Multi-Label Image Classification using Convolutional Neural Network (CNN)

Adithya Nair, Akshay Harikumar, Sreejith Kumara Pai
Department of Computer Science
Amrita Vishwa Vidyapeetham,
Amritapuri, India

***Abstract*—Movie posters encompasses the whole message and feeling of the film. Posters are more than just a promotional material, that captures a viewer's attention. A good poster is able to convey important qualities of a film such as theme and genre to make the movie seem as appealing as for a wide viewership as possible. CNN models are capable enough to solve single label classification of images, however their ability to classify multi label images with acceptable accuracy remains to be tested. This paper proposes an approach to use the pre-trained CNN-ResNet50 model in-order to categorise movie posters into it's corresponding genres. Based on the accuracy rate of learning and their ability to successfully predict the genres of the movie we were able to compare different models. Our CNN model was able to learn the images with a maximum accuracy of 60.21% when compared to our DNN model which gave a maximum accuracy of 23%.**

I. Introduction

A Movie Poster is the first impression of a movie. Posters are the face of a movie when it comes to promoting and advertising. Good posters are able to communicate important aspects of a film such as cast, theme, and elements of plot. Thus, designers have incentive to include salient features in their posters to make their movie attract more viewers. The ability to grasp the concept of the movie in such a way that, a viewer can identify the theme behind it, is crucial to a poster. The film-makers are able to get a feedback on how a viewer is able to visualize their film. Therefore a model which can extract the features of a movie poster and identify the genres can become handy for both the film-makers as well as the designers. In this problem we have used Multi-label Image Classification for identification of Genres from Movie Posters using Convolutional Neural Networks. The success of CNN on single-label image classification also sheds some light on the solution of the multi-label image classification problem. Generally, CNN can well handle images with well-aligned objects, while it is relatively inaccurate in predicting images with objects severely mis-aligned or occluded. Therefore, by relaxing the multi-label problem into several single-label tasks and alleviating the issues of mis-alignment and occlusion,the great discriminating ability of the CNN model can be better exploited. In the past, multi-label classification was mainly motivated by the tasks of text categorization and medical diagnosis. Examples range from news articles to emails. Nowadays, we notice that multi-label classification methods are increasingly required by modern applications. In semantic scene classification, a photograph can belong to more than one conceptual class at the same time, such as sunsets and beaches. Similarly, in music categorization, a song may belong to more than one genre. The same approach can be employed to find the genres that a movie belongs to, based on the posters.

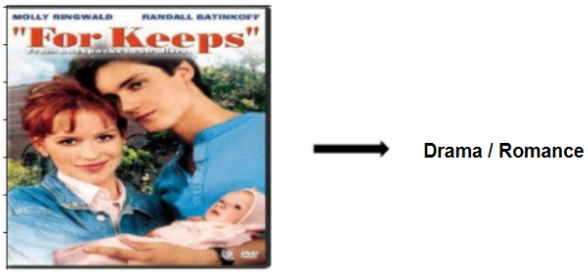

Fig. 1: Prediction Of New Data

II. Related Works

Work from Daniel Gardner (2017) which focused on classification of Satellite Images using multiple models. Early work from Barnard et al. (2002) that focused on identifying objects in particular sub-sections of an image. Research from Ying Hong (2021) involving classification of clothing images, which gave us a broad view on how to approach the problem of Multi-Label Image Classification using Convolutional Neural Network.

The work from Luka Popovic (2020) which gave us a strong platform to build upon, helped us identify the neural network architecture which suits our problem. Work of Hossain et al. (2021), gave us an idea on how to approach multi label classification of movie poster as well as how to preprocess data for more performance. His work also showed the effective use of evaluation metrics like hamming, 0/1 loss etc for getting appropriate accuracy scores from models. There have been mentions about making a custom model for predicting the genre of movie posters. The work from Simões et al. (2016) gave us the motive of approaching an advanced version of our problem, by extracting the genres of a movie through video-analysis of it's scenes.

III. Methodology

Our training model uses the pre-trained CNN Resnet50 architecture, which consists of 50 residual block layers. A Residual Neural Network makes the learning process of a deep neural net much faster when compared to other CNN models. With the additional help of back-propagation and several hyper-parameter tuning techniques, we were able to minimize the loss function and improve our model by increasing the number of epochs and providing optimal learning rates.

