**Movie Genre Classification based on Poster Images with**

**Convolutional Neural Networks**

# Jordan Tangy

### 3rd Year Student in Computer Science and Mathematics at Ariel University, Israel

Ron Sider

### 3rd Year Student in Computer Science and

### Mathematics at Ariel University, Israel

**ABSTRACT**

Over the last few decades, Artificial Neural Networks are showing great performances is many fields such as voice and face recognition, generating new Data, language translation, Image classification and much more. One of the most popular deep neural networks is the Convolutional Neural Network (CNN). The interest in having Convolution connected to a deep NN has recently begun to surpass classical methods performance in different fields; especially in pattern recognition. CNN take this name from mathematical linear operation between matrixes called convolution. They have multiple layers; including convolutional layer, non-linearity layer, pooling layer and fully-connected layer.

Among different type of models, Convolutional neural networks has been demonstrated high performance on image classification. In this project we built a Convolutional neural network for Movie Poster multi-label (genres) classification of Movies. The data set used was taken from Kaggle and originally collected from IMDB website. It contains 7868 different movie Posters released between 1980 and 2015. In this paper we will explain how we used a Convolutional Neural Network in order to perform Movie Genre Classification based on Movie Posters and because a movie may belong to multiple genres, this is a multi-label image classification problem. In addition, we will also state the parameters that effect CNN efficiency and how CNN was implemented to obtain good results to classify the images.

**CCS CONCEPTS**

* **Computing methodologies** → **Image representations**; *Vi- sual content-based indexing and retrieval*; *Neural networks*;

## KEYWORDS

Movie genre classification, movie poster, multi-label classification, Convolutional Neural Network

## INTRODUCTION

Image attribute extraction has been widely studied in recent years, since visual attributes can boost various tasks such as image re- trieval and image captioning . Most prior studies either focus on detecting or recognizing visual entities like object and scene, or extracting semantic concepts embedded in images. These attributes really provide widespread influence on multimedia re-trieval and many computer vision applications. On the other hand, some visual attributes are implicit but can be easily perceived by human beings.

### Une image contenant texte, tableau blanc Description générée automatiquement Une image contenant texte, livre Description générée automatiquement Une image contenant texte, extérieur, personne, homme Description générée automatiquement

(a) (b) (c) (d)

**Figure 1: Sample movie poster images. (a) *Captain America: The First Avenger* (Action , Adventure, Sci-Fi; (b) A Star is Born (Drama, Music , Romance); (c) *Expendables 2* (Adventure, Action, Thriller); (d) *Interstellar* (Sci-Fi , Mystery , Adventure ).**

When studying visual attributes for different types of images, we found that movie poster image contains different types of attributes and according to the genre, we can identify certain patterns. First, movie posters are created to attract people paying time and money to watch the corresponding movie. Information on a movie poster, therefore, should be attractive. Figure [1](#_bookmark0) shows four sample posters. In the figure caption we show movie name and the corresponding genres (in the parentheses), which are obtained from the IMDB1 website. Most of them present the most important imagery of the corresponding movie. For example, Figure [1(a)](#_bookmark1) shows a violent scene with army aircrafts, fire in the background and a person holding a shield. This dramatic scene attracts people who like excitement or combat. Second, different movies target at different populations, and movie posters should concisely present genre information. For example, we can clearly perceive that Figure [1(a)](#_bookmark1) and Figure [1(c)](#_bookmark2) present action elements, Figure [1(d)](#_bookmark3) seems mysterious, and Figure [1(b)](#_bookmark4) describe a love story. Third, movie posters usually present important objects like gun and fire in Figure [1(c)](#_bookmark1) and the shield of Captain America in Figure [1(b).](#_bookmark2) Overall, we can identify different components in one image that give an indication to the people who are about to watch which kind of movie it is. Those component combined create a pattern that we want to detect in order to classify the image.

In this work, we propose to analyze movie poster images and

classify them into movie genres based on a Convolutional neural network.

