Prediction of Media-Memorability task using Captions

Akshai Ramesh
Dublin City University,
Ireland
akshai.ramesh2@mail.dcu.ie

ABSTRACT

In this paper, we implement a set of deep learning techniques to examine their performance in predicting the memorability of videos. We use Multi Layer Perceptron (MLP), Recurrent Neural Networks (RNN) and Convolutional Neural Network (CNN) to foresee short and long haul scores. Our methodology gives a standard examination of utilizing deep learning strategies for the problem statement. We find that customized models give solid outcomes utilizing subtitles and find that neural networks accomplishes great score for short-term memorability.

1 INTRODUCTION

As part of MediaEval Predicting Media Memorability Challenge, we examine the memorability scores and predict short and long term score. Among the parameters provided, we use captions to foresee memorability as past work has demonstrated great outcomes when contrasted with different highlights like InceptionV3, C3D, ColorHistogram and HMP. We train distinctive neural system models to predict the score. The models are assessed utilizing Spearman Rank Correlation Coefficient.

Artificial Neural Networks (ANN) have been recently used to identify and distinguish objects in pictures, for example, locating the traffic signal in a busy road [3]. Likewise, Convolutional Neural Networks (CNN) are advance sort of ANN additionally utilized for object identification in pictures. Moreover, Recurrent Neural Networks (RNN) have been utilized already for natural language handling problems.

This paper makes use of such ANN models using captions from the given data to predict the memorability scores. Based on the results, its clear and inevitable to conclude that short-term prediction scores dominate in all the models.

2 RELATED WORK

In the previous works, enormous measure of research study has begun to advance in anticipating video memorability, and ongoing work [2] [6] investigate the utilization of different parameters like ColorHistogram,C3D, picture and video captions for foreseeing memorability. The key discoveries among the past works recommend that inscriptions gives best individual results.

Moreover, deep-learning strategies have been utilized to learn this type of parameters. Convolutional Neural Networks trained to identify and classify the images have given remarkable outcomes on different object recognition problem statements. [5]

Also, Gupta and Motwani[9] have utilized ElasticNet, Support Vector Regression and Lasso Logistic Regression on various combination of parameters including captions. Results demonstrate accomplished score of 0.5 and 0.26 for short-term and long-term. This was achieved using ResNet Model. In other cases, Despite the fact that, absence of hyper-parameter tuning and train data size, the tests gave poor outcomes.

Significant contribution from their work is that they could recognize the words in captions with positive/negative coefficients. They could identify and classify the words in relation to nature with negative coefficients and the ones in relation with humans to positive coefficients. Therefore, I investigate more on captions and the effects due to occurence of specific words.

3 APPROACH

We propose a methodology where 3 models are trained based on the feature-caption and finally demonstrate the comparison of results obtained from these models. The models we are going to utilize here for preparation are : Multilayer Perceptron (MLP), Repetitive Neural Network (RNN) and Convolutional Neural Network (CNN).

GridSearch was used to select the hyper-parameters for model. Some of the parameters chosen using this are - epochs, L2 Regularization, activation function etc.

For the first model training, we have used 3-layer MLP for training the model.

In the principal approach, we have utilized 3 layer MLP to prepare our model. We have additionally applied Ridge (L2) Regularization for each layer to maintain a strategic distance from over fitting. Also dropout has been included after each layer to further decrease the chances for over-fitting.

RNN model are known for their amazing performance on Natural Language Processing Tasks. Due to this reason, we have selected RNN as the second model. RNNs have embedded layers toward the start, trailed by a layer of Long Short Term Memory neurons.The LSTM layer has 150 concealed neurons followed by another layer of 30 neurons to diminish the dimension of the system. Moreover, dropout and regularization are added to reduce chances for over-fitting.

In the last methodology, we have utilized CNN for preparing the model. Despite the fact that, CNNs are generally utilized [1] for image classification, we have chosen to assess and analyze the results in our utilization case. CNNs likewise have an embedding layer toward the start and followed by a convolutional layer. We have picked 1D Convolutional layer for carrying out the test. Further, to diminish the dimensionality of our system we utilized another two layers of 10 neurons each. Each layer of neurons has dropout and regularization.

