An Exploration of Deep Learning Methods to Predict Media Memorability Task, Using Captions

Harshal Chaudhari
Dublin City University, Ireland
harshal.chaudhari2@dcu.ie

ABSTRACT

In this paper, we implement different deep learning approaches to investigate their performance in predicting media memorability. We use Multi Layer Perceptron (MLP), Recurrent Neural Networks (RNN) and Convolutional Neural Network (CNN) to predict short and long term scores. Our approach provides a baseline comparison of using deep learning methods for the task. We find that customized models provide reliable results using captions and discover that neural networks achieves good score for short term memorability.

1 INTRODUCTION

In this paper, we investigate the video memorability scores and predict short and long term score as part of the MediaEval Predicting Media Memorability challenge. Among the features given, we use captions to predict memorability as previous work has shown good results as compared to other features like HMP, C3D, ColorHistogram and InceptionV3. We train different neural network models to predict the score. The models are evaluated using Spearman Rank Correlation Coefficient.

Artificial Neural Networks (ANN) have been previously used to classify objects in images such as cats, by watching unlabelled images [3]. Similarly, Convolutional Neural Networks (CNN) are advance type of ANN also used for object recognition in images. Furthermore, Recurrent Neural Networks (RNN) have been used previously for natural language processing tasks.

This paper uses such ANN models to predict memorability scores using captions (labelled videos). Short term predictability score dominates in all the models when results are evaluated.

2 RELATED WORK

In the previous years, large amount of research work has started to progress in predicting video memorability, and recent work [2] [6] explore the use of various features like C3D, ColorHistogram, image and video captions for predicting memorability. The key findings among the previous works suggest that captions provides best individual results. Furthermore, deep learning methodologies has been used to learn such features. CNNs trained for image classification has provided unprecedented results on various object recognition tasks. [5]

Furthermore, Gupta and Motwani have used Lasso Logistic Regression, Support Vector Regression and ElasticNet on different set of features including captions. Results indicate achieved score of 0.5 and 0.26 for short term and long term respectively using ResNet model.

MediaEval'18, Dublin City University, Ireland

Smeaton et al [7] have used multi layer perceptron on C3D features, HMP, HOG descriptors and more. They have also used video and image saliency features, lastly have implemented an ensemble approach for prediction. The results by using saliency were better than perceptron approach. Although, lack of hyperparameter tuning and train data size, the experiments provided poor results.

3 APPROACH

We propose an approach where 3 different models are trained on captions and provide a comparison of the score from the models. 1. Multilayer Perceptron (MLP) 2. Recurrent Neural Network (RNN) and 3. Convolutional Neural Network (CNN) are the models we used for training. Hyper-parameters were selected using GridSearch for the model such as activation function, L2 Regularization, epochs, etc.

In the first approach, we have used 3 layer MLP to train our model. We have also applied Ridge (L2) Regularization for each layer to avoid over fitting. Moreover, after each layer dropout has been added to further reduce over fitting of the model.

For the second approach, we have used a RNN as their performance is excellent [4] on Natural Language Processing (NLP) tasks. RNNs have an embedding layer at the beginning, followed by a layer of Long Short Term Memory (LSTM) neurons. The LSTM layer has 150 hidden neurons followed by another layer of 30 neurons to reduce the dimension of the network. Furthermore, dropout and regularization are added to reduce over fitting.

In the last approach, we have used CNN for training the model. Although, CNNs are widely used [1] for image classification tasks, we have decided to evaluate and compare the performance in our use case. CNNs also have an embedding layer at the beginning and followed by a convolutional layer. We have choosen 1D Convolutional layer for the experiment. Further, to reduce the dimensionality of our network we used another two layers of 10 neurons each. Each layer of neurons has dropout and regularization applied to them.

3.1 Data Preprocessing

The feature we chose for training and testing our models are video captions. Each video has been labelled with a short description of what is the content of the videos. To provide these as a feature to our neural network approach. We first convert words into vectors using tokenization. Tokenization of words is performed by mapping words from the corpus to numbers. In our case, we go with one-hot encoding (binary encoding) for MLP and CNN models. In case of RNN, we use sequence encoding.