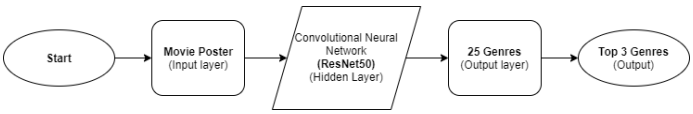

Fig. 2: Block Diagram of a CNN Model

A. DataSet

We have collected a dataset from IMDB which contains Hollywood movies released between the years 1980 to 2015. Dataset includes 7867 images of movie posters with 25 genres where each posters belonging to more than one genre.

Each sample in the dataset consists of the image ID, corresponding categories of genres in text and the labels, where each label in the dataset consists of 25 genres which are one hot encoded according to their respective genres.

B. Data Preprocessing

Our image dataset consists of 3 channel RGB color space images of varying sizes, which had to be resized to a dimension of 350 x 350 pixels for better and efficient training. The resized images were then transformed to tensors. The dataset was then later split into training and test samples of sizes 85% and test 15% respectively, and loaded into training and testing dataloaders in batches of size 33.

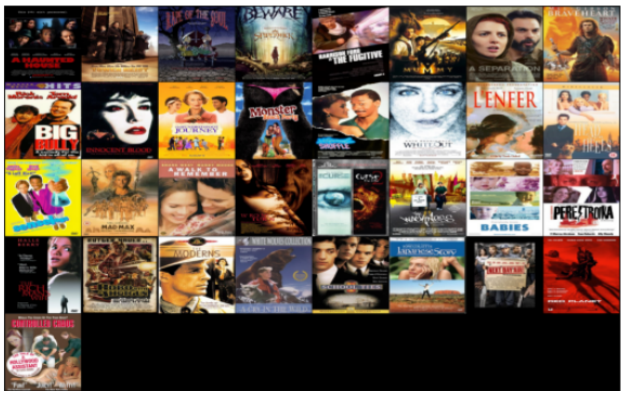

Fig. 3: Image batches of dimension 350 x 350

C. Hyper Parameter Tunings

Our training process makes use of a gradient descent optimization method, which is able to find optimal values for the weights and biases such that the loss function is minimum. The model parameters are updated using Adaptive Movement Estimation algorithm, or Adam which has the capability to adapt to the learning rate of each parameter by keeping track of the exponentially decaying average of the past gradients along with exponentially decaying average of the past squared gradients. A typical ResNet model consists of several residual layers stacked on top of each other, which aims to learn deeper networks without drop in performance. Each residual block consists of two weighted layers, an activation layer(ReLu) and a skip connection between the two weighted layers. The main objective of a Residual Neural Network is to learn the difference between the input and the output and determine how accurate it can be to the output by combining the learning differences and the actual input. Our proposed ResNet50 model consists of 50 such residual layers capable of learning deeper networks in an optimized mode by reducing the number of parameters which in-turn narrows the complexity. At the end use a Sigmoid activation function to classify the values within the final node into a probability score between 0 and 1 for easier classification.

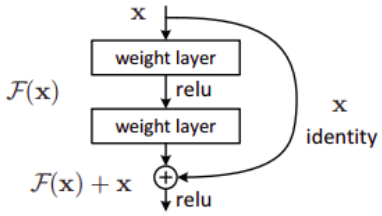

Fig. 4: Residual Block Diagram

D. Loss Function

When coming to loss function, we have used Binary Cross Entropy Loss (BCE Loss) which can be used along with sigmoid activation. Binary cross entropy compares each of the predicted probabilities to actual class output which can be either 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Fig. 5: Binary Cross Entropy Loss

IV. Solution Approach

Since the labels were one hot encoded, in order to test our dataset correctly, we had to transform the model outputs which were in their corresponding probability scores to one

hot encoded values (1's and 0's). Therefore we decided to go with the top three probability scores in each sample, convert those values to 1 and the remaining to 0's and and predict the corresponding three genres.

V. Experimental Results and Performance Evaluation

A. Model Evaluation

Different models were experimented, and their corresponding testing as well as training accuracies were computed.

VGG-16 : VGG16 increased the layer from Alexnet by having 13 convolutional and 3 fully connected layer. Hidden layers were activated by ReLu activation layer.

