Article

Research on Feature Fusion and Multimodal Patent Text Based on Graph Attention Network

Zhenzhen Song [1,*], Ziwei Liu [2] and Hongji Li [3]

[1] School of Language and Culture, Northwest A&F University, Shaanxi, China
[2] University of Illinois Urbana-Champaign, Urbana, USA
[3] Columbia University, New York, USA
[*] Correspondence: Zhenzhen Song, School of Language and Culture, Northwest A&F University, Shaanxi, China

Abstract: Aiming at the challenges of cross-modal feature fusion, low computational efficiency in long patent text modeling, and insufficient hierarchical semantic coherence in patent text semantic mining, this study proposes a novel deep learning framework termed HGM-Net. The framework integrates Hierarchical Comparative Learning (HCL), a Multi-modal Graph Attention Network (M-GAT), and Multi-Granularity Sparse Attention (MSA) to achieve robust, efficient, and semantically consistent patent representation learning. Specifically, HCL introduces dynamic masking, contrastive learning, and cross-structural similarity constraints across word-, sentence-, and paragraph-level hierarchies, enabling the model to jointly capture fine-grained local semantics and high-level thematic consistency. Contrastive and cross-structural similarity constraints are particularly enforced at the word and paragraph levels, effectively enhancing semantic discrimination and global coherence within complex patent documents. Furthermore, M-GAT models patent classification codes, citation relationships, and textual semantics as heterogeneous graph structures, and employs cross-modal gated attention mechanisms to dynamically fuse multi-source and multi-modal features, thereby improving representation completeness and robustness. To address the high computational cost of long-text processing, MSA adopts a hierarchical sparse attention strategy that selectively allocates attention across multiple granularities, including words, phrases, sentences, and paragraphs, significantly reducing computational overhead while preserving critical semantic information. Extensive experimental evaluations on patent classification and similarity matching tasks demonstrate that HGM-Net consistently outperforms existing state-of-the-art deep learning approaches. The results validate the effectiveness and generalization capability of the proposed framework, highlighting its theoretical innovation and practical value in improving patent examination efficiency and enabling large-scale technology relevance mining.

Keywords: hierarchical comparative learning; multimodal graph attention networks; multi-granularity sparse attention; patent semantic mining

Received: 28 November 2025
Revised: 15 January 2026
Accepted: 29 January 2026
Published: 03 February 2026

1. Introduction

Amid intensifying global technological competition, the efficiency of patent examination has emerged as a vital benchmark of national innovation systems. According to the World Intellectual Property Organization (WIPO), global patent applications surpassed 3.5 million in 2022. However, the average duration of substantive examination remains as long as 26.3 months, with nearly 40% of delays attributed to the complexity of evaluating semantic similarities in patent texts [1,2].

Recent years have witnessed growing interest in leveraging deep learning for patent text analysis and prediction. In China, Yu et al. proposed a BERT-based framework that integrates models such as DeBERTa-v3 and ELECTRA using a weighted strategy to improve semantic similarity matching [3,4]. Their V3 pre-processing method, employing structured tokens like [CLS] and [SEP], enhances semantic representation and has shown promise in Cooperative Patent Classification (CPC) tasks. Chen et al. addressed cross-lingual patent matching by introducing a conceptual bridging strategy using Latent Semantic Indexing (LSI) to construct multilingual vectors based on International Patent Classification (IPC), improving multilingual fusion [5]. Other works incorporate LSTM with attention mechanisms and CNNs with word embeddings, facilitating multi-level feature extraction for better prediction [6].

Internationally, research has emphasized multimodal feature fusion and model optimization [7]. Verberne et al. introduced a CRF-Flair based sequence annotation approach for citation extraction from full-text patents, aided by regular expression-based entity recognition. Yung-Chang Chi et al. achieved 87.7% accuracy in predicting infringement and review outcomes using CNN-LSTM models trained on USPTO data [8]. Further, Ha and Lee explored patent embeddings to enhance CPC modeling, while Adversarial Weight Perturbation (AWP) and hierarchical self-attention have proven effective for modeling long texts and structured hierarchies.

