이학석사학위논문

Movie Genre Classification Based on Plot Description

영화 줄거리에 기반한 영화 장르 예측

2019년 2월

서울대학교 대학원

통계학과

이종진

Abstract

Movie Genre Classification Based on Plot Description.

Jongjin Lee
The Department of Statistics
The Graduate School
Seoul National University

Movies plot is designed to provide movie's information to audiences. it means that movie genre information could be inherent in movie plot. Based on this fact, we perform movie genre classification from plot description in this study. To make a genre classifer from movie plot, we consider two requirements for the classifier. First, the classifier has to be capable of extracting features from document. Second, because not all of the sentences and words are related with movie genre, the classifier with attention mechanism would be better. Considering these two aspects, we determined to use Hierarchical Attention Network (HAN, Yang et al. (2016)) as a classifier in this study. It is the document classifier using bidirection GRU, which has attention mechanism. We use HAN architecture with a bidirectional LSTM instead of bidirectional GRU and train it using Wikipedia Movie Plots as a dataset. We evalute trained classifier's performance using test set, and investigate which words are important to determine movie genre, using activation value of attenttion of words and sentences.

Keywords : *Movie genre classification, Movie plot description, Attention mechanism, Hierarchical Attention Network*

Student Number : 2017-28763

Contents

Chapter 1

Introduction

Natural Language Processing(NLP) has been widely studied in recent years, and has shown considerable performance in text analysis like classification, translation and sentiment analysis. Advances in treating with natural language have brought conveniences in our lives, and NLP techniques still have possibility to be applied in many ways. Therefore, in this study, we try to apply up-to-date RNN architecture for real-world problem and our main goal is making a classifer that can automatically tag movie genre based on movie plot.

Movie plot is designed to provide movie's information to audiences, people can easily capture movie's information like genre, through reading a movie plot. If we can extract feature through a proper model, classifying movie genre is possible in a extent. In this study, we applied Hierarchical Attention Network(HAN, Yang et al. (2016)) with bidirectional LSTM as a classifier. We use bidirectional LSTM to extract features from embedded word vectors, and aggregate these into sentence vectors. Similarly, plot vectors are calculated with sentence vectors. Finally, using plot vectors, we classify movie genre. In addition with bidirectional LSTM, Annotation mechanism are also applied to word's and sentence's level to help classifier to focus on genre relatives words and sentences.

We review the concepts of Word2Vec and LSTM which are used in the classfier being used in chapter 2. And, chapter 3,4 discuss about data and model which are used for this study. After that, Chapter 5 is about the performance of trained classifer and analysis of the attention values. Lastly, Conclusion can be found in chapter 7.

Chapter 2

Review of Word2Vec, LSTM

2.1 Word2Vec

Word2Vec (Mikolov et al. (2013)) is a embedding method that gives vector expressions of words. Vector expression of Word2Vec reflects word's semantic relations, because it is based on distributional hypothesis that words that are occur in same contexts tend to possess similar meaning. Because of this property, NLP tasks using Word2Vec as a embeding method shows good performance.
There exist two kinds of Word2Vec. One is Continuous Bag of Words(CBOW) and the other is Skip-gram. CBOW predicts central word using context words, on the other hand, Skip-gram predicts context words using central word. Brief architecture of the both models are in Figure 2.1.

CBOW predicts central word with predefined number of adjacent context words. Sentences are reshape into pairs of context words and central word,

and context word's one-hot vectors are input of the model. Through the matrix W_1, words are transformed into embedded vertors, and its mean and W_2 matrix are used to predict central word. The matrix W_1 and W_2 are parameters to be updated.
On the other hand, Skip-gram predicts context words from central word. Generally, Skip-gram shows better performance than CBOW.

