Abstract

Image memorability, due to its subjective nature, is a very hard problem in Image Processing. But thanks to Deep Learning and large availability of data and GPUs, great strides have been made in predicting the memorability of an image. In this paper, we propose a novel deep learning architecture that is a hybrid of LSTM and CNN that uses information from the hidden layers of the CNN to compute the memorability score of an image. The proposed architecture automatically learns visual emotions and saliency which is shown by our heatmaps generated using GradCam technique. The model is trained and evaluated using the Large-scale Image Memorability dataset (LaMem) from MIT. The results show that the model achieves a rank correlation of 0.679 and a mean squared error of 0.011 which is better than the current state of the art models and also is close to human consistency (p=0.68).

Keywords: Deep Learning, Image Memorability, Visual Emotions, Saliency, Object Interestingness

1 Introduction

In this digital era, data is core to almost every media platform which starts from television to social networks. Every media platform relies on content to engage their users/viewers. This provides a compulsion for these platforms to understand the exponentially growing data in order to serve the right content to their users. Since most of these platforms rely on visual data, concepts such as popularity, emotions, interestingness, aesthetics and most importantly memorability is crucial to be understood to increase viewership. In this paper, image memorability will be the sole focus, which is one of the most underexplored applications of deep learning.

Human beings are very reliant on visual memories to remember, identify and discriminate between objects in real life. However, not all humans remember the same visual information in a common manner. [1] This is a long-standing question that has been asked by neuroscientists for years and research is still underway to explain how exactly the cognitive processes in the brain encode and store certain information in order properly retrieve that information when required. The human brain is capable of encoding intrinsic information about objects, events, words and images just after a single exposure to visual data [13]. Also, in the experiments conducted by Standing et al [24], people were exposed to 10,000 images and it was inferred that the memory performance in the recognition task was extremely high. The people who participated in the experiment were able to remember thousands of images even after being exposed to many similar images later in the experiment.

Image memorability is generally measured as the probability that a person will be able to identify a repeated photograph when he/she is presented with a stream of images [5, 7, 20, 25]. Cognitive psychologists have shown that more memorable images leave a larger trace on the long-term memory of the brain [5]. However, the memorability of a certain image varies from person to person and depends on the context and the previous experiences of the person.

In [7], [8] and [13], researchers have shown that humans show a level of consistency when remembering the same kind of images with a very similar probability irrespective of the time delay with which they look at the same image. This research has led to the inference that it is possible to measure the probability that an individual will remember an image. To measure the probability that a person will remember an image, the user is presented with a stream of images and is checked if the person can detect the repetition of an image as shown in Figure 1. Based on how well an image's repetition is detected by that individual, the memorability is identified.

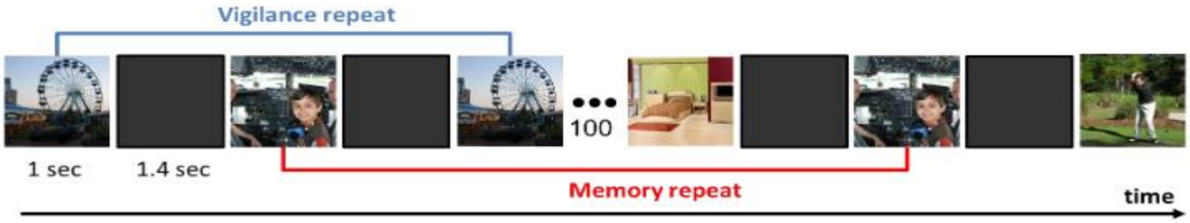


Figure 1: **Visual Memorability Game.** Participants watch for repeats in a long stream of images.

Just like other properties of images such as photo composition, image quality and beauty, image memorability is dependent on the user's opinion and context, hence making it vary from person to person. However, despite this large variability, generally, humans do agree with each other on certain common factors that tend to make an image more memorable. Factors like color harmony and object interestingness are generally agreed upon by people as factors that improve image memorability [12]. Thus, even though the subjectivity of image memorability can make it seem like it's impossible to predict the memorability using a computer program, but the above-mentioned studies and availability of datasets have it made it possible to predict the memorability of an image [7, 10, 11].

In the past decade, several methods have been proposed [] to predict the memorability of an image using deep learning methods. In this paper, we have proposed a novel architecture which uses a hybrid of Convolutional Neural Networks and Long Short-Term Memory Networks (LSTM) to build a deep neural network architecture that uses a memory driven technique to predict the memorability of images. We use the LaMem dataset, which consists of 60,000 images, with each image being labelled with a memorability score. The dataset was tagged using the Visual Memorability Game as mentioned before in this paper.

In this paper, we propose a new general-purpose neural network architecture called V-Net that achieves results that is very close to human performance on the LaMem dataset. Previous works have shown astonishing results that have helped achieve really good results on LaMem, but this architecture has brought it very close to human performance that can have a considerable impact on applications of image memorability prediction in e-commerce platforms and photo-sharing platforms.

