Can Machine Learning Models Predict Movie Genre?
Data Science 2 Final Project Report

Yujiao Chen, Brian Ho, Jonathan Jay
May 3, 2017

Summary

In this project, we evaluated multiple machine learning methods for predicting movies' genres based on widely available attributes such as their titles, crowd-sourced descriptions, and poster designs. For these analyses we limited our prediction classes to romance, horror and science fiction (sci-fi) movies, as classified by The Movie Database (TMDb). Our most effective model used text from movie descriptions and titles, reduced their dimensionality by principal components analysis (PCA), and performed classification with a support vector machine (SVM). This model achieved 80% testing accuracy on the multi-class problem. We also trained models that significantly outperformed random chance using convolutional neural networks (CNNs) on movies' posters and SVM on the color composition of movies' posters. As a secondary analysis we also explored CNNs to predict movies' decades based on posters alone. This report explains our methods and results, discusses our findings and their implications for future research to predict movie genre.

Background: the genre prediction problem

In our own experience as consumers, genre is critical to deciding whether to watch an unfamiliar movie. Genre conveys information about the mood, types of settings and plotlines viewers can expect. Genre classification, therefore, is an important task for services that aim to help consumers find content (including TMDb itself, a user-generated database that crowdsources movie metadata such as genre). Automating this task might be particularly useful for standardizing classification of movies from different regions or eras, or when human observers disagree. Effective genre classification algorithms could be adapted to individual users' preferences or to differing conceptualizations of genre classes.

Classifying genres, however, is a challenging machine learning problem. While many genre categories are easily recognized by movie consumers, distinctions among genres are not clear-cut. First, a single movie may occupy multiple genres. Second, human observers might disagree as to a particular movie's genre. In our preliminary review of genre classifications, we found that even between TMDb and IMDb (a similar movie database) genre classifications often differed, and many movies belonged to multiple genres. To sidestep this issue of genre "fuzziness," for this proof-of-concept analysis we narrowed the classification task to romance, horror, and sci-fi movies. These were identified based on their low correlation in movies with multiple TMDb labels.

Methods

Data sources: From the TMDb and IMDb databases, we downloaded metadata of top 10,000 movies by popularity and dropped movies with any missing information. Using this smaller and cleaner movie database, for our text-based analyses, we created a balanced dataset representing 500 movies from each genre (romance, horror and sci-fi) as assigned by TMDb, from 2006 to present, with high popularity ratings. The fields used were “title” and “description.” For poster-based analyses, we used 330 movies from each genre, including older movies (1960-present), balanced by decade and selected by descending popularity within each genre-year.

I. Text-based analyses

Like genre classifications, the short text descriptions included in TMDb metadata are intended to convey information about a movie’s tone and subject matter. Therefore, we expected correlations between descriptions and genre. We expected that interpreting descriptions, however, might rely on some expert knowledge. For example, a recent “X-Men” offshoot, “Logan”, is described as follows: *“In the near future, a weary Logan cares for an ailing Professor X in a hide out on the Mexican border. But Logan's attempts to hide from the world and his legacy are up-ended when a young mutant arrives, being pursued by dark forces.”* While the description explicitly mentions a future setting and interaction with a mutant (consistent with sci-fi) and “dark forces” (consistent with thriller/horror), some additional knowledge or experience is necessary to infer that this scenario is likely to produce action and adventure, rather than comedy or romance. We sensed the title’s short length might also convey mood.

To test these hypotheses, we represented movie descriptions with a bag-of-words model, dropping common stopwords. Exploratory visualization showed some of the patterns we expected by genre. Word clouds showed high incidence of “love” and “relationship” in romance movies; “Earth” and “alien” in sci-fi movies; and “evil” in horror movies.

