RESEARCH-ARTICLE

A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios

CHRISTIAN GANHÖR, Johannes Kepler University Linz, Linz, Upper Austria, Austria

MARTA MOSCATI, Johannes Kepler University Linz, Linz, Upper Austria, Austria

ANNA HAUSBERGER, Johannes Kepler University Linz, Linz, Upper Austria, Austria

SHAH NAWAZ, Johannes Kepler University Linz, Linz, Upper Austria, Austria

MARKUS SCHEDL, Johannes Kepler University Linz, Linz, Upper Austria, Austria

Open Access Support provided by:

Johannes Kepler University Linz

Published: 08 October 2024

Citation in BibTeX format

RecSys '24: 18th ACM Conference on Recommender Systems
October 14 - 18, 2024
Bari, Italy

Conference Sponsors:
SIGCHI

RecSys '24: Proceedings of the 18th ACM Conference on Recommender Systems (October 2024)
https://doi.org/10.1145/3640457.3688138
ISBN: 9798400705052

A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios

Christian Ganhör*
christian.ganhoer@jku.at
Institute of Computational Perception,
Johannes Kepler University Linz
Linz, Austria

Marta Moscati*
marta.moscati@jku.at
Institute of Computational Perception,
Johannes Kepler University Linz
Linz, Austria

Anna Hausberger
anna.hausberger@jku.at
Institute of Computational Perception,
Johannes Kepler University Linz
Linz, Austria

Shah Nawaz
shah.nawaz@jku.at
Institute of Computational Perception,
Johannes Kepler University Linz
Linz, Austria

Markus Schedl
markus.schedl@jku.at
Institute of Computational Perception,
Johannes Kepler University Linz and
Human-centered AI Group, AI Lab,
Linz Institute of Technology
Linz, Austria

ABSTRACT

Most recommender systems adopt collaborative filtering (CF) and provide recommendations based on past collective interactions. Therefore, the performance of CF algorithms degrades when few or no interactions are available, a scenario referred to as *cold-start*. To address this issue, previous work relies on models leveraging both collaborative data and side information on the users or items. Similar to multimodal learning, these models aim at combining collaborative and content representations in a shared embedding space. In this work we propose a novel technique for multimodal recommendation, relying on a multimodal **Si**ngle-**Bra**nch embedding network for **R**ecommendation (`SiBraR`). Leveraging weight-sharing, `SiBraR` encodes interaction data as well as multimodal side information using the same single-branch embedding network on different modalities. This makes `SiBraR` effective in scenarios of missing modality, including cold start. Our extensive experiments on large-scale recommendation datasets from three different recommendation domains (music, movie, and e-commerce) and providing multimodal content information (audio, text, image, labels, and interactions) show that `SiBraR` significantly outperforms CF as well as state-of-the-art content-based RSs in cold-start scenarios, and is competitive in warm scenarios. We show that `SiBraR`'s recommendations are accurate in missing modality scenarios, and that the model is able to map different modalities to the same region of the shared embedding space, hence reducing the modality gap.

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; *Personalization*; **Recommender systems**; **Multimedia and multimodal retrieval**.

KEYWORDS

Cold-start Recommendation, Recommender Systems, Hybrid Recommender System, Content-based Recommender System, Multimodal Models, Single-Branch Network, Weight Sharing, Missing Modality, Collaborative Filtering, Multimedia Recommendation

ACM Reference Format:
Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl. 2024. A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3640457.3688138

1 INTRODUCTION

Recommender systems (RSs) offer powerful solutions for navigating the vast amounts of multimedia content available today, helping users find new and interesting items. Modern RSs can handle diverse media types such as text, audio, image, and video, both as their output, i. e., items to recommend, and as modalities the items in the catalog are composed of. While the vast majority of RSs still build upon collaborative filtering (CF) techniques, research advancements in multimedia content analysis and neural learning-to-rank models have enabled the development of effective content-based RSs [10] (CBRSs). Since CF techniques solely rely on past collective interaction data, these algorithms often struggle to provide accurate recommendations when users or items do not appear in the past interactions, a scenario referred to as *cold start*. On the other hand, *pure* CBRSs are often effective for cold start, but relying solely on content information might limit the performance of such algorithms in warm scenarios. Therefore, the most effective way to provide accurate recommendations both in warm and cold-start scenarios is to simultaneously leverage collaborative data and side information

[*] These authors are listed alphabetically and contributed equally to the paper.

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688138

on items and users. The resulting hybrid architectures[1] therefore naturally belong to the domain of *multimodal learning*, where one or more content modalities have to be used in combination with interaction data. As such, CBRSs are still prone to decreases in performance when one or more input modalities are missing. To address this issue, we propose the use of a multimodal **Si**ngle-**Bra**nch embedding network for **R**ecommendation (SiBraR, pronounced "zebra"). SiBraR leverages a single-branch network architecture coupled with weight sharing to embed multiple modalities. This allows mapping different modalities of the same entity (i. e., a user or an item) to similar positions in the shared embedding space, therefore mitigating scenarios in which one or more modalities are missing. This way of addressing the missing modality scenario naturally results in an improvement in performance in cold-start scenarios with respect to both CF approaches and other CBRSs. The contributions of this paper can be summarized as follows:

- We propose SiBraR, a novel CBRS which leverages multimodal information for effective recommendations in standard as well as cold-start and missing-modality scenarios.
- We perform extensive quantitative experiments to assess the accuracy of SiBraR's recommendations, and compare it to other traditional and state-of-the-art methods.
- We analyze the impact of missing modalities on the performance of SiBraR.
- We investigate the shared embedding space of SiBraR, showing that the model maps different modalities to the same region of the shared embedding space, reducing the modality gap.

In the remained of this paper, we first review work related to ours (Section 2). We then introduce the notation and describe our method (Section 3), as well as the experiment setup (Section 4). We discuss the evaluation results, the impact of missing modalities, and the modality gap (Section 5), as well as limitations and possible future extensions of the current work (Section 6).

2 RELATED WORK

Relevant previous work falls within two strands of research: CBRSs for targeting cold start and multimodal representation learning. As our main focus are cold start and missing modality, we exclude general CBRSs not designed for these scenarios and refer the reader to Musto et al. [37].

2.1 Content-Based Cold-start Recommendation

Recommending items in cold-start scenarios is one of the main challenges of RSs [53, 57]; For a recent literature review we refer the reader to Panda et al. [42]. Recent approaches [5, 23, 28, 40, 65, 68, 74] use meta-learning for cold-start recommendation, which might not tackle scenarios where no interactions at all are available for certain users or items. When certain users or items completely lack user–item interactions, RSs often rely on side information. This is done by either adding a content loss term, or adapting the learning process without any additional loss term [63]; Models of the first type are often referred to as *explicit*, and those in the second as *implicit* [66] .[2] Within explicit methods, Wei et al. [67] combine CF with contrastive losses between two representations of items, one obtained from collaborative signals and the other from content representations; The resulting deep neural network (NN)-based content encoder is therefore trained to obtain representations that are as similar as possible to the representations of the collaborative signals, and that can be used in their absence. Similarly, Li et al. [25] use contrastive learning losses in graph-based RSs. The contrastive loss aims at encoding content and interaction information using as positive pairs the pairs of items often co-occurring for the same users. Their approach shows an improvement in performance in recommendations for e-commerce. Wu et al. [70] propose the use of three loss terms to address cold start with content: reconstruction loss terms on two autoencoder architectures, one for item content and the other for user interaction data, and a CF loss. Wang et al. [66] address the problem that due to the development of content production over time, recently added items without interactions might have a different distribution of content features with respect to "older" items, and might therefore be underrepresented in recommendations. To solve this issue, they use a sample interpolation strategy and loss functions that align the representations of interpolations of items with the interpolation of the item representations. Barkan et al. [3] and Zhu et al. [75] propose the use of mean square error as additional loss function to minimize the distance between CF and content representations. Wang et al. [64] propose the use of an encoder-decoder architecture taking as input the concatenation of multimodal data of both users and items; The autoencoder is optimized to simultaneously reconstruct missing modalities and predict the user–item interactions, such as the ratings. Cao et al. [9] separately train a graph NN to model item content similarities, and a transformer-based RS, and propose the use of Euclidean distance as loss term to align the resulting representations. Similarly, Zhang et al. [73] model item content similarities as item embeddings representing a modality-aware content graph; The resulting embeddings are added to the item embeddings obtained from established CF algorithms and the result is used to obtain the final recommendation score. Gong et al. [17] augment an interaction-based graph by constructing representations of cold items based solely on content. Pulis et al. [46] propose a query-by-multiple-example [6] music RS based on item representations extracted from the audio signal of music tracks in a way to maximize the similarity of tracks of the same music genre. Shalaby et al. [58] apply transformers [62] to cold-start session-based RSs and include loss terms to predict both the next item and its category. Magron et al. [32] propose two extensions of neural collaborative filtering [20] with the inclusion of a NN that accepts item content as input. In the first variant, which falls into the category of explicit models, the NN is optimized to minimize the Euclidean distance between its output and the item collaborative representation. In the second variant, falling into the category of implicit models, the deep content feature extractor directly predicts the item collaborative embedding. The category of implicit cold-start methods includes the work of Raziperchikolaei et al. [49], who propose the use of item content as input to a NN and to further use the resulting representations as weights in the first layer

