Faculdade de Design,
Tecnologia e Comunicação
Universidade Europeia

2024

Fernando Lamar Corrêa dos Santos

Predicting video memorability using traditional and incremental approaches

Dissertation submitted to IADE - Faculty of Design, Technology and Communication of the European University, in fulfillment of the necessary requirements for obtaining the Master's degree in Creative Computing and Artificial Intelligence, carried out under the scientific supervision of Dr. André Miguel Guedelha Sabino, Assistant Professor at the Faculty of Design, Technology and Communication of the European University, and Dr. Jacinto Paulo Simões Estima, Assistant Professor at the Department of Informatics Engineering of the University of Coimbra.

I would like to express my sincere gratitude to Europeia ID for recognizing the value of this project and approving the necessary financial support. Their approval, following a thorough review of the thesis proposal, was instrumental in facilitating the progress of this research.

Acknowledgements

I am deeply grateful to several individuals whose contributions were integral to the completion of this research project. First and foremost, I extend my heartfelt thanks to Dr. André Miguel Guedelha Sabino and Dr. Jacinto Paulo Simões Estima, my esteemed orientators. Their extensive knowledge, invaluable guidance, and steadfast support have been indispensable throughout every stage of this study. Their constructive feedback, insightful discussions, and encouragement have not only shaped the direction of this research but also nurtured my growth as a researcher. Their dedication to mentoring and their commitment to academic excellence have been truly inspiring. I also wish to thank Nuno Soler Lopes for their expertise and assistance in crafting the graphs and images, which greatly enhanced the clarity and impact of this thesis.
Furthermore, I am grateful to IADE and Europeia ID for their financial support, which provided us with state-of-the-art machines capable of solving our computational challenges. This support was instrumental in conducting the complex simulations and analyses necessary for this study.
Their collective contributions have made this endeavor possible, and I am deeply grateful for their generosity, patience, and unwavering belief in this project.

palavras-chave treino incremental, memorabilidade de vídeo, ViViT, restrições de hardware, modelos de transformadores.

resumo Este estudo investiga a viabilidade do treino incremental como uma alternativa aos métodos tradicionais de treino para a previsão de memorabilidade de vídeos, particularmente em ambientes com limitações de hardware. Utilizando o modelo ViViT, uma arquitetura baseada em transformers, a pesquisa pretende responder à principal questão de saber se o treino incremental pode fornecer desempenho estável e consistente com menores demandas computacionais (RQ1). Foram realizados duas experiencias: um comparando os métodos de treino incremental e tradicional, e outro aplicando o treino incremental ao conjunto de dados completo. Os resultados indicam que o treino incremental é uma alternativa viável, oferecendo desempenho comparável em métricas de erro como o Erro Quadrático Médio (MSE) e o Erro Absoluto Médio (MAE), enquanto reduz significativamente a carga computacional. No entanto, o treino incremental apresentou limitações na precisão de ordenação, medida pela Correlação de Rank de Spearman (SRC), em comparação com os métodos tradicionais. Os resultados sugerem que o treino incremental pode ser uma solução prática para a previsão de memorabilidade de vídeos em cenários com restrições de recursos, mas refinamentos adicionais são necessários para melhorar o desempenho em tarefas de ordenação. Trabalhos futuros devem explorar otimizações arquiteturais, configurações de entrada, expansão de conjuntos de dados, incorporação de dados multimodais e ajuste da arquitetura ViViT para um melhor manuseio de dependências de longo prazo.

Keywords incremental training, video memorability, ViViT, hardware constraints, transformer models

abstract This study investigates the viability of incremental training as an alternative to traditional training methods for video memorability prediction, particularly in hardware-constrained environments. Using the ViViT model, a transformer-based architecture, the research seeks to address the primary question of whether incremental training can provide stable and consistent performance with reduced computational demands (RQ1). Two experiments were conducted: one comparing incremental and traditional training methods and another applying incremental training to the full dataset. The results indicate that incremental training is a feasible alternative, offering comparable performance in error metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE), while significantly reducing the computational load. However, incremental training exhibited limitations in ranking accuracy, as measured by Spearman's Rank Correlation (SRC), when compared to traditional methods. The findings suggest that incremental training can provide a practical solution for video memorability prediction in resource-constrained scenarios, but further refinement is needed to improve its performance in rank-order tasks. Future work should explore architectural optimizations, optimizing input configurations, expanding datasets, incorporating multimodal data, and tuning the ViViT architecture for better long-range dependency handling.

Abbreviations

AI	**A**rtificial **I**ntelligence
ML	**M**achine **L**earning
CNN	**C**onvolutional **N**eural **N**etwork
ViT	**V**ision **T**ransformer
ViViT	**V**ideo **V**ision **T**ransformer
MSE	**M**ean **S**quared **E**rror
MAE	**M**ean **A**bsolute **E**rror
SRC	**S**pearman's **R**ank **C**orrelation

Contents

List of Figures

List of Tables

Acknowledgements

I wish to acknowledge a number of people, whose assistance was important for the accomplishment of this work. For the first place, let me express the sincerest appreciation to my lance orientators Dr. André Miguel Guedelha Sabino and Dr. Jacinto Paulo Simões Estima. For all phases of this study their wide knowledge, precious insight and unfailing support have proved essential. Not only this research has been greatly positively influenced by their constructive comments and productive debates, but also I have added to the development of myself as a researcher. Their willingness to teach and their standards of excellence have been amazing.

I would also like to thank Nuno Soler Lopes for their help and skills with the graphs and pictures, which were an important addition to this thesis.

And lastly, I would like to thank IADE for the financial aid, which allowed them to purchase adequate machines that were capable to answer our computational needs. This assistance proved essential in performing the intricate modeling and analysis required in this study.

As I have said, all these people contributed to making this work possible and I am honestly grateful to them for their generosity, calmness and faith in this idea.

Chapter A

Introduction

This document presents the work developed as part of the Master's dissertation in Creative Computing and Artificial Intelligence, submitted to the faculty of IADE - Faculdade De Design Tecnologia e Comunicação da Universidade Europeia. The dissertation addresses the problem of video memorability prediction.

This chapter provides an overview of the main topic (A.1), explaining its relevance and the underlying factors that influence memorability in videos. Additionally, it explores methods used to overcome computational constraints when training machine learning models, with a specific focus on incremental training (A.2), which helps alleviate hardware limitations by efficiently processing large datasets.

This chapter also delves into the motivation behind the research (A.3), highlighting the significance of understanding video memorability in fields like marketing, education, and content retrieval. It explains how improving the ability to predict memorable content can enhance the effectiveness of video design and user experience across various applications.

In Section A.4 the specific challenges of predicting video memorability are outlined, such as dealing with the temporal dynamics of video, duration, and frame rate. This section addresses the complexities involved in video memorability prediction, especially compared to static image memorability.

Section A.5, outlines the primary goal of this research, which is to propose a novel method for predicting video memorability. This objective focuses on using transformer-based models, specifically ViViT, to address the challenges mentioned earlier and provide more accurate predictions through advanced techniques like regression analysis.

The research questions guiding the study are presented in Section A.6. These questions examine how the ViViT model improves video memorability prediction compared to traditional methods and evaluate the viability of incremental training under hardware constraints.

Finally, Section A.7, summarizes the key contributions of this work, including the

introduction of a novel transformer-based model for video memorability prediction, the submission of a related paper to an international conference, and the release of an open-source code repository for further research. This section sets the stage for the remainder of the thesis, which will explore the methods, experiments, and findings in greater detail.

A.1 Video Memorability

Video memorability refers to the measure of how well individuals remember specific video content over time. Understanding this concept is crucial in various fields such as advertising design, social media recommendation, education, and marketing. Past studies have shown that the memorability of images and videos is an intrinsic feature, with some content being inherently more memorable than others (Bainbridge et al., 2013; Borkin et al., 2013). This intrinsic memorability can be leveraged to predict and manipulate subsequent memories, making it a valuable metric for content creators and advertisers (Shekhar et al., 2017; Bainbridge et al., 2013).

Memorability research initially focused on images before evolving to analysing video content, as researchers recognized that many of the same principles could be applied to both static and dynamic media. Key attributes that make an image memorable (Isola et al., 2011) are a composition of several factors. Memorability needs to be a stable property shared across different viewers and is experimentally measured as the probability of detecting image repeats during a Visual Memory Game as shown in Figure A.1 (Newman et al., 2020), which is a critical experimental setup used to measure the memorability of images. In this game, participants are shown a continuous stream of images and are asked to press a key whenever they recognize a repeated image from the stream. The images are displayed at varying intervals, and the lag between the first and subsequent appearances of an image is adjusted to test the decay of memory over time. The key measure of an image's memorability is the hit rate, the probability that a viewer correctly identifies a repeated image. This method provides robust data on which images are inherently more memorable.