2 Related Work

In [5] and [7] Isola et al. Pioneered the idea of using data driven strategies to predict image memorability. In [5], Isola et al used the Visual Memory Game to prepare a dataset of images and annotate their respective memorability score. The game was run on Amazon Mechanical Turk where users were presented with a stream of images with some images repeating on a random basis. The users were asked to press a key when they believe that the image that is displayed was already seen before. Using this Isola et al collected 2222 images along with the annotated memorability scores. When they analyzed the images and their memorability scores together, they understood that the memorability of an image is highly related to certain object and scene semantics such as 'Labelled Object Counts', 'Labelled Object Areas' and 'Object Label Presences'. Also, when each image was segregated into scene categories, it was inferred that much of what contributes to the image's memorability score was from both the object and scene semantics. In [7], Isola et al followed up on their work in [5] to understand the human understandable visual attributes in order understand memorability as a cognitive process. The same dataset that was used in [5] was also used for [7] and Isola et al also developed a greedy algorithm to perform feature selection in order to understand identify a compact set of image properties that affect image memorability.

Later, following the work of Isola et al, in [12], Khosla et al presented a new dataset, LaMem, a novel and a diverse dataset with 60,000 images, each tagged with a memorability score that is similar to the dataset from [5]. The authors have used Convolutional Neural Networks(MemNet) to fine tune deep features that out perform all other features by a large margin. The analysis made by the author on the on the responses of high-level CNN layers shows which objects are positive. In [17], Celikkale et al proposed a new computational model based on attention mechanism to predict image memorability based on deep learning. In this paper, the authors have shown that the emotional bias affects the performance of the proposed algorithm due to the deep learning framework arousing negative pictures than for positive or neutral pictures.

Recently, in [15] Sathisha et al developed a novel multiple instance based deep CNN for image memorability prediction. The architecture proposed in [15] shows performance levels that is close to human performance on the LaMem dataset. The proposed model in [15], EMNet, automatically learns various object semantics and visual emotions using a multiple instance learning framework in order to properly understand the emotional cues that contribute extensively to the memorability score of an image.

3 Proposed model

This section deals with the proposed Neural Network architecture, the dataset used and the performance of our model. Further, the results obtained from an extensive set of experiments are compared with previous state-of-the-art results which shows the superiority of the proposed architecture. For every problem that is solved by deep learning, there are four core entities that have to be defined, before the results are obtained. They are: the dataset, the neural network architecture, the loss function and the training procedure.

3.1 Datasets

In this paper, two publicly available datasets are used namely, the LaMem dataset and dataset from Philip Isola et al.

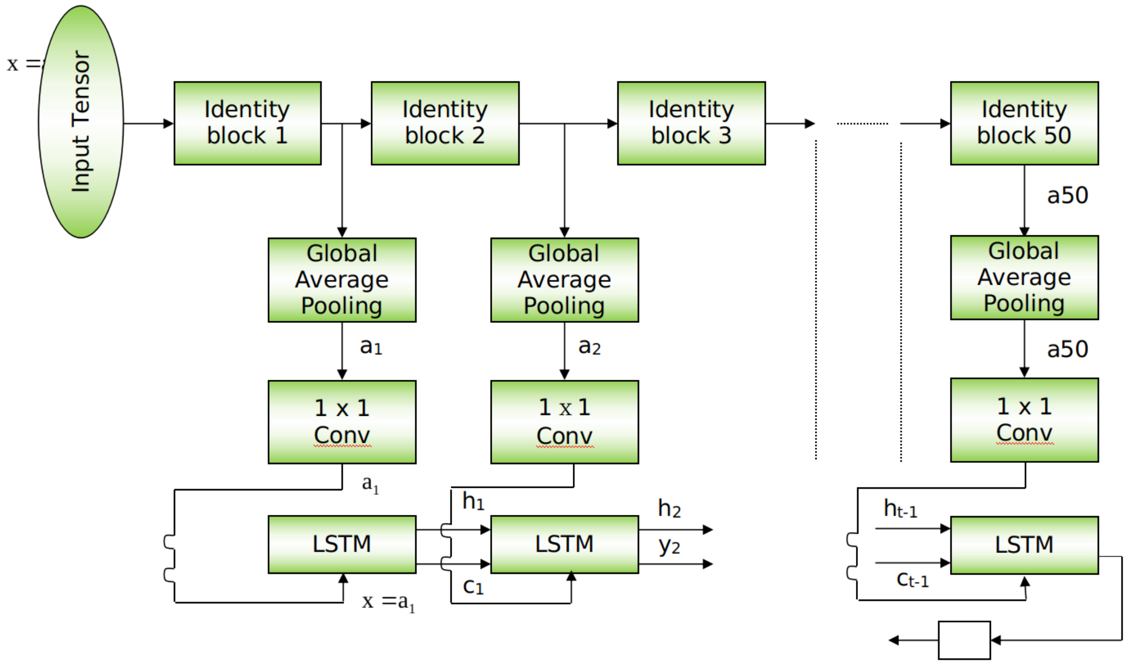
LaMem is currently the largest publicly image memorability dataset that contains 60,000 annotated images [20]. Images were taken from MIR Flickr, AVA dataset, Affective images dataset, MIT 1003 dataset, SUN dataset, image popularity dataset and aPascal dataset. The dataset is very diverse as it includes both object centric and scene centric images that capture a wide variety emotion as well. The images were annotated by users on Amazon Mechanical Turk. Isola et al [7] prepared a dataset containing 2222 images from the SUN dataset and annotated them using the Visual Memorability Game. Amazon Mechanical Turk was used to allow users to view the images and play the game.

Both the datasets were collected with human consistency in mind i.e, the authors ran human consistency tests to understand how consistent the users are when it comes to detecting repetition of images. The consistency was measured using Spearman's rank correlation and the rank correlation for LaMem and Isola et al are 0.68 and 0.75 respectively. This analysis proves that humans are generally consistent when it comes to remembering or forgetting images.

