The story of book covers

Nataljia Gucevska **Sébastien Chevalley** **Alexis Montavon**
e-mail: name.surname@epfl.ch

Abstract

Since there were books, covers have been the one fashionable aspect that could catch the reader's eye at first glance. The aim of this project is to find insights about book covers and how they relate to literary categories, how the covers change over time and to analyze if the sales ranking is somehow related to the visual appearance of a book. Another purpose is to categorize the books using only visual aspect.

1 Introduction

In the following sections we describe our work on analyzing book covers and finding their link with other information. First we discuss how we used the available data of the Amazon dataset and its completion with data from Open Library[1] and the construction of our final dataset. Next we elaborate the different techniques we applied in order to find visually similar book covers, such as **clustering** and machine learning techniques for multi class **classification**. We also describe the analysis made after obtaining the different groups of book covers, and the **image processing** techniques that we applied in order to find their similarities.

2 Related work

Visual design of books is important and often is there to communicate some information to the reader. The link between the book's category and it's cover has already been explored by Iwana, Brian Kenji, et al.[2] using the Amazon dataset and 30 different categories. In their work they strictly focus on the category classification using CNNs.

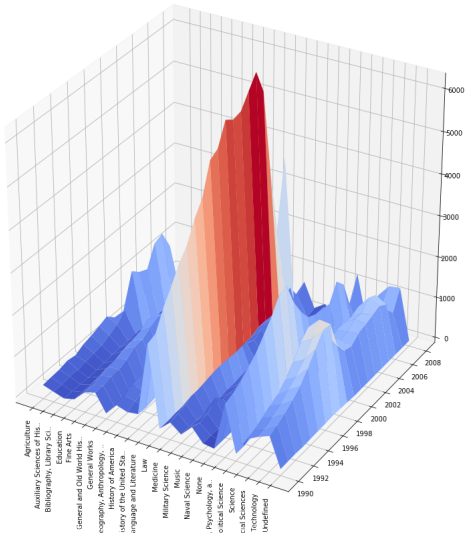

Figure 1: The final distribution of books by year and category. The z-axis represents the number of books per year and category. The category with the most important number of books is ***Language and Literature.***

3 Data collection and processing

3.1 Collecting the data

We started this project by obtaining the Amazon dataset about books and their covers[2]. The information on book covers are given as a 4'096 floats vector representation computed from a Convolutional Neural Network. But not all of the information contained about the books was useful to achieve our goals. The brands and categories given are uninteresting since in 90% of the cases the books were classified under the category "books", the description and prices were sparse. Examples on the amazon's website told us Metadata contained sales rank, categories and unique identifier. We decided to enhance the important data we had (ASIN number, image url, sales information) with another dataset that could give us the publication date, language in the purpose of

[1] https://openlibrary.org
[2] http://jmcauley.ucsd.edu/data/amazon/

filtering and a some better categories. We turned to the OpenLibrary dataset to do so. We chose this second dataset as it allowed us to gain exactly what we were looking for. The categories for the books are the ones given by the Library of Congress Classification[3]. Matching both datasets was done with the ISBN-10 number, which is the number Amazon uses as their ASIN.

We also explored the possibility to work directly with the images instead of using the provided image features in the Amazon dataset. We used the urls of the images that were present in the initial dataset in order to download them. All the images were adapted in order to have the same size by zero padding. Downloading the images was very useful for the image processing part as well.

3.2 Filtering

We first decided to keep only English books, not only were they the most important part of the data but as we were looking for trends, the less cultural mixture we had the better. We also removed all sparse data and were left with a little over 400 000 books.

The final result of the data scraping, merging the two datasets and filtering is showed in Fig. 1.

4 Data analysis

We start the data analysis by clustering the books by cover. Being leaded from the clustering that covers and categories can be correlated we continued by exploring machine learning techniques in order to see with how much accuracy we can predict the category of the book just by looking at its cover without any natural language processing. We also tried to understand why this happens - from visual perspective - by applying image processing techniques to book covers that belong to the same clusters or book covers that were classified under the same category.

4.1 Hierarchical Clustering

In order to have clustering with adaptive number of clusters, we firstly used agglomerative clustering from the scipy library[4]. This allowed us to explore different number of clusters easily, also explore which categories go in the same cluster, the trend over years in a given category and the sales ranking evolution in a specific cluster. The metric we used for the agglomerative clustering is *ward* which minimizes the variance of the clusters being merged[1]. The clustering was done based on the features generated from book covers[5] present in the initial dataset.

4.1.1 Categories by cluster

By exploring the book categories provided from Open Library we noticed that some categories such as *Language and Literature* contain significantly more books compared to the other categories, which leaded us to the conclusion that it contains subcategories which can be very different. From here comes the hierarchical clustering. In order to easily see the categories that split in separate clusters we equalized the number of books by category. That leaded to interesting results, showed in Fig. 2.

