MEMER - Multimodal Encoder for Multi-signal Early-stage Recommendations

Mohit Agarwal*
mohitagarwal@sharechat.co
ShareChat
India

Srijan Saket*
srijan@sharechat.co
ShareChat
India

Rishabh Mehrotra
rishabhmehrotra@sharechat.co
ShareChat
India

ABSTRACT

Millions of content gets created daily on platforms like YouTube, Facebook, TikTok etc. Most of such large scale recommender systems are data demanding, thus taking substantial time for content embedding to mature. This problem is aggravated when there is no behavioral data available for new content. Poor quality recommendation for these items lead to user dissatisfaction and short content shelf-life. In this paper we propose a solution **MEMER** (**M**ultimodal **E**ncoder for **M**ulti-signal **E**arly-stage **R**ecommendations), that utilises the multimodal semantic information of content and uses it to generate better quality embeddings for early-stage items. We demonstrate the flexibility of the framework by extending it to various explicit and implicit user actions. Using these learnt embeddings, we conduct offline and online experiments to verify its effectiveness. The predicted embeddings show significant gains in online early-stage experiments for both videos and images (videos: 44% relative gain in click through rate, 46% relative gain in explicit engagements, 9% relative gain in successful video play, 20% relative reduction in skips, images: 56% relative gain in explicit engagements). This also compares well against the performance of mature embeddings (83.3% *RelaImpr* (RI) [18] in Successful Video Play, 97.8% *RelaImpr* in Clicks).

KEYWORDS

Recommendations, Early stage, Multimodal Semantic

ACM Reference Format:
Mohit Agarwal*, Srijan Saket*, and Rishabh Mehrotra. 2023. MEMER - Multimodal Encoder for Multi-signal Early-stage Recommendations. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3543873.3587679

1 INTRODUCTION

Early stage recommendation plays a pivotal role in the journey of a content during its lifecycle. Traditional recommendation techniques like collaborative filtering, content-based filtering have proven to be useful for a considerable amount of time. Recent studies have shown that embedding based approaches are better for item cold start recommendation than traditional approaches because they can learn low-dimensional representations of items, allowing them to recommend new items to users even when metadata like title, genre is not available. Additionally, these approaches can scale to millions to items and users. But most of such large scale recommender systems are data demanding, thus taking substantial time for content embedding to mature. For example, based on our observations, Figure 1, it takes ~20k views for an item embedding to stabilize for an interaction dependent algorithm. This problem is aggravated when there is no behavioral data available for new content. Due to this, recommender systems which are trained on historical behavioral data are insufficiently trained for new items leading to poor performance in online early stage recommendation [8, 10].

Given the importance, there has been a lot of interest recently both in academia and research divisions of major internet companies to address the item cold-start recommendation. CB2CF [1], Heater [21] are improvements on top of CF based approaches. CLCRec [17] tries to solve early stage ranking by maximizing the mutual dependencies between item content and collaborative signals. All the mentioned approaches are not very flexible because they are specifically designed for CF-based backbone models. There are some model-agnostic approaches like DropoutNet [16] which applies dropout to learn to reproduce the accuracy of the input latent model when preference data is available while also generalizing to cold start. Meta Embedding [12] and MeLU [9] utilizes a meta learning approach to learn initial desirable embeddings for new items. It leverages the representations of previously learnt items using a gradient based approach. MWUF [20] proposes meta scaling and meta shifting networks to warm up cold item embeddings. Even the model-agnostic approaches like DropoutNet, MeLU and MWUF have limitations because of strict data requirements to learn cold item embeddings. CVAR [19] takes a step further by removing extra requirements for data and uses latent variables to learn a distribution over item side information. It generates desirable item ID embeddings using a conditional decoder. But CVAR only utilizes categorical variables in the form of one-hot vectors as item side information. This leads to two problems, (1) Non-linearity is not captured [13] (2) Using only categorical features impacts fairness in early stage recommendation

In this paper, we work on the videos uploaded on the Indian social media platform ShareChat. We propose a multimodal approach to capture the semantic information of content and use it to enhance the early stage recommendation where behavioral information is absent. We leverage the audio, visual, text features extracted through DNNs and club it with other auxiliary features to achieve the desired embeddings. We also prove the efficacy of

