Neural Comput & Applic
DOI 10.1007/s00521-017-3162-x

ORIGINAL ARTICLE

Predicting movie box-office revenues using deep neural networks

Yao Zhou[1] · **Lei Zhang**[1] · **Zhang Yi**[1]

Received: 4 April 2017 / Accepted: 15 July 2017
© The Natural Computing Applications Forum 2017

Abstract In the film industry, the ability to predict a movie's box-office revenues before its theatrical release can decrease its financial risk. However, accurate predictions are not easily obtained. The complex relationship between movie-related data and movie box-office revenues, plus the increasing volume of data in online movie databases, pose challenges for their effective analysis. In this paper, a multimodal deep neural network, incorporating input about movie poster features learned in a data-driven fashion, is proposed for movie box-office revenues prediction. A convolutional neural network (CNN) is built to extract features from movie posters. By pre-training the CNN, features that are relevant to movie box-office revenues can be learned. To evaluate the performance of the proposed multimodal deep neural network, comparative studies with other prediction techniques were carried out on an Internet Movie Database dataset, and visualization of movie poster features was also performed. Experimental results demonstrate the superiority of the proposed multimodal deep neural network for movie box-office revenues prediction.

✉ Zhang Yi
zhangyi@scu.edu.cn

Yao Zhou
zy3381@gmail.com

Lei Zhang
leizhang@scu.edu.cn

[1] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, People's Republic of China

Keywords Movie box-office revenues · Deep neural networks · Convolutional neural networks

1 Introduction

Movie box-office revenues indicate the financial performance of movies. The ability to predict a movie's box-office revenues before its theatrical release can reduce the producers' financial risk. Often, to increase a movie's financial success, large advertising investments are made to promote it before its theatrical release. Reliable predictions can inform production decisions in earlier stages of the production process, as well as providing guidance for movie viewers. Although movie success has been considered an unpredictable problem [23], several studies have attempted to develop approaches for movie box-office revenues prediction [9, 23, 31]. Specifically, the construction of computational models to investigate the relationships among movie-related variables has achieved considerable success. Pre-release movie-related data such as genres, ratings, and participating actors, are already extensively exploited for such predictions. More recently, attention has been drawn to the possibilities for leveraging additional movie-related data to improve the prediction performance [23, 26, 31]. In particular, movie posters contain information that could affect movie box-office revenues. As an advertising medium, movie posters are intentionally designed to convey the content of movies and attract the attention of potential viewers. As a result, movie posters are often recognized by viewers long before the movie's theatrical release. Increasingly, movie posters are made publicly available on online movie databases (e.g., IMDB), so it is important to investigate the influences of movie posters on movie box-office revenues.

Published online: 01 August 2017

Movie posters have attracted research interest [15, 33] in the field of movie genres classification. Since the content of movie posters is highly related to their genres, the inclusion of features of movie posters with other movie-related data could provide more information for predicting movie box-office revenues. However, feature extraction in these methods is typically done by hand, which does not generalize well and relies on prior knowledge.

Recently, deep neural networks (DNNs) have achieved success in various tasks including image classification [18], speech recognition [10], and machine translation [24]. With plentiful, well-labeled data, DNNs are able to learn nonlinear mappings and representations in either a supervised or unsupervised manner [11, 14, 34]. Convolutional neural networks (CNNs) are a type of widely used DNN, consisting of convolutional layers and pooling layers. By incorporating local connections and shared weights, CNNs are able to learn hierarchical features in a data-driven fashion. Also, given the recent advances in computational capacity, CNNs are relatively efficient for processing large images [18]. Moreover, CNNs can be separately pre-trained with visual content, facilitating their use as modular components providing feature outputs to a larger model.

In this paper, we propose a multimodal DNN for movie box-office revenues prediction. In particular, a CNN is first constructed to learn features from movie posters. After being well trained, it is used as a feature extracting module in the multimodal DNN to process movie posters. The proposed multimodal DNN accepts movie poster features and other movie-related data as input and predicts the movie box-office revenues as output. This model is expected to extract movie poster content relevant to financial success so the associated multimodal data are expected to improve prediction performance. As a side benefit, the learned movie poster features can be analyzed with visualization techniques, which aids human understanding of the CNN features.

The rest of this paper is organized as follows. Section 2 briefly reviews related work, and Sect. 3 introduces the characteristics of CNNs. We describe the proposed multimodal DNN method for movie box-office prediction in Sect. 4. Section 5 describes comparative studies carried out using the IMDB dataset, including experimental results and analysis. This paper is concluded in Sect. 6.

2 Related works

Existing methods for movie box-office revenues prediction can be classified according to their choice of movie-related data and prediction techniques. For example, movie trailers were used as input data for a linear support vector machine (SVM) classifier to predict the opening-week box-office revenues [26]. In this study, audiovisual features of movie trailers were identified and extracted manually, and showed that information from movie trailers can improve prediction performance. In another study [21], the impact of movie trailers on box-office revenues was investigated using video view statistics collected from social media. Equation systems were used to analysis the interaction between movie trailer sharing and box-office revenues. In a third study [16], data from social network service were collected for movie box-office forecasting, a genetic algorithm was adopted for selecting predictive variables, and several nonlinear regression algorithms were employed for building forecasting models. All these methods utilized multimodal data and extracted representative features for movie box-office revenues prediction. However, the methods required careful manual feature extraction and are not suitable for exploring complex relationships in movie-related data. Moreover, these methods cannot easily incorporate data with unbalanced dimensions. Using such methods, high-dimensional data would typically contribute more to the prediction result than those of low-dimensional data, thus limiting the ability to exploit high-dimensional data for learning. Another widely used approach to movie box-office revenues prediction involves artificial neural networks (ANNs), which can approximate arbitrary nonlinear functions. In [9, 23, 31], common movie-related data were utilized as input to an ANN, which produced the movie box-office revenue as the prediction result. Results obtained using ANNs were better than those from other classification techniques. Although these studies obtained fairly good prediction performance, they mostly limited their investigation to common movie-related data for movie box-office prediction. The impact of movie posters on movie box-office revenues was not considered.

