Predicting Media Memorability Using Ensemble Models

David Azcona, Enric Moreu, Feiyan Hu, Tomás E. Ward, Alan F. Smeaton
Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland
david.azcona@dcu.ie

ABSTRACT

Memorability, defined as the quality of being worth remembering, is a pressing issue in media as we struggle to organize and retrieve digital content and make it more useful in our daily lives. The Predicting Media Memorability task in MediaEval 2019 tackles this problem by creating a challenge to automatically predict memorability scores building on the work developed in 2018. Our team **ensembled** transfer learning approaches with video captions using embeddings and our own pre-computed features which **outperformed** Medieval 2018's **state-of-the-art architectures**.

1 INTRODUCTION AND RELATED WORK

The MediaEval Predicting Media Memorability Task [8] focuses on predicting how memorable a video is to viewers. It builds on the work developed in 2018 [6] and requires participants to automatically predict memorability scores for videos reflecting the probability that videos will be remembered. The dataset is composed of soundless short videos each with two scores for memorability that refer to the probability of being remembered after two different durations of memory retention: short-term and long-term (after 24-72 hours). Our team participated in 2018 [22] with a range of approaches including video saliency, neural EEG techniques and visual aesthetics but this year we focus on ensemble methods [1].

Media Memorability has attracted research interest recently in the area of Computer Vision [7, 20, 23]. Recently Convolutional Neural Networks (CNNs) trained on large datasets such as ImageNet performed better at predicting memorability scores than using video captions and pre-computed features [10, 25]. In addition, multimodal approaches with textual descriptors or video captions that use state-of-the-art neural network (such as embeddings) approaches have the potential to increase the effectiveness of these models [15].

2 OUR APPROACH

The memorability dataset is composed of 10,000 videos, an official test set of 2,000 videos and a development set of 8,000 videos. Teams were provided with the development set's labels only. We divided the development set into our own training (7,000 videos) and validation (1,000 videos) sets. We leveraged our held-out validation set to choose hyper-parameters and evaluated the performance of our models. Our team's approach is to develop individual models per set of features extracted and to then combine them using ensemble models. First, we developed traditional Machine Learning and highly regularised linear models (following last year's [10]):

i) Support Vector Regression [3]
ii) Bayesian Ridge Regression (probabilistic model) [2].

Second, highly regularized Deep Learning techniques such as:

i) Embeddings as high level representations for words [16]
ii) Transfer Learning by using neural network activations as feature extractors and fine-tuning our own networks.

We decided to manually extract **8 frames** from each source video, the first frame and one frame after each of the seven seconds of the video. The following are the categories and models we built:

a) **Off the shelf pre-computed features:** extracted by the challenge's organisers [8]. First, video specialised features: **C3D** (101 features per video) and **HMP** (6,075 features per video) as a histogram of motion patterns. Then, image features extracted for three key frames in each video, concatenated into a long vector. Frame features we used are: **LBP**, local texture information; **InceptionV3**, output of the FC7 layer; **Color Histogram**; and **aesthetic** visual features.
b) **Our own pre-computed features:** We incorporated visual **aesthetic** features by fine-tuning all layers in Resnet50 [13] for each of 8 frames. We adapted the code in [1] to extract 7 **emotions** (anger, disgust, fear, happiness, sadness, surprise, neutral), gender scores and spatial information for frames.
c) **Textual information:** We processed annotated video **captions** with a bag-of-words [11] approach using TF-IDF [18] and input those into a linear model. A more modern approach was to utilise Embeddings and Neural Networks. We built the following network architecture: an Embeddings layer (marked as non-trainable) by leveraging Glove's pre-trained embeddings [17] with 300 dimensions followed by a Gated Recurrent Unit (GRU) [4] with very high dropout [24] and, optionally some fully connected layers with ReLU activation [9] and dropout.
d) **Pre-trained CNN as a feature extractor:** using transfer learning and a pre-trained model (on ImageNet), we applied global average pooling to the output of the last convolutional block before the fully-connected layers at the top of the network. The pre-trained models used were: VGG16 (4,096 features) [21], DenseNet121 (8,192) [14], **ResNet50** (16,384) and **ResNet152** (16,384) [12].
e) **Fine-tuning our own CNN:** another type transfer learning where we took a ResNet architecture, removed the old fully-connected layers at the top, added some new ones with a sigmoid at the end and trained the network by unfreezing layers iteratively to predict memorability scores.
f) **Ensemble models:** leveraging the predictions for the individual models, we ran all possible combinations of the weights with replacement using 20 bins of 5% each.

