# Single-Branch Architectures for Recommendation

Marta Moscati[1,*]

[1]Johannes Kepler University Linz, Linz, Austria

**Abstract**

Single-branch architectures have been proven effective for several multimodal learning tasks. In this talk, after reviewing the use of single-branch architectures in multimodal learning, I describe their use in multimodal recommendation, showing how they allow to address missing-modality and cold-start scenarios. I then describe the use of single-branch architectures in collaborative filtering, showing how they allow to reduce the number of model parameters without substantially affecting the quality of recommendations.

**Keywords**

Multimodal Learning, Multimodal Recommender Systems, Collaborative Filtering,

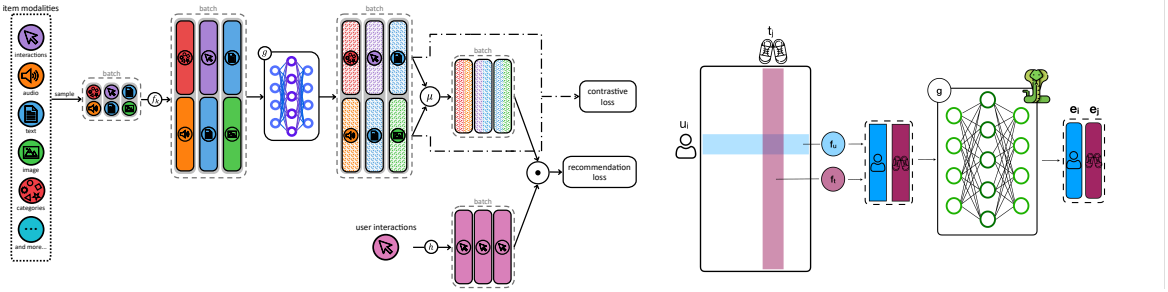
**Single-Branch Architectures in Multimodal Learning.** Multimodal learning leverages information from multiple diverse modalities (e.g., text, image, audio) to improve the performance of machine learning (ML) models on various tasks, such as speech recognition, event detection, and media description [1, 2, 3]. Typically, multimodal models address these tasks by mapping all modalities to a joint embedding space where modality instances are close if they have similar semantic meaning. This approach is referred to as *reducing the modality gap*, i.e., reducing the distance between embeddings of different modalities if they refer to the same content. For instance, text-visual models like CLIP [4] reduce the distance between embeddings of matching images and descriptions (e.g., the picture of a puppy and the text "*A picture of my dog*"). Most common multimodal methods achieve this goal with the use of separate, modality-specific Neural Networks (NN) [5, 6, 7, 8, 9, 10, 11, 12, 2, 13]. In contrast to these networks, recently single-branch architectures [14, 15] have shown promising results in multimodal learning. These architectures use the same NN to map multiple modalities to the joint embedding space. The positive impact of sharing model weights across modalities has also been proven particularly effective in addressing the performance drop experienced in missing-modality scenarios [16, 17] i.e., when certain modalities are not available [18]. Furthermore, sharing the same NN to encode multiple modalities allows reducing the number of model parameters of single-branch architectures compared to architectures using modality-specific branches.

**Single-Branch Architectures in Multimodal Recommendation.** Multimodal recommender systems (RS) [3, 19] apply multimodal learning to the domain of recommendation by providing recommendations to users based on input data of multiple diverse modalities. Extending the approach of collaborative filtering (CF) methods, which rely solely on user–item interaction data, multimodal RS leverage also side information on the users, such as demographic data, or items, such as textual reviews or product images. This renders multimodal RS effective also in cold-start scenarios [20] i.e., cases in which the lack of historical data on past interactions for certain users or items leads to a lower recommendation accuracy. Motivated by the observation that the performance deterioration in cold-start scenarios can be regarded as a missing-modality scenario, we proposed to address cold start by means of a Single-Branch Recommender (SiBraR [21]). The architecture of an instance of SiBraR leveraging side information on the items is shown in Figure 1a and consists of a single-branch encoding network $g$ shared across item modalities. The model relies on two loss function terms: one recommendation loss for Bayesian Personalized Ranking (BPR) [22] aimed at increasing the model recommendation accuracy, and one contrastive loss between pairs of item modalities [23, 24] aimed at further reducing the modality gap. We showed that in cold-start and missing-modality scenarios, SiBraR

*DaQuAMRec 2025: 1st International Workshop on Data Quality-Aware Multimodal Recommendation, September 22nd, 2025, Prague, Czech Republic*

marta.moscati@jku.at (M. Moscati)

0000-0002-5541-4919 (M. Moscati)




(a) SiBraR [21]: Single-branch multimodal RS.

(b) CoBraR [26]: Single-branch for CF.

**Figure 1:** Single-branch architectures for recommendation.

provides more accurate recommendations compared to well-established CF and multimodal RS [21], and that the performance improvement is to be attributed to the simultaneous use of the single-branch architecture and the contrastive loss [25].

**Single-Branch Architectures in Collaborative Filtering.** Modern CF algorithms [27] rely on representing users and items in a joint embedding space where these are close if corresponding to a positive interaction. Motivated by the analogy between this approach and the objective of reducing the modality gap in multimodal learning, as well as by the observation that weight sharing allows to simultaneously reduce the modality gap and to reduce the number of model parameters, we proposed the use of a single-branch architecture for CF. The resulting model, named Collaborative Branch Recommender (CoBraR [26]), is depicted in Figure 1b. The architecture and training paradigm of CoBraR are similar to the well-established two-branch CF algorithm DMF [28], where the user and item representations are obtained by converting their sparse interaction vectors, i.e., the corresponding row or column in the user–item interaction matrix, to dense vectors by mean of two multi-layer perceptrons (MLP) - one MLP for the user and one MLP for the item. The novelty of CoBraR is that the same MLP is used to encode both users and items. We showed that CoBraR reduces the number of model parameters compared to its two-tower counterpart DMF, improving beyond-accuracy aspects of recommendation quality without compromising accuracy.

