Judging a Book by it's Cover and Title

1st Prathu Baronia
14D070046
Indian Institute of Technology, Bombay
India

2nd Balraj Parmar
14D070001
Indian Institute of Technology, Bombay
India

***Abstract*—Don't judge a book by it's cover is a common phrase in English which tells us not to evaluate a book from its cover page only. In this project we aim to do exactly this and show that it is possible to predict the genre of a book from the cover page art and the title of the book with a considerable accuracy. Transfer Learning approach is used to extract features from the dataset using VGG-16 Convolutional Neural Network which was pretrained on the Imagenet Dataset. A simple neural network of three layers is trained using this features as input with a validation accuracy of 72.3%. A Bag of words approach has been adopted to extract features from the book titles post pre-processing the dataset and a validation accuracy of 78.6% was achieved through a Random Forest Classifier.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The cover of the book is the first thing that we observe in a book. It is our first introduction to the book and along with the title functions to communicate to the reader about the style, tone and content of the book. Humans are naturally trained through their life experiences to guess the genre or the style of the book from the cover page and title with reasonable accuracy. In this project, we aim to train a model to predict the genre from the cover page and title using a Image Classifier and a Text Classifier.

The difficulty of the task arises from the fact that there is a large variety in book covers and the title. Also, unlike many other classification problems where there the classes are more starkly defined, There is a lot of ambiguity about how a genre is defined and book covers are sometimes also intentionally misleading. There are innumerable books that have been written till now, and over the times there has been a lot of changes in how the genre is defined, the way artwork is used on the cover page to attract the viewer and so on. And so its important to select our data such that it can be a good representative of different genres, and which can be easily utilized by a Machine Learning model.

Convolutional Neural Networks(CNN) is a class of deep, feed-forward artificial neural networks that has successfully been applied to 0 visual imagery. They are the current state of the art method for image classification. But it has been observed that they generally require hundreds of thousands of images to perform accurately without over fitting. Transfer Learning has been suggested as one of the solution when our dataset is limited. It uses a CNN trained over millions of images to extract features(which are a good, compact representation of the data present in the images) and then use these features to train any classifier like a neural network or SVM etc.

We have used this approach in our solution. We have used VGG16 [2] network pretrained on ImageNet database to extract features from our dataset, We have then created and trained a fully connected neural network on this features to classify our data in five genres.

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). Also known as the vector space model. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model has also been used for computer vision. The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier.

II. Related Works

For the particular problem of genre prediction from book cover, Iwana et.al [3] compared custom LeNet [4] and pre-trained AlexNet network for the genre classification problem. They found that the Top-1 accuracy in the transfer learning approach with AlexNet was twice as good as the custom made LeNet network. They didnt made use of the text titles of book at all. They relied only on images to predict the genre.

Gatys et.al [5] have used deep CNNs to learn and copy the artistic style of paintings. There have been attempts at classifying music by genre, paintings and text. However, a large number of these methods use hand designed features.

III. Image Classification

A. Convolutional Neural Networks

Unlike traditional artificial neural networks for image classification which requires the image to be in a vectorized form (which results in loss of spatial information from the image), CNNs takes input in the matrix form (either 2D if its a gray scale or a binary image, or in multidimensional matrix in case of color spaces like RGB). CNNs comprises of three basic components: convolutional layers, pooling layers, and fully connected layers.

a) Convolutional Layer

: The basic operation that gives the name to the model is the convolution operation. An important feature of CNNs which helps in minimizing the number of parameters compared to a Fully Connected Network of similar architecture is the idea

