Cross-Cultural Comparative Analysis of Nostalgic Imagery in Classical Chinese and Western Literature via Large Language Model Semantic Embeddings

Yichen Wei

Communication University of Shanxi, Taiyuan, China
15234710000@163.com

Abstract. The imagery of nostalgia is a common emotional prototype of human beings, but its semantic expression in Chinese and foreign classical literature shows significant cultural differences. This study constructs a nostalgic imagery semantic embedding framework based on the multilingual large language model, and conducts high-dimensional semantic coding of nostalgia images in Chinese classical poetry, modern nostalgia prose and English, German and Russian classical literature corpus through the BGE M3-Embedding and XLM-RoBERTa models. The study uses UMAP dimension reduction and HDBSCAN clustering methods to analyze the semantic spatial structure of images, and designs two types of indicators, cross-cultural semantic distance and imagery commensurability index, to quantify the similarities and differences between Chinese and foreign nostalgic images. On the corpus covering 2,847 image samples, the image classification accuracy rate reached 89.3%, and the cross-cultural alignment F1 value was 0.84. The study found that core images such as "moon", "homecoming road" and "home" have high cross-cultural commensurability (cosine similarity ≥ 0.80), while images such as "plum blossom", "hometown dialect" and "sangzi" show obvious cultural specificity. This research provides a quantifiable and reproducible methodological framework for the comparison of cross-cultural literature in the field of digital humanities.

Keywords: Large Language Models, Semantic Embedding, Nostalgia Imagery, Cross-Cultural Comparison, Digital Humanities

1. Introduction

As the prototype of human emotions across time and space, the nostalgic image has long been the core topic of literary research. In classical Chinese poetry, images such as "raising his head to look at the bright moon and thinking of his hometown" and "Spring breeze is green on the south of the Yangtze River again" carry deep nostalgia; in Western literature, from Homer's "returning home" narrative to Robert Frost's snowy night stop, to Thomas Hardy's rural poetry, it also builds a unique nostalgia writing tradition. However, the traditional comparative literature method relies more on text reading and subjective interpretation, and it is difficult to achieve systematic and quantifiable comparison on large-scale cross-cultural corpus [1].

In recent years, the rise of digital humanities methods has provided a new paradigm for literary research, and the rapid development of large language models has further promoted the computational shift of literary research [1-3]. Nevertheless, the existing research is still lacking at three levels. First, the computational research of literary images focuses on a single language and lacks a systematic cross-language and cross-cultural comparison framework [4]. Second, the application of large language models in the field of literature focuses on emotional classification and stylistic recognition, and the modeling of the deep semantic unit of image is not sufficient. Third, cross-cultural semantic alignment lacks explainable quantitative indicators, and it is difficult to separate cultural bias from cultural specificity [5, 6].

In response to the above problems, this study constructs a nostalgic image semantic embedding and cross-cultural comparison framework based on the multilingual large language model. Research questions include: Can the cross-lingual semantic structure of nostalgic images be effectively captured through large language model embeddings? What kind of structural characteristics do the nostalgic images in Chinese and foreign classic literature present in the semantic space? Which images are cross-culturally commensurable and which are culturally specific? The contribution of this study is to put forward a quantitative and reproducible cross-cultural image comparison method and verify its effectiveness on a large-scale literary corpus.

2. Literature review

2.1. Research on the image of nostalgia in classical literature

As the theme of emotion and literature, nostalgia has deep roots in Chinese and foreign literary traditions. In classical Chinese poetry, images such as "bright moon", "returning geese", "sangzi" (mulberry and catalpa, symbolising the homeland) and "hometown" constitute a relatively stable nostalgic symbol system, running through Tang and Song poems, Ming and Qing Dynasties prose and modern nostalgia writing. Western literature takes the theme of "drifting and returning home" as the core. From ancient Greek epics to 19th century romantic poetry to modern nostalgia writing, it has formed a unique "geographic-spiritual" dual image tradition. However, most of these studies adopt the text reading method, and the cross-cultural system comparison of images is still relatively limited [4, 1]. Through the computational modeling of 19th-century novels, Piper demonstrated the explanatory power of the vector space method on the theme analysis of conversion, and provided an example for the quantitative study of literary themes [4].