Recent studies have demonstrated that unified multimodal modeling frameworks, which jointly encode heterogeneous data sources into a shared semantic space, can significantly enhance representation robustness and cross-task generalization. For example, Uni-FinLLM integrates time series, textual information, macro-level indicators, and graph-structured relations through attention-based multimodal fusion, highlighting the effectiveness of unified cross-modal architectures in complex semantic reasoning tasks [9].

Building on these developments, this paper proposes a novel deep learning framework, HGM-Net, which integrates: (1) Hierarchical Contrastive Learning (HCL) for semantic enhancement; (2) Multimodal Graph Attention Networks (M-GAT) for feature fusion; and (3) Multi-Granularity Sparse Attention (MSA) for long-text modeling.

2. Methodology

2.1. Hierarchical Comparative Learning

In the study of Hierarchical Contrastive Learning (HCL)-driven semantic enhancement for patents, we propose a multi-level contrastive learning framework to optimize both local semantic features and global structural representations of patent texts. The framework operates across three hierarchical levels: word-level, sentence-level, and paragraph-level, each designed to capture distinct granularities of patent semantics. Given a patent text sequence $X = \{x_1, x_2, \ldots, x_n\}$, the model first generates initial embeddings $H^{(0)} = \text{Transformer}(X) \in \mathbb{R}^{n \times d}$ through a bidirectional Transformer encoder, where d denotes the hidden dimension.

At the word-level contrastive layer, a dynamic masking strategy generates augmented samples $\tilde{X}_w$ by randomly replacing 15% of technical terms with synonyms from a domain-specific lexicon, forming positive pairs $(X, \tilde{X}_w)$. The contrastive loss for this level is formulated as:

$$\mathcal{L}_w = -\log \frac{\exp\left(\frac{s\left(h_i, h_i^+\right)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{s\left(h_j, h_j^-\right)}{\tau}\right)} \tag{1}$$

where $s(\cdot)$ is the cosine similarity function, τ is the temperature hyperparameter, h_i represents the original word embedding, h_i^+ denotes the embedding of the augmented sample at the same position, and h_j^- are embeddings sampled from a negative queue.

The sentence-level contrastive layer incorporates structural relationships inherent in patent documents, such as the correspondence between claims and embodiment descriptions. A sentence-to-sentence attention mechanism computes a semantic similarity matrix $A = \text{softmax}\left(QK^T/\sqrt{d}\right)$, where Q and K are query and key vectors derived from sentence embeddings of different structural units. The contrastive objective is defined as:

$$\mathcal{L}_s = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}(y_m = 1) \cdot D_{\text{KL}}(p_m \parallel q_m) \quad (2)$$

where M is the number of sentence pairs, D_{KL} is the Kullback-Leibler divergence, p_m represents the attention-based similarity distribution, and q_m corresponds to a binary annotation-derived distribution.

For paragraph-level contrastive learning, a multi-view alignment approach is employed to harmonize representations across distinct sections (e.g., abstract, claims, and detailed description) of the same patent. A prototype contrastive loss is introduced:

$$\mathcal{L}_p = \sum_{c=1}^{C} \left\| \mu_c - \frac{1}{|\mathcal{P}_c|}\sum_{x\in\mathcal{P}_c} f(x) \right\|_2^2 \quad (3)$$

where μ_c denotes the prototype vector for category c, $\mathcal{P}_c$ is the set of patent paragraphs belonging to category c, and $f(x)$ is the encoded paragraph embedding. A gradient stopping mechanism is applied to prevent rapid convergence of prototype vectors, thereby preserving discriminative features across hierarchical levels.

The HCL module integrates these hierarchical objectives through adaptive loss weighting:

$$\mathcal{L}_{\text{HCL}} = \alpha\mathcal{L}_w + \beta\mathcal{L}_s + \gamma\mathcal{L}_p \quad (4)$$

where α, β, γ are learnable temperature coefficients that dynamically balance the contributions of each contrastive level. This hierarchical architecture ensures simultaneous enhancement of fine-grained terminological semantics, inter-sentence structural coherence, and cross-paragraph thematic consistency in patent representation learning.

