SAPIENZA
UNIVERSITÀ DI ROMA

Project Report: Book Genre Classification with Multimodal Deep Learning

Noah KAMANGAR, Hanieh MAHDAVI, Zuzana MICIAKOVA,
Sara REIS and Zoé DESPREZ

Fundamentals / Foundations of Data Science 2024
Sapienza Università di Roma

December 2024

Abstract

This report presents the progress of a multimodal deep learning study on book genre classification, combining cover images and textual metadata. A dataset comprising 46,226 books across 29 genres was curated. Initial experiments demonstrate that a multimodal approach integrating ResNet50 and BERT achieves an 83% accuracy, surpassing unimodal baselines. Future efforts will focus on refining fusion techniques and optimizing the model for real-world deployment.

1 Introduction

The book's cover is the first interaction it has with a potential reader, although it is not often enough to convince them to read. The same happens with computers. Although some studies focus on how machine learning can understand a book's genre by identifying the visual cues in its design [1], we wanted to determine if combined visual and written data could provide better results. These findings are especially relevant to understanding if a book is marketed in the correct way to attract its intended target audience. By leveraging combined visual and written data, retailers can optimize categorization strategies, reducing the chances of misplacing books and improving the overall shopping experience.

2 Related Work

Our project's initial idea is based on the aforementioned research to classify a book's genre by its cover analysis [1]. The findings showed that AlexNet could predict the cover genre up to 40% accuracy, offering a promising start to this problem. Another study aligned with our goal, investigating how multimodal networks could improve predictions by considering both the book's visual design and the text on its cover. A multimodal network for visual and textual data was applied. Their approach first utilizes a neural network (NN) to extract features from the visual modality; and another NN to extract features from the textual modality; as the next step a fusion unit decides which of the two modalities is more informative and, finally, they perform the classification by giving more emphasis to the modality that was considered more informative [2]. Various algorithms were tested for the extraction of the features. For the visual data, the article tested many variations of the ResNet algorithms, and, for the textual data, they proposed an EXAN (EXplicit interActive network) architecture.

3 Proposed Method

Our method integrates visual and textual features using a multimodal deep learning approach:

- **Image Processing:** Pre-trained ResNet50 and MobileNet models are employed to extract visual features from resized book covers ($224 \times 224 \times 3$).
- **Text Embeddings:** Book titles and descriptions are embedded using USE and BERT.
- **Feature Fusion:** Visual and textual embeddings are concatenated and passed through fully connected layers for classification.

4 Work Completed So Far

4.1 Dataset Collection and Preprocessing

The dataset comprises 46,226 books across 29 genres. Metadata was enriched using web scraping (Selenium) to include book descriptions. The data preprocessing steps included:

- Resizing images to $224 \times 224 \times 3$.
- Cleaning text by removing stopwords and punctuation.
- Computing embeddings for titles and descriptions using USE and BERT.
- Splitting data into 90% training, 10% testing, and further dividing training data for validation.

The t-SNE visualization (see Figure 1) shows how title and description embeddings cluster by genre, illustrating the effectiveness of the preprocessing step.

4.2 Model Development

Baseline models were trained and evaluated:

- MobileNet + USE achieved 60.1% accuracy.
- MobileNet + BERT scored 56.3%.
- ResNet50 + BERT achieved 83% accuracy, the highest among all models.

4.3 Preliminary Results

Example outputs generated by the model demonstrate its ability to classify book genres accurately (see Figure 2). The model's performance metrics are summarized in Table 1.

5 Challenges Encountered

- **Dataset Quality:** Many books lacked descriptions, necessitating extensive web scraping.
- **Fusion Complexity:** Fine-tuning the fusion of visual and textual features posed challenges.
- **Performance Limitations:** Initial unimodal models showed low performance (e.g., 23% accuracy for MobileNet only).

6 Next Steps

- **Author Feature:** Investigate the impact of including the **author** as a feature. We could explore if it improves, worsens, or has no effect on model performance.
- **Title and Description as Separate Features:** Experiment with feeding **title** and **description** separately to the model, instead of combining them, to see if this improves classification.
- **Hyperparameter Tuning:** Fine-tune the model's hyperparameters (e.g., learning rate, batch size) to optimize performance beyond default settings.