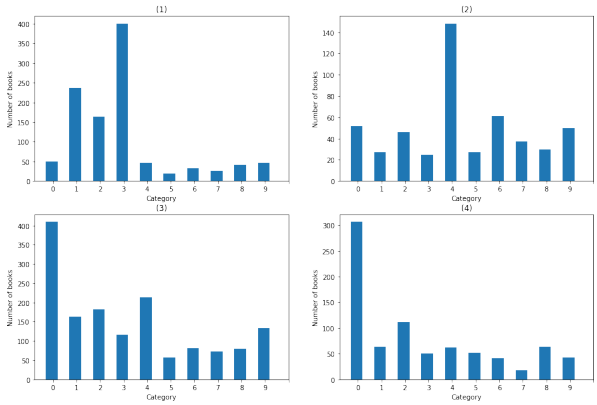

Figure 2: Here we show examples of clusters with dominant category. The clustering is done on features with reduced dimensionality to 100 features using PCA and on randomly chosen books from each category s.t. each category has approximately 4000 books. The dominant categories in (1), (2), (3) and (4) (same as (3)) are "History of the United States and British, Dutch, French, and Latin America", "Language and Literature" and "Auxiliary Sciences of History" respectively.

4.1.2 Years by cluster

Clustering books by covers didn't reveal a lot of information about trends over years. What we expected in this part was to see at least, for fast changing fields, such as books in the category "Technology" to appear in different clusters for distant years such as 2000 and 2007. On the other hand books from the "Auxiliary Sciences of History" have very specific covers in 2003 - 2004 which appear in a separate clusters.

[3] https://en.wikipedia.org/wiki/Library_of_Congress_Classification

[4] https://www.scipy.org

[5] plu

4.1.3 Sales rank by cluster

Using the same covers clustering as before, we tried to analyze if some cluster contained more popular books (based on the amount sold). It turns out that even if the order of magnitude changes, the distribution of `sales_rank` if extremely similar for every cluster. See fig. 3.

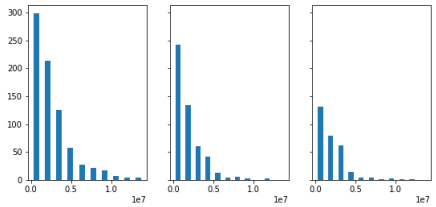

Figure 3: This represents the distribution of sales for 3 random clusters. We noticed that the distribution of sales fallowed the same pattern in every cluster.

4.2 K-Means

With the goal to find more interesting insights for our data we also explored the results with k-means clustering. We varied the dimension of the features $\in \{128, 256, 512, 1024, 2048\}$ and the number of clusters $\in \{10, 20, 30, 70, 80, 100, 200\}$. Category and years distribution in these clusters was not significantly different than the distribution obtained in 4.1. We were able to do this analysis using GPU implementation of K-Means[6].

4.3 Classification

4.3.1 Support Vector Classification

We then try to predict a book's category using only izd cover with the Support Vector Machine algorithm from scikit-learn[7]. Our best approach uses SVC (Support Vector Classifier) with a radial basis function kernel and a one-against-one function for multi-class classification. The main problem with SVM is the time it takes to compute. As a remedy we used PCA to reduce the amount of features we use for each cover. We found out that we could get a descent accuracy with a few categories even if we dropped most of the given features. See Table 1 for the results on a 3-fold cross-validation.

As expected, the accuracy quickly drops when adding more categories. We reached 54.81% using 40'000 books and adding two categories: 'Medicine' and 'Geography, Anthropology, and Recreation'. See the confusion matrix of this model in Fig.4.

Number of features	Accuracy
30	34.03%
300	53.15%
1000	70.17%
2000	71.92%
4096	72.30%

Table 1: The accuracy evaluated while reducing number of features using PCA. The classification is done with SV, with the setting described in 4.3.1. The classification is done for 3 different classes: 'Agriculture', 'Auxiliary Sciences of History' and 'Fine Arts' with total of 23 000 books (balanced in each categories).

Computation over more books and categories was time consuming using this type of classification.