The game's design helps to capture the intrinsic memorability of images, factoring out individual differences in memory by aggregating responses across many viewers. By using large-scale datasets and consistent experimental conditions, researchers can obtain reliable memorability scores for each image. These scores are then analyzed in conjunction with various image features to understand what makes certain images more likely to be remembered. This approach has been essential in advancing our understanding of visual memory and in developing models that can predict the memorability of new images based on their features. Object and scene semantics play a crucial role, with objects

carrying semantic meaning contributing more to memorability. While color and simple image features like mean hue show weak correlations, global image descriptors such as GIST(L. Yang et al., 2009), SIFT(Tao et al., 2010; Isola et al., 2011; Dubey et al., 2015), HOG2x2(Sidorov, 2019; Isola et al., 2011; Dubey et al., 2015), and SSIM(Fajtl et al., 2018) are explored for their potential in predicting memorability, given their effectiveness in scene recognition. Research into video memorability has gained significant attention due

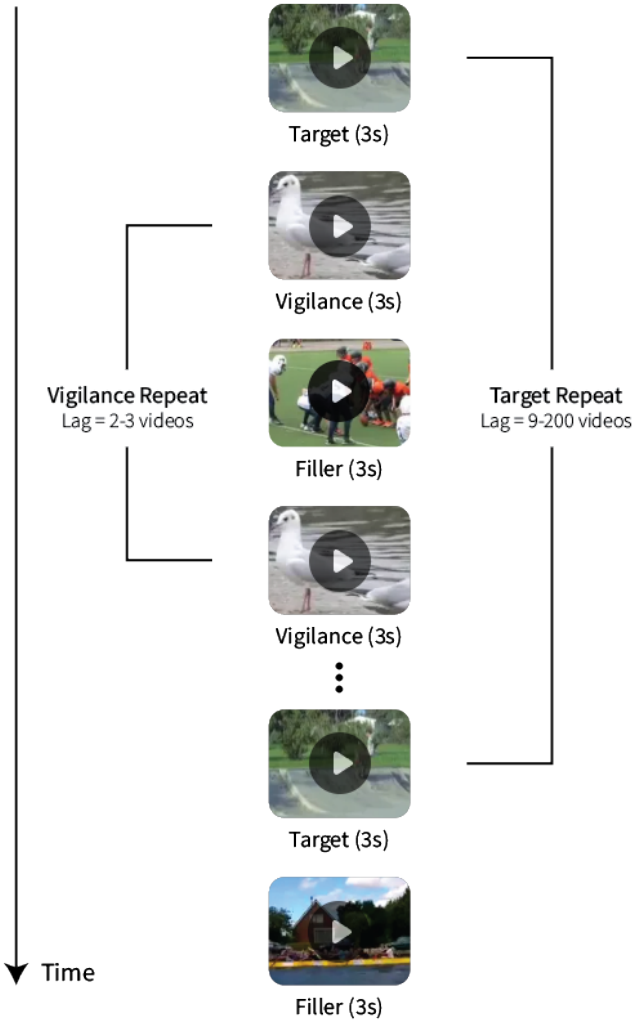

Figure A.1: "Task flow diagram of The Memento Video Memory Game. Participants see a continuous stream of videos and press the space bar when they see a repeat." (Newman et al., 2020).

to its potential applications in content retrieval, filtering, summarization, and storytelling (Lu & Wu, 2022). Algorithms developed to assess the memorability of videos based on their content provide insights that aid in decision-making processes when selecting between competing videos (Cohendet, Yadati, et al., 2018).

Further insights into image and video memorability are provided by Newman et al. (2020). The research explores into additional factors that contribute to memorability,

expanding beyond the previously mentioned attributes. It emphasizes the importance of considering temporal aspects in the case of videos, exploring how the duration of scenes or specific events influences their memorability. Additionally, the study introduces novel metrics and methodologies for assessing memorability, contributing to a comprehensive understanding of the complex interplay of factors that shape the lasting impact of visual content.

A.2 Incremental Training

This work also focuses on the incremental training approach for machine learning tasks, comparing it with the traditional training method, which can be resource intensive, particularly when dealing with heavy multimedia content, such as predicting media memorability, involving high-resolution imagery or video. Incremental training is particularly advantageous when dealing with large datasets and limited computational resources, as it allows for progressive learning without the need for extensive retraining (Y. Wu et al., 2019). Unlike traditional training, where all data is inputted into the model at once for training, incremental training involves dividing the data into smaller sets that train the model incrementally. This study explores how incremental training can improve model performance and efficiency, providing a detailed analysis of its benefits over Traditional training methods in the specific context of video memorability prediction.

Loading a larger dataset during the training process requires suitably large storage and computational resources, which are not always available. For this reason, we want to evaluate the viability of the incremental approach, which splits the dataset into smaller chunks that require less processing power than approaches analyzing the full training dataset at once.

Incremental training methods effectively address hardware constraints by updating models incrementally rather than retraining from scratch. The Train++ algorithm for MCUs reduces memory and computational demands by processing new data in small batches (Sudharsan et al., 2021). Similarly, End-to-End Incremental Learning maintains a small set of exemplars to balance memory use and computational load, preventing catastrophic forgetting (Castro et al., 2018). The Few-shot Class-Incremental Learning (FSCIL) framework optimizes memory and computational resources by using data augmentation and model optimization techniques to learn new classes without forgetting previous ones(Tian et al., 2024). These strategies enable efficient deployment of machine learning models on devices with limited resources.

A.3 The Motivation

Memorability, an essential aspect of human cognition characterized by the ability to recall visual content, is closely linked to the perceptual storage capacity of human memory. For the design of any effective system involving human interaction, it is crucial to consider cognitive and psychological factors (Shekhar et al., 2017).

Video memorability introduces added complexities due to factors such as duration, frame rate, and the temporal structure of videos. The influence of audio on video memorability has also been investigated, with the development of multimodal deep learning-based systems to estimate overall video memorability (Sweeney et al., 2021). Improving video memorability is deemed interesting because memorable videos are more likely to be shared, viewed, and discussed (Mudgal et al., 2024). Memorability prediction is crucial for applications that interact with humans like advertising, film-making, education, and content retrieval, with the task addressing the challenge through a common benchmarking protocol.

The ability of videos to leave a lasting impression on viewers holds the potential to impact society in both positive and negative ways. As such, the prediction of video memorability has become increasingly important. This interest is driven by its wide-ranging applications in content retrieval, filtering, summarization, and storytelling. Understanding which videos are likely to be memorable can help tailor content more effectively to audience needs and preferences, thereby enhancing the overall viewer experience(Lu & Wu, 2022). For instance, in education, designers could use memorability predictions to create educational videos that are more likely to be retained by students. In the field of video summarization, these models could assist in selecting segments that are more memorable, enhancing the effectiveness of summarization algorithms. Additionally, understanding video memorability may have implications for content creators and marketers in designing more impactful and memorable videos for communication and advertising purposes.

A.4 The Problem

The task of predicting video memorability presents several challenges due to the inherent complexities involved in videos as compared to static images. While memorability research has made strides in understanding how visual content is recalled, the prediction of video memorability introduces additional difficulties such as the temporal structure, duration and frame rate. Current approaches, which often rely on deep learning models, still face challenges in effectively capturing these elements to accurately predict memorability.

A.5 The Objective

To contribute to the solution of video memorability prediction, we will seek to propose a novel method for measuring video memorability by using pioneering technologies, specifically focusing on the application of transformer models, to accurately predict a video's memorability score through regression analysis.

A.6 Research Questions

In this study, we aim to address two key research questions that explore the effectiveness of the ViViT model and incremental training methods for video memorability prediction under hardware limitations. The research questions are outlined in Table A.1.

RQ1	Can incremental training be a viable alternative to traditional training for predicting video memorability under hardware constraints, and how do these methods compare in terms of performance, stability, and trade-offs?
RQ2	How does the ViViT model improve the accuracy of video memorability prediction compared to traditional models under hardware-constrained environments?

Table A.1: Research Questions Addressed in this Study

A.7 Contributions

- **Novel model** This thesis introduces ViViT(Arnab et al., 2021), a novel model in the domain of video memorability prediction. ViViT is the first transformer-based architecture applied to this specific task, building on the transformer model's ability to handle long-range dependencies in temporal sequences. By leveraging this advanced architecture, we aim to improve video memorability prediction accuracy over previous models. This contribution is significant in the field of video memorability, where traditional methods such as CNNs and RNNs have been primarily used.

- **WI-IAT 2024** A paper detailing our findings has been submitted to the The 23rd IEEE/WI-IAT 2024 conference, under the workshop titled "WS#20: The International Workshop on AI and Computer Science". The paper is titled "Comparative Analysis of Incremental and Traditional Training Methods under Hardware Limitations: The ViViT Model Case". This submission explores the impact of incre-

mental training methods on transformer models in video memorability prediction, especially under hardware constraints.

- **Code Repository** The code used for the experiments in this thesis—including training, dataset preprocessing, and the generation of images and graphs—will be made available in an open-source GitHub repository. This repository will enable further research and experimentation using our methodology and models. The repository can be accessed at the following link: `https://github.com/ElynoLamar/VideoMemorabilityVivit`.

A.8 Document structure

This thesis is organized as follows:

Chapter 1 - Introduction: This chapter introduces the main topic of video memorability prediction, outlines the research problem and gaps, and presents the contributions made by this thesis. Chapter 2 - Literature Review: In this chapter, we review existing works in the fields of computer vision, deep learning, and video memorability prediction. Key models and techniques, such as CNNs, Transformers, and ViViT, are discussed. Chapter 3 - Methodology: This chapter explains the methodology used for the development and evaluation of the ViViT model, detailing the incremental training setup, evaluation metrics, and data preprocessing. Chapter 4 - Experimental Setup: Here, we discuss the datasets, hardware limitations, and the experimental framework used to train and test our models. Chapter 5 - Results: This chapter presents and analyzes the results obtained from the experiments, comparing incremental and traditional training methods. Chapter 6 - Analysis: In this chapter, we discuss how our results compare with the state-of-the-art models and address the research questions posed in the introduction. Chapter 7 - Conclusion: The final chapter provides a summary of the research, highlights the limitations, and outlines potential directions for future work.

Chapter B

Literature review

The literature review will explore the evolution of computer vision, highlighting key advancements in deep learning techniques like Convolutional Neural Networks (CNNs) and their impact on tasks such as image recognition and segmentation. As research progressed, transformer models, originally developed for natural language processing, have emerged as powerful tools in vision tasks due to their ability to capture long-range dependencies through self-attention mechanisms. This shift has given rise to Vision Transformers(Hagen & Espeseth, 2023) (ViT), which process images as sequences of patches, offering superior performance in image recognition. Further expanding this framework, Video Vision Transformers (ViViT)(Arnab et al., 2021) extend these capabilities to video analysis by addressing both spatial and temporal dynamics. The review will also address existing approaches to predicting video memorability, a growing area of interest in computer vision, and how current models, including ViViT, aim to improve prediction accuracy. Finally, incremental training methods, which address hardware limitations by allowing models to learn progressively, will be explored in the context of transformer-based models for video memorability prediction, identifying gaps in existing research and establishing the relevance of this study.

