Deep Learning for Image Memorability Prediction : the Emotional Bias

Yoann Baveye, Romain Cohendet, Matthieu Perreira da Silva, Patrick Le Callet

Deep Learning for Image Memorability Prediction: the Emotional Bias

Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, Patrick Le Callet
IRCCyN UMR CNRS 6597
Université de Nantes, France
{yoann.baveye, romain.cohendet, matthieu.perreiradasilva,
patrick.lecallet}@univ-nantes.fr

ABSTRACT

Image memorability prediction is a recent topic in computer science. First attempts have shown that it is possible to computationally infer from the intrinsic properties of an image the extent to which it is memorable. In this paper, we introduce a fine-tuned deep learning-based computational model for image memorability prediction. The performance of this model significantly outperforms previous work and obtains a 32.78% relative increase compared to the best-performing model from the state of the art on the same dataset. We also investigate how our model generalizes on a new dataset of 150 images, for which memorability and affective scores were collected from 50 participants. The prediction performance is weaker on this new dataset, which highlights the issue of representativity of the datasets. In particular, the model obtains a higher predictive performance for arousing negative pictures than for neutral or arousing positive ones, recalling how important it is for a memorability dataset to consist of images that are appropriately distributed within the emotional space.

CCS Concepts

•**Computing methodologies** → **Computer vision;** *Scene understanding;* •**Human-centered computing** → *Human computer interaction (HCI);*

Keywords

Image memorability; Affect; Deep learning

1. INTRODUCTION

The study of image memorability has recently attracted the interest of computer vision researchers [4, 13, 22]. Previous work has found that individuals share a tendency to remember and forget the same images [12], which paves the way for the design of frameworks predicting the image memorability from intrinsic information. The first attempts at such a prediction used a handcrafted set of low-level features extracted from images to predict their memorability scores (*i.e.*, the degree to which the image is remembered or forgotten) [13]. Results show that image memorability prediction is possible using intrinsic features only. However, this approach performs moderately and could be improved by using higher-level – *e.g.* semantic – information. In particular, the introduction of deep learning for memorability prediction may disrupt this field of study by substantially increasing the prediction performance of memorability prediction systems.

The second issue addressed in this work is the memorability prediction performance for new images labeled with emotional scores. The emotion a stimulus conveys can be considered as a key element to predict how well it will be memorized. The psychological literature provides evidence that emotional images are generally associated with better memory performance than neutral ones [5, 7, 15]. However, datasets composed of images associated with memorability scores are not designed according to their distribution in the emotional space.

Our contributions mainly focus on these two aspects and can be summarized as follows. In this work, we introduce "MemoNet": a deep learning-based computational model for image memorability prediction significantly outperforming the performance of the models proposed in previous work. The performance of MemoNet is also evaluated on a new dataset of 150 images, for which we collected through a laboratory experiment emotional ratings and memorability scores. Results show that the performance of the model is tied to the emotional scores of images: the deep learning framework obtains a higher predictive performance for pictures inducing arousing and negative emotions than for pictures inducing neutral or positive ones.

The paper is organized as follows. Section 2 provides background material on image memorability prediction work, as well as Convolutional Neural Networks (CNNs). In Section 3, a deep learning-based framework to predict image memorability is introduced and its performance is compared with previous work. The influence of emotions on the performance of image memorability prediction is studied and discussed in Section 4, while the paper ends in Section 5 with conclusions.

2. BACKGROUND

Isola *et al.* pioneered the prediction of image memorability by mapping a combination of global image features using a support vector regression (SVR) [12, 13]. These global features are standard features that have been previously found to be effective at scene and objects recognition tasks. Indeed, Isola *et al.* shown that object and scene semantics

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967269

tends to be a primary substrate of memorability [12]. Thus, the global features are GIST, spatial pyramid histograms of SIFT, HOG2x2, SSIM, and pixel color histograms.

More recently, Mancas and Le Meur showed that attention-related features can advantageously replace low-level features in image memorability prediction by considerably reducing the size of the input features while performing slightly better [22]. Attention-based features perform 2% better by using 17 dimensions instead of the 512 dimensions of the GIST feature, in addition to the other features introduced in [13], namely SIFT, HOG2x2, SSIM, and pixel color histograms.