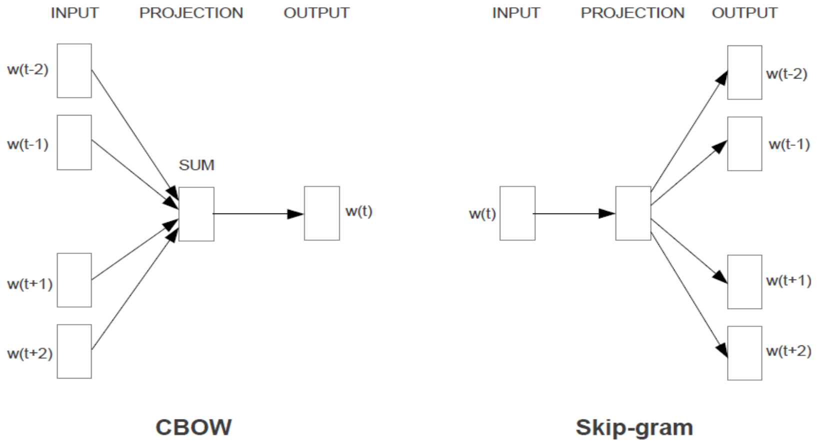

Figure 2.1: Word2Vec

2.2 LSTM

LSTM (Hochreiter, S. and Schmidhulber, J. (1997)) is one of the RNN architectures which are widely used. Unlike RNN, It has advantages that it does not have vanishing gradient problem and long term dependency. It consists of three gates, an input gate, an output gate and a forget gate(f_t). Forget gate(f_t) controls the extent of forgetting memory from the past time. Input gate(i_t) is devised to control the extent of memory to keep flowing. Lastly,

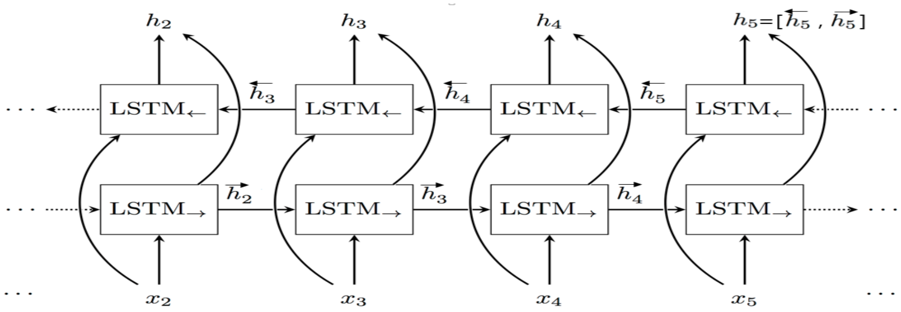

Figure 2.2: Bidirectional LSTM

output gate controls the value of output.

The following is LSTM model at time t. it receives present data x_t and past memory h_{t-1} as a input. ($\odot$ donotes the Hadamard product.)

$$\begin{aligned}
f_t &= \sigma(W_f x_t + U_h h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_h h_{t-1} + b_i) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
g_t &= tanh(W_g x_t + U_g h_{t-1} + b_g) \qquad (2.1) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot tanh(c_t) \\
& x_t \in \mathbb{R}^p,\ W \in \mathbb{R}^{d \times p}\ U \in \mathbb{R}^{d \times d},\ b \in \mathbb{R}^d
\end{aligned}$$

In NLP task, It is known that using features from reverse sequences of words can help to improve its performance. Therefore, in this study, Birectional LSTM (Schuster M. and Paliwal, K. K. (1997)) is used. Figure 2.2 shows the architecture of Bidirection LSTM. The architecture of bidirectional LSTM is simple. It has hidden state which is outcome of concatenation of LSTM and reverse LSTM.

Chapter 3

Data & Data Preprocessing

3.1 Data

For making movie genre classifier, Wikipedia Movie Plots [3.1] is used as a dataset. Wikipedia Movie Plots dataset is collection of plot summary descriptions scraped from Wikipedia, which contain descriptions of 34,886 movies from around world. This dataset contains 8 columns including 'Genre' and 'Plot', and only 'Plot' and 'Genre columns are used in this study. After preprocessing, data are splitted into train set, validatiion set and test set, with a ratio of 3:1:1. Train set and validation set are used for training a classifer, and test set is used for evaluating a trained classifier.

We will discuss about preprocessing of data later, After preprocessing, data have 19247 movies plots and 8 movie genres. 8 movie genres and count are followed.