[1] We refer to hybrid RSs as CBRSs and to RSs leveraging only content information as *pure* CBRSs.

[2] Although the term implicit is used for both CBRSs strategies and the type of user–item interaction data, its meaning should be clear from the context.

of an encoder for the user interaction profiles. Cai et al. [8] and Behar et al. [4] model collaborative as well as content information on heterogeneous graphs to learn embeddings that are representative of both. R et al. [47] use NNs to obtain latent representations of users, items, and their side information; these are further concatenated and fed to a NN to obtain the score of a user–item pair. The use of concatenation of content and interaction representations is also proposed by Volkovs et al. [63], who further apply dropout as regularization technique to mimic the lack of information during training, showing that this leads to a better performance in cold-start scenarios. Gong et al. [16] leverage large-language models (LLMs) to obtain domain-invariant item representations, demonstrating their effectiveness in cold-start scenarios, while Sanner et al. [56] leverage pre-trained LLMs to provide recommendations from both item-based and language-based preferences, demonstrating their better performance w. r. t. standard CF in the near cold start (i. e., when some user or items have few interactions). The novelty of our work relies on the use of a single-branch network to encode multimodal content information as well as collaborative data. Since our proposed SiBraR can be leveraged both with and without the use of a content loss, it does not immediately fall into either categories of explicit or implicit CBRSs, hence establishing a new paradigm.

2.2 Multimodal Representation Learning

Multimodal learning leverages information from multiple modalities, such as audio, image, or text, to improve the performance on various machine learning (ML) tasks, such as classification, retrieval, or verification [2, 71]. Although multimodal tasks differ, the way they are usually addressed is very similar and relies on learning joint representations from multiple modalities. Several multimodal methods explored the use of NNs to map multimodal information to joint representations. For example, multi-branch NNs use separate independent and modality-specific NNs to map each modality to a joint embedding space [1, 38, 39, 50, 54]. Recently, multi-branch architectures have also been extended with the use of transformers [30, 59, 62, 71]. Such multi-branch methods have achieved remarkable performance using modality-complete information, i. e., when all modalities are available for training and evaluation. However, they suffer from performance deterioration if a modality is missing either during training or evaluation [24, 27, 31]. Recently, single-branch models [55], i. e., models that share the same embedding NN across multiple modalities, have shown promising results in multimodal learning. Although effective for other multimodal ML tasks, these models have never been translated to the domain of recommendation. Our work fills this gap in the current status of research by proposing SiBraR, a novel multimodal RS based on a single-branch architecture. Furthermore, we analyze the effectiveness of single-branch architectures in missing modality scenarios, which are related to cold start in recommendation.

3 METHODOLOGY

In this section, we introduce our multimodal **Si**ngle-**Bra**nch embedding network for **R**ecommendation (SiBraR). We introduce the notation and mathematical formulation of the problem. We then describe how SiBraR leverages multimodal information for recommendation in warm and cold-start scenarios.

Notation. We denote with $\mathcal{U} = \{u_i\}_{i=1}^{M}$ the set of M users and $\mathcal{I} = \{i_j\}_{j=1}^{N}$ the set of N items and refer to users and items as *entities.* We consider the scenario of implicit feedback and represent the user–item interactions as a binary matrix in $R \in \mathbb{R}^{M \times N}$ with nonzero entries $R_{ij} = 1$ for a positive interaction of user u_i with item i_j. We refer to rows of R as user profiles $\mathbf{u}_i \in \mathbb{R}^N$; Profiles store information on the items with which user u_i interacted. Analogously, we refer to columns of R as item profiles $\mathbf{i}_j \in \mathbb{R}^M$. Cold-start users (items) are those without interactions, i. e., with an empty profile. Each user u_i is represented by a set of *modalities* $\mathcal{M}_i^{\text{user}}$, consisting of *feature vectors* representing available side information, e. g., gender and country, and their profile if not empty. Empty profiles ($\mathbf{u}_i = \mathbf{0}$) of cold-start users are not included in the set of modalities available for the user and are treated as missing modalities. Analogously, each item i_j is represented by a set of item modalities $\mathcal{M}_j^{\text{item}}$, consisting of feature vectors representing the item's available side information, e. g., image or text data, and their profile if not empty.

SiBraR. The proposed model outputs a recommendation score $\hat{y}_{ij}$ for a given user–item pair (u_i, i_j), by assigning an embedding vector $\mathbf{e}_i$ to user u_i and an embedding vector $\mathbf{e}_j$ to item i_j, and by computing their scalar product $\hat{y}_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$. The ordered list of the top k items that were not already interacted with by user u_i (i. e., such that $R_{ij} = 0$) constitutes the recommendations for user u_i. Based on the assumption that different modalities of the same entity contain similar semantic representations, SiBraR aims at constructing embedding vectors for accurate recommendations by taking any available modality as input and projecting them into the same shared space. Therefore, SiBraR leverages weight-sharing and uses the same deep NN g to embed different modalities. The network g is optimized to provide accurate recommendations with any of the modalities as input. This encourages the model to encode modalities of the same entity to similar embeddings; For instance, for the same user u_i: $g(m_1^i) \simeq g(m_2^i)\ \forall m_1, m_2 \in \mathcal{M}_i^{\text{user}}$.

We now describe our proposed SiBraR architecture in its variant leveraging multimodal item side information. We denote this variant by Item-SiBraR. Item-SiBraR is designed to tackle scenarios of item cold start and missing item modality, while the SiBraR variant leveraging multimodal user side information User-SiBraR is designed for user cold start and missing user modality.