*Equal Contribution

WWW '23 Companion, April 30-May 4, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587679

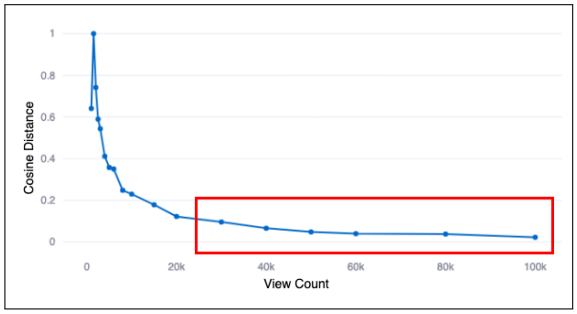


Figure 1: Embedding Maturity Curve (Video Play Signal)

this approach on a variety of explicit and implicit signals [3, 6, 15] in the context of recommendation in social media. We prove this through a series of offline and online experiments where we observe a relative gain of 43.9%, 46.14%, 9.07% in online metrics (CTR, explicit engagements and Successful video play respectively) and gains of similar magnitude in offline metrics in terms of AUC, F1 and *RelaImpr* (RI) [18]. We also show that MEMER achieves fairness by ensuring equitable performance across different categories of content. This provides a level ground to all the content irrespective of historical popularity which would not be the case if only behavioral information is taken into consideration.

2 PROPOSED APPROACH

2.1 Model

2.1.1 Problem Statement. Given the user-item interactions and semantic side-information of items, our goal is to learn a model that can generate embeddings for early-stage items and thus provide personalized recommendations.

2.1.2 Methodology. A base backbone model (Field-aware Factorization Machines [7]) is trained on all user-item interactions. The model is behavioural-based and performs well for items having sufficient data. Such items which have crossed a maturity threshold (M) are considered *Old Items* within the system. Using interactions and semantic understanding of these old items, we train the MEMER model to transform semantic embeddings to the space of base model FFM.

2.1.3 Architecture. The proposed MEMER model (Figure 2b) consists of encoders, decoder and a multimodal semantic module. The encoders and decoder are standard neural networks motivated by the ones proposed in [19]. In addition to this, we propose a semantic module which aims to leverage the spatio-temporal information of items through semantic audio-visual-textual features and generate personalized recommendations for users in the early stage of an item. The proposed model trains on the user-item interactions using mature embeddings of the backbone model with additional semantic audio-visual-textual feature information of the items.

Encoders. Encoder enc_1 takes the mature FFM (base) embedding of an item i (denoted by v_i, where $v_i \in \mathbb{R}^d$) and produces output e_1.

$$e_1 = enc_1(v_i) \tag{1}$$

Encoder enc_2 takes the fused features of item i, generated from the semantic module using multi-modal information (denoted by $feat_i$), and produces output e_2.

$$e_2 = enc_2(feat_i) \tag{2}$$

Decoder. Decoder dec_1 jointly works on the output of both the encoders and produces reconstructed embedding (p_{recon}) and predicted embedding (p_{rec}). It also takes frequency ($freq$) as an input which is the normalized count of item i to make the model aware of the stage of the item. Low $freq$ would represent early stages and so on.

$$p_{recon} = dec_1(e_1, freq) \tag{3}$$

$$p_{rec} = dec_1(e_2, freq) \tag{4}$$

Multimodal Semantic Module. V_i, A_i and T_i represent the visual, audio and textual features extracted from feature extraction models as explained in 2.2. We perform an early-fusion of the embeddings and train a non-linear neural network on it (W and b are the weights and biases of the network). It generates an output $feat_i$ which is a non-linear low-dimensional semantic representation of item i.

$$feat_i = \sigma(W(V_i \oplus A_i \oplus T_i \oplus Aux_i) + b) \tag{5}$$

Training Objective. L_w refers to the wasserstein distance [14] between enc_1, enc_2.

$$L_{encoder} = L_w(enc_1, enc_2) \tag{6}$$

$L_{reconstruction}$ is the mean-squared-error between p_{recon} and v_i. MSE is a measure of the difference between the original and the reconstructed signals, and it quantifies how well the reconstruction process has preserved the original signal.