B.1 Computer Vision

B.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have proven to be highly effective and successful in the field of computer vision, particularly for tasks like image classification, object detection, and segmentation. The integration of CNNs in various applications has made them the de facto standard in the field, achieving state-of-the-art performance and significantly advancing computer vision technology. CNNs are celebrated for their hierarchical

learning capabilities and automatic feature extraction, which have been instrumental in solving complex vision problems and enhancing the accuracy of predicting image memorability with growing precision over time (Bhatt et al., 2021; Needell & Bainbridge, 2021; Squalli-Houssaini et al., 2018; Leonardi et al., 2019). Recent studies have further highlighted the prowess of CNNs, not only in traditional areas but also in competitions related to computer vision and image processing. This has solidified their reputation as a special type of neural network with exceptional capabilities (Khan et al., 2020). The success of CNNs has rendered traditional methods based on handcrafted features obsolete(Vázquez et al., 2017) and has led to substantial improvements in the state-of-the-art of many computer vision tasks (Liu et al., 2021; Wang & Lee, 2021; Dias et al., 2018). However, despite the dominance of CNNs, there is a growing trend towards the emergence of vision transformers (ViTs) that are beginning to rival CNN-based approaches in the field of computer vision (Xue et al., 2022). Nevertheless, CNNs continue to be essential in achieving exceptional performance in tasks such as image recognition, object detection, and semantic segmentation, and remain critical in pushing the boundaries of what is achievable in computer vision (T. Yang et al., 2017). In conclusion, CNNs have not only revolutionized the state of the art in computer vision but have also been fundamental in advancing research in the field of memorability, where researchers have utilized these networks from the early stages to enhance their ability to forecast the memorability of images with ever-growing precision. As technology evolves, CNNs remain an essential component in the toolbox of computer vision, despite the rising interest in alternative architectures like vision transformers.

B.1.2 Transformers

Transformers have emerged as a transformative architecture in deep learning, particularly for natural language processing (NLP) tasks. Originally introduced in the seminal paper by Vaswani et al. (2017), the transformer model leverages attention mechanisms to effectively capture complex relationships within sequential data, surpassing traditional models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) in both efficiency and scalability (Tay et al., 2020). The architecture's ability to perform parallel computations through self-attention mechanisms allows it to process all elements of a sequence simultaneously, which is a significant departure from the sequential processing inherent in RNNs and LSTMs (Cao, 2023). This parallelization not only enhances computational efficiency but also enables the model to learn long-range dependencies that previous architectures struggled to capture (Dai et al., 2019).

The transformer model is fundamentally structured around two primary components: the encoder and the decoder. The encoder transforms the input data—whether it be a

sentence in NLP or patches of an image in computer vision—into a set of continuous representations. Conversely, the decoder utilizes these representations to generate the desired output sequence (H. Lin et al., 2021). This encoder-decoder architecture has proven to be versatile, extending the utility of transformers beyond NLP into fields such as computer vision, where they are referred to as Vision Transformers (ViTs) (Shi et al., 2023). The adaptability of transformers to various domains underscores their significance in the current landscape of deep learning.

Central to the effectiveness of transformers is the attention mechanism, particularly the self-attention variant. This mechanism allows the model to assign varying levels of importance to different parts of the input sequence when processing each element, thereby enabling it to manage long-range dependencies more effectively than RNNs and LSTMs (Lamb, 2021). The self-attention mechanism operates by computing a weighted representation of the input, where the weights are determined by the relevance of each element to the others in the sequence (Chen, 2024). This capability is particularly advantageous in tasks requiring the understanding of context over extended sequences, such as language translation and image recognition (WANG & ZHAO, 2023).

In the realm of computer vision, Vision Transformers have adapted the principles of the transformer architecture to process visual data by treating images as sequences of patches. This pioneering approach allows transformers to be applied to image recognition tasks, demonstrating their flexibility and effectiveness beyond traditional CNNs (Shi et al., 2023). The ability of transformers to handle visual data through self-attention mechanisms has led to significant advancements in image classification and synthesis tasks, showcasing their potential to revolutionize the field of computer vision (Carion et al., 2020).

The evolution of transformers has also prompted research into various enhancements and adaptations of the original architecture. For instance, modifications to the attention mechanism, such as the introduction of competitive ensembles of independent mechanisms, have been explored to improve the performance of transformers (Lamb, 2021). Additionally, the integration of Fourier transforms into the attention mechanism has been proposed as a means to further enhance the efficiency and effectiveness of token mixing in transformer models (Lee-Thorp et al., 2021). These ongoing innovations reflect the dynamic nature of transformer research and its potential for continued advancements in deep learning applications.

Self-Attention

Self-Attention is a core component of the transformer model, unlike traditional approaches like RNNs and CNNs that require data to be processed sequentially or in localized con-

texts, self-attetion alters how we process sequences in neural networks by allowing the model to weigh the importance of different parts of the inputted data. Self-ttention computes the attention scores between all pairs of positions in the input sequence. For a given position, the self-attention mechanism allows the model to consider how relevant every other position is and to aggregate these insights to produce a new representation for that position.(Vaswani et al., 2017). The traditional self-attention approach can be computationally intensive, particularly with high-resolution images, leading to challenges in efficiency and scalability(Venkataramanan et al., 2023).

Multi-Head Attention

Multi-Head Attention is a essential enhancement of the self-attention mechanism utilized in transformer architectures, facilitating the simultaneous capture of diverse relationships within data. This mechanism is particularly significant in the context of natural language processing (NLP) and computer vision, where it allows models to discern various patterns and relationships from multiple perspectives. By partitioning the input into several heads, each with distinct learned weights, multi-head attention enables the model to focus on different aspects of the input sequence, effectively capturing various subspaces of attention. This capability is essential for tasks that require nuanced understanding and representation of complex data structures.

The foundational principle of multi-head attention lies in its ability to apply multiple attention functions in parallel. Given a set of queries (Q), keys (K), and values (V), the multi-head attention mechanism computes self-attention for each head independently. The mathematical formulation of this process is expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k denotes the dimensionality of the key vectors. This formula illustrates how the attention scores are derived from the interactions between queries and keys, scaled by the square root of the key dimension, and subsequently applied to the values to produce the output. The results from each head are concatenated and projected back into the original space, thereby enhancing the model's ability to attend to various positions in the input sequence from multiple viewpoints. This multi-faceted approach significantly enriches the representational capacity of the transformer model, allowing it to learn complex relationships more effectively.

The advantages of multi-head attention extend beyond mere representation; they are particularly pronounced in applications such as NLP and computer vision. In NLP tasks, for instance, different heads can specialize in capturing syntactic relationships, such as

grammar and structure, while others may focus on semantic connections, such as meaning and context. This specialization allows for a more comprehensive understanding of language, leading to improved performance in tasks like machine translation and sentiment analysis (Evstropov, 2023). Similarly, in Vision Transformers, different heads can concentrate on various spatial regions or image features, enabling the model to capture a wide array of patterns across images, which is crucial for tasks like object detection and image segmentation (Carion et al., 2020).

The process of multi-head attention can be succinctly summarized in three primary steps: first, the input data is divided into multiple smaller subspaces, with each head assigned its own query, key, and value matrices. This division allows for a more granular analysis of the input data. Second, each head independently performs the attention operation, focusing on different parts of the input and thereby capturing distinct relationships. Finally, the results from all heads are concatenated and projected into the output space, culminating in a comprehensive representation that integrates the insights gleaned from each head. This structured approach to attention not only enhances the model's ability to learn from data but also contributes to its robustness and generalization capabilities across various tasks and domains (Quinton, 2024).

The significance of multi-head attention in the transformer architecture cannot be overstated. It is a critical component that enables transformers to achieve state-of-the-art performance across a multitude of tasks, from language modeling to image classification. The ability to model multiple relationships simultaneously allows transformers to adapt to the complexities of real-world data, making them indispensable tools in the fields of artificial intelligence and machine learning. As research continues to evolve, the exploration of multi-head attention and its applications will undoubtedly lead to further advancements in these domains (Burtsev, 2020).

Encoder-Decoder Architecture

The encoder-decoder architecture, as stated by Vaswani et al. (2017), is a fundamental framework used in various machine learning tasks. The encoder is responsible for processing the input data and transforming it into a fixed-dimensional representation. In the context of natural language processing, the input data could be a sequence of words or tokens. The encoder analyzes this input sequence and generates a condensed representation, often referred to as a context vector or hidden state, that captures the relevant information from the input. This condensed representation is typically a continuous vector that encodes the input sequence's semantics and context. decoder, on the other hand, takes the encoded representation produced by the encoder and uses it to generate an output sequence.It is typically the same structure to the encoder but mirroed, operating in

reverse. In the context of natural language processing tasks like machine translation or text generation, the output sequence is usually another sequence of words or tokens in a different language or format. The decoder processes the encoded representation and generates the output sequence step by step, often feeding past generated tokens/outputs to the new prediction.

B.1.3 ViT

The emergence of Vision Transformers(Hagen & Espeseth, 2023) (ViTs) bring superiority over Convolutional Neural Networks (CNNs) in image recognition. The Visiopn Transformer architecture involves dividing input images into patches, which are then transformed into vectors and processed through a standard Transformer model.

The key advantage of Vision Transformers lies in their ability to maintain a large field of view from the outset, overcoming the progressively expanding receptive fields of CNNs during training. ViTs' success is due to its capacity to handle substantial amounts of training data and the freedom to explore diverse perspectives on the data without predefined focus patterns. This contrasts with CNNs, which start with a localized focus dictated by convolutional layers(Dosovitskiy et al., 2020).

Vision Transformers are data-hungry and require extensive training to learn how to focus and make accurate predictions(Hagen & Espeseth, 2023). While ViTs excel in heavyweight, state-of-the-art scenarios with significant computational and energy resources (Dosovitskiy et al., 2020), the enduring significance of CNNs and traditional methodologies persists in the realm of lightweight machine learning, particularly for production purposes, computational efficiency, and the promotion of environmentally sustainable AI.

