Deep learning for predicting image memorability

Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaëlle, Claire-Hélène Demarty

DEEP LEARNING FOR PREDICTING IMAGE MEMORABILITY

Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaëlle and Claire-Hélène Demarty

Technicolor
975 Avenue des Champs Blancs, 35576 Cesson-Sévigné, France
Email: hammad.squalli-houssaini@imt-atlantique.net
{quang-khanh-ngoc.duong, gwenaelle.marquant, claire-helene.demarty}@technicolor.com

ABSTRACT

Memorability of media content such as images and videos has recently become an important research subject in computer vision. This paper presents our computation model for predicting image memorability, which is based on a deep learning architecture designed for a classification task. We exploit the use of both convolutional neural network (CNN) - based visual features and semantic features related to image captioning for the task. We train and test our model on the large-scale benchmarking memorability dataset: LaMem. Experiment result shows that the proposed computational model obtains better prediction performance than the state of the art, and even outperforms human consistency. We further investigate the genericity of our model on other memorability datasets. Finally, by validating the model on interestingness datasets, we reconfirm the uncorrelation between memorability and interestingness of images.

Index Terms— Image memorability, computational model, deep learning, interestingness, image captioning

1. INTRODUCTION

In our fast moving world, media platforms such as social networks, media advertisement, information retrieval and recommendation systems need a great power to deal with exponentially growing data day after day. Thus the ability to understand the content plays a key role in such media systems to help them optimize their processing. Different concepts such as visual saliency and aesthetics [1, 2], emotion [3], social popularity [4], interestingness [5–7], and memorability [8, 9] may intervene in the understanding of content. This paper focuses on memorability of images, an emerging and less investigated concept in multimedia and computer vision.

While image memorability has been only considered recently in computer vision, visual memory has been largely studied in psychology since decades. As examples, with experiments on memory capacity psychologists showed that people can remember thousands of pictures they saw only once, even when exposed to many other similar images afterward [10, 11]. However, different images can be more or less remembered depending on numerous factors concerning both intrisic visual appearance and user's context [8]. It is important to note that even if memorability is likely to be influenced by the user context, there is a general agreement (concensus) between groups of people on judging a certain image [12]. Such key finding grounds and suggests for further studies in quantifying how memorable a certain image is for a common observer.

In computer vision, researchers have investigated intrinsic and extrinsic characteristics that make an image memorable [9, 13, 14] and a photograph memorable [15]. It reveals that color, simple image features derived from pixel statistics, and object statistics (number of objects, log mean pixel coverage over present object classes, *etc.*,) do not have strong correlation with memorability. Meanwhile, object and scene semantics offer high correlation with memorability. Other subjective concepts such as aesthetics and interestingness have also been shown to be uncorrelated with memorability. Importantly, experiments in [9,16] have confirmed a sufficient human consistency while they annotated image memory, and thus prove the feasibility of predicting memorability of images. To support research in the field, several small datasets have been publicly released such as those concerning face photographes [17], scene categories [14, 15], visualization pictures [18], and affective impact on image memorability [19, 20]. Especially, thanks to some research at MIT, a first large-scale image memorability dataset (LaMem) contaning roughly 60,000 images annotated by crowdsourcing has been published together with a memorability prediction model (MemNet) for benchmarking the task [16].

In this paper, starting from LaMem, we develop a computational model for predicting image memorability scores and we show that the proposed model obtains better prediction performance than the state-of-the-art MemNet model. It even exceeds human consistency on LaMem. Our computational model differs from existing ones in two main points. First, instead of naturally treating the prediction as a regression problem as in [16], we treat it as a classification problem and aggregate the predicted probabilities of belonging to different classes (corresponding to different degrees of memorability) for the final memorability score. Second, unlike most

of existing works, which used either handcrafted low-level image features [9,21], or features extracted from a pre-trained CNN [22], we propose to exploit additional semantic features related to image captioning to improve the prediction.

The rest of the paper is organized as follows. Section 2 presents the proposed computational model for image memorability prediction. Experimental results are shown in Section 3, including our further study on the genericity of the model and the correlation between the interestingness and memorability concepts. Finally, we conclude in Section 4.

2. PROPOSED COMPUTATIONAL MODEL

The workflow of the investigated prediction systems, which are based on either a regression model (as a baseline) or on a classification model (proposed), is presented in Fig. 1. The feature extraction step and the proposed classification model will be presented in subsections 2.1 and 2.2, respectively.

