WHAT IMAGES ARE MORE MEMORABLE TO MACHINES?

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the problem of measuring and predicting how memorable an image is to pattern recognition machines, as a path to explore machine intelligence. Firstly, we propose a self-supervised machine memory quantification pipeline, dubbed "MachineMem measurer", to collect machine memorability scores of images. Similar to humans, machines also tend to memorize certain kinds of images, whereas the types of images that machines and humans memorize are different. Through in-depth analysis and comprehensive visualizations, we gradually unveil that"complex" images are usually more memorable to machines. We further conduct extensive experiments across 11 different machines and 9 pre-training methods to analyze and understand machine memory. This work proposes the concept of machine memorability and opens a new research direction at the interface between machine memory and visual data.

1 INTRODUCTION

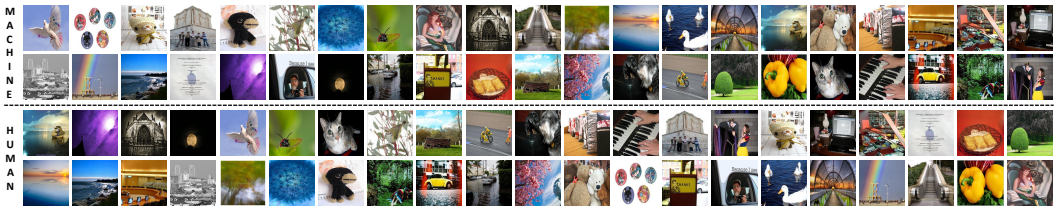


Figure 1: **Sample images** that are sorted from less memorable (left) to more memorable (right) for both machines and humans. The first top two rows are arranged by machine memorability scores and the bottom two rows are sorted by human memorability scores.

The question posed by Alan Turing in 1950 – "Can machines think?"TURING (1950) has underpinned much of the pursuit within the realm of artificial intelligence. Despite the significant strides made in the last few decades, an unambiguous answer to this question remains elusive. The power of pattern recognition machines today is staggering. But due to the nature of back-propagation Rumelhart et al. (1985) and black box models Guidotti et al. (2018), decisions made by machines can sometimes be untrustworthy. Therefore, a deeper comprehension of machine intelligence is a vital step towards crafting machines that are not just powerful, but even more reliable and understandable. In addition, discerning the similarities and differences between artificial and natural intelligence could lay the foundation for the creation of machines capable of perceiving, learning, and thinking in a manner more akin to humans Lake et al. (2017). In this paper, we propose an exploration of machine intelligence from the perspective of memory, a pivotal component in both intelligence and cognition Colom et al. (2010).

When we look at images, we humans can naturally perceive the same types of information, and thus tend to behave similarly in memorizing images Khosla et al. (2015). That is, some images sharing certain patterns are more memorable to all of us Isola et al. (2013); Khosla et al. (2015); Goetschalckx et al. (2019). Recalling our proposal on exploring machine intelligence from a view of memory, we dive into: How well do machines memorize certain types of images? What attributes make an image memorable to a machine? Do different machines exhibit varying memory characteristics?

We begin with a quantification of machine memory. We adopt the human memorability score (HumanMem score) Isola et al. (2011) concept and propose the machine memorability score (henceforth referred to as MachineMem score) as a measure of machine memory. Our next goal is to collect MachineMem scores for a variety of images. To accomplish this, we introduce a novel framework, the MachineMem measurer, inspired by the repeat detection task and visual memory game Isola et al. (2013). This framework produces MachineMem scores for images in a self-supervised manner, allowing us to label the entire LaMem dataset Khosla et al. (2015). Based on this, we train a regression model to predict MachineMem scores in real-time and introduce some advanced training techniques to enhance the performance of both human and machine memorability score predictors.

Armed with collected MachineMem scores, we delve into the investigation of what makes an image memorable to machines. This exploration involves multiple approaches, beginning with a visual teaser (Figure 1), wherein sample images are arranged by their MachineMem scores. A quantitative analysis follows, presenting the correlations between MachineMem scores and 13 image attributes. We then analyze the memorability of different classes for machines. Further, we apply GANalyze Goetschalckx et al. (2019) to visualize how the memorability of a particular image changes with varying MachineMem scores. A comparative analysis between machine memory and human memory helps highlight the differences and similarities. Lastly, we study machine memorability of images for different types of machines.

Lastly, we aim to understand machine memorability more deeply. We conduct two case studies that analyze the MachineMem scores produced by 11 different machines and 9 varying pretext tasks.

2 Related Work

Visual cognition and memory of humans. Pioneering studies Isola et al. (2011; 2013) have systematically explored the elements that make a generic image memorable to humans. They established a visual memory game, a repeat detection task that runs through a long stream of images. This game involves multiple participants, and the averaged accuracy of detecting repeated images provides a quantified HumanMem score for each image. Subsequent research in this area Lahrache & El Ouazzani (2022); Zhang et al. (2020); Bylinskii et al. (2022) has created more datasets Khosla et al. (2015); Bainbridge et al. (2013); Lu et al. (2020); Goetschalckx & Wagemans (2019) and developed more powerful methods for predicting HumanMem scores Kim et al. (2013); Celikkale et al. (2013); Peng et al. (2015); Fajtl et al. (2018); Perera et al. (2019); Lu et al. (2020); Leyva & Sanchez (2021). One of the goals of this work is to compare machine memory and human memory. To this end, we preserve the definition of MachineMem score, mirroring the HumanMem score, and incorporate the key design elements of the visual memory game into our MachineMem measurer.

In psychology and cognition research, memory is broadly divided into sensory Pearson & Brascamp (2008), short-term Cowan (2001), and long-term categories Mandler & Ritchey (1977); Vogt & Magnussen (2007); Brady et al. (2008). The visual memory game primarily captures long-term memory Isola et al. (2013). Yet, given that HumanMem scores remain stable over various time delays Isola et al. (2013); Khosla et al. (2015), they are likely indicative of both short-term and long-term memory Borkin et al. (2015); Cowan (2008). Hence, our MachineMem measurer, which collects MachineMem scores, considers both short-term and long-term memory of machines. These aspects are captured by adjusting the training length in stage (b).

What images are more memorable to humans? Here, we briefly summarize the characteristics typically associated with human-memorable images:

• Images with large, iconic objects, usually in square or circular shapes and centered within the frame, tend to be more memorable. This suggests that a single iconic object makes an image more memorable than multiple objects.

• Images featuring human-related objects (such as persons, faces, body parts) and indoor scenes (like seats, floors, walls) have higher HumanMem scores, while outdoor scenes (such as trees, buildings, mountains, skies) generally contribute negatively.

• Bright, colorful images, especially those with contrasting colors or a predominantly red hue, are more memorable to humans.

