GEMM: A Graph Embedded Model for Memorability Prediction

Tahsin Tariq Banna*, Swakshar Deb*, Sejuti Rahman* and Shafin Rahman†
*Department of Robotics and Mechatrionics Engineering, University of Dhaka, Bangladesh
†Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
{tahsinbanna, swakshar.sd}@gmail.com, sejuti.rahman@du.ac.bd, shafin.rahman@northsouth.edu

***Abstract*—The focus of this study is to predict human memorability - a person's ability to remember previously seen images or objects. Although recent works have employed deep learning-based approaches to address the problem, they do not utilize spatial structural information within the images. This work investigates Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) to approach the problem. The object-centric features within the images are extracted using deep CNN-based models, which contain the structural information of the image. A generic baseline model is created and improved upon iteratively through structural data by constructing graphs and attention mechanisms on the graph edge connections. The constructed graph nodes represent the objects within the image, and the edge connections between the nodes represent the spatial relation to the objects. These graph embeddings are used to train our proposed Graph Embedded Memorability Model (GEMM), which shows significant improvements from the baseline as the attention improves the edge connections of the graph nodes. The model is then evaluated on the LaMem, SUN memorability, and FIGRIM datasets. Although existing state-of-the-art models perform well on one or two datasets, the proposed model generalizes over all three datasets with a Spearman's rank correlation of 0.71 on LaMem, 0.69 on SUN memorability, and 0.59 on the FIGRIM dataset. This model achieves a new state-of-the-art performance compared to the existing literature.**

I. Introduction

In our daily lives, we constantly use visual information to convey a message to other people. We interact and communicate with other humans and read their expressions through the visual senses. Rather, our interaction with anything largely depends on the proper visual information we can process. Examples of this might be a teacher trying to teach a concept to a student, a company trying to communicate a message to a customer through an advertisement, or even a software developer trying to develop a more memorable Graphical User Interface (GUI) for improved user experience. The messages we convey are most effective when we can remember it for a long time. This is where the concept of memorability comes in. Image memorability is a person's ability to later remember or forget a previously shown image. This is quantified as a probability of the person remembering the image and the delay between two observations of an image does not impact this probability. Besides just conveying messages, memorability has a wide area of applications, such as measuring mild cognitive impairments (MCI) and detecting dementia, mental disorders, autism, and other neurodegenerative diseases at an early stage.

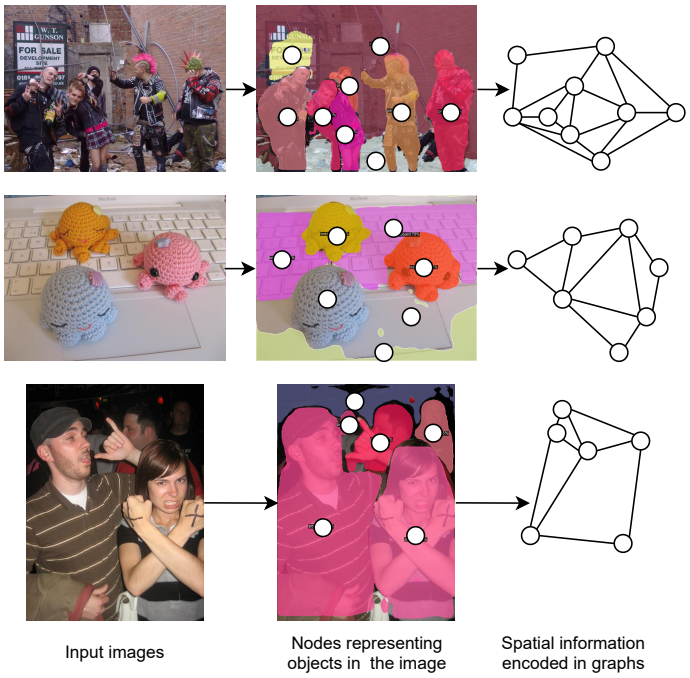


Fig. 1. Spatial structural information of images represented using graphs. It shows a core idea of our model, where we capture the object-centric information of the image by representing each object as nodes and connecting them using edges to create a graph representation. In the images, it can be seen that humans, background structures, and other objects like signboards and dolls are represented as objects with associated nodes. The graph constructed from these nodes contains crucial information about the image and is used to train our proposed model.

Contemporary deep learning methods can handle regularly structured data well, such as images and texts but fail when the data is irregular. For example, Transformers [1], and Recurrent Neural Networks (RNNs) [2], [3] perform well on a deterministic sequence of data in one dimension, while Convolutional Neural Networks (CNN) [4], [5] perform really well on grid-like structures such as images. But from the observations of [6]–[9], It is clear that the memorability of an image is not related to patterns or locations in the image that are well structured. Rather it is heavily correlated to object-centric features and how the object is positioned in the image in relation to its surroundings. As a result, the use of models designed for structured data has not been well generalized among all datasets and has performed poorly on many fronts.

