Judging a book by just its cover (& title)
Project Report

Members	Angel Gantzia	angel_g@mit.edu
	Pavena Vongkhammi	paven189@mit.edu
	Nick Antoniou	nanto@mit.edu
	Dido Stoikou	dido_st@mit.edu

March 11th, 2024

Colab Notebook Link

Colab Notebook Link
Drive Folder (for access to the data): 🗀 HODL_book_covers

I. Problem Description

In the digital age, book covers play a pivotal role in influencing consumer purchasing decisions, particularly in impulsive buying scenarios on platforms such as Amazon. Despite the saying "Don't judge a book by its cover," we propose the hypothesis that the design of a book cover can significantly impact its marketability and sales performance. As such, this project seeks to investigate the correlation between the visual appeal of book covers and its likelihood of becoming a bestseller (using a binary predictor of best-sellers and non best-sellers).

To this end, we have devised a methodology that involves scraping a dataset from Amazon across various categories that usually have visually intriguing covers within Amazon's best selling Kindle books lists. Then, we employ custom convolutional neural networks (CNNs) and use transfer learning from established models like ResNet and VGG to make predictions from the visual elements of book covers. Moreover, in an extension of our original project proposal, this project also explores a multimodal approach by integrating book titles with cover images to analyze our predictions, recognizing the multifaceted nature of consumer choice. This extension of work imitates the consumers' journey as they browse and buy books within Amazon's platform, and provides a more nuanced understanding of the interplay between textual and visual stimuli in consumer behavior.

From a business perspective, the utility of this project lies in offering valuable insights for publishers, authors, and retailers in strategizing more effectively on how to design and market their books. By using our predictive models, publishers and independent authors have a reliable way of informing cover design, potentially transforming conventional practices in the publishing industry. Stakeholders can link factors that resonate most with consumers, and adjust them accordingly by running multiple tests in favor of improving the odds of a book reaching the coveted best-selling status. Lastly, online retailers such as Amazon can factor in their recommendation engine, the visual appeal of books, to position them on the pages and boost their overall sales.

Ultimately, this research bridges the gap between machine learning and market strategy, aiming to predict whether people do indeed let visual determinants affect their buying behavior on platforms like Amazon. We aspire to guide the book industry toward a more data-driven, empirical approach to cover design and marketing, enhancing the commercial success and visibility of books in the digital marketplace.

II. Dataset

Initially, we scraped the Amazon website for the book dataset that was needed, seeing that scraping data is a great way to create your own dataset, especially from a website that has such a plethora of books and their metadata. We scraped the best-sellers lists, aiming to create a dataset with a variety of books, including their metadata, such as the title and- more importantly- their cover. We, also, tried to scrape the descriptions of the books themselves, to increase the training data of our LLMs, but this was met with specific scraping protections, such as CAPTCHA (since that would require visiting thousands of book's web pages). We found ways to circumvent that, but they were out of the scope of this project to implement, because they ranged from relatively ineffective and quick to helpful and expensive. As such, we complemented the Amazon best-sellers dataset with a Kaggle dataset.

Specifically, the initial Kaggle dataset contained 130,000 Kindle ebooks, including their ASIN code, title, author, link to their cover online, their product page on Amazon, and some other metadata. We selected the six categories that we expected to have illustrative covers, that would allow for the training of models to carry out the task at hand. Namely, those genres were: "Mystery, Thriller & Suspense", "Children's eBooks", "Teen & Young Adult", "LGBTQ+ eBooks", "Literature & Fiction", "Science Fiction & Fantasy", "Romance".

A problem we faced was the imbalance present in the dataset, concerning the split between best-sellers (663) and non best-sellers (32,972). We combined our initial 1,200 scraped best-sellers, with the 663, and included an equal number of non best-sellers from the Kaggle dataset. The dataset was slightly imbalanced between the genres, but we did not expect this to pose a problem to the task at hand.

