Mutimodal classification of book covers

Klaudia Biczysko
klaudiateresa.biczysko.4334@student.uu.se

Justyna Sikora
justyna.sikora.3492@student.uu.se

Abstract

In this paper, we decided to explore different ways to classify a book by its cover. First, we classified book covers based solely on textual data, that is, the titles by using a pre-trained BERT model. Second, we used ResNet for the classification of the cover pictures. Finally, we concatenated two models and use them in a multi-modal model. BERT managed to achieve the best accuracy of 58.5%, while ResNet scored 30%. Since the concatenated model achieved an accuracy of 23.4%, the improvement of classification was not reached. We concluded that our concatenation is too complex.

1 Introduction

As the saying goes, don't judge a book by its cover. However, a book cover can often provide information about the content of a book. The purpose of our paper was to explore different methods by which a book can be classified by its cover. In this paper, we tried to answer below research questions:

1. Is it possible to classify the book into the correct genre, given an image or the title of a book?
2. Is it better to use a title or a cover to categorize a book?
3. Can we improve the accuracy of a neural network classifier by concatenating textual and image features?

2 Methodology

In this section, we first discuss the dataset, which was used in the experiments (section 2.1). We then briefly present the supervised classifiers (section 2.2), which we used to perform the classification.

2.1 Dataset

For this task, we used *BookCover30* published by Iwana et al. (2016). The dataset consists of 57,000 books web-scraped from the Amazon.com, Inc.marketplace, which were classified into 30 categories according to genre e.g. *Cookbooks* or *Medical books* (see Tab. 1). Figure 1 presents some sample book covers.

Label Category	Category Name
0	Arts & Photography
5	Comics & Graphic Novels
15	Literature & Fiction
25	Self-Help
29	Travel

Table 1: Sample categories from the *BookCover30* dataset. The dataset contains 30 labels.

The information of each book in the dataset contains its cover, title, author and other subcategories. The training set and test set are split into 90% - 10%. However since our BERT model for the title classification required also a validation set, we divided training set/ test set/ validation set into 80% - 10% - 10% respectively.

2.2 Experimental set up

BERT (KB) BERT is an acronym for Bidirectional Encoder Representations from Transformers and was introduced by Devlin et al. (2018). As ELMo, it is a type of pre-trained deep contextualized word embeddings, however, BERT uses Transformer, which is an attention-based model with positional encodings to represent word positions. The Transformer model is considered to be bi-directional since the encoder reads simultaneously the full word sequence. Hence this feature enables the model to learn the context of a word based on the left and right of the word. While ELMo is a character-based model, BERT encodes

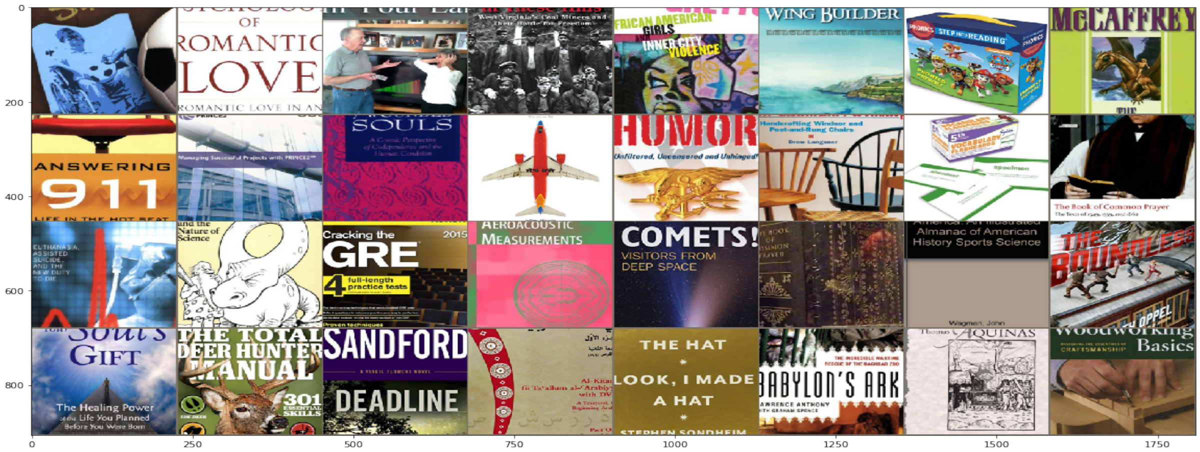

Figure 1: Visualization of the covers used in the multimodal classification task.

input as sub-words and learns embeddings for subwords.