**Future Applications of Single-Branch Architectures to Recommendation.** Both SiBraR and CoBraR translate single-branch architectures to RS and open the way for future applications. First, although SiBraR has been proven effective for tackling scenarios where one modality is missing at inference time, its architecture has not yet been evaluated on scenarios where one modality is also missing during training. As for CoBraR, its architecture and training paradigm are based on DMF. Therefore, there is no evidence yet of whether using weight sharing between users and items would also be effective with other loss functions for recommendation, or for other two-branch CF models as well, such as NeuMF [29]. Finally, SiBraR and CoBraR could be combined in a model where a same multimodal single-branch is applied to all multimodal data of both users and items, including interactions and side-information.

## 1. Acknowledgments

I am grateful to Markus Schedl, Shah Nawaz, Christian Ganhör, and Anna Hausberger for the fruitful collaboration and invaluable discussions that led to the development SiBraR and CoBraR. 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. No generative AI tool was used during the preparation of this work.

## Declaration on Generative AI

The author have not employed any Generative AI tools.

## References


[1] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019).

[2] P. Xu, X. Zhu, D. A. Clifton, Multimodal learning with transformers: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[3] Q. Liu, J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, Q. Li, J. Tang, Multimodal recommender systems: A survey, ACM Computing Surveys 57 (2024).

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proc. of ICML, 2021.

[5] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: Proc. of IEEE CVPR, 2016.

[6] A. Nagrani, S. Albanie, A. Zisserman, Seeing voices and hearing faces: Cross-modal biometric matching, in: Proc. of IEEE CVPR, 2018.

[7] A. Nagrani, S. Albanie, A. Zisserman, Learnable pins: Cross-modal embeddings for person identity, in: Proc. of ECCV, 2018.

[8] M. S. Saeed, M. H. Khan, S. Nawaz, M. H. Yousaf, A. Del Bue, Fusion and orthogonal projection for improved face-voice association, in: Proc. of ICASSP, 2022.

[9] O. Arshad, I. Gallo, S. Nawaz, A. Calefati, Aiding intra-text representations with visual context for multimodal named entity recognition, in: Proc. of IEEE ICDAR, 2019.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. of NeurIPS, 2017.

[11] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Proc. of NeurIPS, 2019.

[12] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, in: Proc. of EMNLP, 2019.

[13] A. Hannan, M. A. Manzoor, S. Nawaz, M. I. Liaqat, M. Schedl, M. Noman, PAEFF: Precise Alignment and Enhanced Gated Feature Fusion for Face-Voice Association, in: Proc. of Interspeech, 2025.

[14] M. S. Saeed, S. Nawaz, M. H. Khan, M. Z. Zaheer, K. Nandakumar, M. H. Yousaf, A. Mahmood, Single-branch network for multimodal training, in: Proc. of ICASSP, 2023.

[15] M. Tschannen, B. Mustafa, N. Houlsby, Clippo: Image-and-language understanding from pixels only, in: Proc. of IEEE/CVF CVPR, 2023.

[16] M. I. Liaqat, S. Nawaz, M. Z. Zaheer, M. S. Saeed, H. Sajjad, T. D. Schepper, K. Nandakumar, M. H. Khan, I. Gallo, M. Schedl, Chameleon: A multimodal learning framework robust to missing modalities, International Journal of Multimedia Information Retrieval 14 (2025).

[17] J. Geiger, M. Moscati, S. Nawaz, M. Schedl, Music4all a+a: A multimodal dataset for music information retrieval tasks, in: Proc. of CBMI, 2025.

[18] M. Ma, J. Ren, L. Zhao, D. Testuggine, X. Peng, Are multimodal transformers robust to missing modality?, in: Proc. of IEEE/CVF CVPR, 2022.

[19] D. Malitesta, G. Cornacchia, C. Pomo, F. A. Merra, T. Di Noia, E. Di Sciascio, Formalizing multimedia recommendation through multimodal deep learning, ACM Transactions on Recommender Systems 3 (2025).

[20] F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, New York, NY, 2022.

[21] C. Ganhör*, M. Moscati*, A. Hausberger, S. Nawaz, M. Schedl, A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios, in: Proc. of ACM RecSys, *Equal Contributions, 2024.

[22] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proc. of UAI, 2009.

[23] Z. Yuhui, Y. Wada, H. Waida, K. Goto, Y. Hino, T. Kanamori, Deep clustering with a constraint for topological invariance based on symmetric infonce, Neural Computation 35 (2023).

[24] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2019. arXiv:1807.03748.
[25] C. Ganhör, M. Moscati, A. Hausberger, S. Nawaz, M. Schedl, Single-branch network architectures to close the modality gap in multimodal recommendation, ACM Transactions on Recommender Systems (2025).
[26] M. Moscati, S. Nawaz, M. Schedl, Parameter-efficient single collaborative branch for recommendation, in: Proc. of ACM RecSys, 2025.
[27] S. Zhang, Y. Tay, L. Yao, A. Sun, C. Zhang, Deep learning for recommender systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, New York, NY, 2022.
[28] H.-J. Xue, X.-Y. Dai, J. Zhang, S. Huang, J. Chen, Deep matrix factorization models for recommender systems, in: Proc. of IJCAI, 2017.
[29] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proc. of WWW, 2017.