Item-SiBraR. Figure 1 sketches the architecture and training strategy of Item-SiBraR, while Algorithm 1 provides the pseudocode for the computation of the batch loss used for its training. In the case of Item-SiBraR, we denote the item embedding function with g and the user embedding function with h. For Item-SiBraR, the item embedding function g is the core component, and is described in detail below, while for the user embedding function h we consider either an embedding lookup table, where for each user a unique embedding is randomly initialized and optimized during training, or an architecture similar to DeepMF [72], employing deep NNs

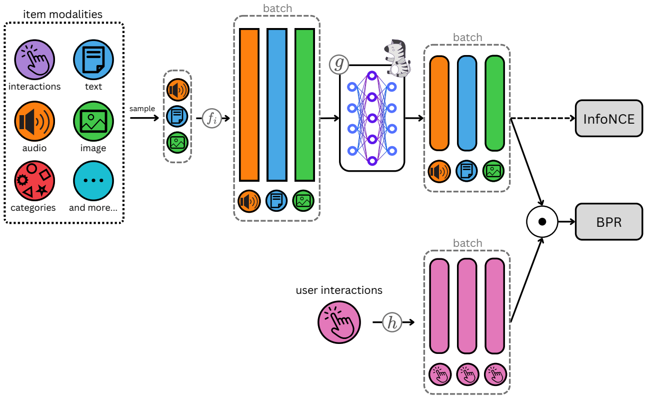

Figure 1: `Item-SiBraR` model and training procedure. The `SiBraR` network represents the single-branch encoding network g shared across modalities. For each user–item interaction pair (u_i, i_j) in the training set, the recommendation loss $\mathcal{L}_{\text{BPR}}$ is computed between positive and negative items. The contrastive loss $\mathcal{L}_{\text{SInfoNCE}}$ is computed for two item modalities and for the set of items consisting of positive item and set of negatives.

to embed the users by taking their profile as input.[3] During training, `Item-SiBraR` samples a set $\mathcal{N}_{ij}$ of $n_{\text{neg}} = |\mathcal{N}_{ij}|$ negative items uniformly at random. For each user–item pair, the user embedding is obtained as $\mathbf{e_i} = h(u_i)$. The model then samples n_{mod} modalities uniformly at random from the set of item modalities $\mathcal{M}^{\text{item}}$. The positive item and each of the negative items are embedded as follows. First, each sampled modality is projected to the input dimension of the shared network g with a shallow network f_i consisting of a linear layer and an activation function. The resulting vectors are then passed through the single-branch network g, i. e., the same network is used irrespective of the modality. The resulting vectors are then averaged to compute the embedding of the item to be used to compute the recommendation score:

$$\mathbf{e}_j = \frac{1}{n_{\text{mod}}}\left(g(m_1) + \cdots + g(m_{n_{\text{mod}}})\right) \qquad (1)$$

The user and item embeddings, $\mathbf{e_i}$ and $\mathbf{e_j}$, respectively, are used to compute the recommendation scores by taking their scalar product. This is done for both the positive item i_j, for which $\hat{y}_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$, and for the negative items $i_k \in \mathcal{N}_{ij}$, for which $\hat{y}_{ik} = \mathbf{e}_i \cdot \mathbf{e}_k$. Based on the scores, the recommendation loss for Bayesian Personalized Ranking (BPR) [52] is computed. Therefore, for a single user–item interaction pair (u_i, i_j), the recommendation loss is given by

$$\mathcal{L}^{(u_i,i_j)}_{\text{BPR}} = \sum_{k\in\mathcal{N}_{ij}} \ln \sigma(\hat{y}_{ij} - \hat{y}_{ik}). \qquad (2)$$

While sharing the embedding network has been proven effective to combine modalities in multimodal ML [55], we also consider the use of a contrastive loss to further align the embeddings from different modalities. When the contrastive loss is not applied, n_{mod} is set to 1 and only one modality is sampled for each user–item interaction pair (u_i, i_j). As the modality sampling is done per interaction pair, over the course of training, all available modalities are used. [4] When the contrastive component is applied, n_{mod} is set to 2 and in addition to the recommendation loss, `Item-SiBraR` models apply the symmetric Info Noise Contrastive Estimation loss [48, 60] $\mathcal{L}_{\text{SInfoNCE}}$ between the two item modality embeddings. To this purpose, we contrast the two modalities of the positive item i_j with the two modalities of the negative samples in $\mathcal{N}_{ij}$. The loss between modalities m_1 and m_2 is therefore computed as $\mathcal{L}_{\text{SInfoNCE}} = \mathcal{L}^{12}_{\text{InfoNCE}} + \mathcal{L}^{21}_{\text{InfoNCE}}$, where $\mathcal{L}^{12}_{\text{InfoNCE}}$ is the InfoNCE [60] contrastive loss between modalities m_1 and m_2:

$$\mathcal{L}^{12}_{\text{InfoNCE}} = -\sum_{(u_i,i_j)} \ln \frac{e^{(g(m_1^j)\cdot g(m_2^j))/\tau}}{\sum_{k\in\{j\}\cup\mathcal{N}_{ij}} e^{(g(m_1^j)\cdot g(m_2^k))/\tau}}, \qquad (3)$$

and τ represents the temperature parameter, considered as a hyperparameter as described in Section 4. Intuitively, $\mathcal{L}_{\text{SInfoNCE}}$ aims at penalizing highly dissimilar embeddings of the same items, as well as similar embeddings of different items.

During inference, for each item $i_j \in \mathcal{I}$, `Item-SiBraR` uses all its available modalities $\mathcal{M}^{\text{item}}_j$, encoding them with the single-branch network g and taking the average of the resulting encoding:

$$\mathbf{e}_j = \frac{1}{\left|\mathcal{M}^{\text{item}}_j\right|}\left(g(m_1) + \cdots + g\left(m_{\left|\mathcal{M}^{\text{item}}_j\right|}\right)\right). \qquad (4)$$

Therefore, the embeddings of cold-start items are obtained from their content modalities only, since their profile is treated as missing modality.

`User-SiBraR`. The training and inference for `User-SiBraR` follows a very similar approach, employing an embedding layer or a simple NN for the item, and a single-branch network shared across modalities for the user.

4 EXPERIMENTAL SETUP

In this section, we describe the experimental setup for our experiments, including the datasets, the baselines used for comparison, the evaluation protocol, the training procedure and hyperparameter tuning. We also provide the code used to carry out our experiments.[5]

4.1 Datasets

We carry out experiments on datasets from three domains: music, movie, and e-commerce. The datasets were selected by considering their popularity for benchmarking RSs, their number of content features, and the multimodality of side information available. We consider an implicit feedback scenario with a binary user–item interaction matrix R: $R_{ij} = 1$ if user u_i interacted with item i_j and $R_{ij} = 0$ otherwise. Table 1 summarizes the characteristics of the datasets after pre-processing (see below) as well as the modalities and dimensionalities of the side information used in our experiments.

[3] In preliminary experiments we compared versions of SiBraR with different underlying RS architectures, including variants similar to DeepMF, i. e., applying a NN to the user and item interaction profile, or similar to MF, i. e., applying an embedding layer on the IDs of users and items. We selected the architecture reaching the highest accuracy of recommendations on the validation set, for the subsequent experiments.

[4] Note that different user–item interaction pairs may correspond to different sampled modalities.