The performance of traditional prediction methods are typically limited in terms of the ability to process raw data and representation learning [2]. Generally, using hand-craft features to represent data requires domain knowledge and a careful engineering designing. Recently, deep learning [19] is getting more common to automatically learn features from raw data by stacking multiple nonlinear modules and achieves superb performance on various tasks. In particular, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have attracted research interests from many domains. RNNs are effective approaches for sequential data modeling by maintaining its hidden states for a long-term history [27], and CNNs can learn abstract and hierarchical features by exploiting the local structure of data. In [30], a character-level sequence-to-sequence learning method is presented for video subtitles understanding, achieves remarkable performance on English-to-Chinese translation. On the other hand, a CNN using face features is designed to recognize human genders within a

video stream [28] and shows its effectiveness in video advertising. In addition, an end-to-end CNN is trained on clinical images directly for automated skin lesions classification and achieves comparable performance to dermatologists [8]. Although there are a large number of successful DNNs based methods in different fields, adopting DNNs for movie box-office revenues prediction still remains to be explored.

3 Preliminaries

In this section, we briefly introduce convolutional neural networks (CNNs). A CNN is a type of DNN that can be used to learn features from images [20]. CNNs typically consist of convolutional layers, pooling layers, and fully connected layers. They have been used with significant success in numerous computer vision problems. A simple CNN with one convolutional layer, one pooling layer, and three fully connected layers is shown in Fig. 1. The weights of the convolutional layers, called kernels, are combined over the input units to produce the feature maps. Pooling layers shrink the feature maps, relieving the computational burden and introducing the property of transformation invariance. Fully connected layers map the activities of the last convolutional layer to an output vector that can be used for categorization. By stacking convolutional layers, CNNs are capable of learning features with hierarchical structures. The lower layers tend to extract structural features such as corners or edges, while the higher layers focus on more semantic features such as object parts.

Specifically, given a CNN with L layers, let l ($1 \leq l \leq L$) denote the index of layers. The weight and bias are represented as w, b, respectively. Let P and Q denote the size of kernels. $f(\cdot)$ is a nonlinear activation function, such as a *sigmoid* or *rectifier* function [12]. Let a denote the activation of a feature map in the lth layer at position (c, r); then its value is computed as follows:

$$a^l_{c,r} = f\left(b^l + \sum_{i=1}^{P}\sum_{j=1}^{Q} w^l_{i,j} \cdot a^{l-1}_{c+i,r+j}\right). \quad (1)$$

There are two main advantages to leveraging CNNs to process movie posters. First, features are learned automatically instead of through manual designation, which requires no prior knowledge of the images. Second, the learned features can be analyzed with visualizing techniques, which can be useful to understand what kinds of movie poster features might influence movies' financial successes.

4 Methodology

The aim of this research was to predict movie box-office revenues given movie-related multimodal data. To achieve this, a multimodal deep neural network for movie box-office revenues prediction is built. A CNN is first constructed as a feature extractor for movie posters, then the movie poster content is combined with other selected movie-related data as input, which is expected to improve the performance of movie box-office prediction. As our main focus is to investigate the impact of movie posters and other movie-related multimodal data on box-office revenues prediction, data selection is discussed first.

4.1 Variables selection

Previous studies have extensively investigated the impact of different variables on movie box-office revenues, yielding advice for choosing movie-related data [9, 23, 31]. The current study copies prior studies in deploying classical factors such as genres, duration, star value and budget.

The movie genre provides important information about a movie, and genres are common as explanatory variables in box-office prediction [1]. Our study included 22 genres: Action, Adventure, Animation, Biography, Comedy,

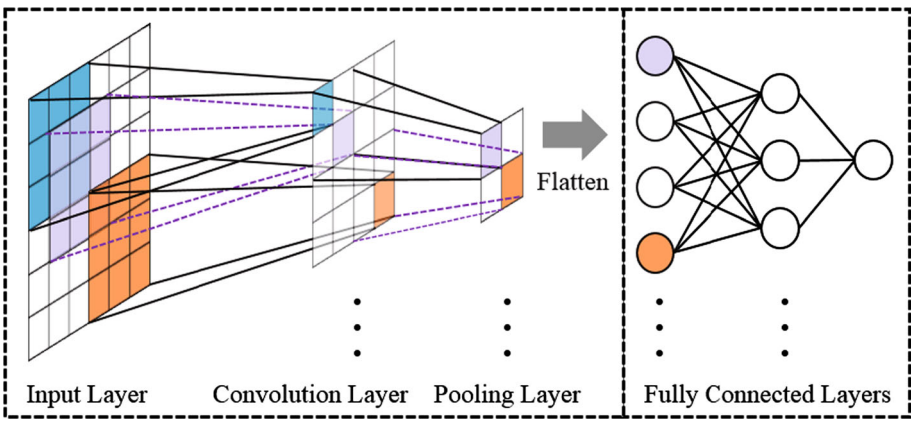

Fig. 1 A simple CNN with one convolutional layer using 3×3 kernel, one pooling layer using 2×2 kernel, and three fully connected layers. Flatten is introduced to reshape the units from matrices to vectors

Crime, Documentary, Drama, Family, Fantasy, Film-Noir, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War, and Western. Since a movie could belong to more than one genre, the genre variable was represented as a binary vector of length 22, which was initialized to all zeros. For each genre a movie belonged to, the corresponding position in the vector was set to one.

