Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/47/20241383

Movie genre prediction based on the bidirectional encoder representations from transformer

Lu Li[1], Jin Lin[2], Tongyu Li[2,3,4]

[1]Department of Geophysics, Tongji University, No.1239, Siping Road, Yangpu District, Shanghai Province, China
[2]Department of Electronic & Electrical Engineering, The University of Sheffield, Sir Frederick Mappin Building, Sheffield, S1 3JD, England

[3]tli79@sheffield.ac.uk
[4]corresponding author

Abstract. The rapid expansion of digital media has underscored the growing significance of predicting genres to target audiences effectively and to enhance filmmakers' understanding of viewer preferences. In this study, we introduced a novel method for forecasting movie genres, leveraging the power of the Bidirectional Encoder Representations from Transformer (BERT) deep learning model. The research team employed a dataset sourced from Douban's website, which featured 5,000 movies, complete with cover images, titles, and genre information. This undertaking tackled several key challenges, including the extraction of features from textual data, the categorization of movie genres, and the incorporation of cultural nuances. BERT's bidirectional representation, especially its variant tailored for Chinese language tasks, 'bert-base-chinese', was employed to extract textual features. Meanwhile, the visual features from movie covers were processed using the Wide ResNet-50-2 architecture. The combined features underwent classification, and the resulting model achieved an accuracy of 34.67%, a recall rate of 74.5%, and an F1 score of 0.4755. The study validates the potential of the BERT model in predicting movie genres and offers significant insights for future research in multimedia content classification.

Keywords: Machine Learning, Deep Learning, Natural Language Processing.

1. Introduction

In the era of burgeoning digital media and the proliferation of streaming platforms, the annual output of movies continues to surge in both quantity and variety. Consequently, it has become progressively more daunting for audiences to make informed choices that resonate with their preferences. Traditionally, viewers have leaned on movie covers and titles as their initial means of filtering and identifying genres. However, this method can be constrained by individual experiences and cultural backgrounds, leading to potential biases and errors in classification. It is widely believed that movie genre prediction is both intriguing and of practical importance. It can provide invaluable insights to filmmakers, assisting them in better understanding viewer preferences. Furthermore, movie genre prediction can enhance film recommendation services for audiences.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/47/20241383

Several researchers have ventured into this domain, achieving notable progress and results. For instance, Makita et al. have applied the Bernoulli event model to estimate the likelihood of a given genre for a movie [1]. Using the Bayesian rule, they evaluated the posterior probability of a film's genre. This method utilised movie ratings as feature vectors and achieved a 59% precision prediction rate and a 69% rate of including correlated genres on the MovieLens 100k dataset. Moreover, Zhang et al. have proposed a multi-modal approach using shot information for movie genre classification [2]. This method incorporated visual modalities and audio and linguistic ones to enhance the recognition of higher semantic genres such as documentaries or musicals. This approach outperformed existing methods by 17%~21% regarding mean Average Precision on datasets like MovieNet and Condensed Movies. The technique also demonstrated superior performance in scene boundary detection tasks, surpassing current methods by an average of 1.1% in AP. Other researchers like Kundalia et al. introduced a novel technique of using neural networks for knowledge transfer learning to generate multi-label movie genre predictions from posters [3].

There are many challenges in completing this task, including:

1) Feature Extraction from Movie Text Information: Merging the heterogeneous data from movie covers (visual) and titles (textual) poses a significant challenge. Our projected dataset comprises publicly available data and information scraped from movie websites. Effective feature extraction and representation from textual movie information in the dataset require techniques like text preprocessing, tokenisation, and semantic understanding. Ensuring the feasibility and effectiveness of the experiments demands the right choice of dataset, model, and optimised model parameters.

2) Definition and Segregation of Movie Genres: The definition and categorisation of movie genres may vary between individuals or organisations. For instance, a movie can belong to multiple broad genres like action, sci-fi, and adventure or specific sub-genres such as martial arts, dark comedy, disaster, etc. While more general categories are easier to label and identify, they might overlook a movie's complexity. On the other hand, sub-genres highlight a movie's distinctiveness but can be harder to differentiate. Clarifying which genres to predict, handling multi-label issues, and balancing granularity are significant challenges. We aim to balance broad and sub-genres or employ multi-label methods to represent movies belonging to multiple genres.

