM, 69–78.

[13] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In Advances in neural information processing systems. 5574–5584.

[14] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In CVPR. IEEE, 7482–7491.

[15] Yingming Li, Ming Yang, and Zhouchen Mark Zhang. 2018. A survey of multi-view representation learning. IEEE Transactions on Knowledge and Data Engineering (2018).

[16] Dawn Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebra. 2018. Variational autoencoders for collaborative filtering. In Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, 689–698.

[17] Tomasz Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).

[18] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In SIGKDD. ACM, 701–710.

[19] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Chi Wang, Kuansan Wang, and Jie Tang. 2019. Netmf: Large-scale network embedding as sparse matrix factorization. In WWW. ACM, 1509–1520.

[20] Jian Tang, Meng Qu, Mingzhe Wang, Huaiyan Wang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In WWW. ACM, 1067–1077.

[21] Jizhe Wang, Piqee Huang, Huan Zhao, Zhbo Zhang, Binqiang Zhao, and Dik Lun Lee. 2018. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In SIGKDD. ACM, 839–848.

[22] Menghan Wang, Xiaolin Zheng, Yang Yang, and Kun Zhang. 2018. Collaborative filtering with social exposure: A modular approach to social recommendation. In Thirty-Second AAAI Conference on Artificial Intelligence.

[23] Wei Xu, Wei Liu, Haoyuan Chi, Xiaolin Huang, and Jie Yang. 2018. Multi-task classification with sequential instances and tasks. Signal Processing: Image Communication 64 (2018), 59–67.
we use 20 PSs and GPU 100 workers. Computing resources: 1) CPU: 4
2) GPU: Tesla P100-PCIE-16GB.

After preprocessing, we construct item graph and category graph. For
Taobao, we use 10 parameter severs (PSs) and 50 GPU workers. For
online experiment, we split a session into two sessions. We use timestamps
to split user behaviors into sessions. We noticed that users’ sequential behaviors are
more sparse in Movielen; the time lag between two consecutive rated films may be larger than multiple years. To get a proper
length of data sessions for M2GRL, we split a session into two sessions if there is a one-year idle period. Moreover, if a session is too long (larger than 50), we split the session
into two sessions.

After preprocessing, we construct item graph and category graph. Take the item graph for example, we assume that two items are
connected if they occur consecutively in one user’s behavior history.

8.2 Running Environment
All the models are experimented on MaxCompute 4, which is a
data processing and computing platform in Alibaba. To be fair, all of these models share the same training and testing datasets. For
Taobao, we use 10 parameter severs (PSs) and 50 GPU workers. For
Movielen, we use 1 PS and 1 GPU worker. For online experiment, we use 20 PSs and GPU 100 workers. Computing resources: 1) CPU: each PS has 20 cores and 30GB memory, and each worker has 8 cores and 10GB memory. 2) GPU: Tesla P100-PCIE-16GB.

8.3 Algorithm settings and parameters
In this section we describe some algorithm settings and parameters
related to the experiments. EGESasy is DeepWalk, Node2Vec, Line,
and Graphsage are tested on the Aligraph, a graph representation
learning platform in Alibaba. We implement distributed versions of
GRU4Rec and YouTube-DNN to support the big data. We mainly use
the original parameters for baselines. Below we list some important
parameters of each algorithm. If not specifically noted, the batch
size is set to 256.

8.4 Code Release
We have prepared an experiment version of M2GRL code to release. It will take sometime to be available online due to company’s policy.

---

4https://cn.aliyun.com/product/odps