Book Genre Classification with Multimodal Deep Learning

Fundamentals of Data Science 2024
Sapienza Università di Roma
December 2024

group 28
Noah KAMANGAR, Hanieh MAHDAVI, Zuzana MICIAKOVA,
Sara REIS, and Zoé DESPREZ

Abstract

A successful book cover is expected to effectively represent the book. But can its genre really be identified just by looking at the design? While prior research classified book genres using covers alone; this study extends that approach by incorporating textual descriptions and titles through multimodal deep learning. Our dataset, scraped from Amazon's book section, spans 29 genres, with balanced representation and preprocessed for consistency and quality. Multiple architectures were experimented with, including Universal Sentence Encoder (USE), Bidirectional encoder representations from transformers (BERT) for text, and ResNet50 and MobileNetV2 for images. Results demonstrate that combining visual and textual data improves classification accuracy compared to unimodal approaches, offering insights into effective automated book categorization and marketing strategies.

1 Introduction

Books are often judged by their covers, which provide the first impression to readers. The cover of a book is the first interaction with a potential reader, although it is often not enough to get them to read and get an idea of the book's content or genre. Therefore, a book's metadata, author, and description are often relevant complementary information to communicate the essence. 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 correctly to attract its intended audience. This research could also help retailers properly shelf books according to their genre, ensuring potential readers can easily find titles that align with their interests. By leveraging combined visual and written data, retailers can optimize categorization strategies, reducing the chances of misplacing books and improving the 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]. In this research, they proposed using CNNs (Convolutional Neural Networks), specifically AlexNet and LeNet, using a dataset with 30 categories that we are also using for this work. The findings showed that AlexNet could predict the cover genre with an accuracy of 25 % on average, offering a promising start to this problem. Another study looked in the direction we were hoping to pursue and investigated how the use of multimodal networks could improve the prediction when focusing both on the book's visual design and also the text on its cover. Aside from the visual cues, the cover also displays informative data such as the title, the author's name, and the publishing house, which can aid in identifying the genre of the book. With this in mind, 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].

3 Dataset

The dataset we used for the project was obtained from GitHub [1]. It included book covers and titles but did not include book descriptions. To address this, we developed crawlers to scrape book descriptions from Amazon. The dataset originally contained 30 genres, but we excluded the category "Calendar" for relevance. Additionally, we dropped rows with empty titles and descriptions. To maintain balance across genres, the dataset was sampled to include an equal number of entries for each category. Further steps included cleaning descriptions to remove control characters and validating the presence of book cover images. To maintain a uniform textual representation, text data, including titles and descriptions, underwent cleaning to remove digits, punctuation, and stopwords.
The labels for book categories were encoded numerically using a label encoder for compatibility with machine-learning models. Additionally, descriptions and titles were converted into dense vector representations using a pre-trained BERT or USE model. Finally, the book cover images were resized to a uniform dimension of 224x224 and transformed into tensors for integration into the dataset.

As a last step, the data was divided for training, testing, and validation as shown in Table 1.

Table 1: Dataset Split

Split	Books /Category	Total Books
Train	1,430	42,900
Test	164	4,920
Validation	143	4,290

3.1 Data Scraping

To scrape descriptions from Amazon, our initial approach involved implementing a crawler using the BeautifulSoup library in Python. However, this approach encountered issues: after 10 requests, subsequent requests were blocked as they were identified as coming from bots. To address this, we developed a scraper using Selenium, which mimicked human behavior. This approach successfully bypassed the blocking, allowing us to scrape about 500 books per session.

3.2 Exploratory Data Analysis

For text embedding, we used Universal Sentence Encoder (USE) for titles and descriptions, then plotted the results using t-SNE. Each data point represents a book, and the color indicates its category. As you can see, the embeddings for descriptions are denser compared to titles. This is because the words used in descriptions are typically more strongly related to a category than those in titles.

We performed a similar analysis for book covers using ResNet-50. In this case, the categories are more overlapped, indicating that book covers alone are not as effective in distinguishing categories compared to combining them with text features.

Another aspect we explored was the psychological perspective: whether there is a relationship between the dominant color of a book cover and its category. The answer is yes as visible in Figure 2. We sampled books in each category from the training set with sample sizes of 1, 10, 100, 500, 1,000, and the total number of books, and calculated the average dominant colors. There are noticeable differences; for example, books in the Children's category tend to have lighter colors, whereas those in the Mystery category are generally darker.