3) Challenges Due to Background and Cultural Differences in Movies: Movies can reflect specific eras, regions, and societal or political influences, resulting in unique styles or themes. Or they might target varied audiences, using distinct narrative techniques or linguistic styles. During experimentation, these factors can lead to classification inaccuracies. Hence, we plan to incorporate additional information and features into the dataset and model to capture this background knowledge and cultural differences.

To achieve more accurate movie genre predictions, we intend to utilize deep learning model in this case due to their excellent performance in many tasks [4-6]. To be more specific, the Bidirectional Encoder Representations from Transformers (BERT) deep learning model was employed to extract features from movie covers and titles. These features are then fed into classifiers for categorization. Many researchers have already used deep learning models for feature extraction from movie covers and labels with promising results. Through this study, we aim to explore the feasibility of movie genre prediction based on the BERT model and provide valuable references for related studies.

2. Method

2.1. Dataset Preparation

The dataset used in this study was sourced from the official Douban website, acquired using web scraping tools. The dataset encompasses 5,000 data entries, and is characterized by three distinct features: the movie's cover image (image), the movie's title (name), and the movie's genre (type). Movie genres were categorized into 33 distinct classes including Comedy (0), Drama (1) and Animation (2) etc. Information regarding the movie's title and genre was compiled into a JSON format,

DOI: 10.54254/2755-2721/47/20241383

exemplified by: “name”: “The Martian”, “type”: “1/25/26”. Concurrently, data pertaining to the movie cover images was organized into a separate directory for ease of reference.

2.2. BERT

BERT is a pre-training model for natural language processing based on the Transformer architecture. It's capable of learning text features for classification predictions. BERT can pre-train vast amounts of unlabeled data and fine-tune for specific downstream tasks, achieving state-of-the-art results. Compared to traditional unidirectional language models or shallow concatenations of two unidirectional models for pre-training, BERT introduces a new Masked Language Model (MLM) to generate deep bidirectional language representations. Hence, using the BERT model for movie genre prediction offers significant advantages.

The BERT has emerged as a significant breakthrough in the realm of natural language processing. One of its variants explicitly tailored for Chinese language tasks is the "bert-base-chinese," accessible via the Hugging Face model [7]. Distinctly, the architecture of bert-base-chinese deploys a bidirectional Transformer, which stands out against traditional unidirectional or bidirectional recurrent neural networks. This approach captures bidirectional contextual information within texts seamlessly.

During its pre-training phase, BERT utilizes two principal tasks: MLM mentioned before and the Next Sentence Prediction (NSP). In the MLM task, BERT randomly masks some words in the input sentence with a [MASK] token to predict these obscured words. Simultaneously, the NSP task trains the model to ascertain if two provided sentences are consecutive.

BERT leverages the WordPiece tokenization strategy on the embedding front, enabling it to dissect vocabulary into smaller units, like subwords or characters. Particularly for bert-base-chinese, this facilitates the model's capability to process both Chinese characters and specific words. Furthermore, considering the inherent nature of the Transformer architecture that does not account for sequence order, BERT integrates Positional Embeddings, encoding each word's position within a sentence. Segment Embeddings are another feature that distinguishes two different input sentences.

Size-wise, the bert-base-chinese aligns closely with the foundational version of BERT, encompassing roughly 110M parameters. The model structure comprises 12 layers (i.e., Transformer encoder layers), 768 hidden units, and 12 self-attention heads. A notable aspect is that the bert-base-chinese model has been pre-trained on Chinese Wikipedia data, ensuring its adeptness at comprehending Chinese texts and generating pertinent contextual word embeddings.

Based on the source related to the bert-base-chinese model [8], the training process could be described as the Figure 1. For practical implementation in this study, it is easy to download the model package from the previously mentioned official website, place it in the project's directory, and then import it into the codebase.

