Improving Image Recommendation via Multimodal Model with Visual and Text Encoders

Joaquín De Ferrari[1], **Nicolás Schiaffino**[1], and **Gabriel Venegas**[1]

[1]Departamento de Informática, Universidad Técnica Federico Santa María, Santiago, Chile
{joaquin.deferrari@sansano.usm.cl, nicolas.schiaffino@usm.cl,
gabriel.venegaso@usm.cl}

Abstract

Image recommendation has become a fundamental task in recommender systems, specially pivotal in platforms that revolve around it, such as social media and digital galleries. Despite their relevance, there has been little research done in the recent years regarding this area. The predominant works revolve around Convolutional Neural Networks for feature extraction and interaction-based recommendation proving good results. However they disregard the semantic richness, the emotion, communicative intent, artistic style, or the inherent alignment that images may possess with other modalities, like text. This lack of semantic understanding limits the performance of the current implementations. To address this, we propose a modular multimodal recommendation framework that integrates visual (e.g., ResNet, CLIP) and text (e.g., Sentence-BERT) encoders with numerical interaction data. Our architecture employs a multi-head self-attention system and a fusion network is used to produce a final recommendation score. This framework allows fast and easy testing through its various components. While our research on the PixelRec dataset highlight the challenges of training complex models from scratch, our work provides an effective framework for fast and easy testing including everything needed for its operation.
Project repository: `https://github.com/Joacodef/PixelRec_Multimodal`

1 Introduction

The exponential growth of user-generated visual content has catapulted the previously existing challenge for image recommendation to the forefront. Platforms from Social networks to e-commerce and art galleries host millions of images, making navigation through this vast amount of information an overwhelming task without powerful personalisation. Recent attempts consist mostly of collaborative filtering, which leverages user interaction patterns, and content-based methods, which use visual features extracted by models like CNNs.

Recent development in large-scale vision-language models like CLIP Radford et al. [2021], alongside architectural innovations like the Transformer's attention mechanism Vaswani et al. [2017] that are able to extract a greater semantic meaning with less data allows the research for a richer, more precise and relevant image recommendation with a far more holistic and semantic understanding of content, moving beyond what an image is to what it means. State of the art methods for image recommendation don't make this distinction. These new developments have impacted other fields in the recommendation context but haven't translated in advancements in image recommendation.

To overcome these challenges and exploit these new technologies we propose a modular, open-source multimodal recommendation framework. Our architecture is designed to systematically integrate heterogeneous data sources by fusing rich embeddings from dedicated visual and textual encoders

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

with numerical interaction data. The core of our model is a multi-head self-attention mechanism that learns to dynamically weigh the importance of each modality for a given user-item context. The weighted features are processed by a final fusion network to generate a recommendation score.

Our main contributions are:

- Modular framework for image recommendation.
- A comprehensive empirical analysis on the PixelRec dataset.
- An in-depth analysis of the practical challenges.

The framework works by combining three distinct data sources, visual from the images themselves, text from within titles, tags or descriptions and numerical interactions metrics. The pipeline has a preprocessing and normalizing module to ensure correct operation, afterwards the data is fed to its corresponding channel. Each encoder transforms its raw input into a high-dimensional feature vector, or embedding. With the embeddings created they are projected into a common latent space to ensure they are compatible for comparison and fusion. The key piece of our work is the multi-head self-attention mechanism where the importance of each component is dynamically assessed, assigning more weight to the features most relevant for a given prediction. Pairwise learning for implicit feedback, specifically Bayesian personalized ranking with matrix factorization (BPRMF), typically outperforms pointwise learning counterparts.

2 Related Work

While collaborative filtering is a standard benchmark in recommendation, adapting it for implicit feedback datasets presents unique challenges, as explicit negative signals are absent. Foundational work in one-class collaborative filtering addressed this by reframing the problem: instead of treating unobserved interactions as truly negative, they were assigned a lower weight compared to observed positive interactions during model optimization Pan et al. [2008]. This weighting strategy was refined further by introducing variable confidence levels, where the model could infer a higher degree of preference from repeated user actions Hu et al. [2008]. Such models are generally categorized as pointwise methods, as they learn to predict the value of each individual user-item interaction. This stands in contrast to later pairwise approaches, like BPR, which instead learn to rank pairs of items for a user Rendle et al. [2009].

