Diffusing Surrogate Dreams of Video Scenes to Predict Video Memorability

Lorin Sweeney[1,*], Graham Healy[1] and Alan F. Smeaton[1]

[1]*Insight Centre for Data Analytics, Dublin City University, Ireland*

Abstract

As part of the MediaEval 2022 Predicting Video Memorability task we explore the relationship between visual memorability, the visual representation that characterises it, and the underlying concept portrayed by that visual representation. We achieve state-of-the-art memorability prediction performance with a model trained and tested exclusively on *surrogate dream* images, elevating concepts to the status of a cornerstone memorability feature, and finding strong evidence to suggest that the intrinsic memorability of visual content can be distilled to its underlying concept or meaning irrespective of its specific visual representational.

1. Introduction and Related Work

The natural world is a tempest of sensory threads—from frenzied photons to odious odourants. As we wade through this storm of complex multi-sensory data, our brain is court master and king—tying threads into an intelligible internal representation, and exiling all that it deems unnecessary. What should be remembered, and what should not? The answer is hidden in the whims of the king. Memorability—the likelihood that a given piece of content will be recognised upon subsequent viewing—can be viewed as the *Rosetta Stone* required to decipher the remembering whims of the brain, which is what ultimately motivates and brings meaning to its exploration. Additionally, its proximity to the essence of human experience, and "what the brain deems to be important", casts it into the territory of proxy measure of human importance and quintessential media metric.

Although much progress has been made thinning the query-saturated haze that conceals the landscape of answers mapped by the seminal question: "What makes an image memorable?" [1, 2, 3, 4], the summit remains out of sight, with 25% of the variance still remaining unaccounted for [5]. The shortest path to understanding is through a hurricane of light. Given that we are visually dominant creatures, with over half of the cortex involved in visual processing [6], we naturally expect visual sensory data to exert the greatest influence on memorability. However, it is important not to be lead awry by our brain's appetite for visual sensory data, as semantic meaning is known to play a critical role in visual memorability. Richer and more conceptually distinctive events last longer in memory, and certain semantic categories are inherently more memorable than others [5, 7]. Even though visual memories are stored with an exceptional fidelity of detail (i.e., configurations and contexts of viewed objects [8]), our performance is poor when it comes to remembering random patterns unless they take on object-like qualities [9], suggesting that visual memory is not driven entirely by visual details. Further evidence suggests

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

*Corresponding author.

lorin.sweeney8@mail.dcu.ie (L. Sweeney); Graham.Healy@dcu.ie (G. Healy); alan.smeaton@dcu.ie (A. F. Smeaton)

0000-0002-3427-1250 (L. Sweeney); 0000-0001-6429-6339 (G. Healy); 0000-0003-1028-8389 (A. F. Smeaton)

that visual data is merely a means to conceptual understanding, which is in turn intimately tied to memory, with conceptual distinctiveness supporting higher fidelity visual long-term memory representations than perceptual distinctiveness, and influencing memory retention in a manner that cannot be accounted for by perceptual distinctiveness alone [10, 7]. Perceptual distinctiveness is typically measured within a given object category, and with reference to variations in low dimensional, knowledge agnostic, perceptual features (i.e., colour, and shape). Unfortunately, the line between perceptual and conceptual features begins to blur as we move into higher dimensional features (e.g., length of torso relative to head size), which become more category specific and likely to be acquired through visual experience [11], making it difficult to probe the depth of connection between concept and memorability. However, with the recent explosion in progress in the image synthesis field, and the release of open-source text-to-image diffusion model *Stable Diffusion* [12], we find ourselves uniquely positioned to assess the impact of conceptual features on video memorability independent of its perceptual features, with the exceptional ability to preserve the depth and richness of information inherent to the visual domain.

We hypothesise that if visual data truly is merely a means to conceptual understanding, and that it is the concept itself—which is conveyed/represented through the visual data—that holds the content's intrinsic memorability, then the inter-video relationship of memorability scores predicted with ground-truth video frames should be observable in the memorability scores predicted with synthetic images predicated on purely conceptual video data.

This paper leverages state of the art image synthesis to facilitate the exploration of our aforementioned hypothesis, which can be concisely captured as the following question: can the intrinsic memorability of visual content be distilled to its underlying concept or meaning?

2. Approach

Figure 1: Images used to fine-tune the Stable Diffusion model and create the mem10kstyle token.

Our experiments were carried out within the purview of subtask 1 of the MediaEval Predicting Video Memorability task [13], with the Memento10k dataset—comprised of 7,000 training videos, 1,500 validation videos, and 1,500 withheld test videos—acting as our data landscape. However, before we could set out on our quest for insight, we had to terraform the landscape by synthesising images that reflect the *conceptual essence* of the original Memento10k videos. In order to do so, we leveraged *Stable Diffusion*, a latent text-to-image diffusion model [12]. Stable Diffusion is pre-trained on the LAION-5B dataset [14], which consists of scraped non-curated image-text-pairs from the internet, and is capable of generating high-resolution images from text input.

