Inferring Movie Genres From Their Poster

Mostafa Dewidar
Stanford University
mdewidar@stanford.edu

Abstract

Classifying movies by their posters into their correct genres is a problem that lends itself to Convolutional Neural Networks and can prove to be of immense benefit to advertisers, producers and viewers as it can help align the poster's perceived genre by the audience with its actual genre making it easier for viewers to browse and select movies. Previous work in movie-genre classification has tried extracting low level features from the posters, using standard resnet50 architecture, using features from the movie itself, as well using KNN and Naïve Bayes. In this paper, I explore the application of deeper versions of pretrained traditional networks like VGG19 and ResNet101 as well as AlexNet to explore the viability of transfer-learning and different architectures for this problem. Using the IMDb movie poster dataset A top accuracy of 42% is achieved using ResNet101. Other networks seem to perform comparably well suggesting that improvements in performance may depend more on balancing the dataset and feature-engineering than on changing between these 3 network architectures.

1. Introduction

The movie industry is a very big industry, with a size of about $91.83 Billion dollars in 2020 in the US alone [8]. Movie posters are one way to encode the most information about a movie in one picture. They serve many functions like telling the audience the title of the movie and relevant cast members, providing visual cues as to what type of movie it is and what genre it may belong to, signaling the cinema or art style predominantly used in the movie, its setting, its main characters and maybe even some of its main events. If broken down successfully, such a dense representation can provide a lot of information about a movie that can be helpful to viewers, producers, and distributors. For example, streaming services (a distributor) can bucket movies according to the genres that their posters signal to the viewers and choose alternative posters that better match the genre of the movie in case the poster is misleading to avoid misleading viewers and losing retention. Film production companies can choose posters that are more appealing and thus more likely to boost ticket and DVD sales. And viewers can get better signals as to what they are about to watch in the absence of a trailer (in the case of in-store purchases). All these reasons combine to make the promise of applying different Convolutional Neural Networks to inferring information from a movie's poster like its genre appealing. In this paper, we attempt to apply many different CNN architectures, namely resnet101, VGG19, and AlexNet to this problem to see if they can solve it and if so, which are better adapted to the problem to suggest methods to improve on performance in the future.

2. Related Work

There have been a few approaches trying to identify movie genres from posters in the literature. For example, Makita and Lenskiy used a multivariate Bernoulli event model to learn the likelihoods of genres based on movie posters. It had a success rate of 50% [7] which, although is much better than random baseline, is still not strong enough for industrial applications. Pobar et al. used ML-kNN, RAKEL and Naïve Bayes to detect 18 genres (merged into 11) using a dataset of around 6739 posters and achieved a top accuracy with Naïve Bayes of 38% [4]. Kundalia et al. Balanced the poster data collected from IMDb to make sure all genres had an equal number of posters associated with them and used a network based on a pre-trained inceptionV3 and were able to report a remarkable 84% accuracy [6]. Sung and Chokshi used modified versions of VGG16, Resnet50, and DenseNet169 and reached a peak validation accuracy of 79% but did not report test accuracy [10]. Wi et al. used a Gram layer in a CNN to extract style features from the poster before feeding it into the network and were able to achieve an accuracy of 45% using resnet34 improving on the previous implementation [11]. Barney and Kaya used a custom resnet34 implementation, a custom CNN architecture as well as ML-KNN and were able to achieve a top accuracy of 38.26% [1]. The best results in genre classification however come from Huang and Wang who didn't just rely on posters but combined both au-

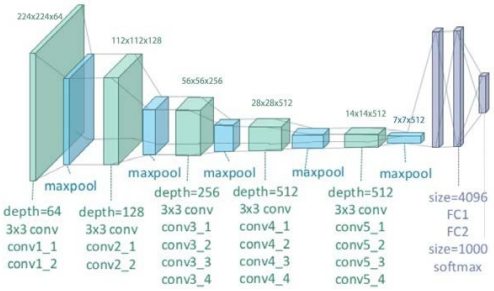

Figure 1. VGG-19 Network Architecture

dio and visual components from the movie to yield a classification accuracy of about 92% [3].

3. Methods

3.1. Transfer Learning

Using versions of the networks that were pre trained on Imagenet data can help with our poster-to-genre classification problem since there are probably features that can transfer (transfer-learning) across different visual recognition datasets and tasks. Thus, I attempted to use pre-trained versions of VGG19, AlexNet and resnet101 to test the efficacy of transfer learning in this problem domain and the properties of each of these networks.