- **Top-5 Accuracy Analysis:** Further investigate why some genres (e.g., *Children's Books*) have higher accuracy than others, and explore methods to improve performance for challenging categories like *Politics*.
- **CNN Architecture:** Test other CNN architectures or deeper models to improve feature extraction, especially for complex genres.

Conclusion

This study highlights the efficacy of multimodal learning in book genre classification. By combining visual and textual features, the proposed model achieves significant improvements over unimodal baselines.

Visualizations

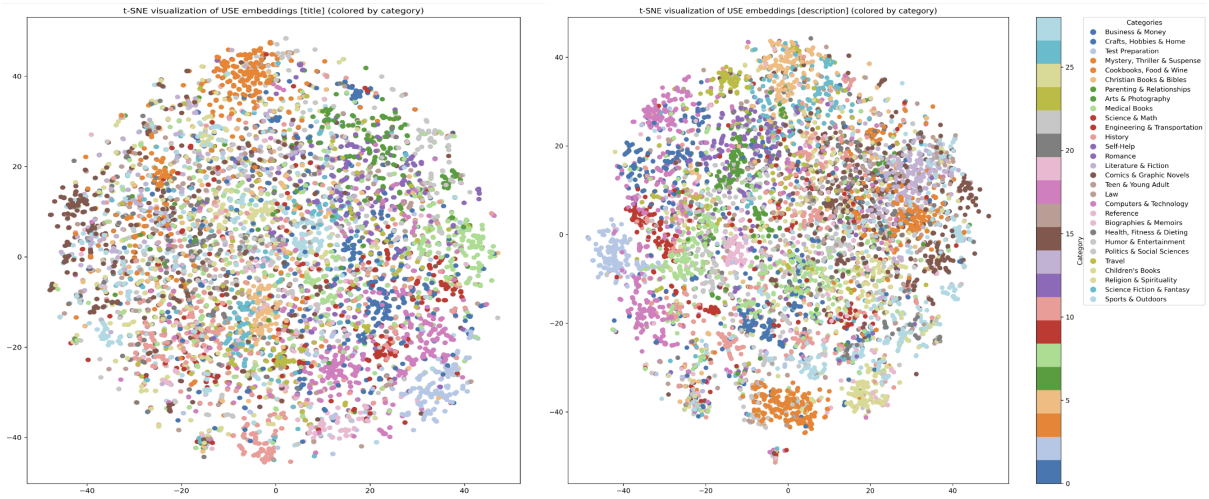

Figure 1: t-SNE Visualization of Title and Description Embeddings, Colored by Category

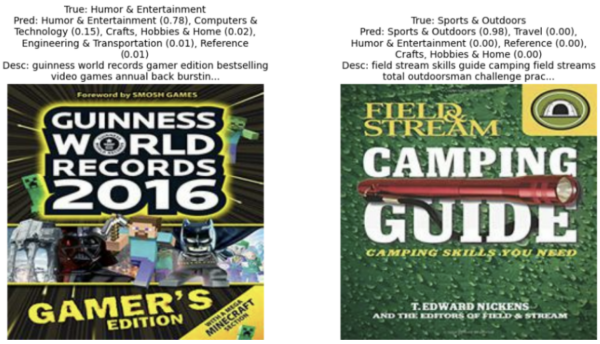

Figure 2: Example output

Model	Accuracy	Top-5 Accuracy	Precision	Recall	F1-Score
MobileNet+USE	60.1%	90%	60%	60.1%	59.6%
MobileNet+BERT	56.3%	89%	57.7%	56.3%	56.2%
ResNet+USE	59.4%	90%	59.8%	59.4%	59.1%
USE	59.2%	90%	59.5%	59.2%	59.1%
BERT	56.5%	89%	56.9%	56.5%	55.9%
MobileNet	23.3%	53%	22%	23.3%	21.2%

Table 1: Model Performance Metrics

References

[1] Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. "Judging a Book by its Cover". In: *arXiv* 1610.09204 (2016). Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan and German Research Center for Artificial Intelligence, Kaiserlautern, Germany and Kaiserslautern University of Technology, Kaiserlautern, Germany.

[2] A. Rasheed, A. I. Umar, S. H. Shirazi, Z. Khan, and M. Shahzad. "Cover-based multiple book genre recognition using an improved multimodal network". In: *International Journal on Document Analysis and Recognition (IJDAR)* 26.1 (2022), pp. 65–88.