Final Project

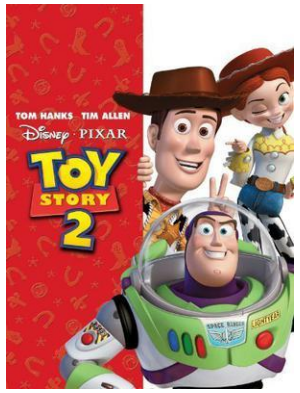
Movie genre classification

Universidad de los Andes

Deep Learning and Neural Network

Ángela Ordoñez C. - Juan Sebastián Garcés C.

Movie genre classification



Objective

Classify a movie genre based on plots and posters.

Methodology

- Machine Learning algorithm using both images and text
- Deep Learning algorithm using both images and text
- Embed results of the best model found in images and the best model found in text.

Universidad de los Andes

Movie genre classification

Using the movie plots - Text

The function TfidfVectorizer, allows summarize the text to a matrix, using the parameters: ngram, max_features and stemmer:

```
vect2 = TfidfVectorizer(ngram_range=(1, 2),max_features=15000,analyzer=split_into_stemmer)
X_dtm2_train = vect2.fit_transform(X_train)
X_dtm2_test = vect2.transform(X_test)
X_dtm2_train.shape
```

This transformation will be use to fit a random forest (Machine Learning) and also to a Neuronal Network (Deep Learning).

Universidad de los Andes

Movie genre classification

Using the movie posters - Image

A PCA with 80 componentes was applied to the flatten grey scale and color images. For the 80 components, fit a random forest as a Machine Learning algorithm.

After split Training color images into train and validation, the pre-trained VGG16 model without de top layer was used as input of a neural network (Deep Learning).

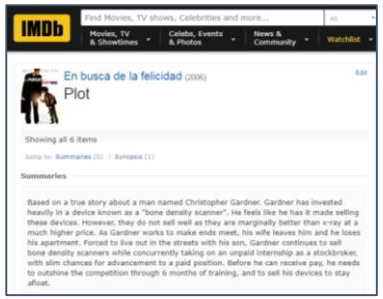
```
from keras.applications.vgg16 import VGG16

model_vgg16 = VGG16(weights='imagenet', include_top=False)
```

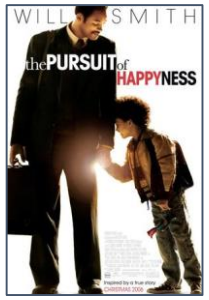
```
from keras.models import Model
model1 = Model(inputs=model_vgg16.input, outputs=model_vgg16.get_layer('block5_pool').output)
```