2.1. Feature extraction

As CNN offers a powerful presentation and has been widely used in the literature for many different tasks, we used the well-known VGG16 network [23] pre-trained on the ImageNet dataset for the extraction of a first image feature. The feature is extracted from the last fully-connected layer before the softmax and has a dimension of 4096.

As scene semantics and high-level visual attributes (such as emotions, actions, movements, appearance of objects, *etc.*,) have been shown to well characterize the memorability of a photo [15, 16], we further investigated the use of some other semantic features derived from an image captioning (IC) system [24]. Such an IC model builds an encoder comprising a CNN and a long short-term memory recurrent network (LSTM) for learning a joint image-text embedding. Thus the CNN image feature and the word2vec representation of the image caption are projected on a 2D embedding space which enforces the alignment between an image and its corresponding semantic caption. We extracted such projected CNN feature as logically it should contain some semantic information expected to be relevant for the prediction of image memorability. This feature has a dimension of 1024.

2.2. Classification model

In order to treat the prediction as a classification task, we first need to convert the ground-truth memorability scores provided with LaMem into K class labels. To ease the classification step, the score thresholds were chosen so that the resulting classes are balanced in terms of number of samples. As an example, for $K = 2$ we split LaMem into two classes of roughly 30,000 images each (for $K = 4$, each class contains roughly 15,000 images). Note that the more classes we have after splitting, the narrower the score range in each class and the smaller the number of images per class. Thus the increase of the number of classes is at the expense of training a good classifier. We tested with different values of K in $\{2, .., 8\}$ and found that $K = 4$ offers the best performance.

Our general classification model contains a branch (one or several MLP layers) for the CNNs, a branch for the IC-based semantic features, a merge layer to concatenate both of them followed by some MLP layers with a softmax on top. As the

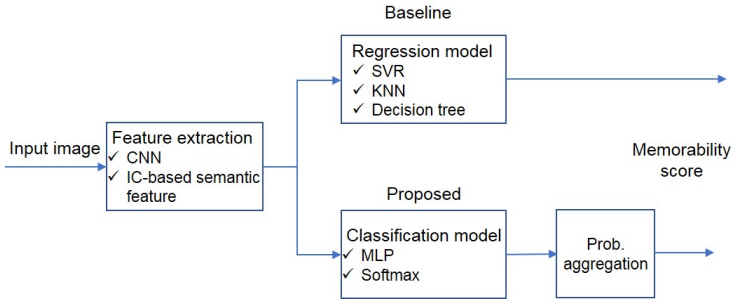

Fig. 1: Workflow of the investigated computational models for image memorability prediction.

softmax gives, in addition to the class label, the probabilities $p(k|i) \in [0, 1]$ that an input image i belongs to each class $k = 1, ..., K$, we take into account these values to compute the final predicted memorability score m_i, in a final probability aggregation step, as

$$m_i = \sum_{k=1}^{K} p(k|i) s_k, \quad (1)$$

where s_k is a threshold value derived from the ground-truth scores of LaMem that separates class k from class $k + 1$, and $s_K = 1$. With $K = 4$, these threshold values are 0.68, 0.77, 0.85, and 1, respectively.

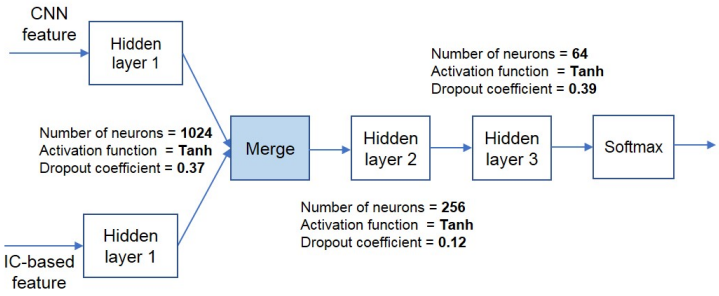

Fig. 2: Best predicting model in terms of average Spearman correlation over a 5-fold validation process.