The authors of ViT model wrote about their model structure:

> "An overview of the model is depicted in Figure B.1. The standard Transformer receives as input a 1D sequence of token embeddings. To handle 2D images, we reshape the image $x \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $xp \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = \frac{HW}{P^2}$ is the resulting number of patches, which also serves as the effective input sequence length for the Transformer. The Transformer uses a constant latent vector size D through all of its layers, so we flatten the patches and map to D dimensions with a trainable linear projection (Eq. 1). We refer to the output of this projection as the patch embeddings."(Dosovitskiy et al., 2020)

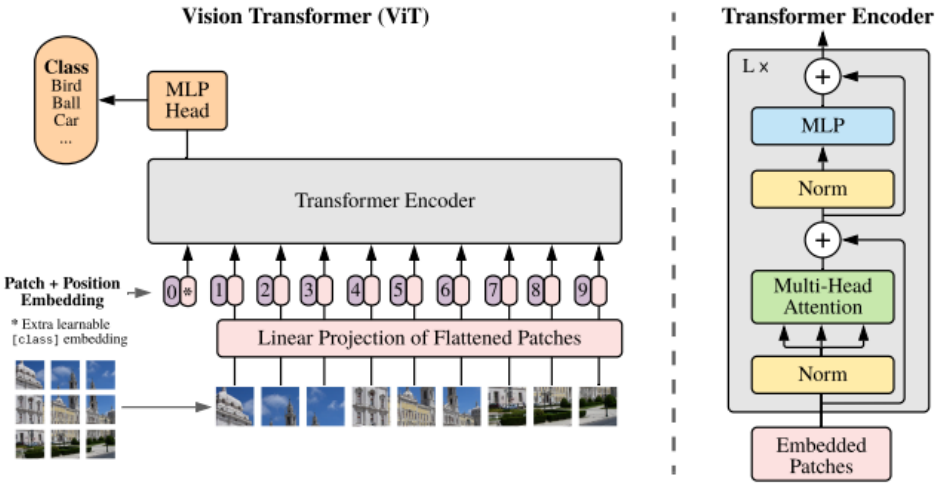

Figure B.1: "Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017)." (Dosovitskiy et al., 2020).

Attention in Vision Transformers

In Vision Transformers (ViTs), the attention mechanism is essential for the model's ability to focus on relevant segments of the input image. This mechanism diverges significantly from traditional convolutional layers, which operate with fixed receptive fields. Instead, the attention mechanism in ViTs is adaptive, allowing the model to attend to relationships between image patches irrespective of their spatial distance. This global attention mechanism operates across all patches of an image concurrently, enabling the model to capture both local and long-range dependencies effectively. The adaptability of this attention mechanism is a fundamental aspect that enhances the model's performance in various vision tasks, as it allows for a more nuanced understanding of the input data (Naseer et al., 2021), (H. Wu et al., 2021).

The formulation of the attention mechanism involves the computation of attention weights, which dictate the focus each patch should have when processing other patches. This process initiates with the generation of three vectors: queries, keys, and values, which are derived from the input patch embeddings. The attention scores are computed by taking the dot product of the query and key vectors, normalized by the dimension of the key vectors, and subsequently processed through a softmax function. These scores are then utilized to weight the corresponding value vectors, culminating in a new representation for each patch that integrates information from all other patches. This method allows the Vision Transformer to learn adaptively which parts of the image are most pertinent

for a specific task, thereby capturing more complex and intricate patterns compared to traditional CNNs, which are limited by the size of their filters (Lee & Kang, 2022), (S. Li et al., 2021).

The flexibility inherent in the attention mechanism of ViTs provides superior performance in tasks requiring an understanding of both fine-grained details and broader contextual information. This adaptability is particularly beneficial in scenarios where the relationships between distant patches are crucial for accurate interpretation, such as in object detection and image classification tasks. The ability of ViTs to process information globally rather than locally enables them to outperform traditional convolutional neural networks (CNNs) in various benchmarks, showcasing their potential in advancing the field of computer vision (J. Li et al., 2022), (S. Li et al., 2021).

Embedded Patches

In the context of embedded patches, Vision Transformers process images by dividing them into smaller, fixed-size patches. These patches serve as the fundamental units of input for the Transformer model, analogous to tokens in natural language processing. Each patch is flattened into a vector and then linearly projected into a higher-dimensional space, resulting in what is known as a patch embedding. This transformation is crucial as it allows the model to handle visual data in a manner similar to how it processes textual data, thus facilitating the application of Transformer architectures to vision tasks (W. Wang, Lu, et al., 2021), (W. Wang, Xie, et al., 2021).

To convert an image into patches, the input image is segmented into non-overlapping square regions of equal size, typically denoted as P×P, where P represents the height and width of each patch. Each patch is then flattened into a one-dimensional vector and passed through a learnable linear layer to produce an embedding of size D, where D corresponds to the dimension of the hidden layers in the Transformer. This transformation effectively converts the two-dimensional spatial information from the image into a one-dimensional sequence of patch embeddings, which can be processed by the Transformer in the same manner as word tokens in text data (W. Wang, Lu, et al., 2021), (W. Wang, Xie, et al., 2021).

In addition to these embeddings, positional encodings are incorporated into each patch to retain spatial information, as Transformers do not inherently capture the positional structure of the input. The inclusion of positional encoding is essential as it allows the model to differentiate between patches based on their spatial locations, thereby enabling it to maintain a sense of spatial context. This step is particularly critical for tasks that necessitate an understanding of spatial relationships, such as object detection and image classification, where the relative positioning of patches can significantly influence the

model's performance (K. Wu et al., 2021), (Chu et al., 2021).

B.1.4 ViViT

ViViT(Arnab et al., 2021) (Video Vision Transformer) is a video classification model and is highlighted as the first pure transformer video classification model(Arnab et al., 2021), outperforming state-of-the-art modelsl(Arnab et al., 2021) like SlowFast(Feichtenhofer et al., 2018), TimeSformer-HR, TSM(J. Lin et al., 2018), STM(Jiang et al., 2019), TEA(Y. Li et al., 2020), and blVNet(Fan et al., 2019). The article explores into the model's contributions, specifically addressing the classical flow of video transformers and proposing a more efficient tokenizer to calculate self-attention.

The ViViT model is detailed across four variations, with Model 2 (Factorised encoder) noted to perform the best. The factorised encoder involves separate transformers for spatial and temporal processing, with a "Late Fusion" of temporal information. Model 3 (Factorised self-attention) and Model 4 (Factorised dot-product attention) are also discussed, each presenting distinct approaches to handling attention mechanisms(Arnab et al., 2021).

With the right use of hyperparameters indicate that ViViT outperforms almost all state-of-the-art models across different datasets, such as embedding methods, input image resolution, the number of views or frames, and input token size(Arnab et al., 2021).

In ViT, the image undergoes a patch-based division, followed by spatial flattening, a procedure referred to as tokenization. When dealing with videos, a similar approach can be applied to individual frames as shown in Figures B.2 and B.3. The authors of ViViT propose a tokenization scheme called uniform frame sampling, where in frames are sampled from the video clip, and a straightforward ViT tokenization process is executed.

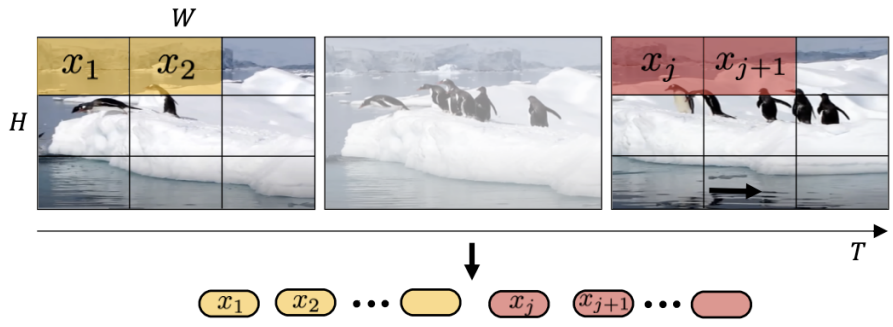

Figure B.2: "Uniform frame sampling: We simply sample nt frames, and embed each 2D frame independently following ViT.Dosovitskiy et al. (2020)." Arnab et al. (2021)

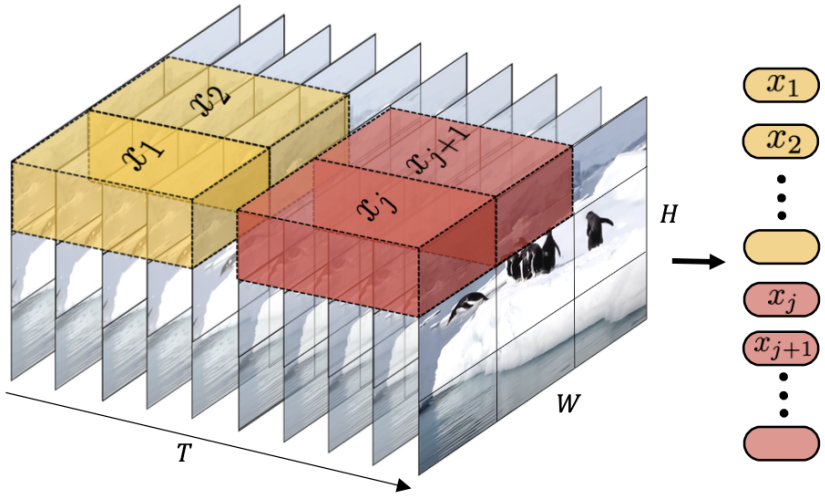

Figure B.3: "Tubelet embedding. We extract and linearly embed nonoverlapping tubelets that span the spatio-temporal input volume" Arnab et al. (2021)

B.1.5 DinoV2- Self-supervised Vision Transformer Model

Dino or the Data-Efficient Image Transformers, has demonstrated remarkable performance in various computer vision tasks, such as image retrieval, copy detection, and video instance segmentation (Wang et al., 2023). Vision Transformer (ViT), a type of DINO, has exhibited superior results compared to state-of-the-art convolutional networks while requiring significantly fewer computational resources for training (Dosovitskiy et al., 2020). Furthermore, DINO has been shown to perform on par with the state of the art on ResNet-50, validating its effectiveness in the standard setting (Caron et al., 2021). Additionally, DINO has been explored as a state-of-the-art self-supervised learning (SSL) algorithm, distilling knowledge from two augmented views of an input image, and has been further extended to joint SAR-optical representation learning (Y. Wang, Albrecht, & Zhu, 2022; Y. Wang, Albrecht, Braham, et al., 2022). Moreover, DINO has been evaluated and compared with other state-of-the-art SSL methods, such as MoCo v2, SimCLR, and SwAV, in the context of surgical computer vision, demonstrating its relevance and applicability in diverse domains (Ramesh et al., 2022).