Afterwards, we scraped the book covers of these observations, to be used in the training of the vision models (this scraping did not meet the same scraping obstacles as described in this section). Lastly, we pre-processed the images, normalizing and resizing them. We experimented with cropping and flipping the images, to enhance our dataset, but we did not expect this to meaningfully affect the accuracy. It would significantly alter the appearance of the images and affect the predictions in a way that is different- and more pronounced- than some other applications, where this image enhancement step yields significantly better results (like object detection).

III. Approach

To approach this problem, we use custom and pre-trained deep learning architectures to dissect the visual influence of book covers (and later also for titles). To begin, we started off with a custom convolutional neural network (CNN), engineered to capture the unique

attributes of book covers. Then, we applied transfer learning techniques, using the ResNet and VGG architectures pretrained on ImageNet. This approach of incorporating deep, layered feature representations, which are learned from a wide range of visual data aimed to improve our accuracy, by potentially generalizing better than our custom model.

Lastly, we implement a multimodal approach, integrating textual data from book titles with visual data from covers. This dual-stream analysis aimed to investigate whether our model's predictive capability would improve with both textual and visual elements and whether they collectively influence consumer purchasing behavior. Across all models, image data preprocessing involved resizing the cover images to a consistent shape of 224x224 pixels and normalizing pixel values to the [0, 1] range. This normalization is critical for deep learning models as it helps in faster convergence.
Training was conducted on a split of the data, reserving 10% as a validation set to monitor the model's performance on unseen data. We utilized Google Colab's GPU acceleration to expedite the training process, significantly reducing the time required for model convergence. During training, we carefully monitored the loss and accuracy metrics for both training and validation sets. The goal was to achieve a balance where both training and validation accuracy were high, indicating good model generalization. To this end, hyperparameters such as the learning rate, batch size, and the number of epochs were tuned based on the model's performance on the validation set.

A. CNN

Unlike conventional object detection tasks, our focus lies in deciphering the intricate visual cues embedded within book cover images to predict their success in captivating consumers. So to benchmark against a baseline of roughly 50% (since the collection of images we created is roughly equal), we adopt a custom CNN architecture to initially assess accuracy and subsequently enhance performance with other techniques. For the CNN, we use three hidden layers for a balance between computational efficiency and model complexity, optimizing our resource utilization.
To expedite model training and leverage our Colab resources efficiently, we employ random search with 10 trials to fine-tune our parameters thoughtfully. We select batch sizes of 64, 128, and 256 during CNN fine-tuning to explore a range of options that balance computational efficiency with model generalization. Smaller batch sizes, such as 64, afford finer adjustments during training, potentially capturing more nuanced patterns in the data. Conversely, larger batch sizes like 256 provide more stable gradient updates but may demand higher memory resources.
Our choice of epochs, ranging from 5, 10 or 15, aims to strike a balance between optimizing computational time and avoiding underfitting or overfitting. We avoid excessively low epochs to prevent underfitting and excessively high epochs to mitigate overfitting risks, thereby optimizing our model's learning process. Regarding learning rates, we explored rates between 1e^-5 and 1e-3, using a log-scale distribution to capture the exponential nature of

learning rate effects. This range is broad enough to include conservative learning that ensures stability and aggressive learning that accelerates training.

Furthermore, we carefully tune weight decay using values of 0, 1e-4, and 1e-3 in our custom CNN to introduce regularization and mitigate overfitting risks. Weight decay acts as a form of regularization by penalizing large weights in the model, effectively constraining its complexity. By fine-tuning this parameter, we strike a balance between preventing overfitting and enabling the model to capture crucial patterns in the data, ensuring robust generalization to unseen scenarios.