ResNet (JS) ResNet is a type of deep neural network architecture introduced in 2016 by Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun (He et al., 2016). It is widely used within computer vision, along with models such as VGG and Alexnet. ResNet was designed especially to deal with the problem of the vanishing gradient, which is achieved by using a shortcut connection that leads to skipping some of the layers.

We used ResNet50 implementation pre-trained on ImageNet and provided by PyTorch.

Multimodal model (KB, JS) We tried to create an end-to-end model by concatenating the output from BERT and ResNet.

Evaluation metrics To assess the quality of the presented models, we used accuracy. For the multimodal model, we also decided to include top-5 accuracies.

3 Experiments

In this section, we present our experiments. First, we describe title classification with BERT (section 3.1), then cover image classification (section 3.2) and finally multimodal classification (section 3.3).

3.1 Title classification (KB)

For preprocessing and title classification, we chose the base BERT model pre-trained on unlabeled data extracted from the BooksCorpus and English Wikipedia. We also experimented with the base multilingual BERT, since some of the titles are not in English.

First, we imported the dataset from the Github repository, since columns did not have any titles, we decided to add them. After that, we created a train set, test set and validation set and divided them into 80% - 10% - 10% respectively. For tokenizing titles, BertTokenizer was used. After the tokenization process, we looked for the longest title. The idea behind it was to set padding to a maximum length of the sentence instead of 512 (the maximum length of a sequence allowed for BERT).

After the preprocessing process, we imported a pre-trained BERT model with 12 layers of Transformer encoder. We used Adam optimizer. The best learning rate for our model was 0.0000001, while we also tested 0.00000001, 0.001, 0.01, 0.1, 0.2 and 0.3. We trained the classifier for 5, 10 and 15 epochs. The best performing model achieved an accuracy of 58.5% for 5 epochs (see Tab. 2). Surprisingly, the monolingual BERT model performed better than the multilingual BERT.

Model	Epochs	Accuracy
BERT	**5**	**58.5%**
BERT	10	56.8%
BERT	20	56.5%

Table 2: The accuracy scores achieved by BERT.

3.2 Cover classification (JS)

In order to prepare the data for the training, the images were augmented and normalized, while the data for validation was only normalized. This included resizing images to 224 px x 224 px.

The data for training and testing was split according to the original division of the dataset provided by the authors - 90% for training and 10% for testing.

Once preprocessed, the data was fed into the model, which consists of a fully-connected layer that outputs predictions for the 30 classes. For the training, we used Adam optimizer with a learning rate of 0.001 and CrossEntropyLoss.

We also experimented with a more advanced model, with two linear layers - 2048 and 512 and a softmax activation function on the output layer, but it did not significantly improve the results.

After training the model for 25 epochs, it obtained 30% validation accuracy.

3.3 Multimodal classification (KB, JS)

To load data for the multi-modal model we modify the dataset class constructed for the text model so that it outputs not only labels and text, but also preprocessed images. Again, augmented and normalized for training and only normalized for validation and testing. Data is then split into train, test and validation sets (80% - 10% - 10%).

After training the models of one modality, both text and image models' parameter dictionaries were saved. We loaded them into a multimodal model, where we concatenate features from both models and finally pass them through the ReLu activation function. The model was trained for 10 epochs with SGD optimizer, learning rate 0.001 and CrossEntropyLoss and then the parameter dictionary based on the loss from the best performing epoch was saved. After 10 epochs, the model gained 34% accuracy on the validation set.

Finally, we evaluated the multimodal model on the test set by calculating top-k accuracy. The model reached 23.9% in top-1 accuracy, 42.2% in top-3 and 53.5% in top-5.