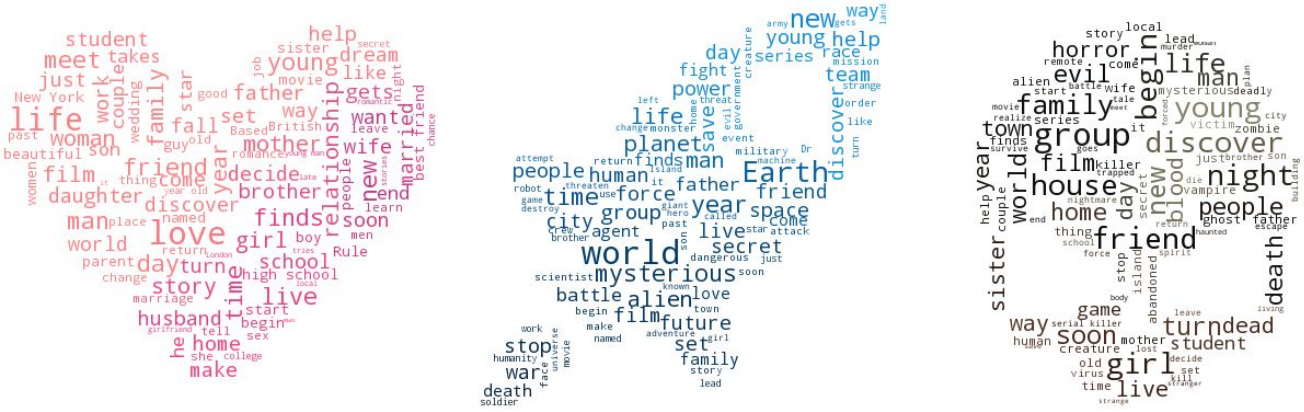


Figure 1: Common words in movie descriptions, by genre (romance, sci-fi and horror)

After dropping stopwords, 48,405 words remained in the descriptions. We reduced our wordlist to the 179 most-common words (those appearing 30+ times) for creating the new data frame, split according to training and testing, then performed PCA on the training set. The first 100 principal components explained 90% of the variance, allowing us to reduce dimensionality by almost half. We tuned a SVM with radial basis function using cross-validation in the training set. Then we performed the same principal components transformation on our testing set and tested the SVM performance. It predicted genre with 80% accuracy.

This model performed particularly well for romance movies, correctly predicting 100% of these movies in testing:

		preds		
		Horror	Romance	Scifi
actual	Horror	**75.0%**	12.5%	12.5%
	Romance	0.0%	**100.0%**	0.0%
	Scifi	17.6%	17.6%	**64.7%**

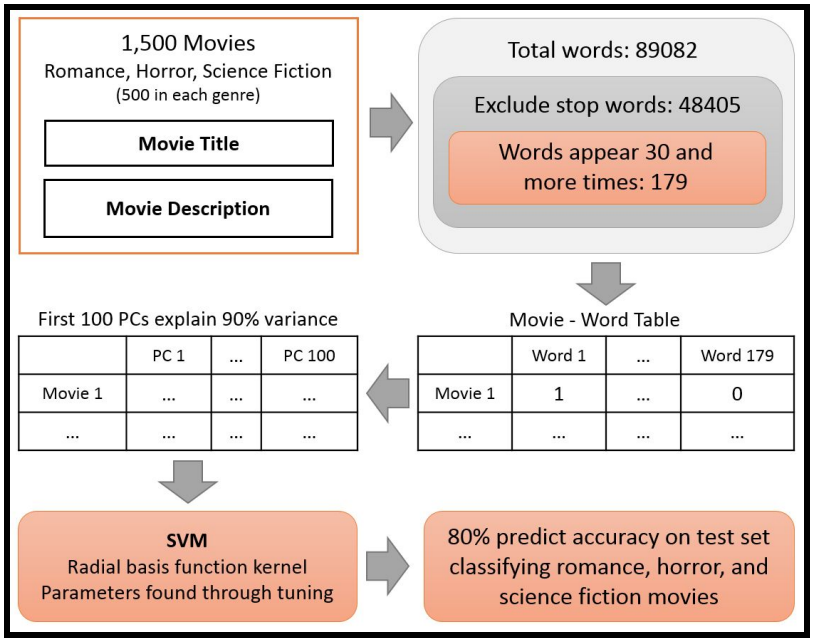


Figure 2: Analysis of movie description text

Our tests confirmed, additionally, that title features predict genre. We created custom features, such as length in words and characters; the number of special characters or digits appearing in the title; and sentiment score (using the "RSentiment" package). SVM models trained on these features distinguished better than chance in a two-class problem between romance and horror (65% for linear kernel and 66% for RBF kernel).