Another variable that can be defined objectively is the duration or running time of a movie. These data can be easily collected from online databases. Statistical analysis of movie data show that running time is, on average, around 100 min. Movie duration was represented as a numerical variable.

The production budget is also considered an important determinant of box-office revenues [31]. This information, representing a film company's investment in a movie, is not typically published. However, it can be estimated from budget information mentioned in advertisements and social media prior to the movie's theatrical release. Production budget was represented as a numerical variable.

Movie participants can also affect a movie's box-office revenues. It has been demonstrated that a star's value can affect the success of movies [7]. A superstar actor/actress can be defined as one who is well-known and well regarded for his/her work and therefore is likely to contribute significantly to the financial success of a movie. A famous director is also associated with high-quality movies. To represent the value of star participants, information about performers and directors was extracted from online movie sites. Only the lead director and the top three performers in the cast were considered. Specifically, the amount of "likes" on Facebook was used to represent star value, which was a numerical variable.

In addition to movie characteristics, social commentary such as professional movie critiques and viewer reviews can also have an important effect on movie box-office revenues [3]. In particular, integrating user comments with other text metadata for content-based movie recommendation can considerably enhance performance of movie recommender systems [29]. Besides, sentiment analysis of movie critiques could be a promising way for incorporating comments to improve box-office revenues prediction performance. For the sake of efficiency, we simply use the amount of movie reviews as representation of social commentary. It ranges from 1 to 800 and has an average value of 140. We employ it as a numerical variable, feed it into a DNN after normalization.

The number of consumers rating movies and the scores they assign are significant indicators of viewers' attitudes and attention to movies. User rating scores and box-office earnings have been found to be highly correlated [22]. Such information is plentiful and can be obtained from online movie sites. In this study, voting users was denoted as the number of users who voted for a movie, and voted score was the average movie score rated by users, which ranged from 0 to 10.

Movie posters contain the primary promotional content for movies, designed to attract the interest and curiosity of viewers. As such, movie poster content potentially contains information useful for box-office prediction [26]. Movie posters attract the interest and curiosity of viewers, which is informative but more difficult to exploit than other numerical or vector data.

Movie box-office revenues are naturally represented as numerical data. To represent each movie as a sample in certain class, we converted them to a discretized form grouped according to different ranges of box-office revenues.

4.2 Multimodal DNN prediction model

The problem of movie box-office revenues prediction is framed as classifying a movie into a category that indicates its level of financial success. A multimodal deep neural network was built to address this classification problem. Specifically, a multimodal input layer was adopted to process data in different modalities. Figure 2 illustrates the architecture of multimodal DNN. It consists of three parts, i.e., a multimodal input layer, hidden layers, and an output layer. In contrast to traditional multilayer neural networks, multimodal deep neural networks can accept multimodal data as input and learn features from that input in a data-driven fashion.

The input units of the multimodal DNN model are determined by the number of data modalities employed, and the output units are decided by the number of levels defined for movie box-office revenues. Formally, assuming the model consists of L layers, and $f(\cdot)$ represents the activation function, the output for each layer is computed as follows:

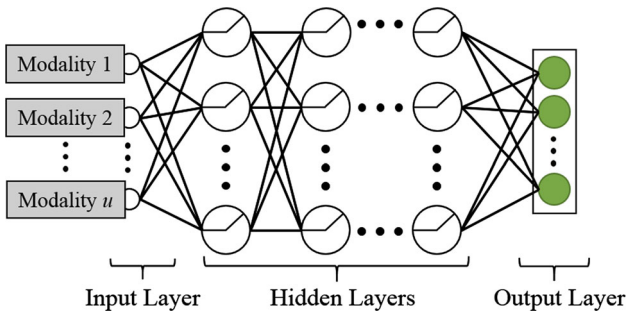

Fig. 2 The architecture of a multimodal DNN model. It accepts multimodal data as input from the *left side* and outputs the success level of movie box-office revenues on the *right side*

$$a_i^{l+1}(\theta,x) = f\left(\sum_{j=1}^{n_l} w_{ij}^{l+1} \cdot a_j^l(\theta,x)\right), \qquad 0 \le l \le L$$

where $a_j^l(\theta,x)$ represents the activation of unit i in lth layer for given sample x and parameters θ, $a_j^l(\theta,x) = x_j$ when $l = 1$, and x_j is the jth input variable. w_{ij}^{l+1} denotes the weight connection between the jth unit in lth layer and the ith unit in $(l+1)$th layer, and n_l is the number of units in lth layer. The hidden layers are used similarly to those of multilayer neural networks.

In the output layer, the activations of output units are sent to a softmax classifier. Therefore, the prediction result can be interpreted as probabilities for each category. The calculation of the output layer is formulated as follows:

$$\Pr_i(\theta,x) = \frac{e^{a_i^L(\theta,x)}}{\sum_{j=1}^{k} e^{a_j^L(\theta,x)}}$$

where L is the number of units in the output layer and k represents the total number of classes. $Pr_i(\theta,x)$ represents the probability of a given input sample x be classified to class i with parameters θ.