4 Results & Analysis

In this section, first, our findings for BERT and ResNet are discussed. Then, we present the analysis of the concatenation model. Figures 2, 3 and 4 illustrate the confusion matrix of the evaluated models on the test data, with the ground truth plotted on the horizontal y-axis, and the predicted genres plotted on the horizontal x-axis.

As it was mentioned in Section 3.1, BERT managed to achieve 58.5% of accuracy. By analyzing the confusion matrix, we can see that among the easiest categories for model to learn were: *Calendars* (92%), *Computers and Technology* (85%), *Test preparation* (84%) *Comics & Graphic Novels* (77%) and *Romance* (72%). Model performed worst on the *Mystery, Thriller & Suspense* (36%) and *Politics & Social Sciences* (38%).

As mentioned in Section 3.2, ResNet achieved the accuracy of 30%. As shown in Figure 3, ResNet correctly classified more than 60% of books in 4 categories: *Cookbooks, Food & Wine* (66%), *Comics & Graphic Novels* (63%), *Test preparation* (61%) and *Romance* (60%). On the contrary, *Teen & Young Adult* and *Literature & Fiction* were the most challenging categories for the classifier, since only 6.3% and 8.4% were properly categorized.

As far as the multimodal model is concerned, the highest accuracy scores were reach on *Test preparation* (57%), *Computers and Technology* (56%), *Romance* (50%) and *Christian Books & Bibles* (53%). However, there were multiple categories that achieved lower accuracy than 10%, including *Arts & Photography* (3.7%), *Literature & Fiction* (4.7%), *Business & Money* (5.8%), *Science & Math* (5.8%), *Teen & Young Adult* (7.4%).

The above-presented results show that there is a correlation between the hardest and easiest categories to predict for the text model and the image model. We can observe similar trends in the multimodal model, with *Literature & Fiction* and *Teen & Young Adult* as one of the most difficult categories for all models and *Test preparation* and *Romance* as the least challenging.

The results from the Bert model, which obtained the best scores of all models, suggest that some categories could be seen as confusing even for humans - 17% of *Children's Books* were labelled as *Teen & Young Adult* as well as 17% of *Medical Books* were categorized as *Health & Fitness & Dieting*, i.e. categories that to some extent are related. This also applies to *Mystery, Thriller & Suspense* which was incorrectly labelled as *Literature & Fiction* (15%) and more surprisingly *Romance* (23%). However, these tendencies are less obvious for the worse performing models.

Moreover, some books may belong to more than one category, which makes it even more difficult to correctly categorize them. For instance books categorized as *Religion& Spirituality* can be also found under *Christian Books & Bibles*.

5 Conclusions & Future work

In this investigation, the aim was to assess whether it is possible to classify a book into the correct genre, given a cover image or a title. This study has shown that a neural network model is able to classify a book into the correct genre.

The second aim of this study was to investigate whether it is more beneficial to take text or image features. The most obvious finding to emerge from this study is that BERT outperformed ResNet by almost 30%, which suggests that it is for the model to predict a category, given a text.

Finally, our final goal was to improve the accuracy by employing a multimodal model. As a result of our findings, we can conclude that our way of concatenating BERT with ResNet produced a very deep and complex model. We were not able to achieve satisfactory results and increase the accuracy scored by BERT or ResNet.

Future research direction may include tuning hyperparameters, as well as finding other ways for concatenating models.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2016. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*.