In this work, we examine the application of GCN [10] to learn image memorability and how memorability correlates to the spatial structural information of the image. The reason behind using GCN is that we intend to handle images in an object-centric way, with the location and relation of objects and their neighbors represented using graphs. For this purpose, we propose the Graph Embedded Memorability Model (GEMM), which uses graph embeddings of object representations to capture the spatial structure of the image. Furthermore, we capture the contextual mismatch (i.e., unusualness) and local contextual information (i.e., facial expressions, anger, happiness, and other emotions) of a scene. Although there have been attempts to correlate local unusualness [7], [9] and emotions [11], [12] with memorability separately, there has not been any unified approach.

To the best of our knowledge, this is the only study on image memorability that utilizes the recent developments in Graph Neural Networks. In summary, the main contributions of this work include:

- Implementation of a fully automated end-to-end deep learning method that leverages spatial information to predict memorability.
- Demonstration of self-attention mechanism in reducing graph complexity.
- Generalization of the model over several datasets and use cases establishing a new state-of-the-art performance.

II. Related Work

A. The Memorability Game

Isola et al. [13], [14] pioneered the field of image memorability in 2011 and introduced the Visual Memorability Game to understand and calculate memorability of images. They created the SUN memorability dataset [14], and since then, other researchers have used similar or slightly modified versions of the game to construct their own datasets. The game's central theme is to show images to the participants and see whether they respond to repetitions of certain types of images. These responses are used to give memorability scores to the images. Studies on memorability have shown that it is a stable property that does not vary over time [15] or context [16]. Research has also been done on images of scene categories and objects, which show that the memorability of images largely depends on the objects present within an image and their relative positions to each other. [13] and [14] used handcrafted features such as pixel histogram, GIST [17], SIFT [18], SSIM [19] and HOG [20]–[22] to predict memorability scores.

B. Using Hand-crafted Features

The memorability of images is affected by several local attributes and features, such as the placement of image regions and local attributes in scenes. Researchers have shown that these features and attributes contribute significantly to the memorability of images. One study proposed by Khosla et al. [6], [23] in 2012 proposed a model to predict the memorability of image regions, as people tend to remember some objects and image regions more than others. They proposed an image encoding process using noisy memory through a probabilistic framework, assuming that different image regions correspond to different object groups and have different probabilities of being remembered or forgotten. Additionally, Kim and Yoon et al. [9] proposed in 2013 two new spatial features that contribute to image memorability: Weighted object Area (WOA) and Relative Area Rank (RAR). WOA considers the size and position of objects in an image, while RAR correlates to the relative unusualness of the size of an object. Studies [11], [16], [24] have also found that naturally pleasing or visually aesthetic scenes lack unusualness and negatively affect memorability. Visual saliency [6], [11], [25], [26], image popularity [11] and emotions [11], [12] on the other hand, correlate positively with image memorability.

C. Deep Learning

Improvements in the field of deep learning helped generate models that predicted memorability much better than hand-crafted models. One such model is MemNet [11], proposed by Khosla et al. in 2015. It is a Hybrid-CNN model trained on the ILSVRC 2012 [27] and Places dataset. They also created a baseline model which used HOG2x2 features and showed that it performed better when accounting for False Alarms (FA), which helps to minimize signal noise. Furthermore, it was found that the model performed worse when fine-tuned on a smaller dataset like the SUN Memorability Dataset, which highlights the need for a larger dataset. This led to them constructing the Large Scale Memorability Dataset (LaMem). Fajtl et al. proposed AMNet [8] in 2018, which used transfer learning from deep models to understand memorability better. They used ResNet-50 [28] trained on ImageNet as their base classification model and found that deep models that perform well on classification tasks also perform well on predicting memorability. Additionally, this study explains why end-to-end deep learning features outperform specially designed visual features and features taken from CNN models. Squalli-Houssaini et al. [29] proposed a deep learning model that uses a combination of CNN and Image Captioning (IC) features to predict image memorability. They use a VGG16 model pre-trained on ImageNet to extract features and convert the memorability scores into four balanced classes. They then use an encoder consisting of a CNN and LSTM to create a text-image embedding of the image captions in 2D. This enforces conformity of the image with its corresponding semantic caption, which contains relevant semantic information about memorability. The final memorability score is calculated through a memorability aggregation step and shows that this method effectively predicts memorability. On the other hand, Perera et al. [30] proposed only training a regression network on the final layer of a strong CNN model. They showed significant improvement in prediction scores using VGG16 as their base CNN model. Even though RAR, WOA, and other object-centric features show a good correlation with memorability, there has been virtually no research on utilizing these features for memorability score prediction. So, in this paper, we attempt

to use Graph Convolutional Networks (GCNs) [10] to build a deep learning model that leverages this spatial structural information.

