Item-Based Variational Auto-Encoder for Fair Music Recommendation

Team ML

Jinhyeok Park, Dain Kim, Dongwoo Kim

POSTECH Graduate School of AI

Machine Learning Laboratory

Contents

Introduction: EvalRS Data Challenge

- The goal of the challenge is to devise a model that satisfies various evaluation metrics comprehensively.
  - Given the history of user, we recommend the top 100 items for each user.

- Metrics

Accuracy metrics
(e.g., Hit Rate)

Accuracy metrics on
a per-group basis
(Fairness)

Behavioral tests
(Diversity)

→ Our methods mainly targets leveraging the **fairness** of the model.

Methods: Overview

- Backbone models
  1. Variational auto-encoders (VAE)
  2. Bayesian personalized ranking matrix factorization (BPRMF)

- Strategies for fairness
  1. Item-based VAE
  2. Popularity-aware training (target: **artist popularity groups)**
  3. Fairness regularizer (target: **item popularity groups)**

- Curating final recommendation result

Methods: Backbone Models

1. Variational Auto-Encoder for Collaborative Filtering
    - It aims to produce a user-item interaction matrix from multinomial distribution by maximizing a likelihood.

2. Bayesian personalized ranking matrix factorization (BPRMF)
    - It is a matrix factorization with pairwise ranking method.
    - We utilized BPRMF to leverage 'Be less wrong', by replacing the top-1 item with the results of BPRMF.

Methods: Strategies for Fairness

- U: the number of users
- I: the number of items

1. Item-based VAE

- Traditional VAE uses implicit feedback of **a user** as a model input. (User-based VAE)
- We propose an item-based VAE which utilizes implicit feedback of **an item**.
- Empirically, we found that the item-based VAE successfully mitigates the popularity bias.

User-based VAE

$$\mathbf{x_u} = [x_{u1}, x_{u2}, \ldots, x_{uI}]$$

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
U_1							
U_2							
U_3							
U_4							
U_5							

$U\ X\ I$

transpose

Item-based VAE

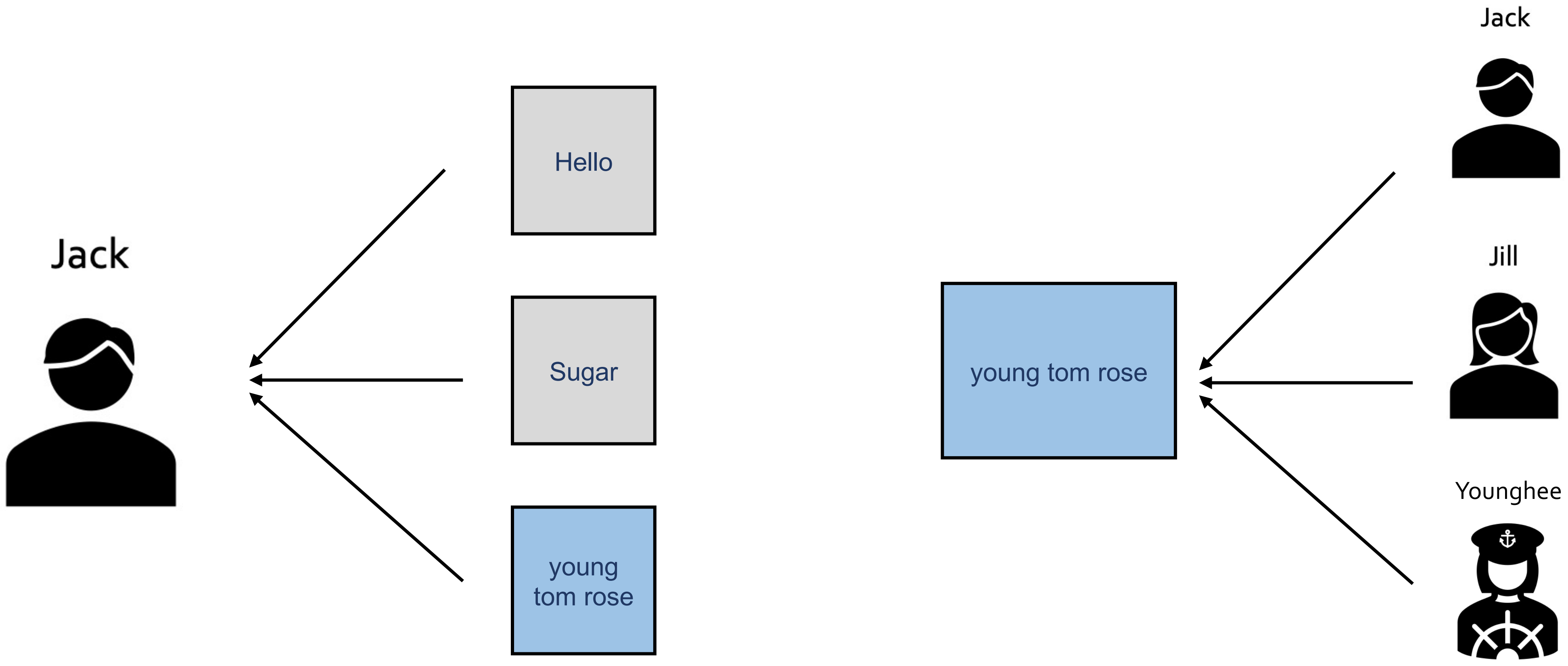
$$\mathbf{x_i} = [x_{i1}, x_{i2}, \ldots, x_{iU}]$$