B.1.6 Swin Transformer Model

The Swin Transformer(Krizhevsky et al., 2012) was developed as a response to these limitations of ViT(Dosovitskiy et al., 2020), introducing several architectural innovations that tailored the Transformer framework more effectively for visual tasks:

Hierarchical Structure: Unlike standard Transformers that treat the image as a flat sequence of patches, the Swin Transformer introduces a hierarchical structure. This

design processes images in stages, effectively reducing dimensionality and aggregating semantic features progressively. This approach not only helps in handling different scales of visual information but also mimics the multi-resolution processing of CNNs, thereby enabling more efficient learning dynamics.

Shifted Window-Based Attention: One of the hallmark features of the Swin Transformer is its shifted window-based attention mechanism. In this configuration, the self-attention is calculated within local windows that shift between successive layers. This technique allows the model to capture both local and long-range dependencies without the prohibitive computational costs of global attention. By alternating the positions of these windows, the model ensures comprehensive coverage of interactions across the entire image, thereby overcoming one of the major drawbacks of localized attention approaches.

Patch Merging: The architecture also incorporates a patch merging strategy that acts similarly to downsampling in CNNs. This process involves merging adjacent patches to reduce the spatial resolution while increasing the depth of feature maps. Such a mechanism is critical in building a multiscale representation that is more adaptable to various visual phenomena and can process higher-level semantic information more effectively.

The Swin Transformer(Krizhevsky et al., 2012) has been rigorously tested across multiple benchmark datasets, showcasing its superior capabilities over both traditional CNNs and other Transformer-based models. In image classification tasks on the ImageNet(Krizhevsky et al., 2012) dataset, Swin Transformers have demonstrated enhanced accuracy and efficiency. In object detection and semantic segmentation tasks using datasets like COCO 2017(T.-Y. Lin et al., 2015) and ADE20K(Zhou et al., 2019), the model not only improved the baseline performance but also showed greater scalability and adaptability across different computational settings.

Specifically, the Swin Transformer excels in handling complex visual tasks that benefit from understanding both fine details and broader contextual information, making it a robust choice for a variety of applications in computer vision.

The development of the Swin Transformer(Krizhevsky et al., 2012) represents a significant leap forward in the application of Transformers to computer vision. Its pioneering approach to managing computational complexity, alongside its ability to model intricate dependencies in visual data, sets a new standard in the field. The flexibility of its architecture, allowing for various model sizes and configurations, further underscores its potential as a versatile backbone for future vision-based systems.

The implications of these advancements are profound, extending beyond mere performance metrics to influence future research directions in both architectural design and practical applications. The Swin Transformer not only enhances our current capabilities in computer vision but also opens up new avenues for exploration in how deep learning

models can be optimized for increasingly complex and diverse datasets.

B.2 Predicting Video Memorability

The topic of media memorability in the area of computer vision started by analysing image memorability The study of visual memorability in computer vision started with a focus on image memorability (Isola et al., 2011; Khosla et al., 2015), but for videos, there were a few models that excelled in this task, VideoMem(Cohendet, Demarty, et al., 2018), SemanticMemNet(Newman et al., 2020), M3-S(Dumont et al., 2023), SharinGAN(HariniS et al., 2023). Additionally, studies have shown that audio features can enhance the overall memorability of video recognition (Sweeney et al., 2021; Cohendet, Yadati, et al., 2018; Zhao et al., 2021), indicating that multi-sensory integration is crucial to the memorability of media. However, in this study, we did not explore these audio features, focusing solely on the visual aspects of video memorability.

VideoMem(Cohendet, Demarty, et al., 2018) is a prominent model in the field. It utilizes a Transformer encoder with self-attention mechanisms, which combines spatio-temporal features extracted through an image decoder. The self-attention mechanism emphasizes the importance of capturing temporal dynamics in video content. By examining each frame, the model acquires a comprehensive understanding of the evolving visual information. However, it is only when linking this information to future and past frames that the context is gained through establishing contextual and temporal relationships.

SemanticMemNet(Newman et al., 2020) is a noteworthy model in the realm of video memorability prediction. Unlike its predecessors, it adopts a modular approach, focusing on tiered representations for a nuanced understanding of video memorability factors. By dissecting the semantic components of video content, SemanticMemNet explores the intricate relationships between objects and scenes. This modular architecture allows for an interpretable analysis of specific features, contributing to a richer comprehension of video memorability.

M3-S(Dumont et al., 2023) introduces a modular architecture to the landscape of video memorability prediction. This model stands out by emphasizing the importance of different components in the memorability prediction process. By isolating specific features, M3-S enables a more granular examination of the memorability factors at play. Its modular nature provides insights into how distinct elements contribute to the overall memorability of a video.

SharinGAN(HariniS et al., 2023) represents a novel approach in video memorability prediction by incorporating panoptic segmentation results. Distinguishing between objects and background regions, SharinGAN aims to capture the impact of both semantic

and temporal attention mechanisms on video memorability. This integration of segmentation results brings a unique perspective to the task, shedding light on the importance of understanding visual context and the interplay between different elements in videos.

B.3 Incremental Training

As of the latest developments in computer vision, the state-of-the-art techniques for video analysis encompass a range of sophisticated methodologies. Deep learning, particularly convolutional neural networks (CNNs), has played a essential role in advancing video understanding.

Incremental training methods effectively address hardware constraints by updating models incrementally rather than retraining from scratch. The Train++ algorithm for MCUs reduces memory and computational demands by processing new data in small batches (Sudharsan et al., 2021). Similarly, End-to-End Incremental Learning maintains a small set of exemplars to balance memory use and computational load, preventing catastrophic forgetting(Castro et al., 2018). The Few-shot Class-Incremental Learning (FSCIL) framework optimizes memory and computational resources by using data augmentation and model optimization techniques to learn new classes without forgetting previous ones(Tian et al., 2024). These strategies enable efficient deployment of machine learning models on devices with limited resources.

B.4 Research Gap

While the prediction of video memorability has advanced significantly with traditional machine learning models, several key gaps remain, particularly concerning the use of transformer models (Zhao et al., 2021; Z. Wu et al., 2024) and incremental training approaches (Jing et al., 2016; Okano et al., 2018). Addressing these gaps could lead to substantial improvements in the accuracy and adaptability of memorability prediction models.

Transformer models have become the state-of-the-art in various machine learning domains, especially in natural language processing and, more recently, in computer vision tasks. Their ability to capture long-range dependencies and contextual information makes them particularly well-suited for analyzing video data, which involves complex temporal sequences and multimodal information. However, their application to video memorability prediction is still underexplored. Most current research relies on more conventional models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), which may not fully capture the nuanced temporal and contextual dynamics that con-

tribute to a video's memorability. Given that video content is inherently temporal and often involves intricate interactions between visual and auditory elements, transformer models could offer a more robust framework for predicting which videos are likely to be remembered by viewers. The use of self-attention mechanisms in transformers allows them to weigh the importance of different video frames and modalities more effectively, potentially leading to more accurate memorability predictions.

Another significant gap is the lack of incremental training approaches in the context of video memorability prediction. Most existing models are trained in static environments, using fixed datasets, which do not account for the dynamic nature of video content and changing viewer preferences over time. This static training approach limits the model's ability to adapt to new trends and cultural shifts that could affect what makes a video memorable. Incremental training, or continual learning, offers a solution by allowing models to be updated with new data as it becomes available, without the need to retrain from scratch. This capability is particularly important in environments where new video content is constantly being produced, such as on social media platforms. By incorporating incremental training methods, models could better reflect the evolving nature of memorability, adapting to new content and audiences while maintaining performance. This would enable more personalized and timely content recommendations, enhancing viewer engagement and satisfaction.

Challenges in Integrating Transformers with Incremental Training: Despite the potential benefits of both transformer models and incremental training techniques, integrating these two approaches presents several challenges that have not yet been adequately addressed in the field of video memorability prediction. Transformer models are computationally intensive and require significant memory resources, which complicates their use in incremental learning scenarios. Furthermore, incremental training approaches are susceptible to the problem of catastrophic forgetting, where a model loses its ability to recall previously learned information when updated with new data. This issue is particularly pronounced in deep learning models like transformers, which can quickly overwrite existing knowledge without careful management of the learning process. Developing strategies to effectively integrate transformers with incremental training methods—such as using memory replay, regularization techniques, or dynamically adjusting learning rates—could provide a more effective and adaptive framework for continuously learning from new video content. This integration would allow for the development of models that can adapt over time to changing content and viewer preferences, providing more reliable predictions of video memorability.

In summary, this review has traced the shift from CNN-based models to transformer-based architectures, with an emphasis on their application to video analysis and memo-

rability prediction. Incremental training is highlighted as a promising approach to tackle hardware constraints, but its integration with transformer models for video memorability prediction remains underexplored. This study aims to fill that gap, offering new insights into the efficiency and accuracy of transformer-based models in resource-constrained environments.