III. Proposed Models

GCN is a natural choice for solving memorability prediction as it can better understand inherent structural data. This is especially useful since the memorability of images depends mainly on the objects within the image and the relations between these objects, which can be represented using graphs. These interconnections form a graph-like structure that contains crucial information about the nature of the image that a GCN-based model can learn far better than any CNN or RNN-based methods. Our final design incorporates GCN to utilize these graph representations of images to find the overall memorability score of the image. We first discuss a baseline model created using generic GCN. We then improve this model by incorporating an attention mechanism with Graph Attention Network (GAT) [31]. Finally, we ensemble it with an existing model, AMNet [8], to propose our novel method. Since the proposed model uses graph embeddings of images to estimate the memorability of images, we name our model the "Graph Embedded Memorability Model (GEMM)".

A. Problem formulation

Suppose, ith image of a dataset $\mathbf{d}$ is represented by $\boldsymbol{V_i}$ of dimension $W_i \times H_i \times D_i$. The ground-truth score of each of these images is, $y_i \in [0, 1]$ which represents the overall memorability score of the image.

A higher y_i score means a highly memorable image and a lower y_i score means the image is less memorable. Within the training dataset a set of tuples $\{(\boldsymbol{V_i}, y_i) : i \in [0, \mathcal{T}]\}$ is included, where the total number of training images within the dataset $\mathbf{d}$ is represented by $\mathcal{T}$. We train the θ_d parameterized model, $\mathcal{F}$ end-to-end, for the dataset $\mathbf{d}$ so that it can predict score, $\hat{y}_j$ very close to the original ground truth score, y_j^* for a given image V_j. So, $\hat{y}_j$ is formulated as follows:

$$\hat{y}_j = \mathcal{F}(V_j; \theta_{\mathbf{d}}), \quad s.t. \quad \hat{y}_j \approx y_j^* \tag{1}$$

We incorporate the spatial structural information of the image in the model. So, we generate a graph from the image that represents the structural information. Let's call this graph, $G = (\mathcal{A}, \mathcal{V}, \mathcal{E})$ where, $\mathcal{V}$ is set of vertex, $\mathcal{E}$ is a set of edges and $\mathcal{A} \in \mathbb{R}^{N \times N}$ is the edge adjacency matrix of the graph G. The vertices are the extracted structural features of the image and the edges are the relationships between these features represented as connections. The adjacency matrix indicates which of the vertices are connected to others using a matrix.

Now, for each of the training images V_i in the dataset $\mathbf{d}$ we have a feature graph representation of G_i. So, the equation 1, for predicting a continuous score, $\hat{y}_j$ near the ground-truth, y_j^* for a given feature G_j becomes,

$$\hat{y}_j = \mathcal{F}(G_j; \theta_{\mathbf{d}}), \quad s.t. \quad \hat{y}_j \approx y_j^* \tag{2}$$

Different parts of the image play essential roles while measuring the overall memorability [6], [7], [9], [23]. Empirical results show that such roles vary from one image to the other. Let, $\mathbf{M}_j \in \mathbb{R}^{W_j \times H_j \times C_j}$ be the self-attention map of the different parts of the image representing their overall importance for jth image. By analyzing $\mathbf{M}_j$ of the image within the dataset $\mathbf{d}$, we can get reasonable insights about prediction scores along with how different objects contribute to the memorability score.

B. Model Overview

Consider we have an image $\boldsymbol{V_i}$ and its graph embedding, G_j consisting of n nodes, which are specified as sets of node features, $(\vec{h_1}, \vec{h_2}, \ldots, \vec{h_n})$ along with an adjacency matrix $\mathbf{A}$, such that $\mathbf{A}_{ij} = 0$ if i and j are not connected, and 1 if they are connected. A GCN layer computes a new set of node features, $(\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_n)$, based on the graph structure and the input features.

To get higher-level representations, every GCN layer starts with shared node-wise feature transformations, which are specified using a weight matrix $\mathbf{W}$. This transforms the feature vectors into $\vec{g}_i = \mathbf{W}\vec{h}_i$. Afterwards, the vectors $\vec{g}_i$ are usually combined at every node in some way.

Generally, to satisfy the property of localization, a graph convolutional operator is defined as a combination of features across the neighborhoods; defining N_i as the surrounding neighborhood of the node i. The final output features of node i are then defined as:

$$\vec{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \vec{g}_j\right) \tag{3}$$

where, σ is the activation function, and α_{ij} specifies the weight factor or importance of node j's features in relation to node i.

We use this layer that maps an input graph embedding G_j to a feature vector output of $\vec{h}'_i$ as the building block for our model. We create a generic GCN architecture with 3 GCN layers. It starts out with the extracted image features as an input to the network. The inputs are a graph embedding of 1000 nodes, each with 1024 features. These are passed through 3 GCN layers with relu activations. The final GCN layer sends the features to a mean pooling layer where the features are converted to a vector of 1000 elements. This is then sent to a network of fully connected dense layers with 1 hidden layer of 64 neurons and a single output neuron. This final dense layer is the regression network that gives us the overall memorability score of the image. GCNs are better at understanding long-ranged interaction between objects better than CNNs. Typically, CNNs need to be very deep in order to encode these connections. A GCN can achieve the same result with a shallower network by simply having an edge connecting the two points.