[5] https://github.com/hcai-mms/SiBraR—Single-Branch-Recommender

Algorithm 1: Pseudocode for computing the batch loss function in `Item-SiBraR` models

```
Input  : $b \subset \{(u_i, i_j) | R_{ij} = 1\}$: Batch of user–item training interactions
         $n_{\text{neg}}$: Number of negatives per interaction
         $\mathcal{M}^{\text{item}}$: Item training modalities
         ctrEmbs: Whether to use contrastive loss or not
         $\lambda, \tau$: Weight and temperature of contrastive loss
Output : $\mathcal{L}$: Loss for the batch
$\mathcal{L}_{\text{BPR}} \leftarrow 0$                                   // Reset BPR Loss
$\mathcal{L}_{\text{SInfoNCE}} \leftarrow 0$                              // Reset contrastive loss
if ctrEmbs then
    $n_{\text{mod}} \leftarrow 2$
else
    $n_{\text{mod}} \leftarrow 1$
for $(u_i, i_j) \in b$ do
    $\mathbf{e_i} \leftarrow h(u_i)$                                      // Embed user $u_i$
    $\mathcal{N}_{ij} \leftarrow$ negSampling($u_i, i_j, R$)              // Sample negatives
    $\mathcal{M}_{ij} \leftarrow$ randomSampling($\mathcal{M}^{item}$)    // Sample $n_{\text{mod}}$ training modalities
    $\mathbf{e_j} \leftarrow \frac{1}{n_{\text{mod}}} \sum_{m \in \mathcal{M}_{ij}} g(m_j)$   // Embed positive item $i_j$ with single-branch $g$ on modalities $\mathcal{M}_{ij}$
    $\hat{y}_{ij} \leftarrow \mathbf{e_i} \cdot \mathbf{e_j}$            // Compute positive logit
    for $i_k \in \mathcal{N}_{ij}$ do
        $\mathbf{e_k} \leftarrow \frac{1}{n_{\text{mod}}} \sum_{m \in \mathcal{M}_{ij}} g(m_k)$   // Embed item $i_k$ with single-branch $g$ on modalities $\mathcal{M}_{ij}$
        $\hat{y}_{ik} \leftarrow \mathbf{e_i} \cdot \mathbf{e_k}$        // Compute negative logit
        $\mathcal{L}_{\text{BPR}} \leftarrow \mathcal{L}_{\text{BPR}} + \ln \sigma(\hat{y}_{ij} - \hat{y}_{ik})$   // Accrete BPR loss
    if ctrEmbs then
        $\mathcal{L}_{\text{SInfoNCE}} \leftarrow \mathcal{L}_{\text{SInfoNCE}} - \ln \frac{e^{(g(m_1^j) \cdot g(m_2^j))/\tau}}{\sum_{k \in \{j\} \cup \mathcal{N}_{ij}} e^{(g(m_1^j) \cdot g(m_2^k))/\tau}} - \ln \frac{e^{(g(m_2^j) \cdot g(m_1^j))/\tau}}{\sum_{k \in \{j\} \cup \mathcal{N}_{ij}} e^{(g(m_2^j) \cdot g(m_1^k))/\tau}}$   // Accrete contrastive loss
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{BPR}} + \lambda \mathcal{L}_{\text{SInfoNCE}}$
```

					User Features				Item Features			
	# users	# items	# inter.	sparsity	Age (**d**)	Gender (**c**)	Country (**c**)	Occupation (**c**)	Text (**v**)	Audio (**v**)	Image (**v**)	Genre (**m**)
Onion [36]	5,192	13,610	407,653	0.99423	–	2	152	–	768	50; 100; 4,800	4,096	853
ML-1M [18]	5,816	3,299	813,792	0.95759	7	2	–	21	768	–	–	18
Amazon [21]	11,454	4,177	87,098	0.99818	–	–	–	–	768; 768	–	2,048	–

Table 1: Summary of datasets. Left side: Characteristics after pre-processing. Right side: Representation type (categorical, discrete, vector, multilabel), dimensionality (for v) or number of options (for c, d, and m) of each content feature available. Dimensionalities separated by a colon indicate that more than one feature is available, – denotes features that are not available or have been neglected (see Section 4.1).

Music4All-Onion (Onion) [6] [36] This large-scale dataset provides user–item interactions, side information on the users and on the items, including several representations of the items content, related to the audio signal, to the lyrics, and to the videoclips. We consider gender and country as user side information. We include representations of the audio signals in terms of i-vectors [13] of dimension 256 and visual representations of the YouTube videoclips of the music tracks obtained with a pre-trained instance of ResNet [19], both as provided by Moscati et al. [36]. Additionally, we extend the set of features provided by Moscati et al. with more recent NN architectures from the domains of music information retrieval and natural language processing. More specifically, we encode the 30*s* audio snippets provided by the Music4All dataset [43] with the NNs MusicNN[7] [44, 45] and Jukebox[8] [12], and the lyrics with the `all-mpnet-v2`[9] pre-trained instance of the SentenceTransformer model [51] provided by HuggingFace [69], after applying the same lyrics-specific pre-processing described by Moscati et al. [36]. We restrict to users and items for which all side and content information is available and to users of age between 10 and 80. Following standard practice in the music RS domain [34, 35], we restrict to the set of listening events over one year, selecting 2018, as this year was not affected by the outbreak of Covid-19, which impacted the listening behavior of users of music streaming platforms [14, 29]. Finally, as commonly done in domains of implicit feedback such

[6] https://zenodo.org/records/6609677

[7] https://github.com/jordipons/musicnn/tree/master

[8] https://github.com/openai/jukebox/

[9] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

as music recommendation [34, 35], we convert the interactions to binary implicit feedback with a threshold of 2 on the interaction counts (reducing false-positive interactions) and perform 5-core filtering for users and items.

MovieLens 1M (ML-1M) [10] [18] This dataset consists of movie ratings collected from the MovieLens website.[11] In addition to the user–item interactions, the dataset provides side information on users (age, gender, and occupation) and items (movie genre). Additionally, we crawl Wikipedia for movie plots and encode the product titles with the `all-mpnet-v2`[12] pre-trained instance of the SentenceTransformer model [51] provided by HuggingFace [69]. We restrict to users and items for which all features are available and to users of age between 10 and 80 years. We apply 5-core filtering and convert the ratings to binary implicit feedback by ignoring the actual values of the ratings, keeping only the interaction information.

Amazon Video Games (Amazon) [13] [21] Amazon Reviews'23 is a large-scale dataset consisting of Amazon Reviews posted from May 1996 to September 2023; This recently released dataset extends the well-established Amazon Reviews'18 dataset [41]. Amazon includes several item features such as title, description, price, and one or more links to images of the products. After considering several categories, we chose to restrict to the category *Video Games*, as this resulted in sets of users, items, and user–item interactions that are large enough to mimic a real-world scenario, and provides characteristics that are different from the two other datasets considered. We restrict to reviews posted between January 2016 and December 2019. [14]We only consider items that include titles, descriptions, and at least one image of the product in high resolution. We convert the ratings to binary implicit feedback with a threshold of 1, i. e., ignoring the actual ratings. We apply 5-core filtering, restricting to users that rated at least 5 different items and to items that have been rated by at least 5 different users. We encode the product titles and descriptions with the `all-mpnet-v2`[15] pre-trained instance of the SentenceTransformer model [51] provided by HuggingFace [69]. We download the first high-resolution image of the product. Following the same approach as Saaed et al. in [55], we resize the image such that the smaller edge has size 256, crop it at the center to a square of size 224 pixels, and normalize each channel to the mean and standard deviation of ImageNet [11]. We encode the resulting image with the pre-trained instance of `ResNet-101` [19] provided by TorchVision[16] [33].

4.2 Dataset Splitting

To evaluate the RSs in warm as well as user and item cold-start scenarios, we split the datasets in three different ways.

Warm split. For every user, we split the data into a training, a validation, and a test set, respectively consisting of 80%, 10%, and 10% of the number of their interactions, randomly selected.

User Cold start. We split the users into disjoint sets of training, validation, and test users, respectively consisting of 80%, 10%, and 10% of the number of users. All interactions of train users are considered during training, and are neglected during validation and testing. All interactions of validation (test) users are considered during validation (testing), and are neglected during training and testing (validation). Since test users are not seen during training, this scenario is apt to measure the performance of RSs in the user cold-start scenario.

Item Cold start. Similar to user cold start, we split the items into disjoint sets of training, validation, and test items, respectively consisting of 80%, 10%, and 10% of the number of items. All interactions of train items are considered during training, and are neglected during validation and testing. All interactions of validation (test) items are considered during validation (testing), and are neglected during training and testing (validation). Since test items are not seen during training, this scenario is apt to measure the performance of RSs in the item cold-start scenario.

4.3 Baselines

We compare the performance of `SiBraR` with two CBRSs for cold-start recommendation. We also include two CF models leveraging only collaborative data. For CBRS we evaluate `SiBraR` against CLCRec [67] and DropoutNet [63], selected since they are often used as baselines for CBRS addressing cold-start scenarios, and since they cover both explicit (CLCRec) and implicit (DropoutNet) approaches.[17] For CF, we include matrix factorization (MF) with BPR [52] and Deep Matrix Factorization (DeepMF) [72]. We select MF since it is the simplest yet effective latent-representation-based model for recommendation, and DeepMF due to its effectiveness in leveraging a two-tower deep NN for recommendation by taking the user and item profiles as inputs. Additionally, we include two naive baselines: Rand, which randomly selects the items to recommend to each user, and Pop, which recommends the same most popular items to all users, measuring popularity in terms of number of interactions in the training set.