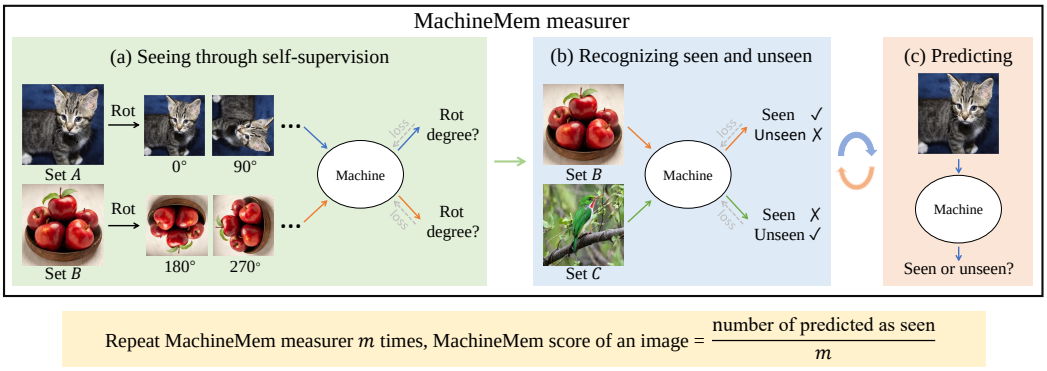
• Simplicity in images often enhances memorability.

Conversely, images that deviate from these trends are generally less memorable. For a more comprehensive understanding, we recommend readers to consult the relevant literature Isola et al. (2013); Khosla et al. (2015); Goetschalckx et al. (2019).

3 MEASURE MACHINE MEMORY

We propose the MachineMem measurer as a pipeline to measure MachineMem scores of images. The design of the MachineMem measurer follows the key idea used in the visual memory game Isola et al. (2011), that is, quantifying machine memory through a repeat detection task. The visual memory game encapsulates three integral components: observe, repeat, and detect. Similarly, we structure the MachineMem measurer as a three-stage process, where each stage corresponds to one of these components. A conceptual diagram of the MachineMem measurer is presented in Figure 2.

Although humans are capable of observing and memorizing images without any feedback mechanism, machines, on the other hand, still lack this ability. Therefore, in stage (a), we adopt a self-supervision pretext task to guide machines to observe images. Following this, stage (b) instructs the machines to distinguish between observed and unobserved images, thereby equipping the machines to execute the repeat detection task in stage (c).



Repeat MachineMem measurer m times, MachineMem score of an image = $\frac{\text{number of predicted as seen}}{m}$

Figure 2: **A realization of the MachineMem measurer**. Our MachineMem measurer has 3 stages: (a) Seeing images through self-supervision, (b) Recognizing seen and unseen images, and (c) Predicting whether an image has been seen .Each image we present (cat, apple, and bird) symbolizes an image set (A, B, and C), each containing n images. We focus on measuring MachineMem scores for set A produced by an identical machine. In every episode of the MachineMem measurer, we randomly select sets B and C from an expansive dataset while keeping the cat set constant. The MachineMem scores of set A are obtained by repeating the MachineMem measurer m times.

Seeing images through self-supervision. In the endeavor to help machines observe and memorize images, supervision is indispensable. However, the supervision signal should be self-sufficient. In light of this, we contemplate employing a pretext task as supervision that satisfies three criteria: (1) it necessitates minimal structural modifications at the machine level, (2) it does not degrade or distort the input data, and (3) it allows machines to observe whole images rather than cropped segments. The rotation prediction self-supervision task Gidaris et al. (2018) fulfills all these prerequisites and is therefore selected as the pretext task for stage (a).

Following common practice Gidaris et al. (2018); Deng et al. (2021), we define a set of rotation transformation functions $G = \{R_r(\boldsymbol{x})\}$, where R_r is a rotation transformation function and r are the rotation degrees $r \in \{0°, 90°, 180°, 270°\}$. Rotation prediction is a multi-class (4 class here) classification task, where the goal is to predict which rotation degree has been applied to an input image $\boldsymbol{x}$. The loss function is formulated as:

$$\mathcal{L}_{\text{rot}} = \frac{1}{4}\left[\sum_{r\in\{0°,90°,180°,270°\}} \mathcal{L}_{\text{CE}}(\boldsymbol{y}_r, \boldsymbol{\theta}_m(R_r(\boldsymbol{x})))\right], \tag{1}$$

where $\boldsymbol{y}_r$ is the one-hot label of $r \in \{0°, 90°, 180°, 270°\}$. $\mathcal{L}_{\text{CE}}$ denotes cross-entropy loss. $\boldsymbol{m}$ is a machine parameterized by $\boldsymbol{\theta}_m$.

Stage (a) uses two sets (sets A and B) of images, where images in both two sets are labeled as seen. Half of them (set B) go to stage (b) and the other half (set A) go to stage (c), as shown in Figure 2. We force machines to achieve good performance (top-1 accuracy ≥ 80 %) on the rotation prediction task. By default, machines without pre-training are trained for 60 epochs in this stage.

Recognizing seen and unseen images. This stage aims to teach machines to recognize seen and unseen images. We use a set of seen images (set B) that has been used in stage (a) and sample a set of unseen images (set C) from a large-scale dataset. A binary classification task targeted at recognizing seen and unseen images is employed. The backbone of the machine remains identical but we replace the 4-way classification layer with a new 2-way linear classification head. The loss function is expressed as:

$$\mathcal{L}_{\text{seen}} = \frac{1}{2}\left[\sum_{r\in\{\text{seen},\text{unseen}\}} \mathcal{L}_{\text{CE}}(\boldsymbol{y}_l, \boldsymbol{\theta}_m(\boldsymbol{x}))\right], \tag{2}$$

where $\boldsymbol{y}_l$ denotes the one-hot label of $r \in \{\text{seen}, \text{unseen}\}$, CE is cross-entropy loss, and $\boldsymbol{m}$ stands for a machine parameterized by $\boldsymbol{\theta}_m$.

In our default setting, stage (b) lasts for 10 epochs, and the machine will enter stage (c) upon finishing each epoch.

Predicting whether an image has been seen. During the final stage, we utilize a set of previously 'seen' images (set A) that were not involved in stage (b) for measurement purposes. Here, the machine's task is to discern whether a given image has been seen before, thereby replicating the repeat detection tasks we use to evaluate human memory capabilities.

Given the inherent uncertainty in learning-based systems Hüllermeier & Waegeman (2021), it is imperative to ensure that the results generated during this stage are both meaningful and reliable. We thus employ the calibration error metric Guo et al. (2017) as an assessment tool to gauge the reliability of the machine's predictions. Lower calibration errors are indicative of models with a higher degree of reliability and result accuracy. Specifically, we employ the RMS calibration error Hendrycks et al. (2018) and adaptive binning Nguyen & O'Connor (2015) techniques to measure this.