1 [http://www.imdb.com](http://www.imdb.com/)

we can determine the genre of a movie poster. This work corre- sponds to the second aforementioned observation, and could be the fundamental module for movie content access, management, and presentation. Estimating subtle attributes from only images is actu- ally not a totally new idea. The work in [[2](#_bookmark24)] shows that a person’s first name can be roughly predicted based only on his/her face im- age. Other attributes like occupation can also be predicted to some extent based on face image [[3](#_bookmark25)]. The reasons for enabling such pre- dictions are that first name is statistically related to genders, races, and when a person borns, and occupation is statistically related to genders, ages, and other body context. These inspiring works motivate us to study implicit factors related to movie genres, and encourage us to build a computational model to do classification.

Contributions of this work are summarized as follows.

To promote and facilitate the proposed movie poster anal- ysis, we collect a large-scale movie poster dataset from the IMDB website. This dataset mainly consists of posters of movies released from 1980 to 2015 in Hollywood, as well as the associated metadata like movie genre, names of the director and main actors, box o@ce, and so on.

•

We construct a computational model based on deep neu- ral network to automatically classify a movie poster into genres. Note that one movie belongs to multiple genres, see Figure [1.](#_bookmark0) This task is therefore a multi-label image classification problem.

•

Note that, although classifying a movie poster image into genres is the main target of this work, potential contributions of this study is to limited to this. With movie genre classification, we may be able to construct a movie recommender system assuming that a person likes movies of similar genres. The relationship between objects and movie genres can be discovered, and can be important clues for amateurs to design posters. Based on the experience of the model construction, we may extend this computational model to estimate other movie properties like box o@ce.

The rest of this paper is organized as follows. Section [2](#_bookmark6) provides brief literature survey on movie genre classification. Section [3](#_bookmark7) de- scribes the proposed model to achieve movie genre classification based on poster images. Details of the multi-label classification prob- lem are provided. Evaluation results and discussion are provided in Section [4,](#_bookmark12) and Section [5](#_bookmark16) concludes this paper.

## RELATED WORKS

There have been a few works on movie genre classification. Wei-Ta Chu and Hung-Jui Guo (Academics from National Chung Cheng University, Taiwan ) proposed an interesting approach to classify movie posters. They used a large movie poster dataset and another file that is mapping the poster according to its genres. For example if the genres associated to movie X are action, comedy and adventure, then the number 1 will be assigned to those genres and to the rest will be assigned the number 0.

Let’s break there approach into three parts.

They ***first*** build a Convolutional Neural Network composed of seven convolutional layers followed by a batch normalization layer. After normalizing, the feature maps are flattened as a vector to be a visual representation. Then this goes through a fully connected layer and another batch normalization layer. The ***second*** part is to input the image into the YOLO (You Only Look Once) algorithm which is an algorithm used for object recognition, i.e. that the algorithm will recognize different kind of objects on the poster e.g. cars, dogs, cats, persons, books, instruments, etc… .

The detected objects are flattened and are going through a fully connected layer. The output of the ***first*** and the ***second*** sub networks are combined and embedded to a fully connected layer, from there to a batch normalization and finally, this goes through softmax function in order to classify the results.

The main idea is to extract objects from the picture and to train the Network to understand that there are relations between certain objects and genres. For instance, if on the picture we see persons holding guns and tanks in the background, the YOLO algorithm will identify those objects. When these objects are embedded with the previous output of the convolution, the NN will start to establish a link between guns/tanks and genre War/Action. It finds out that this method is powerfull, because usually, objects on posters are a good indicator on the genre of the movie..

## MOVIE GENRE CLASSIFICATION

* 1. **Deep Neural Network**

Given a set of training data *D* = *X* , *Y* , where *X* = (*x*1, *x*2, ..., *xN* ) is the set of *N* poster images and *Y* = (*μ*1, *μ*2, ..., *μN* ) is the corre- sponding genre information. The vector *μi* = (*yi*, 1, *yi*, 2, ..., *yi*, *M* ) is a binary vector, where *yi*, *j* = 1 indicates that the *i*th poster belongs to the *j*th genre. Note that a movie may belong to mul- tiple genres, i.e., 1 *j yi*, *j M*. Based on *D*, we would like to construct a computational model that outputs the probabil- ity of a given poster image *xi* belonging to each movie genre. That is, the constructed model acts as a function such that (*x* ) = (*y*ˆ*i*, 1, *y*ˆ*i*, 2, ..., *y*ˆ*i*, *M* ), where 0 *y*ˆ*x*, *j* 1. A good model would output the estimated vector (*y*ˆ*i*, 1, *y*ˆ*i*, 2, ..., *y*ˆ*i*, *M* ) as close to