4.4 Metrics

We evaluate the performance of the algorithms on lists of $k = 10$ recommended items, as common in the recommendation domain. We measure accuracy in terms of Normalized Discounted Cumulative Gain (nDCG)[18] averaged over users in the test set. We test the significance of the best performing algorithm with respect to the others using multiple paired t-tests using Bonferroni correction to account for the multiple comparisons. We consider an improvement significant if $p < 0.05$. During evaluation, for each user in the validation (test) set, we rank all items in the validation (test) set and compute the metrics on the resulting ranking.

4.5 Hyperparameter Tuning

We carry out an extensive hyperparameter optimization to rigorously evaluate the effectiveness of `SiBraR` and of the baselines.

[10] https://grouplens.org/datasets/movielens/1m/

[11] https://movielens.org/

[12] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[13] https://amazon-reviews-2023.github.io/index.html

[14] Although the Covid-19 virus was identified in December 2019, the virus spread to other areas of Asia, and then worldwide in early 2020. Therefore, this selection also excludes interactions affected by the outbreak of Covid-19.

[15] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[16] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet101.html

[17] See Section 2 for a detailed description of explicit and implicit CBRSs for cold start.

[18] We refer the reader to the Additional Material available in the code repository for an evaluation on mean average precision, recall, and catalog coverage.

For all models, we tune the learning rate, the weight decay, the embedding dimension, and – whenever the model employs NNs – the number of layers and the number of nodes of each layer. For CBRSs, we treat the training modalities as hyperparameters.[19] For DropoutNet and SiBraR, which support the use of one or more modalities in addition to interaction data, we consider all possible modality combinations of length 1 up to the number of available modalities. For CLCRec, which only supports the use of one modality at a time in addition to interaction data, we consider all of them separately. For a fair comparison of CLCRec with DropoutNet and SiBraR, which support the use of more than one modality, we separately select the best DropoutNet and SiBraR models leveraging only one, or one or more modalities. The variants leveraging one modality are denoted with DropoutNet$_{\text{one}}$ and SiBraR$_{\text{one}}$, those leveraging one or more with DropoutNet$_{\text{best}}$ and SiBraR$_{\text{best}}$. The algorithms reaching the best evaluation metric on the validation set used for the hyperparameter optimization are again evaluated on the test set. The nDCG@10 computed on the test set is the one reported for comparison in Section 5.[20] For CLCRec and SiBraR we further tune the weight and the temperature of the regularization loss.[21] The hyperparameters are selected through Bayesian optimization, relying on the module Sweep of Weights and Biases platform.[22] For an overview of the hyperparameters and their value ranges, we refer the reader to the configuration files available in the repository.[23] We set the number of negative samples required during training to 10. We use nDCG@10 on the validation set as validation metric and run all experiments for a maximum of 50 epochs, selecting the model weights that reach the maximum validation metric. We arrest training at an earlier epoch if no improvement on the validation metric is observed for 5 consecutive epochs.

5 RESULTS

In this section, we report the performance of SiBraR and of the baselines in standard and cold-start scenarios. We then analyze the effect of the single-branch in mapping multiple modalities to the shared embedding space.

5.1 Performance Comparison

Table 2 shows the nDCG@10 results[24] of the algorithms on three datasets. For Onion and ML-1M, where both user and item information is available, we show the results on the three evaluation scenarios described in Section 4.4. Since user information is not available for Amazon, we restrict ourselves to the random and item cold start for this dataset. Each row refers to a different algorithm. Solid lines divide the algorithms in the three categories described in Section 4.4. For SiBraR (DropoutNet), which allows leveraging more than one modality, we include both SiBraR$_{\text{one}}$ (DropoutNet$_{\text{one}}$) and SiBraR$_{\text{best}}$ (DropoutNet$_{\text{best}}$). The sign † indicates significant improvement of the best performing RS over all other RS (paired t-tests with $p < 0.05$ considering Bonferroni correction to account for multiple comparisons).

Warm split. In the random split scenario, corresponding to the case in which all users and items in the test set are warm, overall MF, CLCRec, and our SiBraR are among the best-performing algorithms, with the best result varying across recommendation domains. Specifically, CLCRec reaches the highest nDCG@10 on ML-1M, slightly outperforming our SiBraR$_{\text{one}}$ and MF. On Onion, both versions of SiBraR outperform MF and CLCRec. Finally, on the Amazon dataset, MF outperforms SiBraR and CLCRec.

Cold start. The limitation of CF techniques in cold-start scenarios is evident from the fact that they are outperformed by CBRSs. This is more evident in the item than in the user cold-start scenario. We attribute the larger improvement of CBRSs with respect to CF in the item cold start to the larger and more informative set of content modalities available for items; For cold-start users, where only demographic information is available, the difference between CBRSs and CFs is therefore less pronounced. Our SiBraR outperforms all other algorithms (CF and CBRSs) on all datasets and cold-start scenarios. The difference in performance is substantial when leveraging all the available modalities (SiBraR$_{\text{best}}$). This shows that by sharing the weights of the single-branch encoder, SiBraR is effectively leveraging the multimodality of the information available during the evaluation phase.

Missing modalities. As discussed in Section 3, the strength of SiBraR, which also makes it effective in cold-start scenarios, is its ability to tackle missing modality scenarios. It relies on the information extracted from the available modalities by means of a single branch, which is trained to encode information on the information shared by all modalities, including the missing one. Figure 2 highlights this by comparing the performance of SiBraR$_{\text{best}}$ on Onion[25] on the random split, corresponding to a warm-start scenario. The values and bars report the test nDCG@10 reached by a single SiBraR instance trained with all modalities, and evaluated leveraging different combinations of modalities. Since the model has been trained with five different modalities (interactions, audio, text, image, and item genre), 31 different modality combinations are evaluated. Each modality corresponds to a different color. If it is used at evaluation, this is reflected by a colored box below the bar plot, while if it is not used, the box is left blank. The grey dashed lines show the nDCG@10 reached by the comparison algorithms (MF, CLCRec, DeepMF, and DropoutNet). Overall, the performance of SiBraR increases with an increasing number of modalities. Moreover, considering the ranking of combinations of the same number of modalities, image is the most informative side-information modality, followed by text, audio, and genre. By only leveraging interaction data

[19] We would like to emphasize that we also investigated combinations of training modalities that did not include the interaction profile as input to the single-branch network. Therefore, interactions are not used as a proxy or anchor for other modalities, as done e. g., in ImageBind [15]. This equal treatment of the modalities in the training phases of our experiments motivates our analogy between cold-start and missing-modality scenarios.

[20] While we select DropoutNet$_{\text{best}}$ and DropoutNet$_{\text{one}}$ from the sets that also comprise DropoutNet$_{\text{one}}$ and SiBraR$_{\text{one}}$, since the optimization and model selection is done on the validation set and the final evaluation on the test set, it may occur that e. g., SiBraR$_{\text{one}}$ is better than SiBraR$_{\text{best}}$ on the test set.

[21] We also carried out a comparison of the performance of SiBraR with and without contrastive loss and an ablation study of the impact of the weight and temperature on recommendation accuracy. Due to the page limit, we refer the reader to the repository for the results of these analyses.

[22] https://wandb.ai/

[23] https://github.com/hcai-mms/SiBraR—Single-Branch-Recommender

[24] We refer the reader to the auxiliary material for similar tables reporting precision@10, recall@10, AP@10 and coverage@10.

[25] We carry out this analysis on Onion since it is the dataset with the highest number of modalities available.