The interchange between stage (b) and stage (c) is iterated, allowing us to gather multiple measurements from the images within set A. The final result from a single MachineMem measurer episode is chosen based on the iteration that yields the lowest calibration error. This approach also captures both the short-term and long-term memory capabilities of the machines.

Obtaining MachineMem scores. Drawing from the HumanMem scores approach Isola et al. (2011), we define the MachineMem score of an image as the ratio of the number of seen predictions to the total number of MachineMem episodes.

During each MachineMem measurer episode, besides set A which is destined for stage (c) for score computation, we randomly select two other sets of images (set B and set C) from a dataset, each containing n images. This randomized selection process is designed to ensure that the MachineMem measurer accurately reflects the machine's memory capabilities rather than fitting specific distributions. We default n to 500. To calculate the MachineMem scores for set A, we repeat the MachineMem process m times, where m is set to 100, mirroring the average number of participants involved in HumanMem score collection. The MachineMem scores for set A are thus obtained after repeating the MachineMem measurer m times.

We have collected and labeled MachineMem scores for all images in the LaMem dataset (totaling 58,741 images). The average MachineMem score is 0.680 (SD 0.070, min 0.39, max 0.91), which contrasts with the average HumanMem score of 0.756 (SD 0.123, min 0.20, max 1.0). By default, we employ a ResNet-50 as our basic machine and use the scores produced by this machine to represent the MachineMem scores. We have also endeavored to evaluate the memory characteristics of various other machine models, with their respective training specifics and analyses discussed in the subsequent sections.

4 PREDICT MACHINEMEM SCORES

Collecting MachineMem scores with the MachineMem measurer can be a time-consuming process, often requiring several hours to generate scores for 500 images. In response to this challenge, we aspire to train a robust regression model capable of predicting MachineMem scores in real-time. This task aligns with the prediction of HumanMem scores, prompting us to revisit and enhance approaches tailored towards predicting HumanMem scores.

Past research Fajtl et al. (2018); Perera et al. (2019) has demonstrated that a modified ResNet-50 regression model (with adjustments to the final layer to accommodate regression tasks) can deliver satisfactory performance in predicting HumanMem scores. This model is trained utilizing dropout Srivastava et al. (2014) and RandomResizedCrop Szegedy et al. (2015). With this training setup, complemented by an ImageNet-supervised pre-training initialization, this simple ResNet-50 regression model can attain a Spearman's correlation, ρ, of 0.63 when predicting HumanMem scores. For comparison, the human consistency Khosla et al. (2015) registers at $\rho = 0.68$, while the state-of-the-art result Perera et al. (2019) reaches $\rho = 0.67$.

We show a superior performance can be accomplished in predicting HumanMem scores by employing self-supervised pre-training and strong data augmentations. Specifically, we transfer the knowledge from the pre-trained MoCo v2 Chen et al. (2020b). At the data level, we substitute RandomResizedCrop Szegedy et al. (2015) with CropMix Han et al. (2022b), while integrating Random erasing Zhong et al. (2020) and Horizontal flip applied in a YOCO manner Han et al. (2022a). This results in a ResNet-50 regressor that attains $\rho = 0.69$ in predicting HumanMem scores, surpassing even human consistency! We refer to this model as the enhanced ResNet-50 regression model.

In the subsequent step, we aspire to train a regression model capable of predicting MachineMem scores that align as closely as possible with those derived from empirical observations of 100 trials from the MachineMem measurer. We employ and train this enhanced ResNet-50 regression model for the task of predicting MachineMem scores. We randomly select 1000 images as the test set, using all remaining images as the training set. This model also achieves a $\rho = 0.69$ in predicting MachineMem scores. We designate our model as the MachineMem predictor.

5 WHAT MAKES AN IMAGE MEMORABLE TO MACHINES?

This section aims to analyze MachineMem scores in order to understand what factors contribute to an image's memorability for machines. We present some sample images in Figure 1, first analyzing the relationship between MachineMem scores and 13 image attributes. With the aid of the MachineMem predictor, we predict MachineMem scores of all ImageNet Russakovsky et al. (2015) training images to analyze which classes are most and least memorable to machines. Additionally, we employ the GANalyze Goetschalckx et al. (2019), capable of adjusting an image to generate more or less memorable versions, as a means to discover hidden trends that could potentially influence MachineMem scores. In conjunction with GANalyze, we conduct a comparative study between machine memorability and human memorability.

5.1 QUANTITATIVE ANALYSIS

Do image attributes adequately determine MachineMem scores? Here we examine 13 image attributes, roughly grouped into 4 categories, each focusing on different measurements. Based on MachineMem scores of images, we sort all LaMem images and organize them into 10 groups, from the group with the lowest mean MachineMem scores to the highest. Each group contains approximately 5870 images. Spearman's correlation results are computed based on all 58741 images. Figure 3 presents plots illustrating the relationship between image groups and varying image attributes.

Image quality. We employ two metrics, NIQE Mittal et al. (2012b) and BRISQUE Mittal et al. (2012a), to assess image quality. Lower NIQE and BRISQUE values suggest better perceptual quality, indicating fewer distortions.

Both BRISQUE and NIQE demonstrate a very weak correlation with MachineMem scores ($\rho = -0.06$ and -0.13, respectively). NIQE shows a relatively stronger correlation, possibly because BRISQUE, which involves human subjective measurements, aligns more with human perception.