C. GEMM

In order to improve the generic GCN model and incorporate graph attention in our model, we define α_{ij} implicitly and use self-attention over the graph node features. Veličković et

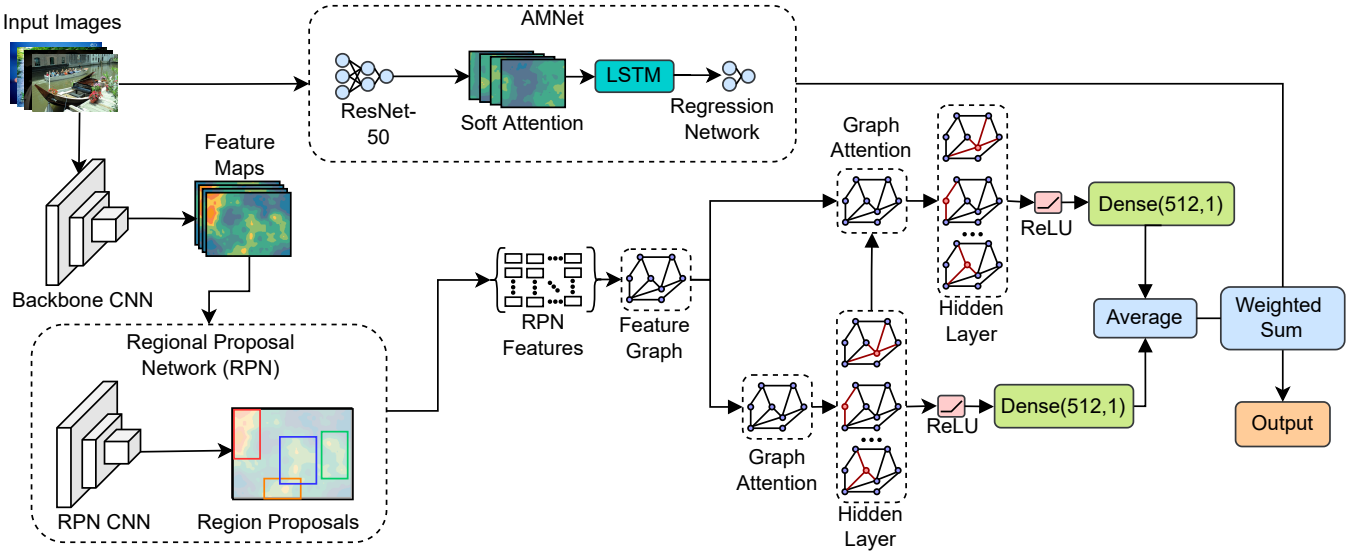


Fig. 2. GEMM: An ensemble of the AMNet [8] model and attention guided GCN for memorability prediction. It takes the final layer output of the ensemble model and passes it through a weighted sum layer to generate the final output. The graph embedding is generated from the Regional Proposal Network (RPN) of the Mask R-CNN model. The output feature matrix is of dimension 1000×1024 with 1000 regions of object proposals and each with a vector of 1024 features.

al. [31] employed this idea as it was in fact self-sufficient for top results on machine translation, as demonstrated by the Transformer architecture [1].

Generally, α_{ij} is computed as a by-product of attention mechanism, $a : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$, which computes unnormalised coefficients e_{ij} across pairs of nodes i, j, based on their features:

$$e_{ij} = a(\vec{h}_i, \vec{h}_j) \tag{4}$$

The graph is injected by only ever allowing node i to attend over the nodes within its neighborhood, $j \in \mathcal{N}_i$. The coefficients for each neighborhood are typically normalized using the softmax function to allow for comparison across neighborhoods:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \tag{5}$$

The framework is not affected by the specific attention mechanism used (represented by the variable a) and can be trained along with the rest of the network in a single, unified training process. Thus, using this update function, the attention-guided block performs better than that of the generic GCN. Although the attention-guided graph convolutional network is implemented from an object-centric view in order to capture the relationships between objects in a scene, it can recognize contextual mismatch (i.e., unusualness). However, this ignores the local contextual information in a scene, such as emotions (e.g., anger, happiness), that can further influence the memorability score. Therefore, to incorporate both the information from a scene, we ensemble AMNet [8] as our local contextual network to improve the performance further. Thus, the proposed model can capture a scene's contextual mismatch and local contextual information.