.≤ ≤

{ }

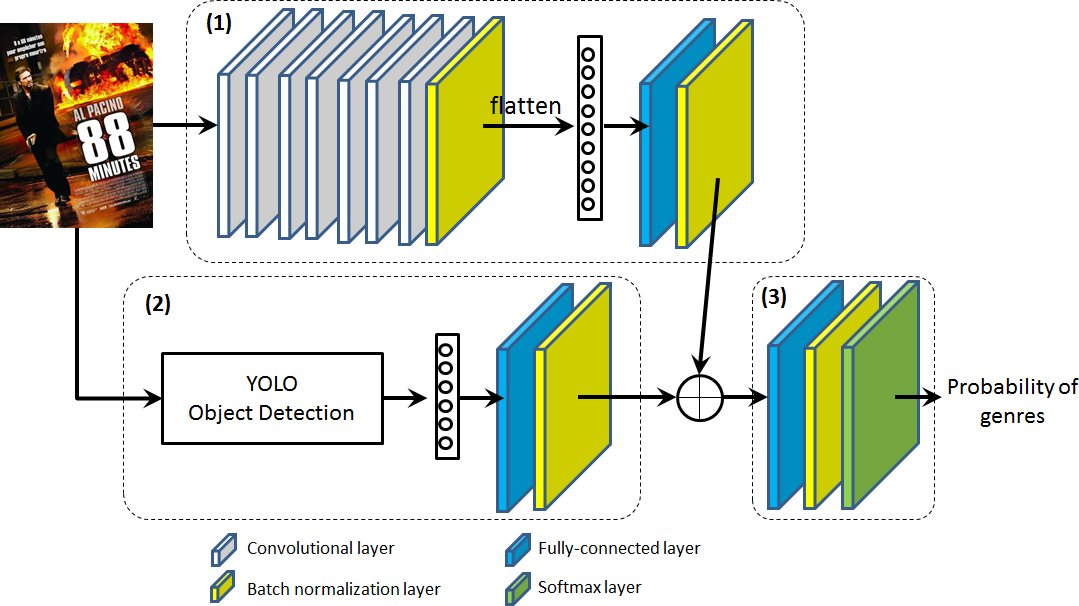
F ≤ ≤

F

the ground truth vector (*yi*, 1, *yi*, 2, ..., *yi*, *M* ) as possible.

In this work, we construct the function by jointly consider- ing visual representation extracted from a convolutional neural network, and object information extracted by one state-of-the-art object detector [[12](#_bookmark36)]. Figure [2](#_bookmark9) illustrates the overall network struc- ture. To extract effective visual representation, we construct the convolutional neural network similar to the convolutional part of AlexNet [[9](#_bookmark31)], as shown in the first part of Figure [2.](#_bookmark9) This network consists of seven convolutional layers, where the seventh convo- lutional layer is followed by a batch normalization layer [[5](#_bookmark27)]. After normalization, feature maps are fiattened as a vector to be the vi- sual representation. The visual representation will be combined with object information in the third part of Figure [2.](#_bookmark9) To reduce heterogeneity and enable feature combination, we would like to

F



#### Figure 2: Structure of the proposed deep neural network.



map the visual representation and the object information into a common space. Therefore, the visual representation is embedded by a fully-connected layer with batch normalization, as shown in the end of Figure [2(1).](#_bookmark9)