		Warm			Cold				
						Item		User	
		ML-1M	Onion	Amazon	ML-1M	Onion	Amazon	ML-1M	Onion
Base	Rand	0.0052	0.0009	0.0013	0.0503	0.0081	0.0110	0.0440	0.0087
	Pop	0.1268	0.0182	0.0255	0.0087	0.0063	0.0210	0.4583	0.0994
CF	MF [52]	0.2589	0.1438	**0.0818**†	0.0300	0.0055	0.0125	0.4613	0.0920
	DeepMF [72]	0.1165	0.1297	0.0556	0.0097	0.0197	0.0210	0.4124	0.0575
CBRS	DropoutNet$_{\text{one}}$ [63]	0.2427	0.0749	0.0374	0.2032	0.1142	0.1033	0.4584	0.0947
	DropoutNet$_{\text{best}}$ [63]	0.2427	0.0935	0.0445	0.2032	0.1689	0.1033	0.4584	0.1086
	CLCRec [67]	**0.2680**†	0.1378	0.0569	0.1762	0.1503	0.1348	0.4616	0.0989
	SiBraR$_{\text{one}}$	0.2593	0.1466	0.0728	0.2592	0.1457	**0.1479**†	**0.4673**	0.0918
	SiBraR$_{\text{best}}$	0.2561	**0.1616**†	0.0728	**0.2994**†	**0.1982**†	**0.1479**†	0.4659	**0.1094**

Table 2: Evaluation results w. r. t. nDCG@10. Bold indicates the best performing RS, † indicates significant improvement over all others (paired t-tests with $p < 0.05$ considering Bonferroni correction).

at evaluation provides better recommendations than leveraging any side-information modality or combinations of two of them. In comparison to the other models, already leveraging only text or image content, SiBraR outperforms DropoutNet, while either the simultaneous use of interaction and one content modality, or 3 or more content modalities are required to outperform DeepMF and CLCRec. Finally, leveraging at least 3 modalities and including the interaction data allows SiBraR to outperform MF.

5.2 Modality Gap

In this section we analyze to which extent SiBraR is able to map different modalities to the same region of the embedding space, hence filling the modality gap often displayed by multimodal models [22, 26]. We consider SiBraR$_{\text{best}}$ trained on the Onion random split, which during training relies on all available modalities. Figure 3 shows the embeddings of the five item modalities (text, audio, image, genres, and interactions) after training. The left plot shows the embeddings used as input to the single-branch network g, while the right one shows the output of the single-branch network. To visualize the high-dimensional vectors in two dimensions, we first select the first 10 principal components through principal component analysis. We then apply T-distributed Stochastic Neighbor Embedding (t-SNE) [61][26] to further reduce the dimensions to 2. To improve clarity, we only show the embeddings of a subset of all items, consisting of 3,000 items sampled uniformly at random. The left plot of Figure 3 shows that before the single-branch network, embeddings of different modalities cover different regions of the embedding space. On the contrary, the right plot shows that the single-branch network maps all the modalities to the same region of the embedding space. This is an indication of SiBraR's ability to extract representations that are similar – and therefore similarly effective for recommendation – from any of the item modalities.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose SiBraR, a novel multimodal RS that leverages a single-branch architecture to encode multimodal user and item information, including collaborative data. The single-branch design allows SiBraR to provide accurate recommendations from any modality. As a result, SiBraR's recommendations are accurate also in scenarios of missing modality, including cold start. We show through extensive quantitative experiments that SiBraR significantly outperforms CF as well as CBRSs in terms of accuracy of recommendations in item cold-start scenarios, and that it is competitive with both CF and CBRSs in warm-start as well as user cold-start scenarios. Furthermore, we analyze the impact of missing modalities on the performance of SiBraR, showing that the proposed model reaches its best performance when all modalities are leveraged, but it is still able to outperform CF and other CBRSs models when input modalities are missing. Moreover, we analyze SiBraR's embedding space shared by the multiple modalities and show that as a result of its design, SiBraR is able to reduce the modality gap. SiBraR's ability to combine multimodal representations relies on the use of the single-branch, which does not exclude the use of a contrastive loss. The impact of the contrastive loss on SiBraR can be analyzed by means of an ablation study; The performance of SiBraR can also be compared with that of multi-branch architectures with contrastive loss. We focused on a version of SiBraR with an underlying RS architecture similar to DeepMF and envision as future work variants of SiBraR based on other core RS. A deeper analysis on the modality gap of SiBraR could analyze the distance in the embedding space between multimodal embeddings of a same user or item, as well as the correlation between modality gap and model performance. Our analysis of missing-modality scenarios was limited to the case in which all modalities are available during training but not during inference, i. e., we did not consider missing modalities at training time. Additionally, our work did not consider the scalability of SiBraR in function of the dimensionality of the input modalities. Although this scenario is partially addressed in the analysis summarized in Figure 2, we did not evaluate the performance of RSs in scenarios where all items in the validation and test sets have no information regarding one of the modalities related to

[26] We use t-SNE with the default parameters reported by the library scikit-learn [7] and refer the reader to its documentation for details, (https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html).

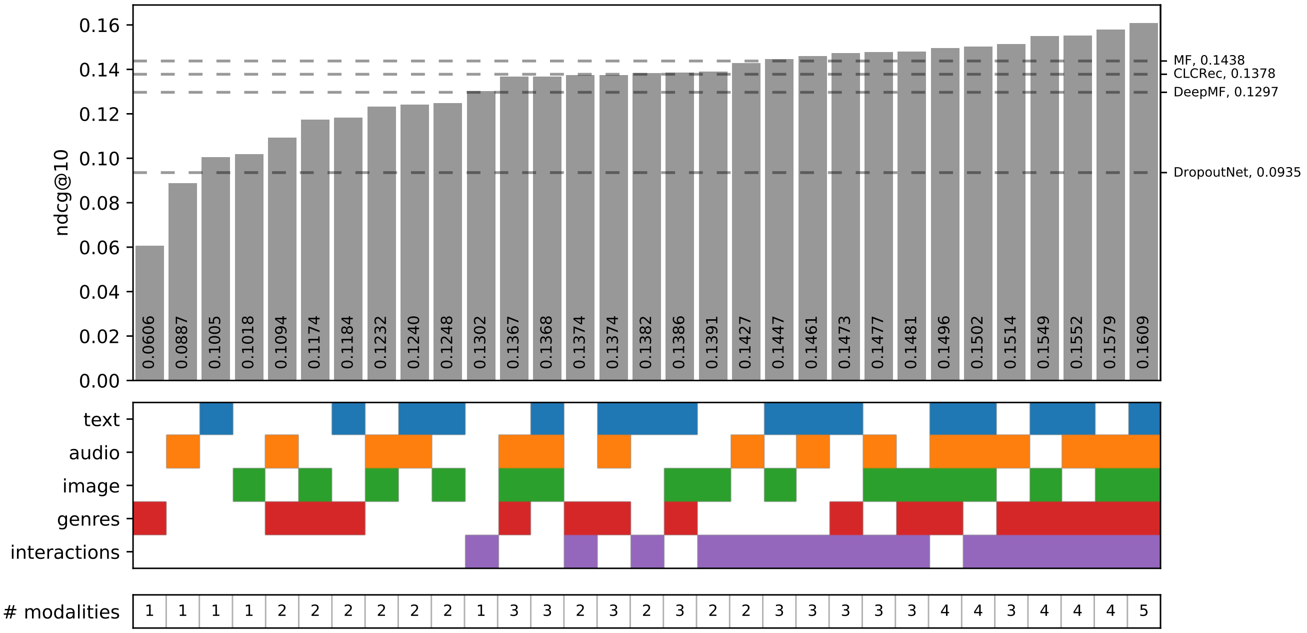

Figure 2: Performance of SiBraR on the test set of warm-start Onion, based on a varying set of modalities used. The bottom integers show the number of modalities. If a modality is used, its block is filled with the corresponding color in the central plot. The bar plot shows SiBraR's performance in terms of nDCG@10 for each set of modalities. The gray dashed horizontal lines show the nDCG@10 of CF and CBRS algorithms.