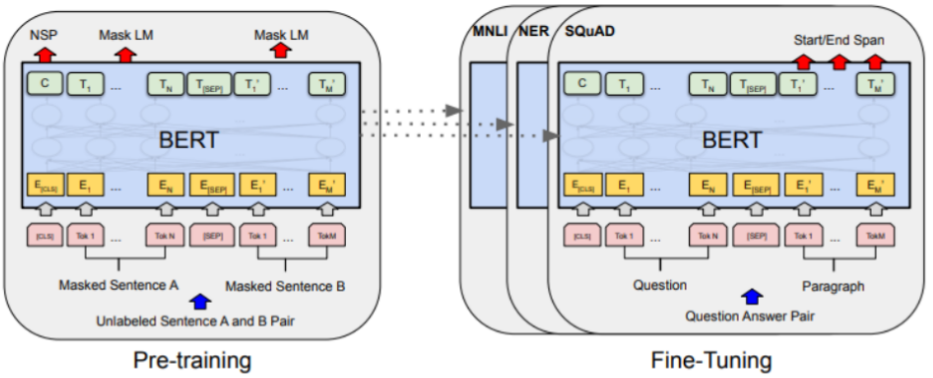

Figure 1. The Bert-base-chinese model framework [8].

DOI: 10.54254/2755-2721/47/20241383

2.3. Gpipe

Gpipe is a library for efficient training of large neural networks using pipeline parallelism. Pipeline parallelism is a technique that splits a neural network into several parts, called cells, and assigns each cell to a different device, such as a GPU or a TPU. The cells are then executed in a sequential order, passing the intermediate outputs from one cell to the next. This way, each device can work on a different part of the network at the same time, reducing the memory consumption and increasing the throughput.

Google Brain introduced GPipe in a paper which claimed that Gpipe can train models that was 25 times larger than the memory limit of a single device, and achieved up to 3.5 times speedup compared to data parallelism [9]. Gpipe also uses checkpointing to reduce the peak memory usage and gradient recomputation to improve the accuracy.

Gpipe is implemented in TensorFlow and can be used for any neural network that can be expressed as a sequence of layers. It supports both synchronous and asynchronous pipeline execution modes, and can handle various types of network architectures, such as convolutional, recurrent, and attention-based models. There are also some open-source implementations of Gpipe in other frameworks, such as PyTorch and JAX [10]. These implementations may have some differences from the original paper, but they share the same core idea of pipeline parallelism.

2.4. Implementation details

Our model was implemented using PyTorch and the Hugging Face's Transformers library. The BERT architecture with a base Chinese pre-trained model (bert-base-chinese) was used for text processing. The images, on the other hand, were processed using the Wide ResNet-50-2 architecture from torchvision. Data augmentation techniques applied to the images include resizing to 256x256 pixels, centre cropping to 224x224 pixels, tensor transformation, and normalisation with given mean and standard deviation values. The final model concatenated the outputs of the BERT and Wide ResNet-50-2 models, which were then fed into a classifier with a linear layer having an output size of 34.

We employed the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 1×10^{-4} and momentum set to 0.9 for the optimisation strategy. As for the loss function, we utilised the Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), suitable for multi-label classification. We implemented the GPipe approach to enhance computational efficiency, splitting the model over two devices and using a chunk size of 2 for pipeline parallelism.

The training process ran for ten epochs with a batch size of 32. During evaluation, the threshold for the output logits was set to 0, after which they were binarised to produce the model's predictions. Performance evaluation was based on the accuracy, recall and F1 score (average over samples) using the true labels and the model's predictions.

3. Results and discussion

The overall assessment results are shown in the Table 1 below.

Table 1. Overall evaluation indicator and values

Evaluation indicators	Value
Accuracy	34.67%
Recall	74.5%
F1 score	0.4755

3.1. Accuracy

In machine learning tasks, especially in classification problems, accuracy is a standard evaluation metric that measures the proportion of correct predictions made by the model. However, in multi-label classification tasks, the interpretation and meaning of accuracy are different. In multi-label classification, a sample is predicted correctly only if all its labels are correctly predicted. In other

DOI: 10.54254/2755-2721/47/20241383

words, if any label of a sample is mispredicted, whether an extra prediction or a missed one, the sample will be considered mispredicted.