3.2 Deep Hybrid CNN for the prediction of memorability scores

3.2.1 Model Overview

This section provides a detailed explanation about the V-Net. A visual depiction is given in Figure 2. The figure shows that there are two distinct portions in the entire architecture. On the top we use Resnet-50 [30] which is the state of art deep learning architecture for many applications. Resnet-50 is a 50-layer deep neural network that contains convolution matrices at each layer. The main innovation in ResNet is the skip connection which helps us to avoid vanishing gradients in very deep neural networks. A single ResNet block (identity block) is shown in Figure 3. The input image is given to ResNet50, and the size of the image used in our experiment is 224x224 px. One of the core features of the proposed architecture is that it is the CNN part of the architecture is fully convolutional in nature and due to the use of Adaptive Average Polling layers, the input image can be of a larger or smaller than 224x224 px size.



At the bottom, a Long Short-Term Memory (LSTM) unit is placed which is responsible for predicting the output, which is the memorability score. LSTM is an enhancement to Recurrent Neural Networks (RNN) and was proposed by Hochreiter & Schmidhuber [21]. RNNs are generally used for sequential data such as text-based data or time series data. However, in RNN we don't use any memory units that can resolve any long-term dependencies, which was solved in [21]. In an LSTM unit, we compute a 'cell state' which can retain information from previous input sequences. LSTM units accept sequential data as inputs and in this architecture, the input to the LSTM unit are the activations of the hidden layers of the Resnet50 model. As the input sequences being sent to the LSTM unit have to be of the same size, we use Global

Average Pooling (GAP) to shrink the activations of the hidden layers to a size of (Cx1x1) where C is the number of channels [26]. Global Average Pooling is very much alike Densely connected Layers in Neural Networks in the sense that they simply perform a linear transformation on a set of feature maps. This allows us to ensure that we don't have to care too much about the size of the output activations at each layer and about the size of the input image. As mentioned in [26], Global Average Pooling also doesn't have any parameter to optimize, thus avoiding overfitting and reducing computational needs. Further, a convolution operation is done on the output of GAP layers to obtain a 128-channel output which can be Flattened to obtain vectors of Rank 128.

The main reason behind passing the hidden layer activations to the LSTM unit, is to ensure that the cell state vector can remember the important information from the hidden layers and when the final layer's activation is passed to the LSTM unit, we'll be able to use the important information of the previous layers along with the final layer's activation and then use all that information to compute the memorability score. The LSTM layer's output is a n-dimensional vector, which is passed to a linear fully connected layer that gives a scalar output, which is the memorability score of the image. This strategy allows us to not just use the final layer's activations alone which is generally done in previous works as shown in MemNet [20], MemBoost [2] and VGGMemNet [15].

3.2.2 Mathematical Formulation

So, the input image is a tensor of size (3, 224, 224), which is denoted by $A_{0.}$ The output of L^{th} identity block is denoted by A_L. At each L^{th} identity block, the output of the identity block is calculated by:

$Z^{[1]}{}_L = W^{[1]}{}_L \text{ (conv) } A_{L-1}$

$Z^{[2]}{}_L = W^{[2]}{}_L \text{ (conv) } Z^{[1]}{}_L$

$A_L = \text{relu}(Z^{[2]}{}_L)$, where $\text{relu}(a) = \max(0,a)$

For all L, A_L is passed through a Global Activation Pooling layer, which converts a (C, W, H) tensor to a (C, 1, 1) tensor by taking the average of each channel in the activation matrix A_L.

At the LSTM layer, the initial cell state is denoted by C_0 and the initial activation is denoted by h_0. Initially, before the hidden layer activations are passed to the LSTM, C_0 and h_0 are initialized as random vectors using He initialization strategy.

The LSTM unit consists of three important gates that form the crux of the model:

1. Update Gate – Decides what information should be remembered and what information should be thrown away
2. Forget Gate – To decide which information is worth storing
3. Output Gate – The output of the LSTM unit

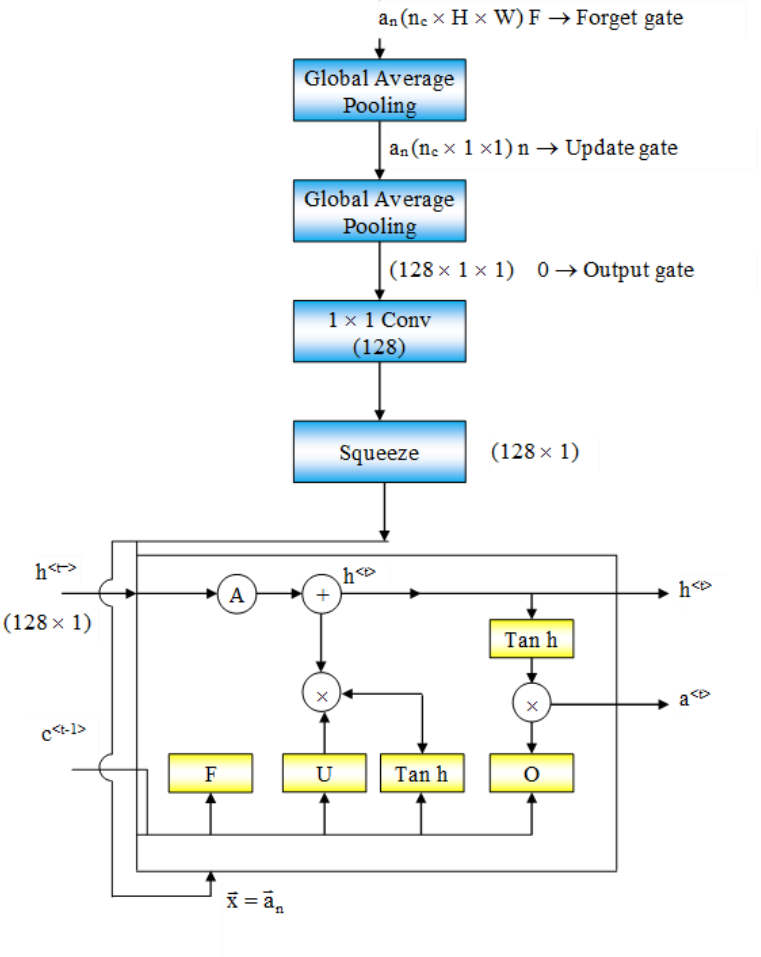
Update Gate: $G_u = \text{sigmoid } (W_{hc}\, C^{<t-1>} + W_{ub}\, x^{<t>} + b_u\,)$