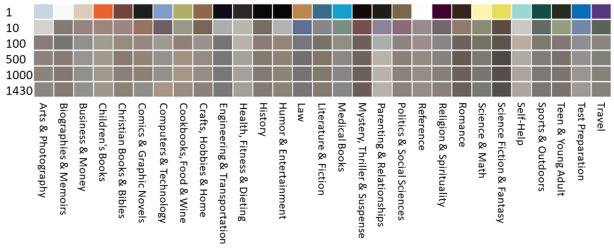

Figure 2: Dominant colors of the cover images

4 Models

As mentioned in Section 3, the dataset used in this project slightly differed from the original paper [1]. Not only in the cover images used but it also included book descriptions as the predictor of the genre, a feature completely omitted in the paper. To quantify the potential improvement in performance and assess the accuracy of the model in a more ample context motivated us to select multiple architectures for both modalities, text, and image, respectively. Although our research was aimed at the potentially most accurate and modern models, the computational cost was an important factor in the ultimate model selection.

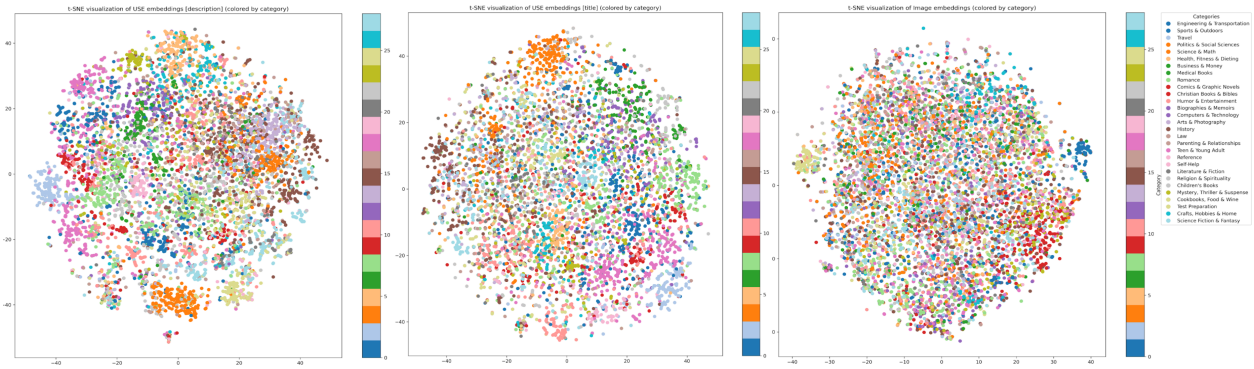

Figure 1: t-SNE visualizations of the Dataset

4.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model and currently, a benchmark for many natural language processing (NLP) tasks. Its repeatedly demonstrated accuracy in numerous papers (e.g. [3], [4]) made it an interesting candidate for text-based training in our project.

4.2 USE

Universal Sentence Encoder (USE) is a lightweight encoder-only model optimized for creating sentence embeddings. Compared to BERT it is very efficient and computationally inexpensive which comes with a certain trade-off in accuracy. Thanks to its properties it was a good counterpart to BERT.

4.3 ResNet50

One of the most frequently cited models used for image classification has proved to be ResNET50. As a part of the ResNet family, it is a deep convolutional neural network with 50 layers with high accuracy for classification tasks and robust architecture. This comes at the cost of higher memory usage and being computationally intensive.

4.4 MobileNetV2

Finally, to alleviate the computational cost present with ResNet50 we used an alternative in the form of MobileNetV2. MobileNetV2 is a highly efficient and lightweight CNN, with a competitive performance on image classification.

4.5 Multimodal Approach

To train a model that uses both image and text data, a multimodal model architecture must be used. We decided to use a similar approach as in the paper by J. Miller et. al [5]. Figure 3 shows the architecture of our training pipeline. The CNN and the text embedding can easily be replaced by another CNN or another text embedding. Both representations are concatenated and used with a softmax activation for classification.