3.2. VGG19

VGG16 and VGG19 [9] architecture was an attempt to see the effect of increasing depth on effectiveness of CNNs. The net contains eight layers with weights; the first five are convolutional and the remaining three are fully- connected. The output of the last fully-connected layer is fed to a 1000-way softmax which produces a distribution over the 1000 class labels for the ImageNet challenge. As shown in fig 1.

3.3. AlexNet

AlexNet [5] was is a large computer vision model with more than 60 million parameters. AlexNet has 650,000 neurons, five convolutional layers followed by 3 max pooling layers and three fully connected layers. AlexNet popularize the regularization technique of dropout whereby neurons are dropped (activations set to zero) during training with random probability so that they have no effect on the outcome for the training batch and then used normally during inference so as to make sure that the network doesn't overweight some neurons while ignoring others thereby nudging it to learn complex features of the data. Architecture for AlexNet can be seen in fig 2.

3.4. Resnet101

The Resnet (Residual Network) architecture was invented to solve the problem of increasing inefficiency of

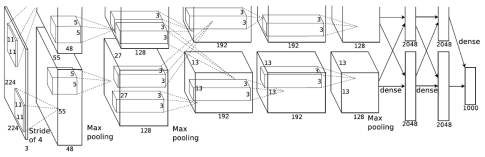

Figure 2. AlexNet Architecture

training and optimization in very deep neural networks, since it had been shown that depth can be very helpful to the accuracy of CNNs. Resnets are built on the corollary of the hypothesis that states that if one can asymptotically approximate a complicated function with multiple stacked non-linear layers, then the same can be said for residual functions. [2]

4. Data and Features

I used the IMDb posters dataset but filtered it down to 8252 images to speed up the training process. All images were resized to be of size (300,180,3) and the data was split into 80:10:10 training:validation:test split to make the most out of the few training examples that exist. Since there are only about 40,000 movies in the IMDb and since the number of movies made in the world in general is a finite and relatively small number, this was another reason why transfer learning made sense to be used for this problem. When it comes to labels, each poster had an associated set of genre labels that ranged from 1 to 3 associated genres. To make this problem simpler, I pre-processed the data to only include one label for each poster to ensure that it's a single label instead of a multi-label problem and reduce the complexity of the features needed to be learned by the model. The label was selected according to ascending alphabetical order of genre names appearing for this particular poster, a random selection scheme that should maintain the distribution of the data. The distribution of the labels for the data can be viewed in Table 1. and Fig 4. Examples from DataSet can be seen in fig 5. and fig 6.

5. Experiments, Results, Discussion

5.1. Resnet101

Training for 4 epochs with a learning rate of 0.01 using the aforementioned loss yielded the following results Table 2:

Training seemed to converge on a result after 4 epochs as we can see that both training a validation loss converge to the same value and training loss starts to go lower than validation loss indicating that we might be over fitting if we train for more epochs.

Confusion matrix shows that the network is skewed to predicting the most popular genres when it's confused or