In addition to visual appearance, we also consider object seman- tics that may highly relate to movie genres. For example, in action movies, *person* and *car* may more often appear on posters; while in romantic movies, the hero and the heroine are often the main objects on posters. To explore how object information provides clues for genre classification, we detect objects by the YOLO system [[12](#_bookmark36)], as shown in Figure [2(2).](#_bookmark9) Most state-of-the-art object detection systems first find region proposals and then classify the proposals into one of the predefined object classes. Instead, the YOLO system formulates object detection as a regression problem. An image is divided into grids, and from each grid the system predicts several bounding boxes and their associated confidence values. The confi- dence value conceptually represents the extent of overlap between the predicted bounding box and the ground truth. The kernel model for bounding box prediction is a convolutional neural network con- sisting of 24 convolutional layers followed by 2 fully-connected layers. This structure is inspired by the GooLeNet model, and pro- vides real-time object detection and classification. Details of the YOLO system please refer to [[12].](#_bookmark36)

We adopted the YOLO version 2 [[13](#_bookmark38)] trained based on the MSCOCO

dataset. Given a poster image, this system outputs bounding boxes of detected objects and their associated confidence values, e.g., *c*1, ..., *cK* if there are *K* detected objects. Figure [3](#_bookmark10) shows sample results of object detection for poster images. To represent object information of a poster, we sum confidence values of objects of the same type. That is, *oi* = *cj* + *ck* if both the *j*th object and the *k*th object are object *i*, say a *person*. A poster image is then represented as the vector *o* = (*o*1, ..., *oB* ). Note that if there are more *i*th objects, the value *oi* tends to be larger; on the other hand, if there is no *i*th object in this poster, the value *oi* would be zero. In our adopted YOLO system, totally 80 object classes are detected, including *per- son*, *car*, *dog*, and so on, and therefore the vector *o* is 80-dimensional (*B* = 80). As shown in Figure [2(2),](#_bookmark9) the vector *o* is also embedded by a fully-connected layer with batch normalization.

{ }

The embedded visual representation and object information are concatenated as the input of the third part of Figure [2.](#_bookmark9) This part consists of one convolutional layer, one batch normalization layer,

#### Figure 3: Sample results of object detection for poster im- ages.

and finally one softmax layer. Given a poster *xi* , this network finally outputs the estimated probability vector *μ*ˆ*i* = (*y*ˆ*i*, 1, *y*ˆ*i*, 2, ..., *y*ˆ*i*, *M* ), where *y*ˆ*i*, *j* indicates the probability of the poster *xi* being the *j*th

movie genre.

Detailed configurations of the network are shown in Table [1.](#_bookmark11) The item conv3-100, for example, means that the convolution kernel is

1. 3, and the number of filters is defined as 100. The activation

×

function of each layer is ReLU, the objective function is the mean square error between estimated probability vector *μ*ˆ and ground truth vector *μ*, and the optimization algorithm is SGD with the

learning rate 0.1. The training process was conducted in 20 epochs, with mini-batch size 128.

## Multi-label Classiftcation

The output of the network shown in Figure [2](#_bookmark9) is an *M*-dimensional probability vector, where each dimension indicates how likely a given poster belongs to a movie genre. As shown in Figure [1,](#_bookmark0) a movie usually belongs to multiple genres. In addition, the number of genres a movie belongs to would be varied. One intuitive way to solve this multi-label classification problem is thresholding each dimension. If the value of the *i*th dimension of *μ*ˆ is larger than a threshold, we say the given poster belongs to the *i*th movie genre. However, how to define the threshold for each dimension (different dimensions may be associated with different thresholds) obviously is the main problem.

#### Table 1: Detailed conftgurations of the proposed network.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | input (100 × 100 poster images) | | |  |  |
| Part 1 | conv3-100 maxpooling | conv3-32 | conv3-64 | conv3-128 | conv3-256 | conv3-128 | batch norm. maxpooling fiatten | fully-connected (80 nodes) batch norm. |
| Part 2 | YOLO object detection  fully-connected (80 nodes) batch norm. | | | | | | | |
| Part 3 | fully-connected (256 nodes)  batch norm. softmax | | | | | | | |