IV. Experiemnts

Datasets: We have used three datasets that have been utilized by current state-of-the-art methods. They are the LaMem dataset [11], the SUN memorability dataset [14] & the FIGRIM dataset [15]. These datasets were created with versions of the original memorability game. All the data were collected from human participants and the scores were annotated by the statistical performances of the participants. *(1) LaMem* The images for the LaMem dataset was sampled from other datasets such as MIR Flickr [32], AVA dataset [33], SUN [20], image popularity dataset [34], Abnormal Objects dataset [35], aPascal dataset [36], affective images dataset [37], MIT1003 [38] and NUSEF [39] dataset. It contains 58,741 images, which is approximately 27 times the size of the next largest dataset, containing large variance within its images. The images are split into train, test, and validation sets for easier implementation. Out of the 58,741 images, the train splits contain 45,000 images, the validation set contains 3741 and the test set contains 10,000 images. *(2) SUN Memorability Dataset* The SUN dataset [20] is a large dataset created for the sole purpose of providing an annotated collection of images for various research purposes such as neuroscience, robotics, computer vision, human cognition, perception, etc. It consists of 108,753 images in 397 different categories. The SUN Memorability dataset [14] was created by randomly sampling the SUN dataset [20]. The images were cropped and scaled to a 1:1 proportion of 256 pixels. 8220 images were selected as filter images and 2222 as target images. Among these, there was a 50-50 split for the training and testing images; both sets contained 1111 images. *(3) FIGRIM* The Fine Grained Image Memorability (FIGRIM) [15] Dataset was also subsampled form the SUN dataset [20]. The images were sampled from

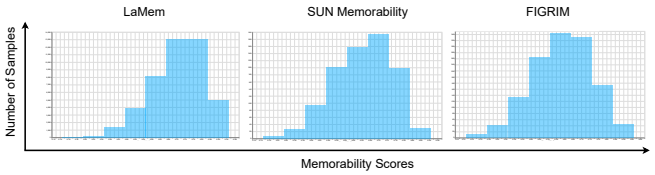


Fig. 3. Visualization showing the histogram of memorability scores distributed over the images in the three datasets. It shows that most of the images within the dataset have a memorability score between 0.6 to 0.8.

21 different scene categories and near duplicate images were manually removed. The images were then cropped to 1:1 resolution of 700 pixels. A total of 9428 images were selected and a quarter of the images were chosen as targets while the rest were fillers. The target images consisted of 1754 images, and the filler images consisted of 7674 images.

Finally, a histogram of memorability score distributed on the datasets is shown in Figure 3. It can be seen that most of the images within the datasets are within the 0.60 to 0.80 score range.

Evaluation metric: Prediction of image memorability falls under the category of supervised regression tasks. Usually, regression models are evaluated using error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc. But most of the studies on image memorability [8], [11], [13], [14], [29], [30] evaluate their model on Spearman's rank correlation score which measures the monotonic relation between two functions. So, even if the two functions are not equal or are even linear, if the data points have greater input (i.e., the domain) values than that of a given data point, they will also have a greater output (i.e., range) value. In a way, this is similar to the Pearson correlation. Still, the Pearson correlation only calculates the linear relation between two functions, whereas Spearman's rank correlation calculates the monotonic relation no matter whether they are linearly related or not. RMSE is used in [8], [14] in order to find the overall errors of the models. This often gives empirical results to show if a model overfits or underfits to a dataset.

Compared methods: We compare our work to four categories of methods: *(1) Hand-crafted Feature Models*: Some of the earliest methods [13], [14], [16] used pixel histograms, HOG, SSIM, etc features to predict memorability scores. *(2) Image Features*: Most deep learning methods [6], [8], [11], [30] either learn features from the images or use features from existing CNN to predict scores. Some of these methods [8], [30] proposed learning only the final regression output of a pre-trained classification model. *(3) Image Captioning (IC) Features*: One of the methods [29] uses CNN features from a pre-trained model along with IC features from the word2vec representation of manually annotated objects. *(4) Baseline*: We propose a baseline GCN model for comparison with our method. We create a generic and attention-guided GCN model and compare both with our proposed method.

TABLE I

THE TABLE SHOWS THE SPEARMAN'S RANK CORRELATION (ρ) SCORE OF DIFFERENT MODELS COMPARED ON THE THREE DATASETS. IT CAN BE SEEN THAT THE PROPOSED FINAL METHOD, GEMM-2 OUTPERFORMS ALL OTHER MODELS ACHIEVING STATE-OF-THE-ART PERFORMANCES ON IMAGE MEMORABILITY PREDICTION. THE BEST-PERFORMING MODEL IS SHOWN IN **GREEN** AND THE SECOND BEST IN **BLUE**.

Models	Datasets		
	LaMem	SUN memorability	FIGRIM
AMNet [8]	0.68	0.65	0.49
MemNet [11]	0.64	0.63	-
Isola et al [14]	-	0.46	-
MemBoost [30]	0.67	**0.66**	**0.57**
Squalli-Houssaini et al [29]	**0.72**	0.59	0.48
Human consistency	0.68	0.75	0.74
Baseline	0.58	0.54	0.46
GCN (with attention)	0.62	0.61	0.54
GEMM	**0.71**	**0.69**	**0.59**

Implementation details[1] Since the datasets consisted of images but the model utilized graph structures, the data had to be preprocessed before the features were extracted and sent to train the model. The images were loaded one by one and reshaped into a size of 250×250px. They were randomly cropped and rotated about their centers.