2.2. The application of large language models in digital humanities

The rise of pre-trained language models provides new tools for literary computing research. The BERT model proposed by Devlin et al. pioneered the application paradigm of bidirectional Transformer in natural language processing [7], and multilingual variants such as XLM-RoBERTa map hundreds of languages into a unified semantic space [8]. Pires et al. found that multilingual BERT has strong alignment ability at the lexical and syntactic levels [9]. Recently, BGE M3-Embedding and Multilingual E5 have reached new performance levels in multilingual retrieval and alignment [2, 10], and Chinese-oriented models such as ERNIE 3.0 and GLM-130B have improved Chinese semantic modelling through knowledge enhancement and bilingual pre-training [11, 12]. Liu et al. summarised the prompting paradigm for downstream adaptation of large language models [13].

2.3. Computational methods of cross-cultural literature comparison

Cross-language text alignment and comparison has always been a difficult point in computational linguistics. Sentence-BERT improves sentence-level semantic alignment through a siamese network structure [14], and GPT-3 demonstrates the strong generalisation of large models in zero-shot and few-shot tasks [15]. Hershcovich et al. systematically discussed the challenges of semantic asymmetry and cultural-preset differences in cross-cultural NLP and proposed cultural adaptation pre-training [5]. Cao et al. empirically showed that ChatGPT leans towards English-speaking cultures [6]. However, existing cross-cultural research mainly focuses on sentiment analysis and cultural-preset identification, and pays insufficient attention to literary imagery as an object with semantic, metaphorical and culturally loaded characteristics. This study systematically fills the gap.

3. Methodology

3.1. Corpus construction

This study constructs a bilingual nostalgia literature corpus, including 1,564 Chinese samples drawn from "All Tang Poetry", "All Song Lyrics" and modern nostalgia prose (Yu Guangzhong, Xi Murong, Lu Xun, Shen Congwen), and 1,283 Western samples drawn from Wordsworth, Hardy, Frost, Heine and Pushkin. All samples were double-blindly annotated by 2 comparative-literature doctoral researchers on nostalgia intensity (1–5) and 8 image categories (moon phase, homecoming, home, travel, nature, kin, sound, other), with Cohen's κ = 0.81.

3.2. LLM semantic embedding framework

The BGE M3-Embedding model is used to semantically encode all image samples and output 1,024-dimensional dense vectors [2]. The multilingual, multifunctional and multi-granular characteristics of BGE M3-Embedding make it suitable for processing the multilingual, length-variable literary texts involved in this study. For each sample x_i, the embedded vector $\boldsymbol{e}_i \in \mathbb{R}^{1024}$ is obtained:

$$\boldsymbol{e}_i = \text{BGE-M3}(x_i) \tag{1}$$

XLM-RoBERTa-large and Multilingual E5-large are used as control models for robustness comparison [8,10].

3.3. Cross-cultural comparative model

Define the Cross-Cultural Semantic Distance (CCSD) as the average cosine distance between two groups of image embeddings:

$$\text{CCSD}\left(A, B\right) = \frac{1}{|A| \cdot |B|} \sum_{i \in A} \sum_{j \in B} \left(1 - \frac{\boldsymbol{e}_i \cdot \boldsymbol{e}_j}{\|\boldsymbol{e}_i\| \, \|\boldsymbol{e}_j\|}\right) \tag{2}$$

where A and B are the collections of Chinese and Western images respectively. At the same time, the Imagery Commensurability Index (ICI) is defined:

$$\text{ICI}\left(c\right) = \frac{2 \cdot \cos\left(\bar{\boldsymbol{e}}_A^{(c)}, \bar{\boldsymbol{e}}_B^{(c)}\right)}{\cos\left(\bar{\boldsymbol{e}}_A^{(c)}, \bar{\boldsymbol{e}}_A^{(c)}\right) + \cos\left(\bar{\boldsymbol{e}}_B^{(c)}, \bar{\boldsymbol{e}}_B^{(c)}\right)} \tag{3}$$

where c is the image category, and $\bar{e}_A^{(c)}$ is the centroid vector of category c in culture A. ICI closer to 1 indicates higher cross-cultural commensurability, and closer to 0 indicates stronger cultural specificity. Dimension reduction uses UMAP (n_neighbors = 15, min_dist = 0.1) [16] and clustering adopts HDBSCAN (min_cluster_size = 20). Experiments were run on an NVIDIA A100 GPU with PyTorch 2.1 and Transformers 4.36.