Previous work thus relies on a predefined set of handcrafted features extracted from images. However, building complex handcrafted features requires strong domain knowledge and is highly problem-dependent. Obtaining a satisfying feature set is thus not a trivial issue. In contrast with previous work using handcrafted features, in this work we focus on CNNs to predict image memorability scores. Beginning with LeNet-5 [20], CNNs followed a classic structure: they are composed of stacked convolutional layers followed by one or more fully-connected layers. So far, best results on the ImageNet classification challenge have been achieved using CNN-based models [17, 24]. However, large labeled datasets are crucial to train such large CNN frameworks and there is currently no large image datasets with memorability labels. Nevertheless, fine-tuning is an effective paradigm for learning high-capacity CNNs when data is scarce [9]. The fine-tuning strategy consists in pre-training a deep CNN on a large-scale external dataset (*e.g.*, ImageNet) and fine-tuning the pre-trained network by continuing the back-propagation on the small-scale target data to fit the specific classification task.

3. MEMORABILITY PREDICTION

Fine-tuning pre-trained models originally predicting object and scene semantics is particularly suitable for image memorability prediction since both concepts are inherently related to each other [12, 16]. Neural networks have also successfully modeled the biologically inspired memory processes [1, 3, 23]. Consequently, we fine-tune the GoogleNet model introduced by Szegedy *et al.*, that became in 2014 the new state of the art performance on the ImageNet dataset [24].

3.1 Fine-tuned Convolutional Neural Network

The GoogleNet architecture is a concatenation of nine similar "Inception" networks. An Inception network consists in 1×1, 3×3, and 5×5 convolutions stacked upon each other, with max-pooling layers to reduce the resolution. Given the depth of the network, two auxiliary losses are connected to intermediate layers to increase the back-propagated gradient and to prevent the vanishing gradient problem. During training, the auxiliary losses are added to the total loss of the network with a discount weight. Training is stopped after a specific number of training iterations is reached. At test time, the auxiliary losses are no longer needed and are thus removed from the network. In our fine-tuning approach, the two auxiliary and the final softmax activations are replaced by a fully-connected layer composed of a unique neuron. The loss functions associated to the model are the Euclidean loss.

All the layers of the pre-trained model are fine-tuned, but the learning rate associated to the original layers is ten times smaller than the one associated with the new last neuron. Indeed, we want the pre-trained layers to change very slowly, but let learn faster the new layer. The weights of the new layer are initialized using the xavier algorithm that automatically determines the scale of initialization based on the number of input and output neurons [10]. The biases are all initialized as constant, with the default filling value 0.

The proposed fine-tuned model for image memorability prediction is denoted MemoNet in the remaining of this paper.

Table 1: Global performance for image memorability prediction models in terms of Spearman's Rank Correlation Coefficient (ρ) and Mean Square Error (MSE)

	ρ	**MSE**
Isola *et al.* [12]	0.462	0.017
Mancas and Le Meur [22]	0.479	X
MemoNet 1k	0.522	0.017
MemoNet 10k	0.620	0.012
MemoNet 30k	0.636	0.012

3.2 Experimental Results

In order to obtain results that are fully comparable with previous work, we use in this paper the same data and the same train/test protocol used by Isola *et al.* [12] and Mancas and Le Meur [22]. The dataset is composed of 2,222 images with memorability labels collected by Isola *et al.* using crowdsourcing [13]. The images were randomly selected from the SUN dataset [26] and represent various scene categories. Object and scene categories are thus available for each image (see [8] for details). A memorability score, used as "ground truth" to train and test MemoNet, is defined as the percentage of correct detections of an image when it is repeated once in a stream of images by a set of participants. MemoNet is trained 25 times using the training sets defined by Isola *et al.* composed of one half of the images, and tested on the other half of the images.

Similarly to Isola *et al.*, global performance is defined as the mean of the Spearman's Rank Correlation Coefficient (ρ) between the ground truths and the predictions obtained for each of the 25 trained models [13]. In addition, global Mean Square Error (MSE) is assessed. The performance of MemoNet for several training iterations numbers (*i.e.*, 1k, 10k, 30k), as well as the performance of previous work, are indicated in Table 1.

Results show that the performance of MemoNet significantly outperforms the performance obtained by both Isola *et al.* [12] and Mancas and Le Meur [22]. More particularly, MemoNet 30k obtains a 32.78% relative increase in term of Spearman's Rank Correlation Coefficient compared to the approach of Mancas and Le Meur. Table 2 shows the best performing object and scene categories for categories composed of over 100 pictures. Interestingly, best performing object categories are similar to the best performing ones obtained by Isola *et al.* [12]. No significant linear correlation was found between the size of the object categories and their performance ($r = -0.018, t(25) = -0.092, p = 0.46$).