Early work in image recommendation relied mostly on content-based methods using visual features extracted by Convolutional Neural Networks (CNNs). Capable of identifying visually similar items, these approaches often fail to comprehend content at a deeper semantic level, relying solely on visual characteristics. To circumvent this, the field moved toward multimodal recommendation, with a focus on creating richer and more complete item representations by leveraging multiple data types Liu et al. [2024]. In this context, the use of pre-trained encoders to capture high-level semantic information becomes a natural choice.

For visual representation, ResNet He et al. [2016] has become a foundational architecture, introducing residual learning to effectively train extremely deep neural networks. More recent works have improved performance further. For instance, ConvNeXt Liu et al. [2022] modernized ConvNets to rival the performance of Transformers at a fraction of the complexity, while self-supervised methods like DINO Caron et al. [2021] learn powerful visual features from unlabeled data through self-distillation. On the textual side, models like BERT Devlin et al. [2018] and RoBERTa Liu et al. [2019] provide deep language understanding. Meanwhile, architectures such as Sentence-BERT Reimers and Gurevych [2019], an adaptation of BERT, are specialized in producing high-quality and semantically meaningful sentence embeddings. The strategy in multimodal systems is to project these diverse feature vectors into a shared latent space to enable their effective fusion, a technique popularized by models like CLIP Radford et al. [2021].

A central challenge in multimodal systems is how to effectively combine the features extracted from different encoders. Simple concatenation is a straightforward method, but it treats every data modality as equally important and, as such, is mostly used as a baseline. A more sophisticated approach is to use attention mechanisms, originating from the Transformer architecture Vaswani et al. [2017]. In the context of recommendation, attention can dynamically learn the importance of different features in relation to each other, assigning more weight to the modalities most relevant for a given prediction.

The multi-head attention variant allows this process to capture a wide range of complex inter-feature relationships from several perspectives simultaneously. In such architectures, an attention layer processes the combined feature vectors to produce a refined, context-aware representation, which is then passed to a final prediction network.

3 Multimodal framework

To address the challenges of multimodal image recommendation, we designed and implemented a modular framework that consists of a robust data processing pipeline and a flexible, attention-based neural architecture.

3.1 Data Preprocessing and Feature Caching

To handle a large and diverse dataset such as PixelRec, a systematic and efficient data handling module plays a fundamental role. The first step is validating and standardizing the raw data. An image processor handles all visual data, first checking for corrupted files and then resizing all images to a uniform dimension. Similarly, a text processor is implemented to normalize all text, such as titles and descriptions, preparing it for the text encoder. To handle data sparsity, a filter is applied, removing users and items that do not meet a minimum threshold of interactions. This step ensures that the model trains on a denser subset of the interaction graph, which can help in learning more meaningful patterns. A feature caching system is implemented to mitigate a significant computational bottleneck in multimodal systems: the repeated extraction of features from powerful, pre-trained encoders. After the initial preprocessing, the feature vectors for each item are extracted once for each combination of encoders and stored on disk. During training and evaluation, the framework retrieves these pre-computed embeddings from the cache instead of running the encoders again, dramatically accelerating experimentation cycles.

3.2 Multimodal Recommendation Architecture

The core of our framework is the `MultimodalRecommender`, a neural network designed to integrate and weigh features from different modalities. For any given user u and item i, the model generates a prediction score.