[3.1] https://www.kaggle.com/jrobischon/wikipedia-movie-plots

Table 3.1: Genres

Genre	Count
Comedy	7217
Romance	2670
Action	2449
Thriller	1781
Crime	1607
Horror	1587
Western	974
Science Fiction	962
Total	19247

3.2 Data Preprocessing

Wikipedia Movie Plots dataset is not a well-refined dataset for training a classifier. Before training a classifier, preprocessing procedures are need to be done.

For movie genre, Misspelled genre are corrected and irrelevant sentences are deleted first. After that, we unify expression of genres. For example, we unify 'science fiction, 'sci-fi', 'science-fiction' to 'science_fiction' and seperate 'romcom' to 'romance' and 'comedy'. Subordinate genres are also unified into semantically parent genres. Lastly, Movies which have multiple genres are romoved from dataset and a genre 'Drama' is deleted to get enough samples of movies which do not have multiple genres. 8 movie genres which have enough samples are used. (Despite of all the preprocessings, It still has problem because of relation between genres.)

For movie plot, all of the capitalized words are converted into lowercase and the words containing apostrophe are converted into unabbreviated form. All of punctuation mark are deleted. Lastly, movies whose plot have more than two sentences are used for this study.

Chapter 4

Methodology

4.1 Embedding

To make a classifer using natural language, a embeding method is needed. In this study, Word2Vec is used. However, because of the shortage of the samples, pre-trained Word2Vec model has to be used. Google's pre-trained model [4.1] is used, which contains 3 millions of words that they trained on roughly 100 billion words from a Google News dataset. The word embedding dimension we use is 300. When we embeding words in the dataset, we skip the words which are not in Word2Vec model.

[4.1] https://code.google.com/archive/p/word2vec/

4.2 Method

We use Hierarchical Attention Networks (Yang et al. (2016)) as a classifier. HAN is a document level classifer which consists of GRU-based sequence encoder. It extract features from word vectors, and aggregate these into sentence vectors and aggregate sentence vectors again into document vectors. It also has two attention mechanism that lead the classifier to focus on class related words and sentences. We apply HAN with bidirectional LSTM instead of bidirectional GRU as a classifier. The HAN arhchitecture we used are as followed

Assume that the dataset has M movies, each movies has S sentence, and each sentences has W words. $w_{ijk}, i = 1, ..., M, j = 1, ...S, k = 1, ..., W$ represent embedding vector expression of ith word in the jth sentence in the kth movie. The classifer calculate sentences vector s_{ij} and movies vector m_i sequentially.

Equation (4.1) shows the process that sentences vectors are calculated with a bidirectional LSTM, and attention mechanism.

$$\begin{aligned}
&h^w_{ijk} = (\overrightarrow{h}_{ijk}, \overleftarrow{h}_{ijk}) = Bi\_LSTM_w(w_{ijk}) \\
&\alpha^w_{ijk} = tanh(W_w h^w_{ijk} + b_w) \\
&z^w_{ijk} = \frac{\exp({\alpha^w_{ijk}}^T c^w)}{\sum \exp({\alpha^w_{ijk}}^T c^w)} \\
&s_{ij} = \sum z^w_{ijk} h^w_{ijk}
\end{aligned} \tag{4.1}$$

In the same way, plot vectors are calculated with the sentence vectors.

$$
\begin{aligned}
h_{ij}^{s} &= Bi\_LSTM_s(s_{ij}) \\
\alpha_{ij}^{s} &= tanh(W_s h_{ij} + b_s) \\
z_{ijk}^{s} &= \frac{\exp({\alpha_{ij}^{s}}^T c^s)}{\sum \exp({\alpha_{ij}^{s}}^T c^s)} \\
m_i &= \sum z_{ij}^{s} h_{ij}^{s}
\end{aligned}
\tag{4.2}
$$

And, the probabilities of ith movie plot being classified into each genre are calculated with plot vector and softmax function.

$$
p_i = softmax(W_m m_i + b_m) \tag{4.3}
$$

Chapter 5

Results

Using the proposed architecture, we fit a classifier for movie genre. Hidden states of word-LSTM and setence-LSTM are 30-dimension, respectively. And, mini-batch size of 10 is used. Number of words and Number of sentences are used a maximum number of words and sentences in each mini-batch. Stochastic gradient with momentum 0.9 is used for updating parameters.