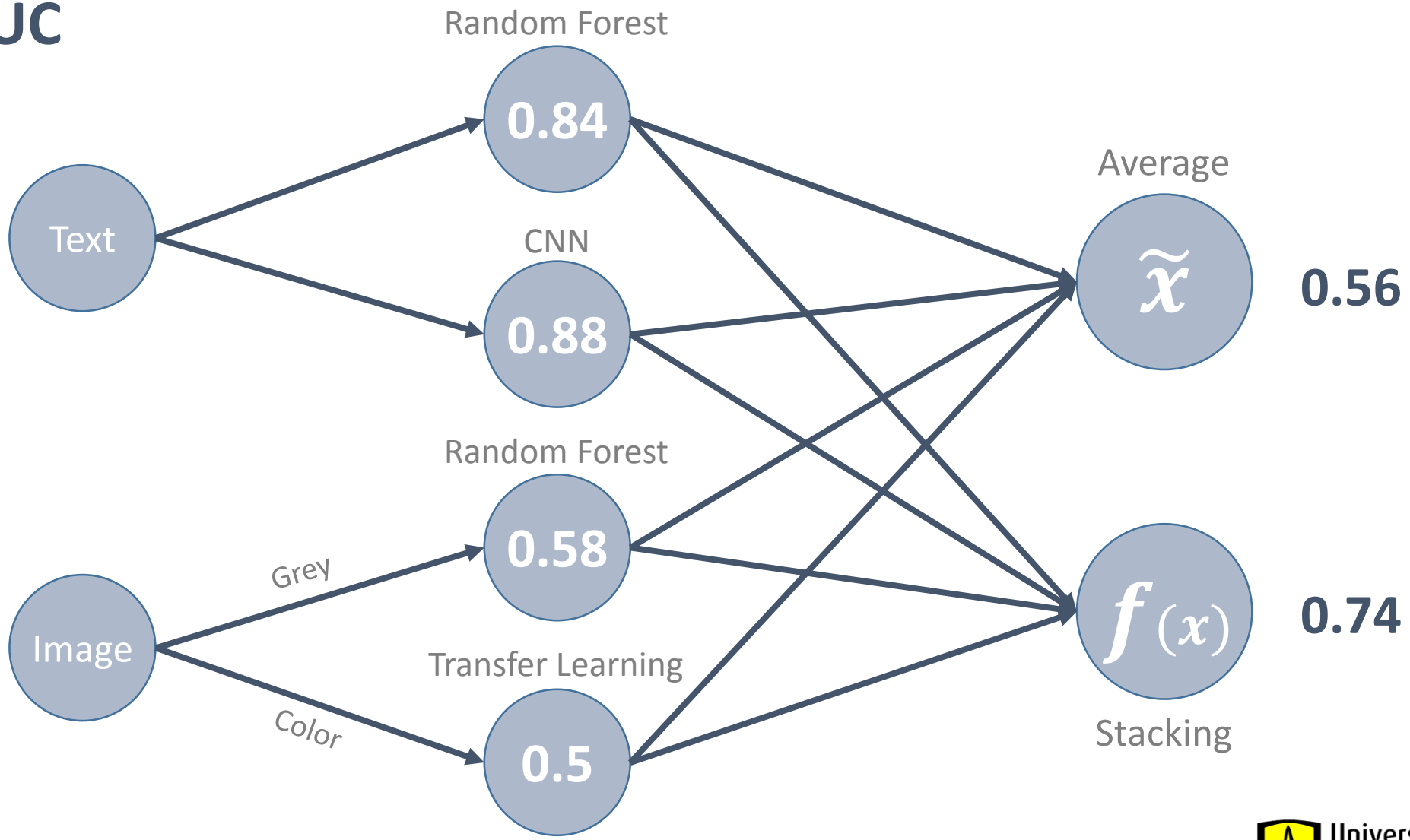
Universidad de los Andes

Movie genre classification

Results - AUC







Universidad de los Andes

Movie genre classification

Random Forest – Text

```
[ ]  Modelo2.fit(X_dtm2_train, y_train)
```

```
OneVsRestClassifier(estimator=RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                    max_depth=10, max_features='auto', max_leaf_nodes=None,
                    min_impurity_decrease=0.0, min_impurity_split=None,
                    min_samples_leaf=1, min_samples_split=2,
                    min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=-1,
                    oob_score=False, random_state=50, verbose=0, warm_start=False),
                  n_jobs=1)
```

```
[ ]  Modelo2=clf.fit(X_dtm2_train, y_train)
     y_pred2 = clf.predict_proba(X_dtm2_test)
     roc_auc_score(y_test, y_pred2, average='macro')
```

```
0.83952160425667
```

Universidad de los Andes

Movie genre classification

Random Forest – Images

```
[ ]  clf.fit(X_train, y_train_genres)
```

```
OneVsRestClassifier(estimator=RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=10, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=-1,
            oob_score=False, random_state=42, verbose=0, warm_start=False),
          n_jobs=1)
```

```
[ ]  y_pred_genres = clf.predict_proba(X_test)
     roc_auc_score(y_test_genres, y_pred_genres, average='macro')
```

```
0.5798856252910195
```

Universidad de los Andes

Movie genre classification

CNN – Text

```
[ ]  max_words=15000

     Modelo4= Sequential()
     Modelo4.add(Dense(300, input_shape=(max_words,)))
     Modelo4.add(Activation('relu'))
     Modelo4.add(BatchNormalization())
     Modelo4.add(Dropout(0.7))
     Modelo4.add(Dense(24, activation='sigmoid'))
```

```
[ ]  Modelo4.compile(loss = 'categorical_crossentropy',
                     optimizer = Adagrad(),
                     metrics = ['accuracy'])

     Modelo4.fit(X_dtm2_train, y_train, epochs=10, verbose=2)
```

```
Epoch 1/10
 - 2s - loss: 7.0963 - acc: 0.2532
Epoch 2/10
 - 2s - loss: 5.7879 - acc: 0.3139
```

```
[ ]  y_pred4=Modelo4.predict_proba(X_dtm2_test)
     roc_auc_score(y_test, y_pred4, average='macro')
```

```
0.875589276062426
```

Universidad de los Andes

Movie genre classification

Transfer Learning – Images

```
model2.summary()
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
flatten_1 (Flatten)          (None, 25088)             0
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              25691136
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_3 (Dense)              (None, 24)                12312
=================================================================
Total params: 26,228,248
Trainable params: 26,228,248
Non-trainable params: 0
_________________________________________________________________
```

```
roc_auc_score(np.array(y_valCN), np.array(preds.tolist()), average='macro')
```

```
0.487433200743367
```

Universidad de los Andes