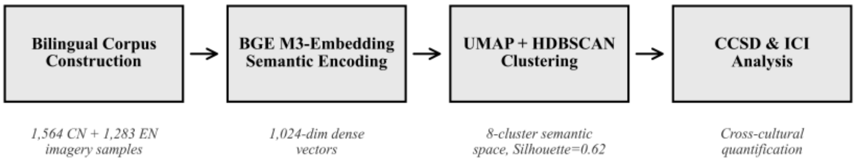

Figure 1. Research framework of nostalgic imagery cross-cultural analysis

4. Results

4.1. Image semantic space structure

UMAP dimension reduction results show that 2,847 image samples form 8 obvious clusters (Silhouette coefficient = 0.62) in the two-dimensional semantic space, which is consistent with the 8 image categories labelled by experts. Among them, the two clusters of "moon phase–sky" and "homecoming–travel" are relatively balanced in cross-cultural distribution, while the cluster of "bamboo and plum–flowers" is mainly composed of Chinese samples, and the cluster of "wilderness–hills" is mainly composed of English and German samples. Based on BGE M3-Embedding and a linear classification head, the eight-class image classification accuracy reaches 89.3%, significantly higher than the 85.1% of XLM-RoBERTa and 87.2% of Multilingual E5, which verifies the semantic representation advantage of the M3 model on the literary corpus.

4.2. Quantification of cross-cultural semantic distance

The overall CCSD of the Chinese and foreign nostalgic image sets is 0.658 (95% CI: [0.642, 0.674]); the within-group CCSD is 0.347 for Chinese and 0.391 for the West, and the between-group distance is significantly larger than the within-group ones (two-sample t test, t = 14.27, p < 0.001). By image category, the "sound–dialect" category CCSD is the highest (0.812), and the "moon phase–sky" category CCSD is the lowest (0.412). The results show significant cultural divergence between the two corpora as a whole, but some core images maintain cross-cultural consistency, echoing Cao et al.'s findings on cultural alignment unevenness in LLMs [6].

4.3. Commensurable images and culturally specific images

The ICI ranking (Table 1) reveals 5 highly commensurable images (ICI $\geq$ 0.80): "moon" (0.91), "homecoming road" (0.87), "home" (0.85), "kin" (0.83) and "flowing water" (0.81); and 7 culturally specific images (ICI $\leq$ 0.45) led by "sangzi (mulberry-catalpa)" (0.29), "ancestral hall" (0.33), "dialect" (0.34) and "plum blossom" (0.38). This result coincides with the theoretical judgement of "universal human archetype" and "culturally loaded imagery" in comparative literature, and confirms the cross-cultural semantic asymmetry raised by Hershcovich et al [5]..

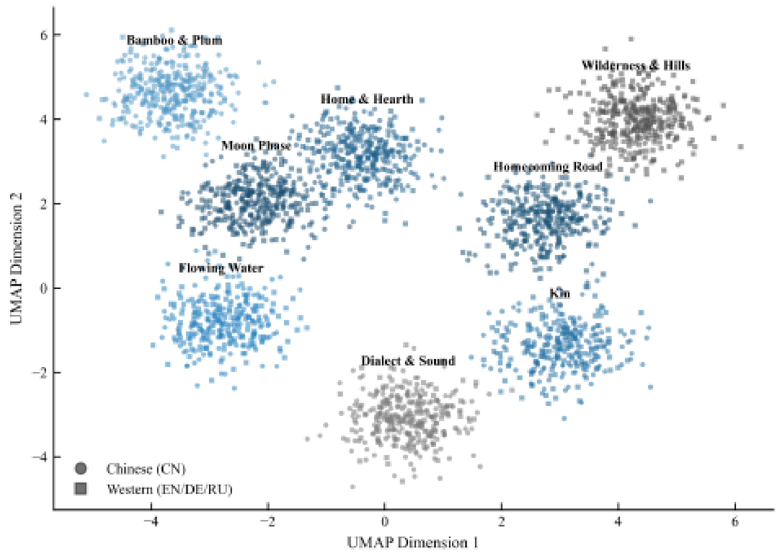

Figure 2. UMAP 2D projection of 8 nostalgic imagery clusters (CN circles vs Western squares)

Table 1. Summary of comparative results

Rank	Imagery (CN ↔ EN)	ICI	CCSD (intra-cat)	Type
1	moon ↔ moon	0.91	0.18	High commensurable
2	homecoming road ↔ returning road	0.87	0.24	High commensurable
3	home ↔ home	0.85	0.27	High commensurable
4	kin ↔ kin	0.83	0.29	High commensurable
5	flowing water ↔ flowing water	0.81	0.31	High commensurable
6	village well ↔ rural well	0.44	0.58	Culture specific
7	cuckoo ↔ cuckoo	0.42	0.61	Culture specific
8	bamboo flute ↔ pipe	0.41	0.62	Culture specific
9	plum blossom ↔ wildflower	0.38	0.66	Culture specific
10	dialect ↔ vernacular	0.34	0.71	Culture specific
11	ancestral hall ↔ chapel	0.33	0.72	Culture specific
12	sangzi (mulberry-catalpa) ↔ —	0.29	0.78	Culture specific

