Video Memorability - Sentiment Analysis, Video Motion and Gestalt’s Principle

Joel Peter Henry Rozario
Msc in Computing, Data Analytics
joel.henryrozario2@mail.dcu.ie

ABSTRACT

This paper describes the 3 different approaches adopted to predict the long-term and short-term video memorability scores for the given videos. This predictability task is based on three propositions: Gestalt’s principle on Visual Memory, Frames with Human/object motion, and by analyzing the impact of sentimental analysis on captions using NLTK VADER. Ensemble techniques, XGBoost and Random Forest, and Artificial Neural Network - Multi-layer Perceptron were used to predict video memorability scores.

1. INTRODUCTION

Even though we have managed to identify near-human consistency features that impact image memorability [1], a human’s ability to memorize or forget visual video content is still beyond comprehension. As per a study by Nielsen Norman Group, ‘visual content has witnessed a 56.3% YoY increase in 2019’. Thus, it is crucial to find factors that make a visual content interesting or desirable to the human mind.

This paper approaches the memorability task by focusing on 3 propositions that influence the neuro and psychological factors in making a video memorable [5].

1. It has been studied that, emotional content is an influential factor in cognitive processing which includes perception, memory, and attention [2].
2. Secondly, humans have shown a tendency to better memorize videos with human or object motion than a static image. [3]
3. Furthermore, this paper has extended a segment of work [5] on gestalt’s principle in predicting video memorability Dublin University as four Gestalt’s principles: similarity, connectedness, proximity, and common region are reported to have a significant bearing on Visual Working Memory [4].

2. RELATED WORK

A research done on the impact of visual features [6] [8] shows high-level features extracted by CNN, to have a better impact on predictability scores. Additionally, encoded captions gave a higher score for both long-term and short-term videos comparatively [6][7].

Emotions can either impair or enhance our short-term and long-term memory retention.[2] There have been substantial reports that emotions can trigger longer retention of episodes thus enhancing semantic and visual working memory. A paper by Rohit Gupta et. al [7] reports that high memorability score videos are dominated by semantics including a living presence or an activity, whereas low memorability scores can be associated with natural scenery and negative words.

3. APPROACH

3.1 Sentiment Analysis of Caption

To implement the first proposition, Stanford’s Natural Language Toolkit (NLTK) VADER [9] is used to perform sentiment analysis on the captions. To reduce the morphological variation, each token was lemmatized and further cleaned and only the compound feature was used, which is an overall descriptor of the sentiment analyzer, while the positive, negative, and neutral descriptors were removed due to high multicollinearity. Three models were run with the VADER descriptor.

- MLP with One hot encoded caption – captions were tokenized and padded
- MLP with TF-IDF of captions to find the importance of keywords [7]
- An embedded Neural Network with Word embeddings – A vector representation using Global Vectors (GloVe)

MLP architecture is designed with 3 layers, with Rectified Linear Unit (ReLu) as the activation function which overcomes the vanishing gradient drawback in sigmoid. L2 Regularization, Ridge Regression, is applied to the activation function on the input and hidden layers to reduce overfitting of the training data. Adamax (Adaptive Movement Estimation) is used as data is encoded and sparse in nature.

3.2 Predicting with living presence or motion

High-level features such as Histogram of Moving Pattern's and C3D were used as it reveals key information on moving objects and object detection. Ensemble learning techniques such as Random Forest Regressor with Decision Trees and XGBoost is used to predict both the short-term and long-term memorability. A test was performed with C3D and HMP's of the middle frame independently and then a combined test was performed to evaluate the second proposition. Due to the high dimensionality of HMP, PCA was performed to reduce the dimensions. Finally, a grid search was performed to fine-tune the parameters.

3.3 Applying Gestalt's Principle – Aesthetics and Color Histogram.

By extending the work done by Dublin University [5], gestalt's principle is applied by using Aesthetics and Color Histogram independently and as a combination. Aesthetics features were normalized and like the previous step, Ensemble learning – Random Forest and XGBoost are preferred. A grid search was performed to find the best parameters

4. RESULTS AND ANALYSIS

This paper presents the best results acquired by running models based on all 3 propositions. 20% of the data is used as the validation set to predict the short-term and long-term memorability. The evaluation metric is Spearman's Correlation and Mean Squared Error as the loss function.

Table 1: Short-term memorability scores

Model	***Spearman Score***	***MSE***
Gestalt's Principles		
Aesthetics	0.318	0.005533
Color Histogram	0.211	0.005939
Aesthetics + color Histogram	0.289	0.005702
Principle of Human Presence or Movement		
C3d	0.332	0.00547
C3d + HMP	0.308	0.005594
HMP	0.302	0.005597
Sentiment Analysis with NLTK VADER		
Encoded + VADER	***0.456***	***0.0049***
TF-IDF + VADER	0.454	0.00505
GloVe + VADER	0.376	0.005238
MediaEval Benchmark 2019	**0.528**	

Sentiment Analysis with One hot encoding of captions yielded the best results for short-term memorability with 3-layer Multi-layer Perceptron. The figure below shows a steady increase in accuracy for each epoch. Callback function was used to determine the optimum number of epochs to avoid overfitting. Aesthetics had a decent memorability score compared to other independent features implying the significance of gestalt's principle on visual memory [4].