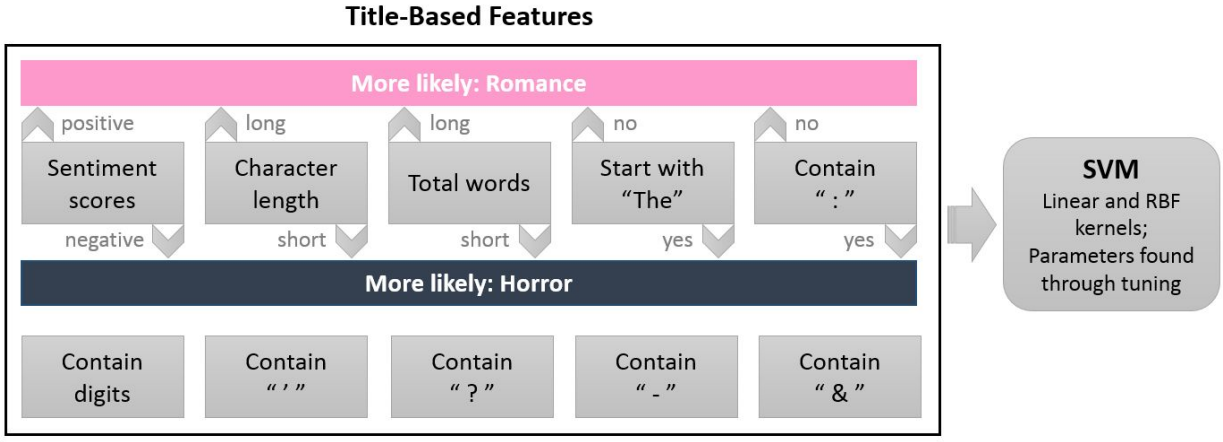


Figure 3: Analysis of movie title text features

II. Image-based analyses

While our predictive accuracy using movie descriptions was high, crowdsourced descriptions may not be available for every genre classification use case. Movie posters, however, are often available even before the movie has been released. Posters often depict movie characters and/or settings, and convey a movie's mood. We tested two hypotheses: (1) that we could train convolutional neural networks (CNNs) to predict genre and (2) that traditional machine learning models could predict genre based on posters' color patterns alone.

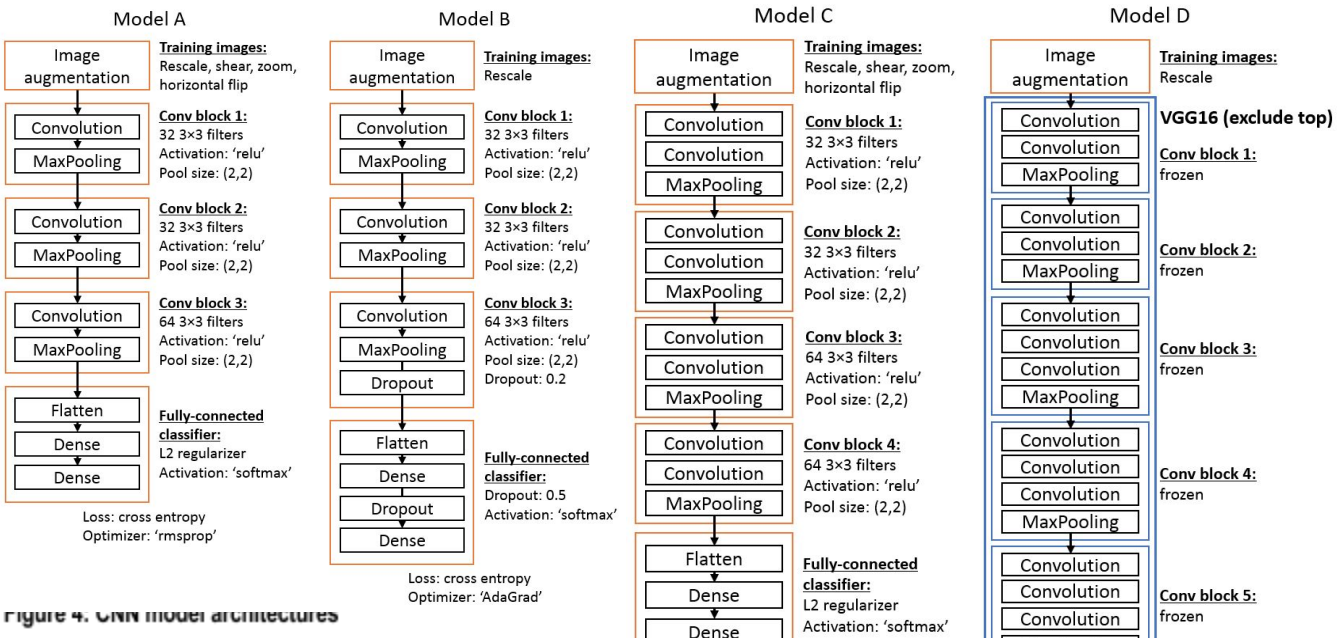
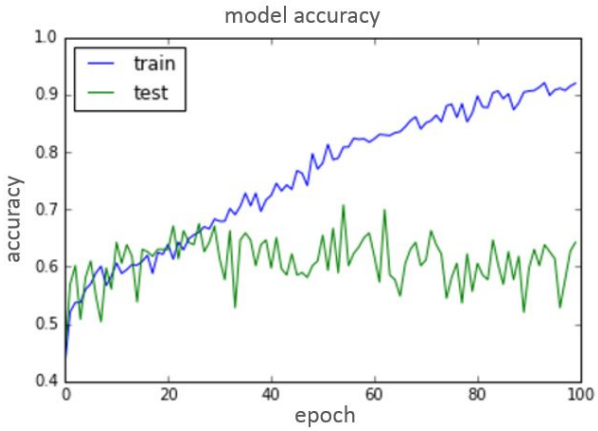


