Multimodal Deep Features Fusion For Video Memorability Prediction

Roberto Leyva[1,2], Faiyaz Doctor[1], Alba G. Seco de Herrera[1], Sohail Sahab[2]
[1] University of Essex, Colchester, UK [2]Hub Productions, London, UK
{r.leyva,fdocto,alba.garcia}@essex.ac.uk,sohail@hub.tv

ABSTRACT

This paper describes a multimodal feature fusion approach for predicting the short and long term video memorability where the goal to design a system that automatically predicts scores reflecting the probability of a video being remembered. The approach performs early fusion of text, image, and video features. Text features are extracted using a Convolutional Neural Network (CNN), an FBResNet152 pre-trained on ImageNet is used to extract image features and video features are extracted using 3DResNet152 pre-trained on Kinetics 400. We use Fisher Vectors to obtain a single vector associated with each video that overcomes the need for using a non-fixed global vector representation for handling temporal information. The fusion approach demonstrates good predictive performance and regression superiority in terms of correlation over standard features.

1 INTRODUCTION

Remembering videos is a key aspect of advertising, entertainment, and recommendation systems [3]. We are more influenced by videos that remain fresh in our memory and subsequently share their contents with others. Creating memorable video content is crucial for generating consumer impact and engaging entertainment and profitable marketing campaigns. Understanding and predicting memorability as a function of video features is therefore important for computational video analysis tasks. In this work, we propose a method for video memorability prediction [4] keeping in mind that the videos are not necessarily attractive or interesting. Thus, we explore which features provide better regression results. No assumptions are made on the task's structure, and we proceed to analyze text, image, and video features in combinations to determine their ability to predict long terms and short term memorability using different machine learning based regression techniques. Our findings show that long and short term memorability share the same feature structure giving better accuracy when fusing features of a different type for the short memorability task. These outcomes also leave room for future improvements.

The works that precede this study have addressed the memorability tasks mainly using the provided features or replacing them [2, 6, 7, 12, 25, 26, 26]. The memorability task can be done using single-source or multi-source feature information to train a regression model. Gupta *et al.* [7] propose using images information source via linear highly regularized models to prevent over-fitting using the provided features, Residual Network (ResNet) features and Dense Network (DenseNet) features. Over-fitting is potentially a primary concern in the memorability task. They use Least Absolute Shrinkage and Selection Operator (LASSO) [23], Support Vector Regression (SVR), and Elastic Network (ENet) for their experiments.

Savii *et al.* [20] propose using only the video temporal information employing video features for the memorability task. Here the method is passing Convolution 3D (C3D) [24] and Histogram of Motion Patterns(HMP) [1] features to a Deep Neural Network (DNN) where the final score is obtained using a DNN+ k-Nearest Neighbour (k-NN) regressor. In similar work, Tran-Van *et al.* [25] proposes a solution to capture the temporal information where they combine Image features IV3 with an Long Short Term Memory (LSTM) to produce the memorability score.

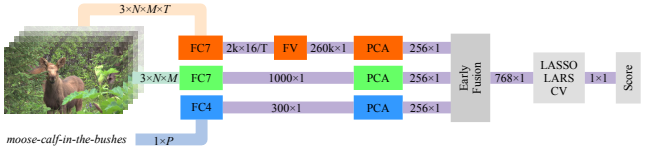

Figure 1: Video memorability prediction pipeline via three-stream media source information. We early fuse text, image and video features to create a memorability score.

2 APPROACH

Multi-source feature fusion usually gives improved results over isolated modeling of features as has been shown in [6, 7, 12, 25, 26]. Chaudhry *et al.* [2, 26] models used image, text, and video features and achieved better results when fussing them as compared to modelling them individually [22]. However, fusing multiple features from the same information source, e.g., image source, can increase complexity while giving little improvements to the tasks' performance [6]. For instance, Joshi *et al.* [12] propose using the Memorability Network [13] along with Hue Saturation and Value (HSV) 3D [6], colorfulness [10], aesthetics [8], saliency Net [18] , C3D [24], and Global Vectors (GloVe) of text features [19]. This approach gives little gains over single-feature source selection. For this reason, we deem appropriate extracting only one feature from each of the following information sources: text, image, and video. Secondly, modeling the Spatio-temporal domain via recurrent networks may become very computational costly [25]. Because we are targeting large-scale video analysis, we consider a less complex approach. Thirdly, to generate the memorability score, we explore linear regularized methods and deep learning models. This consideration rests on the assumption that the latter techniques do not necessarily achieve better generalization, as mentioned in [7]. Finally, we can improve the provided features' performance [17].