To train this network, discretized movie box-office revenues are adopted. For each input data modality, the training can be conducted separately, and then the pre-trained model can be incorporated to utilize features from multimodal data.

4.3 Prediction with movie poster features

To exploit movie poster content for movie box-office revenues prediction, a CNN was constructed to extract representative features. The CNN was first pre-trained with movie posters as its input and movie box-office revenues as its output. To make the training of the CNN more efficient and less vulnerable to overfitting, only one fully connected layer was adopted. That is, feature maps of the last convolutional layer were processed with a global average pooling layer [32], and then mapped to prediction results directly. Subsequently, the CNN was incorporated into the multimodal DNN. To regularize the training of the multimodal DNN, the original numerical and discretized form of movie box-office revenues were both adopted as output, and parameters of the multimodal DNN were updated according to the cost functions of these two outputs. The architecture of the multimodal DNN prediction model is shown in Fig. 3.

In the output layer of the multimodal DNN, softmax activation was adopted for the classification task to predict the financial success level. To evaluate how well the model fits the training data, a cross-entropy function was employed. Therefore, parameters of the multimodal DNN model were updated by minimizing the cost function as follows:

$$\arg\min_{\theta} -\frac{1}{N}\left[\sum_{n=1}^{N}\sum_{k=1}^{K}\left(c_k^n \cdot \log \Pr_k(\theta,x^n)\right)\right] \tag{2}$$

where N is the number of samples and K is the number of total categories. $\Pr_k(\theta,x^n)$ denotes the probability of nth sample to class k, and θ is set of parameter of the multimodal DNN. c_k^n is the true class of that sample.

The regression in the multimodal DNN is a subsidiary task, which acts as a training regularizer. Mean square error is utilized as the cost function, and the optimization is achieved as follows:

$$\arg\min_{\theta} \frac{1}{2N}\left[\sum_{n=1}^{N}\left(\|y^n - a^L(\theta,x^n)\|^2\right)\right] \tag{3}$$

where θ represents parameters of the multimodal DNN. $a^L(\theta,x^n)$ denotes the activation of nth sample in the output layer, and y^n is the original box-office revenues of sample n.

The main advantage of this model is that it utilizes features from multimodal data, which are more informative and thus can improve the prediction performance. Moreover, data with unbalanced dimensionality can be incorporated efficiently.

5 Experiments

This section describes experiments carried out on the *IMDB* dataset. These included comparisons between the multimodal DNN prediction model and baseline methods, evaluation of the impacts of different activation functions, and visualization of learned features.

5.1 Dataset

The experimental dataset was obtained from the Internet Movie Database website (imdb.com), while box-office revenues for each movie were obtained from the-numbers.com. A set of 3807 movie samples were collected. The sample movies were categorized into six groups according to their box-office revenues, with the average amount for each group, sorted by box-office revenues. This was reasonable because we obtained the samples from a ranked list. All the selected numerical data in each sample item were normalized as follows:

$$\phi(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

where $\phi(\cdot)$ denotes the normalization function and x_i denotes the ith data. After this, the data for each sample

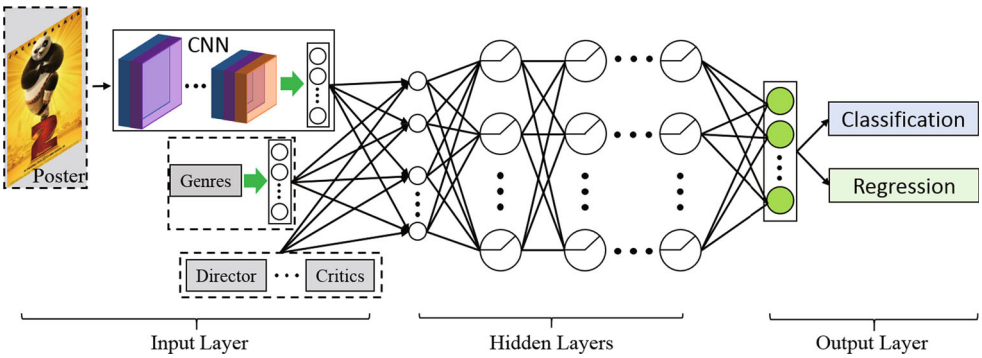

Fig. 3 The architecture of a multimodal DNN prediction model. It accepts movie posters and other movie-related data as multimodal data input. At the output layer, training signals are generated from both classification and regression loss functions

were converted to the same scale, a step that can accelerate the convergence of DNNs.

5.2 Evaluation metrics

To measure the performance of the proposed method, average percent hit rate (APHR) [1] was deployed as the evaluation metric. It indicated the percentage of correct classifications relative to the total number of samples. Two types of APHR were employed to judge the accuracy with respect to different classes:

- First, the absolute accuracy. This metric indicated the exact (Bingo) hit rate, meaning that only classifications to the correct class would be considered.
- Second, the relative accuracy. This metric is also based on Bingo, but included consideration of classification results in adjacent classes (1-Away).

Algebraically, the APHR can be calculated as follows:

$$\text{APHR} = \frac{\text{Number of samples correctly classified}}{\text{Total number of samples}} \tag{4}$$

$$\text{APHR}_{\text{Bingo}} = \frac{1}{n}\sum_{i=1}^{K} c_i \tag{5}$$

$$\text{APHR}_{1-\text{away}} = \text{APHR}_{\text{Bingo}} + \frac{1}{n}\sum_{i=1}^{K} c_{i-1} + c_{i+1} \tag{6}$$

where n represents the total number of samples, K is the total number of classes, and c_i denotes the total number of samples correctly classified as class i. When $i \leq 1$ or $i \geq K$, $c_i = 0$.