Forget Gate: $G_f = \text{sigmoid } (W_{fc}\, c^{<t-1>} + W_{fx}\, x^{<t>} + b_f)$

Output Gate: $G_o = \text{sigmoid}(W_{oc}\, C^{<t-1>} + W_{ox} x^{<t>} + b_u)$

Hidden cell state: $h^{<t>} = G_u + h^{<t>} + G_f * h^{<t-1>}$

LSTM output: $c^{<t>} = G_0 * h^{<t>}$



3.3 The Loss Function

The scores of the images in both the mentioned datasets are continuous valued outputs, making this entire task a regression task. To understand how good our model is at predicting the memorability, loss functions are used which is can approximate the divergence between the target distribution and the predicted distribution. Generally, for regression tasks, the L2 loss function, also known as the Mean Squared Error (MSE) is used as the loss function for the proposed model.

(EQN)

Where the (y-hat) represents the predicted value, while (y) represents the ground truth value of the nth image in the dataset. The second term is added to the existing loss function to prevent the model from overfitting. The regularization procedure is known as L2 regularization, which multiplies a weight decay (hyperparameter) and the summation of all the weights used in the Neural Network. The weight decay prevents the weights from being too big which ultimately prevents the model from overfitting.

3.4 Training Procedure

3.4.1 Optimization

The loss function is actually differentiable and is also a function of the parameters of the Neural Network. The gradient of the loss function with respect to the weights can guide us through a path to that'll allow us to identify the right set of parameters that yield a low loss using gradient descent based methods. In our experiments, we used ADAM which is a combination of Stochastic Gradient Descent with Momentum and RMSprop [23]. The loss function of Neural Networks is very uneven and sloppy due to the presence of too many local minima and saddle points. ADAM uses exponentially weighted moving averages to ensure that our learning path is much smoother and can actually converge at a good minimum.

3.4.2 Learning Rate and One Cycle Learning Policy

The learning is one of the most important hyper parameters in deep learning as it decides how much we want to change the weights to move towards to a minimum in the loss function's surface. Basically, a smaller learning rate causes our model to converge slowly, while a large learning rate can cause the model to diverge. That's what makes the learning rate so important. In [27], the author describes an approach to set the learning rate and batch size. The author recommends to do one cycle of learning rate of 2 steps of equal length. We choose maximum learning rate using range test. We use lower learning rate as 1/5th or 1/10th of maximum learning rate. We go from lower learning rate to higher learning rate in step 1 and back to lower learning rate in step 2. We pick this cycle length slightly lesser than total number of epochs to be trained. And in the remaining iterations, we annihilate learning rate way below lower learning rate value (1/10th or 1/100th). The motivation behind this is that, during the middle of learning when learning rate is higher, the learning rate works as regularization method and keep network from overfitting. This helps the network to avoid steep areas of loss and land better flatter minima.

3.4.3 Pseudocode

```
Procedure mem (images):
    Cache = []
    A[0] = images[0]
    For i=1 to n_layers:
        A[i] = W[i] (conv) a[i-1] + b[i]
        A[i] = relu(A[i])
        A[i] = A[i] + A[i-1]
        Cache[i] = A[i]
    For i = 1 to n_layers:
        S = cache[i]
        S = globalAveragePooling(S)
        S = W[i] (conv) S
        h, c = LSTM_CELL(s, h, c)
    L = w_l * h + b_l
    Return L
Procedure LSTM_CELL(x, $h_{t-1}$, $c_{t-1}$):
    it = sigmoid ( $W_{xi}$ * x + $W_{hi}$ *  $h_{t-1}$ + $W_{ci}$ *  $c_{t-1}$ + $b_i$)
    ft = sigmoid (  $W_{xf}$ * x + $W_{hf}$ *  $h_{t-1}$ + $W_{cf}$ *  $c_{t-1}$ + $b_f$ )
    ct = ft* $c_{t-1}$ + it * tanh($W_{hc}$ * $h_{t-1}$ + $W_{xc}$ * x + $b_c$)
    ot = sigmoid( $W_{xo}$ * x + $W_{ho}$ * $h_{t-1}$ + $W_{co}$ * ct + $b_o$ )
    ht = ot * tanh(ct)
    Return ht, ct
Procedure globalAveragePooling(tensor):
    c, h, w = dimensions(tensor)
    for i in range(c):
        Avg = (1/h) * (Σtensor[i])
        tensor[i] = Avg
    return tensor
```