B. Resnet152

Our task is not to detect objects found in images, but rather analyze book covers, and from a dataset of very diverse visual information to find whether they will succeed in making consumers buy them. As such, ResNet152, with its 152 layers, provides a depth that helps in capturing a wide variety of patterns and features in book cover images, ranging from simple textures and colors to complex compositional elements. This depth is particularly beneficial in our context, where the visual appeal of a book cover can significantly influence consumer purchasing decisions. The model's ability to learn such detailed representations makes it an ideal choice for discerning the nuanced visual cues that could indicate a book's bestseller potential.
When training the model, we decided to freeze the earlier layers due to considerations of our computational constraints, as well as the nature of our task. By freezing these layers, we could leverage the pre-trained weights that are already adept at capturing generic image features, thus conserving computational resources and focusing the model's learning capacity on the upper layers more relevant to our specific task. This approach not only optimizes our computational efficiency, especially important given the limited computing power available on Colab Pro, but also tailors the model's learning process to focus more on the advanced features that are likely to be more predictive of a book's bestseller status.

The rationale behind our choices for hyperparameter space was to align our objective to optimize our model's performance with the computational limits of Colab Pro. For batch sizes, as we mentioned above, smaller ones can navigate the training landscape with more granularity, while larger sizes offer more stability in the gradient updates but increase memory requirements. For our ResNet model, we lowered the batch sizes to 32 and 64, because they fit better to our problem, balancing the computational demand with the ability to generalize from training data. For the number of epochs, we used the same range as before—5, 10, and 15—to determine the appropriate duration for model training, and the same learning rates range and weight decay for regularization.

Lastly, the choice of criterion—BCEWithLogitsLoss or CrossEntropyLoss—was based on the nature of our binary classification task. Usually, BCEWithLogitsLoss is the most suitable

option for binary outcomes and provides a stable version of the BCE loss by combining it with a sigmoid layer. CrossEntropyLoss is generally used for multi-class classification problems where the outputs are mutually exclusive, but we tested its application in our binary context for a comprehensive understanding of its effects. In implementing random search, we varied these hyperparameters within their defined spaces to identify the combination that yielded the best validation accuracy. We did that so we could survey a wide array of configurations, an essential step given that even small changes in hyperparameters can significantly affect model performance.

C. VGG16

VGG16 is a convolutional neural network architecture developed by the Visual Geometry Group (VGG) at the University of Oxford. It gained widespread popularity after achieving exceptional performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014. The network comprises 16 layers, including 13 convolutional layers and 3 fully connected layers. Its uniform architecture, consisting of small 3x3 convolutional filters stacked one after another, enables it to capture intricate features within images effectively. Due to its simplicity and effectiveness, VGG16 remains a popular choice for various image classification tasks, making it well-suited for our book cover analysis.

To enhance the effectiveness of our model, we decided to freeze the earlier layers of the VGG16 architecture, focusing our fine-tuning efforts solely on the last layer. By doing so, we aim to capitalize on the pre-trained weights of the earlier layers, while allowing the model to learn dataset-specific patterns and nuances more effectively. For hyperparameter tuning of this model, we kept the same hyperparameter tuning space as Resnet, so as to have a fair comparison between the two.

D. Multimodal Classifier (Vision + text)

Apart from the above, we decided to also fit a predictive model that would determine the bestseller status of books using a multimodal approach, integrating both text (book titles) and image (book covers) data. Our task was conducted using the TensorFlow library.

The initial step involved preprocessing the textual and visual data to make it suitable for the neural network. The dataset consisted of book titles and corresponding cover images, stored in a CSV file and a directory of images, which we scraped previously. For text preprocessing, book titles were first converted to lowercase to standardize the text input. We then employed the Tokenizer class from Keras to tokenize the titles, converting them into sequences of integers, where each integer represented a specific word. These sequences were padded to ensure uniform length, facilitating batch processing during model training.

The multimodal architecture was designed to process textual and visual inputs separately before merging. For text, an Embedding layer followed by an LSTM layer was used to capture the sequential nature of text data. Since it had mostly superior performance, for images, we utilized the VGG16 model as a feature extractor, with its output flattened and fed

into subsequent layers. The features extracted from both modalities were then concatenated and passed through fully connected layers, culminating in a single neuron with a sigmoid activation function for binary classification (bestseller or not). To address potential overfitting, we considered several strategies, including dropout layers and regularization techniques, although the specific implementation details of these were dependent on model performance during the validation phase.