While the images synthesised using Stable Diffusion are generally high quality in terms of image resolution, if left unspecified in the input text prompt, the compositional construction of the synthesised images is often quite unpredictable and hyper-stylised/unrealistic (i.e., cartoonish, painted, rendered). With the aim of combatting this and guiding the style of synthesised images,

we created a style token (mem10kstyle) that could be appended to prompts by fine-tuning the *stable-diffusion-v1-5* checkpoint on 20 real world photographs (depicted in Figure 1.) which reflect the "in the wild" nature of Memento10k videos, and used 1,500 Memento10k video frames as regularisation images, training for a total of 2,200 steps.

Stable Diffusion requires input prompts in order to generate images, so using each video's first caption as a foundation, we build a textual prompt by pre-appending video action labels, appending one of three custom prompt modifiers, and finishing with our *mem10kstyle* token. Our custom prompt modifiers are tailored to the content depicted in the video to further guide the image generation process. We then create a dataset we call *"Memento10k Surrogate Dream"*—acknowledging that the synthesised images are in fact dream-like surrogates for the videos—by passing each prompt to our fine-tuned Stable Diffusion model (Figure 2.)

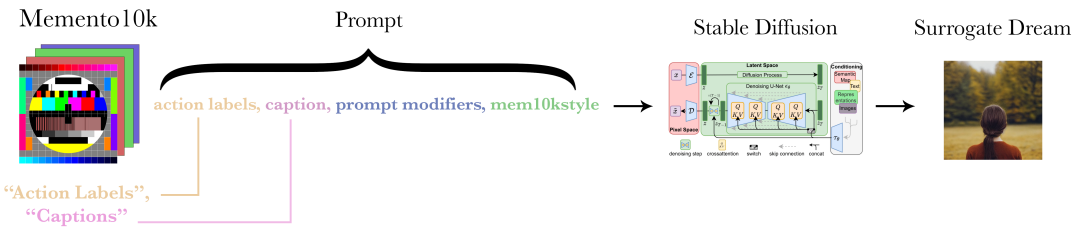

Figure 2: Surrogate Dream Pipeline to synthesize images.

We submitted 5 runs for evaluation in the Predicting Video Memorability task. Each run falls into one of two categories: *Genesis* or *Surrogate Dream*.

Genesis: Approaches trained on vanilla Memento10k data are considered to be *Genesis*, and were trained on visual features extracted from unaltered Memento10k video frames. The runs entitled **Mem10k_DenseNet121** and **Mem10k_DenseNet121_Dream** are ImageNet-pretrained DenseNet121 models fine-tuned (for 50 epochs, with a maximum learning rate of 1e-3, and weight decay of 1e-2) on the middle frame of the Memento10k training videos. The run **Mem10k_CLIP_Ridge_Regression_Mem10k** is a Bayesian Ridge Regressor (BRR) fit with default sklearn [15] parameters on stacked CLIP visual embeddings (extracted from the first, middle, and last video frames) [16].

Surrogate Dream: Approaches trained on images generated with our fine-tuned Stable Diffusion model are considered to be *Surrogate Dream*, and with the exception of memorability scores, were trained exclusively on surrogate visual data. The runs entitled **Dream_DenseNet121_Mem10k** and **Dream_DenseNet121_Dream** are ImageNet-pretrained DenseNet121 models fine-tuned (for 50 epochs, with a maximum learning rate of 1e-3, and weight decay of 1e-2) on our Memento10k Surrogate Dream dataset.

3. Discussion and Outlook

Table 1 shows the Spearman scores for our runs from subtask 1, with **Genesis/Surrogate Dream** indicating whether the approach was trained on ground-truth video frames, or synthesized images respectively, and the final token Mem10k/Dream of each approach indicating whether it was tested on ground-truth video frames, or synthesised images respectively. In the broader context of memorability prediction, all of our runs sit firmly in state-of-the-art territory, with two of our runs marginally outperforming the hitherto state-of-the-art memorability prediction model SemanticMemNet [2]. Although our run entitled **Mem10k_CLIP_Ridge_Regression_Mem10k** achieved the highest Spearman score, the most

Table 1
Official results on the test-set for each of our approaches.

Approach	Run Name	Spearman
Genesis	Mem10k_DenseNet121_Dream	0.583
	Mem10k_DenseNet121_Mem10k	0.645
	Mem10k_CLIP_Ridge_Regression_Mem10k	**0.667**
Surrogate Dream	Dream_DenseNet121_Mem10k	0.625
	Dream_DenseNet121_Dream	**0.664**

notable aspect of our results centres around our run entitled **Dream_DenseNet121_Dream**, which was both exclusively trained and tested on surrogate dream images, and not only outperforms our control run entitled **Mem10k_Dense121_Mem10k**, but achieves an impressive better than state-of-the-art score of 0.664.