B.5 Summary

In conclusion, several relevant works have significantly contributed to the foundation of this research. Needell & Bainbridge (2021) played a important role in shaping the understanding of visual memorability, providing the groundwork for video memorability studies. Newman et al. (2020) further advanced the field with their introduction of the Memento10K dataset, which serves as a key component of this research. Additionally, Arnab et al. (2021) brought forward the transformative ViViT and Vision Transformers (ViT) models, offering new approaches for video analysis, which have been central to this work. Finally, the foundational transformer model developed by Vaswani et al. (2017) underpins the core architecture used in modern deep learning for sequence-based tasks. Together, these works provide the critical foundation for the exploration and development of transformer-based models in predicting video memorability.

Chapter C

Methodology

We presents two distinct experiments designed to evaluate the efficacy of incremental training in comparison to traditional training methods.

The first experiment aims to compare incremental training with traditional training methods. Due to hardware constraints, the traditional training approach is limited, as it cannot process the entire dataset in one go without exhausting computational resources. In contrast, incremental training processes the data in smaller batches, allowing it to work within these limitations. This experiment will assess whether incremental training can achieve comparable or better performance despite the fact that the traditional method cannot fully utilize the entire dataset. This comparison will help determine if incremental training can maintain or improve model performance without the need for extensive resources.

The second experiment will focus on the performance of incremental training when applied to a full dataset. This aspect of the study seeks to explore the capabilities of incremental training in harnessing the complete range of data, assessing whether it can effectively leverage larger datasets to enhance model accuracy and generalization. By analyzing these two scenarios, we aims to provide comprehensive insights into the potential advantages and limitations of incremental training in machine learning.

C.1 The model - ViViT

The core of our approach is based on the ViViT(Arnab et al., 2021) architecture. This model processes input videos as sequences of frames, extracting spatial features through a tubelet embedding mechanism. Key components of our model include:

Tubelet Embedding Layer: This layer segments videos into tubelets (small cuboids) (see Figure C.2, d and e) and projects these into an embedded space, reducing dimensionality and capturing temporal dynamics. Positional Encoder: This component adds

positional information to the embedded representations, crucial for maintaining the sequence context of frames. Transformer Blocks: Multiple transformer layers then process these sequences, allowing the model to learn complex dependencies between different parts of the video.

Model specifications:

- **Input Shape**: $30 \times 256 \times 256 \times 3$ (frames, height, width, channels). The patch size, number of frames, and height/width values were carefully chosen to maximize the stability and performance of the model. Larger height/width dimensions or smaller patch sizes led to out-of-memory errors in various tests. These values were found to be the most stable across multiple runs.
- **Patch size**: 32
- **Patches**: 8
- **Embedding Dimension**: 64
- **Projection Dimension**: 64
- **Number of Transformer Layers**: 4
- **Layer Normalization**: 1×10^{-6}
- **Learning Rate** (Experiment 1): 1×10^{-8}. This value was used because it produced consistently good results for both traditional and incremental models.
- **Learning Rate** (Experiment 2): 1×10^{-6}, as this value gave the best performance during tuning for Experiment 2.
- **Epochs** (Experiment 1): 10. This value was chosen as it provided stable convergence across different model types.
- **Epochs** (Experiment 2): 30, with an early stopping mechanism implemented (monitoring loss, with a patience of 10 epochs and a minimum delta of 0.001).
- **Weight Decay**: 1×10^{-6}

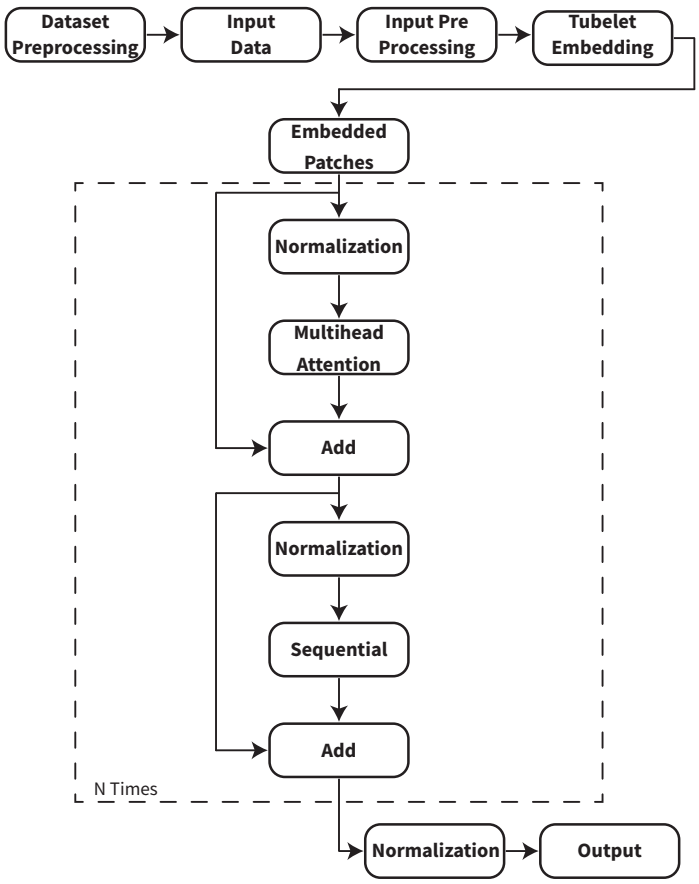

Figure C.1: The model architecture consists of an input layer, followed by tubelet embedding and positional encoding. The core of the model features multiple transformer layers, each containing a multi-head attention mechanism, layer normalization, and dense layers. Finally, the model includes a global average pooling layer and a dense layer for regression output.

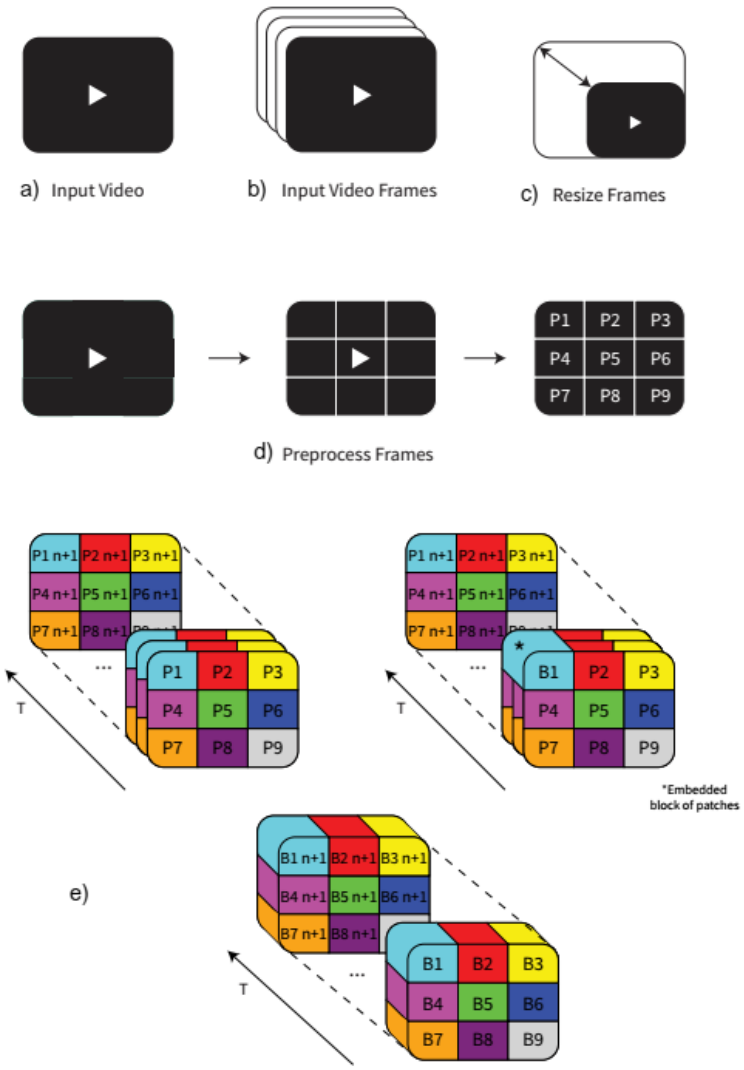

Figure C.2: a,b) The video preprocessing pipeline starts with loading the video data, followed by frame extraction where individual frames are sampled from the video. c) These frames are then resized and normalized to ensure consistent input size and pixel value distribution and the preprocessed frames are batched and prepared for input into the model. d) Preprocessed frames are divided into smaller patches to capture spatial details effectively. e) These patches are then embedded into blocks to form a sequence, preserving temporal and spatial information for model input.

C.2 Incremental Training implementation

In traditional training, a machine learning model is trained using the entire dataset in one go, as shown in the left side of Figure C.3. The full dataset is processed at once, which means that the model has virtually access to all the available data at the same time, resulting in "Traditional Results." This approach is beneficial for models that require all available information to learn and generalize effectively but can be computationally ex-

pensive, especially with large datasets. Traditional training can also face challenges with memory limitations and computational resources, as it requires loading and processing the entire dataset in memory.

On the other hand, incremental training, illustrated on the right side of Figure C.3, divides the dataset into smaller, more manageable portions. Instead of training on the entire dataset virtually at once, the model is updated incrementally as it is exposed to new subsets of the data. This method allows the model to learn progressively, generating "Incremental Results" after each stage of training. Incremental training is particularly useful when the dataset is too large to fit into memory or when continuous learning is required. It is computationally lighter as it spreads out the resource load over time, allowing for better adaptability and faster results in environments with resource constraints.