In this work, we adopt a grid search scheme based on Matthews correlation coe@cients to determine the (nearly) best thresholds for each dimension. Details of the best threshold finding algorithm is shown in Algorithm [1.](#_bookmark13) Basically, the idea is adjusting the thresh- old for each dimension separately. Every time when we adjust the threshold for a dimension, we binarize the predicted probability vector with the adjusted threshold, and calculate the Matthews cor- relation coe@cient (the *MC* function shown in line 8) between the thresholded vector and the ground truth vector. Note that the nota- tion *μ*[1 : *i*] denotes the vector with values of the first *i* dimensions are from the vector *μ*, and values of the remaining *M i* dimensions are set as 0. That is, when we try to find the best threshold for the *i*th dimension, all dimensions from 1 to *i* 1 are jointly considered. For the *i*th dimension, the Matthews correlation coe@cients for all thresholds are stored in the vector *ρ*, from which we find the index of threshold that causes the largest correlation (line 13). The best threshold for the *i*th dimension is then determined by the cor- responding value (line 14). After checking all dimensions, the set of thresholds *T* = *t*1, ..., *tM* that yields the largest correlation is found.

−

−

{ }

Given a test poster image, the framework shown in Figure [2](#_bookmark9) out- puts the probabilities *y*ˆ1, *y*ˆ2, ..., *y*ˆ*M* of this poster being a specific genre. If the estimated probability *y*ˆ*i* is larger than the correspond- ing threshold *ti* , we say that this poster belongs to the *i*th genre.

{ }

## EVALUATION

* 1. **Databases**

To evaluate the proposed study, we collect a movie poster image dataset from the IMDB website. We crawled one poster per Holly- wood movie released from 1980 to 2015, and collected 8,191 poster images in total. Resolutions of these poster images range from

89 132 to 300 581. The genre information associated with each

× ×

movie is also collected. There are totally 23 different genres. Fig- ure [4(a)](#_bookmark17) shows the distribution of the numbers of posters in different movie genres. Note that the total number of posters shown in this figure is much larger than 8,191, because one movie usually belongs to multiple genres. From Figure [4(a),](#_bookmark17) we see that the numbers of movies in different genres are quite imbalanced.

To diminish the infiuence of imbalance in training, we conduct different extents of augmentation for different genres. The genres *Drama* and *Comedy* are not augmented, because they already have a

**ALGORITHM 1:** Find Best Threshold

**Input:** Predicted probability vector *μ*ˆ = (*y*ˆ1, ..., *y*ˆ*M* ), truth probability vector *μ* = (*y*1, ..., *yM* ), threshold upper bound *u*, and threshold stride *s* ,

**Output:** The set of best thresholds *T* = {*t*1, ..., *tM* }

**1** Initialize the set of threshold *T* = {*t*1 = 0, *t*2 = 0, ..., *tM* = 0}

**2 for** *i* = 1 *to M* **do**

**3** *j* = 0

**4** Initialize an empty vector *ρ*

**5** Initialize an empty vector *θ*

**6 for** *j* < *u* **do**

**7** Initialize an *M* -dim binary vector

*b* = (*b*1 = 0, *b*2 = 0, ..., *bM* = 0)

**8** if *y*ˆ*i* > *j* , *bi* = 1. Otherwise, *bi* = 0

**9** *ρ* = *ρ* .*append* (*MC* (*μ*[1 : *i* ], *b* ))

**10** *θ* = *θ* .*append* (*j* )

**11** *j* = *j* + *s*

**12 end**

**13** *k* ∗ = arg max*k ρ* = (*ρ*1, ..., *ρk* , ...)

**14** *ti* = *θ* [*k* ∗]

**15 end**

**16** Output a set of threshold *T* = {*t*1, ..., *tM* }.

large number of posters. For the genres *Action*, *Romance*, *Crime*, *Ad- venture*, *Thriller*, and *Documentary*, we randomly crop two 150 150 regions from the original images. Therefore, together with the orig- inal images, the volume of data for these genres is increased twice. Similarly, the volume of data in genres *Horror*, *Biography*, *Family*, *Fantasy*, and *Sci-Fi* is increased seven times; the volume of data in genres *Short*, *Music*, *Animation*, *History*, and *Sport* is increased fifteen times; and the volume of data in genres *War*, *Musical*2, and *Western* is increased thirty times. After data augmentation, the total number of distinct images is 16,997. By summing the number of images in all genres, the total number is 36,295.

