Linear Models for Video Memorability Prediction Using Visual and Semantic Features

Rohit Gupta, Kush Motwani
Conduent Labs, India
rohit.gupta@conduent.com,kush.motwani@conduent.com

ABSTRACT

Memorability is defined as the quality or state of being easy to remember. In the context of videos today, with billions of hours of user generated video content on online platforms like social media, prediction of a cognitive measure like memorability has many potential applications. We investigate the use of various visual and semantic features in building models for video memorability prediction. Along with the features provided as part of the Predicting Media Memorability Task at MediaEval 2018, we utilize generic descriptors extracted from the feature map of Convolutional Neural Networks. We are able to discover intrinsic semantic factors affecting memorability based on our analysis of models that use video captions. Finally, we are able to build an ensemble of models that outperforms models built using a single feature type.

1 INTRODUCTION

In this work, we investigate the use of various visual and semantic features to predict video memorability, and conduct an extensive analysis over the chosen features, to develop a robust video memorability predictor. Among the features provided by challenge organizers [1] we train models over the HMP, LBP and ColorHistogram visual features and InceptionV3-Predictions & C3D-Predictions semantic features. In addition, we train models over the video captions provided, and over features extracted from the last convolution layers of Convolutional Neural Networks trained for image classification [6] [5] [8] applied to frames of the video. The models are evaluated using Spearman's rank correlation as the metric. Our key findings and contributions are as follows:

(1) Models for short term memorability perform much better than models for long term memorability.
(2) Models based on InceptionV3-Preds, LBP and ColorHistogram do not work well and are outperformed by those based on C3D-Preds and HMP.
(3) Models based on the video captions provided outperform models trained on all above mentioned visual features.
(4) Analysis of models trained on video captions allows us to identify semantic factors affecting video memorability.
(5) We find that models based on high level representations learned by CNNs trained on image classification tasks outperform both models based on visual features and captions.

2 RELATED WORK

Work on video memorability has recently begun to generate a lot of interest, and recent works [2] [9] investigate the use of various low and high level visual features, deep learning based action recognition representations (C3D-Preds), and image and video captions for memorability prediction. The major findings on memorability from these papers are that models using captions give the best individual results, and how features learned using deep learning improve those results. Additionally, researchers have found that high level semantic features learned by CNNs trained for image classification achieve state of the art performance on a variety of computer vision tasks [8].

3 APPROACH

3.1 Models

Since most of the features provided are very high dimensional and the number of videos is of the same order of magnitude as the dimensionality of the features, high variance and over-fitting are a major potential concern in this task. As a result we preferred simple, linear, highly regularized models, namely:

(1) LASSO (L1) regularized Logistic Regression [10]
(2) Linear Support Vector Regression [3]
(3) ElasticNet (L1 and L2 Regularized Linear Regression) [4]

For each set of features we tried each of these 3 models and select the best one. Values for various model hyper-parameters controlling the extent of regularization (such as C, alpha and L1-Ratio) were picked by using grid search over the dev set.

In order to improve prediction accuracy, we build ensemble models using some of our best models. We use a simple weighted averaging technique, where we blend the outputs of the best single models developed previously. Weights for ensemble averaging are also picked using grid search over the dev set.

3.2 Features and Data Pre-Processing

Video level features like HMP and C3D-Preds are used as-is, while frame level features such as ColorHistogram and LBP are concatenated across frames. We pre-process the text captions into Bag-of-Words features using CountVectorizer from Scikit-Learn. We use word unigrams and bigrams, remove English stopwords and pick the optimal vocabulary size by cross-validation.

ResNet and DenseNet features are extracted for the 1st, 56th, 112th frames of the video from the penultimate layer of ResNet50 and DenseNet121 models respectively. These features are then averaged across frames and L2-Normalized. This results in a 1024 and 2048 dimensional feature vector for DenseNet and ResNet respectively.

The memorability scores is pre-processed in 2 different ways depending on the model:

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

- Z-score Normalization for Regressors: carried out in order to make the scores scale-invariant, this results in a significant increase in model accuracy.
- Sampling for Logistic Regression Classifier: We train the Logistic Regression Classifier using binary labels sampled from a binomial distribution parameterized by the memorability scores. This is motivated by the fact that the memorability score is the percentage of subjects who can recall the contents of the video [2].