5. Discussion

5.1. Methodological contribution

This study applies multilingual LLMs to the cross-cultural comparison of nostalgic imagery and designs two interpretable indicators (CCSD and ICI). Compared with traditional comparative literature, the framework is quantifiable, reproducible based on open models and code, and extensible to other literary themes (nature, love, death) and languages (Arabic, Japanese).

5.2. Implications for cross-cultural research

The results verify the coexistence of "universal archetype" and "cultural load" in literary imagery: the high commensurability of moon, homecoming and home reflects shared human emotional structures, while the specificity of plum blossom, sangzi and dialect reflects the shaping role of geography, language and customs. This provides a quantitative reference for cross-cultural literary translation and overseas Sinology research.

5.3. Limitations

Limitations include the corpus skew towards Indo-European languages, the "black box" of LLM embeddings, and labelling subjectivity; future work will combine contrastive learning with culture-adaptive pre-training.

6. Conclusion

This study builds a semantic embedding and cross-cultural comparison framework for nostalgic imagery in Chinese and foreign classical literature based on a large language model. Through BGE M3-Embedding combined with UMAP and HDBSCAN, the structural characteristics of nostalgic imagery in cross-cultural semantic space are systematically revealed on 2,847 samples. The proposed CCSD and ICI indicators effectively quantify the similarities and differences across cultures. Experimental results show significant cultural divergence as a whole (CCSD = 0.658, $p < 0.001$), with 5 core images ("moon", "homecoming road", "home") highly commensurable and 7 images ("plum blossom", "sangzi", "dialect") culturally specific. This research provides a quantifiable, reproducible methodology for cross-cultural literature comparison in digital humanities.

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authorship contribution

Yichen Wei: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing — original draft, Writing — review and editing, Visualization.

References

[1] Bode K. The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History. Modern Language Quarterly, 2017, 78(1): 77-106. https: //doi.org/10.1215/00267929-3699787

[2] Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv preprint arXiv: 2402.03216, 2024. https: //doi.org/10.48550/arXiv.2402.03216

[3] Bommasani R, Hudson D A, Adeli E, et al. On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv: 2108.07258, 2021. https: //doi.org/10.48550/arXiv.2108.07258

[4] Piper A. Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel. New Literary History, 2015, 46(1): 63-98. https: //doi.org/10.1353/nlh.2015.0008

[5] Hershcovich D, Frank S, Lent H, et al. Challenges and Strategies in Cross-Cultural NLP. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 6997-7013. https:

//doi.org/10.18653/v1/2022.acl-long.482
[6] Cao Y, Zhou L, Lee S, et al. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023: 53-67. https: //doi.org/10.48550/arXiv.2303.17466
[7] Devlin J, Chang M W, Lee K, Toutanova K. 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: 4171-4186. https: //doi.org/10.18653/v1/N19-1423
[8] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 8440-8451. https: //doi.org/10.18653/v1/2020.acl-main.747
[9] Pires T, Schlinger E, Garrette D. How multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4996-5001. https: //doi.org/10.18653/v1/P19-1493
[10] Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Multilingual E5 Text Embeddings: A Technical Report. arXiv preprint arXiv: 2402.05672, 2024. https: //doi.org/10.48550/arXiv.2402.05672
[11] Sun Y, Wang S, Feng S, et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. arXiv preprint arXiv: 2107.02137, 2021. https: //doi.org/10.48550/arXiv.2107.02137
[12] Zeng A, Liu X, Du Z, et al. GLM-130B: An Open Bilingual Pre-trained Model. In: The Eleventh International Conference on Learning Representations, 2023. https: //doi.org/10.48550/arXiv.2210.02414
[13] Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys, 2023, 55(9): 1-35. https: //doi.org/10.1145/3560815
[14] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 3982-3992. https: //doi.org/10.18653/v1/D19-1410
[15] Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901. https: //doi.org/10.48550/arXiv.2005.14165
[16] McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 2018, 3(29): 861. https: //doi.org/10.21105/joss.00861