$$L_{reconstruction} = \sum_{j=1}^{d} (p_{recon}^j - v_i^j)^2 \tag{7}$$

S^{pred} is the dot product between the user embedding and the predicted item embedding; z^{pred} is the sigmoid of S^{pred}; $L_{recommendation}$ is the binary cross entropy loss which optimizes the log-likelihood, is easy to compute, penalizes confident wrong predictions, handles class imbalance, and works well for probabilistic models

$$S^{pred} = v_{early}^{pred} \cdot v_u \tag{8}$$

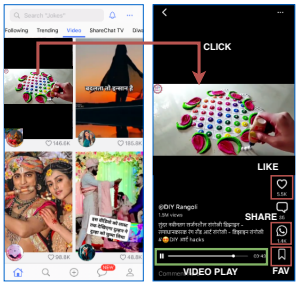
$$z^{pred} = \frac{1}{1 + e^{-S^{pred}}} \tag{9}$$

$$L_{recommendation} = -(y \log(z^{pred}) + (1 - y) \log(1 - z^{pred})) \tag{10}$$

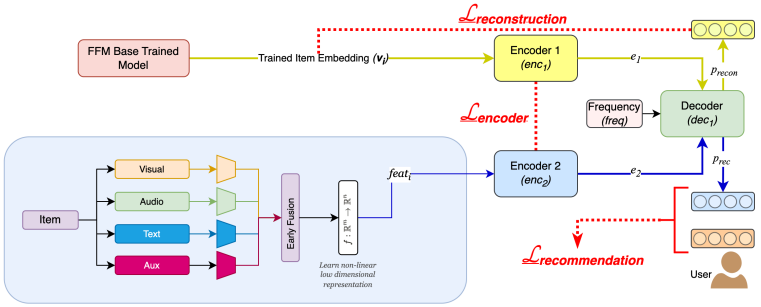
L_{MEMER}: Combined weighted loss of all the individual losses defined above

$$L_{MEMER} = \alpha L_{encoder} + \beta L_{reconstruction} + \gamma L_{recommendation} \tag{11}$$

α, β and γ are hyper-parameters and we select the set of values that produce the best overall performance on our test dataset using random search. We also take into account the relative scales of the individual losses and the desired trade-off between them.



(a) Platform Signals, Red- explicit, Green- implicit



(b) MEMER Model Overview

Figure 2: Platform Signals and Model Overview

2.2 Feature Extraction

2.2.1 Visual Embeddings. For videos, we leverage the pretrained Resnext3d-101 model [11] for visual feature extraction. Our approach involves sampling frames at 18FPS, followed by averaging to generate a single 2048-D embedding per second. We further average pool across seconds to get one representative 2048D embedding per video. Resnext3d-101 model is trained on Kinetics 400 dataset and can effectivey capture spatiotemporal information in video sequences. For images, we use Densenet161 [5] model and extract the 2208-D penultimate layer for features. This model has been trained on ImageNet dataset and performs robustly well on a range of image based computer vision applications.

2.2.2 Audio Embeddings. We extract the audio channel from our videos and it is sampled at 16kHz to get the output audio stream. This is then fed into a pre-trained VGG model [4] to extract the audio embeddings. VGGish is trained for sound classification and produces high-level semantically meaningful 128-D embeddings.

2.2.3 Text Embeddings. We also extract text from image frames and then use a pre-trained Fasttext [2] model to generate 100-D embeddings for the extracted text. The model is fast and efficient in capturing subword information.

2.3 Inference

During inference, the required features of an early-stage item are extracted and passed through the trained model to generate a fixed 32-dimension embedding vector. This vector is then used to calculate the dot product with the user embedding (v_u), producing a score that indicates the likelihood of the user interacting with the item. This score is then used for personalized item suggestions at user-level. For an early-stage item,