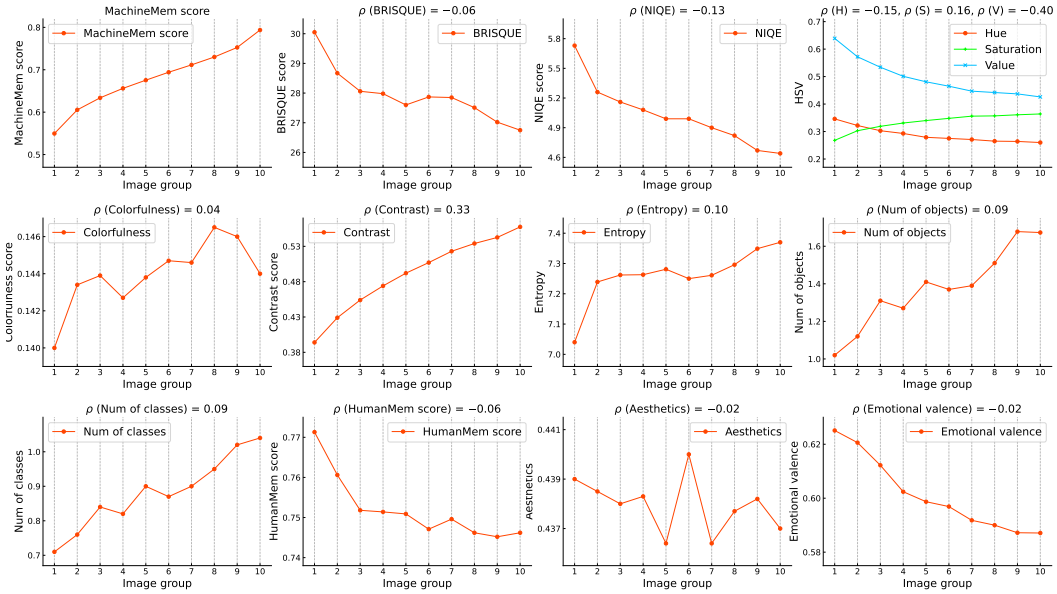


Figure 3: **Relation between image groups and varying image attributes**. We find value and contrast to be the two most significant attributes that correlate moderately ($\rho \geq 0.3$) with MachineMem scores. On the other hand, attributes such as hue and saturation show a weak correlation ($0.15 \leq \rho < 0.3$) with MachineMem scores. Other factors, such as NIQE, entropy, and number of objects, demonstrate a very weak correlation ($0.08 \leq \rho < 0.15$) with MachineMem scores. These findings are based on Spearman's correlation (ρ) computed from the entire data set.

Pixel Statistics. We investigate the correlation between MachineMem scores and basic pixel statistics. Hue, saturation, and value from the HSV color space Agoston (2005) are measured, along with colorfulness Hasler & Suesstrunk (2003), contrast Matkovic et al. (2005), and entropy.

Interestingly, value and contrast show substantial correlations with MachineMem scores ($\rho = -0.40$ and -0.33, respectively). Deep color and strong contrast are two significant factors that make an image memorable to machines. Hue and saturation are weakly correlated with MachineMem scores ($\rho = -0.15$ and 0.16, respectively). Entropy exhibits a very weak correlation with MachineMem scores ($\rho = 0.10$). However, as presented in Figure 3, the group with the lowest MachineMem scores, *i.e.*, group 1, displays very low entropy. Images with very low MachineMem scores often lack contrast or have a light color background (see Figure 1), and therefore tend to have low entropy. Furthermore, colorfulness seems to have no clear correlation with MachineMem scores ($\rho = 0.04$), except for the fact that group 1 scores very low in terms of colorfulness.

Object Statistics. We measure the number of objects and the number of classes (unique objects) within an image. A YOLOv4 Bochkovskiy et al. (2020) model is employed as the object detector.

Both metrics are very weakly correlated with MachineMem scores (same $\rho = 0.09$). By excluding data with 0 objects, their correlations with MachineMem scores are still very weak (same $\rho = 0.10$).

Cognitive Image Property. We employ HumanMem scores, aesthetics, and emotional valence as cognitive image properties. The HumanMem scores are obtained from the LaMem dataset. To measure aesthetics, we utilize a pre-trained LIMA model Talebi & Milanfar (2018). We use the emotional valence predictor from the GANalyze for determining emotional valence.

The correlation between HumanMem scores and MachineMem scores is very weak ($\rho = -0.06$). Similarly, other cognitive image properties, such as aesthetics and emotional valence, exhibit negligible correlation with MachineMem scores (both with $\rho = -0.02$). These findings suggest that MachineMem scores represent a unique image property that is distinct from other image properties.

In conclusion, unlike human memory, which is largely driven by semantics, machines, devoid of such common sense, tend to emphasize more on basic pixel statistics.

5.2 WHAT CLASSES ARE MORE OR LESS MEMORABLE?

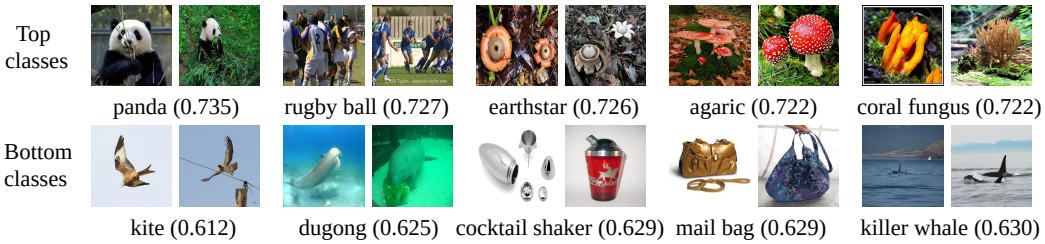


Figure 4: **ImageNet classes sorted by their mean MachineMem scores**. We report the top-5 and bot-5 classes and their mean MachineMem scores. It appears that the classes ranking highest commonly exhibit lower values paired with pronounced contrast. To illustrate, images of pandas consistently feature a mix of both white and black hues, often juxtaposed against a green background, thus enhancing the overall contrast. Conversely, classes ranked at the bottom predominantly showcase lighter backgrounds occupying a substantial proportion of the pixel space.

Do images belonging to certain classes tend to be more or less memorable to machines? We use the MachineMem predictor to predict MachineMem scores of all ImageNet training images to obtain mean MachineMem scores of 1000 ImageNet classes. Figure 4 summarizes the top and bottom classes. By analyzing gained results, we find the answer to be yes: Classes containing light backgrounds are usually less memorable to machines, for instance, classes related to sea or sky. Classes that have strong contrast, high value, and multiple objects tend to have high MachineMem scores.

