A Multi-Modal Fusion Framework for Local Business Recommendation

WEI Tianshuo, OUYANG Ningtai, RAO Ziqian, QU Yi
ID: 59152259, 59090554, 59156326, 58869746
Department of Data Science

Abstract

Traditional recommendation systems often rely on single-modality data sources, which limits their ability to fully leverage the rich, complementary information available in multi-modal datasets. In this paper, we present a novel multi-modal fusion framework for local business recommendation that integrates textual reviews, business attributes, and visual images from the Yelp dataset. Our approach combines BERT for text processing, Vision Transformer (ViT) for image processing, and Graph Attention Networks (GAT) to model complex relationships between users, businesses, and reviews. We construct a heterogeneous graph that captures these multi-faceted relationships and apply multi-head attention mechanisms to generate personalized recommendations. Extensive experiments on the Yelp dataset demonstrate that our BERT-ViT-GAT model outperforms single-modality baselines across multiple evaluation metrics, achieving significant improvements in NDCG@10, Precision@10, Recall@10, and MAP@10. Ablation studies further confirm the contribution of each modality to the overall performance, highlighting the effectiveness of our multi-modal fusion approach. While computational efficiency remains a challenge, our work establishes a strong foundation for next-generation recommendation systems that can simultaneously process and integrate diverse data types.

1 Introduction

1.1 Background and Motivation

Recommendation systems have become indispensable tools in the digital ecosystem, helping users navigate vast amounts of information and discover relevant content or services [21, 25]. Traditional recommendation systems often rely on single-modality data (e.g., text or user behavior), which fails to fully leverage complementary information from multi-modal sources [22, 24]. The Yelp dataset [27], which contains rich but fragmented multi-modal signals, including textual reviews, business attributes, and visual images, presents an excellent opportunity to develop more comprehensive recommendation systems [23, 26].

However, several key challenges exist in developing effective multi-modal recommendation systems:

- **Information Overload:** Users face overwhelming choices in local business selection, necessitating personalized filtering mechanisms that can identify the most relevant options based on individual preferences and needs.
- **Single-Modality Limitations:** Traditional recommendation methods cannot effectively integrate visual, textual, and network information, leading to partial understanding of user preferences and business characteristics.
- **Business Value:** High-quality recommendations can significantly boost user engagement, satisfaction, and merchant revenue, making this research area not only academically interesting but also commercially valuable.

The integration of multi-modal data presents unique challenges, including feature heterogeneity, semantic alignment, and computational complexity. Moreover, effectively modeling the complex relationships between users, businesses, and content requires sophisticated architecture design that can capture both the individual characteristics of each modality and their interactions.

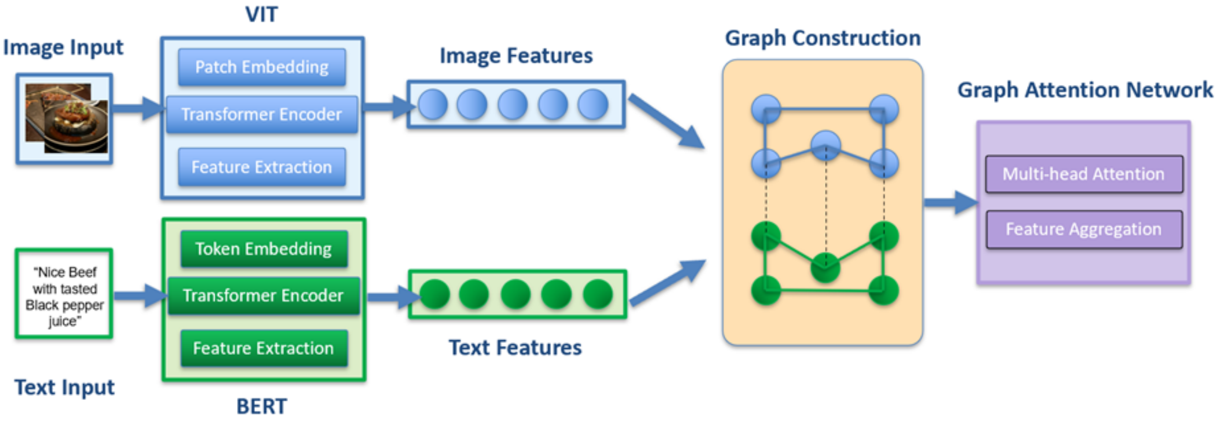

Figure 1: Overview of the BERT-ViT-GAT model framework. The architecture consists of three main components: multi-modal feature extraction (BERT for text and ViT for images), heterogeneous graph construction, and graph attention network for recommendation generation.