4 Results

4.1 Evaluation Metric

The L2 loss function is a good metric to tell how well the proposed model performs, but here the Rank Correlation method is also used to evaluate the proposed model. The Spearman Rank Correlation (p) is computed between the predicted score and target score, is used to tell the consistency between the predicted scores and target score from the dataset. The value of (p) ranges from –1 to 1, where 1 tells us that there is complete agreement between the predicted and the target memorability score, while –1 tells that there is complete disagreement. The rank correlation between predicted and target memorability score is given by the formula below:

{Formula}

Where, {explain formula}

4.2 Results

V-Net was trained on the LaMem dataset which contains 60,000 images. For cross validation purposes, the authors divided the dataset into 5 sets for cross validation purposes where each set contains 45,000 images for training, 10,000 images for testing and 3,741 images for validation purposes. In [list of related work], the authors have developed 5 models and averaged the results of the models to compare with previous works. So, to ensure that the comparison of our results is fair, we training 5 models on the 5 sets and averaged the results.

As mentioned previously, the metric used to compare the proposed architecture with the existing models is the Spearman Rank Correlation metric that is computed between the predicted and ground truth scores. Table 1 represents the results of various models on the LaMem dataset. In table 1, it can be observed that V-Net attains a rank correlation of 0.679 which is a 6.09% increase from MemNet, 35% increase from CNN-MTLES, a 2% increase from MCDRNet and an 1.2% increase from EMNet. The human level accuracy on LaMem is 0.68 and V-Net has brought us extremely close to human accuracy with a difference of just 0.001.

In table 2, the same model trained on the LaMem was used to make predictions on the dataset from Isola et al which consists of 2,222 images. Without any retraining on the new dataset, V-Net showed better results than the other previously proposed architecture. From table 2 it can be inferred that V-Net attains a rank correlation of 0.671 which is a 10.33% increase from MemNet, 5.48% increase from MCDRNet, 1.4% increase from EMNet and a 45.67% increase from SVR. The human accuracy for this dataset hasn’t been provided by the authors and hence we are not sure how close {our model} is close to human accuracy, but it is clear that V-Net has outperformed all other previous works.

Method (LaMem dataset)	**(Roe)**
V-Net	0.679
MemNet	0.640
MCDRNet	0.663
EMNet	0.671
CNN-MTLES (different dataset split) [28]	0.5025

Method (Isola et al dataset)	**(Roe)**
V-Net	0.673
MemNet	0.61
MCDRNet	0.638
EMNet	0.664
SVR	0.462

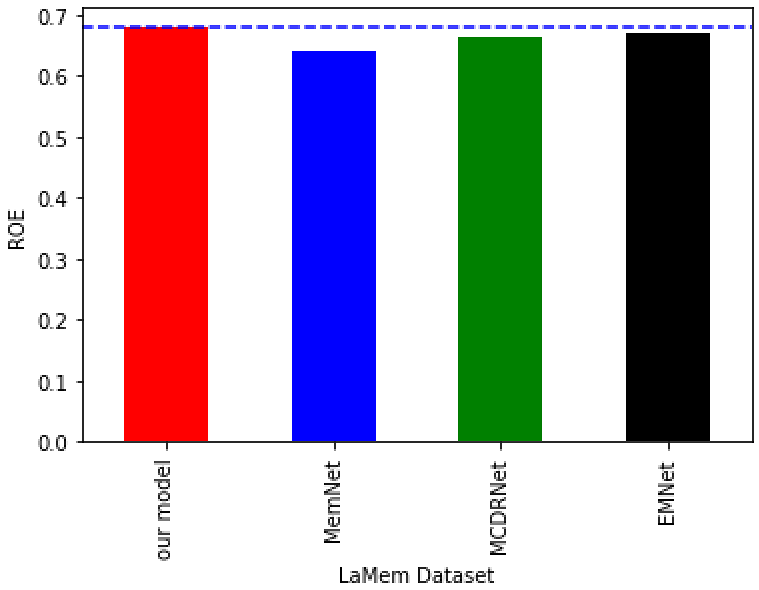


Fig 1. Bar Chart showing the superiority of V-Net over the state-of-the-art EMNet on the LaMem dataset

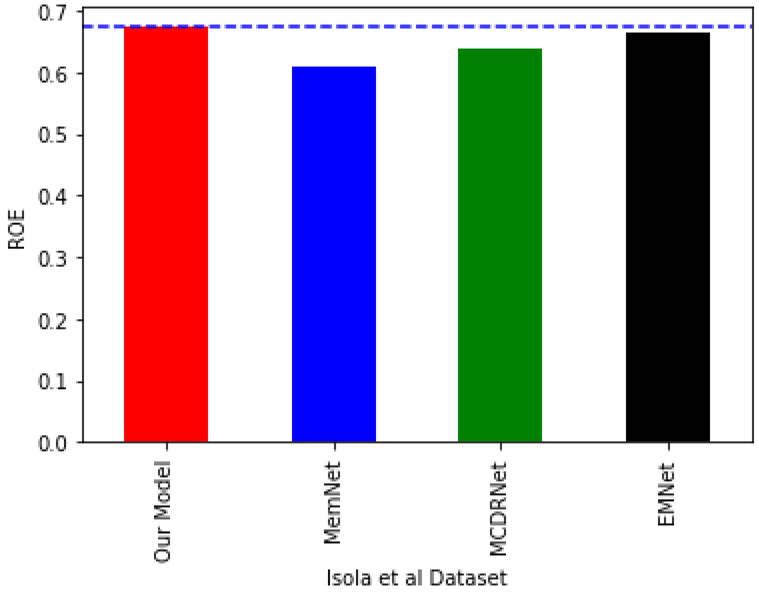


Fig 2. Bar Chart showing the superiority of V-Net over the state-of-the-art EMNet on the Isola et al dataset