The model was compiled with the Adam optimizer and binary cross-entropy loss function. After training, the model's effectiveness was evaluated based on its accuracy in classifying books as bestsellers or not on the validation set. This evaluation provided insights into the model's ability to generalize and its potential applicability in real-world scenarios.

IV. Results

A. Results Table

Best Model	Best Hyperparameters	Test Accuracy
Multimodal	Batch size: 64 Number of epochs: 50 Learning Rate: 0.001 Weight Decay: - Criterion: Binary Crossentropy	80.81%
VGG	Batch size: 64 Number of epochs: 15 Learning Rate: 0.0008043319923577275 Weight Decay: 0.0001 Criterion: CrossEntropyLoss	80.19%
ResNet152	Batch size: 32 Number of epochs: 10 Learning Rate: 0.0007351570049329399 Weight Decay: 0.001 Criterion: CrossEntropyLoss	79.87%
CNN	Batch size: 128 Number of epochs: 10 Learning Rate: 0.00039046224929057674 Weight Decay: 0.001 Criterion: CrossEntropyLoss	78.62%

B. Results discussion

With CNN, we achieved a test accuracy of 78.62% using the aforementioned hyperparameters. This marks a significant improvement compared to the baseline accuracy of 49.37% (due to the negligible imbalance of classes). By implementing a custom CNN, we have enhanced the model's performance by approximately 59.25%, demonstrating its promising potential for our task.

On the other hand, with ResNet, we achieved a test accuracy of 78.62%. This marks a significant improvement compared to the baseline accuracy of 49.37%. Contrary to our expectations, the accuracy of the model only had a marginal improvement compared to our custom model. Further investigation is needed, but one explanation for this could be that ResNet152, though very powerful, is a more general architecture. Sometimes, a well-optimized smaller network can outperform a larger, more complex one depending on the specific data and task.

With VGG16, we achieved a test accuracy of 80.19%. This represents a remarkable improvement over the baseline accuracy of 49.37%. By leveraging VGG16, we have boosted the model's performance by approximately 62.43%, showcasing its efficacy for our task. Additionally, VGG16 exhibited a slight 0.4% improvement over ResNet152, reaffirming the effectiveness of both models in our analysis.

Lastly, our multimodal deep learning model was trained and validated over 50 epochs, demonstrating promising results in the task of predicting bestseller books from their titles and cover images. The model achieved a validation accuracy of 80.81%, indicating a robust ability to generalize and accurately classify unseen data. The training process utilized a batch size of 64, which was determined to be optimal for the available computational resources while still maintaining efficient learning dynamics. The learning rate was set at the default value of 0.001, a common choice for the Adam optimizer that balances rapid convergence with the stability of the training process. The model was not explicitly regularized with weight decay, suggesting potential for further performance enhancements. This default learning rate, along with the binary cross-entropy criterion for the loss function, enabled the model to effectively learn from the multimodal dataset.

Comparing our multimodal approach to the image-only models gives us a clear picture of the benefits of using both text and images. Our multimodal model, which combines book titles and cover images, achieved an accuracy of about 80.81%. This is slightly better than the image-only models we tried, where the VGG model reached 80.19% accuracy and the ResNet152 model was close behind with 79.87% accuracy. This shows us that including the book titles gave our model a small but valuable boost. It suggests that the titles of the books carry important information that helps the model make better predictions about whether a book will be a bestseller. Using both types of data—text and images—seems to give us a more complete understanding than using just one type on its own.