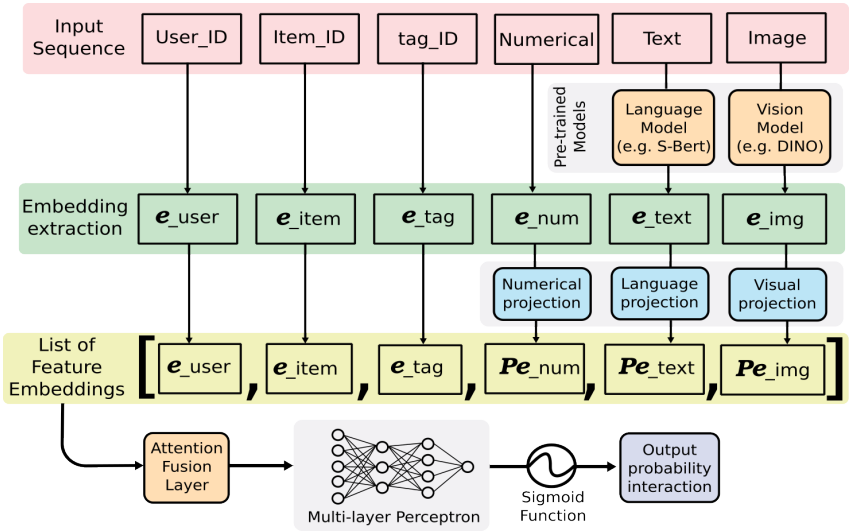


Figure 1: Architecture of the proposed multimodal recommendation framework. The model processes input IDs and raw text/image data, generates embeddings, projects them into a common latent space, fuses them with an attention layer, and predicts an interaction probability.

First, the model processes features from multiple sources simultaneously. It maps the user u and item i to dense embedding vectors, $\mathbf{p}_u$ and $\mathbf{q}_i$. Concurrently, it processes the item's metadata, including

pre-computed **visual features** ($\mathbf{e}_v$), **textual features** ($\mathbf{e}_t$), **categorical tags** ($\mathbf{e}_{cat}$), and **numerical features** ($\mathbf{e}_{num}$). As these features originate from different encoders (e.g., ResNet, Sentence-BERT), each is passed through a distinct projection layer (MLP) to map them into a common latent space of dimension d:

$$\mathbf{p}_v = \text{MLP}_v(\mathbf{e}_v), \quad \mathbf{p}_t = \text{MLP}_t(\mathbf{e}_t), \quad \mathbf{p}_{cat} = \text{MLP}_{cat}(\mathbf{e}_{cat}), \quad \mathbf{p}_{num} = \text{MLP}_{num}(\mathbf{e}_{num})$$

The resulting set of embeddings, $\{\mathbf{p}_u, \mathbf{q}_i, \mathbf{p}_v, \mathbf{p}_t, \mathbf{p}_{cat}, \mathbf{p}_{num}\}$, forms an input sequence $\mathbf{S}$ for the fusion module. Our framework supports multiple fusion strategies, including simple concatenation, a gated mechanism, and a more sophisticated **multi-head self-attention** mechanism that learns the contextual relationships among modalities. The attention function is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In this formulation, Q, K, and V represent linear projections of the input sequence $\mathbf{S}$, and d_k is the dimension of the keys. Multi-head attention enhances this by performing the attention function multiple times in parallel:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

where each head is defined as $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. This fusion step produces a single, context-aware vector, $\mathbf{f}_{ui}$. Finally, this fused vector is passed through a terminal Multi-Layer Perceptron (MLP), which acts as the prediction head. A sigmoid function, σ, is applied to the MLP's output logit to produce the final interaction probability, $\hat{y}_{ui}$:

$$\hat{y}_{ui} = \sigma(\text{MLP}_{\text{final}}(\mathbf{f}_{ui})) \tag{3}$$

3.3 Training Objective

We formulate the recommendation task as a binary classification problem for implicit feedback. The model is trained by optimizing a composite loss function, $\mathcal{L}_{\text{total}}$. The primary objective is the standard **Binary Cross-Entropy (BCE) Loss**, $\mathcal{L}_{\text{BCE}}$:

$$\mathcal{L}_{\text{BCE}} = -\sum_{u,i \in \mathcal{D}} (y_{ui} \log(\hat{y}_{ui}) + (1 - y_{ui}) \log(1 - \hat{y}_{ui})) \tag{4}$$

To explicitly encourage the model to learn aligned representations across modalities, an optional **Contrastive Loss** term, $\mathcal{L}_{\text{contrastive}}$, can be included. This loss pushes the projected visual embedding $\mathbf{p}_v$ and textual embedding $\mathbf{p}_t$ of the same item closer together in the latent space. The final loss function is a weighted sum:

$$\mathcal{L}_{\text{total}} = (1 - \alpha)\mathcal{L}_{\text{BCE}} + \alpha\mathcal{L}_{\text{contrastive}} \tag{5}$$

Here, α is a non-negative hyperparameter that controls the balance between the primary recommendation task and the auxiliary alignment task.