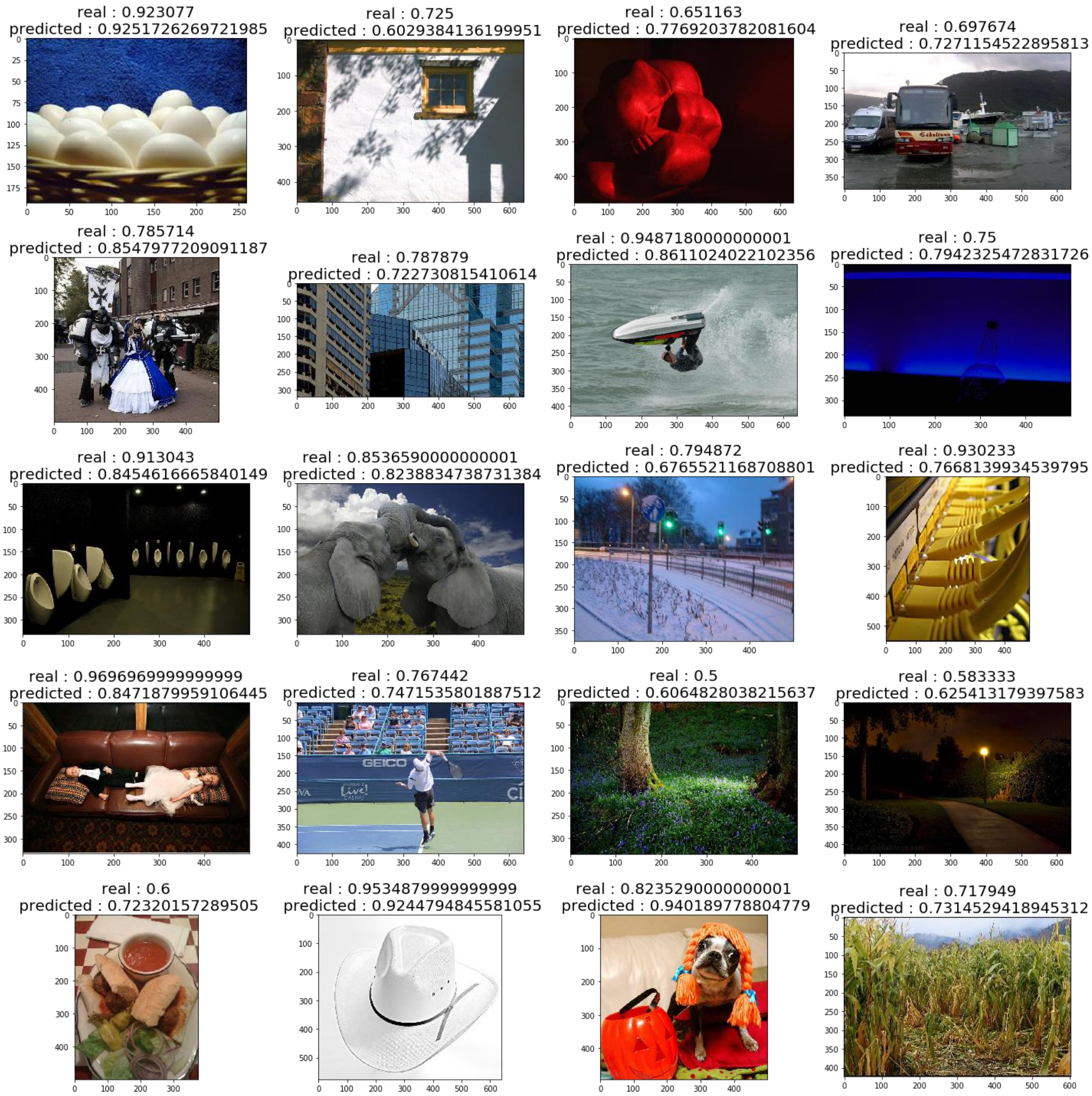

real : 0.923077
predicted : 0.9251726269721985
real : 0.725
predicted : 0.6029384136199951
real : 0.651163
predicted : 0.7769203782081604
real : 0.697674
predicted : 0.7271154522895813
real : 0.785714
predicted : 0.8547977209091187
real : 0.787879
predicted : 0.722730815410614
real : 0.9487180000000001
predicted : 0.8611024022102356
real : 0.75
predicted : 0.7942325472831726
real : 0.913043
predicted : 0.8454616665840149
real : 0.8536590000000001
predicted : 0.8238834738731384
real : 0.794872
predicted : 0.6765521168708801
real : 0.930233
predicted : 0.7668139934539795
real : 0.9696969999999999
predicted : 0.8471879959106445
real : 0.767442
predicted : 0.7471535801887512
real : 0.5
predicted : 0.6064828038215637
real : 0.583333
predicted : 0.625413179397583
real : 0.6
predicted : 0.72320157289505
real : 0.9534879999999999
predicted : 0.9244794845581055
real : 0.8235290000000001
predicted : 0.940189778804779
real : 0.717949
predicted : 0.7314529418945312