4. THE ROLE OF AFFECT

Affect is a key factor in determining the memorability of images [2, 5]. This section thus investigates if the performance of trained models is the same no matter the emotion elicited by the image given as input.

Table 2: Influence of the object and scene semantics composed of over 100 pictures on the prediction performance of MemoNet 30k. The size and the mean of the memorability scores (GT) for each category are also indicated.

Rank	Category	Size	GT	ρ
1	person sitting	165	0.753	0.655
2	person	554	0.725	0.628
3	pole	108	0.667	0.613
4	mountain	272	0.593	0.606
5	painting	101	0.696	0.593
6	wall	989	0.718	0.581
7	window	589	0.662	0.580
8	table	212	0.703	0.580
9	sign	147	0.664	0.572
10	door	361	0.668	0.570
11	chair	268	0.712	0.564
12	fence	181	0.656	0.554
13	sky	1080	0.628	0.550
14	tree	814	0.630	0.548
15	plant	417	0.640	0.543
16	floor	766	0.727	0.537
17	ground	269	0.637	0.521
18	ceiling	571	0.713	0.514
19	water	151	0.631	0.511
20	road	297	0.647	0.497
21	sidewalk	163	0.643	0.493
22	building	699	0.630	0.492
23	ceiling lamp	289	0.713	0.491
24	box	121	0.719	0.489
25	steps	115	0.659	0.477
26	grass	341	0.630	0.464
27	car	192	0.649	0.464

4.1 Memorability for IAPS Images

A subset of 150 images randomly selected from the International Affective Picture System (IAPS) dataset [19] is used in this work to analyze the performance of MemoNet for emotional images. Affective scores are available for these images in terms of valence, arousal, and dominance [14, 19]. Valence ranges from negative (*e.g.*, sad, disappointed) to positive (*e.g.*, joyous, elated), whereas arousal can range from inactive (*e.g.*, tired, pensive) to active (*e.g.*, alarmed, angry), and dominance ranges from dominated (*e.g.*, bored, sad) to in control (*e.g.*, excited, delighted).

The experimental protocol presented by Isola *et al.* [13] is reproduced in a laboratory environment to collect memorability scores for each of the 150 selected images. Participants had to perform a memory task and an emotional rating task.

4.1.1 Memory task

The memory task consists of a memory encoding phase interlaced with a recognition memory test. The task instruction is to press the space bar whenever an image reappears in a sequence of images. A black frame is displayed between each image for 1 second. During the task, 50 targets (*i.e.*, images repeated once selected from the subset of 150 images) and 200 fillers are displayed, each of them being displayed for 2 seconds. The fillers, composed of other images randomly selected from IAPS, provide spacing between the first display of a target image and its repetition. Images are displayed pseudo-randomly: the spacing between a target image and its repetition has to be separated by at least 70 images (*i.e.*, 3.30 min) in order to measure memorability corresponding to a long-term memory performance. Whenever the space bar is pushed, the image is framed by a green rectangle to show the participants that their answer is taken into account. Responses that may occur during the 1 second-long inter-stimuli black frame following the target image are also considered. The task is preceded by a training phase to familiarize the participants with the task.

4.1.2 Emotional rating task

An emotional rating task was set up to collect arousal and valence scores for 100 images displayed for 6 seconds. Before the display of each image, the participant is invited to get prepared for the rating process of the next image. The ratings are collected on the 9-point Self-Assessment Manikin (SAM) scales for arousal and valence [6], which is a powerful and easy to use pictorial system regardless of age, educational or cultural background due to its non-verbal design. The images are randomly displayed, for a total task duration of about 30 minutes. Similarly to the memorability task, the emotional rating task starts with instructions, followed by a training phase composed of training images spanning the entire arousal-valence emotional space to familiarize the participants with the task but also with the rating scales.

4.1.3 Procedure

The images were displayed on a 40 inch monitor (TV-LOGIC LVM401) with a display resolution of 1,920 × 1,080. The participants were seated at a distance of 150 centimeters from the screen (three times the screen height). The 1,024 × 768 images were centered on a black background; at a viewing distance of 150 cm, the stimuli subtended 18.85 degrees of vertical visual angle. Fifty participants (18-41 years of age; $mean = 22.54$; $SD = 5.01$; 60% of them female) compensated for their participation were recruited in Nantes, France. All participants have either normal or corrected-to-normal visual acuity. Correct visual acuity was assured prior to this experiment through near and far vision tests using Parinaud and Monoyer charts respectively. The first experimental phase was then launched, corresponding to the memory task. The next day, participants performed the emotional rating task. For each task, displayed images were selected to ensure that each memorability score is generated from at least 16 annotations and that each affective score is generated from at least 32 annotations.