×

In addition to genre information, we also collect other metadata for future study, including Box O@ce, Rated, Awards, Director, Writer, imdbVotes, imdbRating, Actors, and Metascore. For example, Figure [4(b)](#_bookmark18) shows the distribution of the number of posters with

2 A *Music* movie presents content related to music, like biography of a musician or a story of a band; while in an *Musical* movie, actors largely sing songs to present narrative.

different ratings. Ratings in the IMDB website range from 0 to 10. As can be seen, most movies get ratings ranging from 5 to 8, i.e., a nearly normal distribution can be seen. Figure [4(c)](#_bookmark19) shows the distribution of the number of movies in different box o@ce ranges. As we expect, most movies are not sold well. The box o@ce of around 80% of the movies is less than 50 millions US dollars. Note that the values of box o@ce are adjusted for infiation according to statistics provided in [[1].](#_bookmark23)

Combining poster images with these metadata, we believe this is the first heterogeneous movie poster dataset that may facilitate not only genre classification, but also several other studies such as the relationship between actors and genres, the relationship between director and box o@ce, and so on. This database will be released in public soon.

* 1. **Performance of Movie Genre Classiftcation** To evaluate movie genre classification based on poster images, we randomly select 75% of images from each genre for training the proposed model, and the remaining images are used for testing. The performance metric is accuracy defined as

*Accuracy* = ǁ*μ* AND *b* ǁ × 100%, (1)

ǁ*μ* OR *b* ǁ

where *μ* and *b* are truth labels and predicted labels (in binary vector form), respectively. The notation denotes L1-norm of a vector. We compare performance obtained by one baseline system and several neural network variants. Detailed settings of these methods

ǁ·ǁ

are described as follows.

* + 1. Baseline system [[7](#_bookmark29)]: To our best knowledge, this is one of the few prior works on movie poster classification. We implement the best setting reported in [[7](#_bookmark29)] to be the base- line system. They represented a poster image by low-level features including dominant colors, edge-based features, and number of faces. They first find feature centroids of im- ages in each genre. Given a test poster image, the distance between it and centroid of each genre is calculated, and the label of each poster is determined as the genre with the nearest feature centroid.

•

* + 1. Fully-connected neural network: We first try a simple fully-connected neural network to do poster classification. Five fully-connected layers with each consisting of 512 nodes are connected, followed by one softmax layer. Other training settings, such as activation function and mini- batch size, are the same as the proposed method mentioned in Section [3.1.](#_bookmark8)

•

* + 1. Convolutional neural network: We also try a CNN to do poster classification. Eight convolutional layers are con- nected, with the numbers of filters 100, 32, 64, 128, 256, 256, 512, and 512. Output of the final convolutional layer is fiattened, and then fed to one fully-connected layer (512 nodes) followed by a softmax layer. Settings for training are the same as the aforementioned.

•

* + 1. Object information only: Only the second part and the third part of Figure [2](#_bookmark9) are considered. This setting is used to demonstrate how well genre classification can be done if only object information is adopted.

•

* (5) The proposed structure illustrated in Figure [2.](#_bookmark9)

#### Table 2: Accuracy of movie genre classiftcation based on dif- ferent variants (%).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Accuracy | 7.31 | 14.05 | 14.34 | 15.79 | 18.73 | 17.70 | 15.36 |

1. The proposed structure without batch normalization layers: Comparing with the setting in (5), this setting is used to demonstrate the effectiveness of batch normaliza- tion.

•

1. The proposed structure with the third part replaced by SVM: The third part of Figure [2](#_bookmark9) acts as a classifier. It is interesting to investigate a classification scheme differ- ent from neural network. Therefore, we try to construct an SVM classifier based on the fused feature vectors, and evaluate performance obtained by this approach.