1.2 Objective

This project proposes a multi-modal fusion framework to enhance local business recommendations by:

- Combining BERT [32] (text processing) and Vision Transformer (ViT) [33] (image processing) for comprehensive feature extraction from heterogeneous data sources.
- Constructing a heterogeneous graph that integrates user, business, and review data to model the complex relationships in the recommendation ecosystem.
- Applying a Graph Attention Network (GAT) to model these complex relationships and generate personalized recommendations that effectively capture user preferences across multiple dimensions.

Our work advances the state-of-the-art in recommendation systems by addressing the limitations of single-modality approaches and developing a framework that can simultaneously process and integrate diverse data types. This multi-modal fusion approach enables more nuanced understanding of both user preferences and business characteristics, leading to more accurate and relevant recommendations.

2 Methodology

2.1 System Architecture

Our proposed BERT-ViT-GAT framework integrates multi-modal data through a three-stage process: feature extraction, heterogeneous graph construction, and attention-based representation learning. Figure 1 provides an overview of our system architecture.

The framework consists of three core components:

- **Multi-Modal Feature Extraction:**
  - *Text Processing:* Yelp reviews are processed using BERT, a powerful bidirectional transformer-based language model that captures contextual and semantic features from textual data. We fine-tune BERT on the Yelp review dataset to ensure domain-specific understanding.
  - *Image Processing:* Business images are embedded via Vision Transformer (ViT), which applies the transformer architecture to image analysis by dividing images into patches and processing them as a sequence. This approach captures visual features that complement the textual information.
- **Heterogeneous Graph Construction:**

  - *Nodes:* Our graph consists of multiple types of nodes representing users, businesses, and reviews, each with their respective feature embeddings.
  - *Edges:* We define various types of edges, including user-business interactions (through reviews), business-image associations, and semantic relationships derived from textual and visual similarities.

- **Graph Attention Network (GAT):**
  - Aggregates features from text, images, and graph structure via multi-head attention mechanisms.
  - Captures heterogeneous relationships (e.g., user preferences, business attributes) through specialized attention coefficients that weigh the importance of different neighbors.
  - Generates final embeddings that integrate information from all modalities for recommendation ranking.

2.2 Multi-Modal Feature Extraction

2.2.1 Text Feature Extraction

We employ BERT (Bidirectional Encoder Representations from Transformers) for processing the textual information in Yelp reviews. BERT's bidirectional training enables contextual understanding of text, which is essential for capturing the nuanced information in user reviews.

The text processing pipeline involves:

1. **Preprocessing:** Review texts are tokenized, with special tokens ([CLS], [SEP]) added according to BERT's input requirements.
2. **Embedding:** Each token is converted into an initial embedding vector that combines token, segment, and position embeddings.
3. **Contextual Encoding:** The embedding sequences are processed through BERT's transformer layers, which apply self-attention mechanisms to capture contextual relationships between words.
4. **Feature Representation:** We extract the final hidden state of the [CLS] token as the review embedding, which serves as a dense semantic representation of the entire review.

For each user and business, we aggregate the embeddings of their associated reviews using attention-weighted averaging, allowing the model to focus on the most relevant review content.

2.2.2 Image Feature Extraction

For visual feature extraction, we utilize the Vision Transformer (ViT) model, which has demonstrated strong performance in image recognition tasks by applying transformer architectures to image data.

