Article

Journal of Integrated
SCIENCE& TECHNOLOGY

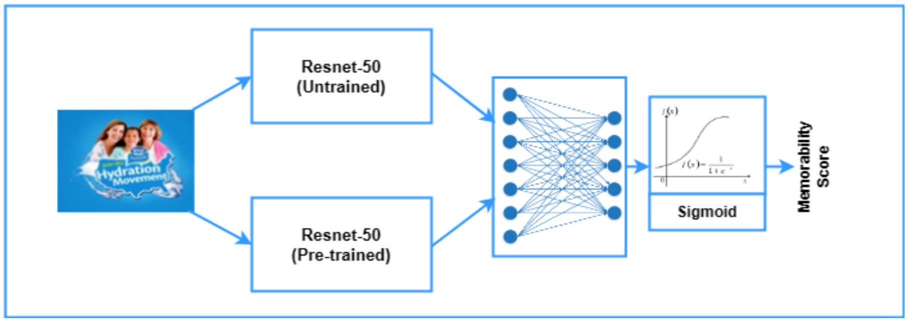
Modular Deep Learning for advertisement image memorability: Object and text bias

Amit Kumar Mandal,* Firos A., Satish Kumar Das

Department of Computer Science & Engineering, Rajiv Gandhi University, Arunachal Pradesh, India

Received on: 04-Aug-2023, Accepted and Published on: 12-Oct-2023

ABSTRACT



Hundreds of advertisement images are generated daily around us. They are composed of with different contents. A few of these are only with object(s), some of these are only with text and rest is with both object(s) and text as content. Out of these some are remembered and rest of them goes out of mind. More memorable advertisement images convey their messages more conveniently to the end users. The degree or extend to which these advertisement images are remembered or forgotten is matter of concern for the product owners or sellers. In this paper we try to analyze the correlation between the image memorability and the image content at object and text level, and also try to predict the advertisement image memorability using modular deep learning. For these purposes we conducted a memory game and proposed a modular neural network using ResNet-50. Result analysis from our proposed model MResNet and ResMem revealed that images with only text as content are more memorable and text as content modifies the memorability of an image. Also from the memorability game, it was found that memorizing ability decreases when the task of memorization increases.

Keywords: Image Memorability, ResMem, Text, Object, Modular Approach

INTRODUCTION

People generally remember what is interesting to them in an advertisement. The creative elements in advertisement, those leave long-lasting memories, need to deliver on intended purpose of the advertisement image. We come across hundreds of advertisement images daily generated by different electronic and print media. They can have different contents. Out of these, a few are with only object(s), some of these are with only text and rest are with both object(s) and text as content. Out of these some are remembered and rest of them goes out of mind. So, we need a measure to quantifying the remembrance or forgetfulness of an image. Image memorability can address this.

Image memorability can predict which images are memorable. It is defined as the degree to which an image is remembered in future after exposition. Memorability is an intrinsic feature of an image and independent of observers[1,2]. Memorability is a parameter which can be measured, described, predicted, or even can be changed[3].

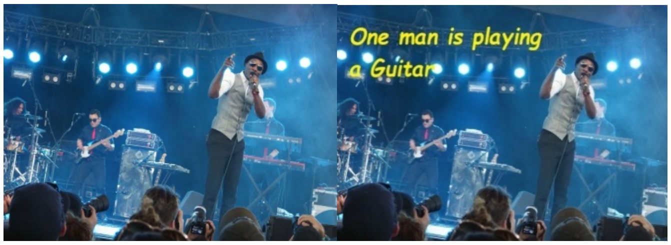


Figure 1. Image without text, image with embedded text[4] (Downloaded and edited).

But what makes an image memorable? Is it only a matter of attention and context? Consider the images in Figure 1. Both the images have some distinct objects: a singing man, some audiences, some lights, etc. All the objects are almost distinct and draw almost equal attention in the first image. But when we embed some

*Corresponding Author: Amit Kumar Mandal, Department of Computer Science & Engineering, Rajiv Gandhi University, Arunachal Pradesh, India. Email: amit.mandal@rgu.ac.in

Cite as: *J. Integr. Sci. Technol., 2024, 12(2), 740. URN:NBN:sciencein.jist.2024.v12.740*

ISSN: 2321-4635
http://pubs.thesciencein.org/jist

informative text on the image (example "one man is playing a Guitar"), all the objects except the man who is playing a guitar get little attention and are ignored to some extent. Viewers mainly concentrate on the textual area of the image and the man with guitar. This suggests that text along with object have impact in perceiving content of an image.