$$v^{pred}_{early} = dec_1(enc_2(feat_{early}), freq_{early}) \tag{12}$$

$$S^{pred}_{early} = v^{pred}_{early} \cdot v_u \tag{13}$$

3 EXPERIMENTS

3.1 Offline

3.1.1 Dataset. We collect the video and image post interactions data in Hindi language for a period of 40 days for multiple implicit (Video Play - if a user successfully watches a recommended video beyond a certain threshold, Skip) and explicit (Click, Like, Share, Favorite) signals. Along with the interactions, we also have the user embedding, item embedding and item semantic features. We also have tag field available corresponding to all the posts. For the train split, we take the posts created in the starting 33 days which have crossed the required threshold (M) to be considered as old items. To get an estimate of M, we evaluate the running cosine distance across view counts for embeddings at t and t-1, as shown in Figure 1. We observe the value on x-axis when the distance between embeddings starts tending to 0 and use this as the threshold (M). Similarly, posts created in the last 7-day period are taken in the test split. We randomly sample negatives per each positive interaction of a post in training. To compare our early-stage embeddings with the matured embeddings of the base model, we take test posts that have also crossed required threshold (M). Dataset details are summarized in Table 1.

Table 1: Dataset Statistics

Signal		Split	No. of Users	No. of Items	No. of Interactions
VIDEO	Click	Train	30,647,285	154,644	1,030,517,743
		Test	21,746,963	21,011	1,072,692,829
	Video Play	Train	43,939,529	179,804	7,246,081,748
		Test	10,739,212	17,191	79,036,346
	Skip	Train	29,864,325	179,824	5,613,192,373
		Test	8,594,674	3,977	39,615,433
	Like	Train	28,987,643	154,969	1,523,268,936
		Test	6,863,130	2,897	30,066,371
	Share	Train	21,236,448	138,869	487,987,856
		Test	5,697,818	2,572	24,292,208
	Favorite	Train	28,389,195	151,385	1,326,606,212
		Test	7,122,308	2,831	30,703,720
IMAGE	Click	Train	10,593,267	117,696	54,908,003
		Test	2,663,538	14,120	146,456,845
	Like	Train	15,892,357	128566	265,046,064
		Test	3,537,010	15,243	202,432,864
	Share	Train	14,360,774	98,627	243,284,943
		Test	3,683,436	11,975	175,449,454
	Favorite	Train	20,960,965	126,368	610,157,595
		Test	6,883,538	14,582	294,611,311

3.1.2 Baselines. We compare the proposed embeddings with some of the most commonly used initialization methods [Fig1].

(1) Random init, where values are sampled from a normal distribution.
(2) Global average [20], $X^{signal} = (\sum^{N} x^{signal})/N$ which is the average of all post embeddings in the training corpus.

Table 2: Offline Results

	Methods	Click			Video Play			Skip			Like			Share			Favorite		
		AUC	F1	RI(%)	AUC	F1	RI(%)	AUC	F1	RI(%)	AUC	F1	RI(%)	AUC	F1	RI(%)	AUC	F1	RI(%)
VIDEO	Random	0.533	0.039	8.70	0.500	0.151	-0.15	0.500	0.430	-0.30	0.503	0.071	0.81	0.505	0.045	1.65	0.499	0.080	-0.36
	Global Avg	0.763	0.056	69.0	0.604	0.197	66.2	0.546	0.481	42.1	0.885	0.159	98.6	0.753	0.076	86.0	0.715	0.133	84.4
	Tag Avg	0.724	0.049	58.6	0.605	0.190	66.7	0.559	0.501	54.5	0.855	0.136	91.1	0.733	0.078	79.3	0.698	0.134	77.6
	MEMER (Video)	0.867	0.066	96.1	0.627	0.205	81.0	0.595	0.505	87.4	0.886	0.158	98.9	0.760	0.081	88.5	0.728	0.139	89.3
	MEMER (Visual + Audio)	**0.873**	**0.067**	**97.8**	**0.631**	**0.207**	**83.3**	**0.600**	**0.508**	**91.8**	**0.887**	**0.160**	**99.1**	**0.766**	**0.086**	**90.7**	**0.731**	**0.144**	**90.7**

	Methods	Click			Like			Share			Favorite		
		AUC	F1	RI(%)	AUC	F1	RI(%)	AUC	F1	RI(%)	AUC	F1	RI(%)
IMAGE	Global Avg	0.631	0.017	50.2	0.843	0.113	93.9	0.701	0.118	63.1	0.701	0.130	64.9
	Tag Avg	0.574	0.015	28.3	0.856	0.091	97.4	0.741	0.135	75.9	0.745	0.153	78.9
	MEMER (Visual + Text)	**0.685**	**0.019**	**70.9**	**0.876**	**0.107**	**102.8**	**0.778**	**0.139**	**87.4**	**0.768**	**0.148**	**86.3**

(3) Tag average, $X_{tag}^{signal} = (\sum^{N_{tag}} x_{tag}^{signal})/N_{tag}$ which is the average of post embeddings within a particular tag.