The distributions of memorability score predictions for vanilla and surrogate dream approaches are shown in Figure 3. When combined with the evaluation scores, this provides the first of its kind strong evidence that visual data is merely a means to conceptual understanding, and that it is the concepts themselves—which are conveyed/represented through the visual data—that hold the content's intrinsic memorability.

Graph B in Figure 3 tentatively suggests that surrogate dream images are more memorable than ground-truth video frames by virtue of the left skew in predicted scores from our run trained on Mem10k frames and tested on surrogate dream images. However, detailed exploration and investigation into the nature and composition of images in our Memento10k Surrogate Dream dataset is warranted and should be a focus of future research.

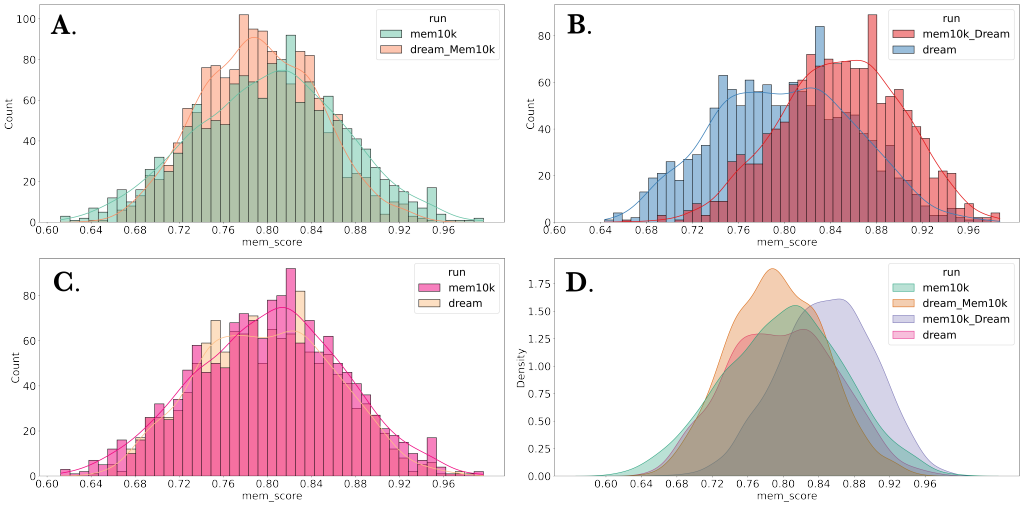

Figure 3: Distribution of run predictions on official test set. Legend format: (trained on)_(tested on).

Acknowledgements

Science Foundation Ireland under Grant Number SFI/12/RC/2289_P2, cofunded by the European Regional Development Fund.

References

[1] P. Isola, J. Xiao, A. Torralba, A. Oliva, What makes an image memorable, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 145–152.

[2] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 223–240.

[3] L. Sweeney, G. Healy, A. F. Smeaton, Leveraging audio gestalt to predict media memorability, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2020. URL: http://ceur-ws.org/Vol-2882/.

[4] L. Sweeney, G. Healy, A. F. Smeaton, The influence of audio on video memorability with an audio gestalt regulated video memorability system, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2021. URL: http://ceur-ws.org/Vol-3181/.

[5] P. Isola, J. Xiao, D. Parikh, A. Torralba, A. Oliva, What makes a photograph memorable, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2013) 1469–1482.

[6] R. Snowden, R. J. Snowden, P. Thompson, T. Troscianko, Basic vision: an introduction to visual perception, Oxford University Press, 2012.

[7] T. Konkle, T. F. Brady, G. A. Alvarez, A. Oliva, Conceptual distinctiveness supports detailed visual long-term memory for real-world objects., Journal of Experimental Psychology: General 139 (2010) 558.

[8] T. F. Brady, T. Konkle, G. A. Alvarez, A. Oliva, Visual long-term memory has a massive storage capacity for object details, Proceedings of the National Academy of Sciences 105 (2008) 14325–14329.

[9] S. Wiseman, U. Neisser, Perceptual organization as a determinant of visual recognition memory, The American Journal of Psychology (1974) 675–681.

[10] G. M. Huebner, K. R. Gegenfurtner, Conceptual and visual features contribute to visual memory for natural images, PLoS One 7 (2012) e37575.

[11] P. G. Schyns, R. L. Goldstone, J.-P. Thibaut, The development of features in object concepts, Behavioral and Brain Sciences 21 (1998) 1–17.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[13] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. García Seco de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, M. Sultana, Overview of the MediaEval 2022 predicting video memorability task, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2023.

[14] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, arXiv preprint arXiv:2210.08402 (2022).

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.