	U_1	U_2	U_3	U_4	U_5
I_1	logit_11	logit_12	logit_13	logit_14	logit_15
I_2		logit_22			
I_3		logit_32			
I_4		logit_42			
I_5		logit_52			
I_6		logit_62			
I_7		logit_72			

$I\ X\ U$

Methods: Strategies for Fairness

1. Item-based VAE

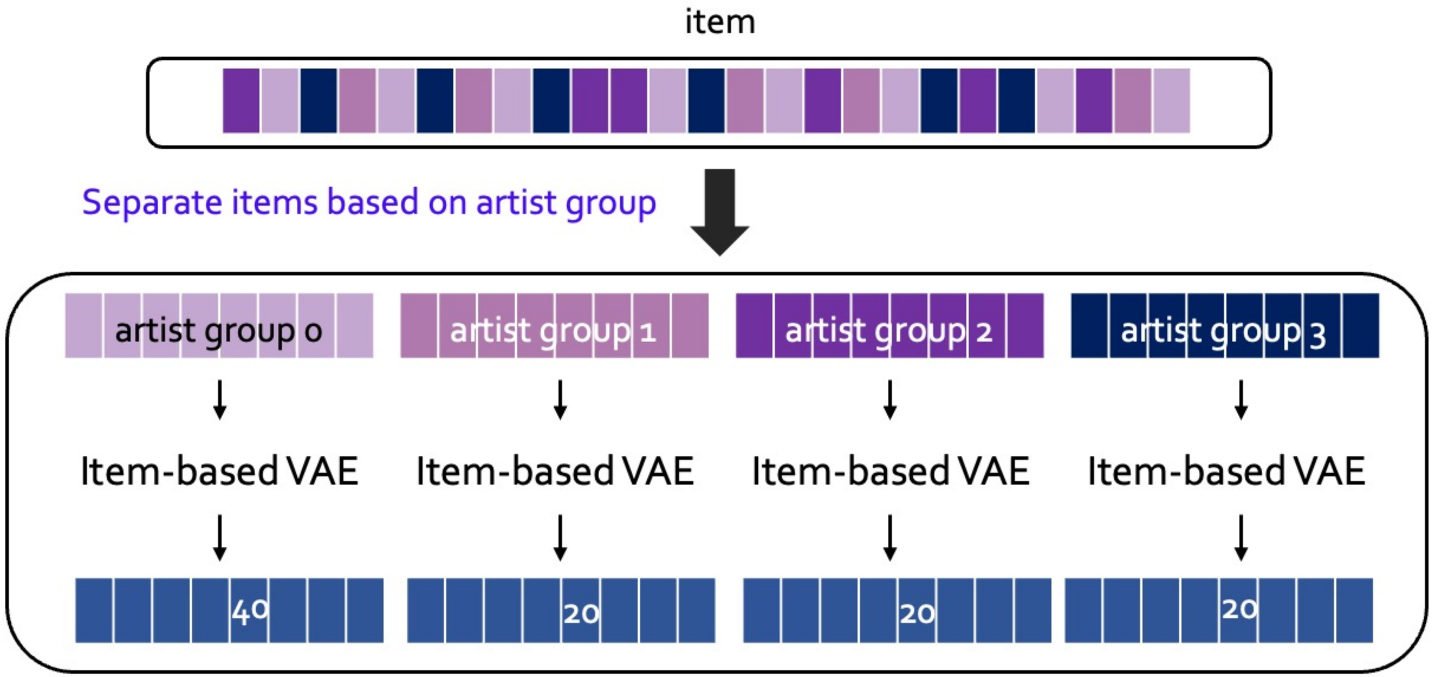


Item-based VAE focus on who would like to hear this **song?**

Methods: Strategies for Fairness