5.3 GANALYZE

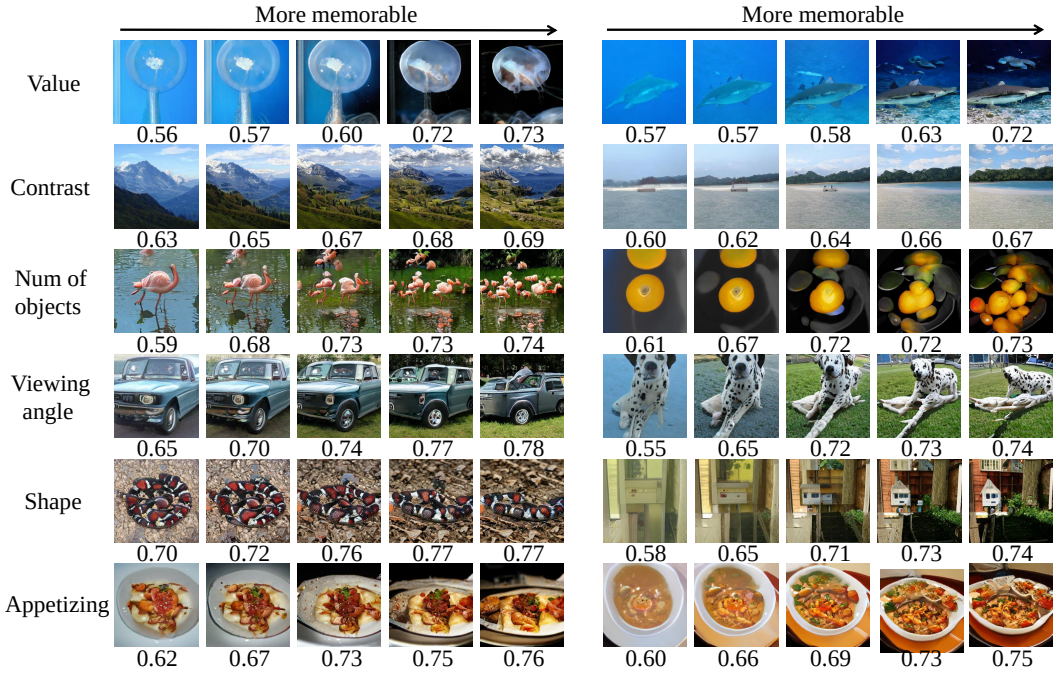


Figure 5: **Images generated by GANalyze**. We visualize what will happen if we make an image more or less memorable to machines. We summarize 6 trends, where the first 3 of them (value, contrast, and number of objects) are previously found in the quantitative analysis part. We show the results of GANalyze as further confirmations. For certain objects, viewing angle, shape, and appetizing are hidden trends that are unveiled by GANalyze. An overall trend is that GANalyze is often complexifying images to make them more memorable to machines.

While the relationship between MachineMem scores and various image attributes has been established, certain concealed factors that potentially enhance an image's memorability for machines remain elusive. Therefore, we leverage the potent capabilities of GANalyze to uncover these hidden elements that could influence MachineMem scores. More specifically, we employ the MachineMem predictor as the Assessor within GANalyze to guide the model in manipulating the latent space to change an image's machine memorability. The results of this investigation are depicted in Figure 5.

In the process, we also utilize GANalyze to provide additional validation for the correlations between MachineMem scores and image attributes. We elucidate three observable trends: Value, Contrast, and Number of Objects. In terms of concealed trends, we identify three standout candidates. The newly unearthed trends are 1: Viewing angles that yield more information, such as side views, are typically more memorable than angles that provide less information, 2: Objects that adhere to standard shapes, such as squares or circles, appear less memorable to machines. In contrast, objects with irregular shapes tend to be more memorable, and 3: Regarding food-related objects, images with high MachineMem scores frequently appear more appetizing.

A critical overarching trend observed across almost all images is the machine tend to memorize complex images. Factors such as contrast, number of objects, viewing angle, shape, and appeal to taste can all be considered manifestations of this pervasive trend.

5.4 Human memory vs. machine memory

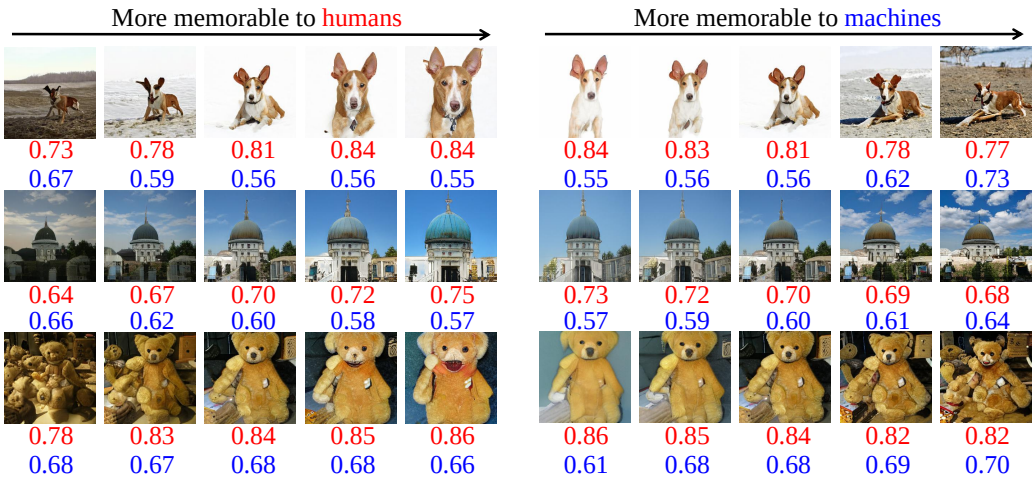


Figure 6: **A comparison between human memory and machine memory**. We employ GANalyze to make an image more as well as less memorable to both humans and machines. HumanMem scores are labeled in ered while blue indicates MahinceMem scores. Generally speaking, simple images are more memorable to humans while complex images are more memorable to machines.

As presented in Figure 3, MachineMem scores and HumanMem scores are very weakly correlated ($\rho = -0.06$). But in GANalyze, which is good at showing global trends, we find machines tend to memorize more complex images, which is on the reverse side of humans that are usually better at memorializing simple images. Such results are presented in Figure 6. Other than ResNet-50, we also explore the correlations between multiple machines (10 other different machines that will be presented in the next section) and humans, however, none of these machines show clear correlations ($|\rho| \geq 0.15$) with humans. This further suggests MachineMem is very distinct from HumanMem.

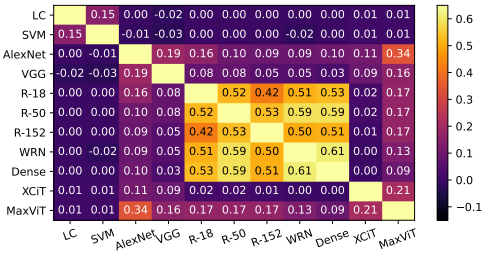
6 Understanding Machine Memory

HumanMem, as an inherent and consistent attribute of images, is universally recognized by individuals, transcending their diverse backgrounds Isola et al. (2013). This implies that, despite varying human experiences, there exists a shared element in how humans remember visual data. But do machines exhibit a similar principle? Here we delve into two key questions: Will MachineMem scores remain consistent across different machines? What role does varying pre-training knowledge play?

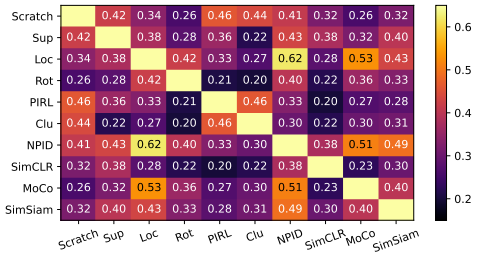
6.1 MEMORY ACROSS MACHINES

We scrutinize the relationships among 11 distinct machines, grouping them into four categories: conventional machines (linear classifier and SVM Cowan (2001)), classic CNNs (AlexNet Krizhevsky et al. (2012), VGG Simonyan & Zisserman (2014)), modern CNNs (ResNet-18, ResNet-50 He et al. (2016), ResNet-152, WRN-50-2 Zagoruyko & Komodakis (2016), and DenseNet121 Huang et al. (2019)), and Vision Transformers (ViTs) (XCiT-T Ali et al. (2021) and MaxViT-T Tu et al. (2022)). We examine and evaluate the MachineMem scores of 10000 LaMem images as produced by these varying machines. Due to the inherent constraints of conventional machines, we employ a binary classification task (0 °and 90 °comprising one class, and 180 °and 270 °forming the other) as their pretext tasks in the initial stage (a) of the MachineMem measurement process. The training parameters are identical for each machine within the same category, although slight variations exist across different categories.