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

To this end, we use other feature representations following authors [20, 26] using ConceptNet [21], skip-thought [15]. Thereby, we consider other deep learning approaches for feature extraction giving particular importance to the spatio-temporal domain as [20, 25].

Our proposed method uses three primary feature modalities (text, image, and video) for predicting the memorability score, Figure 1 shows the pipeline in detail.

Text Features: we use the provided video captions as an input text to a Convolutional Neural Network (TCNN). The text is vectorised via tokenization and word embedding into 100 dimensions to feed the network using the IMDB dataset for sentiment analysis [14]. We use this dataset because of the high accuracy of the network on this task ultimately gave us confidence that the model is adequately trained and can be trusted as a feature generator. We use the last Fully Connected (FC) layer as a feature generator resulting from the concatenation of the text input convolution embedding. This process results in a 300-dimensions feature vector, i.e., 3× embedding size.

Image Features: We extract the middle frame of each video clip and apply FBResNet152 [11] pre-trained on ImageNet. To this end, we feed the model the middle frame to extract a 1000-dimensions feature vector from the last FC layer. We also explored selecting other frames from the sequences without achieving better correlation values.

Video Features: To extract video features, we use 3DResNet152 [9] pre-trained on Kinetics-400. We feed the video sequence to retrieve a feature vector for every 16 frames producing a 2048-dimensions feature vector. Although for this particular case we may have fixed-length video clips, in practice the number of frames is not fixed and stacking the produced features may become very computationally complex. Inspired by the work of Girdhar *et al.* [5] using Vector of Locally Aggregated Descriptors (VLAD) vectors for action recognition, we follow a similar approach using Fisher Vectors (FV) to address this problem. The technique then creates a single feature vector for each video sequence using Fisher Vectors. The method is to generate a Gaussian Mixture Model (GMM) model from the 16-frame collection features and project them into a high dimensional space via the soft assignment. As the resulting feature space is considerably high, we reduced the dimensions via Principal Component Analysis (PCA) following an FV-GMM-PCA fashion [16]. This last step provides a single feature vector for each video sequence capturing the motion information from the clips.

Feature Fusion: We combine the text, image and video features via early fusion. Prior to this step, we reduce the features' dimensionality using PCA with 256 components aiming for better feature representation. The vectors are then stacked as $3 \times 256 = 768$ and feed into the regression model, as Figure 1 illustrates. The last step is to perform the regression using a regularized method. To this end, we used LassoLarsCV [23] in the pursuit of cross folding that gives the best regression parameters for the final model automatically.

3 RESULTS AND ANALYSIS

The memorability dataset comprises 10000 short soundless videos split into 8000 videos for the development set and 2000 videos for the test set [4]. The videos are varied and contain different scene types. Also provided are some pre-computed content descriptors. Table 1 shows that our approach performs better on STM than on LTM. We experimentally found that the regression model has a significant impact on the correlation values. This selection requires further analysis in terms of features as well. Perhaps unsupervised models may reveal more about the nature of the tasks.

Table 1: Memorability task evaluation using Spearman's rank correlation for different models.

Task	Run	Validation	Test
STM	TCNN/FBRN152/3DRN152/LassoLarsCV	0.5149	0.459
	TCNN/FBRN152/3DRN152/LassoCV	0.4987	0.463
	FBRN152/LassoLarsCV	0.4936	0.445
	TCNN/FBRN152/3DRN152/DNN	0.4837	0.436
	TCNN/DN201/3DRN152/LassoLarsCV	0.5185	**0.467**
LTM	TCNN/FBRN152/3DRN152/SVR	0.2394	**0.203**
	TCNN/FBRN152/3DRN152/LassoCV	0.2321	0.185
	TCNN/FBRN152/3DRN152/DNN	0.2104	0.159
	FBRN152/SVR	0.2491	0.189
	DN152/SVR	0.2612	0.196