2. Popularity-aware training

- Goal: mitigate unfairness in artist popularity groups
- We divide items by artist popularity groups and train VAEs separately on each group.

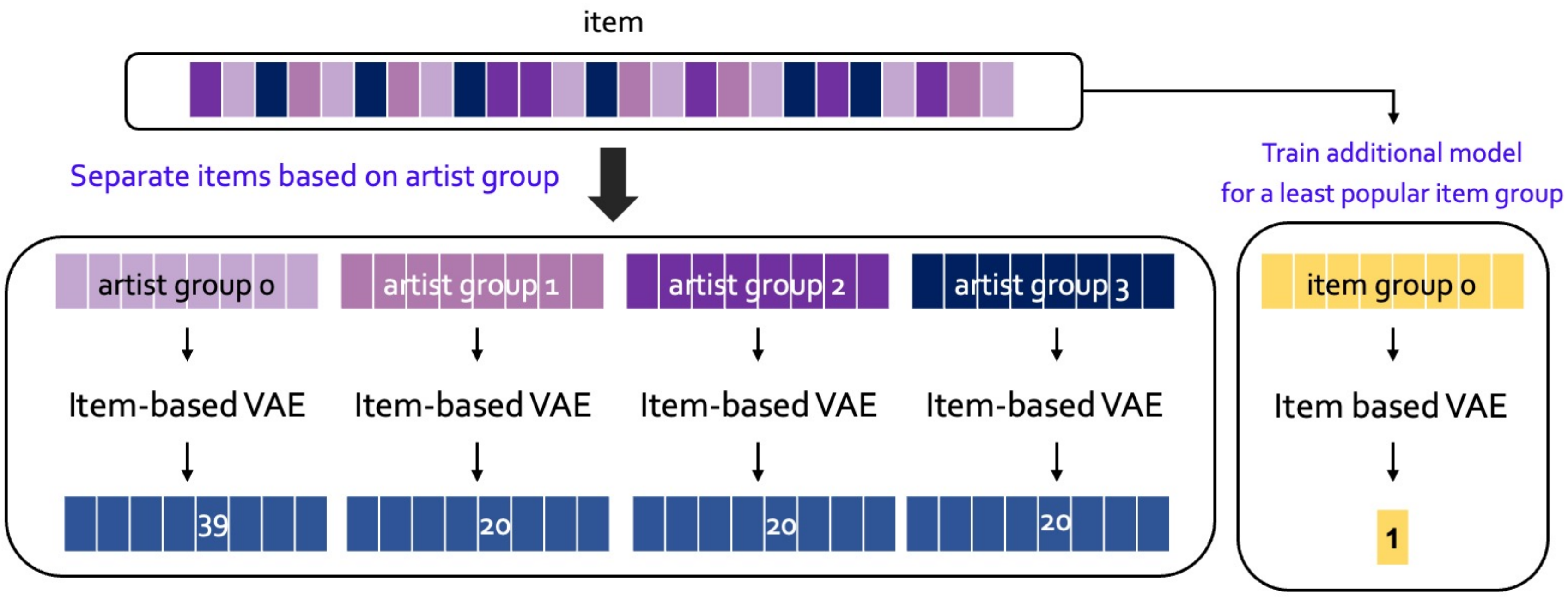


Methods: Strategies for Fairness

2. Popularity-aware training

- Moreover, we found that the number of items in the least popular item group is relatively small and not recommended well.