5.3 Experiments settings

In the multimodal DNN model, the hidden layers were composed of five fully connected layers. Rectified linear unit (*ReLU*) was adopted as the activation function of each hidden layer for nonlinear transformation, *dropout* was added after every fully connected layer to reduce the co-adaption of hidden units, and *softmax* was applied at the last layer to produce categorical results. The architecture of the multimodal DNN model for movie box-office prediction is shown in Table 1.

To incorporate information from movie posters in the multimodal deep neural network, a CNN was built. The CNN consisted of 10 convolutional layers and 4 pooling layers, *ReLU* was adopted as the activation function, and global average pooling was employed following the last convolutional layer. In the pre-training step, two units were employed in the last layer to predict whether a movie achieves financial success or failure. Details of the CNN architecture is shown in Fig. 4. The number of movie poster dataset may insufficient for fully training the CNN, and result in overfitting which hampers the performance of CNNs. To alleviate this issue, the CNN is firstly pre-trained on ImageNet dataset [6]. Although it mostly contains natural images, the objects in movie posters are quite similar to them and the high level features of CNN can be relevant to our movie posters dataset.

A k-fold cross-validation scheme was employed for evaluating the multimodal DNN. The ratio of validation data was 20% in the experiments, and k was set to 5.

Table 1 Architecture of the multimodal DNN model for movie box-office prediction

Layer	Type	Units
1	Multimodal input	–
2	FC + relu	72
3	Dropout (0.6)	–
4	FC + relu	128
5	Dropout (0.6)	–
6	FC + relu	256
7	Dropout (0.6)	–
8	FC + relu	128
9	Dropout (0.6)	–
10	FC + relu	72
11	Dropout (0.6)	–
12	Softmax	6

FC represents fully connected layers

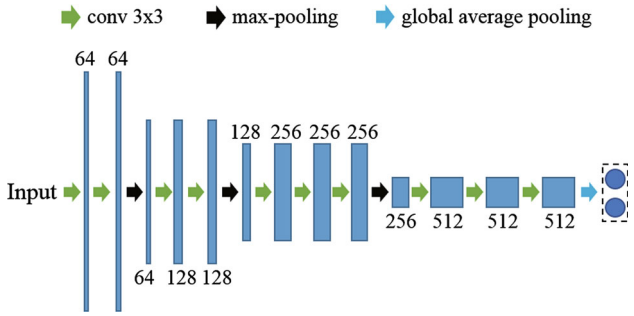

Fig. 4 The CNN architecture for movie posters classification. Numbers at the *top* or *bottom* of each layer represent the amount of convolutional kernels

Parameter optimization was achieved using the Adam [17] algorithm, and was used for minimizing both cost functions for classification and regression with learning rates of 5e−3 and 5e−5, respectively. The experiments were implemented within the MXNet [4] framework, running on a computer with a GPU of NVIDIA GeForce GTX TITAN X.

5.4 Results and analysis

To evaluate the performance of the multimodal DNN for movie box-office prediction, comparison to competitive methods including *SVM* [5], Random Forest (*RF*) [25], and state-of-the-art multilayer BP neural network (*MLBP*) [31] was employed. The network structure we used in experiments is exactly as same as the MLBP described in [31] except slight difference of number of input variables. The parameters of each method are carefully tuned.

To investigate the impact of movie posters on movie box-office revenues, the first comparison excluded movie poster content. To balance the dimensionality of numerical and vector data as input for *SVM* and Random Forest, principal component analysis (*PCA*) was adopted. In the multimodal prediction model, the dimension reduction was achieved by fully connected layers. In particular, data of dimension larger than one (e.g., movie genres) was reduced to 1. The performance of these methods is shown in Table 2.

The results in Table 2 show that, for both 'bingo' and '1-away' metrics, the multimodal DNN (highlighted in bold) based metadata prediction model performed significantly better than the other methods on each fold and on average, while *MLBP* technique performed slightly better than *RF*, and *SVM* achieves lower performance than others. The good performance of DNNs mainly attributed to network depth and *dropout* technique. Increasing network depth has been demonstrated to be a efficient way to improve the classification performance [13], while *dropout* provide an effective way to avoid overfitting problem when training a network with large scale parameters and achieves a better generalization. Additional experiments, deploying the original data without dimension reduction to these classifiers directly, led to a worse performance. This poor performance is mainly due to the difficulty of learning an effective representation of sparse binary vectors, which typically represented movie-related data such as genres.

To further investigate the influence of the activation function on the multimodal DNN, experiments using different activation functions such as *sigmoid*, *tanh*, and *ReLU* were conducted. The results are presented in Fig. 5.

The results show that *ReLU* performs slightly better than *sigmoid* and *tanh*. This improvement may be attributed to the local linearity of the *ReLU* activation function, which alleviates gradient vanishing issues. It also implies that building DNNs with *ReLU* would be a better choice for movie box-office revenues prediction.

To investigate the impact of movie poster content, a comparison study of different methods including the extra movie poster features was conducted. The dimensions of the movie posters features were determined by the output units of the CNN, which was set to 2 in our experiment. All other movie-related data were used exactly as in the first experiment. To keep the dimensionality balance for each data, *PCA* was again deployed to reduce the dimension to *1* in the *RF* and *SVM* cases. In the *MLPBP* and multimodal DNN (highlighted in bold) cases, this was achieved by mapping of fully connected layers. Results are shown in Table 3.

