Classification of Book Genres By Cover and Title

Yash Goyal(201502181), Mayank Garg (201530097), Kumar Abhishek(201502172), Dhruval Jain(201530109)

Problem addressed:

In our project the problem we are addressing is to use book cover and its title to classify the book to a genre. It has been shown that the cover design has a significant impact on the sales of a book, with book sales often shooting up after a change in design. Our goal is to create a model that can determine how representative a cover is of its genre, as a method to later evaluate if the more a book cover resembles others in its genre, the higher the book sales.

Major challenges:

In this project we faced was

1. Dataset generation, by scraping the website Openlibrary.org.
2. How to use feature extraction from images and text collectively, that is what weights we should give to features extracted from images and text and how to concatenate the th features.

Work Flow:

- Identification of book genres.
- Dataset generation of selected book genres.
- Pre-processing of book covers.
- Extracting Image (Book Cover) Features using Imagenet.
- Extracting Features from Book Title using Word2Vec.
- Combining both the features to feed them into a classifier for final prediction.

- Using different types of classifier and analysis of the results.

Identification of book genres:

We tried to choose genres which are not overlapping in nature. By overlapping we mean the features which define the book genres are different in nature.
Hence we chose, Business, Fantasy, Textbooks, Science-Fiction and Romance as book genres.

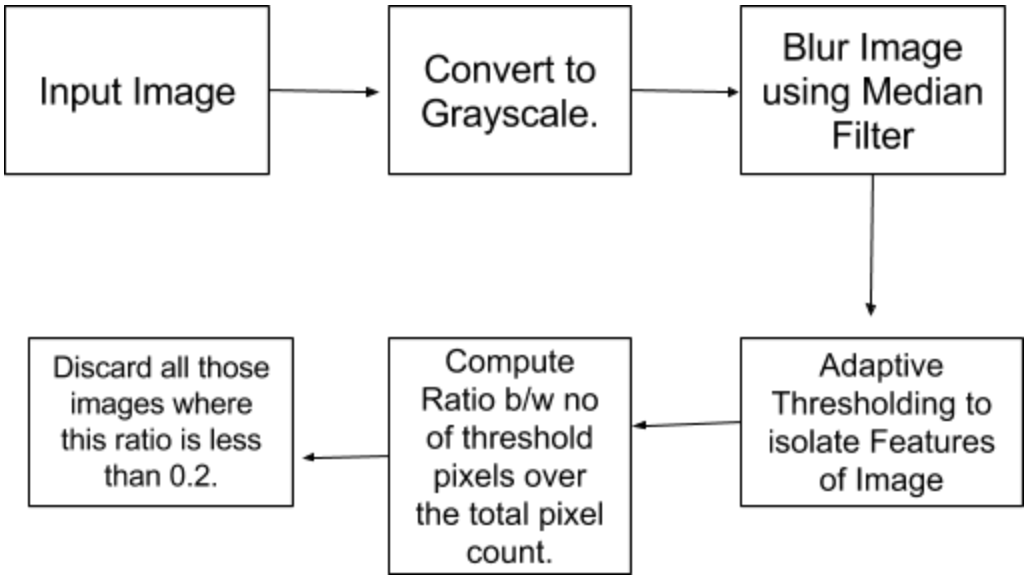
Dataset generation of selected book genres:

The dataset obtained from OpenLibrary.org consists of a total of 6,185 images from the above stated five genres.

We used *Selenium* and *BeautifulSoup* library in python to scrape the book-cover images and titles to generate the dataset.

We encoded the titles in UTF-8 to bring all the words in plain english.
Training and Testing data distribution is in the ratio 83:17.

Preprocessing of book covers:



Initially, we used ‘History’ as a genre but the book-covers were mostly plain color in texture and text on it also didn’t provide much of the information. So, most of the books in this genre were discarded in the pre-processing step where the ratio was less than 0.2.

Feature Extraction from Images

Number of image features used are 4096.Various CNN architecture were considered to extract features from the images:

1. AlexNet
2. VCGNet-16
3. VCGNet-19

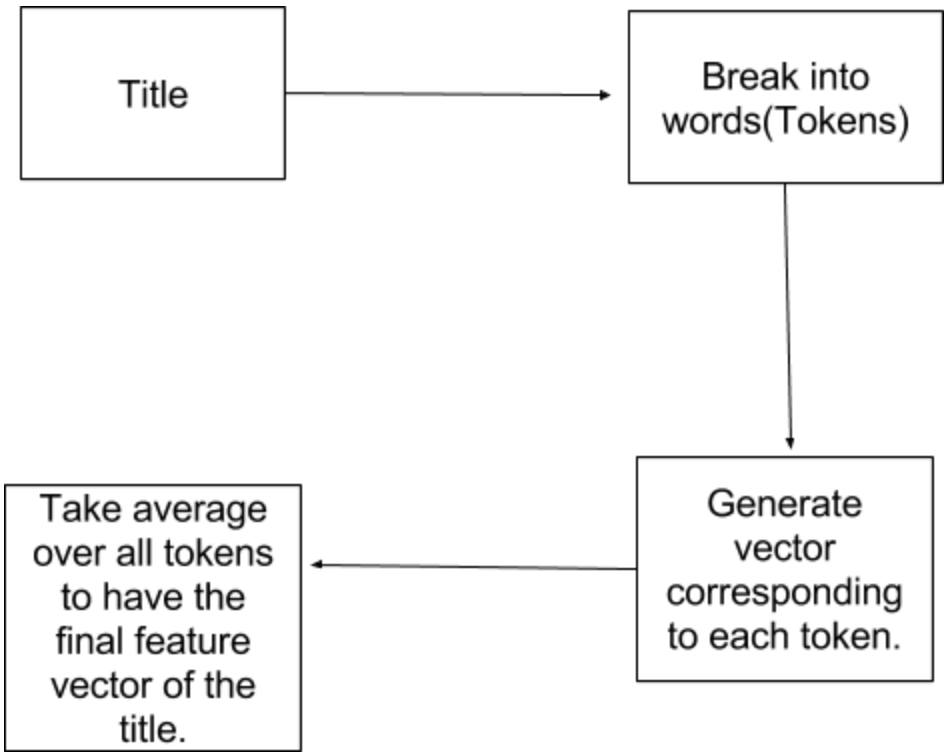
We chose Alexnet because it gave the maximum accuracy on our dataset.

Architecture	Accuracy(Using only CNN for feature extraction)

AlexNet	45.66%
VCGNet-16	39.17%
VCGNet-19	36.52%

Feature Extraction from text:

Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word Vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.
We have use Word2Vec that is trained on GoogleNews dataset.
It maps each word in the corpus to a 300 dimensional vector in the vector space of real numbers.



Combining the features:

Features from both the attributes are appended alternatively and title features are re-used when they are exhausted.

Total features thus obtained= 4096 + floor(4096/300)

= 7996

This ensures uniformity and equality of both features since title features are very less than image features.

Classification into genres:

We have used various classifiers into predict the genres like:

1. Multi-class SVM
2. AdaBoost using Random Forest
3. Softmax