•

Table [2](#_bookmark15) shows performance comparison of different variants. The first column shows that simple low-level features and nearest neighbor classifiers do not work well when the number of genre is increased to 23 (in [[7](#_bookmark29)], only a limited poster dataset consisting of 6 genres was used). Table [2(2)](#_bookmark15) and Table [2(3)](#_bookmark15) show that neural network-based methods provide much better performance than [[7](#_bookmark29)]. This confirms to the current research trend, especially when a large number of training data are available. Table [2(4)](#_bookmark15) shows that the effectiveness of object information, even if the object detectors in YOLO are not specially constructed for movie posters. The pro- posed model combines visual appearance and object information, and the best performance can be achieved, as shown in Table [2(5).](#_bookmark15) Comparing with Table [2(5)](#_bookmark15) with Table [2(6),](#_bookmark15) we verify the effective- ness of batch normalization [[5](#_bookmark27)]. Last, Table [2(7)](#_bookmark15) shows that the constructed SVM classifier does not work better than the neural network classifier in our case.

Figure [5](#_bookmark20) shows sample classification results, including two suc- cess cases (top two) and two failure cases (bottom two). Figure [5(a)](#_bookmark21) shows warm colors, which are highly related to romance. For Fig- ure [5(b),](#_bookmark22) the YOLO object detection module works quite well in detecting animals, which often appear in animation movies. On the other hand, both Figure [5(c)](#_bookmark32) and Figure [5(d)](#_bookmark33) are presented in a more abstract way. The visual appearance of Figure [5(c)](#_bookmark32) looks a little bit horrible, and Figure [5(d)](#_bookmark33) looks like showing an animation. The truth labels of Figure [5(c)](#_bookmark32) are crime, drama, and mystery, where crime is actually related to horror. The truth labels of Figure [5(d)](#_bookmark33) are short, comedy, and drama, which is quite di@cult to infer if we just look at one poster image without watching the movie or the movie trailer.

## 5 CONCLUSION

We have presented a system to classify movie poster images into genres. A deep neural network is proposed to jointly consider visual appearance and object information, and a classifier is constructed to estimate the probabilities of a poster belonging to different genres. Multi-label classification is achieved by thresholding the estimate probabilities, with the thresholds adaptively determined by a grid search scheme. To evaluate the proposed system and facilitate future research, we collect a large-scale movie poster dataset consisting of

**Box Office**

3000

4000

2500

4000

3000

2000

Number

3000

2000

Number

1500

2000

Number

1000

1000

500

1000

0 0 0

0~1

1~2

2~3

3~4

4~5

5~6

6~7

7~8

8~9

Rating

Drama Comedy Action Romance Crime Thriller Adventure Document. Horror Biography Mystery Family Fantasy Sci-Fi Short Music Animation History Sport

War Musical Western Other

9~10

0~50M

50~100M

100~150M

150~200M

200~250M

250~300M

300~350M

350~400M

400~450M

450~500M

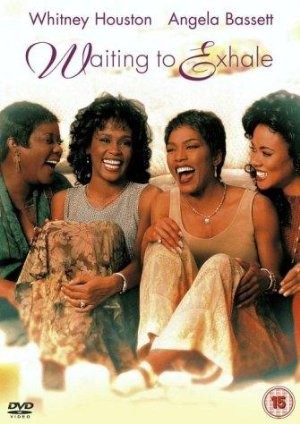
>500M

(a)

(b)

(c)

#### Figure 4: (a) The distribution of the numbers of posters in different movie genres. (b) The distribution of the numbers of posters with different ratings. (c) The distribution of the numbers of posters in different box oÆce ranges.

1. Predicted: Romance, Drama, Comedy; (b) Predicted: Adventure, Animate,

8,191 distinct movies belonging to 23 different genres. The evalua- tion results show that promising performance over previous works can be achieved by the proposed model. In the future, we plan to improve classification performance by incorporating more infor- mation, and utilize the rich metadata to investigate more movie properties.

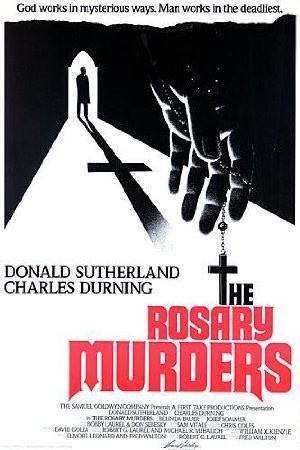
## ACKNOWLEDGMENTS

This work was partially supported by the Ministry of Science and Technology of Taiwan under the grant MOST 105-2628-E-194-001- MY2 and MOST 106-3114-E-002-009.