4.2 Results

From Section 4.1, a set of 150 images selected from IAPS with memorability and emotional scores is created. The 150 images are pre-processed in order to be given as input to MemoNet 30k. Indeed, MemoNet 30k is fed with the resized 224 × 224 center crop of each image. Black bands added to several pictures by Lang *et al.* [19] to obtain the same ratio for each image are removed before cropping the image. As expected, the global performance of MemoNet 30k for this new dataset is lower than the performance of the model measured on the dataset created by Isola *et al.* [13] ($\rho = 0.251$; $MSE = 0.033$).

A local performance of MemoNet 30k is defined as the difference between the ground truth memorability score and the mean of the predictions from the 25 trained models for an image. Table 3 shows the rank correlation between the local performances of MemoNet 30k and the affective scores

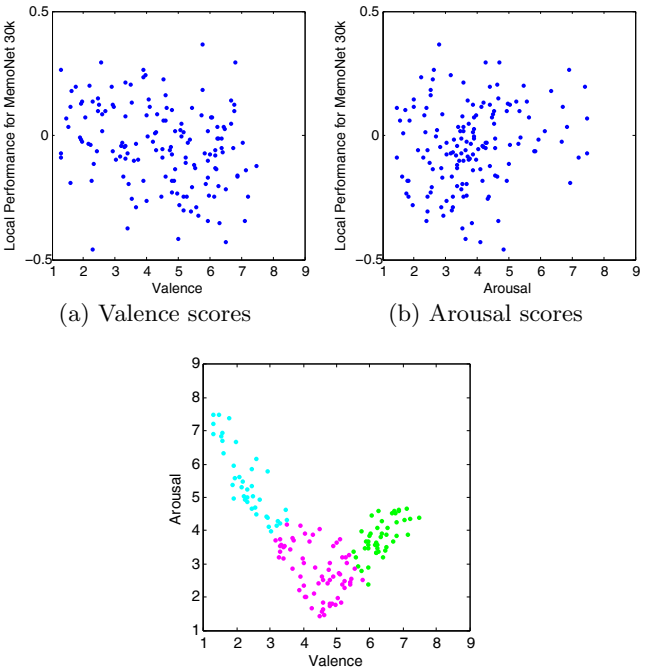
Table 3: Spearman's Rank Correlation Coefficient (ρ) between the local performances of MemoNet 30k for the subset of IAPS images and their affective labels ($^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$)

Dimension	Data origin	ρ
Valence	Lang *et al.* [19]	-0.285***
	Ito *et al.* [14]	-0.248**
	Ours	-0.284***
Arousal	Lang *et al.* [19]	0.096
	Ito *et al.* [14]	0.222*
	Ours	0.198**
Dominance	Lang *et al.* [19]	-0.221**

collected either in the experiment detailed in Section 4.1, or in previous work [14, 19]. These correlations measure the relation between the considered emotional dimension and the fact that the model predicts a memorability score higher or lower than the ground truth. Please note that the local performances using Ito *et al.*'s data are computed on the subset of images from IAPS selected both in this work and Ito *et al.*'s work, corresponding to 104 images. The local performances of MemoNet 30k and the emotional scores collected in this work are shown in Figure 1. The rank correlations exhibit a moderate, but coherent, relationship between the performance of MemoNet 30k and the valence, arousal and dominance scores for the emotional ratings collected in both previous work [14, 19] and our work. Valence is negatively correlated with the local performances of MemoNet 30k, while arousal is positively correlated. Similarly to valence, dominance exhibits a negative correlation with the local performances of MemoNet 30k. Valence and arousal account for most of the independent variance [11, 18]. Consequently, dominance is not taken into account in the following analysis.

To provide a deeper analysis of the relationship between the accuracy of the predicted memorability scores and the affective properties of the pictures collected in this work, the k-means clustering algorithm is used to separate the 150 images into three clusters in the valence-arousal space (see Figure 1(c)). The three clusters separate the images inducing arousing and negative emotions (cluster 1) from the images inducing neutral emotions (cluster 2) and from those inducing moderately arousing and positive emotions (cluster 3). It is important to note that arousal and valences scores of the 150 images are significantly correlated ($r = -0.522, t(148) = -7.445, p < .0001$). A one-way ANOVA revealed that the local performances of MemoNet 30k is significantly different for the three clusters ($F(2, 147) = 5.82; p < .005$). The Tukey's multiple comparison test confirmed that the local performances for the pictures in the first cluster is in average closer to zero – *i.e.* the optimal performance – ($mean = 0.0298$) than for images in the second ($mean = -0.0536$) or third ($mean = -0.0847$) clusters. In other words, this result shows that MemoNet 30k has the highest predictive performance for arousing negative pictures. For the other groups (*i.e.*, neutral and positive) the model is less reliable to predict the memorability.