Image influences memory. Textual information, which describes an image or part of an image, may enhance the memorability. Mostly the advertising industries use images those have text embedded on them. They use text in such a way that they convey optimum information to the end users. The informative text along with object influences the mind and has an impact on the memorability.

BACKGROUND

Different people have different memorizing capabilities for different visual cognitive events[5]. Though there are differences, Isola et al.[6] made the first attempt to prove computationally that people consistently remember some images with details and forget the rest. They designed a memorability game and presented a series of images in a sequence to the observer and asked them to detect repetitive images in that sequence. Then they calculated a memorability score for each image, which is the rate an image was being remembered in a sequence after the image was shown for a single time. They used GIST, SIFT, HOG and SSIM features to train the SVR. Khosla et al.[7] used different image attributes such as gradient, color, texture, shape, saliency and semantic to represent more memorable or forgettable regions within an image.

Khosla et al.[8] designed the first deep learning based model, MemNet, for predicting image memorability. They used AlexNet[9] for their model and used LaMem dataset which consists of 60,000 images. Kim et al.[10] investigated the correlation between image memorability and the spatial features of object such as location and size, and relative unusualness of object's size. Moreover, Basavaraju et al.[11] proposed three models: SVR_OMP, DCNN_OMP_I and DCNN_OMP_II which utilized object's spatial-size and spatial-location to predict the object's memorability. The DCNN_OMP_II model used modular approach using AlexNet. Basavaraju et al.[12], in another work, designed FOD-MemNet which utilized depth and motion features to predict the memorability.

Fajtl et al.[13] proposed an image memorability estimation model, AMNet, to investigate the influence of attention mechanism on image memorability estimation. Yoon et al.[14] investigated the relationship between the object spatial composition and memorability of image using deep neural network based on coarse scene parsing. Praveen et al.[15] designed ResMem-Net which combines ResNet-50 and a LSTM to predict image memorability. Recently, Needell et al.[16] designed ResMem, ResMemRetain and M3M architectures based on ResNet[17] and AlexNet[9] to utilize semantic information to predict memorability. Hagen et al.[18] used vision transformer to develop ViTMem to analyze relation between semantic content and image memorability. Banna et al.[19] developed a graph embedded model to utilize the spatial structural features found within the image.

Words which are emotionally aroused and easily visualized are recalled better[20]. According to Tuckute[21] words are as memorable as picture and have intrinsic properties. Also Borkin et al.[22] found that text and human recognizable objects are key in visualization and help in recalling the visualization. Some other works[23–26] studied the influence of text along with other features in determining the video memorability. But according to our best knowledge, the influence of text in determining the image memorability has not been studied yet.

EXPERIMENT

Creation of Dataset

To conduct our experiment we collected images from Pitt Image Ads Dataset[27]. Our dataset consists of 2000 images, out of which 400 images were used as target images and rest were used as filler images, and named it as Adv_Text_Object Dataset. Then we divided the collected images into 4 content categories:

1. Category 1: 100 target images with only object(s) as content (may contain very few text which can be ignored).
2. Category 2: 100 target images with only text as content. Considered both decorative and plain text (may contain negligible amount of object(s), shapes, lines, etc.).
3. Category 3: 100 target images with both object(s) and text as content.
4. Category 4: Same 100 target images from Category 1, but text are added manually