4 RESULTS

Tables 1 and 2 give an overall summary of our experimental results. Results for the best model for each feature are presented. For HMP, LBP and ColorHistogram, Lasso Logistic Regression is the best model, while ElasticNet is the best model for the other features. For ensembling, we discover the optimal weights are as follows:

(1) **Ensemble1:** 0.25*Caption score + 0.75*ResNet score
(2) **Ensemble2:** 0.1*HMP score + 0.1*Caption score + 0.8*ResNet score

Table 1: Long Term Memorability Scores

Model	Validation	Test		
	Spearman		Pearson	MSE
C3D-Preds	0.153			
HMP	0.136			
InceptionV3-Preds	0.092			
LBP	0.098			
Color Histogram	0.048			
Captions	0.208	0.213	0.228	0.0194
ResNet	**0.261**	0.234	0.259	0.0190
DenseNet	0.259	**0.245**	**0.269**	**0.0189**
Ensemble1	0.235	**0.253**	**0.278**	**0.0188**
Ensemble2	0.257	0.244	0.268	0.0189

Table 2: Short Term Memorability Scores

Model	Validation	Test		
	Spearman		Pearson	MSE
C3D-Preds	0.311			
HMP	0.281			
InceptionV3-Preds	0.150			
LBP	0.254			
Color Histogram	0.105			
Captions	0.438	0.402	0.408	0.00560
ResNet	**0.501**	**0.488**	**0.524**	**0.00487**
DenseNet	0.491	0.473	0.504	0.00500
Ensemble1	0.484	**0.497**	**0.530**	0.00488
Ensemble2	**0.508**	0.495	0.528	0.00489

5 ANALYSIS AND DISCUSSION

To analyze our interpretable captions based model we look at vocabulary terms corresponding to the most positive (Figure 1) and negative (Figure 2) coefficients of the caption based models (averaged over 100 models using different validation splits). This reveals

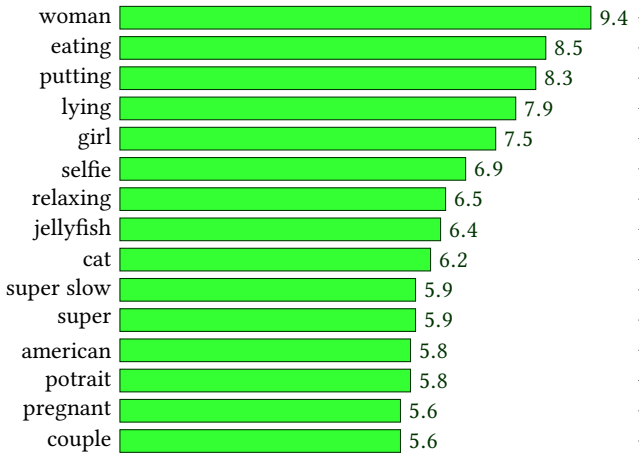

Figure 1: Terms for the most positive coefficients

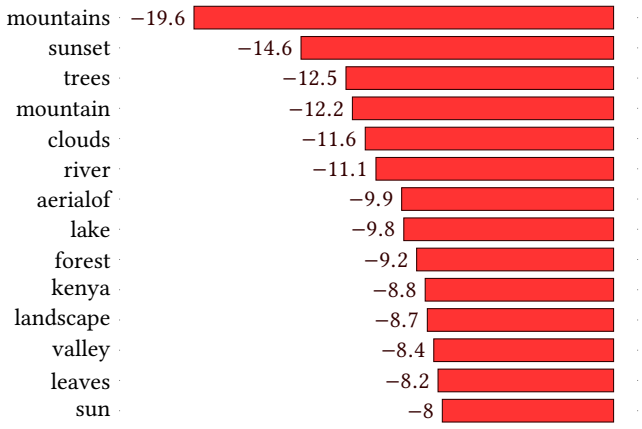

Figure 2: Terms for the most negative coefficients

that the highest negative coefficients are all dominated by terms related to natural scenery; whereas the highest positive coefficients are dominated by terms related to people and indoor actions & objects. Contrary to intuition, videos of nature are not memorable, a result that has also been observed in the context of image memorability. [7]

Like in past work on video memorability, our caption based models give superior performance than the C3D features, and low-level visual features. In contrast to past work however, we also examine the use of features extracted from the penultimate layer of CNNs and observe that models trained on these features outperform the models based on captions. Coupled with the fact that models trained over InceptionV3 predictions give poor results, we infer that the representation learned by CNNs capture additional semantic information relevant to predicting memorability beyond simply the category the image belongs to.

REFERENCES

[1] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q.K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. MediaEval 2018: Predicting Media Memorability Task. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[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*. ACM, 178–186.

[3] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. 2005. Working Set Selection Using Second Order Information for Training Support Vector Machines. *Journal of Machine Learning Research* 6, Dec (2005), 1889–1918.

[4] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33, 1 (2010), 1.

[5] 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.

[6] 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*.

[7] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1469–1482.

[8] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[9] Sumit Shekhar, Dhruv Singal, and Harvineet Singh. 2017. Show and Recall: Learning What Makes Videos Memorable. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2730–2739.

[10] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. 2012. An Improved GLMNET for L1-regularized Logistic Regression. *Journal of Machine Learning Research* 13, Jun (2012), 1999–2030.