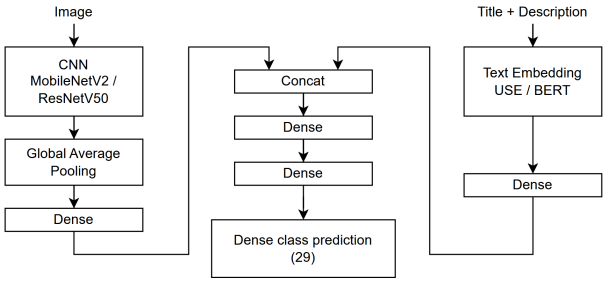

Figure 3: Multimodal model architecture

5 Results

We used different combinations of CNNs and text embeddings to compare them. We also tried out different sizes of validation sets and different batch sizes. Initial tests showed that a validation set size of 10 % of the training data and a batch size of 16 provided the best test accuracy. The results in Table 2 were all obtained with the final hyperparameters. Furthermore we experimented with different levels of finetuning of the CNNs. Due to time and computational constraints and also because of good performance we decided to freeze all layers of the CNN and only train the parameters of the Global

Average Pooling and Dense Layer. In order to better evaluate the results, it should be noted that a random classifier with 29 classes would have an accuracy of 3.4 %.
In addition, to evaluate the multimodal approach, we also trained unimodal classifiers on the data set. These take into account either only text data or image data, but apart from that they use the exact same data set and also use the same seed for generating the validation set.
For the evaluation, we take accuracy as the main benchmark. Furthermore, the top 5 accuracy in order to have a less stringent assessment of the model, precision, recall and F1-score.

The results are provided in Table 2. The best performance is achieved by the combination of the MobileNetV2 and USE model. An Accuracy of 60 % and a Top-5 Accuracy of 90 % are very good results, considering that a random classifier would have an accuracy of 3.4 %. The precision, recall and F1 score show no particular anomalies and can also be interpreted as expected and good.
It is more interesting to compare the results of the multimodal approach with the unimodal approaches. A classification that only uses USE embedding is almost as good as the results of the multimodal approaches with an accuracy of 59.2 %. MobileNetV2 alone, on the other hand, performs rather poorly with an accuracy of 23.3 %. These results suggest that the classification is mainly based on the text data. This confirms the assumption made in 3.2 that the task of classifying the genre of books on the basis of the book cover alone is very difficult. Figure 5 also shows that the classification with the book cover only works well in some categories and very poorly in most of them.
Furthermore, the MobileNetV2 performance of 23.3 % is reasonable, considering that in previous work an accuracy of 24 % was achieved on almost the same data set [1]. In addition, as described in Section 3, we had to decrease the size of our dataset because we could not scrape a description of some books and had to balance the dataset.
Two other interesting observations are that BERT performs significantly worse than USE and ResNet also performs worse than MobileNetV2, even though they are both the more powerful models. It should be noted that we did not finetune BERT with our dataset and it is unclear what performance we would achieve with it. A possible explanation why the performance of ResNet is worse than MobileNetV2 could be that our dataset is too small and also very likely is that there is still room for improvement when fine tuning ResNet and we have not found the optimal settings.

It is also very important to note that BERT alone performed better than MobileNet + BERT. This illustrates how difficult a multimodal approach can be. One possible explanation is that feature fusion is very difficult and simple concatenation is not sufficient. Furthermore, the features of MobileNet are perhaps rather noisy when combined with BERT embedding and therefore reduce the overall performance.

Figure 4 shows the Confusion Matrix of our best performing model, MobileNetV2 + USE. It shows that there is a significant difference in the performance of categories. While a category like “Comics & Graphic Novels” seems to be rather easy with an accuracy of 86.6 %, a category like “Teen & Young Adult” seems to be very hard to get correct with an accuracy of 32.3 %.

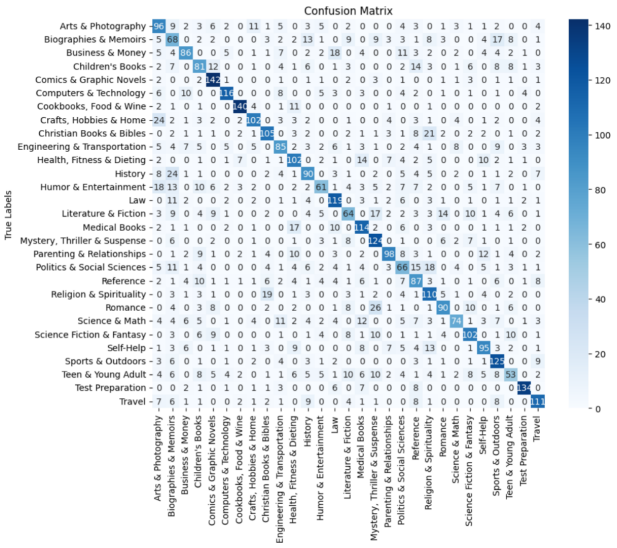

Figure 4: Confusion Matrix of MobileNetV2 + USE

As already mentioned before and visible in Figure 5, classifying books only based on their cover works well for only a few categories. Only 4 categories have an accuracy over 50 %. Still the results seem reasonable since e.g. the categories “Children’s Books” and “Comics & Graphic Novels” are getting confused visually.