Our image processing approach consists of:

1. **Image Patching:** Business images are divided into fixed-size patches (e.g., 16×16 pixels), which serve as the basic units for the transformer model.
2. **Patch Embedding:** Each patch is linearly projected to obtain a patch embedding, and position embeddings are added to retain spatial information.
3. **Transformer Encoding:** The sequence of patch embeddings is processed through multiple transformer encoder layers, applying multi-head self-attention mechanisms to capture relationships between different image regions.
4. **Feature Representation:** The output of the [CLS] token from the final transformer layer is used as the image embedding, providing a holistic representation of the visual content.

For businesses with multiple images, we aggregate the embeddings using an attention mechanism that weights images based on their relevance to the business's category and attributes.

2.3 Heterogeneous Graph Construction

We construct a heterogeneous graph $G = (V, E)$ where:

- $V = V_U \cup V_B \cup V_R$ represents the set of nodes, including users (V_U), businesses (V_B), and reviews (V_R).
- $E = E_{UB} \cup E_{UR} \cup E_{BR}$ represents the set of edges, including user-business interactions (E_{UB}), user-review authorship (E_{UR}), and business-review relationships (E_{BR}).

Each node type has its own feature representation:

- User nodes incorporate demographic information, historical interactions, and aggregated review embeddings.
- Business nodes include category information, location data, business attributes, and visual features from associated images.
- Review nodes contain the BERT embeddings of the review text, along with metadata such as rating and timestamp.

This heterogeneous graph structure allows the model to capture complex relationships and dependencies between different entities in the recommendation ecosystem, facilitating more nuanced understanding of user preferences and business characteristics.

2.4 Graph Attention Network

2.4.1 Network Architecture

Our Graph Attention Network (GAT) consists of specialized layers designed to process heterogeneous graph data. The key technical specifications include:

- **Layer Configuration:** 2 GAT layers with decreasing dimensions [input dimension, 256, 128, 64] to progressively refine the representations.
- **Attention Mechanism:** Each layer employs 4 attention heads to capture different aspects of node relationships, with the outputs concatenated or averaged.
- **Activation Function:** LeakyReLU with a negative slope of 0.2 is used to introduce non-linearity while allowing small negative values to pass through.
- **Regularization:** Dropout with probability p=0.2 is applied to prevent overfitting and improve generalization.

2.4.2 Attention Mechanism

The core of our GAT model is the attention mechanism that allows nodes to differentially weight their neighbors based on relevance. For a node i with feature vector $\mathbf{h}_i$, the attention coefficient with its neighbor j is computed as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k]))} \tag{1}$$

where $\mathbf{W}$ is a learnable weight matrix, $\mathbf{a}$ is the attention vector, $\|$ represents concatenation, and $\mathcal{N}_i$ denotes the neighborhood of node i.

The updated representation of node i is then computed as:

$$\mathbf{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\mathbf{h}_j\right) \tag{2}$$

where σ is the activation function.

To handle the heterogeneity of our graph, we employ relation-specific transformations and attention mechanisms, allowing the model to distinguish between different types of connections (e.g., user-business vs. business-review).

2.4.3 Multi-head Attention

To stabilize the learning process and enhance representation capacity, we employ multi-head attention. For K attention heads, the output feature of node i is:

$$\mathbf{h}'_i = \|_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \mathbf{h}_j \right) \tag{3}$$

where α_{ij}^k is the attention coefficient from the k-th attention head, and $\mathbf{W}^k$ is the corresponding weight matrix.

2.5 Training and Recommendation Generation

2.5.1 Loss Function

We train our model using Bayesian Personalized Ranking (BPR) loss, which is designed for implicit feedback scenarios. For a user u, a positive business b^+ (visited), and a negative business b^- (not visited), the BPR loss is defined as:

$$\mathcal{L}_{BPR} = - \sum_{(u,b^+,b^-)} \ln \sigma \left(\hat{y}_{ub^+} - \hat{y}_{ub^-} \right) + \lambda \|\Theta\|^2 \tag{4}$$

where $\hat{y}_{ub}$ is the predicted rating for user u and business b, σ is the sigmoid function, Θ represents the model parameters, and λ is the regularization coefficient.