→ We adopted an additional VAE that is specifically trained on that group.



Methods: Strategies for Fairness

3. Fairness Regularization

- It aims to introduce an additional regularizer term to the objective to narrow the gap between group losses.

$$L_\beta^R(x_i;\theta,\phi) = \underbrace{L_\beta(x_i;\theta,\phi)}_{\text{beta VAE loss}} - \gamma \cdot \boxed{F_\phi}$$

- When it comes to VAE, the regularizer computes the average difference between the **group reconstruction loss** and the **entire reconstruction loss**.
  - Groups are divided into 0, 1, 2, and 3 based on the track popularity.

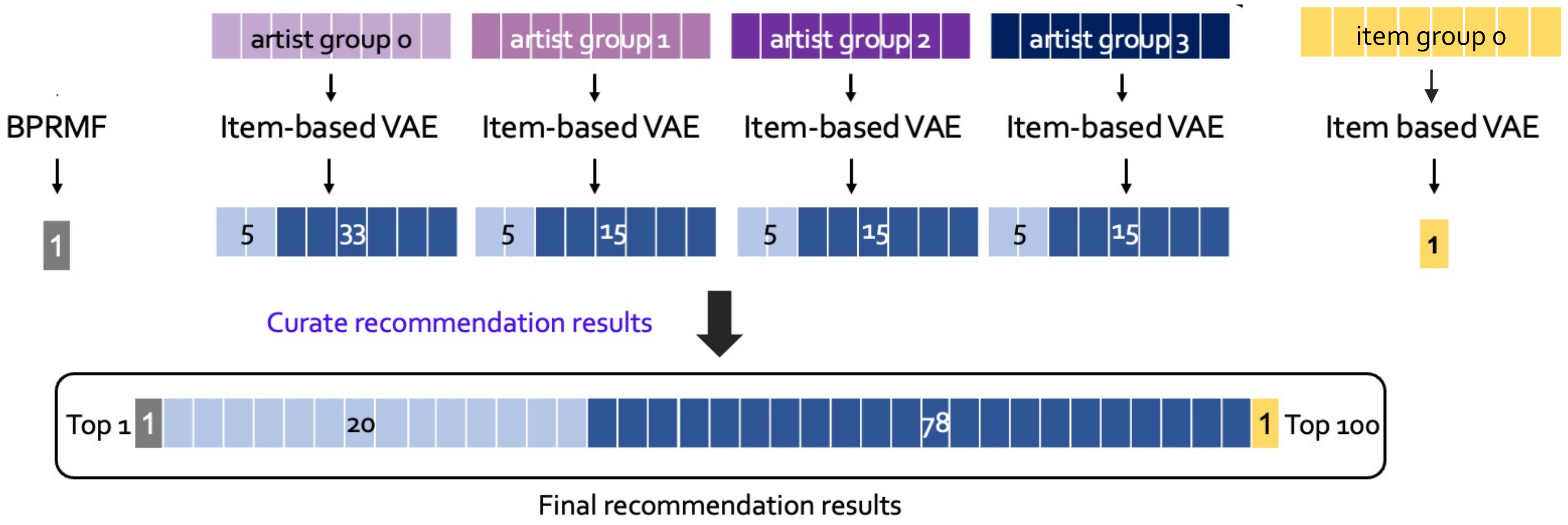
- I: the number of items
- G_j: a set of items in group j

$$F_\phi(x_i;\theta,\phi) = \mathbb{E}\left[\left|\underbrace{\frac{1}{|G_j|}\sum_{c\in G_j}\mathbb{E}\left[\log p_\theta\left(x_c \mid z_c\right)\right]}_{\text{group reconstruction loss}} - \underbrace{\frac{1}{I}\sum_{i=1}^{I}\mathbb{E}\left[\log p_\theta\left(x_i \mid z_i\right)\right]}_{\text{entire reconstruction loss}}\right|\right]$$