The results suggest that affect should be taken into account in datasets of images labeled with memorability scores to ensure they induce a large variety of emotions. Indeed, emotion and memorability being related, the performance of a memorability model depends on the emotions induced by the images used to train the model. The results also suggest that emotional information could be a valuable feature to increase the performance of the model for neutral and positive pictures, especially as it is possible to computationally infer emotional information from pictures [21, 25].



(a) Valence scores (b) Arousal scores

(c) k-means clustering result

Figure 1: Local performance of MemoNet 30k with (a) valence and (b) arousal scores collected in our experiment for each of the 150 images selected from IAPS, as well as (c) the result of the k-means clustering ($k = 3$).

5. CONCLUSIONS

The study reported in this paper focuses on the image memorability prediction using deep learning. The proposed model significantly outperforms previous work and obtains a 32.78% relative increase in performance compared to the best-performing model from the state of the art. An experimental protocol has also been set up to collect memorability and emotional scores in a laboratory environment. However, the generalization of the performance of the deep learning model to this new dataset is a mitigated success. In particular, an emotional bias appears to influence the performance of the proposed model: the deep learning framework obtains a higher predictive performance for arousing negative pictures than for neutral or positive ones. This underlines the importance for an image dataset used for memorability prediction to consist in images appropriately distributed within the emotional space.

Because memorability is also subjective, the memorability prediction is doomed to inaccuracy if one is only interested in the intrinsic information of the images. Our current work focuses on the integration of context-dependent and observer-dependent information for the purpose to personalize the memorability prediction.

6. REFERENCES

[1] W. C. Abraham and A. Robins. Memory retention–the synaptic stability versus plasticity dilemma. *Trends in neurosciences*, 28(2):73–78, 2005.

[2] J. Abrisqueta-Gomez, O. F. A. Bueno, M. G. M. Oliveira, and P. H. F. Bertolucci. Recognition memory for emotional pictures in alzheimerŠs patients. *Acta Neurologica Scandinavica*, 105(1):51–54, 2002.

[3] B. Ans and S. Rousset. Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connection science*, 12(1):1–19, 2000.

[4] W. A. Bainbridge, P. Isola, and A. Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323–1334, 2013.

[5] M. M. Bradley, M. K. Greenwald, M. C. Petry, and P. J. Lang. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition*, 18(2):379–390, 1992.

[6] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[7] L. Cahill and J. L. McGaugh. A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and Cognition*, 4(4):410–421, 1995.

[8] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 129–136, June 2010.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, June 2014.

[10] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

[11] M. K. Greenwald, E. W. Cook, and P. J. Lang. Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of psychophysiology*, 3(1):51–64, 1989.

[12] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.

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

[14] T. A. Ito, J. T. Cacioppo, and P. J. Lang. Eliciting affect using the international affective picture system: Trajectories through evaluative space. *Personality and Social Psychology Bulletin*, 24(8):855–879, 1998.

[15] E. A. Kensinger, B. Brierley, N. Medford, J. H. Growdon, and S. Corkin. Effects of normal aging and alzheimer's disease on emotional memory. *Emotion*, 2(2):118–134, 2002.

[16] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Scene memory is more detailed than you think the role of categories in visual long-term memory. *Psychological Science*, 21(11):1551–1556, 2010.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[18] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, pages 39–58, 1997.

[19] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual. *Technical report A-8*, 2008.

[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[21] N. Liu, E. Dellandréa, B. Tellez, and L. Chen. Associating textual features with visual ones to improve affective image classification. In *4th International Conference on Affective Computing and Intelligent Interaction*, pages 195–204, Oct 2011.

[22] M. Mancas and O. L. Meur. Memorability of natural scenes: The role of attention. In *2013 20th IEEE International Conference on Image Processing (ICIP)*, pages 196–200, Sept 2013.

[23] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–457, 1995.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[25] W. Wang and Q. He. A survey on emotional semantic image retrieval. In *2008 15th IEEE International Conference on Image Processing*, pages 117–120, Oct 2008.

[26] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, June 2010.