Fine-Tuning CLIP model with Classified Book Cover Images in PySpark

Danila Rozhevskii
Syracuse University
drozhevs@syr.edu

Kanishk Gupta
Syracuse University
kgupta07@syr.edu

Chinmay Maganur
Syracuse University
camaganu@syr.edu

Abstract

The cover of a book serves as the initial point of contact with potential readers and provides crucial insights into its content. The classification of book genres based on the information can provide immense benefits to modern retrieval systems as well as improve the generation of book covers. Our research aims to classify books into related genres and generate book covers using CLIP. We train and adopt a classification model to predict genres for the books in Pyspark, and feed the resulting dataset of book title prompts and predicted genres into CLIP to generate book covers. CLIP is a generative model that creates images based on textual descriptions. The model employs a vanilla transformer to generate embeddings and a vanilla diffusion model to produce images from those embeddings.

Keywords: NLP, Classification, CLIP, Neutral Networks, Transformers

1. Introduction

Books have been an essential tool for preserving and disseminating information and knowledge throughout the course of human history. We concentrate on the task of categorizing books by genre using details like the author and book title and use that method to improve book cover generation by the fine-tuned CLIP model. Contrary to the adage that you shouldn't judge a book by its cover, book covers frequently influence readers' first impressions of a book and reveal key details about its content. According to research, a book's cover design can have a big impact on how many copies it sells, with book sales usually rising after a design change. We used a machine learning model to categorize the books into different genres to help with our book classification efforts. For the purpose of creating book covers utilizing titles, we integrated the CLIP model. Contrastive Language-Image Pre-training, or CLIP for short, is a successful technique for learning from natural language supervision. The model is made up of two submodels, one for text encoding and the other for image encoding, which embeds text and images into different mathematical spaces. Our study demonstrated that CLIP was four times more effective at zero-shot ImageNet accuracy than competing techniques when tested by OpenAI.

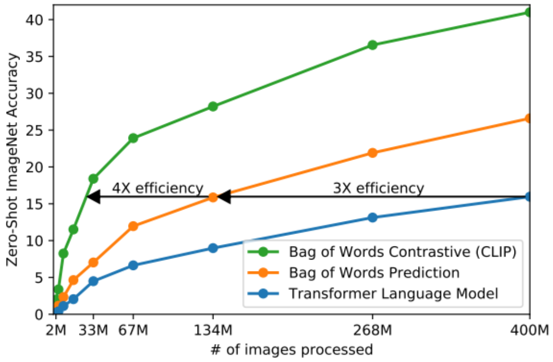


Figure 1: Effectiveness of CLIP Model

2. Related work

Several studies have attempted to classify book genres using machine-learning techniques. Chiang et al., 2015 used transfer learning with convolutional neural networks on book cover images and natural language processing on book titles to classify five genres. Iwana et al., 2016 curated a massive dataset of 57,000 book covers belonging to 30 different genres and managed to achieve a Top 1 prediction accuracy rate of 24.7% and a Top 3 prediction accuracy rate of 40.3%. respectively. Buczkowski et al., 2018. crawled 160k book covers from GoodReads.com and grouped them into 14 classes, achieving an accuracy of 67% and 74% using two different convolutional neural networks. Another study utilized a logistic regression model to categorize book genres by using both image and title information and achieved an accuracy rate of 87%.. Finally, Lucieri et al., 2020 benchmarked deep learning models on a dataset of 55,100 samples, achieving the highest accuracy of 55.7% using a text-image model. Text-to-image generation has garnered significant interest and research lately. Previous work, such as. Mansimov et al., 2015, Reed et al., 2016 application of cGAN, and H. Zhang et al., 2017 stage-wise image generation approach, has primarily focused on generating high-quality images from text descriptions. VirTex Desai and Johnson, 2021, ICMLM Bulent Sariyildiz et al., 2020, and ConVIRT Y. Zhang et al., 2022 used transformer-based language modeling, masked language modeling to learn image embeddings. However, the major difference is related to the scale of the models or the amount of data they use.

3. Method

3.1. Approach

The main method of generating the most closely related images of book covers from the book titles was to add a category description to those titles. That is why, the first part of the project was to find the best-performing classification algorithm that could catch the relationship of the book title and its category. The second part of the project is to use the pre-trained classification model to attach the category description to book covers and fine-tune the ViT-B/32 CLIP model with respective book covers to generate a variety of (image, text) pairs.

3.2. Dataset

There are two datasets used in the project; one is used for classification model training and another one for fine-tuning the CLIP model.

The first dataset should have been rich and large enough to make a good generalization of the relationship between book titles and author names. The data used for the classification process had 207,572 books from the Amazon.com, Inc. marketplace. After cleaning and pre-processing steps, the dataset size shrunk to around 193,000 books.