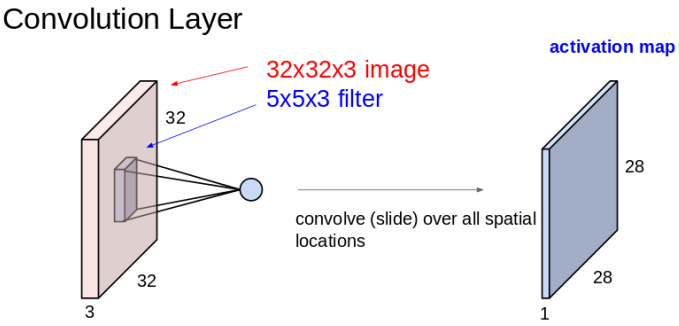


Fig. 1: Unit of a convolution layer

of weight sharing. All the neurons/kernels in a particular convolutional layer have the same weights which are learned during training. The output of this layer would be equivalent to what you get if you would have convolved the input matrix with a kernel equivalent to one of the neurons.

b) Pooling Layer

: Another important feature of CNNs is its pooling layer, it is a down-sampling operation which serves two purposes: it helps in reducing number of parameters(and hence amount of computation) and it helps in firming the fact that the exact location of a particular value in the feature map is less important.

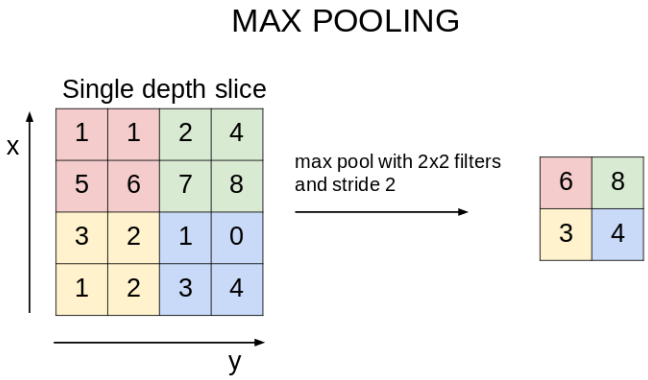


Fig. 2: Max Pooling

c) Fully Connected Layer

: All the neurons in this layer are connected to all the neurons in the previous layer just like a traditional neural network architecture.

d) Activation Functions

: As far as activation functions are concerned, Activation functions like sigmoid and tanh suffered from the problem of vanishing gradients and the training time increased. Recently, ReLu (Rectified Linear Unit) (ReLu(x) = max(0, x)) has been the most favorable activation function among researchers owing to the fact that it speeds up the training time significantly without deteriorating the accuracy.

The CNNs have been back in limelight after AlexNet [1] showed remarkable performance in ImageNet challenge 2012.

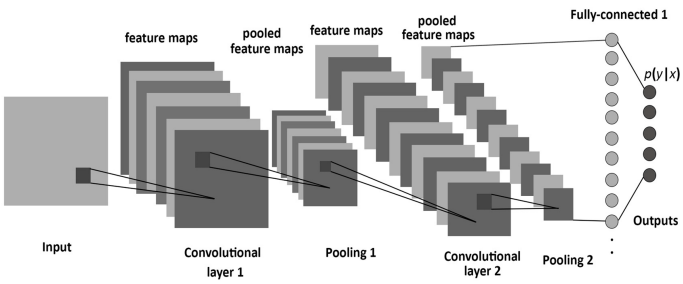


Fig. 3: Fully connected layer in a CNN

B. Transfer Learning

One caveat when using CNNs is that, it often requires hundred of thousands of training samples to give good accuracy without over fitting. But often we find that, our dataset has only few hundred or thousand samples. Transfer Learning has been suggested as one of the methods to use in such scenarios. It has been found that features learned by a deep CNNs on one dataset of images with millions of images are highly transferable to another problem. This is especially true for natural images. So a pretrained network is used to extract features and these features are used to train a classifier like a neural network or SVM.

C. VGG16 network

VGG16 networks philosophy is uniformity and deep layers. It uses a uniform 3x3 filters in all the layers. It showed that increasing the depth indeed results in better performance over ImageNet dataset. A 16 layer and a 19 layer version are publically available for academic use.