We performed some hyper-parameter selection thanks to the use of the Bayesian optimization library Hyperas [1]. The selection was done over the following set of parameters: number of neurons per layer (from 32 to 1024, depending on the layers), dropout coefficient, activation function (ReLU, tanh) and optimizer (adam, rmsprop, sgd) so as to maximize the average Spearman correlation coefficient between the predicted scores and the ground-truth scores in a 5-fold validation process. Before that, the number of layers (3), batch size (16) and number of epochs (10) were manually set once again while targeting the maximisation of the Spearman correlation coef-

[1] https://github.com/maxpumperla/hyperas

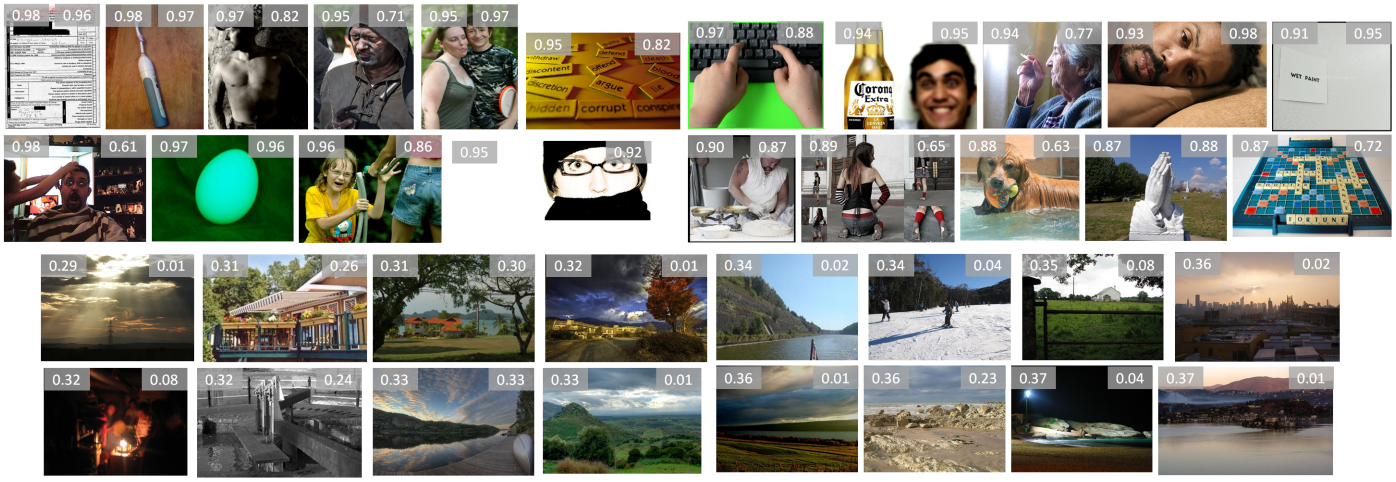

Fig. 3: Example images from LaMem together with their ground-truth values (top left corner) and their predicted scores by our best model (top right corner). Predicted score values were normalized between 0 and 1, after probability aggregation.

ficient. Fig. 2 depicts our best model that offers an average Spearman correlation of 0.72.

3. EXPERIMENT AND INVESTIGATION

Section 3.1 presents the memorability prediction results on LaMem. We investigate the genericity of the model on other memorability datasets in Section 3.2, and the uncorrelation between memorability and interestingness in Section 3.3.

3.1. Prediction results on Lamem dataset

We evaluated our memorability prediction performance on LaMem [16] with the use of different features and different prediction models. The obtained results, measured by the mean Spearman correlation over a 5-fold validation process, are presented in Table 1. As a baseline investigation (see Fig. 1), we tested the use of CNN and IC-based semantic features with different regression methods (support vector regression (SVR), KNN, Decision Tree) where the hyper-parameters for each model were optimized by the Bayesian optimization algorithm implemented in the Hyperopt[2] library. We found that SVR performs better than other regression techniques (thus we mention only SVR in Table 1), and offers similar performance compared to the proposed MLP 2-class model and the state-of-the-art MemNet model. Fig. 3 shows LaMem example images of both some most and least memorable images, together with their predicted scores for our best 4-class model, and their ground-truth values.

As it can be seen, IC-based semantic features, when used with either SVR or MLP, offer better prediction performance than conventional CNN features even though their dimension is 4 times smaller. This reveals that the scene semantics is very important for the task of memorizing images and thus such attributes help better predicting the memorability. Overall, the combination of CNN and IC-based features offers the best mean Spearman correlation (0.72), higher than MemNet (0.64) and human consistency (0.68) on Lamem [16]. This result confirms the complementary between IC-based semantic features and CNNs in predicting image memorability.

Feature	Prediction model	mean Spearman correlation
CNN	SVR	0.64
CNN	MLP, 2 classes	0.64
CNN	MLP, 4 classes	0.65
IC-based feature	SVR	0.65
IC-based feature	MLP, 4 classes	0.67
CNN + IC-based feature	**MLP, 4 classes**	**0.72**
Hybrid-CNN (MemNet)	*SVR*	*0.64*

Table 1: Prediction results on LaMem dataset obtained by different methods.