As represented in Figure 7a, machines within the same category generally exhibit strong correlations (average ρ of modern CNNs = 0.53), indicating a tendency to memorize similar images. However, less apparent correlations are observed among machines from distinct categories. For instance, conventional machines, due to their limited modeling capabilities, do not correlate with machines from other categories.



(a) **Memory across machines**. Each off-diagonal corresponds to Spearman's correlation (ρ) of two machines. Machines within each category are usually strongly correlated, but this trend does not scale to machines across categories.



(b) **Memory across pre-training methods**. Spearman's correlation (ρ) of two pre-training methods is presented at each off-diagonal. Though having different prior knowledge, an identical structured machine tends to memorize similar images.

6.2 MEMORY ACROSS PRE-TRAINING METHODS

We explore nine pre-training methods applicable to a ResNet-50 model. This includes supervised ImageNet classification pre-training and eight unsupervised methods, such as relative location Doersch et al. (2015), rotation prediction Gidaris et al. (2018), PIRL Misra & Maaten (2020), DeepCluster-v2 Caron et al. (2021), and four instance discrimination approaches (NPID Wu et al. (2018), SimCLR Chen et al. (2020a), MoCo v2 Chen et al. (2020b), and SimSiam Chen & He (2021)). The analysis is conducted on 10,000 LaMem images.

Figure 7b summarizes our findings. The highest correlation in MachineMem scores (ρ = 0.62) is observed between location prediction and NPID, while the weakest correlation (ρ = 0.20) emerges between PIRL and SimCLR. In general, the memory capabilities of a ResNet-50 model are not significantly influenced by its pre-training knowledge (average ρ = 0.35). This observation suggests that MachineMem, like HumanMem, can be considered an intrinsic and stable attribute of an image, shared across different models despite variations in their pre-training knowledge.

7 CONCLUSION

We propose and study a property of images, *i.e.*, machine memorability. Machine memorability shows a cognitive property of machines and can serve as a pathway to help us to further explore machine intelligence. We hope our findings could provide insights into fundamental advances in computer vision, machine learning, natural language processing, and general artificial intelligence.

REFERENCES

Max K.. Agoston. *Computer Graphics and Geometric Modeling: Implementation and Algorithms*. Springer, 2005.

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021.

Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323, 2013.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, 22(1):519–528, 2015.

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.

Zoya Bylinskii, Lore Goetschalckx, Anelise Newman, and Aude Oliva. Memorability: An image-computable measure of information utility. In *Human Perception of Visual Information*, pp. 207–239. Springer, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2021.

Bora Celikkale, Aykut Erdem, and Erkut Erdem. Visual attention-driven spatial pooling for image memorability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 976–983, 2013.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Roberto Colom, Sherif Karama, Rex E Jung, and Richard J Haier. Human intelligence and brain networks. *Dialogues in clinical neuroscience*, 2010.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001.

Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008.

Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning*, pp. 2579–2589. PMLR, 2021.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6363–6372, 2018.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations*, 2018.

Lore Goetschalckx and Johan Wagemans. Memcat: a new category-based image set quantified on memorability. *PeerJ*, 7:e8169, 2019.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Junlin Han, Pengfei Fang, Weihao Li, Jie Hong, Mohammad Ali Armin, , Ian Reid, Lars Petersson, and Hongdong Li. You only cut once: Boosting data augmentation with a single cut. In *International Conference on Machine Learning (ICML)*, 2022a.

Junlin Han, Lars Petersson, Hongdong Li, and Ian Reid. Cropmix: Sampling a rich input distribution via multi-scale cropping. In *arXiv preprint arXiv:2205.15955*, 2022b.

David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pp. 87–95. SPIE, 2003.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. Convolutional networks with dense connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pp. 145–152. IEEE, 2011.

Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1469–1482, 2013.

Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pp. 2390–2398, 2015.

Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. Relative spatial features for image memorability. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 761–764, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Souad Lahrache and Rajae El Ouazzani. A survey on image memorability prediction: From traditional to deep learning models. In *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp. 1–10. IEEE, 2022.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Roberto Leyva and Victor Sanchez. Video memorability prediction via late fusion of deep multi-modal features. In *2021 IEEE international conference on image processing (ICIP)*, pp. 2488–2492. IEEE, 2021.

Jiaxin Lu, Mai Xu, Ren Yang, and Zulin Wang. Understanding and predicting the memorability of outdoor natural scenes. *IEEE Transactions on Image Processing*, 29:4927–4941, 2020.

Jean M Mandler and Gary H Ritchey. Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 3(4):386, 1977.

Kresimir Matkovic, László Neumann, Attila Neumann, Thomas Psik, Werner Purgathofer, et al. Global contrast factor-a new approach to image contrast. In *CAe*, pp. 159–167, 2005.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

Joel Pearson and Jan Brascamp. Sensory memory for ambiguous vision. *Trends in cognitive sciences*, 12(9):334–341, 2008.

Houwen Peng, Kai Li, Bing Li, Haibin Ling, Weihua Xiong, and Weiming Hu. Predicting image memorability by multi-view adaptive regression. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1147–1150, 2015.

Shay Perera, Ayellet Tal, and Lihi Zelnik-Manor. Is image memorability prediction solved? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *ECCV*, 2022.

A. M. TURING. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 10 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433.

Stine Vogt and Svein Magnussen. Long-term memory for 400 pictures on a common theme. *Experimental psychology*, 54(4):298, 2007.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Hao-Yang Zhang, Jie Liu, and Dou-Dou Wang. Review of the application of deep learning in image memorability prediction. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, pp. 142–146. IEEE, 2020.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13001–13008, 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A Appendix

B Training Details of MachineMem Measurer

In our research, we investigate the memory characteristics of 11 distinct machines. These machines are categorized into four groups, namely conventional machines (comprising linear classifier and SVM Cowan (2001)), classic CNNs (such as AlexNet Krizhevsky et al. (2012) and VGG Simonyan & Zisserman (2014)), modern CNNs (including ResNet-18, ResNet-50 He et al. (2016), ResNet-152, WRN-50-2 Zagoruyko & Komodakis (2016), and DenseNet121 Huang et al. (2019)), and ViTs (like XCiT-T Ali et al. (2021) and MaxViT-T Tu et al. (2022)).