Furthermore, we notice for sequence encoding the stored captions have uneven length. Therefore, we add padding at the start of

each feature to make up for the uneven length of the captions. We add padding of zeros at the beginning of each caption.

3.2 Data Cleaning

Captions are stored as a string of multiple words. We remove any punctuations and replace those with spaces. Along with that every word in the corpus is converted to a consistent case.

3.3 Training Phase

3.3.1 ***Multi Layer Perceptron:*** For training our model, we split the dataset into 80% for training and 20% for testing and validation. As mentioned previously, we begin our training phase by building a Multi Layer Perceptron model. We are using a 3 layer MLP model with first two layers consisting of 10 neurons each and last output layer of 2 neurons to predict short and long term scores.

We also add dropout layers in between to reduce the risk of overfitting the data. Along with dropout, we use Ridge regularization to regularize the neural network and drop dimensions of lower weights.

For hyper parameter selection, we use Gridsearch to find the optimum parameters. Using GridSearch, we find and chose activation function as selu for two layers and sigmoid for the output layer. For optimizer, we choose adamax as the best optimizer function and Mean Squared Error as loss function. Finally, we train the model for 20 epochs which during training phase fits the data very well as can be seen in Fig 1. the training and validation loss graph.

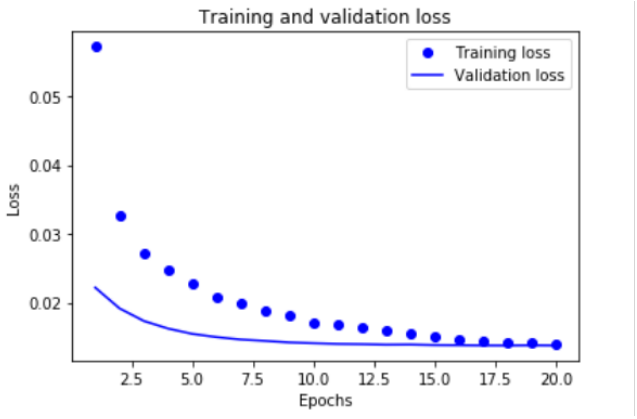

Figure 1: Training and Validation Loss for MLP

3.3.2 ***Recurrent Neural Network:*** For RNN, we build our model using LSTM which perform very well in natural language processing tasks. For LSTM, we design our model with 150 hidden layers followed by another layer of 30 neurons. For output layer, we go with two neurons each for short and long term score.

Similar to MLP approach dropout layers are added to reduce over fitting the data. We also use ridge regularization here to drop dimensionality corresponding to lower weights.

For hyper parameter selection, we use GridSearch to find the best parameters. We find that activation function selu performs the best for initial layers and sigmoid for output layer. For optimizer function, we go with adamax and Mean Squared Error as loss function. The model is trained for 10 epochs and performance is evaluated using training and validation loss graph shown below in figure 2.

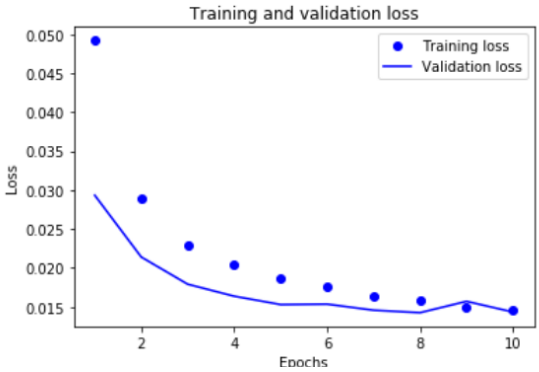

Figure 2: Training and Validation Loss for MLP

3.3.3 ***Convolutional Neural Networks:*** We choose CNN for our next model. CNNs are known to perform very well [1] on image classification tasks. For CNN approach, we use sequence encoding for training our model. Train and test split are kept the same for consistency purpose. We design the CNNs using an embedding layer and 1D convolutional layer. We create a convolutional layer of 128 filters with window size of 5. We add another layer of 10 neurons to reduce the dimensions of the network.