These images were then sent to a Mask R-CNN [40] for object segmentation. It uses regional proposal networks (RPNs) to efficiently generate a set of object proposals, or regions of interest, that are likely to contain objects in an image. It uses the feature pyramid network (FPN) of a Mask R-CNN with ResNet-50 [28] pre-trained weight as the base model. The features are extracted from the 2nd fully connected layer of the RPN from the model. This returns a 1000×1024 sized matrix of 1000 regions containing possible important objects, each with 1024 features. Each of these 1000 regions is treated as an object considered a node for the graph. Initially, the 1000 nodes are fully connected with each other forming a graph. This graph from a single image is considered the graph representation or embedding of the image. We then send this graph embedding to the model for training a GCN model and a final regression network to predict the memorability score of the input image. The model would still perform well if objects in the images are sparse because the number of proposed regions stays the same regardless of object sparsity. Memorability is not correlated with object sparsity. Rather, it is dependent on the unique number of fixation points [41] in the image.

A. State-of-the-art results

In Table I, we compare our model with the existing methods, showing the performance of different datasets on different models. The performance of the models is shown using Spearman's rank correlation coefficient. The table shows existing models along with the human consistency on the top and then shows the proposed baseline models and GEMM below. Our observations from the results are as follows: *(1)* Using hand-crafted features is not a good idea as it does not capture even

[1]Codes and evaluations are available in https://github.com/TahsinTariq/GEMM

TABLE II
THE ROOT MEAN SQUARED ERROR (RMSE) OF THE PROPOSED MODELS COMPARED ALONG WITH THE EXISTING METHODS. THE BEST-PERFORMING MODEL IS SHOWN IN BOLD. IT CAN BE SEEN THAT RMSE IS AN INDICATOR OF THE OVERALL ERROR OF THE MODEL AND CAN EXPLAIN THE PERFORMANCE OF THE MODEL WHEN OVERFITTING.

Models	Datasets		
	LaMem	SUN memorability	FIGRIM
AMNet [8]	0.008	0.011	0.016
Isola et al [14]	-	0.017	-
Baseline	0.011	0.013	0.021
GCN (with attention)	0.010	**0.005**	0.014
GEMM	**0.008**	0.021	**0.013**

TABLE III
THE BASE MODEL WAS A MODEL WITH 3 GCN LAYERS AND 4 FULLY CONNECTED HIDDEN LAYERS. THE FULLY CONNECTED LAYERS WERE RELU ACTIVATED WITH THE OPTIMIZER BEING ADAM AND THE LOSS FUNCTION BEING HUBER LOSS. THE EXPERIMENTS WERE CONDUCTED ON THE SUN MEMORABILITY DATASET.

GCN layers	Loss Function		Activation Function		
	Huber	**MSE**	**ReLU**	**Sigmoid**	**Tanh**
1	0.685	0.682	0.682	0.641	0.641
2	0.675	0.678	0.669	0.661	0.681
3	**0.691**	0.689	0.689	0.641	0.671
4	0.672	0.683	0.669	0.666	0.685

the entirety of the information within the image. Because of this, the model by Isola et al. [13], [14] performs the worst among all the models compared. *(2)* When comparing with methods that utilize learned features or features from CNN models, we see that even our baseline GCN with attention achieves similar results while GEMM achieves state-of-the-art performance on all three datasets. It is also interesting to note that on the FIGRIM dataset, our baseline generic GCN model achieves similar performance on models that have outperformed human consistency on the LaMem dataset. Furthermore, this is done so without any finetuning of the model. *(3)* Using IC features, Squalli-Houssaini et al. [29] reported the highest performnce on the LaMem dataset. However, his model did not generalize over the other two datasets and in fact performed significantly worse than all other methods. While our proposed GEMM did not outperform his method on the LaMem dataset, it was well within the margin of error. On the contrary, GEMM not only performed better than his method, but it also did so by a very large margin. It is also seen that the attention-guided GCN also outperformed his models on both the SUN memorability and the FIGRIM dataset. *(4)* GEMM outperforms the baseline model by a huge margin. This can be attributed to the fact that even though the baseline has all the available spatial structural information of the image, it does not know which part of the image to give more importance over the other. This is because, in generic GCNs, the weight of the edges connecting the nodes is set to one. Because these nodes represent objects in the image, the model gives the same importance to all the objects. But since memorability is largely dependent on specific objects in the image, this method does not perform that well. This however changes when we introduce the attention-guided GCN. This model significantly increases the performance of the generic GCN. This is because the graph attention mechanism is better capable of understanding which object nodes consist of features that influence memorability. This performance is improved even further by ensembling it with the AMNet model, which we propose as GEMM.