## REFERENCES

1. 2017. *Box O¶ce Mojo*. [http://www.boxo@cemojo.com.](http://www.boxofficemojo.com/)

Ground truth: Romance, Drama, Com- edy



Comedy; Ground truth: Adventure, Animate, Comedy

1. Huizhong Chen, Andrew C. Gallagher, and Bernd Girod. 2013. What’s in a Name? First Names as Facial Attributes. In *Proceedings of IEEE International Conference* *on Computer Vision and Pattern Recognition*.
2. Wei-Ta Chu and Chih-Hao Chiu. 2017. Predicting Occupation from Images by Combining Face and Body Context Information. *ACM Transactions on Multimedia* *Computing, Communications, and Applications* 13, 1 (2017), Article No. 7.
3. Wei-Ta Chu and Yi-Ling Wu. 2016. Deep Correlation Features for Image Style Classification. In *Proceedings of ACM International Conference on Multimedia*. 402–406.
4. Sergey Ioffe and Christian Szegedy. 2014. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of* *International Conference on Machine Learning*.
5. Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. 2014. Automatic Movie Posters Classification into Genres. In *Proceedings of ICT Innovation*. 319–328.
6. Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. 2014. Movie Posters Classifi-

cation into Genres based on Low-level Features. In *Proceedings of International Convention on Information and Communication Technology, Electronics and Mi-* *croelectronics*.

1. Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. 2016. *Judging a Book By its Cover*. [https://arxiv.org/abs/1610.](https://arxiv.org/abs/1610.09204) [09204.](https://arxiv.org/abs/1610.09204)
2. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classi-

(c) Predicted: Action, Horror, Documen- (d) Predicted: Adventure, Documentary,

fication with Deep Convolutional Neural Networks. In *Proceedings of Advances*

tary; Ground truth: Crime, Drama, Mystery

Animate; Ground truth: Short, Com- edy, Drama

*in Neural Information Processing Systems*.

1. Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James. Z. Wang. 2015. Rating Image Aesthetics Using Deep Learning. *IEEE Transactions on Multimedia* 17, 11

#### Figure 5: Success (top two) and failure (bottom two) classift- cation cases.

(2015), 2021–2034.

1. Zeeshan Rasheed and Mubarak Shah. 2002. Movie Genre Classification by Exploiting Audio-Visual Features of Previews. In *Proceedings of International* *Conference on Pattern Recognition*.
2. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of IEEE*

*Conference on Computer Vision and Pattern Recognition*.

1. Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *arXiv* *preprint arXiv:1612.08242* (2016).
2. Behjat Siddiquie, Rogerio S. Feris, and Larry S. Davis. 2011. Image Ranking and Retrieval based on Multi-Attribute Queries. In *Proceedings of IEEE International* *Conference on Computer Vision and Pattern Recognition*. 801–808.
3. Gabriel S. Simoes, Jonatas Wehrmann, Rodrigo C. Barros, and Duncan D. Ruiz. 2016. Movie Genre Classification with Convolutional Neural Networks. In *Proceedings of International Joint Conference on Neural Networks*.
4. Kazuma Takahashi, Keisuke Doman, Yasutomo Kawanishi, Takatsugu Hirayama,

Ichiro Ide, Daisuke Deguchi, and Hiroshi Murase. 2016. A Study on Estimating the Attractiveness of Food Photography. In *Proceedings of IEEE International* *Conference on Multimedia Big Data*.

1. Jonatas Wehrmann and Rodrigo C. Barros. 2017. Convolutions through Time for Multi-label Movie Genre Classification. In *Proceedings of the Symposium on* *Applied Computing*. 114–119.
2. Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2016. *Boosting* *Image Captioning with Attributes*. [https://arxiv.org/abs/1611.01646.](https://arxiv.org/abs/1611.01646)
3. Howard Zhou, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. 2010. Movie Genre Classification via Scene Categorization. In *Proceedings of ACM International Conference on Multimedia*. 747–750.