We evaluate performance of the embeddings using *AUC score* and *F1 score*. Since we have taken test posts which have sufficient interactions, we also compare the performance of our early stage embeddings with matured embeddings using *RelaImpr* (RI) [18] $\frac{AUC(model)-0.5}{AUC(matured)-0.5} * 100\%$. This gives an indication of the similarity of the predicted embeddings with respect to the matured embeddings.

3.1.3 Results and Analysis. The offline results in Table 2 shows that MEMER is outperforming the other initialisation approaches by a significant margin. The AUC is better compared to others both in explicit and implicit signals. Additionally, we have performed ablative experiments for videos to prove the effectiveness of each component of the semantic module. The results demonstrate that MEMER (visual and audio features) does better than MEMER (only visual features) across all metrics, underlining the importance of incorporating an additional modality to the semantic module. Furthermore, to gauge the generalizability of our proposed approach, we have conducted experiments on two different datasets that use different types of multimodal features. Our model achieves significantly better results in both of them, highlighting the robustness and general usability of MEMER.

Comparison with matured embeddings: The gains in RI suggests that MEMER is able to come closest compared to others when the performance is compared against the matured embeddings. This also indicates that the initialisation is happening much closer to the convergence point leading to comparable performance even with no behavioral feedback.

3.1.4 Implementation Details. The embedding size of user and item embeddings has been fixed to 32. Learning rate is set to 0.001. Training is done using Adam optimizer with shuffled mini-batches of size 16384. We use PyTorch for all the training and experiments and is done on Tesla T4 GPUs.

3.2 Online

3.2.1 Experiment Setup. The experiment was setup as an AB test where the control set of users were shown new content using the tag-based approach and the test set was shown new content where the embeddings were initialised by MEMER predictions. Both the experiment groups had around 150k users. A dot product between user embedding and content embedding was taken to generate recommendation. The list was sorted by scores and the topK content was shown to the users.

Table 3: Online Results

User Metrics		Control (Tag Avg)	Test (MEMER)	Relative Gain (%)
VIDEO	CTR	0.0459	0.0613	**43.90**
	Engagements/Views	0.0099	0.0145	**46.14**
	Successful Video Play	0.2633	0.2871	**9.07**
	Skips	0.1583	0.1255	**-20.75**
	Interactions/Views	0.0220	0.0334	**52.07**
IMAGE	Engagements/Views	0.0291	0.0454	**56.33**

3.2.2 Results & Analysis. The online results in Table 3 suggest that the explicit actions including likes, shares, favorites (save to gallery), engagement (likes+shares+favorites) and CTR (click through ratio) is better than control by ~50%. We also observe significant gains in implicit user feedback signals including skips (20% lesser compared to control) and successful video watch (9% better compared to control). Both the implicit actions are mapped to a binary outcome on the basis of video watch time and duration of the video.

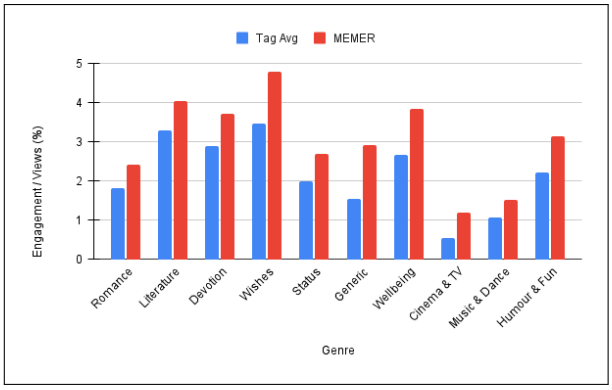


Figure 3: Online Results across Genres

Figure 3 shows that the online gains we observe with MEMER cannot be attributed to better performance in a specific category. For eg. One can come up with an approach that performs extremely well for an engagement heavy category like "Wishes" and the overall numbers for this approach could be better than the control variant.