A Appendix

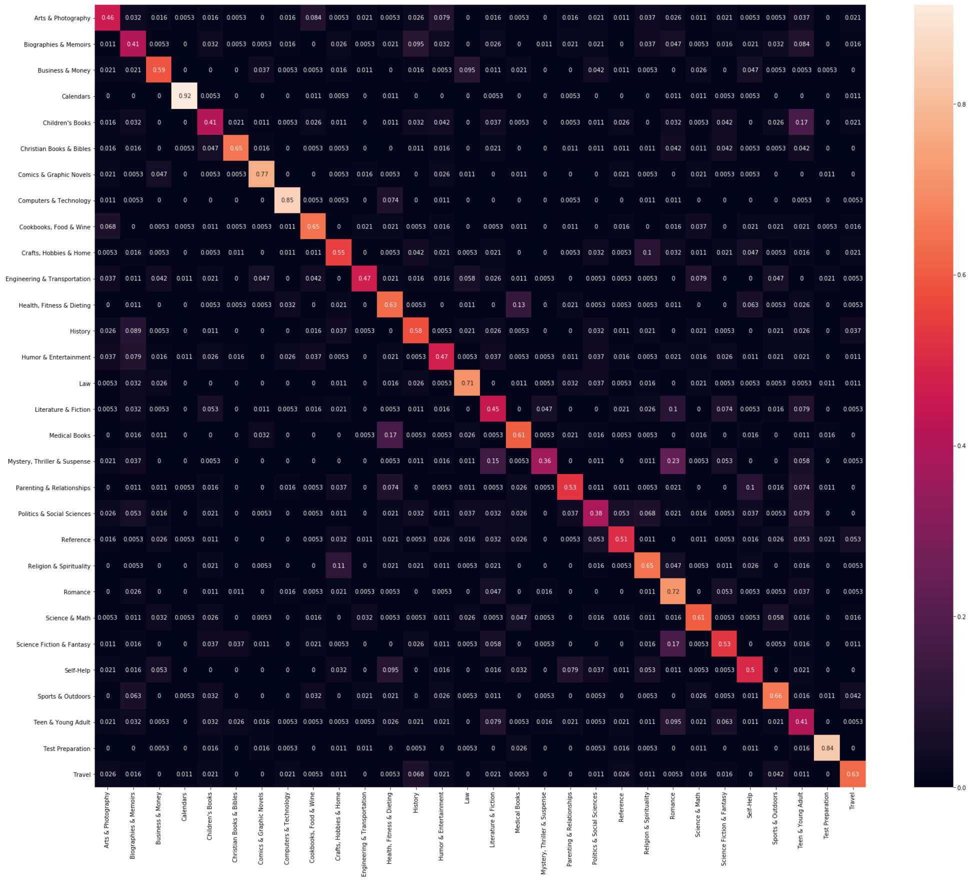

Figure 2: Confusion matrix of the BERT model on the test data. The horizontal axis represents the ground truth and the vertical axis represents the predicted genres.