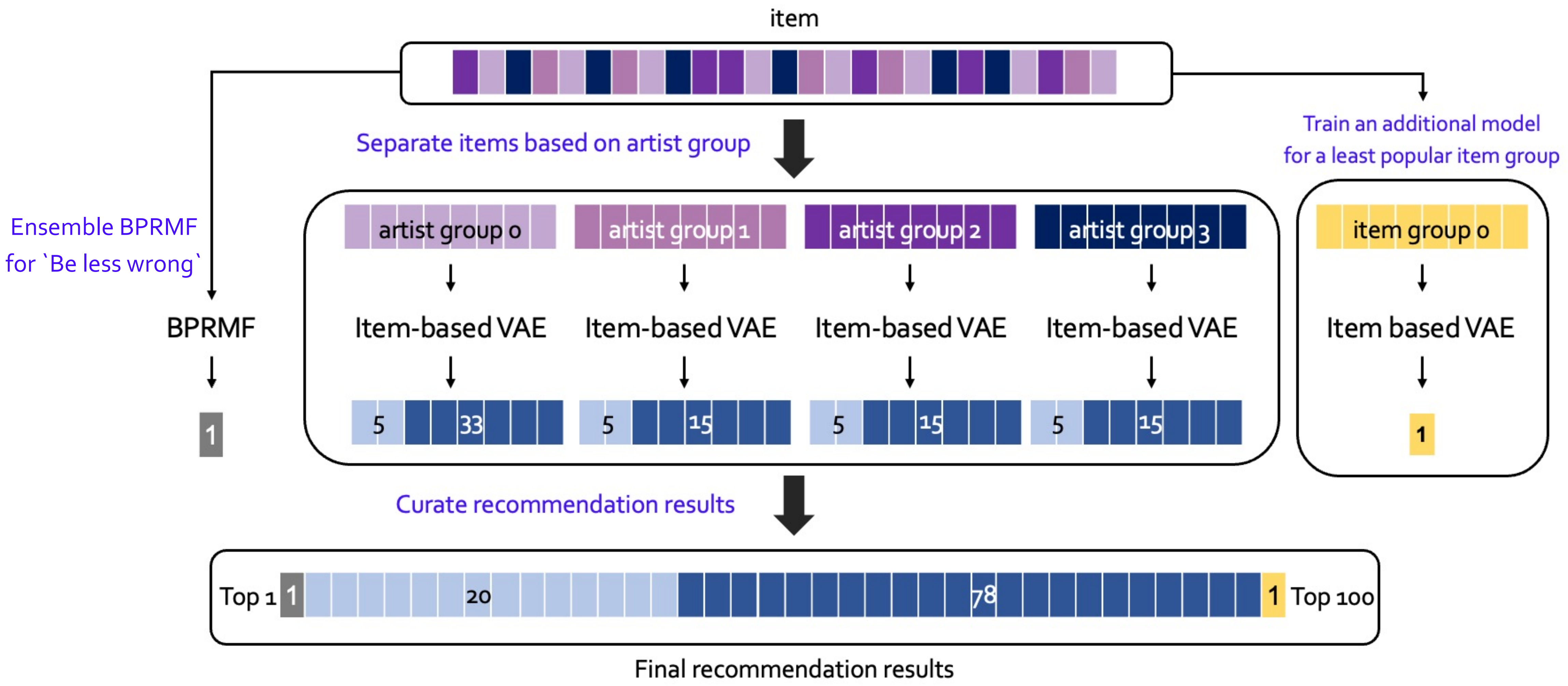
Methods: Final Recommendation

- We organize the outputs of item-based VAEs and BPRMF to produce the final recom mendation results.

1. Curate top-20 items with 5 most probable items in each group.
   - Order: group 2 ⟶ group 1 ⟶ group 3 ⟶ group 0
2. List remaining 78 items with the same order.
3. Replace top-1 & top-100 item.



Methods: Summary



Experiments

Analysis between user-based and item-based VAE

Phase 1 results of the baseline models

	Hit Rate	MRR	User (MRED)	TrackPop (MRED)	ArtistPop (MRED)	Score Phase1
VAE(item)	**0.2121**	**0.0399**	-0.0287	-0.0529	-0.0216	**0.0138**
VAE(user)	0.1593	0.0256	-0.0323	-0.0937	-0.0430	0.0082

- Item-based VAE outperforms user-based VAE not only in terms of the accuracy but also in terms of the MRED between various groups.

- Item-based VAE also produces the best performance in terms of phase 1.