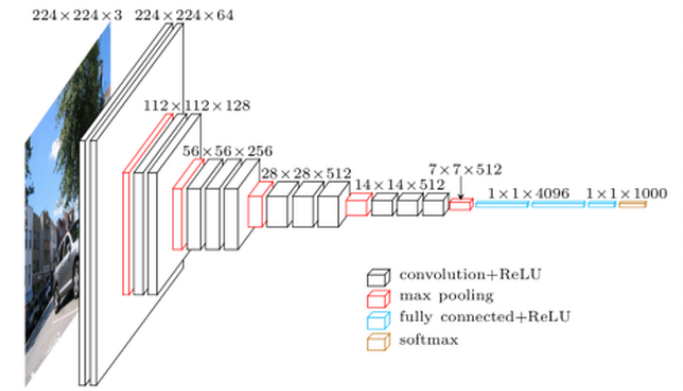


Fig. 4: Architecture of VGGNet

Here in our solution to the problem, we have used pretrained VGG16 network to extract features from our dataset. A three layer network is then trained using these features to classify into 5 genres.

IV. Bag of Words

A problem with modeling text is that it is messy, and techniques like machine learning algorithms prefer well defined fixed-length inputs and outputs. Machine learning algorithms

cannot work with raw text directly; the text must be converted into numbers. Specifically, vectors of numbers.In language processing, the vectors x are derived from textual data, in order to reflect various linguistic properties of the text.This is called feature extraction or feature encoding. A popular and simple method of feature extraction with text data is called the bag-of-words model of text.

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

- A vocabulary of known words.
- A measure of the presence of known words.

It is called a bag of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. The intuition is that documents are similar if they have similar content. Further, that from the content alone we can learn something about the meaning of the document.The bag-of-words can be as simple or complex as you like. The complexity comes both in deciding how to design the vocabulary of known words (or tokens) and how to score the presence of known words.

A. Example of Bag of Words

TABLE I: Example Sentences

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness,

Here in TABLE I we present a set of example sentences taken from a popular book.
Vocabulary is the list of unique words in the input. Here the vocabulary is : { 'it', 'was', 'the', 'best', 'of', 'times', 'worst', 'age', 'wisdom', 'foolishness'}

The objective is to turn each document of free text into a vector that we can use as input or output for a machine learning model. Because we know the vocabulary has 10 words, we can use a fixed-length document representation of 10, with one position in the vector to score each word. The simplest scoring method is to mark the presence of words as a boolean value, 0 for absent, 1 for present.

TABLE II: Encoded Sentences

[1, 1, 1, 1, 1, 1, 0, 0, 0, 0] [1, 1, 1, 0, 1, 1, 1, 0, 0, 0] [1, 1, 1, 0, 1, 0, 0, 1, 1, 0] [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

In II we present encoded sentences which we can input to a classifier for training and testing purposes.

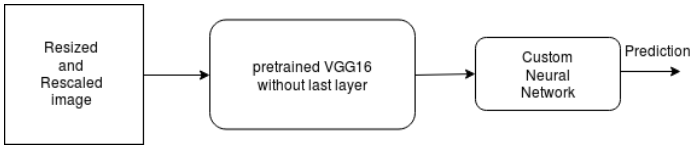


Fig. 5: Image Classifier Flow

V. Implementation

A. Image Classifier

a) Dataset

: Our dataset was downloaded from [7]. The dataset initially had a lot of faulty images. Some were empty, some werent originally in the jpeg format but were forcibly were stored at .jpeg which resulted in an error when those were read by the code. There were also some images in formats which werent supported by the module. Our Dataset after this process had total 30000+ images(6000 images per class) which were separated into 80% training set and 20% validation set.