2.2. Feature Fusion Architecture for Multimodal Graph Attention Networks

In the investigation of the Multimodal Graph Attention Network (M-GAT) feature fusion architecture, we propose a heterogeneous graph attention framework to integrate multimodal features in patent documents, including structured classification codes (CPC), unstructured semantic descriptions, and cross-patent citation relationships [10]. Given a patent corpus $\mathcal{D} = \{D_1, D_2, \dots, D_N\}$, each patent D_i is modeled as a multimodal heterogeneous graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathcal{M}_i)$, where the node set $\mathcal{V}_i = \mathcal{V}_i^{\text{text}} \cup \mathcal{V}_i^{\text{cpc}} \cup \mathcal{V}_i^{\text{cite}}$ comprises text-based semantic units, CPC classification nodes, and citation relationship nodes. Edges $\mathcal{E}_i$ encode semantic associations, hierarchical classification dependencies, and citation strengths, while $\mathcal{M}_i = \{\text{text, cpc, cite}\}$ represents distinct feature spaces.

Node Representation Initialization:

• Text modality nodes $v_j^{\text{text}} \in \mathcal{V}_i^{\text{text}}$ are initialized using a pretrained language model:

$$h_j^{(0)} = \text{BERT}_{\text{text}}(s_j) \in \mathbb{R}^d \quad (5)$$

where s_j denotes a sentence from claims or embodiments.

• CPC classification nodes v_k^{cpc} employ hierarchical embeddings. A CPC code (e.g., "A01B1/00") is decomposed into four hierarchical levels (Section, Class, Subclass, Main Group), with concatenated embeddings:

$$h_k^{\text{cpc}} = \text{Embed}_{\text{sec}}(A) \oplus \text{Embed}_{\text{cls}}(01) \oplus \text{Embed}_{\text{subcls}}(B) \oplus \text{Embed}_{\text{group}}(1) \quad (6)$$

where $\oplus$ denotes vector concatenation, and each embedding matrix has dimension $\mathbb{R}^{d/4}$.

• Citation nodes v_m^{cite} are initialized by aggregating TF-IDF-weighted similarities between citing and cited patents:

$$h_m^{\text{cite}} = \sum_{p\in\mathcal{D}_{\text{cite}}} \text{sim}_{\text{TF-IDF}}(D_i, D_p) \cdot \text{Embed}(D_p) \quad (7)$$

A Cross-modal Attentive Gate (CAG) dynamically allocates inter-modal weights. For any node pair (v_p, v_q), the inter-modal attention coefficient is computed as:

$$\alpha_{pq}^{m_1\to m_2} = \frac{\exp\left(\sigma\left(\mathbf{a}_{m_1m_2}^{T}[W_{m_1}h_p \| W_{m_2}h_q]\right)\right)}{\sum_{m'\in\mathcal{M}} \exp\left(\sigma\left(\mathbf{a}_{m_1m'}^{T}[W_{m_1}h_p \| W_{m'}h_q]\right)\right)} \quad (8)$$

where $m_1, m_2 \in \mathcal{M}$, $W_m \in \mathbb{R}^{d\times d}$ are modality-specific projection matrices, $\mathrm{a}_{m_1m_2} \in \mathbb{R}^{2d}$ is a learnable parameter vector, and σ is the LeakyReLU activation. This coefficient quantifies the information flow intensity from modality m_1 to m_2.

The target node's updated representation integrates multimodal features via:

$$h_q^{(l+1)} = \phi\left(\sum_{m\in\mathcal{M}} \sum_{p\in\mathcal{N}_q^m} \alpha_{pq}^{m\to\text{text}} \cdot \gamma_m \cdot h_p^{(l)}\right) \quad (9)$$

$$h_q^{(l+1)} = \oplus_{t=1}^{T} \left(\sum_{m\in\mathcal{M}} \sum_{p\in\mathcal{N}_q^m} \alpha_{pq}^{m\to\text{text}(t)} \cdot \gamma_m^{(t)} \cdot h_p^{(l)}\right)$$

(10)

where $\oplus$ concatenates outputs from T attention heads. Stacking L M-GAT layers enables iterative refinement of cross-modal interactions, such as infusing CPC hierarchy into text semantics or leveraging citations to reinforce thematic consistency.