Figure 4. CNN model architectures

Using CNN design principles and trial and error, we developed four CNNs that significantly outperformed random chance in classifying by genre. Three (Models A-C below) were trained from scratch and applied to the two-class and three-class problem; a fourth (Model D, below) was trained by replacing the top layer of the pre-trained VGG-16 model[1] and solely applied to the two-class problem (romance-horror).

Most promising



These models all significantly outperformed random chance in cross-validation. Models

[1] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556

A-C achieved predictive accuracy between 75-80% for two classes and between 60-65% for three classes; Model D approached 70% accuracy on the two-class problem. The most promising results were from Model C (see above).

Instead of analyzing details in an image using CNN, other noticeable features of a poster, such as hues, tones and shades, can be good indicators for genre. For example, the poster of romance movie tends to have bright colors, while the poster of horror movie usually has a dark background and the poster of sci-fi movie often shows high contrast:

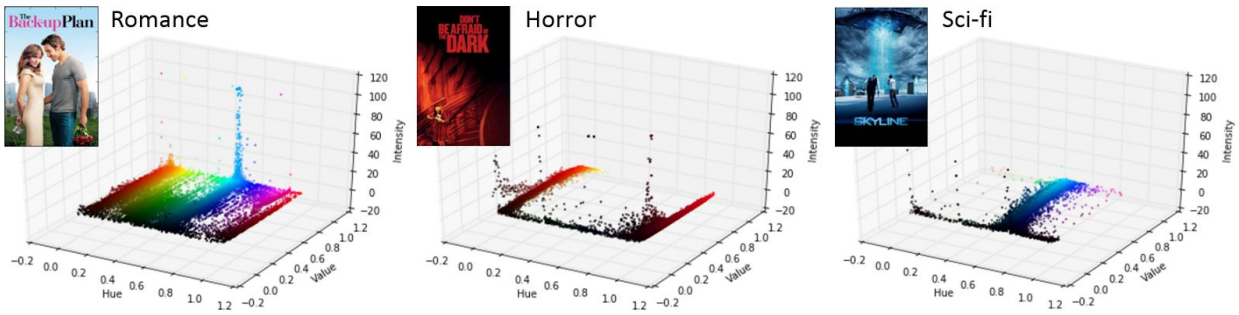
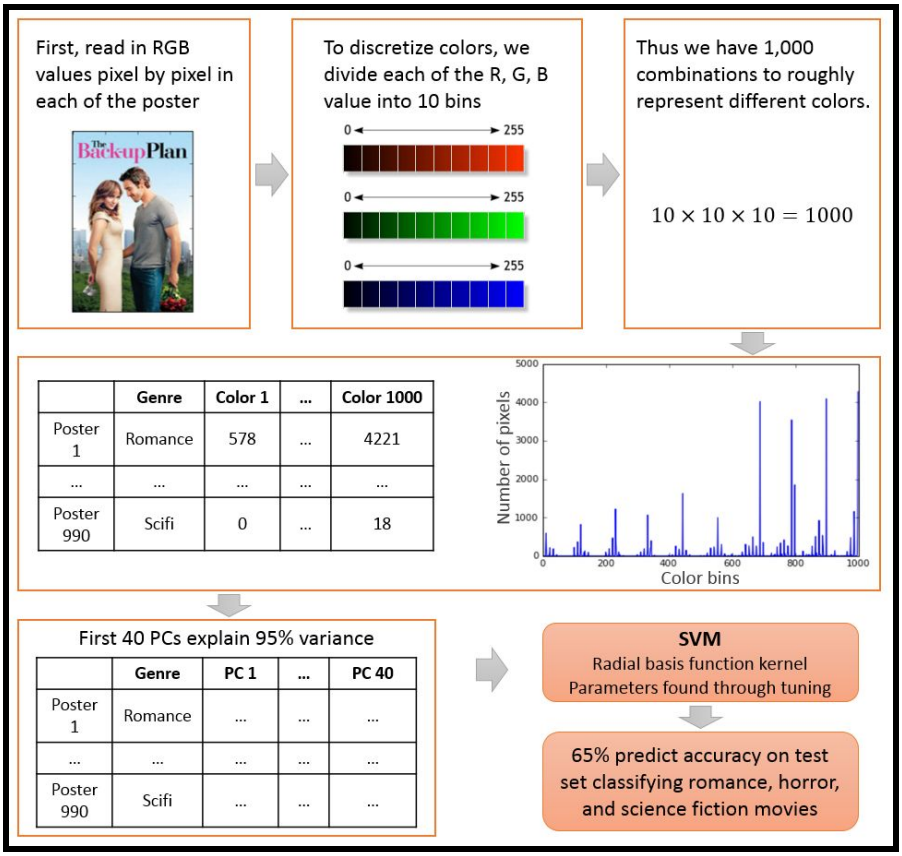