Comparison of the results shown in Tables 2 and 3 indicates that incorporating movie poster features learned by CNN as input into the multimodal DNN can improve the model's prediction performance. Moreover, the low-dimensional representation of movie posters extracted by CNNs can boost the performance of the *SVM*, *RF*, and *MLPBP* prediction methods as well.

To intuitively understand predictive ability of different variables on DNN output, sensitivity analysis with different combination of variables is employed. In particular, we mainly focus on the input variables including score, rating, comments, participants, budget, duration, genres. Experiments are processed with a specified variable disabled while the rests available. In addition, the impacts of whether movie poster features is used is investigated as well. Hence, the DNNs accept a subset of variables, and the performance degradation caused by a disabled variable would reflect its contribution.

Sensitivity analysis results shown in Fig. 6 indicate that rating and budget affects the performance more significant than others, due to the accuracy degrades obviously when one of them is disabled. Comparing with average performance of DNNs in Tables 2 and 3, prediction using movie poster features achieves better results than not using it, even though one of the variables is disabled.

Table 2 Comparison of different methods versus the proposed multimodal DNN method for movie box-office prediction

Classifier	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Average (%)
SVM						
Bingo	32.01	32.54	33.47	32.27	35.45	33.15
1-Away	62.17	62.43	63.23	60.84	64.68	62.67
RF						
Bingo	47.35	45.32	46.14	46.95	46.16	46.38
1-Away	84.58	80.21	81.93	82.80	81.61	82.23
MLBP						
Bingo	48.11	47.78	50.56	48.67	49.33	48.89
1-Away	86.11	83.67	82.67	84.34	84.70	84.30
DNN						
Bingo	**52.44**	**49.12**	**53.89**	**50.13**	**51.67**	**51.45**
1-Away	**87.45**	**87.33**	**85.44**	**86.22**	**85.78**	**86.44**

Movie-related data excludes movie poster content. Numbers in first row represent different folds

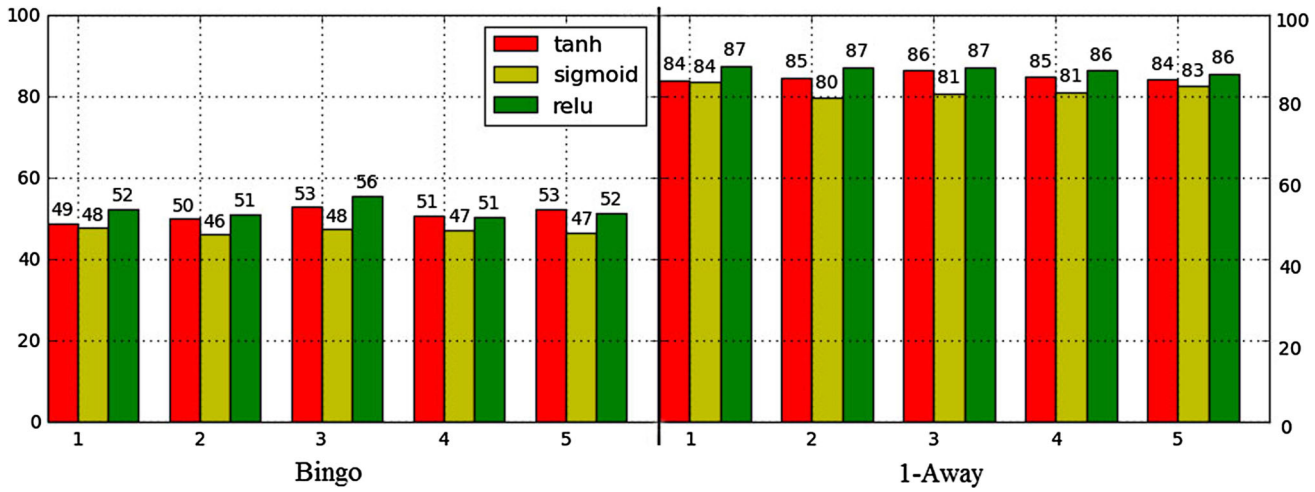

Fig. 5 Performance of the multimodal DNN with different activation functions for movie box-office prediction (Numbers in the *bottom* represent different folds)

Table 3 Comparison of different methods and the proposed multimodal DNN method for movie box-office prediction

Classifier	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Average (%)
SVM						
Bingo	33.07	35.19	34.39	33.33	35.85	34.37
1-Away	63.62	65.74	65.61	64.02	65.21	64.84
RF						
Bingo	48.28	45.76	47.35	49.21	48.28	47.78
1-Away	85.19	81.09	82.01	83.86	82.14	82.86
MLBP						
Bingo	49.38	48.56	51.78	49.45	51.14	50.06
1-Away	88.05	85.13	84.89	85.22	86.36	8 5.93
DNN						
Bingo	**53.22**	**50.00**	**54.56**	**50.44**	**52.78**	**52.20**
1-Away	**90.33**	**88.11**	**87.00**	**88.78**	**88.78**	**88.60**

Movie-related data includes movie poster content. Numbers in first row represent different folds