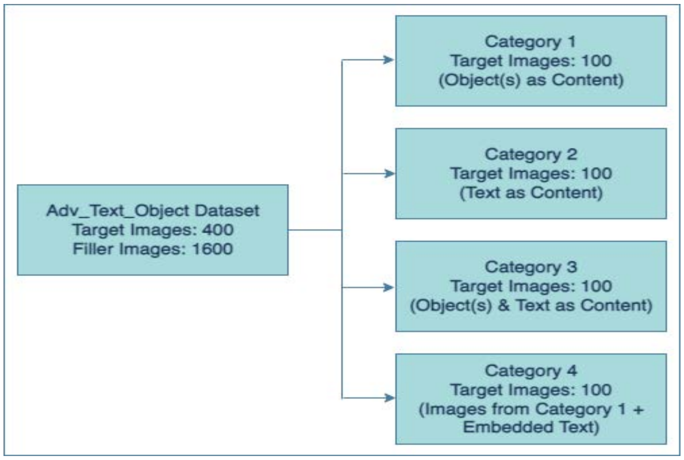


Figure 2. Structure of Adv_Text_Object Dataset.

Calculation of Ground-Truth Memorability

To calculate the ground-truth memorability score, total 50 undergraduate students of Moridhal College (Dhemaji - 787057, Assam, India) were engaged in a memory game. There were 30 male and 20 female participants aged between 18 and 21 years. We divided the game in 4 sessions.

Session 1: Each participant got 5 target images,
Each image repeated 3 times,
Game size: 200 image (target + filler)
Session 2: Each participant got 5 target images,
Each image repeated 4 times,

Game size: 300 image (target + filler)
Session 3: Each participant got 5 target images,
Each image repeated 5 times,
Game size: 400 image (target + filler)
Session 4: Each participant got 5 target images,
Each image repeated 6 times,
Game size: 500 image (target + filler)

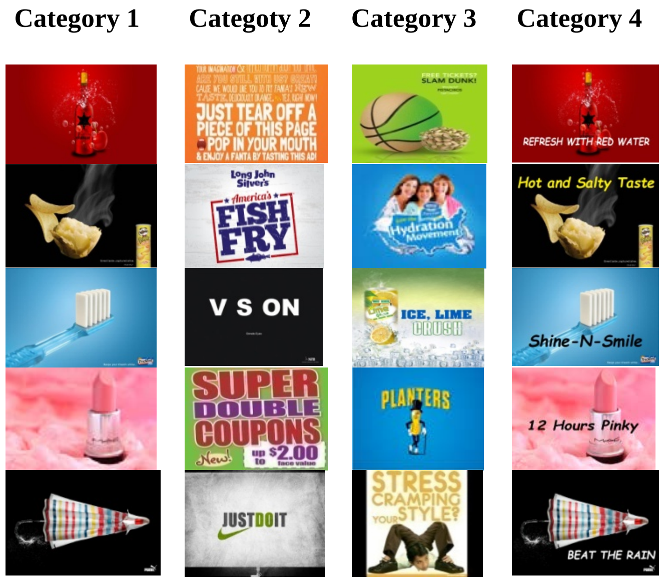


Figure 3. Sample images from each category of Adv_Text_Object.

As the session progresses, the number of repetition of target images and game size were increased. This was done to analyze the participants' memorizing ability when the task of memorization of image increases. Each target image was presented for duration of 1 second with an interval of 1 second. Each participant got total 20 (4 sessions x 5 target images) images for memorability task. Each target image got repeated 18 (3 times in session 1 + 4 times in session 2 + 5 times in session 3 + 6 times in session 4) times in the whole game and repeated after minimum 20 images in each session.

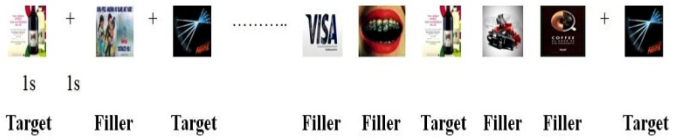


Figure 4. Memorability game for ground-truth memorability calculation.

To calculate the memorability score, we followed the work by Isola et al.[6]. Memorability score was calculated by

$$MS = \frac{H}{N} \tag{1}$$

Where,

MS = memorability score,

H = number of correctly detection of repeated image,

N = number of repetition i.e. the number of exposition of image to the participants. In our work N = 18 (3+4+5+6)

To evaluate the human consistency of our dataset, we used Spearman's Rank Correlation[28]. At first, we randomly divided the memorability scores of target images into two equal sets. Then we calculated the Spearman's Rank Correlation between the two sets of memorability scores. We repeated this process for 10 random splits and averaged the Spearman's Rank Correlation to get the final score.