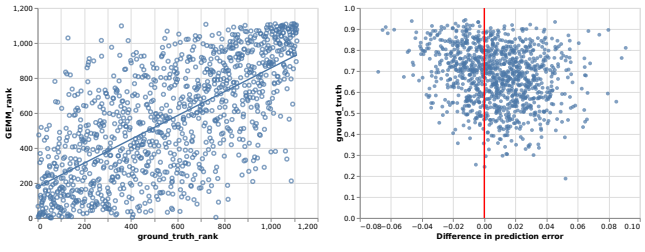
The Spearman's rank correlation is a good indication of the monotonic relation between the reference and observations. But it is unable to indicate the numeric error within them. This error is better represented by the RMSE error values. In Table II we show the RMSE error of the available models on the three datasets. The table shows that as the models perform better and better, their RMSE error decreases. This is expected as the model better approximates the memorability scores and is clearly illustrated in the RMSE errors on the FIGRIM dataset. However, oftentimes models may overfit to a dataset and fail to produce a reasonable performance even though it has a lower overall RMSE error. This is evident with the SUN memorability dataset, where even though attention guided GCN has a very low RMSE error, it does not reflect so in with its Spearman's rank correlation score. It is clear that the model overfits on the dataset to some degree. This can be seen alleviated in GEMM as the RMSE error increases and so does the performance. As a consequence, the model does not overfit to the data and achieves state-of-the-art on the dataset.

B. Ablation study

In Table III we perform ablation studies on different parts of the GEMM architecture. We show the performances of different components as we increase the GCN layers of the model on the SUN memorability dataset. In all cases, the performance of the model increases with the GCN layers up to 3 layers. It adds a fourth layer showing a significant performance drop in the performance. This is due to the message-passing layer of the GCN model becoming deeper and deeper. Nodes in the graph aggregate the messages from their neighboring nodes which in turn gather messages from even deeper neighbors. At some point, all of the nodes in the graph aggregate messages from all other nodes because of the depth traversed. This smoothens out the messages and the performance no longer improves as expected. In our model, this GCN smoothing happens at the fourth GCN layer. We also show that using Huber loss is preferable to MSE loss as Huber loss is less sensitive to outliers in the data and is generally used for robust regressions. We also compare the different activation functions which show that the sigmoid function does not perform that well while ReLU and $\tanh$ perform very similarly to each other. ReLU is preferred over $\tanh$ as it not only performs better, it has a more consistent performance.

C. Qualitative results

Figure 4 shows two plots that explain why GEMM performs better over other methods on the SUN memorability dataset. In particular, we compare it with the AMNet [8] model.



(a) (b)

Fig. 4. (a) Ground truth and GEMM prediction ranks on the SUN memorability dataset. The rank correlation score is 0.69. (b) Each point represents an image with the Y axis showing the ground truth memorability score and the X axis showing the difference between the AMNet and GEMM prediction errors. The points with positive values (right of the red line) show a win for GEMM and negative values show AMNet wins. If the error distribution was symmetric, their performances would have been similar. But since more points are on the positive side, it indicates GEMM performing better.

The first plot shows Spearman's rank correlation of GEMM predictions compared to ground truth ranks. The points in the plot represent images, with the X-axis being the rank of the ground truth score and the Y-axis representing the rank of the predicted score of GEMM. This gives an overall Spearman's rank correlation score of 0.69 for the dataset. We can see how GEMM performs better than AMNet in the second plot, which shows the difference in prediction errors for the image samples. The red line represents zero prediction error between the two models. The images on the right side of the red line are closer to the ground truth for GEMM predictions and the images on the left for AMNet predictions. The plot is not evenly distributed about the center red line, which means that there is a clear difference between the two model performances. This difference is that the plot leans heavily to the right-hand side of the red line, representing more prediction wins for GEMM over the AMNet model.

We show some images with varying memorability from the three datasets in Figure 5. The memorability of shown images increases from left to right in each dataset. The AMNet and GEMM prediction is also shown for the images. The images with high memorability scores contain features that stand out from the scene. Most of them seem to be images of humans, animals, or human-like depictions. Fewer features stand out for images with medium memorability scores, but still, the different parts of the images are clearly defined. There are hardly any distinguishing features for images with low memorability. The scenes are generic and flat, with few objects standing out. There are not any humans or animals present in these images either. The prediction scores show that GEMM is almost always closer to the ground truth performance than AMNet. This performance is just for one dataset; it is consistent over all the datasets.

V. Conclusion

In this work, we propose GEMM, an end-to-end automated deep learning model for predicting memorability. A lot of the current literature use hand-crafted features and those that have developed an end-to-end model do not leverage the contemporary developments in deep learning techniques. Their limitation can be summarized in a number of ways. Firstly, even though memorability is largely dependent on the relative positioning of the objects within the image [6], [7], [9], these models fail to capture their spatial information in a meaningful way. Secondly, these methods are specialized to one or a few datasets and do not generalize over a large amount of data. This is rather important as the variability in real-world data can drastically change the overall perception of memorable images [11]. Finally, only a handful of studies have focused on using the attention mechanism, which has been proven to be very effective in predicting memorability scores [8], [29]. To overcome these limitations we have introduced a Graph Convolutional Network (GCN) based approach leveraging Graph Attention (GAT) mechanism for memorability predication. We extract the spatial information of an image and employ an attention mechanism to learn the importance of the edges of the graph in order to understand the object-centric relations. We use these attentions to train our GCN-based model end-to-end to predict the overall memorability of images. We show that our model can generalize over a large number of datasets and validate our performance claims by comparing our results to current literature. We have found that on the two of three most widely used datasets (SUN memorability dataset [14] and FIGRIM dataset [15]), our model achieves state-of-the-art performance. Furthermore, it is on par with existing works on the third dataset, the LaMem dataset [11].