4 DISCUSSION AND OUTLOOK

From Table 1, we can see that the best regression model is not the same for both tasks. For the STM task, LassoLarsCV achieves the best results while SVR for the LTM task, respectively. Although it is not the same regression model, we achieve the best correlation results for the memorability tasks when fusing all three types of features. It is worth noticing that image-based features achieve the second-best results. Regarding the frame selection criterion, i.e., the middle frame, we observed no significant difference by selecting other frames in the Spearman's rank correlation. This aspect may be linked to the short length of the videos. We can quickly inspect that there is a strong visual relationship between the first and the last frame. Perhaps longer sequences may require more elaborate temporal analysis. Thus, for practical purposes, we prefer to incorporate specific video-designed features. We also verified the PCA effectiveness before the early fusion and by individual feature selection. We observed an improvement c.a. 4-7% in Spearman's rank correlation, thus it is a good practice to project the features into a lower dimensional space before feed the regression model. The proposed method enables us to capture the memorability associated with videos comprising multimedia features. With this in mind, it is possible to create models for similar tasks in video content for other computer vision applications. The memorably test, then, can extrapolate multimedia analysis for other case studies, e.g. video summarization where the scores can be treated as features weights, where, naturally, the features are not necessarily visual.

ACKNOWLEDGMENTS

This study has been funded through an Innovate UK Knowledge Transfer Partnership between Hub Productions Limited and the School of Computer Science & Electronic Engineering, University of Essex, Partnership No: 11071.

REFERENCES

[1] Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. 2011. Comparison of video sequences with histograms of motion patterns. In *2011 18th IEEE International Conference on Image Processing*. IEEE, 3673–3676.

[2] Ritwick Chaudhry, Manoj Kilaru, and Sumit Shekhar. 2018. Show and Recall@ MediaEval 2018 ViMemNet: Predicting Video Memorability. *Group* 1 (2018), G1.

[3] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, and France Rennes. 2018. MediaEval 2018: Predicting Media Memorability Task. *CoRR* abs/1807.01052 (2018). arXiv:1807.01052 http://arxiv.org/abs/1807.01052

[4] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc Q. K. Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. Predicting Media Memorability Task at MediaEval 2019, Sophia Antipolis, France, Oct. 27-29, 2019(2019). *Proc. of MediaEval 2019 Workshop* (2019).

[5] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. 2017. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 971–980.

[6] Ankit Goyal, Naveen Kumar, Tanaya Guha, and Shrikanth S Narayanan. 2016. A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2822–2826.

[7] Rohit Gupta and Kush Motwani. 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features.. In *MediaEval*.

[8] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1633–1640.

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6546–6555.

[10] David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, Vol. 5007. International Society for Optics and Photonics, 87–95.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Tanmayee Joshi, Sarath Sivaprasad, Savita Bhat, and Niranjan Pedanekar. 2018. Multimodal Approach to Predicting Media Memorability.. In *MediaEval*.

[13] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*. 2390–2398.

[14] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[15] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. 3294–3302.

[16] R. Leyva, V. Sanchez, and C. Li. 2019. Compact and Low-Complexity Binary Feature Descriptor and Fisher Vectors for Video Analytics. *IEEE Transactions on Image Processing* 28, 12 (Dec 2019), 6169–6184. https://doi.org/10.1109/TIP.2019.2922826

[17] Yang Liu, Zhonglei Gu, and Tobey H Ko. 2018. Learning Memorability Preserving Subspace for Predicting Media Memorability.. In *MediaEval*.

[18] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2016. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 598–606.

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[20] Ricardo Manhães Savii, Samuel Felipe dos Santos, and Jurandy Almeida. 2018. GIBIS at MediaEval 2018: Predicting Media Memorability Task.. In *MediaEval*.

[21] Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[22] Wensheng Sun and Xu Zhang. 2018. Video Memorability Prediction with Recurrent Neural Networks and Video Titles at the 2018 MediaEval Predicting Media Memorability Task.. In *MediaEval*.

[23] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[25] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. 2018. Predicting Media Memorability Using Deep Features and Recurrent Network. In *MediaEval*.

[26] Shuai Wang, Weiying Wang, Shizhe Chen, and Qin Jin. 2018. RUC at MediaEval 2018: Visual and Textual Features Exploration for Predicting Media Memorability.. In *MediaEval*.