3.1 Data Preprocessing

Caption was chosen as a feature, based on which we train and test our model. Every video has been named with a short depiction of what is the substance of the recordings. To give these as an element to our neural system approach. We first transform words into vectors utilizing tokenization.The words from corpus are mapped to numbers for carrying out tokenization. For our situation, we go with one-hot encoding technique for MLP and CNN models. For RNN we made use of sequence based encoding strategy.

It was observed that stored captions were of variable length, thus sequence encoding was used to rectify this. As part of this, paddings were added at beginning of each feature to resolve the variable length case. Zero was used as the padding-value.

NLTK libraries and defined methods were used for carrying out feature extraction. At first the subtitles were extracted from the given file and afterward were handled to discard all the regex leaving about just significant English words. Alongside this, the stop words were removed and stemming of words was accomplished for standardization. Later the words were vectorized utilizing TF-IDF scores.

3.2 Data Cleaning

Before training the model with caption as a feature, the caption content were stored as a sentence with more than one word. All the captions were converted to a consistent lower case. The punctuations were removed and replaced with spaces.

3.3 Training Phase

3.3.1 Multi-Layer Perceptron : The dataset was split into 80:20 ratio with 80% of data for training and 20% of data for development set. After splitting the data, a Multi-Layer Perceptron model is built. The MLP model consists of 3 layers with 10 neurons each in first two layers and 2 neurons in the last layer to denote as the output label for predicting short and long term scores. The dimensions of lower weight were dropped. This was carried out by making use of dropout and ridge regularization to regularize the neural network.

For choosing the hyper parameters, Gridsearch was used to discover and select the optimum parameters. Selu was used as an activation function for first two layers and the sigmoid function was used for the output layer. For loss estimation, Mean Squared Error function was used. Adamax optimizer was chosen for the optimization. The model was trained for 20 epochs with the training stage fitting the data very well as can be inferred from Fig 1. the training and validation loss graph.

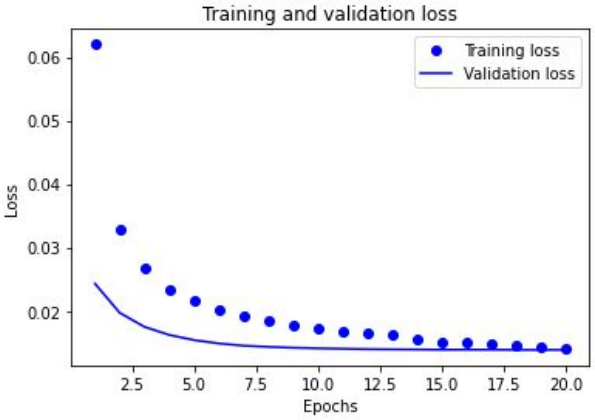

Figure 1: Training and Validation Loss for MLP

3.3.2 Recurrent Neural Network: LSTM (Long short-term memory) is known for amazing performance with natural language processing tasks. With this observation, for RNN, the decision was made to build a LSTM model. Here the model was designed containing 150 hidden layers along with 30 neurons in another layer. Two neurons are used in the output layer to represent short and long term score.

In this scenario its very important to take care and avoid over-fitting. To prevent over-fitting the data, dropout layers are made use of. Along with this, ridge regularization is used to drop dimensionality related to lower weights.

For hyper parameter determination, we use GridSearch to locate the best parameters. We find that selu activation function plays out the best for beginning layers and sigmoid for yield layer. Similar to MLP model, here also we make use of Adamax Optimizer and data loss is calculated using Mean-Squared Error function. The model is trained for 10 epochs and the model is then evaluated using training and validation loss graph as shown in figure 2.

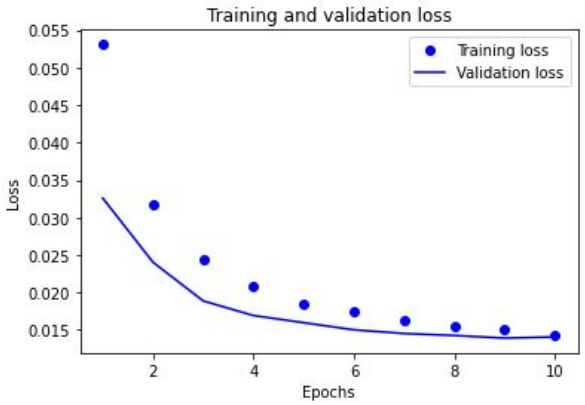

Figure 2: Training and Validation Loss for RNN

3.3.3 Convolutional Neural Networks: We pick CNN for our next model. CNNs are known to perform very well [1] on picture arrangement errands. For CNN approach, we use succession encoding for preparing our model. Train and test split are saved the equivalent for consistency reason. We plan the CNNs utilizing an inserting layer and 1D convolutional layer. We make a convolutional layer of 128 channels with window size of 5. We include another layer of 10 neurons to diminish the components of the system.