3.2. Genericity of the model on other memorability datasets

In order to investigate the genericity of our computational model, we tested it on several existing memorability datasets. This also helped us understand more about the differences in content between these datasets. For each of them we computed the Spearman correlation between the provided ground-truth scores and the predicted ones by our model. Example images for each dataset are shown in Fig. 4 and the obtained Spearman correlation values, together with information about the datasets (*i.e.,* number of images in each dataset and type of images), are detailed in Table 2.

As expected, our model generalizes quite good to Isola's [15] and Bylinskii's [14] datasets with a Spearman correlation of 0.59 and 0.48, respectively. This can be explained by the

[2]https://github.com/hyperopt/hyperopt

fact that these two datasets contain images related to scene categories, a topic quite close to the one of LaMem on which we trained our model as shown in Fig. 4. Our model performs poorly on two datasets containing specific images of human faces [17], and visualizations [18]. This is understandable as images in these datasets are very different from those in LaMem, again as presented in Fig. 4. For datasets containing images dedicated to the rating of emotions, our model performs reasonably on Cohendet's [20] while poorly on Libkuman's [19] dataset. Actually, the images in [20] were also in Libkuman's dataset, however the former was carefully annotated through lab experiments, leading to the conclusion that the ground-truth scores may be more reliable.

Dataset's first author	Size	Type of images	Spearman correlation
Isola [15]	2222	Scene categories	0.59
Bylinskii [14]	1754	Scene categories	0.48
Dubey [25]	850	Object segmentation	0.36
Cohendet [20]	150	Rating of emotions	0.39
Libkuman [19]	703	Rating of emotions	-0.02
Bainbridge [17]	2222	Face pictures	-0.06
Borkin [18]	410	Visualizations	-0,27

Table 2: Prediction results obtained by the proposed computational model on other memorability datasets.

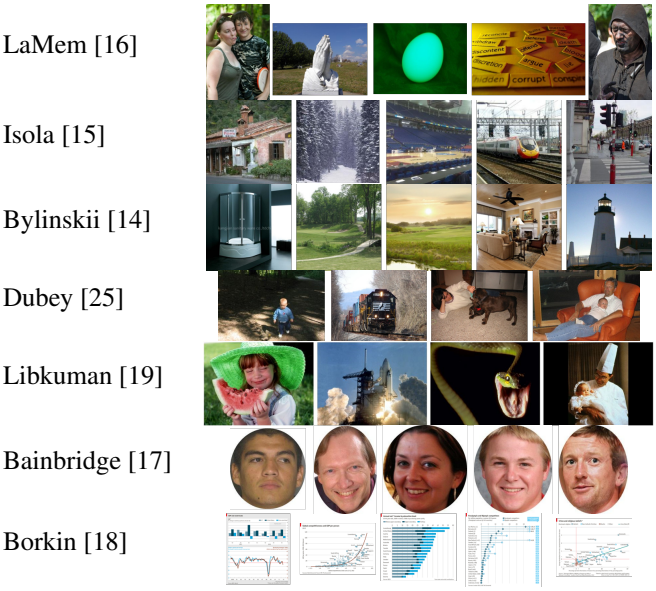

Fig. 4: Example images from different memorability datasets.

3.3. Correlation between memorability and interestingness

As memorability and interestingness are two emerging and important concepts for understanding images, we also investigated their correlation by applying our computational model for memorability prediction on existing interestingness datasets and again computing the Spearman correlation between the predicted scores and the ground-truth interestingness scores. The results are shown in Table 3, together with information about the tested interestingness datasets. Note that among these datasets, Flickr dataset was collected from Flickr Interestingness API [3] so the collected interestingness scores are somewhat *social-driven*. Meanwhile other datasets (*e.g.,* MediaEval and Gygli) provides *content-driven* interestingness as the labels were annotated by human annotators.

Dataset	Size	Type of images	Spearman correlation
MediaEval [7]	7396	Movie frames	0.07
Flickr [26]	123185	Social network	-0.05
Gygli 1 [5]	2688	Scene categories	-0.21
Gygli 2 [5, 15]	2222	Scene categories	-0.19
Gygli 3 [5]	3180	Webcam pictures	0.01

Table 3: Prediction results obtained by the proposed computational model on interestingness datasets.

As detailed in Table 3, all Spearman correlation values are very low, meaning that our memorability prediction model can not be used for interestingness prediction, even for datasets [5, 15] containing the same type of scene category images. This is even more visible when one focuses on Isola's dataset [15] for which ground-truth labels are available for both memorability and interestingness: our model is doing well for memorability prediction (Spearman correlation $= 0.59$, see Table 2), while it fails to predict interestingess. These results re-confirm the uncorrelation between memorability and interestingness that was already stated in the literature.