4 Experiments

To evaluate the effectiveness of our proposed model, we conducted a series of experiments on the large-scale public dataset PixelRec, which is available at `https://github.com/westlake-repl/PixelRec.git`. The framework was compared to several baseline implementations, and its performance was analyzed using several recommendation metrics. For our experiments, which were constrained by computational and hardware limitations, we used the PixelRec50K subset.

Statistic	**interactions.csv**
Number of Users	50000
Number of Items	82865
Total Interactions	989494
Average Number of Interactions per User	19.79
Average Number of Interactions per Item	11.94
Highest Number of Interactions by a User	434
Highest Number of Interactions for an Item	146
Density (%)	0.0239

Table 1: Dataset interaction description for **pixelRec50K**

4.1 Experimental Setup

For our main experiments, we followed the preprocessing steps outlined in the previous section. Filtering was applied to create a denser subset with at least 5 interactions per user and item. The resulting subset was subsequently split into training (70%), validation (15%), and testing (15%) sets. To assess the performance of the framework, we performed a Top-K retrieval task using a random negative sampling strategy of 20 items per positive instance. Performance was measured on a random 30% subset of the users.

Four baseline models were tested to establish a reference point:

- **Random:** Recommends items randomly. This serves as a lower-bound sanity check.
- **Most Popular:** Recommends the same global top-K most interacted-with items to all users.
- **User-kNN:** A collaborative filtering method that recommends items based on the preferences of similar users.
- **Item-kNN:** A collaborative filtering method that recommends items similar to those a user has previously interacted with.

Based on our framework, we established two hypotheses to guide our experiments:

1. The multimodal models will significantly outperform traditional methods like **Item-kNN**, **User-kNN**, and **Most Popular** in standard recommendation metrics.
2. While both visual and textual information can contribute to better recommendations, the textual data will provide a stronger and more representative signal for making effective recommendations.

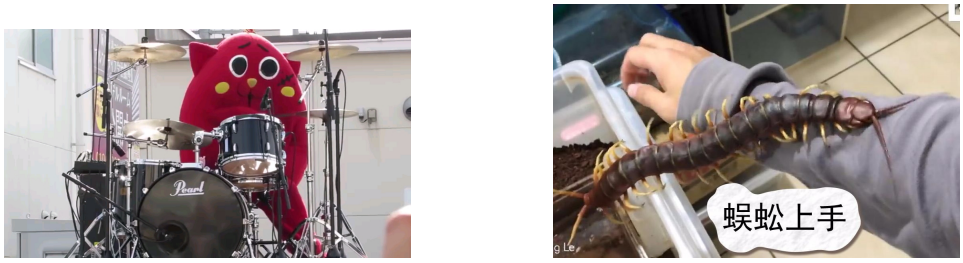
5 Results & Analysis

The performance of our multimodal framework compared to the baselines and across different encoder combinations is presented in Table 2. The results confirm our first hypothesis, with the multimodal models demonstrating superiority in Recall, NDCG, and MRR for both $k = 5$ and $k = 10$ compared to the baselines. The only metric where they did not lead was Novelty, narrowly losing to the **Random** baseline, which is expected given its nature. While the **Most Popular** baseline achieves high recall, it does so by sacrificing personalization, which is where the multimodal models excel.

Among the various combinations of visual and textual encoders, the pairing of **CLIP** for visual features and **Sentence-BERT (SBERT)** for textual features emerges as the clear winner. This model, highlighted in bold in the table, achieves the highest scores across nearly all metrics, with an NDCG@10 of 0.4247 and an MRR of 0.3178. This superior performance suggests a strong synergistic alignment between the vision-language pre-training of CLIP and the rich, sentence-level semantic understanding of SBERT, allowing the model to capture user preferences with greater nuance.