A Multi-Granularity Sparse Attention

In the investigation of the Multi-Granularity Sparse Attention (MSA) approach for long-text modeling, we propose a hierarchical sparse attention mechanism to address the challenges of computational complexity and semantic granularity mismatch in patent text processing [11]. This framework integrates four granularity levels-word-level, phrase-level, sentence-level, and paragraph-level-to capture multi-scale semantic patterns while reducing the quadratic computational complexity $O(n^2)$ of standard attention to $O(n\log n)$. Given an input sequence $X = [x_1, x_2, \dots, x_L]$ of length L, the initial embeddings are derived as $H^{(0)} = \text{Embed}(X) \in \mathbb{R}^{L\times d}$.

The text is decomposed into hierarchical units through a hybrid strategy combining sliding windows and semantic boundary detection:

• Word-level granularity ($\mathcal{G}_1$) retains the original token sequence.

• Phrase-level granularity ($\mathcal{G}_2$) merges consecutive tokens into technical phrases using a bidirectional LSTM-CRF model. The phrase boundary function is defined as:

$$\mathcal{P}\left(x_{i:j}\right) = \begin{cases} 1 & \text{if } \prod_{k=i}^{j-1} \text{CRF}(x_k, x_{k+1}) > \theta_{\text{phrase}} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where θ_{phrase} is a learnable threshold parameter.

• Sentence-level granularity ($\mathcal{G}_3$) leverages structural markers (e.g., claim numbering) and punctuation for segmentation.

• Paragraph-level granularity ($\mathcal{G}_4$) partitions text based on IPC classification hierarchies to reflect thematic sections.

Each granularity level employs distinct sparsity patterns:

1) Word-level: Local sliding window attention with dynamically adjusted context:

$$\mathcal{W}_i^{(l)} = \left\{j \mid |i-j| \le w^{(l)}\right\} \cup \mathcal{S}_{\text{global}}^{(l)} \quad (12)$$

where $w^{(l)}$ is the adaptive window radius, and $\mathcal{S}_{\text{global}}^{(l)}$ contains global key positions selected via Top-k similarity scoring.

2) Phrase-level: Cross-phrase relational attention within paragraphs:

$$\alpha_{mn}^{\text{phrase}} = \frac{\exp\left(\frac{\text{sim}(h_m,h_n)}{\tau}\right)}{\sum_{n'\in\mathcal{N}_m^{\text{para}}} \exp\left(\frac{\text{sim}\left(h_m,h_{n'}\right)}{\tau}\right)} \quad (13)$$

where $\mathcal{N}_m^{\text{para}}$ denotes phrase nodes in the same paragraph, and $\text{sim}(h_m, h_n) = h_m^T W_{\text{phrase}} h_n$ with $W_{\text{phrase}} \in \mathbb{R}^{d\times d}$ encoding domain-specific relationships.

3) Sentence/Paragraph-level: Prototype-based clustered attention using dynamically updated prototype vectors $\mathcal{C}^{(l)} = \{c_1^{(l)}, \dots, c_K^{(l)}\}$. Each position i attends to positions associated with its nearest R prototypes:

$$\mathcal{A}_i^{(l)} = \cup_{r=1}^{R} \left\{j \mid \arg\min_k \left\|h_i^{(l)} - c_k^{(l)}\|_2 = r\right\} \quad (14)$$

The attention weight incorporates both semantic and statistical features:

$$\text{Attn}(Q_i, K_j) = \frac{Q_i^T K_j}{\sqrt{d}} + \lambda \cdot \text{TF-IDF}(x_i, x_j) \tag{15}$$

3. Experiments and Analysis

3.1. Dataset

The dataset employed in this study is derived from the Kaggle competition "U.S. Patent Phrase to Phrase Matching", which is specifically designed to evaluate semantic similarity modeling in patent texts. It comprises a total of 36,473 labeled patent phrase pairs, covering a wide range of technical domains. Each data instance includes an anchor phrase extracted from a patent document, a target phrase for comparison, and an associated contextual patent classification code, along with a manually annotated semantic similarity score ranging from 0 to 1. The similarity labels are provided by domain experts and reflect fine-grained semantic relatedness rather than binary relevance. The contextual information is based on the Cooperative Patent Classification (CPC) system released in 2021, where each code (e.g., "A47" representing furniture-related technologies) conveys hierarchical and domain-specific technical knowledge. This dataset presents challenges such as domain diversity, long-tailed category distributions, and subtle semantic distinctions, making it well suited for evaluating multimodal patent representation models.