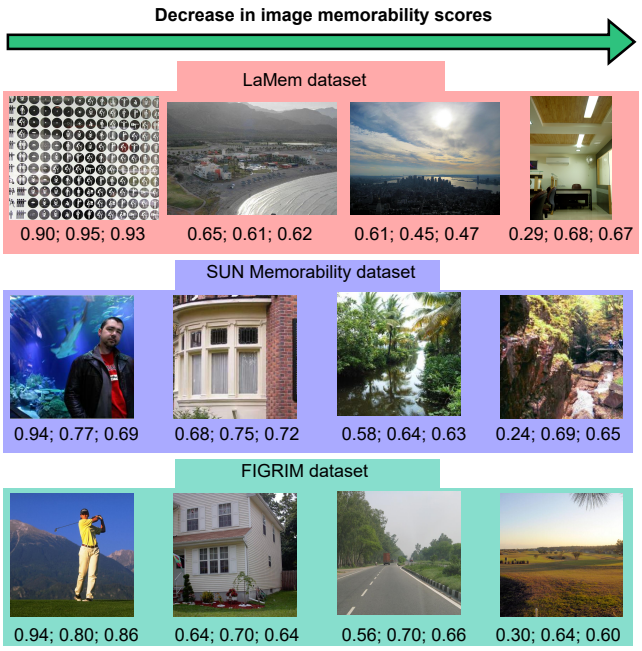


Fig. 5. Qualitative results of image memorability. The high memorability images contain distinguishable objects within the scenes, whereas as the memorability decreases, so do features that "stand out". For each of the images, the numbers indicate the following: Ground Truth Score; Predicted scores from AMNet [8]; Predicted scores from GEMM.

Acknowledgement: This work was supported by the Ministry of Science and Technology Special Research Grant Project 2022-23 and the North South University (NSU) Conference Travel and Research Grants (CTRG) 2021–2022 (Grant ID: CTRG-21-SEPS-10).

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[2] J. Chung, Çaglar Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *ArXiv*, vol. abs/1412.3555, 2014.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural computation*, vol. 9, pp. 1735–80, Dec. 1997.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017.

[6] A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva, "Image memorability and visual inception," in *SIGGRAPH Asia 2012 Technical Briefs*, ser. SA '12. New York, NY, USA: Association for Computing Machinery, 2012.

[7] S. Yoon and J. Kim, "Object-centric scene understanding for image memorability prediction," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 305–308.

[8] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, "AMNet: Memorability Estimation with Attention," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 6363–6372.

[9] J. Kim, S. Yoon, and V. Pavlovic, "Relative spatial features for image memorability," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 761–764.

[10] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2017.

[11] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and Predicting Image Memorability at a Large Scale," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2390–2398, iSSN: 2380-7504.

[12] Y. Baveye, R. Cohendet, M. Perreira Da Silva, and P. Le Callet, "Deep learning for image memorability prediction: The emotional bias," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 491–495.

[13] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," p. 2429–2437, 2011.

[14] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" *CVPR 2011*, pp. 145–152, 2011.

[15] Z. Bylinskii, P. Isola, C. M. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision Research*, vol. 116, pp. 165–178, 2015.

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

[17] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2004.

[18] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 2169–2178.

[19] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[20] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[22] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[23] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, USA, December 2012.

[24] W. A. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs." *Journal of Experimental Psychology: General*, vol. 142, no. 4, pp. 1323–1334, Nov. 2013.

[25] B. Celikkale, A. Erdem, and E. Erdem, "Predicting memorability of images using attention-driven spatial pooling and image semantics," *Image and vision Computing*, vol. 42, pp. 35–46, 2015.

[26] S. Rahman and N. D. Bruce, "Factors underlying inter-observer agreement in gaze patterns: Predictive modelling and analysis," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 155–162.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[29] H. Squalli-Houssaini, N. Q. K. Duong, M. Gwenaelle, and C.-H. Demarty, "Deep learning for predicting image memorability," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2371–2375.

[30] S. Perera, A. Tal, and L. Zelnik-Manor, "Is image memorability prediction solved?" in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 800–808.

[31] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018.

[32] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ser. MIR '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 39–43.

[33] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[34] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 867–876.

[35] B. Saleh, A. Farhadi, and A. Elgammal, "Object-centric anomaly detection by attribute-based reasoning," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 787–794.

[36] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[37] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 83–92.

[38] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2106–2113.

[39] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 30–43.

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[41] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem, "What makes an object memorable?" in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1089–1097.