Table 2: Performance Metrics on Test Set

Model	Accuracy	Top-5 Accuracy	Precision	Recall	F1-Score
MobileNetV2 + USE	60.1%	90%	60%	60.1%	59.6%
MobileNetV2 + 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%
MobileNetV2	23.3%	53%	22%	23.3%	21.2%
ResNet	11.2%	37%	7.6%	11.2%	7.5%

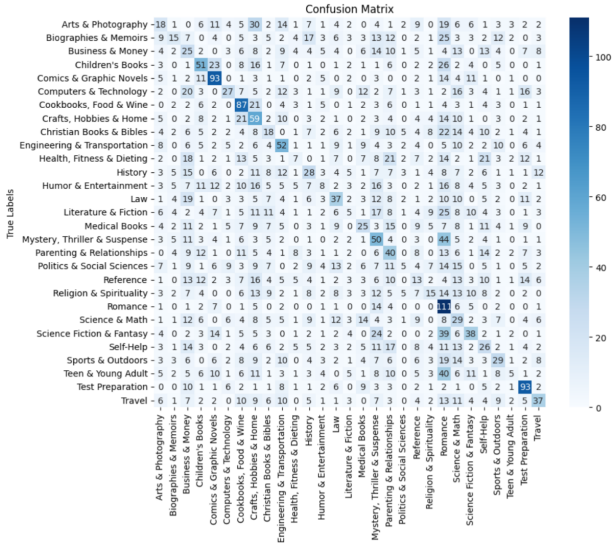

Figure 5: Confusion Matrix of MobileNetV2

6 Conclusion and Future Work

This study highlights the efficacy of multimodal learning in book genre classification by combining visual and textual features since we achieved the best performance by using a multimodal approach. At the same time, we were able to show that a multimodal approach alone is no guarantee of achieving better results than a unimodal approach, as seen with MobileNet + BERT and BERT.

Future work could explore adding the author as a feature, experimenting with separate title and description inputs, and fine-tuning hyperparameters to optimize performance especially with respect to ResNet. Further analysis of top-5 accuracy could help improve results for challenging genres while testing different CNN architectures may enhance feature extraction for complex categories.
Regarding multimodal learning, testing out different feature fusion approaches could be promising. Previous work by Wang et. al [6] presented the use of cross-modal attention mechanisms and hierarchical feature fusion strategies, which is expected to be also beneficial for our task.

Roles

Data Scraping: Hanieh, Noah
Exploratory Data Analysis: Hanieh
Training Pipeline + Training & Evaluation: Noah
First Presentation: Zuzana
Final Presentation: Hanieh, Noah
Research on multimodal algorithms: Sara
Progress Report: Zoé
Report: Sara, Hanieh, Noah

References

[1] Iwana, Kenji, B., Rizvi, R., & Tahseen, S. (2016). *Judging a Book by its Cover*. arXiv. https://arxiv.org/abs/1610.09204

[2] Rasheed, A., Umar, A. I., Shirazi, S. H., Khan, Z., & Shahzad, M. (2022). *Cover-based multiple-book genre recognition using an improved multimodal network*. International Journal on Document Analysis and Recognition (IJDAR), 26(1), 65–88. https://doi.org/10.1007/s10032-022-00413-8

[3] Niimi, J. (2024). *An Efficient Multimodal Learning Framework to Comprehend Consumer Preferences Using BERT and Cross-Attention*. https://doi.org/10.48550/arXiv.2405.07435

[4] Prasanna, K. R., Bharathi, M. G., Parthasarathy, S., & Venkatakrishnan, R. (2023). *Transformer-Based Models for Named Entity Recognition: A Comparative Study*. https://doi.org/10.1109/ICCCNT56998.2023.10308039

[5] Miller, S. J., Howard, J., Adams, P., Schwan, M., & Slater, R. (n.d.). *Multi-Modal classification using images and text.* SMU Scholar. https://scholar.smu.edu/datasciencereview/vol3/iss3/6/

[6] Advanced multimodal deep learning architecture for Image-Text matching. (2024, May 24). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10594167