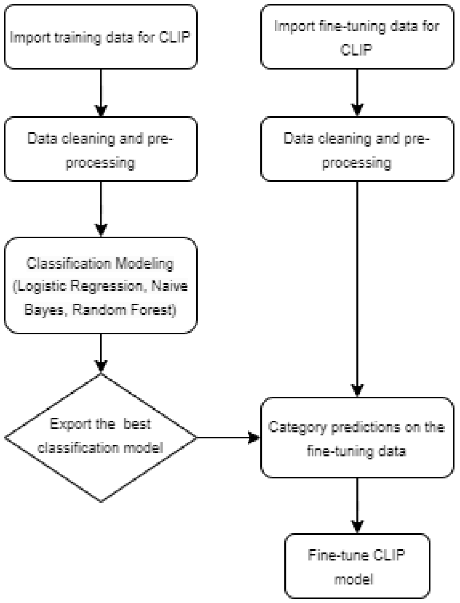
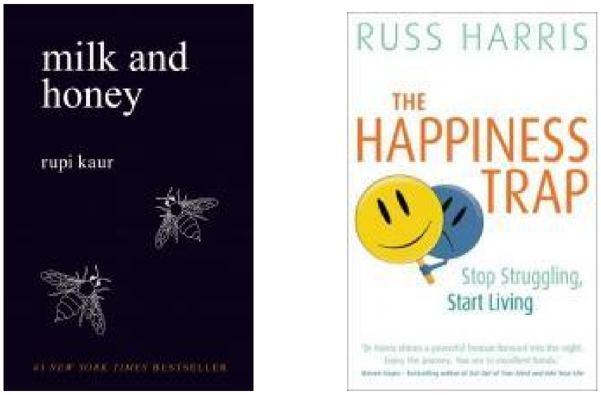

Import training data for CLIP
Import fine-tuning data for CLIP
Data cleaning and pre-processing
Data cleaning and pre-processing
Classification Modeling (Logistic Regression, Naive Bayes, Random Forest)
Export the best classification model
Category predictions on the fine-tuning data
Fine-tune CLIP model

Figure 2: Project workflow schema.

The second dataset consists of 20,590 book cover images and their titles. The idea is to train the classification algorithm on the first dataset and then apply the model to the second one, which is later used for fine-tuning CLIP and the generation of covers given the book title prompt with the predicted genre.



(a) Example 1.

(b) Example 2.

Figure 3: Book covers for fine-tuning CLIP

3.3. Classification model

This step consisted of researching the best classification technique. The models tested included: Logistic Regression, Naive Bayes, and Random Forest. After performing cross-cross validation on all three, the best configurations were chosen and are present in the table below.

Model	Accuracy (ROC)
Naive Bayes	%59.8
Logistic Regression	%58.1
Logistic Regression with CV	%56.9
Random Forest	%54.3

From the table above, the model with the highest performance is the Naive Bayes. The low accuracy on all the models indicates that there is a weak correlation between a book’s title and its category, which makes sense in real life. Some books’ titles do not give any clues about the category they belong to. However, since it is a hard task even for a human brain, we believe that this approximation is still enough to make a good guess about a book’s category based on its title. We will use the pre-trained Naive Bayes model to make category predictions for the book title prompts in the second dataset used for fine-tuning the model.

3.4. Fine-Tuning CLIP

With the dataset of pairs of images and book titles with predicted categories, we built the pipeline to transform and feed data into a pre-trained CLIP model to fine-tune for the task of relating book cover images and book titles with categories. Since CLIP is a large model and works with visual data, we utilize PySpark’s TorchDistributor package to create a series of TorchDistriubutor objects and parallelize the fine-tuning (training) process of the CLIP model. It allows us to save computational resources and fine-tune the model on large-scale data

4. Results

After fine-tuning CLIP on a part of the dataset, the results show the model’s capability to recognize and associate appropriate book cover to the title prompt provided. In the future, we plan to add a predicted category to the prompt to increase the similarity accuracy of book cover generation.