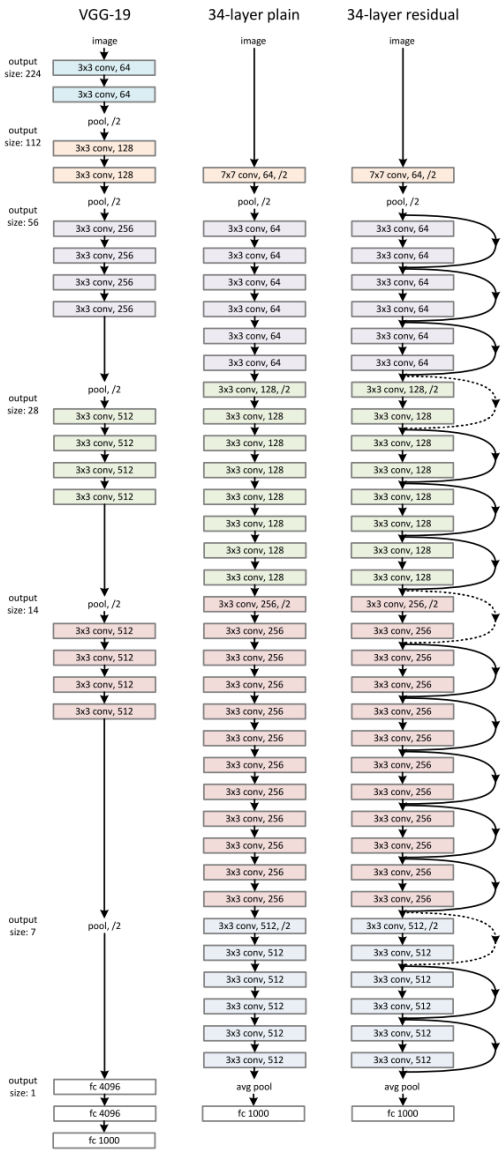

Figure 3. Resnet34 Architecture, Resnet101 extends to 101 layers

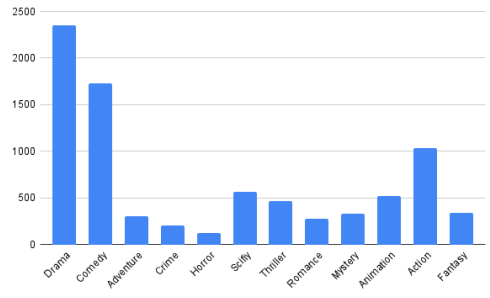

Figure 4. Data Distribution by Genre

Genre	Examples
Drama	2351
Comedy	1731
Adventure	306
Crime	204
Horror	121
Sci-Fi	567
Thriller	467
Romance	276
Mystery	334
Animation	522
Action	1035
Fantasy	338

Table 1. Data Distribution by Genre

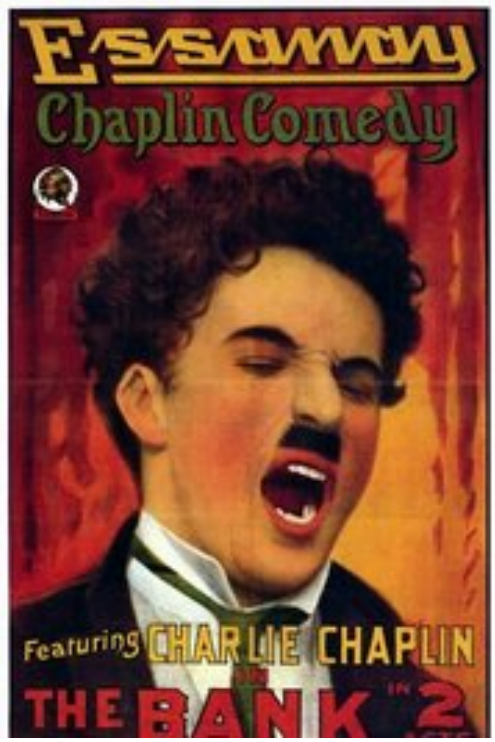

Figure 5. Chaplin Comedy - The Rank 2: Comedy

Epoch	Accuracy	Precision	Recall
1	0.3727	0.2136	0.1947
2	0.3848	0.1922	0.1712
3	0.4218	0.2034	0.1905
4	0.4212	0.2273	0.1872

Table 2. Training Results for ResNet 101

Figure 6. Metropolis - 1927: A Sci-Fi Drama

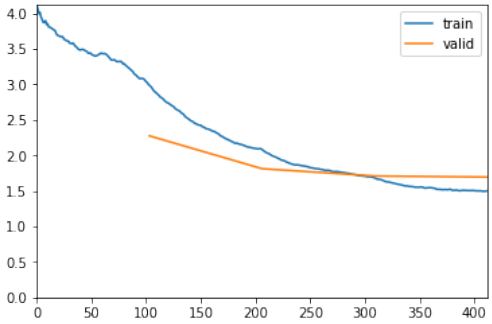

Figure 7. Training and Validation Losses over 4 epochs for ResNet101

prompted by posters that are towards the edges of the distribution. This may signal that additional pre-processing of the data could have yielded better results.

5.2. AlexNet

Training for 4 epochs with a learning rate of 0.01 using the aforementioned loss yielded the following results Table 3:

Training seemed again to converge on a result after 4 epochs as we can see that both training a validation loss converge to the same value and training loss starts to go lower than validation loss indicating that we might be over

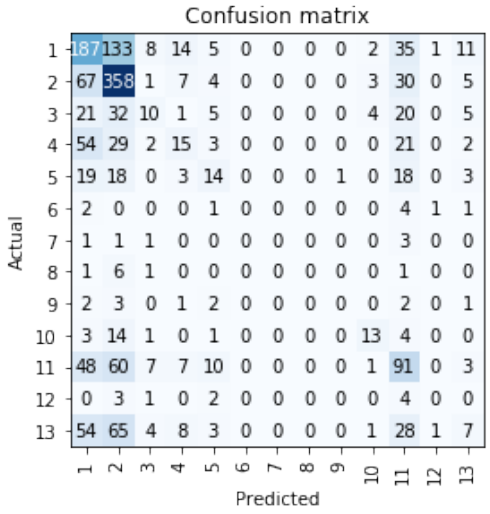

Figure 8. Confusion matrix for resnet101. Each genre is denoted by a class number. Refer to Table 1 for ordering of Genres (Drama: 1, Comedy: 2.. etc.)