Except for the number of training epochs in stage (a) and the corresponding learning rate, all training hyperparameters remain consistent across all machine types and pre-training methods. We have made these adjustments to ensure machines are able to achieve satisfactory performance levels (top-1 accuracy $\geq 80\%$) during stage (a).

Our MachineMem measurer is trained solely on a single GPU, with a batch size of 1 to parallel the visual repeat game settings. We employ SGD as our optimization algorithm and use a cosine learning schedule for our training process. The settings for weight decay and momentum are 0.0001 and 0.9, respectively. The specifics for the training epochs in stage (a) and learning rates for all machine models are as follows:

Conventional machines: Training epochs for stage (a) are set at 60, with a learning rate of 0.01.

Classic CNNs: Training epochs for stage (a) are set at 70, with a learning rate of 0.0005.

Modern CNNs: Training epochs for stage (a) are set at 60, with a learning rate of 0.01.

ViTs: Training epochs for stage (a) are set at 70, with a learning rate of 0.0005.

ResNet-50 with pre-training: Training epochs for stage (a) are set at 30, with a learning rate of 0.01.

C Training Details of MachineMem Predictor

MachineMem predictor and HumanMem predictor share identical training settings. We use a MoCo v2 model (800 epochs pre-training) as initialization. The prediction model is trained for 30 epochs with an SGD optimizer. Weight decay and momentum are set as 0.0001 and 0.9, respectively. The batch size is 100. The initial learning rate is 0.01 with a cosine decay schedule. For CropMix, the crop scale is set as (0.8, 1.0). We use CutMix as the mixing operation and color permutation as the intermediate augmentation.

D Details and Analysis of Calibration

In stage (c), we adopt the calibration error metric to enhance the reliability and validity of our results by measuring images in set A. Regarding particular methods, RMS calibration error Hendrycks et al. (2018) and adaptive binning Nguyen & O'Connor (2015) are employed.

To improve the robustness of our method, we have further enhanced our original approach to incorporate a held-out set of images for the assessment of calibration quality in both seen and unseen image categories. We have validated this enhanced strategy across several neural network models and found that the MachineMem scores produced align closely with our original results, which were based solely on seen images. This was supported by a strong Spearman's correlation ($\rho > 0.6$) across all tested machines. These findings substantiate that assessing calibration quality using seen images alone is adequate for reliable measurements.

E What scenes are more or less memorable?

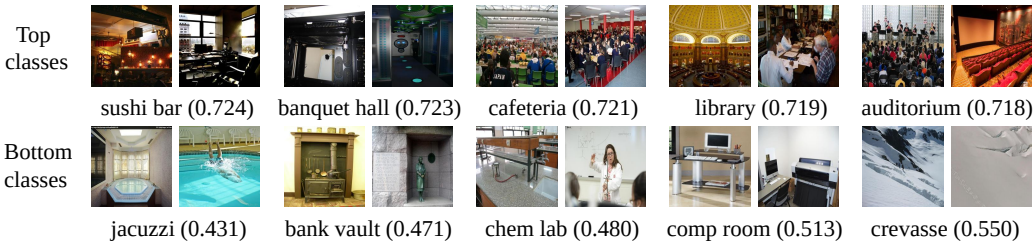


Figure 8: **Places scenes sorted by their mean MachineMem scores**. The top-5 and bot-5 scenes and their mean MahcineMem scores are reported. As observed in the ImageNet classes, the scenes with higher MahcineMem scores generally exhibit lower value, higher contrast, and contain multiple objects. Alternatively, the scene with lower MahcineMem scores often feature white walls and while outdoor scenes.

Do the trends observed from ImageNet classes apply to scenes as well? To answer this, we present the top and bottom Places scenes in Figure 8. We utilized the MachineMem predictor to estimate MachineMem scores for all the training images in Places365 Zhou et al. (2017), thus enabling us to compute average MachineMem scores for 365 scenes. The results indicate that the patterns identified in classes/objects are also evident in scenes, suggesting that this trend is broadly applicable to visual data.

F Can more or less memorable classes be semantically grouped according to a hierarchical structure?

We utilized ImageNet's supercategories to delve into this question. Table 1 outlines the top-five and bottom-five ImageNet supercategories, together with their average MachineMem scores. These findings align with our class-level observations, confirming that memorable classes can indeed be semantically grouped according to a hierarchical structure.

Top-5	basidiomycete 0.722	procyonid 0.720	player 0.716	marketplace 0.716	fungus 0.714
Bot-5	rescue equipment 0.606	computer 0.607	reservoir 0.608	sailing vessel 0.612	hawk 0.612

Table 1: ImageNet supercategories sorted by their mean MachineMem scores. We report top-5 and bot-5 supercategory and their mean MahcineMem scores.

G IS MACHINEMEM CONSISTENT ACROSS TRAINING SETTINGS?

Human memory remains consistent over time Isola et al. (2013). In a visual memory game, an image that is notably memorable after a few intervening images retains its memorability even after thousands of intervening images. We evaluate a ResNet-50 model across different training settings to discern if MachineMem shares this consistency over varying training configurations.

Number of samples. By default, we select the image set size n as 500. We further test this setting with sizes of 50, 1000, and 5000. The Spearman's correlation between the default setting $n = 500$ and these variations are $\rho = 0.05$, $\rho = 0.66$, and $\rho = 0.43$ respectively. A very small n is insufficient to train a stable model and thus fails to accurately reflect memory characteristics. As n increases, the correlations between different n values become increasingly strong.

Number of epochs in stage (a). The default number of epochs in stage (a) is set to 60. We test four additional settings: 15, 30, 45, and 75. The Spearman's correlation between the default setting and these variants are $\rho = 0.21$, $\rho = 0.42$, $\rho = 0.55$, and $\rho = 0.57$ respectively. The correlation becomes stronger as the number of epochs in stage (a) increases. Once the model undergoes sufficient training (30 epochs), its memory characteristics stabilizes over multiple epochs in stage (a).

Number of epochs in stage (b). We default the number of epochs in stage (b) to 10. Four other settings (1, 4, 7, and 13) are tested. The Spearman's correlation between the default setting and these variants are $\rho = 0.31$, $\rho = 0.54$, $\rho = 0.59$, and $\rho = 0.57$. As with human memory, machine memory also appears consistent over time/delay. This finding suggests that for both humans and machines, the correlation between short-term and long-term memory remains strong in such rank memorability measurements.

H VALIDATING TRENDS THAT CHANGE MACHINEMEM SCORES

In section 5.3 of the main paper, 3 newly discovered hidden trends (Viewing angle, shape, and tasty) are unveiled. However, in the GANalyze framework, semantics are not disentangled, *i.e.*, when transferring an image to more or less memorable versions, many semantics are changing together. In this section, we validate these newly discovered trends.