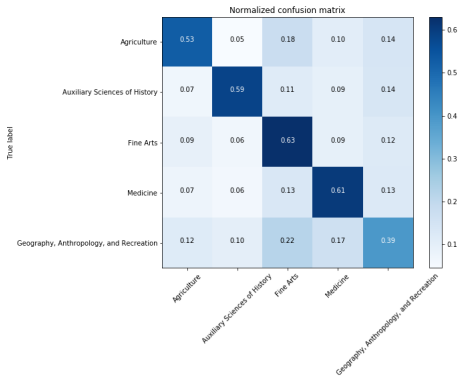

Figure 4: Normalized confusion matrix from the SVC model with 5 categories

4.3.2 Neural Networks

After the encouraging results in 4.3.1 and inspired by Iwana, Brian Kenji, et al. [2] we proceed with Neural Networks. We explored two possibilities for the training dataset: using as training set the precomputed features from the Amazon dataset or using directly the images in a Convolutional Neural Network.

Predicting the category from the cover: To overcome the computational limit of the method presented in 4.3.1 we used neural networks and the advantages of GPU computation. We used StratifiedKFold and StandardScaler from Sklearn library, the former to automatically create a balanced dataset given the labels and to choose correctly hyperparameters for our model. The StandardScaler is recommended for neural network to have every feature normalized between

[6] https://github.com/src-d/kmcuda
[7] http://scikit-learn.org/stable/modules/svm.html

0 and 1 as initially it was not the case. The best results of the model chosen with cross-validation are presented in 5 for 9 different categories. In Fig. 7 we show 10 representative images that are classified with high certitude as part of the given category (with confidence greater than 0.5). We can clearly see that the common pattern that appear on the images are the objects and looks like the classification with high certitude is independent of the colors present on the book cover.

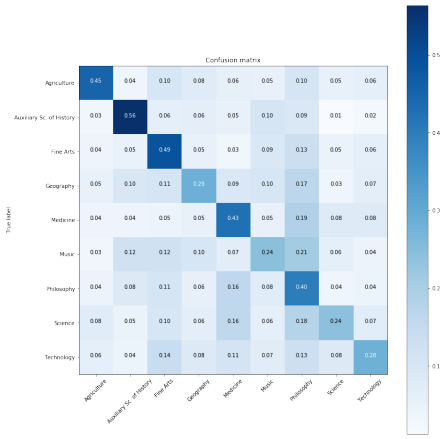

Figure 5: Normalized confusion matrix for predicting categories using neural networks and the features provided in the Amazon dataset.

Predicting the publication date from the cover: We tried to go one step further and predict the publication date from the book cover. For this purpose we used a pretrained CNN model (Xception[8]) with images from ImageNet[9] provided by Keras. The idea was to freeze the already trained part and add two fully connected layers at the output. It relies on transfer learning. The input of this model were directly the images balanced by year. The results of the best model are presented in Fig. 6.

5 Conclusion and future work

As already discussed in other research works [2], the book covers are correlated to the category of the book. We were able to reproduce that in our project and validate it on different categories and on bigger number of books by category. However that cannot be said for the cover change over years. With the features available from the Amazon dataset, as discussed in 4.3.2 and 4.1, we are

[8]https://arxiv.org/abs/1610.02357
[9]http://www.image-net.org/

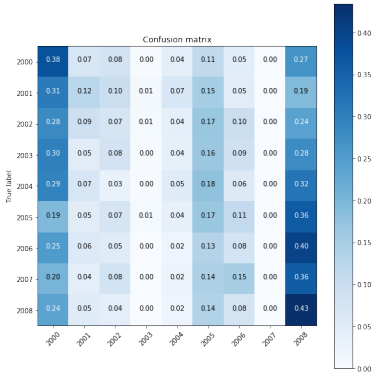

Figure 6: Normalized confusion matrix from the Xception model with 9 categories trying to predict the publication date by the cover of the book, using as input the preprocessed images instead the features provided by the Amazon dataset

not able to predict the publication date or cluster books with close publication date. That can be related to the fact that the features are more concentrated on the objects that appear on the covers and doesn't take in account the colors as shown Fig. 7. It can be possible that with different way of generating the features we can obtain better result on predicting the publication year of the book cover or even the sales ranking.

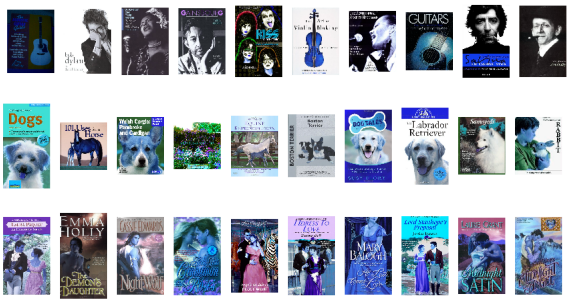

Figure 7: Images classified with high confidence under the category Music (top row), Agriculture (middle row) and Auxiliary sciences (bottom row).

References

[1] Ward, J. H., Jr. 1963. *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, 58, 236–244.

[2] Iwana, Brian Kenji, et al. 2016. *Judging a Book by its Cover*