Table 1: Memorability scores of Adv_Text_Object dataset. We combine the Category 3 and Category 4 as images from both the categories have object(s) and text as their contents.

	Category 1	Category 2	Category (3+4)	Whole Dataset
Average Memorability Score	0.7744	0.7394	0.7878	0.7724
Spearman's Rank Correlation	-	-	-	0.51

Methodological Approaches

For our experiment we consider two deep neural networks: MResNet (proposed) and ResMem. Since our training dataset is small, we used modular approach proposed by Anderson et al.[29] in our model.

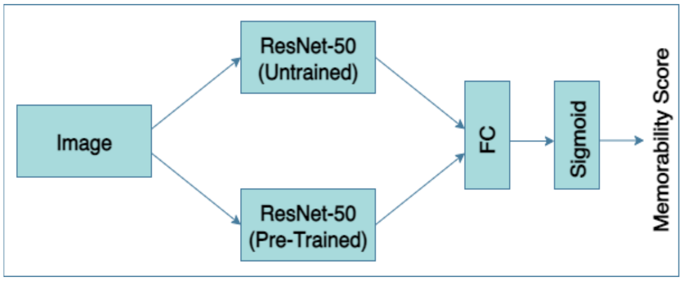


Figure 5. Architecture of MResNet.

To predict the memorability score, we proposed a modular approach named MResNet (Modular ResNet). In this approach, we use two ResNet-50 networks in parallel. The ResNet-50 in the lower stack is a pre-trained residual network trained on ImageNet[30] dataset. During the training process of our approach, this pre-trained ResNet-50 is not allowed to being learned. So this ResNet-50 retains the previously learned features. The ResNet-50 in the upper stack is allowed to learn new features during training process. The features from the both pre-trained and newly-trained networks are feed to the fully connected layer. This layer then passes the output to Sigmoid function which generates the memorability prediction score ranges between 0 and 1. Mathematical representation of the approach is

$$Y = Sigmoid([NN(x, w = \{PTRes\}), NN(x, w^* = \{PTRes\})]) \tag{2}$$

Where,

Y = predicted class level,

NN = ResNet-50,

x = input for the ResNet-50 network,

w = trainable weight initialized with the weights of the pre-trained ResNet-50 (PTRes)

w^* = non-trainable weightinitialized with the weights of the pre-trained ResNet-50 (PTRes) i.e. kept frozen from being learned during the training process.

The training process is carried out by fine-tuning MResNet on the Adv_Text_Object Dataset. The dataset is divided into training and testing sets in the ratio of 8:2 respectively. 1600 images are used for training purpose and 400 images are used for testing purpose. We only train the network in the upper stack of the MResNet.

The model is set for 20 epochs with batch size 32. Since prediction of memorability is a regression task, we used L2 (Least Square Error) loss function. Equation (3) represents the mathematical definition of L2 function:

$$L2 = \sum_{i=1}^{n}(GMS_i - PMS_i)^2 \qquad (3)$$

Where,

GMS_i = Ground-truth memorability score of ith image

PMS_i = Predicted memorability score of ith image

Pseudo Code: MResNet training process

```
**Data:** Image
**Result:**$PMS$ ⟵ Predicted Memorability Score
***while*** not conversed ***do***
    $LF_{un\_train\,i} = f_{un\_train}(\alpha_i, \beta_i) \rightarrow true$
    $LF_{pre\_train\,i} = f_{pre\_train}(\alpha_i, \beta_i) \rightarrow false$
    $PMS_{i'} = w_i LF_{un\_train\,i} + w_i^* LF_{pre\_train\,i}$
    $PMS_i = \frac{1}{1+e^{-PMS_{i'}}}$
    $\min_{w_i,\, w_i^*} L = L2 + \lambda \sum_{j=1}^{n} \omega_{ij}^2$
***end***
```