But it fails the objective of providing a fair ground to all content by preferentially treating one category over the others. A tag-based approach leverages user's current preferences for targeting and is over-indexed on popularity. Compared to this, MEMER is learning the embeddings at an item level and then using it to map it to the right set of users. This ensures that even for a niche category the approach is able to find the interested users within the category better.

4 CONCLUSION

In conclusion, we propose a novel solution, MEMER, to tackle the problem of generating high-quality embeddings for early-stage content. Our framework effectively utilizes the multimodal semantic information of content to generate embeddings that perform significantly better in offline and online experiments compared to conventional methods. The offline experiments also demonstrate the effectiveness of our approach in generating better-quality embeddings. Our framework's flexibility allows us to extend it to different media formats and various explicit and implicit user actions, further improving the quality of generated embeddings. Our results demonstrate the potential of MEMER to significantly improve user engagement and content shelf-life in large-scale recommender systems.

5 LIMITATIONS & FUTURE WORK

The limitations of the MEMER model are primarily due to its reliance on the quality of embeddings generated by the underlying algorithm. The use of historical data and base model embeddings for training can limit the model's effectiveness. To improve the model's performance, future work could focus on adding features such as creator and location information, as well as more detailed content data. SOTA feature fusion techniques could also be employed for model enhancement. Additionally, online training of the base model could lead to improved performance in both early and later stages of content recommendation. Architectural improvements, such as the use of noise aware losses, could be employed to address the issue of noise and biases in the underlying model and reduce direct dependency.

REFERENCES

[1] 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 *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 228–236. https://doi.org/10.1145/3298689.3347038

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5 (2017), 135–146.

[3] Rod Ellis, Shawn Loewen, and Rosemary Erlam. 2006. IMPLICIT AND EXPLICIT CORRECTIVE FEEDBACK AND THE ACQUISITION OF L2 GRAMMAR. *Studies in Second Language Acquisition* 28, 2 (2006), 339–368. https://doi.org/10.1017/S0272263106060141

[4] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[6] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. 2010. Comparison of Implicit and Explicit Feedback from an Online Music Recommendation Service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems* (Barcelona, Spain) *(HetRec '10)*. Association for Computing Machinery, New York, NY, USA, 47–51. https://doi.org/10.1145/1869446.1869453

[7] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM conference on recommender systems*. 43–50.

[8] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*. 208–211.

[9] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1073–1082.

[10] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert systems with applications* 41, 4 (2014), 2065–2073.

[11] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*. 181–196.

[12] Feiyang Pan, Shuokai Li, Xiang Ao, Pingzhong Tang, and Qing He. 2019. Warm up cold-start advertisements: Improving ctr predictions via learning to learn id embeddings. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 695–704.

[13] Cedric Seger. 2018. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

[14] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[15] Masahiro Takimoto. 2006. The effects of explicit feedback on the development of pragmatic proficiency. *Language Teaching Research* 10, 4 (2006), 393–417. https://doi.org/10.1191/1362168806lr198oa arXiv:https://doi.org/10.1191/1362168806lr198oa

[16] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. 2017. Dropoutnet: Addressing cold start in recommender systems. *Advances in neural information processing systems* 30 (2017).

[17] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[18] Ling Yan, Wu-Jun Li, Gui-Rong Xue, and Dingyi Han. 2014. Coupled group lasso for web-scale ctr prediction in display advertising. In *International Conference on Machine Learning*. PMLR, 802–810.

[19] Xu Zhao, Yi Ren, Ying Du, Shenzheng Zhang, and Nian Wang. 2022. Improving Item Cold-start Recommendation via Model-agnostic Conditional Variational Autoencoder. *arXiv preprint arXiv:2205.13795* (2022).

[20] Yongchun Zhu, Ruobing Xie, Fuzhen Zhuang, Kaikai Ge, Ying Sun, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1167–1176.

[21] 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 *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1121–1130. https://doi.org/10.1145/3397271.3401178