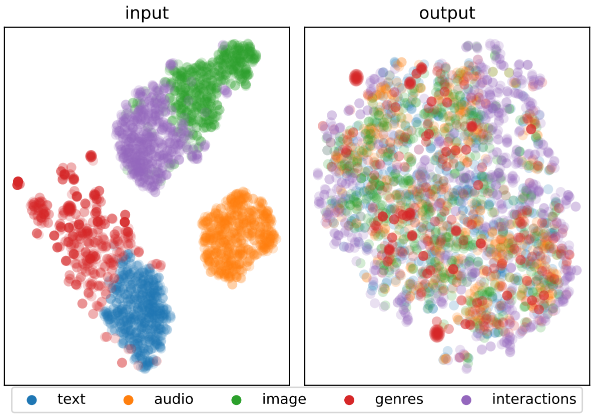

Figure 3: t-SNE projected embeddings before and after SiBraR. While different modalities can be differentiated at SiBraR input (left), modalities overlap substantially in the shared embedding space after applying SiBraR (right).

side information. Finally, reducing the amount of collaborative data while increasing content information might substantially impact the performance of SiBraR in terms of beyond-accuracy metrics. We leave these extensions of the current work for future research.

ACKNOWLEDGMENTS

This research was funded in whole or in part by the Austrian Science Fund (FWF) https://doi.org/10.55776/P33526, https://doi.org/10.55776/DFH23, https://doi.org/10.55776/COE12, https://doi.org/10.55776/P36413. For open access purposes, the authors have applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

Christian Gahnör and Marta Moscati would like to thank Alessandro Melchiorre for his invaluable work in developing the initial version of the framework used to benchmark SiBraR with other RSs.

REFERENCES

[1] Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding Intra-text Representations with Visual Context for Multimodal named Entity Recognition. In *Proc. of ICDAR* (Sydney, Australia). 337–342.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.

[3] Oren Barkan, Noam Koenigstein, Eylon Yogev, and Ori Katz. 2019. CB2CF: a Neural Multiview Content-to-Collaborative Filtering Model for Completely Cold Item Recommendations. In *Proc. of ACM RecSys* (Copenhagen, Denmark). 228–236.

[4] Eric Behar, Julien Romero, Amel Bouzeghoub, and Katarzyna Wegrzyn-Wolska. 2023. Tackling Cold Start for Job Recommendation with Heterogeneous Graphs. In *Proc. of HR Workshop at ACM RecSys* (Singapore, Singapore).

[5] Homanga Bharadhwaj. 2019. Meta-Learning for User Cold-Start Recommendation. In *Proc. of IJCNN* (Budapest, Hungary). 1–8.

[6] Dmitry Bogdanov, Martin Haro, Ferdinand Fuhrmann, Emilia Gómez, and Perfecto Herrera. 2014. Content-based Music Recommendation Based on User Preference Examples. In *Proc. of WOMRAD Workshop at ACM RecSys* (Foster City, CA, USA). 4–9.

[7] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API Design for Machine Learning Software: Experiences from the scikit-learn Project. In *Proc. of LML Workshop at ECML PKDD* (Prague, Czech Republic). 108–122.

[8] Desheng Cai, Shengsheng Qian, Quan Fang, Jun Hu, and Changsheng Xu. 2023. User Cold-Start Recommendation via Inductive Heterogeneous Graph Neural Network. *ACM Transactions on Information Systems* 41, 3 (2023).

[9] Yuwei Cao, Liangwei Yang, Chen Wang, Zhiwei Liu, Hao Peng, Chenyu You, and Philip S. Yu. 2023. Multi-task Item-attribute Graph Pre-training for Strict Cold-start Item Recommendation. In *Proc. of ACM RecSys* (Singapore, Singapore). 322–333.

[10] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender Systems Leveraging Multimedia Content. *Comput. Surveys* 53, 5 (2020).

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of IEEE CVPR* (Miami, FL, USA). 248–255.

[12] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. arXiv:2005.00341

[13] Hamid Eghbal-Zadeh, Bernhard Lehner, Markus Schedl, and Gerhard Widmer. 2015. I-Vectors for Timbre-Based Music Similarity and Music Artist Classification. In *Proc. of ISMIR* (Málaga, Spain). 554–560.

[14] Mona Ghaffari, Gohar Feroz Khan, Shivendu Pratap Singh, and Bruce Ferwerda. 2024. The Impact of COVID-19 on Online Music Listening Behaviors in Light of Listeners' Social Interactions. *Multimedia Tools and Applications* 83, 5 (2024), 13197–13239.

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. In *Proc. of IEEE CVPR* (Vancouver, Canada). 15180–15190.

[16] Yuqi Gong, Xichen Ding, Yehui Su, Kaiming Shen, Zhongyi Liu, and Guannan Zhang. 2023. An Unified Search and Recommendation Foundation Model for Cold-Start Scenario. In *Proc. of ACM CIKM* (Birmingham, United Kingdom). 4595–4601.

[17] Zhen Gong, Xin Wu, Lei Chen, Zhenzhe Zheng, Shengjie Wang, Anran Xu, Chong Wang, and Fan Wu. 2023. Full Index Deep Retrieval: End-to-End User and Item Structures for Cold-start and Long-tail Item Recommendation. In *Proc. of ACM RecSys* (Singapore, Singapore). 47–57.

[18] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2015).

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. of IEEE CVPR* (Las Vegas, NV, USA). 770–778.

[20] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proc. of WWW* (Perth, Australia). 173–182.

[21] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. arXiv:2403.03952

[22] Christina Humer, Marc Streit, and Hendrik Strobelt. 2023. *Amumo (Analyze Multi-Modal Models).* https://github.com/ginihumer/Amumo

[23] Minchang Kim, Yongjin Yang, Jung Hyun Ryu, and Taesup Kim. 2023. Meta-Learning with Adaptive Weighted Loss for Imbalanced Cold-Start Recommendation. In *Proc. of ACM CIKM* (Birmingham, United Kingdom). 1077–1086.

[24] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal Prompting with Missing Modalities for Visual Recognition. In *Proc. of IEEE CVPR* (Vancouver, Canada). 14943–14952.

[25] Guohui Li, Li Zou, Zhiying Deng, and Qi Chen. 2024. Neighborhood-Enhanced Multimodal Collaborative Filtering for Item Cold Start Recommendation. In *Proc. of ICASSP* (Seoul, Korea). 7815–7819.

[26] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In *Proc. of NeurIPS* (New Orleans, LA, USA). 17612–17625.

[27] Ronghao Lin and Haifeng Hu. 2023. MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis. *Transactions of the Association for Computational Linguistics* 11 (2023), 1686–1702.

[28] Han Liu, Hongxiang Lin, Xiaotong Zhang, Fenglong Ma, Hongyang Chen, Lei Wang, Hong Yu, and Xianchao Zhang. 2023. Boosting Meta-Learning Cold-Start Recommendation with Graph Neural Network. In *Proc. of ACM CIKM* (Birmingham, United Kingdom). 4105–4109.

[29] Meijun Liu, Eva Zangerle, Xiao Hu, Alessandro Melchiorre, and Markus Schedl. 2020. Pandemics, Music, and Collective Sentiment: Evidence from the Outbreak of Covid-19. In *Proc. of ISMIR* (Virtual). 157–165.

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 32 (2019).

[31] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. SMIL: Multimodal Learning with Severely Missing Modality. In *Proc. of AAAI Conference on Artificial Intelligence* (Vancouver, Canada). 2302–2310.

[32] Paul Magron and Cédric Févotte. 2022. Neural Content-aware Collaborative Filtering for Cold-start Music Recommendation. *Data Mining and Knowledge Discovery* 36, 5 (2022), 1971–2005.