Convolutional Neural Network was chosen as the final third model. CNNs are known for their great performance in image-classification tasks. Sequence Encoding is used as Encoding technique for building CNN. The training and test data are equally split to account for the consistency.

To decrease the over-fitting problem, we include dropout layer and every layer utilizes ridge regularization to diminish dimensions with lower weights. For this model also, GridSearch was used to find out the optimal parameters. We find that actuation function selu plays out the best for introductory layers and sigmoid for yield layer. We maintain the standard setup with Adamax optimizer and Mean-Squared Error loss function. After training the model for 20 epochs, inference is drawn using the training and validation loss graph as shown in Fig 3.

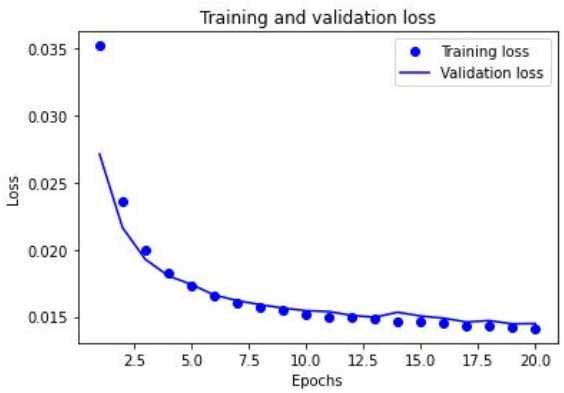

Figure 3: Training and Validation Loss for CNN

4 TESTING PHASE

During the testing stage, each model predicts the staying 20% of the information. The outcomes are assessed utilizing Spearman Rank Cor-connection. In our investigations we find that MLP plays out the best among all followed by RNNs and CNN individually.

After training the models, the test set is given to each model and predictions are made. To evaluate the predictions, Spearman Rank Correlation is used. From the models that we trained, it can be inferred that RNN has the best performance among the lot followed by MLP and CNN respectively.

5 RESULTS AND ANALYSIS

Although CNN is known to be best suitable for image-classification tasks, the results showed that CNN performance was very bad and not on par with the other 2 models. The evaluations showed that RNN model performs the best for short-term score and long-term score, closely followed by MLP model.

We can also infer that for the given task, the performance of RNNs was best, this is due to the LSTM layers which consists of memory segment in them. Also it can be seen that a MLP model with right choice of hyper parameters yields in a very promising result, compared to that from a perplexity model.

The table below shows the clear picture of memorability scores obtained from different models after evaluation of the predicted results :-

Table 1: Spearman Rank Score

Model/Score	CNN	MLP	RNN
Short Term	0.172	0.415	0.424
Long Term	0.092	0.192	0.208

6 CONCLUSION AND FUTURE SCOPE

The decision to use captions as the feature to train the model upon yielded in very satisfactory and useful results. Also to train better model, realizing the need for attention to rare words and their inter-connection was useful to extract new features.

Our models using neural network and deep learning approach provided a baseline analysis for different models and their presentation for anticipating the memorability scores. We can presume that utilization of a basic modified model like MLP for such use case can be as solid as mind boggling models, for example, RNNs.To add, our results (especially RNN and MLP) are very good and comparable with the existing works utilizing caption feature.

As for future additions, long-term memorability can be considered as the significant area of concentration, also an attempt to obtain more features and data can be made to improvise the classifier 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, 1237–1242.5591/978-1-57735-516-8/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, Gra- ham Healy, Feiyan Hu, Kevin McGuinness, Eva Mohedano, and Tomás Ward. 2018. Dublin's Participation in the Predicting Media Memora- bility Task at MediaEval 2018. (2018), 3.

[8] "Stack Overflow" [Online] Available : https://stackoverflow.com/questions/42689066/convolutional-neural-net-keras-val-acc-keyerror-acc

[9] R. Gupta, "Linear Models for Video Memorability Prediction usingVisual and Semantic Features," MediaEval, 2018.