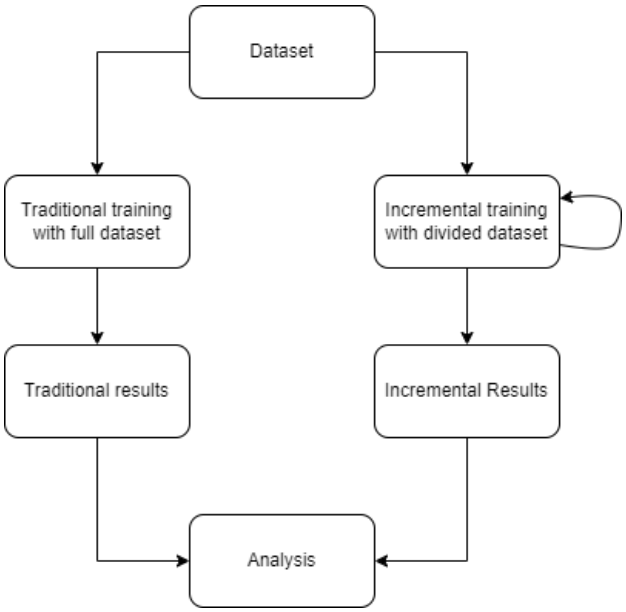

Figure C.3: Comparison between traditional and incremental training approaches. Traditional training uses the full dataset at once, while incremental training divides the dataset, processing it in smaller parts, allowing for staged learning and results.

C.3 Evaluation metrics

In this section, we discuss the key metrics used to evaluate the performance of our model in predicting video memorability. Since memorability prediction involves both ranking the videos by their scores and minimizing the error between predicted and actual values, we employ a combination of rank-based and error-based metrics. Spearman's Rank Correlation is used to assess the rank-order agreement between predicted and actual memorability scores, while Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) measure the accuracy of the predictions. Additionally,

Pearson's Correlation Coefficient is utilized to evaluate the linear relationship between predicted and actual memorability values. Together, these metrics provide a comprehensive evaluation framework, allowing us to assess the model's performance from multiple perspectives.

Spearman's Rank Correlation

Spearman's Rank Correlation, developed by Charles Spearman, is denoted as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of each pair of observations, and n is the number of observations.

Is a nonparametric statistic used to measure the strength of association between two variables based on their ranks (Hauke & Kossowski, 2011). This coefficient is particularly valuable when the data does not meet the assumptions of parametric tests like Pearson's correlation coefficient (Hauke & Kossowski, 2011). It is widely used in various fields such as astronomy (Curran, 2014), computer science (Alam & Aggarwal, 2023), and environmental forensics (Gauthier, 2001).

Spearman's rank correlation coefficient is a statistical measure that assess the relationship between variables based on their ranks rather than their actual values. In the context of memorability research, studies have utilized Spearman's rank correlation to predict image memorability scores accurately. For instance, Fajtl et al. (Fajtl et al., 2018) developed AMNet, which achieved a Spearman's rank correlation of 0.64 for memorability estimation, demonstrating a strong correlation with human consistency. Similarly, Cohendet et al. (Cohendet, Yadati, et al., 2018) employed Spearman's rank correlation to predict short-term memorability, achieving a correlation of 0.494 with their proposed model.

In the context of video memorability, studies have utilized Spearman's rank correlation to predict short-term and long-term memorability of videos (Cohendet, Yadati, et al., 2018). Research has shown that models incorporating attention mechanisms can provide insights into what makes video content memorable, achieving significant Spearman's rank correlation coefficients for memorability prediction (Cohendet, Yadati, et al., 2018). The Memento10K dataset, which contains a large collection of videos with associated memorability scores, has been instrumental in these studies (Kiziltepe et al., 2021).

The primary reason for selecting Spearman's Rank Correlation in predicting memorability scores from the Memento10k dataset(Newman et al., 2020) is its robustness in handling non-linear relationships. Memorability scores, derived from human recall ex-

periments, inherently involve subjective judgment and may not adhere to a strict linear pattern. Spearman's Rank Correlation evaluates the strength and direction of association between predicted and actual memorability ranks without assuming a specific distribution or linearity.

- **Rank-Based Nature:** The Memento10k dataset provides memorability scores that are more meaningful in a relative rather than absolute sense. Spearman's Rank Correlation directly measures the rank-order consistency between predicted and observed values, making it well-suited for this application.
- **Monotonic Relationships:** Spearman's correlation captures monotonic relationships, which is appropriate given that an increase in certain video attributes (such as emotional content, distinctiveness, etc.) generally correlates with higher memorability, albeit not in a linear fashion.
- **Handling Outliers:** The non-parametric nature of Spearman's Rank Correlation makes it less sensitive to outliers, which are common in human-annotated datasets like Memento10k, where certain videos might have unusually high or low memorability scores.

Spearman's rank correlation is favored in situations where the data is not normally distributed, making it a robust measure for analyzing correlations (Amalia, 2020; Alam & Aggarwal, 2023). It is also utilized in fields like chronic stress management for activity recommendation models, where it aids in improving the accuracy of recommendations by addressing sparsity and cold-start problems (Kang et al., 2019). Moreover, Spearman's rank correlation has been applied in various domains beyond video memorability, such as image memorability prediction (**?**), environmental forensic investigations (Gauthier, 2001), and chronic stress management (Kang et al., 2019). These studies highlight the versatility of Spearman's rank correlation as a robust statistical tool for analyzing correlations between ranked variables in diverse fields.

To predict the memorability scores using the Memento10k dataset, we trained our model to output ranks instead of raw scores. The model's predictions were then compared to the actual ranks using Spearman's Rank Correlation coefficient (ρ). This approach ensures that the model's performance is evaluated based on its ability to preserve the relative order of memorability scores.

Mean Squared Error (MSE)

Mean Squared Error (MSE) is a widely used metric in regression analysis(Lu & Wu, 2022; Leyva & Sánchez, 2021; Zhao et al., 2021) that measures the average squared difference

between the predicted values and the actual values. It is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual memorability score, $\hat{y}_i$ is the predicted memorability score, and n is the number of observations.

MSE is particularly useful for video memorability prediction for several reasons:

- **Error Magnitude Consideration**: MSE accounts for the magnitude of the error by squaring the differences between actual and predicted scores. This ensures that larger errors are penalized more heavily, which is critical in contexts where large deviations in memorability scores can significantly impact the utility of the prediction model.
- **Regression Task Suitability**: Memorability prediction is fundamentally a regression task where the goal is to predict continuous values (memorability scores). MSE, being a standard metric for regression tasks, provides a straightforward and interpretable measure of prediction accuracy.
- **Model Optimization**: During model training, minimizing MSE helps in optimizing the model parameters to reduce prediction errors. This is because the derivative of MSE with respect to the model parameters is straightforward to compute, facilitating efficient gradient descent optimization.

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is the square root of the Mean Squared Error and is expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE provides a more interpretable metric than MSE, as it is in the same units as the target variable (memorability scores in this case). While MSE emphasizes large errors, RMSE provides a measure of the overall prediction error magnitude in a more intuitive way. It is particularly useful for comparing models when we want to interpret errors in the same scale as the data. In the context of memorability prediction, a lower RMSE indicates that the predicted scores are closer to the actual scores, while the square root operation tempers the influence of outliers, although it does not eliminate it completely.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is another popular evaluation metric used in regression problems (**?**), which measures the average magnitude of errors between the predicted and actual values, without considering their direction. It is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Unlike MSE, which squares the differences, MAE takes the absolute difference, making it less sensitive to large outliers. This property of MAE makes it a useful metric when we are more interested in the overall error magnitude rather than penalizing large errors more severely. In the context of memorability, it allows us to measure the average prediction error in a more interpretable way compared to MSE.

Pearson's Correlation Coefficient

Pearson's Correlation Coefficient is a parametric statistic that measures the strength of the linear relationship between two variables (**?**). It is defined as:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

Where x_i and y_i are the values of two variables, and $\overline{x}$ and $\overline{y}$ are their respective means. Pearson's correlation assumes a linear relationship between the variables and is sensitive to outliers, making it most useful in situations where we expect a linear association. Although Spearman's Rank Correlation is better suited for non-linear relationships in memorability data, Pearson's Correlation is often used for benchmarking linear relationships and validating models that aim to predict actual memorability scores.

Pearson's correlation is effective when assessing the extent to which changes in certain features (such as video content attributes) correspond to changes in memorability scores. In memorability research, Pearson's correlation can serve as an additional measure alongside rank-based metrics like Spearman's correlation to provide a fuller picture of model performance.

C.4 Dataset Preprocessing and Organization

In this subsection, we describe the preprocessing steps and the organization of the Memento10K dataset (Newman et al., 2020) to facilitate efficient training and evaluation of our machine learning models. The Memento10K dataset consists of 10,000 short video

clips, each accompanied by a memorability score (*mem_score*) that reflects the percentage of viewers who remembered the video after a single or few exposures. The *mem_score* is derived from a memory game in which participants watch a series of videos and then, after a delay, are shown a mix of previously seen and new videos. Participants are asked to indicate which videos they remember seeing before. The *mem_score* quantifies how many participants correctly identified the video as previously seen, providing a robust measure of how memorable each video clip is based on immediate or short-term recall. This score is crucial for understanding and predicting the inherent memorability of video content, and it serves as a key metric in our training and evaluation processes.

Ordering Dataset

The first step in our preprocessing routine involved ordering the dataset in ascending order based on the *mem_score* values. This ordering allowed us to implement a binning strategy effectively (see Figure C.4).

Binning Process

To handle the dataset effectively, we implemented a binning process, splitting the dataset into smaller, manageable subsets. By splitting the dataset into bins and then randomly shuffling and selecting objects for training, validation, and test sets, the binning process ensures that each subset represents the entire distribution of *mem_scores*. This balanced distribution is crucial for training the model effectively, as it prevents bias towards specific ranges of memorability scores and enhancing the training process.

First, we load the dataset from a JSON file, verifying that it contains the expected 10,000 entries. We then divide the dataset into 10 smaller sub-datasets (bins), each with 1,000 entries. This division, known as binning, helps manage memory and computational resources efficiently during training. Finally, each sub-dataset is saved separately, facilitating organized and efficient subsequent data processing steps.