Where,

$LF_{un\_train\,i}$ = Learned image features by untrained module

$LF_{pre\_train\,i}$ = Learned image features retained from pre-trained module, but kept frozen from being learned

α_i, β_i = Image features, parameter

$\frac{1}{1+e^{-PMS_{i'}}}$ = Sigmoid function

L = Loss function

$\lambda \sum_{j=1}^{n} \omega_{ij}^2$ = L2 regularization

RESULT AND DISCUSSION

The performances of our proposed model (MResNet) and ResMem were analysed using Average Memorability Score (AMS) and Spearman's Rank Correlation (ρ). Average Memorability Scores were calculated for each Category 1, Category 2, (Category 3+4) and for whole dataset. Table 2 represents the comparison among the Average Memorability Scores (AMS) of Ground-truth memorability, MResNet and ResMem. The AMS value of Category 2 of MResNet and ResMem indicate that images with only textual information are most memorable while the AMS value of Category 2 of Ground-Truth memorability indicates that images with only text are least memorable.

Table 2: Comparison of Average Memorability Scores

	Category 1	Category 2	Category (3 + 4)	Whole Dataset
Ground-Truth Memorability	0.7744	0.7394	0.7878	0.7724
MResNet	0.7844	0.8220	0.7970	0.8001
ResMem	0.8436	0.8976	0.8538	0.8625

On the other hand, Average Memorability Scores (AMS) of Category 1 calculated by MResNet and ResMem are the lowest. It indicates that image with only object(s) as content are least memorable. While according to Ground-Truth memorability, AMS of images from Category (3 + 4) is the highest i.e. images which contain both object and text are most memorable.

Consistency performances were evaluated using Spearman's Rank Correlation (ρ). It was first calculated between ground-truth memorability scores and predicted memorability scores by MResNet, then calculated between ground-truth memorability scores and predicted memorability scores by ResMem and at last calculated between predicted MResNet scores and ResMem scores. From Table 3 it is found that in first two cases Spearman's Rank Correlations are negative and in the last case it is positive. It indicates that both MResNet and ResMem models behave similarly.

Table 3: Comparison of Spearman's Rank Correlation

	Spearman's Rank Correlation (ρ)
Between Ground-truth Scores and MResNet Scores	-0.03724
Between Ground-truth Scores and ResMem Scores	-0.05115
Between MResNet Scores and ResMem Scores	0.942185

Table 4 and Table 5 represent the comparison between the memorability scores of images without text content and with text content. Memorability scores of images from Category 1 and Category 4 are compared here. It is seen from the Table 4 and Table 5 that if we add text to an image, it modifies the memorability scores. In our most cases, text increases the memorability scores.

Table 4: Comparison of memorability scores of some images without and with embedded text

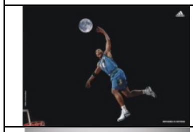
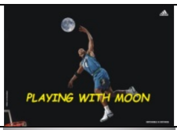
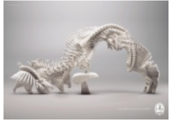
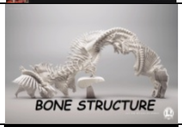
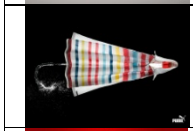
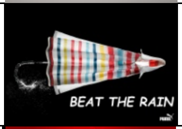
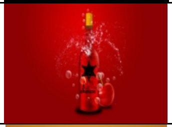
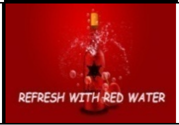
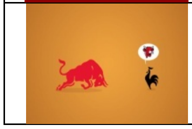
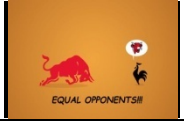
Images		Memorability Score	
Without Text	With Text	Without Text	With Text
		0.7498	0.7989
		0.8902	0.908
		0.811	0.8471
		0.7574	0.8239
		0.8803	0.8568

Table 5: Comparison of average memorability scores of images without text (Category 1) and with text (Category 4)

	Without Text (Category 1)	With Text (Category 4)
Average Memorability Scores	0.8436	0.8515

Table 6: Comparison of average hit percentage against image repetition

No. of Repetition of Target Images in Each Session (NR)	Average Hit in Each Session (AH)	Average Hit Percentage (AH/NR) x 100%
3	2.6	90
4	3.5	87.5
5	4.1	82
6	3.7	61.67

Table 6 represents the participants' memorizing ability when the task of memorizing image increases. The result shows that average hit percentage is highest (AH%= 90%) when the number of repetition of image is 3 and average hit percentage is lowest (AH% = 61.67%) when the target image repetition is 6. It is noted that average hit percentage significantly decreases when each target image was repeated for 6 times.