Viewing angle. We study two scenes (lego and mic) from the nerf-synthetic Mildenhall et al. (2021) dataset. For each scene, we show the top-10 and bottom-10 images/views in Figure **??**, where the top views usually contain more information (more parts of objects presented) than the bottom views. Results here further validate our finding on viewing angles, *i.e.*, viewing angles that provide more information are usually more memorable to machines.

Shape. For both regular and irregular shapes, we draw 25 images as test sets. We transform every image using 4 rotation degrees $\{0°, 90°, 180°, 270°\}$ to extend the number of images to 100. The mean MachineMem score of the regular shapes image set and the irregular image set is identical, 0.61. Though this trend is shown in GANalyze, results here suggest that shape along might not be able to determine MachienMem scores.

Tasty. We collect 100 images from the internet with the keyword "tasty food" as an image set of tasty food. Another image set, not tasty food, which also contains 100 images, is collected using two keywords "disgusting food" and "overcooked food". The mean MachineMem score of the tasty image set is 0.70 while the non-tasty image set is 0.66. This further suggests that food-related images with a higher MahcineMem score often look tastier.

I WHAT IMAGES ARE MORE MEMORABLE TO OTHER MACHINES?

We further incorporate three additional models including AlexNet Krizhevsky et al. (2012), DenseNet121 Huang et al. (2019), and MaxViT-T Tu et al. (2022) to visualize the types of im-

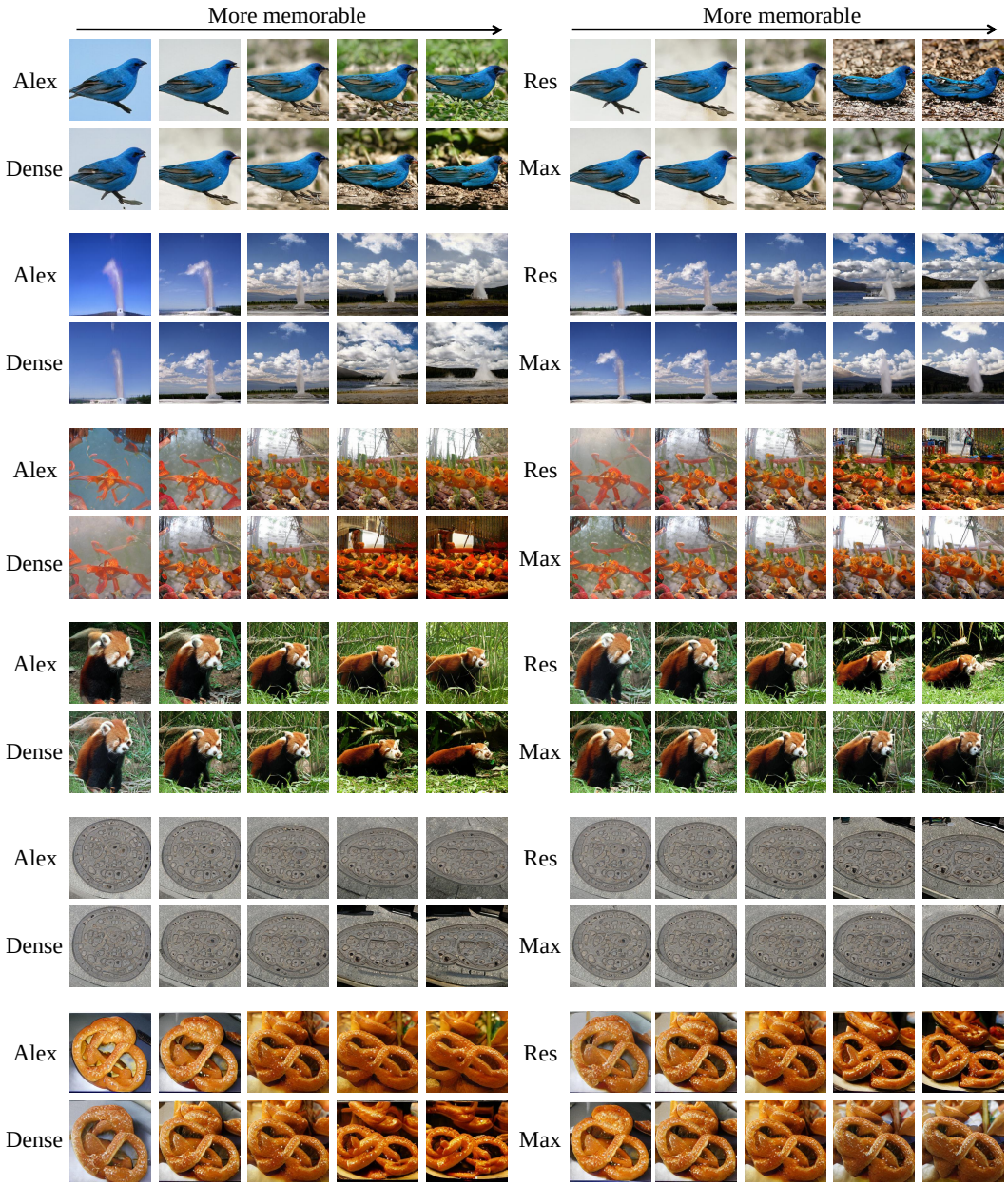


Figure 9: **A comparison between multiple machines using GANalyze**. Alex, Res, Dense, and Max are abbreviations representing AlexNet, ResNet-50, DenseNet, and MaxViT respectively. Each pair of rows represents a distinct trend, in the following order from top to bottom: value, contrast, number of objects, viewing angle, shape, and appetizing. We continue to discern patterns that recur across diverse machine models. For instance, images securing higher MachineMem scores typically exhibit lower value and strong contrast. Most trends identified within the ResNet-50 translate to other machines.

ages they find more memorable. As depicted in Figure 9, we employ GANalyze to facilitate a comparative visual representation across multiple machine models. Factors such as value, contrast, viewing angle, appetizing, and complexity continue to serve as reliable predictors of MachineMem scores across various machine models. However, metrics like the number of objects and shape do not consistently apply to all machines. For instance, MaxViT does not display a preference for images containing a larger number of objects. Despite this, the overarching trend remains, that is, machines generally find more complex images memorable. In the following section, we will present a more thorough quantitative analysis concerning machine memorability across various machine models.

J IMPLICATIONS FOR COMPUTER VISION COMMUNITY AND POTENTIAL APPLICATIONS.

Identifying visually memorable data can lead to practical applications in areas such as data augmentation, continual learning, and generalization. For example, we might design a new data augmentation strategy that can make data more memorable to machines to assist the training of neural networks. In the context of continual learning, it may be advantageous to give greater attention to data that is less memorable.

The way artificial intelligence operates is still vastly different from natural intelligence and creating machines that mimic human behavior remains challenging. A deeper comprehension of how pattern recognition machines work can facilitate the development of more intelligent machines.