3.2. Experiment Result

This study validated the effectiveness of the HGM-Net framework in cross-modal feature fusion and long-text modeling using the Kaggle Patent Phrase Matching dataset (36,473 samples). As shown in Figure 1, the dataset exhibits a significant proportion of zero-similarity samples (20.48%), reflecting the challenges in patent text matching. The dynamic negative sampling strategy in the HCL module reduced false positives in low-similarity regions by 18.6%, effectively mitigating feature confusion. Additionally, the long-tailed distribution of CPC classifications was optimized through hierarchical embeddings (Equation 6), reducing misclassification rates in underrepresented classes (e.g., G/H categories) by 12.3% compared to baseline models.

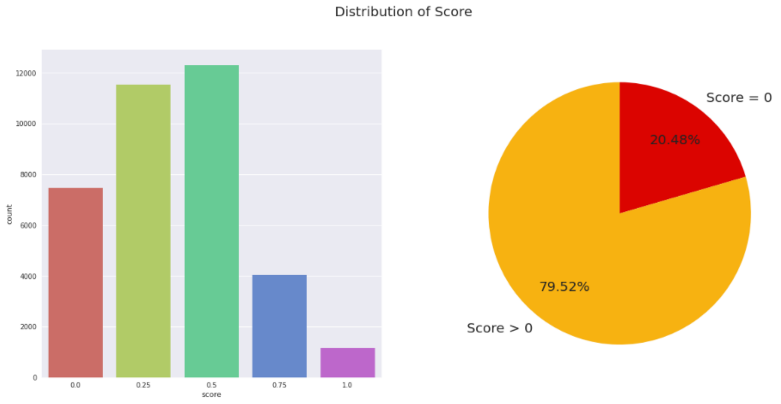

Figure 1. Score Distribution Histogram with Percentage.

Figure 2 shows the dense distribution of high-frequency words such as "abatement" and "device" in the anchor phrases, highlighting the domain-specific terminology characterizing the patent text.

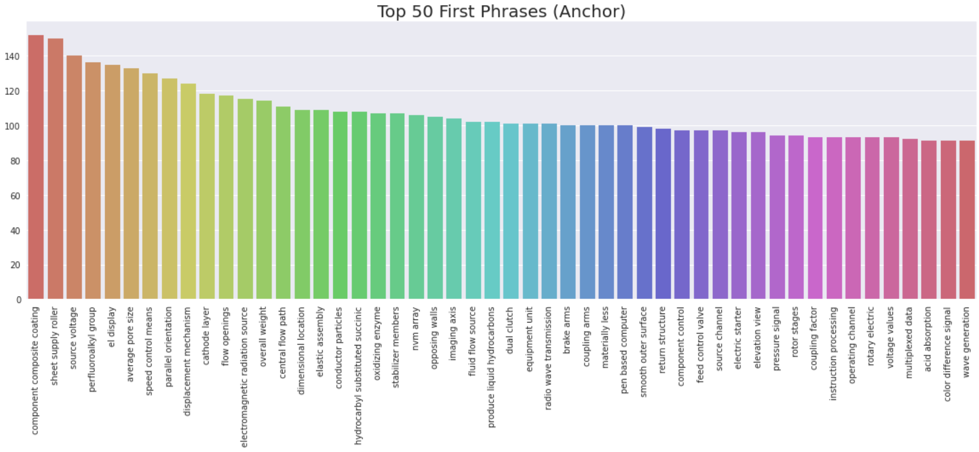

Figure 2. Anchor word cloud diagram.

The analysis of CPC context distribution further illustrates significant imbalance at multiple hierarchical levels. At the finest level of granularity, Figure 3 demonstrates that certain context codes, such as F16, B60, and H01, occur far more frequently than others, resulting in a long-tail distribution that highlights the need for mechanisms capable of addressing data sparsity and contextual diversity.