b) Model

: Input images are resized to 150 $\times$ 150 size. The are also normalized so that the image intensity values lies in the range 0 to 1. These pre-processed image is then passed to VGG Network. Features are extracted from the second last layer of VGGNet. These features are then used to train a neural network of three layers. Number of neurons in the first layer are 256, number of neurons in the second layer are 128, number of neurons in the final(classification) layer are 5. Dropout of 0.5 is used to prevent over fitting. RMSProp is used as the optimizer. Sparse Categorical Cross Entropy is used as the loss function. Relu is used as the activation for intermediate layer. Softmax is used for the classification layer. While testing on a new image, the image is passed to the preprocessing part. It is then passed to VGGNet to extract features. These features are then used as input to the trained neural network and it is classified in one of the genres.

c) Resource used for training

: A VM instance was created on Google Cloud. We werent able to get an access to GPU which hindered the progress and limited the scope of our project to some access. Google Cloud helped us to some extent but wasnt as efficient as expected.

d) Modules

: The code was written in Python3. The code structure was adopted from [8]
Module dependencies: Tensorflow, Keras, Numpy, h5py

B. Text Classifier

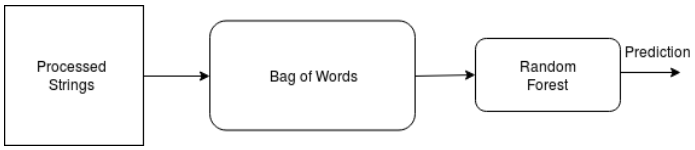


Fig. 6: Text Classifier Flow

Starting off with pre-processing, the dataset contained a lot of blank spaces which had to be replaced with *numpy.nan* so that the classifier can identify it as a NaN value. For Bag of Words implementation the sklearn module of python has inbuilt Vectorizers which allows the user to convert the input title strings into numerical binary arrays. Here we have used the **Count Vectorizer** which after initialization needs to be fitted to the input string list before transforming it into a binary array. After obtaining a binary array out of the titles we input them to a **Random Forest Classifier** with a parameter value of 100 which implies a forest of 100 trees.

a) Modules

: The code as been written in Python.
Module dependencies : Numpy, Sklearn, Pandas

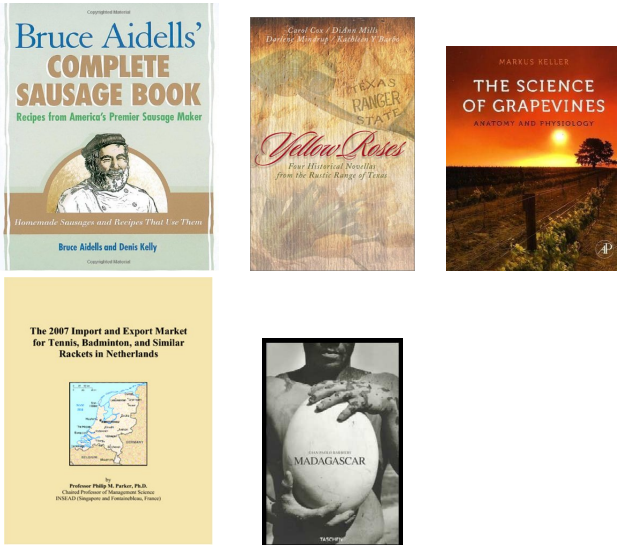
VI. Results & Discussion

A. Image Classifier

TABLE III: Test Samples for Cookbook, Romance, Science, Sports & Travel in clockwise order

We observed a training accuracy of 75% and a validation accuracy of 69.04% on the dataset. Some of the misclassification happened because of the vagueness in the cover. For example in TABLE IV the Food figure, in it there is no indication of a food item and perhaps due to presence of a human in a particular pose(which is prevalent in sports biographies) it is classified as Sports & Outdoors instead of Cookbooks,Food, and Wine. Simple, plain covers with few text have a tendency to be classified as Science & Math. This type of problem is seen in the case of the romance figure in TABLE IV. In case if a natural scenery is present in the image like mountains, sunrise, ice, forest etc, Then the image has tendency to be classified as Travel. This can be seen in the science figure, which actually belongs to Science & Math but is classified as Travel. For sports figure, the presence of a person without clothes perhaps in a posture forces it to be categorized as

TABLE IV: Misclassified Samples for Cookbook, Romance, Science, Sports & Travel in clockwise order



Romance instead of Travel. Its clear from the misclassified images that the machine is able to predict what an average human would have done on observing the cover. We are able to show that the machine can be trained to come closer to human level performance on this problem.