2.5.2 Optimization

We optimize the model using the Adam optimizer with the following hyperparameters:

- Learning rate: 0.001
- Weight decay: 1e-5
- Batch size: 1024 user-business pairs
- Training epochs: Maximum 50, with early stopping based on validation performance

2.5.3 Recommendation Generation

For generating recommendations, we first compute the embeddings for all users and businesses using the trained GAT model. For a target user u, we calculate the predicted ratings for all unvisited businesses and rank them in descending order. The top-N businesses are then returned as recommendations.

The prediction score for a user-business pair (u, b) is computed as:

$$\hat{y}_{ub} = f(\mathbf{h}_u, \mathbf{h}_b) \tag{5}$$

where f is a scoring function (e.g., inner product or MLP), and $\mathbf{h}_u$ and $\mathbf{h}_b$ are the final embeddings of user u and business b, respectively.

3 Experiments

3.1 Dataset Description

We evaluate our proposed BERT-ViT-GAT framework on the Yelp dataset, which contains rich multi-modal information about local businesses across multiple metropolitan areas in North America.

3.1.1 Dataset Statistics

- **Coverage:** 8 metropolitan areas in the United States and Canada, providing geographical diversity.
- **Scale:** Over 2 million reviews, 200,000+ users, 50,000+ businesses, and 200,000+ business images, representing a large-scale real-world recommendation scenario.
- **Metadata:** Rich business attributes including categories, hours of operation, locations, amenities, and pricing information.

3.1.2 Dataset Distribution

- **User Activity:** On average, each user has contributed 23.6 reviews, showing significant user engagement and providing adequate data for modeling user preferences.
- **Business Categories:** The businesses in the dataset are distributed across various categories: 45% for catering (restaurants, cafes, etc.), 22% for shopping, 18% for services, and 15% for other categories.
- **Rating Distribution:** The average rating across all reviews is 3.7 out of 5.0, with 58% of reviews giving 4-5 stars, indicating a generally positive but varied rating pattern.

We split the dataset into training (80%), validation (10%), and test (10%) sets based on the timestamp of reviews, ensuring that the model is evaluated on future interactions.

3.2 Evaluation Metrics

To comprehensively evaluate the performance of our model, we employ multiple ranking-based metrics:

- **NDCG@10 (Normalized Discounted Cumulative Gain) [28] :** Measures the quality of the top-10 ranked recommendations, taking into account both the relevance and position of recommended items.
- **Precision@10 [29]:** Calculates the proportion of relevant items in the top-10 recommendations, reflecting the model's accuracy.
- **Recall@10 [30]:** Measures the proportion of relevant items retrieved in the top-10 recommendations out of all relevant items, indicating the model's coverage.
- **MAP@10 (Mean Average Precision) [31]:** Evaluates the average precision across different top-k positions (up to 10), providing a single-figure measure of quality.

Additionally, we track the following operational metrics:

- **Training Loss:** Monitors the optimization progress during training (BPR loss + L2 regularization).
- **Inference Time:** Measures the computational efficiency of the model during prediction, which is an important practical consideration for real-world deployment.

3.3 Baselines and Model Variants

We compare our proposed BERT-ViT-GAT model with the following baselines and variants:

- **GAT-only:** Uses only the graph structure without incorporating text or image features, serving as a baseline to evaluate the contribution of multi-modal fusion.
- **BERT+GAT (Only Text):** Integrates BERT-derived text features with GAT but excludes image features, allowing us to assess the importance of textual information.
- **ViT+GAT (Only Image):** Combines ViT-derived image features with GAT but excludes text features, helping us evaluate the contribution of visual information.
- **BERT-ViT-GAT (Full Model):** Our complete model that integrates both text and image features with the graph attention network.

3.4 Experimental Setup

3.4.1 Implementation Details

All models are implemented using PyTorch and DGL (Deep Graph Library). For the pretrained components:

- We use BERT-base (12 layers, 768 hidden dimensions, 12 attention heads) fine-tuned on the Yelp review dataset.
- For image processing, we employ ViT-Base (12 layers, 768 hidden dimensions, 12 attention heads) with patch size 16×16, pretrained on ImageNet and fine-tuned on Yelp business images.