4. CONCLUSION

Focusing on image memorability, an emerging and important concept for high-level understanding of images, we present a deep learning-based computational model for predicting such memorability levels. Experiment results on the benchmarking LaMem dataset prove that the proposed system outperforms both the state of the art and human consistency on the task. We also investigated the genericity of our model on other memorability datasets and further re-confirmed the uncorrelation between image memorability and interestingness, another important concept. As future work, we might want to focus on image editing and investigate how such manipulations affect the image memorability of content.

5. REFERENCES

[1] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Percept.*, vol. 7, no. 1, pp. 1–39, Jan. 2010.

[3] https://www.flickr.com/services/api/flickr.interestingness.getList.html

[2] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. of ACM International Conference on Multimedia (MM), Florence, IT*, 2010, pp. 271–280.

[3] U. Rimmele, L. Davachi, R. Petrov, S. Dougal, and E. A. Phelps, "Emotion enhances the subjective feeling of remembering, despite lower accuracy for contextual details," *Psychology Association*, 2011.

[4] A. Khosla, A. Sarma, and R. Hamid, "What makes an image popular?," in *Proc. International conference on World Wide Web*, 2013, pp. 867–876.

[5] M. Gygli, H. Grabner, H. Riemenschneider, F. Fabian, and L. Van Gool, "The interestingness of images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1633–1640.

[6] C.-H. Demarty, M. Sjöberg, G. Constantin, N. Q. K. Duong, B. Ionescu, T. T. Do, and H. Wang, "Predicting interestingness of visual content," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Springer, 2017.

[7] C.-H. Demarty, M. Sjöberg, B. Ionescu, T. T. Do, M Gygli, and N. Q. K. Duong, "Mediaeval 2017 predicting media interestingness task," *MediaEval 2017 Workshop*, September 2017.

[8] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, "Visual long-term memory has a massive storage capacity for object details," in *Proc. Natl Acad Sci*, 2008.

[9] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 145–152.

[10] L. Standing, "Learning 10000 pictures," *Quarterly Journal of Experimental Psychology*, vol. 25, no. 2, 1973.

[11] T. Konkle, TF. Brady, GA. Alvarez, and A. Oliva, "Scene memory is more detailed than you think: The role of categories in visual long-term memory," *Psychological Science*, vol. 21, no. 11, pp. 1551–1556, 2010.

[12] R. R. Hunt and J. B. Worthen, *Distinctiveness and Memory*, Oxford University Press, 2006.

[13] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," in *Advances in Neural Information Processing Systems*, 2011.

[14] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralbaa, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision*, 2015.

[15] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1469–1482, 2014.

[16] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva, "Understanding and predicting image memorability at a large scale," in *International Conference on Computer Vision (ICCV)*, 2015.

[17] W. A. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs.," *Journal of experimental psychology. General*, pp. 1323–34, 2013.

[18] A. B. Michelle, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister, "What makes a visualization memorable?," *IEEE Transactions on Visualization & Computer Graphics*, vol. 19, no. 12, pp. 2306–2315, 2013.

[19] T. M. Libkuman, H. Otani, R. Kern, S. G. Viger, and N. Novak, "Multidimensional normative ratings for the international affective picture system," *Behavior Research Methods*, vol. 39, no. 2, pp. 326–334, May 2007.

[20] R. Cohendet, A. L. Gilet, M. P. Da Silva, and P. Le Callet, "Using individual data to characterize emotional user experience and its memorability: Focus on gender factor," in *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[21] B. Celikkale, A. Erdem, and E. Erdem, "Predicting memorability of images using attention-driven spatial pooling and image semantics," *Image Vision Comput.*, vol. 42, no. C, pp. 35–46, Oct. 2015.

[22] Y. Baveye, R. Cohendet, D. S. Perreira, and P. Le Callet, "Deep learning for image memorability prediction: the emotional bias," in *Proc. ACM Int. Conf. on Mumtimedia (ACMM)*, 2016, pp. 491–495.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[24] R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *CoRR*, vol. abs/1411.2539, 2014.

[25] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem, "What makes an object memorable?," in *International Conference on Computer Vision (ICCV)*, 2015.

[26] Y. Shen, C.-H. Demarty, and N. Q. K. Duong, "Technicolor@MediaEval 2016 Predicting Media Interestingness Task," in *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.