Movie genre Classification

VADYM DUDARENKO

444820

Problem

Movies are one of the most popular means of entertainment. There are large volumes of movie data being generated and shared on the internet every second. The genre of a movie can be deciphered from its synopsis much of the time

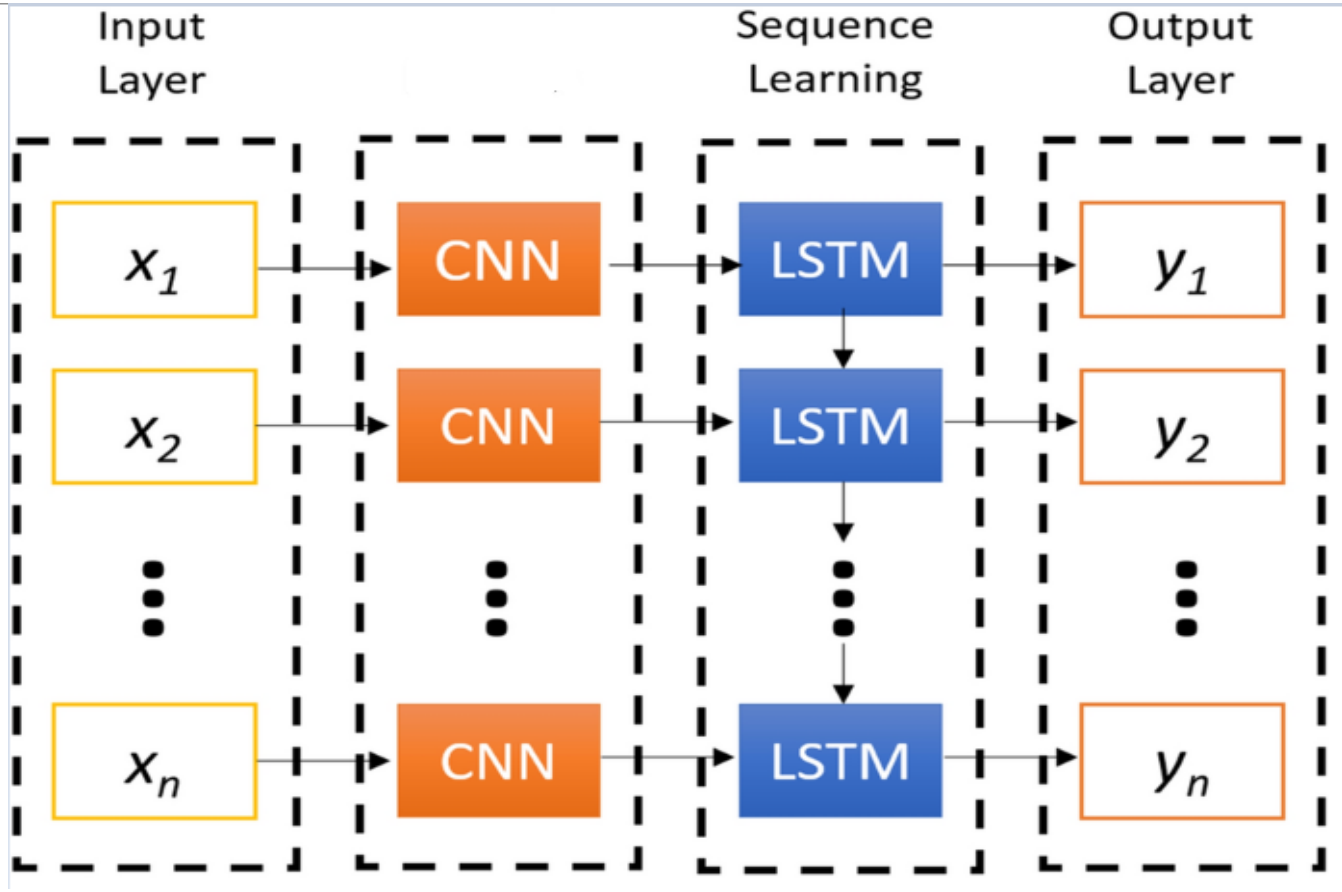
Dataset

	name	genre	released_at	poster	language	director	domain	duration	synopsis	trailer
0	10 Days, 10 Years: Nicaraguan Elections of 1990	Culture & Society	1990-01-01				https://www.allmovie.com/	0H54M		
1	1-2-3 Magic: Effective Discipline for Children	Education	1990-01-01		English		https://www.allmovie.com/	2H0M	Presented by clinical psychiatrist Thomas Phel...	
2	10 Keys to Personal Power	Business	NaT				https://www.allmovie.com/	1H4M		
3	10,000 Maniacs: Time Capsule 1982-1990	Music	1990-01-01	https://cps-static.rovicorp.com/1/avg/cov310/d...	English		https://www.allmovie.com/	0H58M	With their thoughtful folk rock sound and lyri...	
4	10 Rillington Place	Crime, Drama	1971-02-10	https://cps-static.rovicorp.com/2/Open/Sony%20...	English	Richard Fleischer	https://www.allmovie.com/	1H51M	10 Rillington Place is the true story of Briti...	https://video.internetvideoarchive.net/video.m...

Dataset contains next column: name, genre, released_year, poster, language, director, domain, duration, synopsis (description), trailer, cast, url, id.