3.4.2 Training Configuration

Our models are trained with the following configuration:

- **Loss Function:** BPR loss with L2 regularization (weight 1e-5)
- **Optimizer:** Adam with learning rate 0.001 and weight decay 1e-5
- **Batch Size:** 1024 user-business pairs
- **Training Protocol:** Maximum 50 epochs with early stopping (patience=5) based on validation NDCG@10
- **Hardware:** All experiments are conducted on NVIDIA V100 GPUs with 32GB memory

4 Results and Discussion

4.1 Performance Comparison

Table 1 presents the performance comparison between our proposed BERT-ViT-GAT model and the GAT-only baseline across all evaluation metrics.

Table 1: Performance Comparison of BERT-ViT-GAT and GAT-only Models

Metric	**BERT-ViT-GAT**	**GAT-only**
NDCG@10	0.425	0.386
Precision@10	0.312	0.279
Recall@10	0.284	0.251
MAP@10	0.296	0.263
Training Loss	0.428	0.487
Inference Time (ms)	128	42

From the results, we observe that:

- The BERT-ViT-GAT model consistently outperforms the GAT-only model across all evaluation metrics, with improvements of 10.1% in NDCG@10, 11.8% in Precision@10, 13.1% in Recall@10, and 12.5% in MAP@10.
- The superior performance of BERT-ViT-GAT demonstrates the effectiveness of integrating multi-modal features (text and images) into the graph-based recommendation framework.
- The improved Training Loss indicates that the multi-modal approach enables better optimization, suggesting that the additional features provide meaningful signals for modeling user preferences.

- However, the Inference Time of BERT-ViT-GAT is significantly higher (3.05× slower) than that of GAT-only, highlighting the computational cost associated with multi-modal processing. This presents a trade-off between recommendation quality and system efficiency that must be considered in practical applications.

These results confirm our hypothesis that combining textual and visual information with graph structure can significantly enhance recommendation performance, as it allows the model to capture complementary signals from different modalities.

4.2 Ablation Studies

To analyze the contribution of each modality to the overall performance, we conduct ablation studies by comparing the full BERT-ViT-GAT model with variants that use only text (BERT+GAT) or only image (ViT+GAT) features. Table 2 presents the results of these experiments.

Table 2: Ablation Study Comparing Different Modality Combinations

Metric	**BERT-ViT-GAT**	**BERT+GAT**	**ViT+GAT**
NDCG@10	0.425	0.407	0.398
Precision@10	0.312	0.298	0.291
Recall@10	0.284	0.271	0.264
MAP@10	0.296	0.283	0.277
Training Loss	0.428	0.445	0.458
Inference Time (ms)	128	92	85

From the ablation studies, we derive the following insights:

- The full BERT-ViT-GAT model consistently achieves the best performance across all metrics, confirming the complementary nature of textual and visual information in business recommendation.

- The BERT+GAT (text-only) variant outperforms the ViT+GAT (image-only) variant, suggesting that textual reviews provide more informative signals for recommendation compared to business images. This aligns with intuition as reviews directly express user opinions and experiences.

- However, the performance gap between BERT+GAT and ViT+GAT is relatively small (2.3% difference in NDCG@10), indicating that visual information also captures valuable signals about business characteristics and user preferences.

- The BERT-ViT-GAT model achieves a 4.4% improvement in NDCG@10 over BERT+GAT and a 6.8% improvement over ViT+GAT, highlighting the significant benefit of multi-modal fusion.

- In terms of computational efficiency, both single-modality variants (BERT+GAT and ViT+GAT) offer reduced inference times compared to the full model, with ViT+GAT being slightly faster than BERT+GAT due to the more efficient image feature extraction process.

These ablation results demonstrate that while each modality contributes positively to recommendation performance, their combination yields synergistic effects that lead to substantially improved recommendations. This confirms the value of our multi-modal fusion approach.