5.1 Performance of the classifier

We evaluate the performance of a trained classifier with precision, recall and F1 score by class. The following table shows Precision, Recall and F1 Score by class which are calculated with test set.

For 'Western', 'Comedy', a trained classifier shows great performance, but for some genres, especially 'Thriller', the classifier shows poor performance. However, it is inevitable result, because of ploblematic dataset. Even if, we

Table 5.1: Measures

Genre	Precision	Recall	F1 score
Comedy	0.8315	0.7160	0.7694
Romance	0.4947	0.5890	0.5377
Action	0.5313	0.5294	0.5303
Thriller	0.2115	0.3221	0.2553
Crime	0.4444	0.5679	0.4986
Horror	0.6694	0.6255	0.6467
Western	0.8500	0.8500	0.8500
Science Fiction	0.5956	0.6328	0.6136

have deleted movies with multiple genres, movies remained in data do not purely have one main genre. We can also verify it, through analyzing the values of attention.

5.2 Analysis of Attention values

With the values of attention, we can investigate which words are highly relative with movie genre. First, we normalized word's attention with sentence's attention, then we examine movie plots with five words which have the highest values of attention. The followings are three example of analysis. Five bold words have highest activaiton values and the number in parenthesis indicates the rank's of normalized attention values.

1. True Genre : Science Fiction / Predicted Genre : Science Fiction

Plot 1.

. . .

a quasi-racist human organization named the order of flesh and blood is opposed to the **humanoids(2)**, which the members disparagingly refer to as ”clickers”. the order believes the **humanoids(1)** are planning to take over the world and are a threat to the very survival of the human race.

. . .

ironically, the ”real” maxine had died in a bomb attack which the order intended to harm only **robots(3)**. dr. raven, a once-human **replica(4)** himself, explains to cragis and maxine that not only are they practically immortal in their new forms, they can also be theaded to the highest possible level: after an alteration, they will be able to reproduce. finally, dr. raven looks directly into the **camera(5)** and tells the viewer, ”of course, the operation was a success...or you would not be here.” this final line implies the story was set in the distant past.

The trained model correctly classifies movie genre. Above plot is about Science Fiction, we can easily know that the words have the highest value of attention, closely related with movie genre. Humanoids, Robots, and (Human) replica are one of the main topics of science fiction movie. We can check the attention mechanism is helpful to make movie genre classification.

2. True Genre : Comedy / Predicted Genre : Horror

> Plot 2.
>
> ...
>
> but instead prepares poison for dusty together with mona. dusty shows up to take the money and drinks the poisoned drink. don and mona put an unconscious **dusty(3)** in the car and take her home. dusty wakes up, so mona **kills(4)** her with a hammer. later she saws dusty body into pieces and buries it in the garden, and reveals to don that dusty was not actually pregnant.murphy and his friends, freeman (kevin mckidd) and benji (heath freeman), discover dusty is missing, and suspect that something went wrong and attack les.
>
> ...
>
> don and mona dig out the body of **dusty(2)**and go to the place where the gang lives. while mona is trying to hide body parts in the **freezer(5)**,
>
> ...
>
> the movie ends with murphy going after their car. the screen turns black and the end credits roll. two shots are heard followed by a prolonged honk and children screams, indicating that murphy had **killed(1)** don.

The trained model misclassifies movie genre. The genre of Plot 2. is a 'Comedy', but the classifier classifies it as a 'Horror'. Howerver, Anyone can knows that this movie genre is not a 'Comedy'. Some mistakes might be happen when the dataset are made.This plot should be classified as a 'Horror' or 'Thriller' rather than 'Comedy'. If Plot 2. has correct genre, words with high value of attention are related with genre.