There are **10,254** movies (observations). Genre of movies are multi-label. So, this classification is **Multi-label Classification problem**

Hybrid model



Model performance metrics

I check the result of confusion matrix, but for assessment. I use **F1-score**. Calculate metrics globally by counting the total true positives, false negatives and false positives. This is a better metric when we have **class imbalance.**

List of optimized hyper parameters

LSTM:

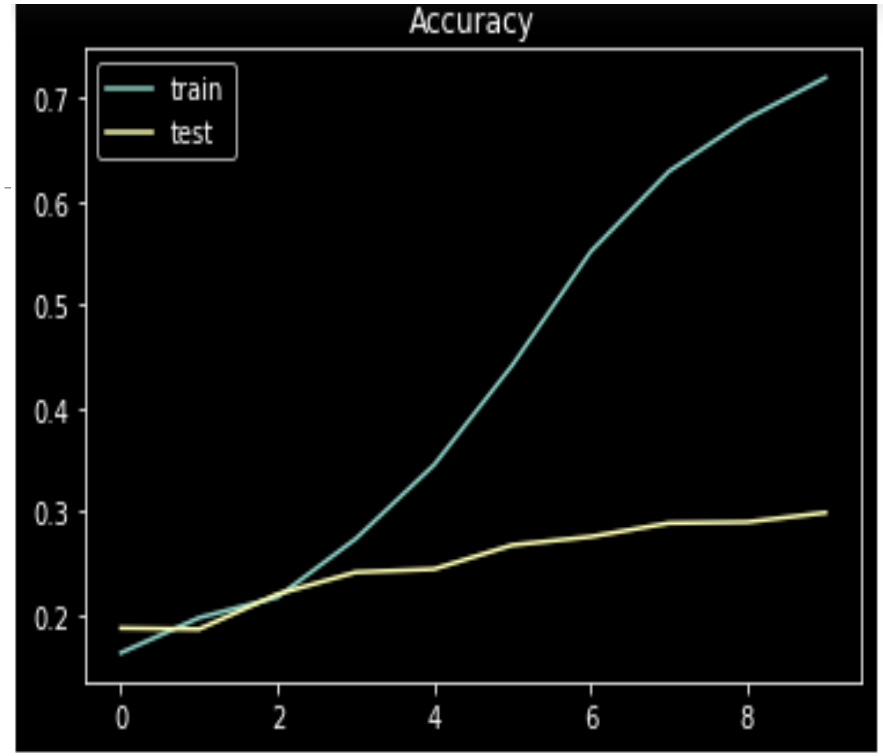
- NUMBER OF NODES AND HIDDEN LAYERS: 30
- NUMBER OF UNITS IN A DENSE LAYER: 41
- DROPOUT: 0.2
- OPTIMAZER: rmsprop
- ACTIVATION FUNCTION: sigmoid
- LEARNING RATE: 0.1
- MOMENTUM: 0.2
- NUMBER OF EPOCHS: 5
- BATCH SIZE: 32

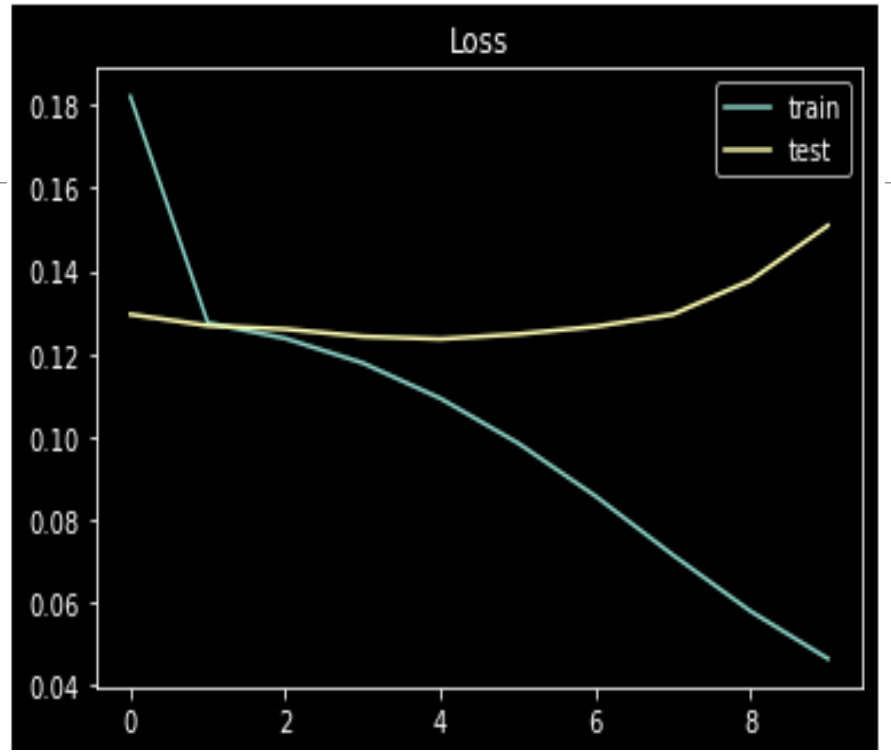
CNN:

- Convolution activation: tanh
- Pooltype: average
- Dense Layer: 41
- Dropout: 0.2

Results

	LSTM	LSTM +	CNN	**CNN+**	LSTM+CNN
F1	0.215	0,261	0.29	**0.34**	0,23
Recall	0.2565	0.274	0.31	**0.51**	0,35
Precision	0.187	0.202	0.28	**0.25**	0.17
Loss	0.12	0.121	0.012	**0,1531**	0,13






```
1/1 [==============================] - 0s 100ms/step
Original tags --> ['Action,Comedy']
Predicted tags --> ['action' 'comedy' 'comedy drama']
```

```
1/1 [==============================] - 0s 38ms/step
Original tags --> ['Drama']
Predicted tags --> ['drama']
```

```
1/1 [==============================] - 0s 31ms/step
Original tags --> ['Mystery']
Predicted tags --> ['mystery' 'science fiction' 'thriller']
```