[1] Code developed for this work has been made publicly available as a repository on Github at https://github.com/dazcona/memorability where further details such as how to deploy, models, hyper-parameters and visualizations can be found.

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

3 RESULTS

The first table shows the performance of individual models, the second shows the weights for the 5 runs submitted and the third shows the final scores on our validation and on the official test data.

Model	Validation (1,000 videos)	
	Spearman	
	Short-term	Long-term
Off the shelf pre-computed features		
C3D	0.32241	0.14113
HMP	0.28583	0.10767
LBP	0.29215	0.13291
Color Histogram	0.12846	0.02069
InceptionV3	0.12280	0.01144
Aesthetics	0.25311	0.09517
Our own pre-computed features		
Aesthetics*	0.41875	0.20361
Emotions*	0.13846	0.08780
Textual information		
Captions w/ TF-IDF	0.42294	0.19344
Captions w/ Embeddings	0.49540	0.23655
Pre-trained CNN as feature extractor		
ResNet50	0.50780	0.20801
ResNet152	0.52278	0.21488
Fine-tuning our own CNN		
Network w/ Transfer Learning	0.40256	0.21451

Model	C3D	LBP	Aest*	Emots.*	Capts.	ResNet152
Short-term ensembles						
1	**0**	**0**	**0**	**0.05**	**0.40**	**0.55**
2	0.05	0	0	0	0.40	0.55
3	0	0	0	0.10	0.40	0.50
4	0.05	0.05	0	0.05	0.35	0.50
5	0.05	0.05	0.05	0.05	0.35	0.45
Long-term ensembles						
1	0	0.25	0	0.25	0.30	0.20
2	**0.05**	**0.20**	**0**	**0.25**	**0.30**	**0.20**
3	0	0.35	0	0.25	0.25	0.15
4	0	0.25	0	0.25	0.30	0.15
5	0.05	0.25	0.05	0.20	0.30	0.15

Model	Validation	Test		
	Spearman		Pearson	MSE
Short-term ensembles				
Ensemble 1	**0.55353**	**0.528**	**0.566**	**0**
Ensemble 2	0.55352	0.527	0.566	0
Ensemble 3	0.55314	0.527	0.565	0
Ensemble 4	0.55230	0.526	0.564	0
Ensemble 5	0.55024	0.525	0.563	0
Long-term ensembles				
Ensemble 1	0.27322	0.269	0.299	0.02
Ensemble 2	**0.27294**	**0.27**	**0.3**	**0.02**
Ensemble 3	0.27285	0.261	0.293	0.02
Ensemble 4	0.27246	0.265	0.298	0.02
Ensemble 5	0.27198	0.266	0.299	0.02

4 DISCUSSION AND OUTLOOK

Our findings and contributions to this area are the following:

i) Deep Learning CNN models will typically outperform models trained with captions and other visual features for short-term memorability; however, techniques such as embeddings and recurrent networks can achieve very high results for captions.

ii) We believe fine-tuned CNN models will outperform pre-trained models as feature extractors given enough training samples and iterations although we could not prove that in this paper.

iii) Ensembling models by using predictions instead of training models with very long vectors of features is an alternative we used to counteract memory limitations.

iv) Ensembling models with different modalities such as emotions with captions, high-level representations from CNNs and visual pre-computed features achieve the best results as they represent different high-level abstractions.