Fig 3. Predicted and actual scores for various images from the LaMem dataset

5. Analysis of the results

In this section, we talk about inferences and patterns that we were able to identify after analyzing the results of V-Net. In order to aid us with this process, we used GradCam [29] technique to understand which part of the images were focused by {our architecture} or in other words, which part of the image gave larger activations. GradCam is an extension of Class Activation Maps (CAM), which uses the gradient obtained during backpropagation process, to generate heatmaps. The heatmaps as shown in Fig 4 shed light on which part of the image enhance the memorability of the image. The Red portion shows the most memorable portion of the image, while the Blue portion shows the least memorable portion of the image. The colors in between i.e, yellow and green show the portions that don’t affect the memorability score too much. Based on the heatmaps and careful manual analysis of the results, the following inferences were made:

1. The object in the image contributes more to the memorability score than the scene in which the object is placed. In almost every heatmap, it is observable that the portion of the image that contains the main object provides higher activations in comparison to rest of the image. Also, images with no objects are predicted to be less memorable in comparison to images containing objects (both living and non-living). Also, images containing a single central object is seen to be more memorable than images with multiple objects. Also, presence of humans in the image contributes to a better memorability score.
2. Using a model pretrained on object classification datasets provide better results and trains faster than using a model pretrained on scene classification datasets. This can be attributed to the fact that memorability scores are directly related to the presence of objects and their interestingness. Thus, a model whose weights contain information about objects take lesser time to converge to a minima.
3. Image Aesthetics doesn’t have much to do with Image memorability. Some images that contain content related to violence aren’t aesthetically good, but the memorability score of the image is predicted to be high.

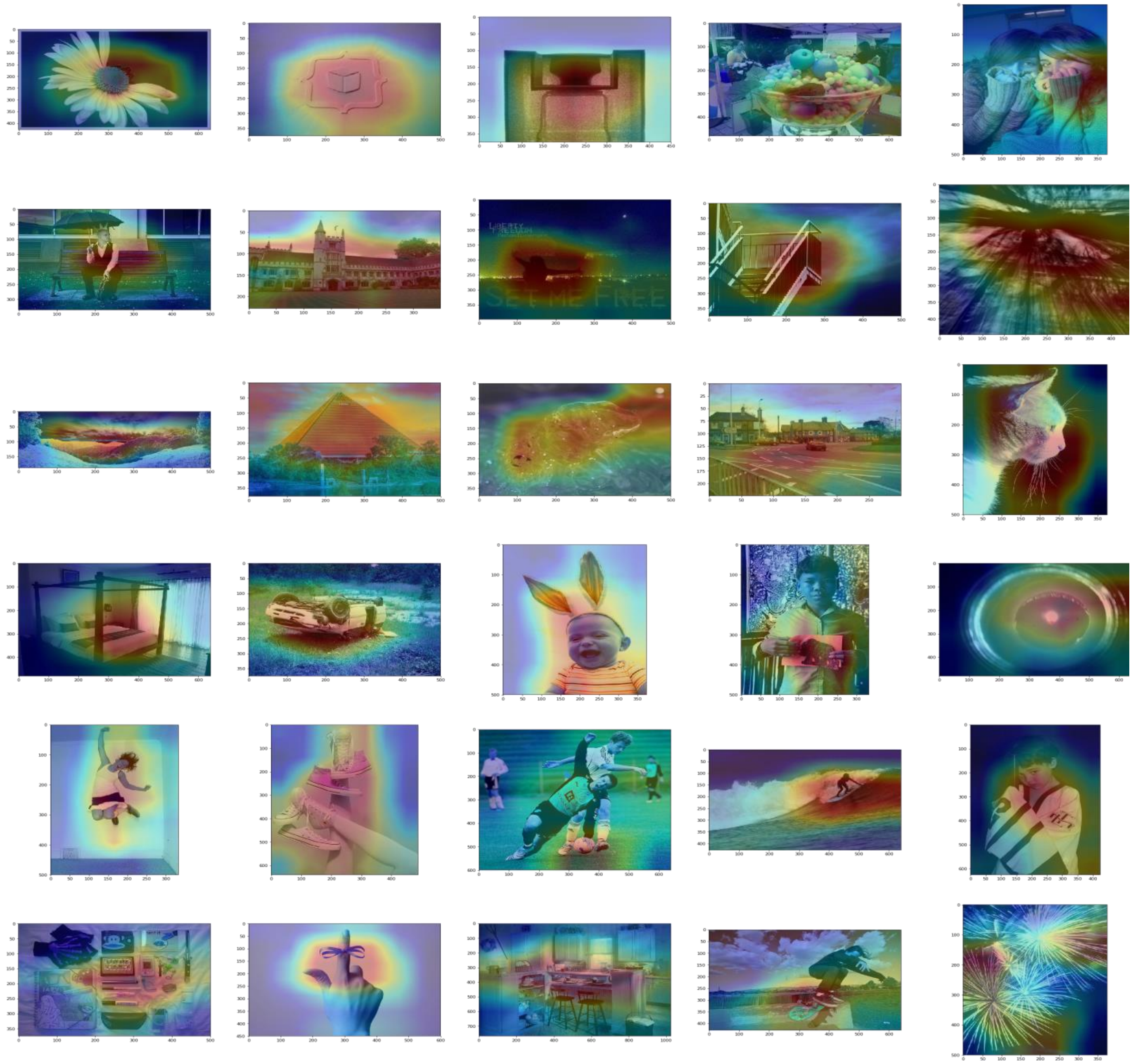


Fig 4. Heatmaps showing the region of the image that provided higher activations in V-Net