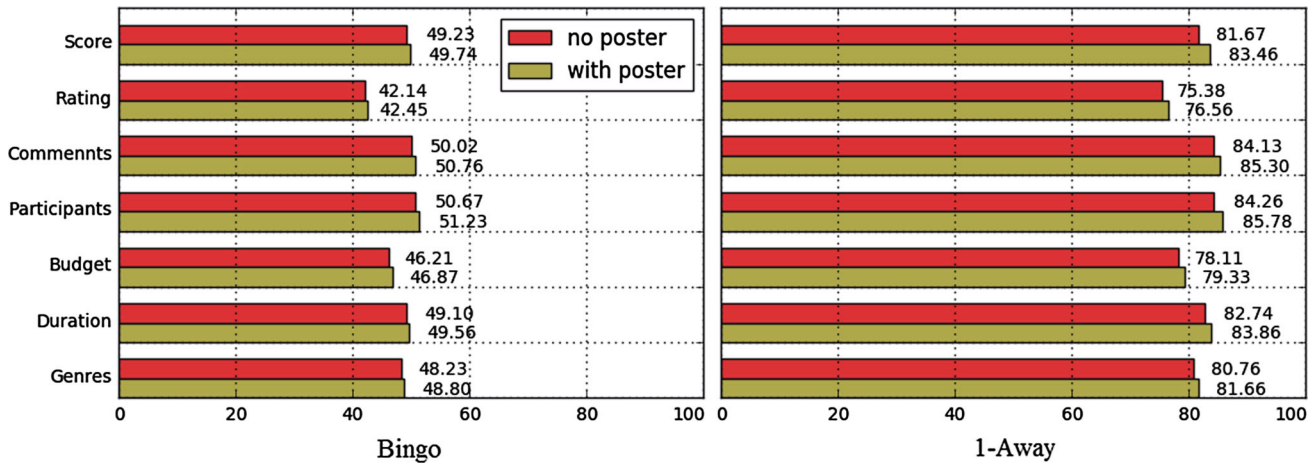

Fig. 6 Sensitivity analysis results for different combination of variables (Variable name on the *left* represent which one is disabled, comparisons with no movie poster features and with movie poster features are shown for each variable)

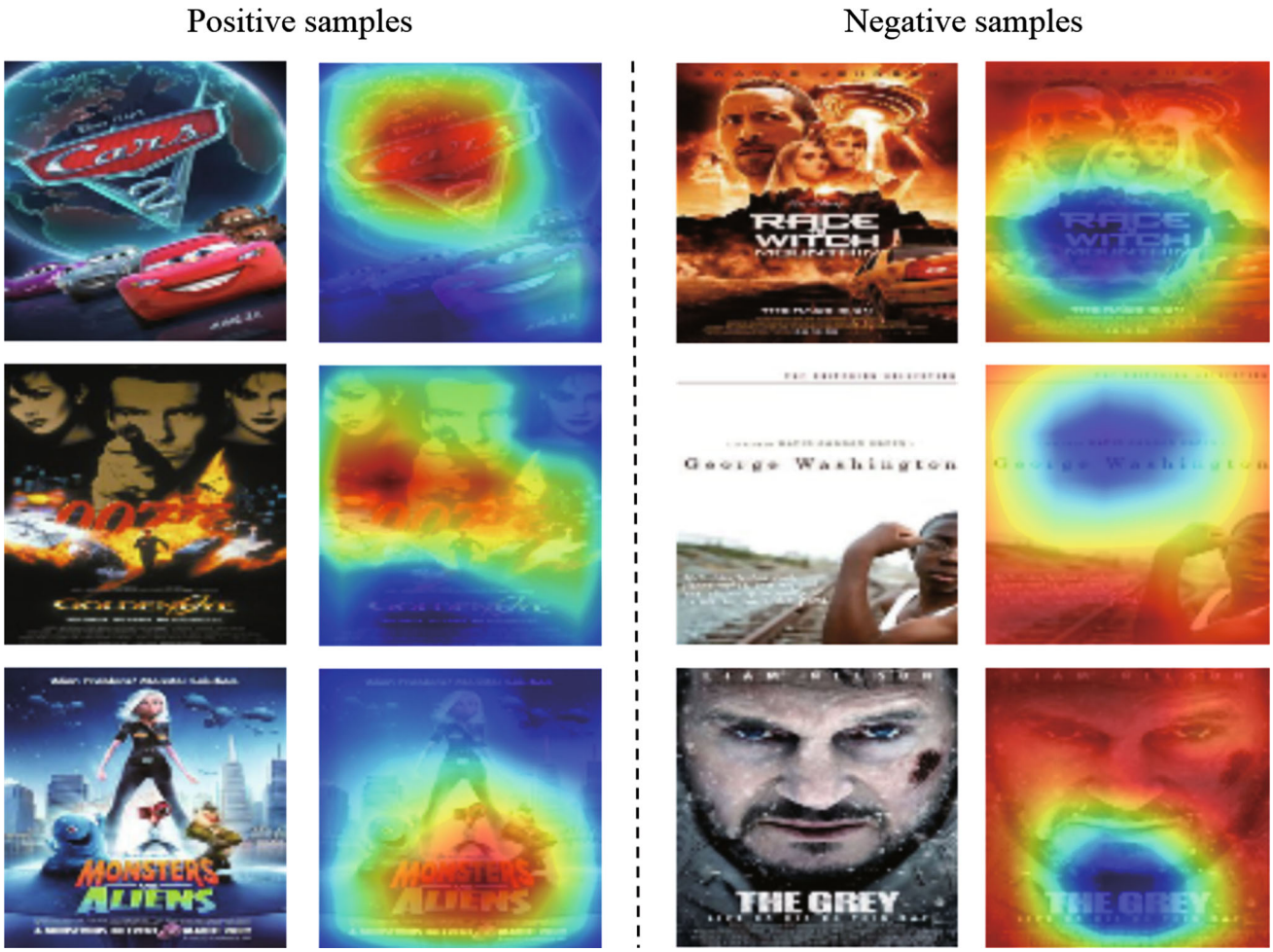

Fig. 7 Visualization of learned features in CNNs by highlighting discriminative regions. For both positive and negative samples, the text part are highlighted with *red* region and *blue* region, respectively, while the former contributes to a positive classification result, and the later to negative

5.5 Feature visualization

To visualize the learned features, the CNN in the multimodal DNN prediction model was first pre-trained with a subset of the entire dataset. It consisted of the top 2 groups and the bottom 2 groups in the IMDB dataset according to box-office revenue. The top and bottom sets were used as positive samples and negative samples, respectively. After the training of the CNN converged, it achieved an accuracy of 63.15% on the validation set which was 20% of the subset. This outcome indicated that the CNN had learned the features of movie posters relevant to movie box-office revenues.