Figure 5: Sample poster color decomposition results

In order to make classifications of movie genre based on color of posters, we first read in the RGB value of each pixels in a poster, then counted the frequency of those RGB values falling into each of the 1,000 discretized color bins. In this way, each poster can be transformed into a 1*1000 dimensional array. By applying principal component analysis, we found that the first 40 PCs explain 95% variance of the data, which helps us greatly reduce the dimensionality of feature space from 1,000 to 40. Built upon this data frame, we tuned a SVM with radial basis function using cross-validation in the training set. Then we performed the same principal components transformation on our testing set and tested the SVM performance. It predicted genre with 65% accuracy (among romance, horror, and sci-fi), which is a surprisingly good result comparable to CNN models described above.

Figure 6: Genre prediction on poster color



Finally, we explored whether CNNs could learn to identify movies' *decades* based on posters alone. While this task might also represent a distinct use case, we intended this analysis primarily to assess variation in posters over time, with implications for tasks such as genre classification.

To isolate the effect of movie age alone, this analysis used only movies categorized in TMDb as sci-fi. We trained a CNN model with 3 convolutional and 3 pooling layers (see Model A architecture above) to predict across 5 classes corresponding to decade ("1960s", "1970s" through "2000s") from 300 training posters. We treated the exercise as a classification task rather than a regression task, refraining from the assumption that variation would be continuous over decades. This model achieved validation set accuracy approaching 40%, significantly outperforming random chance (20%). Next, we restricted samples to "1950s" vs. "2010s": in this two-class problem, the model predicted decade with 85% accuracy. These results suggest high within-genre variation in movie poster features. This variation could compromise how neural networks learn genre, particularly in small training sets such as the ones we used above. Models trained and tested within narrower time bandwidths might perform better.

Limitations

While a PCA/SVM model using description text offered higher predictive accuracy than the other methods we tested, the generalizability of this result is not certain. For example, our CNN models might have performed better with more training data and/or more computationally expensive network architectures. Our decision to limit the predictive task to three highly distinctive genre classes allowed us to demonstrate the feasibility of genre prediction, but does not necessarily generalize to other genre classification tasks, such as $k > 3$ or assigning multiple classes per movie.

Conclusions

Despite these limitations on generalizability, a few key lessons stood out. First, the higher accuracy of a scratch model, compared to the pre-trained VGG-16 model, suggests that genre classification with posters does not amount to an object recognition task. The scratch models may have been better able to learn attributes such as mood--perhaps through color features, which we found highly predictive--than the VGG-16, trained to excel at multiclass object recognition. Second, deep learning did not guarantee better results: SVM on colors alone performed about as well as CNN, and a more traditional machine learning toolkit, using bag-of-words, PCA and SVM, performed significantly better for our task and dataset. Our most predictive model, however, depended on inputs (crowdsourced descriptions) that are not always available. These findings provide a reminder that the right model selection is highly dependent on the problem at hand and the resources available.