Title Prompt	Similarity Score
”Lord of Rings”	%12.4
”The Power of Nowww”	%86
”random book”	%0.8

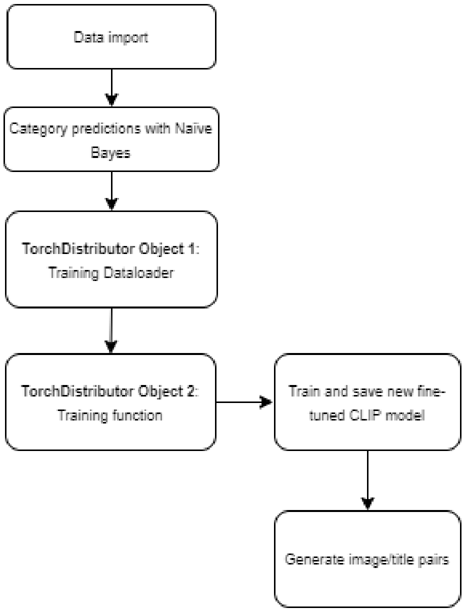

Data import
Category predictions with Naive Bayes
TorchDistributor Object 1: Training Dataloader
TorchDistributor Object 2: Training function
Train and save new fine-tuned CLIP model
Generate image/title pairs

Figure 4: CLIP’s parallel fine-tuning.

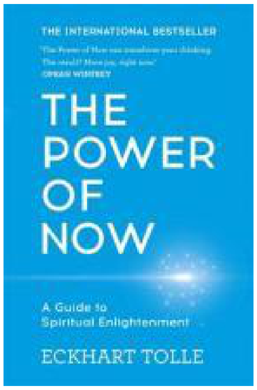

THE INTERNATIONAL BESTSELLER
THE POWER OF NOW
A Guide to Spiritual Enlightenment
ECKHART TOLLE

Figure 5: The predicted book cover from the fine-tuned CLIP with title similarities generated.

In Figure 5, we use the example book cover and generate a few book title prompts to test for similarity with the book cover. From the table above, we can see that the closest title is estimated to be ”The Power of Nowww” which almost says the same thing as the actual book cover, and the model is able to understand the similarity. In the future, if we add a predicted category to each prompt and fine-tune the model again, we will be able to increase the similarity scores.

5. Conclusion

In this study, we evaluated three classification algorithms for the categorization of book genres based on cover images: Naive Bayes, Random Forest, and Logistic Regression. With an accuracy of 58%, our results showed that Logistic Regression with default settings and title/author information performed the best. Logistic Regression and Random Forest also performed well with ROC of 58% and 54%, respectively. Building on these results, we employed the CLIP model to generate images based on the book's title, author, and genre information. In the future, we plan to improve the classification model and use that to predict book categories and use the updated training data to fine-tune CLIP. This approach presents a novel solution to the problem of generating images for book covers based on their genre and provides a useful tool for publishers and authors to visualize the potential covers for their books.

6. Discussion

We evaluated the feasibility of three classification methods for book genre categorization based on cover images: Naive Bayes, Random Forest, and Logistic Regression. However, all models showed relatively low accuracy, which could be attributed to possible class imbalance and the limited number of genres, leading to underfitting, as well as the fact that book titles often do not directly correspond to the genre they are in. To boost the classification models' accuracy, future research can focus on addressing these issues. We found that scraping data from websites like amazon.com, Google Books, and Goodreads.com can provide additional features that can help improve the accuracy of the model. If we are able to add more information as a description for each book title, we will be able to increase the accuracy of category prediction.

For now, we fine-tuned a CLIP model only on a fraction of the data, as an experiment. We only use pairs of book titles with author and book cover images. However, in the future, we plan to use the improved classification model to predict category labels and increase the accuracy of generated CLIP's similarity scores between book cover images and book titles with author and category information. The final product of this research is supposed to take a new book title prompt and its category and be able to generate a book cover image based on the similarity scores between the prompt and its training data.

Furthermore, to improve the user experience and code, we suggest developing a user interface for uploading book titles and receiving real-time category predictions along with book covers generated to closely relate to the user's input.

References

Buczkowski, P., Sobkowicz, A., & Kozlowski, M. (2018). Deep learning approaches towards book covers classification. *ICPRAM*, 309–316.

Bulent Sariyildiz, M., Perez, J., & Larlus, D. (2020). Learning visual representations with caption annotations. *arXiv e-prints*, arXiv–2008.

Chiang, H., Ge, Y., & Wu, C. (2015). Classification of book genres by cover and title. *Computer science: class report*.

Desai, K., & Johnson, J. (2021). Virtex: Learning visual representations from textual annotations. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11162–11173.

Iwana, B. K., Rizvi, S. T. R., Ahmed, S., Dengel, A., & Uchida, S. (2016). Judging a book by its cover. *arXiv preprint arXiv:1610.09204*.

Lucieri, A., Sabir, H., Siddiqui, S. A., Rizvi, S. T. R., Iwana, B. K., Uchida, S., Dengel, A., & Ahmed, S. (2020). Benchmarking deep learning models for classification of book covers. *SN computer science*, *1*, 1–16.

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015). Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. *International conference on machine learning*, 1060–1069.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2022). Contrastive learning of medical visual representations from paired images and text. *Machine Learning for Healthcare Conference*, 2–25.