In addition, we used a visualisation called class activation map, useful for understanding which parts of an image led a CNN to its final classification decision [5, 19]. Figure 2 shows ResNet152 (trained with ImageNet) was leveraged for most memorable videos.

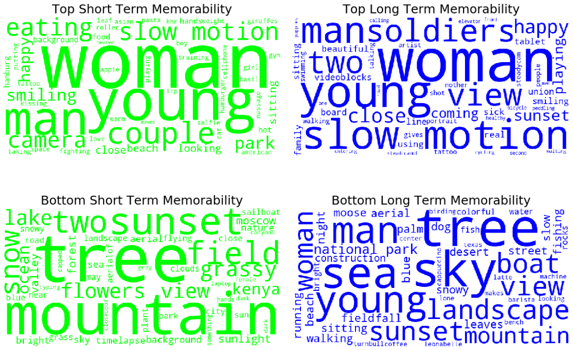
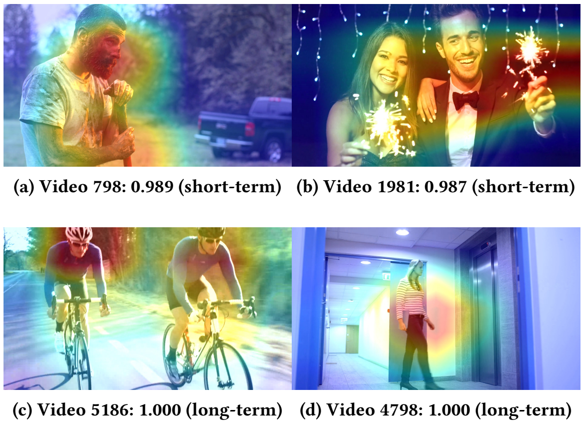


Figure 1: Captions for Most & Least Memorable Videos



(a) Video 798: 0.989 (short-term) (b) Video 1981: 0.987 (short-term)

(c) Video 5186: 1.000 (long-term) (d) Video 4798: 1.000 (long-term)

Figure 2: Class Activation Maps for Most Memorable Videos

ACKNOWLEDGMENTS

The work in this paper is supported by grant number SFI/12/RC/2289. We thank Aishwarya Gupta, Cristiano Moretti, Kevin Callan, Kevin McGuinness and M.SC. students in "CA684: Machine Learning" in 2018/9 at DCU.

REFERENCES

[1] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557* (2017).

[2] Christopher M Bishop. 2006. *Pattern recognition and machine learning.* springer.

[3] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[5] Francois Chollet. 2018. *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek.* MITP-Verlags GmbH & Co. KG.

[6] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. Mediaeval 2018: Predicting media memorability task. *arXiv preprint arXiv:1807.01052* (2018).

[7] Romain Cohendet, Karthik Yadati, Ngoc QK 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.* ACM, 178–186.

[8] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc Q. K. Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. The Predicting Media Memorability Task at Mediaeval 2019. (2019).

[9] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics.* 315–323.

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

[11] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[12] 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.

[13] Feiyan Hu and Alan F Smeaton. 2018. Image aesthetics and content in selecting memorable keyframes from lifelogs. In *International Conference on Multimedia Modeling.* Springer, 608–619.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4700–4708.

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

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems.* 3111–3119.

[17] 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

[18] Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets.* Cambridge University Press.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision.* 618–626.

[20] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision.* 2730–2739.

[21] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[22] Alan F Smeaton, Owen Corrigan, Paul Dockree, Cathal Gurrin, Graham Healy, Feiyan Hu, Kevin McGuinness, Eva Mohedano, and Tomás E Ward. 2018. Dublin's participation in the predicting media memorability task at MediaEval 2018. (2018).

[23] Hammad Squalli-Houssaini, Ngoc QK Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2018. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2371–2375.

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

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