ResNet50 : ResNet50 took note of saturation of accuracy due to vanishing/ exploding gradient affecting the performance of the model. Hence, Researchers introduced skip connection for deeper neural network.

InceptionV3 : InceptionV3 mainly concentrated on making the model computationally efficient in terms of number of parameters and economical cost.

EfficientNet : EfficientNet proposed a new scaling method for unifromly calculating the depth, width and resolution of network. It designed a new baseline neural network and eventually scale it up to a deeper neural network.

Model	Test Accuracy	Training Accuracy
VGG-16	54.98%	60.78%
ResNet50	60.21%	61.56%
InceptionV3	46.27%	46.85%
EfficientNet	45.19%	74.48%

B. Result

We are convinced with the ResNet50 model, which gave an efficient result for our dataset at 25 epochs. We observed that, further increase in epochs would result in higher accuracies.

```
Epoch: 26/30, Test acc: 60.21, Train acc: 61.56
Epoch: 27/30, Test acc: 59.41, Train acc: 61.53
Epoch: 28/30, Test acc: 59.49, Train acc: 61.61
Epoch: 29/30, Test acc: 59.41, Train acc: 61.97
```

Fig. 6: Accuracy of ResNet50 model

Fig. 7: Prediction Of New Data

VI. Data Visualization

The below graph depicts the loss variation with increasing number of epochs.It is observed that when we increase the epoch at first, the loss increases but eventually it decreases as the number of epochs increases further.

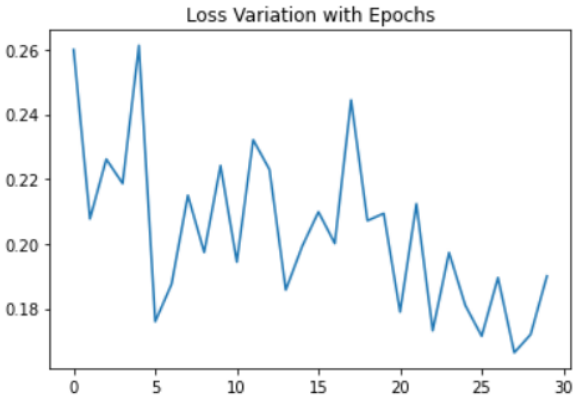

Fig. 8: Loss Variation with Epochs

The graph below depicts the decrease in loss with respect to increase in Number of Iteration.During initial iterations, the loss remains to be high but it decreases as number of iterations increase. Resulting to a loss of around 0.20 - 0.23.

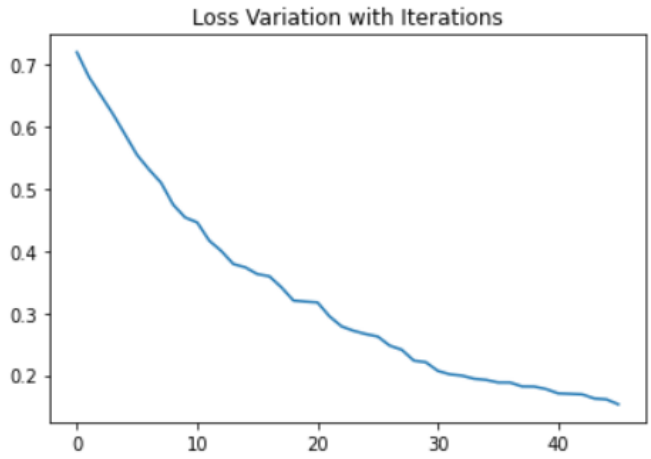

Fig. 9: Loss Variation with Iterations

VII. References

Barnard, Forsyth, Duygulu, P., Kobus, Freitas, J., and David (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary.

Daniel Gardner, D. N. (2017). Multi-label classification of satellite images with deep learning.

Hossain, N., Ahamad, M. M., Aktar, S., and Moni, M. A. (2021). Movie genre classification with deep neural network using poster images.

Luka Popovic, Santiago Cepeda, N. S. (2020). Movie genre classification using convolutional neural networks.

Simões, G. S., Wehrmann, J., Barros, R. C., and Ruiz, D. D. (2016). Movie genre classification with convolutional neural networks.

Ying Hong, N. (2021). Research on multi-label clothing image classification based on convolutional neural network.