Observations from the above result analysis:

1. Images which contain only text as content are most memorable at machine level (MResNet and ResMem) (from Table 2).
2. Images which contain both object(s) and text as content are most memorable at human level (Ground-Truth memorability); which contradicts the result obtained from MResNet and ResMem (from Table 2).
3. At machine level images which contain only object(s) as content are least memorable (from Table 2).
4. At human level (Ground-Truth memorability), images which contain only text as content are least memorable; which contradicts the result obtained from MResNet and ResMem (from Table 2).
5. Machine may not behave similarly as human (from Table 3)
6. Text as content modifies the memorability of image (from Table 4 and Table 5).
7. Memorizing ability deceases when the task of memorization increases (from Table 6).

Our work has some limitations. Our proposed model indicated overfitting. We combined the weights from both pre-trained and untrained modules and to increase the training data we used data augmentation. We applied vertical_flip for data augmentation. We also tried using dropout from 0.2 to 0.5. Also we used L2 regularization to reduce overfitting. But overfitting problem was reduced to only some extend. We used ResMem which has a different complexity than our proposed model. Since ResMem is a well-established model to predict the image memorability we used ResMem along with our model. Also for Category 3 we did not consider images those have long-sentence (more than 15 words) text along with object as we assumed in the beginning of the research that long sentences negatively impact attention and may diminish the object's presence.

CONCLUSION

In this paper, we analyzed the correlation between image memorability and image content at object and text level. We have found that text and object are correlated to image memorability and can modify the image memorability. More memorable advertisement images can be created by modifying the objects and text contents within the images. Along with the increase in the investment in advertising, it is important to test the effectiveness of visual content of an advertisement image using object and text before publish.

ACKNOWLEDGMENTS

This study was supported by Moridhal College, Dhemaji - 787057, Assam, India by providing us the infrastucture to conduct the memory game among the students. Permission letter no. MC/Admn.28(c)/89-90/2022/1017(A).

CONFLICT OF INTEREST

Authors declare no conflict of interest is there for publication of this work.

REFERENCES AND NOTES

1. W.A. Bainbridge, P. Isola, A. Oliva. The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.* **2013**, 142 (4), 1323–1334.
2. D. Parikh, P. Isola, A. Torralba, A. Oliva. Understanding the intrinsic memorability of images. *J. Vis.* **2012**, 12 (9), 1082–1082.
3. O. Sidorov. Changing the Image Memorability: From Basic Photo Editing to GANs. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE, Long Beach, CA, USA, **2019**; pp 790–799.