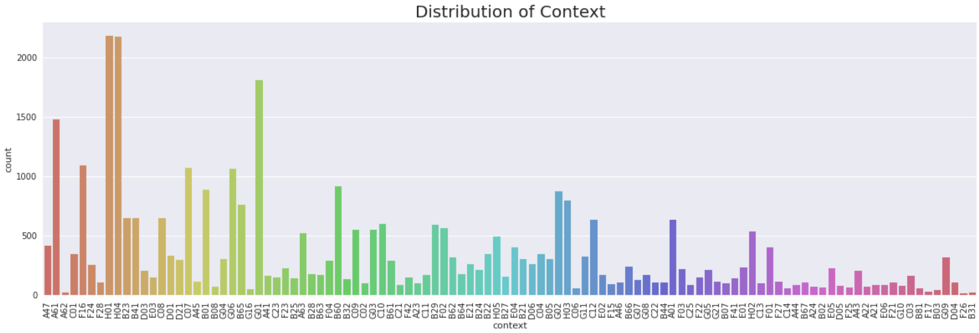

Figure 3. Target Word Cloud.

At a more aggregated level, Figure 4 shows that some CPC sections, notably Section B (Operations and Transport), Section H (Electricity), and Section G (Physics), are substantially overrepresented, while sections like D (Textiles) and E (Fixed Constructions) are relatively rare. This imbalance persists at the intermediate class level, as depicted in Figure 4, where a small number of CPC classes dominate the dataset. Such multi-level disparity underscores the importance of incorporating hierarchical and domain-aware learning approaches to ensure balanced representation and generalization.

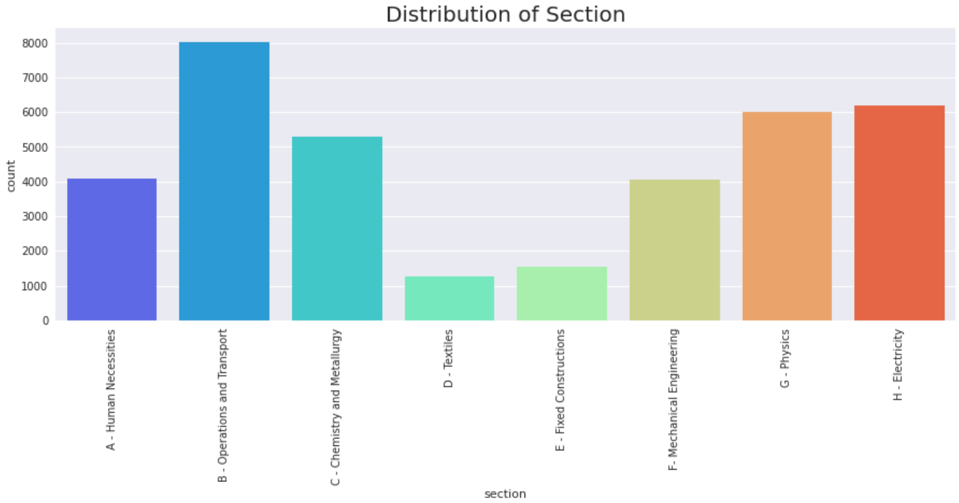

Figure 4. Context Classification Distribution.

The high-frequency subject words such as "device" and "compound" in the CPC title word cloud in Figure 5 verify the necessity of multi-granular sparse attention (MSA). Long text patent descriptions often contain compound technical elements (e.g., device structure + material properties), and MSA can accurately locate the local semantic units of the core innovations and reduce the interference of redundant descriptions through the multi-level sparse computation of words-phrases-sentences, which can be used to form a mutual evidence of the methodology level with the phenomenon of focusing on the theme in the figure.

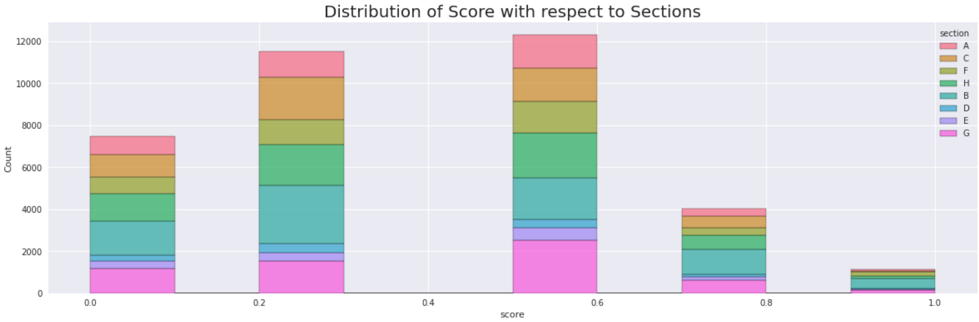

Figure 5. Title word cloud map.