TABLE V: Input to Text Classifier

Category	Title
Sports & Outdoors	Sports Training Notebook: Badminton: For Coaching Instruction On All Levels Of Sport
Travel	Antarctica Satellite (Laminated) (National Geographic Reference Map)
Travel	Eyes to See: U.S. Volunteers in Nicaragua
Cookbooks, Food & Wine	Pizza, A Slice of American History
Cookbooks, Food & Wine	Build Your Own underground Root Cellar
Cookbooks, Food & Wine	Brown Eggs and Jam Jars: Family Recipes from the Kitchen of Simple Bites
Travel	Aruba 1:50,000 & Oranjestad 1:10 000 Travel Map, waterproof, BORCH
Sports & Outdoors	The Fundamentals of Hogan
Sports & Outdoors	Alone: The Triumph and Tragedy of John Curry
Travel	Virgin Islands Reef Creatures Guide Franko Maps Laminated Fish Card 4" x 6"

B. Text Classifier

Inputs to the classifier have been presented in TABLE V in the form of book title and the corresponding book genre. TABLE VI shows the output of the classifier and as can be seen that the classifier predicts the genre labels perfectly because of the small size of the test dataset. When tested fora large dataset i.e 20,000 labels we get a validation accuracy of $\approx 79\%$.

TABLE VI: Output of Text Classifier

Category	Title
Sports & Outdoors	Sports Training Notebook: Badminton: For Coaching Instruction On All Levels Of Sport
Travel	Antarctica Satellite (Laminated) (National Geographic Reference Map)
Travel	Eyes to See: U.S. Volunteers in Nicaragua
Cookbooks, Food & Wine	Pizza, A Slice of American History
Cookbooks, Food & Wine	Build Your Own underground Root Cellar
Cookbooks, Food & Wine	Brown Eggs and Jam Jars: Family Recipes from the Kitchen of Simple Bites
Travel	Aruba 1:50,000 & Oranjestad 1:10 000 Travel Map, waterproof, BORCH
Sports & Outdoors	The Fundamentals of Hogan
Sports & Outdoors	Alone: The Triumph and Tragedy of John Curry
Travel	Virgin Islands Reef Creatures Guide Franko Maps Laminated Fish Card 4" x 6"

VII. Conclusion & Future Work

We conclude that its possible to predict the genre of the book from its cover and the title. As expected, the text classifier performed better than the image classifier because of less ambiguity in the titles compared to the cover art.

For image classification, we might want to try ResNet or InceptionNet CNNs to get the feature vectors for training in future as they show better performance on ImageNet compared to VGGNet which was used here. We can also ensemble the text and image classifiers by using something like mixture of experts model or a softmax gating network. Its expected that the ensemble will show better performance than the individual classifiers.

For text classification we might want to try Hash Vectorizer or TfIdf Vectorizer to increase the speed and to consume less memory space.

References

[1] Krizhevsky et.al, ImageNet Classification with Deep Convolutional Neural Networks, 2012
[2] Simonyan et.al, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014
[3] Iwana et.al, Judging a Book by its cover, 2016
[4] LeCun et.al, Gradient-Based Learning Applied to Document Recognition, 1998
[5] Gatys et.al, A neural algorithm of artistic style, 2015
[6] Bay Area Deep Learning presentation by Andrej Karpathy
[7] Link to the Dataset
[8] Keras tutorial
[9] Info on Pretrained Networks and their usage
[10] Bag of Words Tutorial
[11] CNN Wiki
[12] Bag of Words Wiki