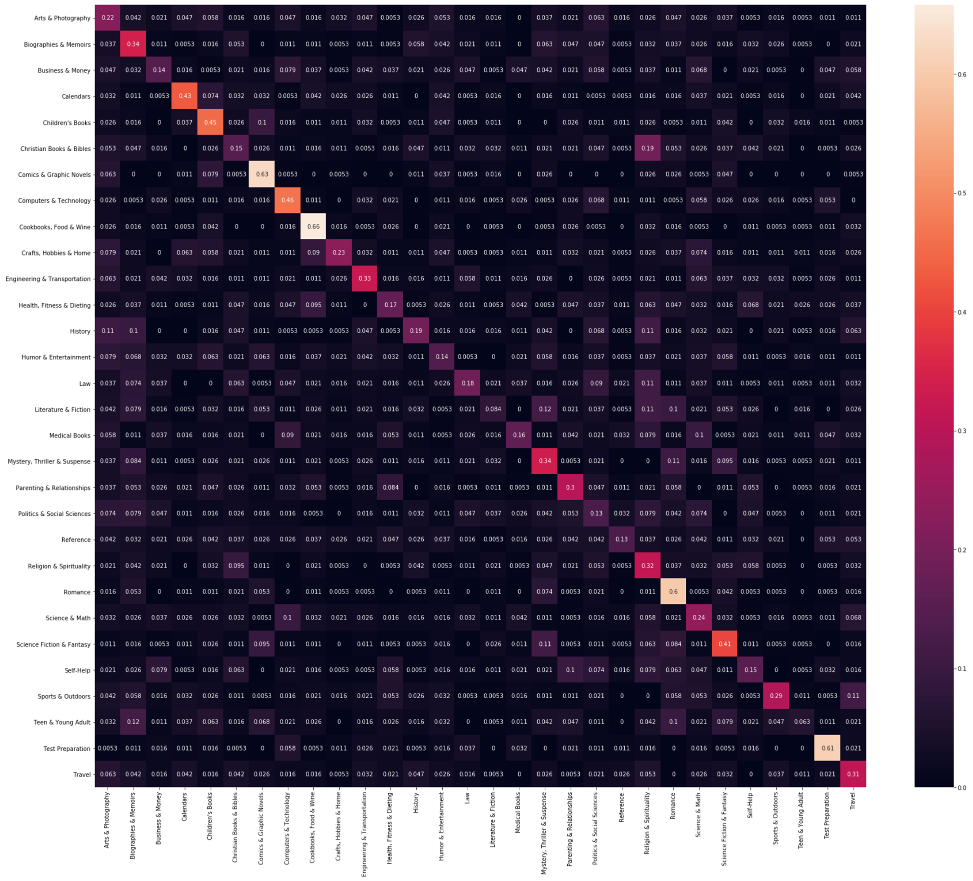

Figure 3: Confusion matrix of the ResNet model on the test data. The horizontal axis represents the ground truth and the vertical axis represents the predicted genres.

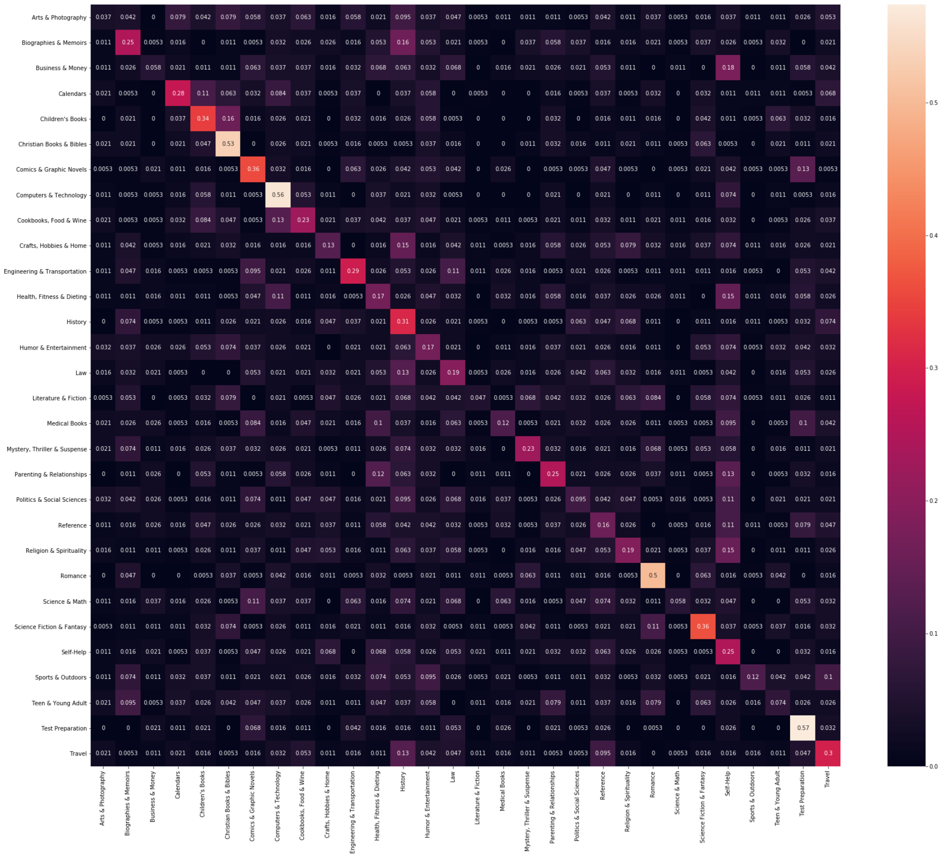

Figure 4: Confusion matrix of the multimodal model on the test data. The horizontal axis represents the ground truth and the vertical axis represents the predicted genres.