Epoch	Accuracy	Precision	Recall
1	0.3472	0.1693	0.1412
2	0.3648	0.1704	0.1421
3	0.3854	0.1744	0.1515
4	0.3896	0.1756	0.1588

Table 3. Training Results for Alexnet

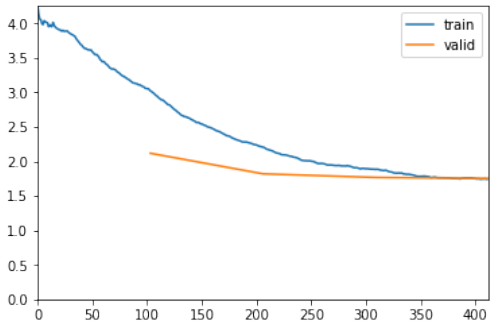

Figure 9. Training and Validation Losses over 4 epochs for AlexNet

fitting if we train for more epochs.

Confusion matrix shows that the network is skewed to predicting the most popular genres when it's confused or prompted by posters that are towards the edges of the distribution. This may signal that additional pre-processing of the data could have yielded better results.

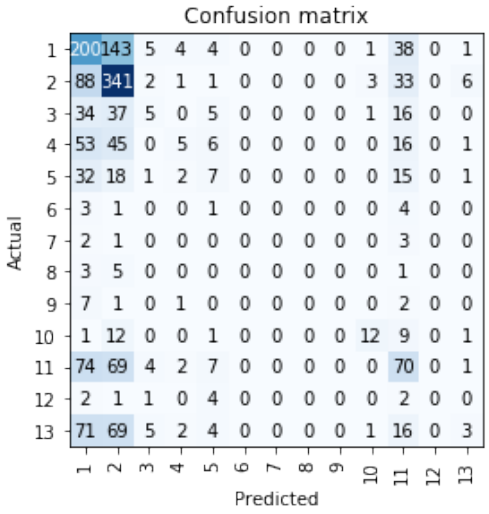

Figure 10. Confusion matrix for AlexNet. Each genre is denoted by a class number. Refer to Table 1 for ordering of Genres (Drama: 1, Comedy: 2.. etc.)

Epoch	Accuracy	Precision	Recall
1	0.3539	0.1510	0.1546
2	0.3787	0.1546	0.1531
3	0.4115	0.1747	0.1848
4	0.4175	0.1811	0.1837

Table 4. Training Results for VGG-19

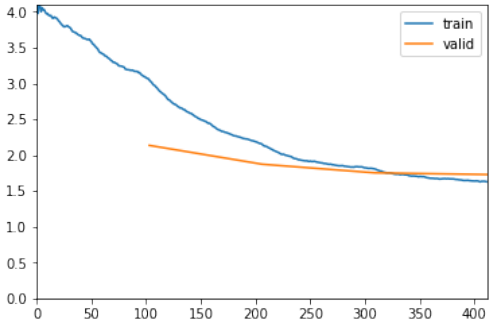

Figure 11. Training and Validation Losses over 4 epochs for VGG-19

5.3. VGG 19

Training for 4 epochs with a learning rate of 0.01 using the aforementioned loss yielded the following results Table 3:

Confusion matrix shows that the network is skewed to predicting the most popular genres when it's confused or prompted by posters that are towards the edges of the distribution. This may signal that additional pre-processing of the data could have yielded better results.

Some of the most confusing examples across networks can be found in fig.

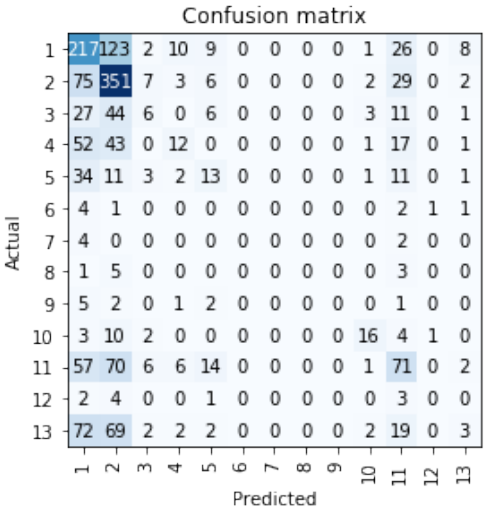

Figure 12. Confusion matrix for VGG-19. Each genre is denoted by a class number. Refer to Table 1 for ordering of Genres (Drama: 1, Comedy: 2.. etc.)