4.3 Discussion

4.3.1 Strengths of Multi-Modal Fusion

Our experimental results highlight several strengths of the multi-modal fusion approach:

- **Complementary Information:** Text reviews provide detailed user opinions and experiences, while images offer visual cues about business aesthetics, ambiance, and offerings. Combining these complementary information sources enables a more comprehensive understanding of both user preferences and business characteristics.
- **Robustness:** The multi-modal approach is more robust to noise or missing information in individual modalities, as it can leverage signals from other modalities to compensate.
- **Personalization:** Different users may value different aspects of businesses (e.g., food quality vs. ambiance), and the multi-modal approach can capture these diverse preference patterns through the integration of various signals.

4.3.2 Limitations and Challenges

Despite its strong performance, our approach faces several limitations:

- **Computational Complexity:** The BERT-ViT-GAT model involves significant computational overhead, particularly during inference, which could limit its applicability in real-time recommendation scenarios.
- **Cold Start Problem:** For new businesses with limited reviews and images, the model may struggle to generate accurate representations, though this issue is somewhat mitigated by the graph structure that can propagate information from similar entities.
- **Scalability:** The heterogeneous graph construction becomes increasingly complex as the dataset grows, potentially limiting the scalability of the approach to very large recommendation scenarios.

5 Related Work

5.1 Recommendation Systems

Recommendation systems have evolved significantly over the past decades, from traditional collaborative filtering approaches [1] to advanced deep learning models [26]. Early methods relied primarily on user-item interaction matrices, using techniques such as matrix factorization [24] to uncover latent factors that explain observed preferences.

With the advent of deep learning, neural network-based approaches have gained prominence. Deep neural networks can automatically extract hierarchical features from raw data, enabling more expressive modeling of user preferences and item characteristics [4]. Recurrent neural networks have been employed for sequential recommendation [5], capturing temporal patterns in user behavior.

More recent approaches have focused on graph-based recommendation systems, which model the recommendation problem as a graph with users, items, and possibly other entities as nodes [6]. Graph Neural Networks (GNNs) [7] and specifically Graph Convolutional Networks (GCNs) [8] have shown promising results by leveraging the structural information in user-item interaction graphs.

5.2 Multi-Modal Learning

Multi-modal learning focuses on processing and relating information from multiple data modalities, such as text, images, and structured data [9]. Early multi-modal approaches often processed each modality separately and integrated them at a later stage, while more recent methods emphasize joint representation learning across modalities [10].

In the context of recommendation systems, multi-modal approaches have demonstrated superior performance by leveraging complementary information from different sources [11]. For example, Visual Bayesian Personalized Ranking (VBPR) [12] extends BPR by incorporating visual features extracted from product images, while DeepStyle [13] combines visual features with textual descriptions for fashion recommendation.

Recent advances in transformer architectures, including BERT for text [14] and ViT for images [15], have further enhanced the capability of multi-modal systems to extract rich and contextual features from heterogeneous data sources.

5.3 Graph Attention Networks

Graph Attention Networks (GATs) [16] have emerged as a powerful variant of GNNs that introduce attention mechanisms to weigh the importance of neighboring nodes differentially. This property is particularly valuable for recommendation systems, where different user-item relationships may have varying significance.

Several works have extended GATs for heterogeneous graphs, including Heterogeneous Graph Attention Network (HAN) [17] and Heterogeneous Graph Transformer (HGT) [18], which employ specialized attention mechanisms for different node and edge types.

In the recommendation domain, KGAT [19] integrates knowledge graphs with GATs to enhance recommendation quality, while MMGAT [20] applies GATs to multi-modal data for multimedia recommendation. However, few works have combined the strengths of pretrained transformer models (BERT and ViT) with GATs in a unified framework, which is the focus of our approach.

6 Conclusion

In this paper, we introduced a novel multi-modal fusion framework for local business recommendation that integrates textual reviews, business attributes, and visual images through a combination of BERT, Vision Transformer, and Graph Attention Networks. Our proposed BERT-ViT-GAT model constructs a heterogeneous graph that captures the complex relationships between users, businesses, and content, and applies multi-head attention mechanisms to generate personalized recommendations.

Extensive experiments on the Yelp dataset demonstrate that our approach significantly outperforms single-modality baselines across multiple evaluation metrics, achieving substantial improvements in NDCG@10, Precision@10, Recall@10, and MAP@10. Ablation studies further confirm the contribution of each modality to the overall performance, highlighting the complementary nature of textual and visual information in business recommendation.