3. True Genre : Horror / Predicted Genre : Thriller

Plot 3.

the film follows max parry (kevin howarth), a disturbed wedding video cameraman, and his unnamed assistant (mark stevenson) as they perform several **murders(4)** that they have videotaped. the two have used a video store tape in order to record the proceedings, breaking the fourth wall and insinuating that the copy of the film being watched is the only existing version of the tape. throughout the film max uses meta-references in order to show off his **gruesome(5)** activities as a serial **killer(3)**. the film raises questions surrounding visceral pleasure, this can be seen in one scene in particular during which the audience cannot see the victims (two at once) being **murdered(2)**, max parry then asks the audience "i bet you wanted to see that, and if you did not , why are you still watching?" the end of the film the audience is left to believe that since they are watching the only copy of the film, that they will potentially become one of max **victims(1).**

The trained model misclassifies 'Horror' as'Thriller'. After examining of test set, we can find that the classifier has difficuty to distinguish 'Horror' and 'Thriller' And that's why the measures of 'Thriller' are poor.

Chapter 6

Conclusion

With a Wikipedia Movie Plots dataset, we apply HAN architecture to classify movie genre based on plot description. Trained model is not a superb classifier, but shows moderate performance. Also, through analyzing of values of attention mechanism, we identify that its attention mechanism is higly effective. However, there are also limitations. Fisrt, we do not consider the case which movies have multiple genres. To overcome relation between movies, movies with multiple genres are need to be used. Second, detailed preprocessings are needed. Lastly, comparison of perfromance with other classifiers is needed. To verify trained classifier's performance, comparison with other classifiers is required. However, there exist lots of work to classify movie genre using other data source like posters. If other sources are used, It surely shows better performance.

Reference

Yang, Z., Yang., Dyer, C., He, X., Smola, A. and Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Associatioin for Computational Linguistics: Human Language Technologies.*

Mikolov, T., Chen, K., Corrado, G. and Dean, J.(2013). Efficient Estimation of Word Representations in Vector Space *ICLR Workshop*

Schuster M. and Paliwal, K. K. (1997). Bidirecdtional recurrent neural networks. *Signal Processing, IEEE Transactions on*, **45**, 2673-2681.

Hochreiter, S. and Schmidhulber, J. (1997). Long short-term memory. *Neural Computation.*

Ertugrul, M., A. and Karagoz, P.(2018). Movie Genre Classification from Plot Summaries Using Bidirectional LSTM *International Conference on Semantic Computing(ICSC).*

국 문 초 록

영화 줄거리에 기반한 영화 장르 예측

영화 줄거리는 청중에게 영화에 대한 정보를 제공하기 위한 목적으로 만들어졌다. 그렇기 때문에, 영화 줄거리에는 영화의 장르에 대한 정보도 자연스럽게 포함되어 있을 것이며, 이를 이용해, 본 연구에서 우리는 영화 줄거리를 기반으로 영화 장르를 예측해보고자 한다. 줄거리를 기반으로 분류모형을 만들기 위해서, 분류모형의 두 가지 필요조건을 고려하였다. 첫 번째로 문서에 적용하여 정보를 추출할 수 있는 분류모형이 필요하다. 두 번째로, 영화줄거리의 모든 단어와 문장이 영화 장르와 관련된 것은 아니므로, 관련 단어에 집중할 수 있는 분류모형이 더 나을 것이다. 이 두 가지 측면을 고려하여 분류모형으로 Hierarchical Attention Network(HAN, Yang et al. (2016))을 분류모형으로 선택하였다. 그것은 Bidirectional GRU를 기반으로 한 분류모형으로 관련 단어 집중할 수 있는 Attention mechanism을 가지고 있다. 우리는 bidirectional GRU 대신 bidirectional LSTM을 사용한 HAN 모형을 이용하였고, Wikipedia Movie Plots을 이용해 모형을 적합해보았다. 그리고 훈련된 모형을 시험자료를 이용해 성능을 살펴보았고, 더 나아가서, Attention의 활성값을 통해 영화 장르 결정에 중요한 역할을 하는 단어와 문장을 본 연구에서 살펴보았다.

주요어 : (영화장르 분류, 영화 줄거리, Attention, Hierarchical Attention Network)

학 번 : 2017-28763