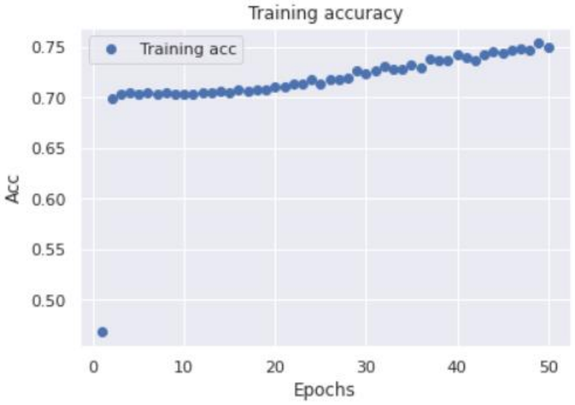

Figure 1: showing accuracy increase over each epoch.

Sentiment analysis with captions running on a MLP yielded better score relatively but with TF-IDF vectorization.

Table 2: long-term memorability scores

Model	***Spearman Score***	***MSE***
Gestalt's Principles		
Aesthetics	0.144	0.0213329
Color Histogram	0.104	0.0215842
Aesthetics + color Histogram	0.139	0.0213318
Principle of Human Presence or Movement		
C3d	0.145	0.0213791
C3d + HMP	0.13	0.0214571
HMP	0.018	0.0215302
Sentiment Analysis with NLTK VADER +		
Encoded + VADER	0.215(RFR)	0.0243405
TF-IDF + VADER	***0.229***	***0.0211602***
GloVe + VADER	0.166	0.0282383
MediaEval Benchmark 2019	**0.277**	

The low scores in both the task for high-level features might be due to the high dimensionality of the data. HMP with Principal Component Analysis gave less score even after retaining 95% variance. Furthermore, with better captions for videos, much efficient sentiment analysis can be performed. Another shortcoming of this paper is the failure to include the weightage of substantial words as studied by Rohit Gupta et. al [7] For example, extra weights to words involving people, activities, and so on.

5. DISCUSSION AND OUTLOOK

Captions with sentiment analysis performed considerably better than the other two propositions on gestalt's and Human/object motion. With a customized sentiment analysis, including additional weights to impactful words such as people and animals, a much better score can be achieved with a Neural Network. Regularized Neural Networks- Multi-Layer Perceptron performed better than

Ensemble techniques including both Random Forest and XGBoost.

REFERENCES

[1] E. A. Khosla, A. S. Raju, A. Torralba and A. Oliva, "Understanding and Predicting Image Memorability at a Large Scale," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 2390-2398.

[2] Tyng, C. M., Amin, H. U., Saad, M., & Malik, A. S. (2017). The Influences of Emotion on Learning and Memory. *Frontiers in psychology*, *8*, 1454. https://doi.org/10.3389/fpsyg.2017.01454Tran-Van, D., Tran, L., & Tran, M. (2018). Predicting Media Memorability Using Deep Features and Recurrent Network. *MediaEval*.

[3] Larson, Martha and Arora, Piyush and Demarty, Claire-Hélène and Riegler, Michael and Bischke, Benjamin and Dellandrea, Emmanuel and Lux, Mathias and Porter, Alastair and Jones, Gareth J.F., (eds.) *Working Notes Proceedings of the MediaEval 2018 Workshop*. CEUR-WS Proceedings.

[4] Peterson, Dwight J, and Marian E Berryhill. “The Gestalt principle of similarity benefits visual working memory.” *Psychonomic bulletin & review* vol. 20,6 (2013): 1282-9. doi:10.3758/s13423-013-0460-x.

[5] Larson, Martha and Arora, Piyush and Demarty, Claire-Hélène and Riegler, Michael and Bischke, Benjamin and Dellandrea, Emmanuel and Lux, Mathias and Porter, Alastair and Jones, Gareth J.F., (eds.) *Working Notes Proceedings of the MediaEval 2018 Workshop*. CEUR-WS Proceedings . CEUR-WS.

[6] Joshi, T., Sivaprasad, S., Bhat, S., & Pedanekar, N. (2018). Multimodal Approach to Predicting Media Memorability. *MediaEval*.

[7] Gupta, R., & Motwani, K. (2018). Linear Models for Video Memorability Prediction Using Visual and Semantic Features. *MediaEval.*

[8] Tran-Van, D., Tran, L., & Tran, M. (2018). Predicting Media Memorability Using Deep Features and Recurrent Network. *MediaEval*.

[9] Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.