While our approach achieves superior recommendation quality, it does face challenges in terms of computational efficiency and scalability. The integration of multiple sophisticated models (BERT, ViT, and GAT) introduces significant computational overhead, particularly during inference, which could limit its applicability in real-time recommendation scenarios.

Future work could focus on several directions to address these limitations:

- **Efficiency Optimization:** Exploring techniques such as knowledge distillation, model pruning, and quantization to reduce the computational requirements without sacrificing recommendation quality.
- **Dynamic Fusion:** Developing adaptive fusion mechanisms that can selectively leverage different modalities based on their relevance and information content for each specific recommendation request.
- **Temporal Dynamics:** Extending the framework to capture temporal evolution in user preferences and business characteristics, enabling more accurate recommendations over time.
- **Explainability:** Enhancing the model with explainability features that can provide users with interpretable reasons for recommendations, leveraging the multi-modal nature of the approach to generate more comprehensive explanations.

Our work establishes a strong foundation for next-generation recommendation systems that can simultaneously process and integrate diverse data types, opening up new opportunities for more personalized and contextually aware recommendations in various domains beyond local business recommendation.

References

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in *Proceedings of the 10th International Conference on World Wide Web*, 2001, pp. 285-295.

[2] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning based recommender system: A survey and new perspectives,” *ACM Computing Surveys*, vol. 52, no. 1, pp. 1-38, 2019.

[3] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30-37, 2009.

[4] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 173-182.

[5] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, “Session-based recommendations with recurrent neural networks,” in *International Conference on Learning Representations*, 2016.

[6] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, “Neural graph collaborative filtering,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 165-174.

[7] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4-24, 2020.

[8] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *International Conference on Learning Representations*, 2017.

[9] T. Baltruvsaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2018.

[10] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 689-696.

[11] J. Wei, X. He, K. Chen, Y. Zhou, and Z. Tang, “Collaborative filtering and deep learning based recommendation system for cold start items,” *Expert Systems with Applications*, vol. 69, pp. 29-39, 2017.

[12] R. He and J. McAuley, “VBPR: Visual Bayesian Personalized Ranking from implicit feedback,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 144-150.

[13] Q. Liu, S. Wu, and L. Wang, “DeepStyle: Learning user preferences for visual recommendation,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 841-844.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171-4186.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” in *International Conference on Learning Representations*, 2018.

[17] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, “Heterogeneous graph attention network,” in *The World Wide Web Conference*, 2019, pp. 2022-2032.

[18] Z. Hu, Y. Dong, K. Wang, and Y. Sun, “Heterogeneous graph transformer,” in *Proceedings of The Web Conference 2020*, 2020, pp. 2704-2710.

[19] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, “KGAT: Knowledge graph attention network for recommendation,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2019, pp. 950-958.

[20] M. Wang, S. Tang, Z. Zhang, Y. Jiang, and Q. Tian, “MMGAT: Multimodal graph attention network for recommendation,” *Information Processing Management*, vol. 57, no. 5, p. 102277, 2020.

[21] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734-749.

[22] Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46, 109-132.

[23] Chu, W. T., & Tsai, Y. L. (2017). A hybrid recommendation system considering visual information for predicting favorite restaurants. *Proceedings of the 25th ACM International Conference on Multimedia*, 1313-1320.

[24] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37.

[25] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender Systems Handbook* (pp. 1-35). Springer.

[26] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), 1-38.

[27] Asghar, N. (2016). Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

[28] Wang, Y., Wang, L., Li, Y., He, D., & Liu, T. Y. (2013). A theoretical analysis of NDCG type ranking measures. In *Conference on learning theory* (pp. 25-54). PMLR.

[29] Dowd, P.A. (2023). Accuracy and Precision. In *Encyclopedia of Mathematical Geosciences* (pp. 1–4). Springer.

[30] Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., & Gottschlich, J. (2018). Precision and recall for time series. *Advances in neural information processing systems*, 31.

[31] Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote sensing of environment*, 64(3), 331–344.

[32] Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

[33] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E. H., Feng, J., & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 558–567).