4. L. Goetschalckx, J. Wagemans. MemCat: a new category-based image set quantified on memorability. *PeerJ* **2019**, 7, e8169.
5. E.B. Hunt, J. Davidson, M. Lansman. Individual differences in long-term memory access. *Mem. Cognit.* **1981**, 9 (6), 599–608.
6. P. Isola, J. Xiao, A. Torralba, A. Oliva. What makes an image memorable? In *CVPR 2011*; IEEE, Colorado Springs, CO, USA, **2011**; pp 145–152.
7. A. Khosla, J. Xiao, A. Torralba, A. Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems 25*; Curran Associates Inc., Lake Tahoe, Nevada; Vol. 1, p 3328.
8. A. Khosla, A.S. Raju, A. Torralba, A. Oliva. Understanding and Predicting Image Memorability at a Large Scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*; IEEE, Santiago, Chile, **2015**; pp 2390–2398.
9. A. Krizhevsky, I. Sutskever, G.E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, 60 (6), 84–90.
10. J. Kim, S. Yoon, V. Pavlovic. Relative spatial features for image memorability. In *Proceedings of the 21st ACM international conference on Multimedia*; ACM, Barcelona Spain, **2013**; pp 761–764.
11. S. Basavaraju, S. Gaj, A. Sur. Object Memorability Prediction using Deep Learning: Location and Size Bias. *J. Vis. Commun. Image Represent.* **2019**, 59, 117–127.
12. S. Basavaraju, P. Mittal, A. Sur. Image Memorability: The Role of Depth and Motion. In *2018 25th IEEE International Conference on Image Processing (ICIP)*; IEEE, Athens, **2018**; pp 699–703.
13. J. Fajtl, V. Argyriou, D. Monekosso, P. Remagnino. AMNet: Memorability Estimation with Attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE, Salt Lake City, UT, **2018**; pp 6363–6372.
14. S. Yoon, J. Kim. Object-Centric Scene Understanding for Image Memorability Prediction. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*; IEEE, Miami, FL, **2018**; pp 305–308.
15. A. Praveen, A. Noorwali, D. Samiayya, et al. ResMem-Net: memory based deep CNN for image memorability estimation. *PeerJ Comput. Sci.* **2021**, 7, e767.
16. C.D. Needell, W.A. Bainbridge. Embracing New Techniques in Deep Learning for Estimating Image Memorability. *Comput. Brain Behav.* **2022**, 5 (2), 168–184.
17. K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, Las Vegas, NV, USA, **2016**; pp 770–778.
18. T. Hagen, T. Espeseth. Image Memorability Prediction with Vision Transformers. arXiv January 20, 2023.
19. T.T. Banna, S. Deb, S. Rahman, S. Rahman. GEMM: A Graph Embedded Model for Memorability Prediction. In *2023 International Joint Conference on Neural Networks (IJCNN)*; IEEE, Gold Coast, Australia, **2023**; pp 1–8.
20. M. Bock. The influence of emotional meaning on the recall of words processed for form or self-reference. *Psychol. Res.* **1986**, 48 (2), 107–112.
21. K. Mahowald, P. Isola, E. Fedorenko, E. Gibson, A. Oliva. Memorable words are monogamous: The role of synonymy and homonymy in word recognition memory. PsyArXiv.
22. M.A. Borkin, Z. Bylinskii, N.W. Kim, et al. Beyond Memorability: Visualization Recognition and Recall. *IEEE Trans. Vis. Comput. Graph.* **2016**, 22 (1), 519–528.
23. R. Kleinlein, C. Luna-Jiménez, D. Arias-Cuadrado, J. Ferreiros, F. Fernández-Martínez. Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability. *Appl. Sci.* **2021**, 11 (16), 7406.
24. H. Ali, S.O. Gilani, M.J. Khan, et al. Predicting Episodic Video Memorability using Deep Features Fusion Strategy. In *2022 IEEE/ACIS 20th International Conference on Software Engineering Research, Management and Applications (SERA)*; IEEE, Las Vegas, NV, USA, **2022**; pp 39–46.
25. R. Leyva, V. Sanchez. Video Memorability Prediction Via Late Fusion Of Deep Multi-Modal Features. In *2021 IEEE International Conference on Image Processing (ICIP)*; IEEE, Anchorage, AK, USA, **2021**; pp 2488–2492.
26. J. Li, X. Guo, F. Yue, F. Xue, J. Sun. Adaptive Multi-Modal Ensemble Network for Video Memorability Prediction. *Appl. Sci.* **2022**, 12 (17), 8599.
27. Z. Hussain, M. Zhang, X. Zhang, et al. Automatic Understanding of Image and Video Advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, Honolulu, HI, **2017**; pp 1100–1110.
28. C. Spearman. The proof and measurement of association between two things. By C. Spearman, 1904. *Am. J. Psychol.* **1987**, 100 (3–4), 441–471.
29. A. Anderson, K. Shaffer, A. Yankov, C.D. Corley, N.O. Hodas. Beyond Fine Tuning: A Modular Approach to Learning on Small Data. **2016**.
30. J. Deng, W. Dong, R. Socher, et al. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE, Miami, FL, **2009**; pp 248–255.