[33] Sébastien Marcel and Yann Rodriguez. 2010. Torchvision the Machine-vision Package of Torch. In *Proc. of ACM Multimedia* (Firenze, Italy). 1485–1488.

[34] Alessandro B. Melchiorre, Navid Rekabsaz, Christian Ganhör, and Markus Schedl. 2022. ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations. In *Proc. of ACM RecSys* (Seattle, WA, USA). 246–256.

[35] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating Gender Fairness of Recommendation Algorithms in the Music Domain. *Information Processing & Management* 58, 5 (2021), 102666.

[36] Marta Moscati, Emilia Parada-Cabaleiro, Yashar Deldjoo, Eva Zangerle, and Markus Schedl. 2022. Music4All-Onion - A Large-Scale Multi-faceted Content-Centric Music Recommendation Dataset. In *Proc. of ACM CIKM* (Atlanta, GA, USA). 4339–4343.

[37] Cataldo Musto, Marco de Gemmis, Pasquale Lops, Fedelucio Narducci, and Giovanni Semeraro. 2022. Semantics and Content-Based Recommendations. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 251–298.

[38] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Learnable PINs: Cross-modal Embeddings for Person Identity. In *Proc. of ECCV* (Munich, Germany). 71–88.

[39] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing Voices and Hearing Faces: Cross-modal Biometric Matching. In *Proc. of IEEE CVPR* (Salt Lake City, UT, USA). 8427–8436.

[40] Krishna Prasad Neupane, Ervine Zheng, and Qi Yu. 2021. MetaEDL: Meta Evidential Learning For Uncertainty-Aware Cold-Start Recommendations. In *Proc. of ICDM* (Auckland, New Zealand). 1258–1263.

[41] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proc. of EMNLP* (Hong Kong, China). 188–197.

[42] Deepak Kumar Panda and Sanjog Ray. 2022. Approaches and Algorithms to Mitigate Cold Start Problems in Recommender Systems: a Systematic Literature Review. *Journal of Intelligent Information Systems* 59, 2 (2022), 341–366.

[43] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Yandre Maldonado e Gomes da Costa, Valéria Delisandra Feltrim, and Marcos Aurélio Domingues. 2020. Music4All: A New Music Database and Its Applications. In *Proc. of IWSSIP* (Niteroi, Rio de Janeiro, Brazil). 399–404.

[44] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. 2018. End-to-end Learning for Music Audio Tagging at Scale. In *Proc. of ISMIR* (Paris, France). 37–644.

[45] Jordi Pons and Xavier Serra. 2019. musicnn: Pre-trained Convolutional Neural Networks for Music Audio Tagging. In *Proc. of ISMIR LBD* (Delft, Netherlands).

[46] Michael Pulis and Josef Bajada. 2021. Siamese Neural Networks for Content-based Cold-Start Music Recommendation.. In *Proc. of ACM RecSys* (Amsterdam, Netherlands). 719–723.

[47] Kiran R, Pradeep Kumar, and Bharat Bhasker. 2020. DNNRec: A Novel Deep Learning Based Hybrid Recommender System. *Expert Systems with Applicationse* 144, C (2020).

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. of ICML* (Virtual). 8748–8763.

[49] Ramin Raziperchikolaei, Guannan Liang, and Young joo Chung. 2021. Shared Neural Item Representations for Completely Cold Start Problem. In *Proc. of ACM RecSys* (Amsterdam, Netherlands). 422–431.

[50] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning Deep Representations of Fine-grained Visual Descriptions. In *Proc. of IEEE CVPR* (Las Vegas, NV, USA). 49–58.

[51] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of EMNLP* (Hong Kong, China). 3973–3983.

[52] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proc. of UAI* (Montreal, Quebec, Canada). 452–461.

[53] Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 2022. *Recommender Systems Handbook.* Springer US, New York, NY.

[54] Muhammad Saad Saeed, Muhammad Haris Khan, Shah Nawaz, Muhammad Haroon Yousaf, and Alessio Del Bue. 2022. Fusion and Orthogonal Projection for Improved Face-Voice Association. In *Proc. of ICASSP* (Singapore). IEEE, 7057–7061.

[55] Muhammad Saad Saeed, Shah Nawaz, Muhammad Haris Khan, Muhammad Zaigham Zaheer, Karthik Nandakumar, Muhammad Haroon Yousaf, and Arif Mahmood. 2023. Single-branch Network for Multimodal Training. In *Proc. of ICASSP* (Rhodes Island, Greece). 1–5.

[56] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large Language Models are Competitive Near Cold-start Recommenders for Language- and Item-based Preferences. In *Proc. of ACM RecSys* (Singapore, Singapore). 890–896.

[57] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[58] Walid Shalaby, Sejoon Oh, Amir Afsharinejad, Srijan Kumar, and Xiquan Cui. 2022. M2TRec: Metadata-aware Multi-task Transformer for Large-scale and Cold-start free Session-based Recommendations. In *Proc. of ACM RecSys* (Seattle, WA, USA). 573–578.

[59] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proc. of EMNLP* (Hong Kong, China). 5100–5111.

[60] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748

[61] Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proc. of NeurIPS* (Long Beach, CA, US). 6000–6010.

[63] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. 2017. DropoutNet: Addressing Cold Start in Recommender Systems. In *Proc. of NeurIPS* (Long Beach, CA, USA). 4964–4973.

[64] Cheng Wang, Mathias Niepert, and Hui Li. 2018. LRMM: Learning to Recommend with Missing Modalities. In *Proc. of EMNLP* (Brussels, Belgium). 3360–3370.

[65] Li Wang, Binbin Jin, Zhenya Huang, Hongke Zhao, Defu Lian, Qi Liu, and Enhong Chen. 2021. Preference-Adaptive Meta-Learning for Cold-Start Recommendation. In *Proc. of IJCAI* (Montreal, Quebec, Canada). 1607–1614.

[66] Wenjie Wang, Xinyu Lin, Liuhui Wang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2023. Equivariant Learning for Out-of-Distribution Cold-start Recommendation. In *Proc. of ACM Multimedia* (Ottawa, Ontario, Canada). 903–914.

[67] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-Start Recommendation. In *Proc. of ACM Multimedia* (Virtual Event, China). 5382–5390.

[68] Jingxuan Wen, Huafeng Liu, and Liping Jing. 2023. Modeling Preference as Weighted Distribution over Functions for User Cold-start Recommendation. In *Proc. of ACM CIKM* (Birmingham, United Kingdom). 2706–2715.

[69] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proc. of EMNLP System Demos* (Virtual Event). 38–45.

[70] Hanrui Wu, Chung Wang Wong, Jia Zhang, Yuguang Yan, Dahai Yu, Jinyi Long, and Michael K. Ng. 2023. Cold-Start Next-Item Recommendation by User-Item Matching and Auto-Encoders. *IEEE Transactions on Services Computing* 16, 4 (2023), 2477–2489.

[71] Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal Learning with Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023).

[72] Hong-Jian Xue, Xin-Yu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep Matrix Factorization Models for Recommender Systems. In *Proc. of IJCAI* (Melbourne, Australia). 3203–3209.

[73] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proc. of ACM Multimedia* (Virtual Event, China). 3872–3880.

[74] Xuhao Zhao, Yanmin Zhu, Chunyang Wang, Mengyuan Jing, Jiadi Yu, and Feilong Tang. 2023. Task-Difficulty-Aware Meta-Learning with Adaptive Update Strategies for User Cold-Start Recommendation. In *Proc. of ACM CIKM* (Birmingham, United Kingdom). 3484–3493.

[75] Ziwei Zhu, Shahin Sefati, Parsa Saadatpanah, and James Caverlee. 2020. Recommendation for New Users and New Items via Randomized Training and Mixture-of-Experts Transformation. In *Proc. of ACM SIGIR* (Virtual Event, China). 1121–1130.