To visualize the features of movie posters that the CNN was looking for, samples of movie posters in the validation set were used as input to the well trained CNNs, and we adopted deep feature localization [32] to identify the discriminative regions. Samples with corresponding discriminative regions are shown in Fig. 7.

The results show that the CNN mainly focused on text regions for discrimination. Positive samples are posters of financially successful movies; the texts in the poster are more colorful than those of negative samples. The visualized features in the CNN also indicate that there is an implicit relationship between the design of text on movie posters and box-office revenues. Moreover, this method can be generalized to other culture product analysis, such as novels and dramas.

6 Conclusion

In this paper, a multimodal deep neural network for movie box-office revenues prediction was proposed. First, a CNN was built for extracting features from movie posters. Then, a multimodal deep neural network was built to leverage both movie poster features and other movie-related data for movie box-office revenues prediction. In addition, the features of CNN learned from movie posters were analyzed. In contrast to neural network-based methods, the multimodal DNN exploited movie poster content rather than only using traditional data for movie box-office prediction, taking advantage of multimodal data of movies. Experiments on an *IMDB* dataset demonstrated that the proposed method achieves better performance than other approaches. In addition, the ability to visualize the learned features of movie posters offers a guide for the design of future movie posters. Future investigations will focus on building additional multimodal DNNs to incorporate movie-related audio and video data.

Acknowledgements This work was supported by the National Science Foundation of China (Grant Numbers 61432012, U1435213).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ainslie A, Drèze X, Zufryden F (2005) Modeling movie life cycles and market share. Mark Sci 24(3):508–517
2. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
3. Brewer SM, Kelley JM, Jozefowicz JJ (2009) A blueprint for success in the us film industry. Appl Econ 41(5):589–606
4. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z (2015) Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274
5. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
6. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE, pp 248–255
7. Elberse A (2007) The power of stars: do star actors drive the success of movies? J Mark 71(4):102–120
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118
9. Ghiassi M, Lio D, Moon B (2015) Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Syst Appl 42(6):3176–3193
10. Graves A, Mohamed AR, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp 6645–6649
11. Guo Q, Jia J, Shen G, Zhang L, Cai L, Yi Z (2016) Learning robust uniform features for cross-media social data by using cross autoencoders. Knowl Based Syst 102:64–75
12. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1026–1034
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
14. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
15. Ivasic-Kos M, Pobar M, Mikec L (2014) Movie posters classification into genres based on low-level features. In: 37th international convention on information and communication technology, electronics and microelectronics (MIPRO), 2014. IEEE, pp 1198–1203
16. Kim T, Hong J, Kang P (2015) Box office forecasting using machine learning algorithms based on SNS data. Int J Forecast 31(2):364–390
17. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
18. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the advances in neural information processing systems (NIPS), pp 1097–1105
19. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

20. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
21. Oh S, Ahn J, Baek H (2015) Viewer engagement in movie trailers and box office revenue. In: 48th hawaii international conference on system sciences (HICSS), 2015. IEEE, pp 1724–1732
22. Qin L (2011) Word-of-blog for movies: a predictor and an outcome of box office revenue? J Electron Commer Res 12(3):187
23. Sharda R, Delen D (2006) Predicting box-office success of motion pictures with neural networks. Expert Syst Appl 30(2):243–254
24. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of the advances in neural information processing systems (NIPS), pp 3104–3112
25. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43(6):1947–1958
26. Tadimari A, Kumar N, Guha T, Narayanan SS (2016) Opening big in box office? Trailer content can help. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2777–2781
27. Wang J, Zhang L, Guo Q, Yi Z (2017) Recurrent neural networks with auxiliary memory units. IEEE Trans Neural Netw Learn Syst. doi:10.1109/TNNLS.2017.2677968
28. Zhang H, Cao X, Ho JK, Chow TW (2016) Object-level video advertising: an optimization framework. IEEE Trans Ind Inf 13(2):520–531
29. Zhang H, Ji Y, Li J, Ye Y (2016) A triple wing harmonium model for movie recommendation. IEEE Trans Ind Inf 12(1):231–239
30. Zhang H, Li J, Ji Y, Yue H (2016) Understanding subtitles by character-level sequence-to-sequence learning. IEEE Trans Ind Inf 13(2):616–624
31. Zhang L, Luo J, Yang S (2009) Forecasting box office revenue of movies with BP neural network. Expert Syst Appl 36(3):6580–6587
32. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2921–2929
33. Zhou H, Hermans T, Karandikar AV, Rehg JM (2010) Movie genre classification via scene categorization. In: Proceedings of the 18th ACM international conference on multimedia. ACM, pp 747–750
34. Zhou JT, Pan SJ, Tsang IW, Yan Y (2014) Hybrid heterogeneous transfer learning through deep learning. In: AAAI, pp 2213–2220