II. Lessons Learnt

A. Business insights

The project shed light on the correlation between book cover design and consumer purchasing behavior, offering valuable insights into the digital marketplace. By understanding the visual appeal of book covers and its impact on sales performance, businesses can refine their understanding of consumer decision-making processes. Publishers can use the predictive models to inform cover design decisions for upcoming book releases. By understanding which visual elements resonate most with consumers, publishers can create compelling book covers that attract attention and drive sales. Independent authors can also benefit from the insights provided by the project to optimize their book cover designs. By aligning cover designs with consumer preferences and market trends, authors can enhance the marketability of their books and improve their chances of success. Online retailers such as Amazon can leverage the findings to enhance their recommendation engines and product positioning strategies. By considering the visual appeal of books alongside other factors, retailers can optimize product placement on their platform to drive sales and improve the overall shopping experience of their customers.

B. Approach lessons

1. Greyscale book covers

Throughout the training of our models, we plateaued on the accuracy at around 80%. To deal with that we reflect on the avenues not explored that could potentially enrich our understanding of consumer behavior. One intriguing aspect that warrants further investigation is the examination of book covers in greyscale. When we consider how readers buy e-books to read on their Kindle devices, they can do it in one of two ways. Particularly, they can buy them from Amazon's website, or they can buy them from their e-reading devices. If they take the latter approach, the book covers are predominantly displayed in greyscale. That is why, understanding the impact of a cover's visual appeal when viewed in greyscale could offer additional insights into the purchasing decisions of e-reader users.

While our study focused on full-color images, mimicking the typical online shopping experience, we recognize that the monochromatic display of Kindle devices could alter which visual elements catch a user's eye. By analyzing book covers in greyscale, we could identify which features remain salient without color cues and potentially uncover design elements that are compelling in the absence of color. Such findings could guide publishers and authors in creating cover designs that are not only attractive in color but also have the visual contrast and composition to stand out on e-readers, potentially influencing the purchasing decisions of a significant segment of readers.

2. Lack of explainability

One of the major challenges in deploying deep learning models, such as the ones used in our project, is the lack of explainability. Deep neural networks, known for their "black-box" nature, make it inherently difficult to trace how exactly they make their predictions. This opaqueness stands in contrast to more interpretable models like decision trees, which provide clear paths and rules that lead to their output. In a visiolinguistic task, where both image and text data play a crucial role, understanding the model's reasoning can be as important as the accuracy of its predictions, especially when the outcomes impact real-world decisions and policies.

The limitations of this lack of transparency can be significant. For instance, without insight into the model's decision-making process, we cannot easily detect if the model is relying on spurious correlations or biased data. This is particularly crucial in book prediction models, where biased decisions could lead to certain types of books being consistently favored or overlooked, perpetuating existing market trends without a basis in individual merit.

Moreover, in the context of user trust and regulatory compliance, explainable AI (XAI) is gaining prominence. Users and regulators may require clear explanations for automated decisions, particularly when these decisions affect authors, publishers, and the broader literary market. The absence of interpretability poses a significant hurdle; even if a model accurately predicts bestsellers, the lack of clear reasoning behind its choices means we cannot confidently offer actionable recommendations. Without understanding the "why" behind a model's predictions, we cannot guide authors or designers on how to craft book covers and titles that might increase the likelihood of becoming a bestseller. This gap in interpretability limits the practical utility of our model as a tool for informed decision-making in the publishing industry. The nuanced art of cover and title design remains, to a degree, a matter of creative intuition rather than analytical deduction when solely relying on these predictive models.

To address this gap, we propose employing techniques like Optimal Classification Tree (OCT) mirroring. This technique aims to create a more interpretable parallel model that approximates the decision-making process of the neural network. By fitting the outputs of the neural network to an OCT, we can analyze which features are given more importance by the neural network. The OCT mirror can then offer a simplified, yet insightful, view of the decision process, highlighting which parameters—such as specific visual elements of the cover or keywords in the title—most influence the prediction.

While OCT mirroring can enhance interpretability, it is important to acknowledge that the simplification necessary for explainability may not capture all nuances of the deep learning model. Thus, it offers a trade-off between interpretability and the model's complex capabilities. Despite this, enhancing model transparency is a step forward, potentially leading to more trust in the model's decisions, easier identification of biases, and a better understanding of the visiolinguistic factors that contribute to a book's success.