6. Conclusion:

Capturing memorable pictures is a very challenging task and it requires an enormous amount of creativity. However, just like any other phenomenon in nature, humans' capability to remember certain images more follows a pattern. Deep Learning based strategies haven been proven to be excellent when it comes to solving problems that involve pattern recognition. Thus, this paper introduced V-Net, a novel neural network architecture that provides close to human performance on predicting the memorability of an image using the largest publicly available dataset for image memorability, LaMem. The rank correlation of V-Net is 0.679 which is extremely close to human accuracy which is 0.68. This result is a 6.09% increase from the performance of MemNet, 35% increase from CNN-MTLES, a 2% increase from MCDRNet and an 1.2% increase from EMNet.

As mentioned in section 5, it was inferred that the object plays a bigger role in enhancing the memorability of the image and that a pretrained model that consists of weights from an object classification dataset converges quickly than a model pretrained on scene classification. These results were observed by manually looking through the highly rated images and lowly rated images. Heatmaps generated using GradCam method were also used to analyze and obtain the above inferences.

References:

(1) https://www.sciencedirect.com/science/article/pii/S1053811917300861 Memorability: A stimulus-driven perceptual neural signature distinctive from memory,

(2) Is Image Memorability Prediction Solved? http://openaccess.thecvf.com/content_CVPRW_2019/papers/MBCCV/Perera_Is_Image_Memorability_Prediction_Solved_CVPRW_2019_paper.pdf

[3] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. In Proceedings of the National Academy of Sciences, 2008

[4] R. R. Hunt and J. B. Worthen. Distinctiveness and memory. In NY: Oxford University Press, 2006.

[5] Understanding the Intrinsic Memorability of Images, http://web.mit.edu/phillipi/www/publications/UnderstandingMemorability.pdf

[6] **S. Kong**, X. Shen, Z. Lin, R. Mech, C. Fowlkes, "Photo Aesthetics Ranking Network with Attributes and Content Adaptation", *ECCV, Amsterdam, the Netherlands, (Oct. 2016).*

[7] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages145–152. IEEE, 2011.

[8] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007

[9] R. Hunt and J. Worthen Distinctiveness and Memory. New York, NY: Oxford University Press, 2006.

[10] Learning Color Names from Real-World Images https://ieeexplore.ieee.org/abstract/document/4270243

[11] L. Liu, R. Chen, L. Wolf and D. Cohen-Or Optimizing Photo Composition. EUROGRAPHICS, 2010

[12] Understanding and Predicting Image Memorability at a Large Scale, http://openaccess.thecvf.com/content_iccv_2015/papers/Khosla_Understanding_and_Predicting_ICCV_2015_paper.pdf

[13] Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. Journal of Experimental Psychology: General, 139(3), 558–578. https://doi.org/10.1037/a0019165

[14] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences, 105(38):14325–14329, 2008

[15] Multiple instance learning based deep CNN for image memorability prediction

https://doi.org/10.1007/s11042-019-08202-y

[16] AMNet: Memorability Estimation with Attention : https://arxiv.org/pdf/1804.03115.pdf

[17] Baveye, Y., Cohendet, R., Perreira Da Silva, M., & Le Callet, P. (2016). Deep Learning for Image Memorability Prediction. Proceedings of the 2016 ACM on Multimedia Conference - MM '16. doi:10.1145/2964284.2967269

url to share this paper:
sci-hub.tw/10.1145/2964284.2967269

[18] DEEP LEARNING FOR PREDICTING IMAGE MEMORABILITY

https://hal.archives-ouvertes.fr/hal-01629297/document

[19] Celikkale, B., Erdem, A., & Erdem, E. (2015). *Predicting memorability of images using attention-driven spatial pooling and image semantics. Image and Vision Computing, 42, 35–46.* doi:10.1016/j.imavis.2015.07.001

url to share this paper:
sci-hub.tw/10.1016/j.imavis.2015.07.001

[20] Understanding and Predicting Image Memorability at a Large Scale
Aditya Khosla, Akhil S. Raju, Antonio Torralba, Aude Oliva; The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2390-2398

[21] Long short term memory https://www.bioinf.jku.at/publications/older/2604.pdf

[22] Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio

https://arxiv.org/abs/1406.1078

[23] Diederik P. Kingma and Jimmy Lei Ba. Adam : A method for stochastic optimization. 2014. arXiv:1412.6980v9

[24] Standing, L. (1973). Learning 10,000 pictures.Quarterly Journal of Ex-perimental Psychology, 25,207–222.

[25] The Intrinsic Memorability of Face Photographs Wilma A. Bainbridge, Phillip Isola, and Aude Oliva Massachusetts Institute of Technology https://www.wilmabainbridge.com/papers/jepg-2013.pdf

[26] https://arxiv.org/pdf/1312.4400.pdf

[27] A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay
[28] P. Jing, Y. Su, L. Nie, and H. Gu. Predicting image mem-orability through adaptive transfer learning from externalsources.IEEE Transactions on Multimedia, 19(5):1050–1062, 2017.

[29] Grad-CAM: Visual Explanations from Deep Networksvia Gradient-based Localization

[30] Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun