MOVIE GENRE CLASSIFICATION

Angélica María Gómez Buitrago
Juan Camilo Castillo Castro

Tools and Approach

CPU GPU

Output
Probabilities Average

ML Images

DL Images

ML Text

DL Text

Output
Probabilities Average

Machine Learning Models

ML Images

Preprocessing

- Greyscale Images
- Apply PCA (40 pc)

ExtraTree Classifier

- Number of estimators (220), the depht (10) and the n-jobs (-5).

Fit and Test

```
roc_auc_score(y_test_genres, y_pred_genres, average='macro')
```

```
0.615380091120266
```

Slightly better than RF Classifier

ML Text

Preprocessing

- Countvectorizer: Ngrams(1,2)
- Countvectorizer Ngrams (1,2) + stop_words
- ★ TfidfVectorizer

RandomForest Classifier

- We changed the number of estimators (220), the depht (10) and the n-jobs (-5).

Fit and Test

ROC

```
roc_auc_score(y_test_genres, y_pred_genres, average='macro')
```

```
0.8238598812838025
```

Deep Learning Models (1)

DL Images

Preprocessing

Build CNN

Fit and Test (5 epochs)

- Greyscale images
- Data Generator

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 254, 158, 32)	320
conv2d_2 (Conv2D)	(None, 252, 156, 32)	9248
max_pooling2d_1 (MaxPooling2	(None, 126, 78, 32)	0
dropout_1 (Dropout)	(None, 126, 78, 32)	0
conv2d_3 (Conv2D)	(None, 124, 76, 64)	18496
conv2d_4 (Conv2D)	(None, 122, 74, 64)	36928
max_pooling2d_2 (MaxPooling2	(None, 61, 37, 64)	0
dropout_2 (Dropout)	(None, 61, 37, 64)	0
flatten_1 (Flatten)	(None, 144448)	0
dense_1 (Dense)	(None, 128)	18489472
dropout_3 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 24)	3096

Total params: 18,557,560

```
roc_auc_score(y_test_genres, y_pred_cnn, average='macro')
```

```
0.49899397414360563
```

Deep Learning Models (2)

DL Text

Basic NN

Preprocessing

Build Basic Neural Network

Fit and Test (50 epochs)

- Countvectorizer: Ngrams(1,2)
- ★ Countvectorizer Ngrams (1,2) + stop_words
- TfidfVectorizer

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_40 (Dense)             (None, 256)               8218368
_________________________________________________________________
activation_34 (Activation)   (None, 256)               0
_________________________________________________________________
batch_normalization_18 (Batc (None, 256)               1024
_________________________________________________________________
dropout_24 (Dropout)         (None, 256)               0
_________________________________________________________________
dense_41 (Dense)             (None, 24)                6168
_________________________________________________________________
activation_35 (Activation)   (None, 24)                0
=================================================================
Total params: 8,225,560
Trainable params: 8,225,048
Non-trainable params: 512
_________________________________________________________________
```

ROC

```
roc_auc_score(y_test_genres, model.predict(X_test_dtm), average='macro')
```

```
0.8057582253354049
```

LSTM NN

Preprocessing

Build LSTM Neural Network

Fit and Test (3 epochs)

- Tokenizer: max_words = 10000 / max_len = 1000

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 1000, 32)          320000
_________________________________________________________________
lstm_3 (LSTM)                (None, 100)               53200
_________________________________________________________________
dropout_4 (Dropout)          (None, 100)               0
_________________________________________________________________
dense_5 (Dense)              (None, 256)               25856
_________________________________________________________________
dropout_5 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_6 (Dense)              (None, 24)                6168
=================================================================
Total params: 405,224
Trainable params: 405,224
Non-trainable params: 0
_________________________________________________________________
```

```
[ ]  roc_auc_score(y_test_genres, model.predict(sequences_matrix_test), average='macro')
```

```
0.5416351929307012
```

! *Even with GPU it took a lot of time to run*

THANK YOU