To reduce over fitting the data, we add dropout layer and each layers uses ridge regularization to reduce dimensions with lower weights. For hyper parameter selection, we use GridSearch to find the best parameters. We find that activation function selu performs the best for initial layers and sigmoid for output layer. For optimizer function, we select Adamax and Mean Squared Error as loss function. The model is trained for 20 epochs and performance is evaluated using training and validation loss graph shown in figure 3.

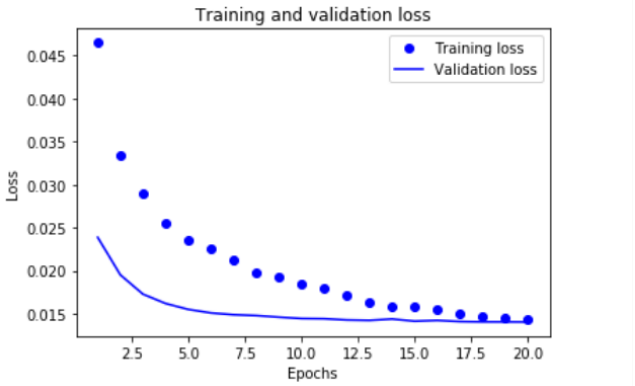

Figure 3: Training and Validation Loss for MLP

4 TESTING PHASE

During the testing phase, each model predicts the remaining 20% of the data. The results are evaluated using Spearman Rank Correlation. In our experiments we find that MLP performs the best among all followed by RNNs and CNN respectively.

5 RESULTS AND ANALYSIS

During the evaluation of the models, we find that MLP model performs the best for short term score and RNN model performs the best for long term score. CNN model perform dramatically poor as compared to the other two. Although, the results of were expected as CNNs are suitable for image classification task.

Moreover, we find that RNNs perform very well for our task as they use LSTM layers which have memory component in them. Further, we can see that a custom MLP model when designed correctly and fine tuned with parameters can provide reliable results as compared to a complex model.

Below Table 1 shows the scores for memorability for the different models used during our experiments.

Table 1: Spearman Rank Score

Model/Score	MLP	RNN	CNN
Short Term	0.417	0.414	0.17
Long Term	0.192	0.2	0.08

6 CONCLUSION AND FUTURE SCOPE

For the results and analysis of using different neural networks and deep learning approach, our approach provides a baseline analysis of different models and their performance for predicting memorability scores. We can conclude that usage of a simple customized model like MLP for such use case can be as reliable as complex models such as RNNs. Furthermore, as compared to previous work done on other features our approach provides good score using captions.

As for future additions, an ensemble approach can be used to combine the results of different models by using techniques like blending and stacking . Also, images from the videos can be used to get more features and data to train using image classification models.

REFERENCES

[1] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. 2011. Flexible, High Performance Convolutional Neural Networks for Image Classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two (IJCAI'11)*. AAAI Press, 1237–1242. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-210

[2] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. 2018. Annotating, Understanding, and Predicting Long-term Video Memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. ACM, New York, NY, USA, 178–186. https://doi.org/10.1145/3206025.3206056

[3] Quoc V. Le, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Marc'Aurelio Ranzato, Jeffrey Dean, and Andrew Y. Ng. 2011. Building high-level features using large scale unsupervised learning. *CoRR* abs/1112.6209 (2011). arXiv:1112.6209 http://arxiv.org/abs/1112.6209

[4] Xiangang Li and Xihong Wu. 2014. Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition. *CoRR* abs/1410.4281 (2014). arXiv:1410.4281 http://arxiv.org/abs/1410.4281

[5] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *CoRR* abs/1403.6382 (2014). arXiv:1403.6382 http://arxiv.org/abs/1403.6382

[6] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (2017), 2730–2739.

[7] Alan F Smeaton, Owen Corrigan, Paul Dockree, Cathal Gurrin, Graham Healy, Feiyan Hu, Kevin McGuinness, Eva Mohedano, and Tomás Ward. 2018. Dublin's Participation in the Predicting Media Memorability Task at MediaEval 2018. (2018), 3.