As mentioned earlier, accuracy pertains to the percentage of samples in which all labels were predicted correctly. In the context of this multi-label classification, it was observed that 34.67% of samples had all their labels accurately predicted. However, this does not imply that more than 65% of the samples were completely mispredicted in a multi-label scenario. It might indicate that some labels within these samples needed to be accurately predicted. Given the complexity of multi-label tasks, this accuracy might be reasonable, but there is room for improvement.

3.2. Recall

Recall signifies the proportion of true labels that were predicted correctly. In this context, it indicates the percentage of samples that were actually of a specific label and were correctly identified as such. A recall of 74.5% indicates that the model does an excellent job of capturing the proper labels. However, a high recall suggests that the model is trying to predict as many labels as possible, which can increase false positives. This is why there is often a trade-off between recall and precision.

3.3. F1 score

The F1 score is the harmonic mean of precision and recall, trying to strike a balance between the two. An F1 score of 0.4755 indicates that the model has not found a perfect balance between precision and recall. This might imply that while the model is good at capturing actual labels (as indicated by recall), it might need some help to ensure that the labels it predicts are accurate, leading to false positives or negatives.

4. Conclusion

Within the dynamic landscape of digital media and streaming platforms, the anticipation of movie genres takes on a pivotal role for both filmmakers and audiences. This research endeavors to meet the rising demand for precise genre prediction by harnessing the cutting-edge BERT model, specifically adapted for extracting textual features. It also incorporates visual elements from movie covers into the prediction process, recognizing the multifaceted nature of this task. The data extracted from the Douban website was robust, covering 5,000 movies and a diverse range of 33 genres. Using the BERT model, particularly the 'bert-base-chinese', proved advantageous in capturing the nuanced information in movie titles. At the same time, incorporating the Wide ResNet-50-2 for image processing ensured a holistic feature extraction mechanism.

Although the accuracy of the multi-label classification system was 34.67%, it is essential to interpret this in the context of multi-label tasks. The achieved accuracy indicates not the system's complete mispredictions but perhaps a few labels within multi-label samples. The model showcased a high recall rate, capturing a significant portion of the proper labels, albeit at the cost of some false positives, as often seen in the trade-off between recall and precision. The future of movie genre prediction seems promising, with the blend of deep learning techniques and rich datasets. While this study made notable progress, there remains scope for enhancement, especially in balancing precision and recall for optimal results. Integrating cultural and contextual nuances, refining model parameters, and exploring other deep learning models can further push the boundaries in movie genre prediction, making it more nuanced and accurate. This research lays a solid foundation and offers a comprehensive blueprint for future endeavours in this domain.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Makita E and Lenskiy A 2016 CoRR abs/1604.08608 (Preprint 1604.08608) URL http://arxiv.org/abs/1604.08608

DOI: 10.54254/2755-2721/47/20241383

[2] Zhang Z Gu Y Plummer B A Miao X Liu J and Wang H 2022 Effectively leveraging Multi-modal Features for Movie Genre Classification (arXiv Preprint 2203.13281)
[3] Kundalia K Patel Y and Shah M 2019 Augmented Human Research 5 11 ISSN 2365-4325 URL https://doi.org/10.1007/s41133-019-0029-y
[4] Suzuki K 2017 Overview of deep learning in medical imaging. Radiological physics and technology 10(3) 257-273
[5] Qiu Y et al 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72 103323
[6] Huang J Chai J and Cho S 2020 Deep learning in finance and banking: A literature review and classification Frontiers of Business Research in China 14(1) 1-24
[7] Bert-base-chinese 2021 https://huggingface.co/bert-base-chinese
[8] Devlin J Chang M W Lee K and Toutanova K 2019 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv Preprint 1810.04805)
[9] Huang Y Cheng Y Chen D Lee H Ngiam J Le Q V and Chen Z 2018 CoRR abs/1811.06965 (Preprint 1811.06965) URL http://arxiv.org/abs/1811.06965
[10] Kim C Lee H Jeong M Baek W Yoon B Kim I Lim S and Kim S 2020 (arXiv Preprint 2004.09910)