Analysis between user-based and item-based VAE

Miss Rate (MR) of each item popularity groups

Model	1	10	100	1000	total
VAE (item)	0.8946	0.7865	0.7770	0.8803	0.7879
VAE (user)	0.9398	0.8861	0.8062	0.6448	0.8407
BPRMF	0.9965	0.9830	0.9387	0.9487	0.9628

- Unlike the user-based VAE, the item-based VAE successfully **mitigates the popularity bias.**
- Item-based VAE shows lower MR for groups of unpopular items.

Experiments

- Results

Performance of our model in four folds

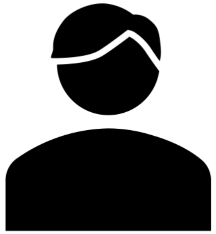
	Hit Rate	MRR	Country (MRED)	User (MRED)	TrackPop (MRED)	ArtistPop (MRED)	Gender (MRED)	Be less Wrong	Latent Diversity	Score Phase2
Fold1	0.0154	0.0015	-0.0030	-0.0035	-0.0021	-0.0007	-0.0003	0.3661	-0.2924	
Fold2	0.0151	0.0016	-0.0036	-0.0021	-0.0024	-0.0021	-0.0012	0.3602	-0.3000	
Fold3	0.0169	0.0021	-0.0047	-0.0044	-0.0023	-0.0005	-0.0004	0.3685	-0.2948	
Fold4	0.0169	0.0017	-0.0036	-0.0017	-0.0024	-0.0010	-0.0008	0.3609	-0.2984	
Average	0.0161	0.0017	-0.0037	-0.0029	-0.0023	-0.0010	-0.0007	0.3639	-0.2964	1.553
Baseline	0.0363	0.0037	-0.0090	-0.0224	-0.0111	-0.0072	-0.0061	0.3758	-0.3080	-1.212

- Our model outperforms the baseline model.
- The strategies for fairness successfully reduces the gap in various groups.

Proposed metrics

Motivation

Jack



$1,000

Jill

$10,000

Can we conclude that the value of $100 accounts the same?

Motivation

Model 1

Hit Rate:0.2

Model 2

Hit Rate:0.02

Can we conclude thats the MRED of 0.01 accounts the same?

Coefficient of Variance based Fairness

- Coefficient of Variance

$$\mathrm{CV} = \frac{\sigma}{m} * 100$$

  - It measures the relative ratio of deviation to performance.

- Coefficient of Variance based Fairness

$$\mathrm{CV}_{\mathrm{HR}} = (\mathrm{HR}_{\mathrm{avg}})^{-1} \sqrt{\frac{\sum_i (\mathrm{HR}_{\mathrm{avg}} - \mathrm{HR}_{\mathrm{group}_i})^2}{N_{\mathrm{groups}}}}$$

  - The proposed metric quantifies the fairness of the model, considering the **average of HR** when measuring the deviation.

Discussion: Coefficient of Variance based Fairness

- Example

	Group1	Group2	Group3	Group4	Hit Rate	MRED	CV ↓
Model1	0.001	0.002	0.003	0.004	0.0025	**-0.001**	**0.447**
Model2	0.019	0.02	0.021	0.022	0.0205	**-0.001**	0.055

- Model 1 and Model 2 show the same MRED, but the Hit Rate of model 2 is much higher.
- Thus, the return from the additional Hit Rate should be perceived differently.
- Our metric reasonably reflects this perceived difference.

Discussion: Coefficient of Variance based Fairness

- Results

Results of proposed metric

	Group0	Group1	Group2	Group3	Hit Rate	MRED	CV ↓
VAE(Item)	0.1741	0.1893	0.2312	0.2058	**0.2121**	-0.0216	**0.1019**
BPRMF	0.0150	0.0279	0.0371	0.0454	0.0372	**-0.0102**	0.3070

- BPRMF shows higher MRED compared to the item-based VAE.
- However, it has a relatively high deviation between groups.

- Our metric evaluates fairness well through the relative MRED to the Hit Rate.

Reflection

- Item-based VAE has a limitation that the model needs to infer all instances to make an inference to one user.

- Thus, it is valuable to develop a better method that focuses on the item's side.

- Moreover, our metric can be further elaborated in a way that reflects a user experience.

Thank you ☺