Method	R@5	NDCG@5	R@10	NDCG@10	MRR	Novelty
Baselines						
Random	0.2360	0.1373	0.4728	0.2166	0.1409	**0.9523**
MostPopular	0.4617	0.2913	0.7212	0.3786	0.2745	0.9271
Item-kNN	0.2348	0.1387	0.4932	0.2406	0.1659	0.9501
User-kNN	0.2580	0.1637	0.4860	0.2357	0.1616	0.9509
Multimodal Models						
CLIP+BERT	0.5105	0.3321	0.7670	0.4147	0.3076	0.9227
CLIP+SBERT	**0.5247**	**0.3435**	**0.7754**	**0.4247**	**0.3178**	0.9219
DINO+BERT	0.5172	0.3368	0.7685	0.4180	0.3112	0.9225
DINO+SBERT	0.5025	0.3260	0.7626	0.4098	0.3026	0.9232
ResNet+BERT	0.5123	0.3318	0.7669	0.4140	0.3066	0.9227
ResNet+SBERT	0.5186	0.3349	–	0.4159	0.3082	0.9224

Table 2: Top-k retrieval results comparing baselines with various multimodal encoder configurations on the **PixelRec50K** dataset. The best performing model is in **bold**.



(a) Image **i213993**, our **best result**, recommended with a **91.16%** probability for a true interaction.

(b) Image **i76860**, our **worst result**, acting as a strong distractor.

Figure 2: Examples of best and worst case predictions from the model.

5.1 Ablation Study

To better understand the individual contributions of the visual and textual modalities, we conducted an ablation study. We used the 'DINO+SBERT' configuration as a representative model and compared its performance against two variants: one without visual features (text-only with SBERT) and one without textual features (visual-only with DINO). The results are shown in Table 3.

The findings prove our second hypotheses correct. When the textual modality is removed, the metrics plummeted across all accuracy-based metrics. This is theorised its due to the human synthesis associated with the video title and description. The goal when writings this is to best summarise and describe the contents at its core. This human preprocessing and its consequential rich text is our theory for this stark result. More surprisingly, the visual modally decreases performance across the board. This suggest that the images add noise to the system, where the information is not specially relevant. This could be due to the fact that the dataset provides the thumbnail of videos. Where the image present in it could not be totally representative of the actual video, whereas the text is. Furthermore the image from the thumbnail is susceptible of clickbait. In summary the visual features from DINO, when paired with SBERT, appear to introduce more noise than valuable information, slightly degrading the overall performance.

Model Variant	R@5	NDCG@5	R@10	NDCG@10	MRR	Novelty
DINO+SBERT (Full)	0.5025	0.3260	0.7626	0.4098	0.3026	0.9232
w/o visual (Text-only)	**0.5088**	**0.3318**	**0.7715**	**0.4167**	**0.3088**	0.9222
w/o text (Visual-only)	0.3879	0.2363	0.6542	0.3219	0.2220	**0.9339**

Table 3: Ablation study on the DINO+SBERT model configuration to assess the impact of each modality. The best result in each column is in **bold**.

6 Conclusions and Future Work

The results validate our initial hypotheses, demonstrating the superiority of a multimodal model over traditional baselines. The fusion of features using a Multi-Head Self-Attention mechanism proved effective, with the combination of CLIP and Sentence-BERT encoders yielding the best performance. Our ablation study further reveals that textual features were the most influential modality in this dataset, providing a stronger signal for recommendation than visual features.

As future work, we plan to conduct a more comprehensive ablation study is needed to analyse the impact of feature dimensions, the inclusion of tag data, and other metadata. Furthermore, evaluating the framework on different datasets is crucial to test the generalizability of our findings regarding the dominance of textual features. The implement more sophisticated negative sampling strategies like hard negative sampling on content similarity to further strengthen the training process. A qualitative analysis could investigate the "clickbait thumbnail" hypothesis by examining cases where visual and textual information conflict, providing deeper insights into model failures and successes.

In conclusion, this work provides a robust framework for building multimodal recommenders and underscores the critical importance of high-quality textual data, which can often be the deciding factor in creating truly personalized visual content experiences.

References

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining (ICDM '08)*. IEEE, 2008.

Q. Liu, Y. Zhang, C. Chen, Z. Wang, G. Li, and Q. Lu. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 56(9), 2024.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang. One-class collaborative filtering. In *2008 Eighth IEEE International Conference on Data Mining*, pages 502–511, 2008. doi: 10.1109/ICDM.2008.16.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR, 2021.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*. AUAI Press, 2009.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.