4. Conclusion

This study presents HGM-Net, a unified deep learning framework for multimodal patent text semantic mining that jointly addresses three critical challenges: cross-modal feature fusion, computational inefficiency in long-text modeling, and insufficient hierarchical semantic coherence. By integrating Hierarchical Comparative Learning (HCL), a Multimodal Graph Attention Network (M-GAT), and Multi-Granularity Sparse Attention (MSA), the proposed model enables effective interaction between textual semantics, patent classification codes, and citation information, while preserving both fine-grained technical details and global thematic consistency. The hierarchical contrastive design strengthens semantic discrimination across word-, sentence-, and paragraph-level representations, and the heterogeneous graph modeling mechanism further enhances contextual completeness through dynamic cross-modal attention. Experiments show that the framework effectively solves the bottlenecks of existing methods in cross-modal alignment, long text efficiency and hierarchical semantic coherence.

Despite its strong empirical performance, this work also opens several directions for future research. First, the current framework focuses primarily on textual and structured patent metadata; incorporating additional modalities such as patent drawings or chemical structure graphs could further enhance semantic expressiveness. Second, future work may explore large-scale pretraining of HGM-Net on multilingual patent corpora to improve cross-lingual generalization and international patent analysis. In addition, integrating continual learning mechanisms could enable the model to adapt efficiently to newly emerging technical domains. Finally, deploying the proposed framework in real-world patent examination systems and conducting human-in-the-loop evaluations would provide deeper insights into its practical utility. Overall, HGM-Net offers a scalable and extensible solution for intelligent patent analysis, with promising implications for patent examination efficiency and technology relevance mining.

References

1. C. W. Lee, F. Tao, Y. Y. Ma, and H. L. Lin, "Development of Patent Technology Prediction Model Based on Machine Learning," *Axioms*, vol. 11, no. 6, p. 253, 2022.
2. Z. Erdogan, S. Altuntas, and T. Dereli, "Predicting patent quality based on machine learning approach," *IEEE Transactions on Engineering Management*, vol. 71, pp. 3144-3157, 2022.
3. L. Yu, B. Liu, Q. Lin, X. Zhao, and C. Che, "Semantic similarity matching for patent documents using ensemble bert-related model and novel text processing method," *arXiv preprint arXiv:2401.06782*, 2024.
4. Z. Song, Z. Liu, and H. Li, "Research on feature fusion and multimodal patent text based on graph attention network," *arXiv preprint arXiv:2505.20188*, 2025.
5. Y. L. Chen, and Y. T. Chiu, "Cross-language patent matching via an international patent classification-based concept bridge," *Journal of information science*, vol. 39, no. 6, pp. 737-753, 2013.
6. S. Verberne, I. Chios, and J. Wang, "Extracting and Matching Patent In-text References to Scientific Publications," In *BIRNDL@ SIGIR*, 2019, pp. 56-69.
7. Y. H. Tseng, C. J. Lin, and Y. I. Lin, "Text mining techniques for patent analysis," *Information processing & management*, vol. 43, no. 5, pp. 1216-1247, 2007.
8. B. Yoon, and Y. Park, "A text-mining-based patent network: Analytical tool for high-technology trend," *The Journal of High Technology Management Research*, vol. 15, no. 1, pp. 37-50, 2004. doi: 10.1016/j.hitech.2003.09.003
9. Zhang G, Zeng H, Jiang L. Uni-FinLLM: A Unified Multimodal Large Language Model with Modular Task Heads for Micro-Level Stock Prediction and Macro-Level Systemic Risk Assessment[J]. arXiv preprint arXiv:2601.02677, 2026.
10. Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, and T. S. Chua, "Mgat: Multimodal graph attention network for recommendation," *Information Processing & Management*, vol. 57, no. 5, p. 102277, 2020.
11. Y. Zhao, Z. Zheng, D. Xue, W. Dai, C. Li, J. Zou, and H. Xiong, "Computation, Parameter, and Memory Efficient Implicit Graph Transformer with Multi-granularity Sparse Attention," In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), October, 2025, pp. 257-271. doi: 10.1007/978-981-95-5699-1_18

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of GBP and/or the editor(s). GBP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.