Figure 13. Confusing examples for all the networks top is prediction class number and bottom is correct class number. Refer to Table 1 for ordering of Genres (Drama: 1, Comedy: 2.. etc.)

5.4. Discussion

By looking at the training and validation lossess, we can see that our networks have converged on a result and that further training would have most likely not improved performance but led to over fitting. By looking at the confusion matrix, however, we can start to see what the networks are learning. Note that the 13th category is an extra category in the data that I created to label the posters that have no associated Genre or for which there are very few training examples ¡10 that the networks won't realistically be able to learn from. Notice how the 13th cateogry gets mislabelled as either Comedy or Drama (the two most prevalent labels) It seems that the network is fitting everything unfamiliar into buckets of what its most likely to find in the dataset. Furthermore, we can see that a good number of the Drama movies are being classified as Comedy and vice versa. When the networks are confused about one label, they just tend to predict one of those two. We can also see that Action is another label that gets thrown around because it's over-represented in the data. All of this suggests that the networks are learning the distribution of the data to a good degree but haven't seen enough examples from all the Genres to be able to make predictions with high precision.

Our highest accuracy is 42%. This is in line with previous attempts using pre-trained models that did not attempt to modify the dataset. This shows that the specific network out of the standard image recogntion models that you choose may not matter as much as pre-processing the dataset to make sure its balanced and that all labels are represented by enough training examples that allow the network to learn their features and distinguish them from other labels.

6. Conclusion & Future Work

VGG19 and Resnet101 performed better than AlexNet, suggesting that depth is more important than other factors when it comes to Movie-to-Genre classification. My best model accuracy was 42% which is substantially better than random guessing and is similar to State of the Art performance when no pre-processing of the dataset is involved. As seen in the confusion matrices, most of the error can be attributed to the network guessing a popular genre when it's confused about a less prevalent genre which suggests that providing more training examples of less-prevalent genre posters could boost the performance of our networks.

In the future, creating a better dataset should be strongly considered. The total number of movies on IMDb is less than 40,000 which is still a small dataset. Creating a bigger dataset might involve creating more than one poster for existing movies or creating fictional poster movies for each genre to augment the data. Another thing that needs to be done is to balance the dataset so that each Genre has equal representation. At the moment some Genres (Comedy, Drama, and Action) dominate the training examples more than other genres which means that the network won't learn the features of the less-prevalent Genres as well. One more thing to consider is feature engineering. Using segmentation and object recognition, or text-recognition on the posters and feeding the outputs of these networks as inputs to the genre-classification network will add valuable information that is hard to learn if not fed explicitly to the network and could prove very valuable to its learning.

7. Acknowledgments

I want to thank the CS231n teaching team for clear instructions and smooth logistical support throughout the course, Pranav Hari for generating the Kaggle dataset that saved me a lot of time web-crawling IMDb to download posters, and my friends and family for emotional support.

References

[1] Gabriel Barney and Kris Kaya. Predicting genre from movie posters. *Stanford CS 229: Machine Learning*, 2019. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[3] Yin-Fu Huang and Shih-Hao Wang. Movie genre classification using svm with audio and video features. In *International Conference on Active Media Technology*, pages 1–10. Springer, 2012. 2

[4] Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. Movie posters classification into genres based on low-level features. In *2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1198–1203. IEEE, 2014. 1

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 2

[6] Kaushil Kundalia, Yash Patel, and Manan Shah. Multi-label movie genre detection from a movie poster using knowledge transfer learning. *Augmented Human Research*, 5(1):1–9, 2020. 1

[7] Eric Makita and Artem Lenskiy. A multinomial probabilistic model for movie genre predictions. *arXiv preprint arXiv:1603.07849*, 2016. 1

[8] Mel-Leo Rosal. U.s. film industry statistics [2022]: Facts about the u.s. film industry, May 2022. 1

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 2

[10] Samuel Sung and Rahul Chokshi. Classification of movie posters to movie genres. In *Proceedings of the Workshop on*

Multimodal Understanding of Social, Affective and Subjective Attributes, 2017. 1

[11] Jeong A Wi, Soojin Jang, and Youngbin Kim. Poster-based multiple movie genre classification using inter-channel features. *IEEE Access*, 8:66615–66624, 2020. 1