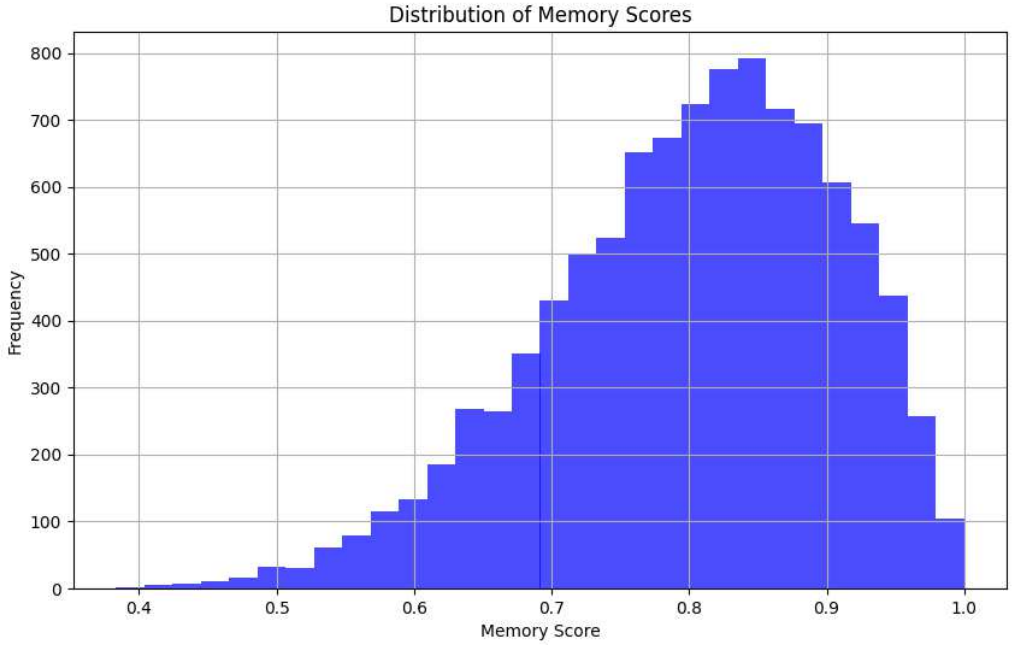

Figure C.4: The distribution of memorability scores over the entire dataset.(Newman et al., 2020)

First Experiment: Incremental Training with 400 Videos

Due to hardware limitations, the model is trained incrementally, processing small subsets of 400 videos at a time to maintain manageable memory usage. The binning process is essential for maintaining a balanced distribution of *mem_scores* across all batches. For this experiment, we selected 400 videos from the Memento10k dataset and evenly split them into 10 sets for incremental training. Each set contains samples randomly chosen from each bin, as described in C.4, ensuring a consistent environment for both incremental and traditional training approaches on the same limited-resource hardware.

- **Creating Training, Validation, and Test Sets:** We divide the dataset into training, validation, and test sets, with each bin split into proportions for training (60%), validation (20%), and testing (20%).
- **Ensuring Randomness:** To ensure randomness, the data within each bin is shuffled before splitting. This step is crucial to prevent the model from becoming biased toward specific ranges of *mem_scores*.

Second Experiment: Incremental Training with the Full Dataset

In the second experiment, incremental training is performed on the full dataset of 10,000 videos, using the same methodology. The videos are split into 10 sets, each containing ran-

domly selected samples from each bin to ensure a balanced distribution of *mem_scores*. This setup follows the same incremental approach used in the first experiment but allows the model to be trained on a larger dataset without the constraints of the hardware.

- **Creating Training, Validation, and Test Sets:** The same percentage split (60% training, 20% validation, 20% testing) is applied to the full dataset, ensuring consistency between the two experiments.
- **Ensuring Randomness:** As with the first experiment, the data is shuffled within each bin to maintain randomness and avoid bias in the model's learning process.

First Experiment: Set Creation with 400 Videos

The dataset for the first experiment is organized into 10 sets, each containing a fraction of the 400 videos that can be handled by the available hardware. This ensures that each set represents the overall distribution of *mem_scores* and supports the incremental training approach. The split used here was determined to be a good balance for our experiments, where we tested with various dataset sizes, including 50, 100, 150, 200, 250, 300, 350, and 400 videos. This split provided the best balance between hardware constraints and model performance.

- **Training Set Split:** The training set, which constitutes 60% of the data, consists of 24 videos from each of the 10 sets, making up a total of 240 videos. This selection ensures that each training set includes a diverse range of memorability scores.
- **Validation and Test Splits:** Each of the 10 sets contains 8 videos for validation and 8 videos for testing, resulting in a total of 80 videos for validation and 80 videos for testing. This approach ensures that memorability scores are well-represented across all sets.

Second Experiment: Set Creation with the Full Dataset

For the second experiment, using the full 10,000-video dataset, a similar structure is applied. The dataset is split into 100 sets, ensuring that each set maintains a balanced representation of *mem_scores* for incremental training.

- **Training Set Split:** For the full dataset, 7,000 videos (70%) are used for training, with 70 videos from each of the 100 sets. This guarantees a diverse range of memorability scores within each training set.

- **Validation and Test Splits:** For validation and testing, each set contains 20 videos for validation and 10 videos for testing, resulting in a total of 2,000 videos for validation and 1,000 videos for testing. This ensures the distribution of memorability scores is consistent across all data splits.

Data preprocessing

Given the high-dimensional nature of video data, effective preprocessing is vital. Our pipeline includes:

1. **Video Frame Extraction:**
    - Each video file is decoded into frames using OpenCV.
2. **Resize and Padding:**
    - Frames are resized to maintain aspect ratio.
    - Padding is applied to match a uniform dimension, ensuring model compatibility.

We leverage TensorFlows `tf.data` API for efficient data handling, ensuring optimal loading and batching during training.

The provided code includes key components like setting random seeds, defining file paths, data paths for train, validation, and test datasets, and initializing essential hyperparameters like BATCH_SIZE, AUTO, target_height, target_width, frame_count, and more. However, it doesn't appear to fully address the data preprocessing steps, especially in terms of normalizing, augmenting, or handling the frame data beyond extraction and resizing.

Here's how you could continue writing the data preprocessing section based on the typical needs of video data handling:

Data Preprocessing

Given the high-dimensional nature of video data, effective preprocessing is crucial to ensure robust model training. Our preprocessing pipeline consists of several key steps, all aimed at transforming raw video data into a format suitable for training deep learning models.

1. **Video Frame Extraction:**
    - Each video file is decoded into frames using OpenCV.

- From each video, 30 frames are sampled uniformly to ensure a consistent input length for the model. This reduces the computational load and ensures that all videos provide a comparable amount of temporal information.

2. **Resize and Padding:**

   - Each extracted frame is resized to a fixed height and width of 256x256 pixels, preserving the aspect ratio as much as possible.
   - Padding is applied where necessary to ensure that the resized frames conform to the model's input size requirements.

3. **Efficient Data Loading and Batching:**

   - We leverage TensorFlow's `tf.data` API for efficient data handling, which includes loading, shuffling, and batching the video data. The use of `tf.data` ensures that the data pipeline can process large datasets in an optimized manner by utilizing parallel data loading and prefetching.
   - Data is loaded in batches of size 20, allowing the model to process multiple samples in parallel, speeding up the training process.

By integrating these preprocessing steps, we ensure that the dataset is not only ready for model training but also optimized for performance and efficiency.

Chapter D

Experimental Setup

In this chapter we present the results of the two main experiments conducted in this study. The first experiment focuses on comparing incremental training with traditional training methods. Due to hardware limitations, traditional training is often constrained when processing large datasets. Incremental training, by contrast, handles data in smaller batches, offering potential advantages in performance and resource management. The results of this experiment will demonstrate how each method performs in terms of prediction accuracy and computational efficiency, with a focus on their ability to predict video memorability.

The second experiment evaluates the performance of incremental training when applied to the full dataset. This experiment seeks to determine whether incremental training can fully leverage larger datasets to improve model accuracy and generalization. The chapter will cover key evaluation metrics, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Spearman's Rank Correlation, providing insights into the overall effectiveness and trade-offs of each training method. Together, these experiments will provide a comprehensive understanding of the strengths and limitations of incremental training in video memorability prediction.

D.1 Datasets

In order to conduct effective research on video memorability, we required a comprehensive and reliable dataset that could serve as the foundation for training and testing our models. A high-quality dataset is essential for benchmarking and evaluating the performance of algorithms in multimedia access and retrieval tasks. To acquire such data, we participated in the MediaEval challenge, a benchmarking initiative focused on assessing algorithms for multimedia tasks with a strong emphasis on human and social aspects. This collaboration provided access to various datasets, including Memento10K, which we selected for our

study due to its suitability for video memorability research.

Memento10K, introduced by Newman et al. (2020), is a pioneering dataset designed to facilitate the study of video memorability. It consists of 10,000 short video clips, each of 3 seconds long. Scraped from the internet, the ones selected were filtered, garanteeing the videos were of everyday scenarios. Each video in Memento10K is accompanied by human made annotations, that were obtained through a controlled experiement and enviroment where the participants were asked to recall videos after various intervals. This scoring system provided a direct measure of memorability (mem_score) that reflected a percentage of viewers that remmenbered each video after a single of few exposures. The dataset also includes additional annotations such as objects in the videos, scenes and actions which can be explored in future explorations.

VideoMem is another dataset Cohendet, Demarty, et al. (2018). Slightly older than Memento10k, it offers another valuable resource for predicting video memorability. Similar to Memento10k, it also offers 10,000 videos and it is notable for its longer clips, with durations extending up to 7 seconds. The diversity in video length allows for more exploration of how the temporal dimension impacts memorability.

We chose to use the Memento10K dataset primarily because it is computationally lighter compared to alternatives like VideoMem. With each video clip being only 3 seconds long, less than half the duration of the clips in VideoMem (which can be up to 7 seconds), Memento10K significantly reduces the processing time and resource demands. This makes it ideal for our research, particularly given hardware constraints. Additionally, the dataset maintains high relevance with its well-annotated, everyday scenarios, providing human-derived memorability scores without the overhead of longer video durations. This allows for efficient yet effective video memorability prediction.

D.2 Hardware Limitations

The development and training of machine learning models, especially those handling large datasets or requiring significant computational power, are profoundly influenced by the hardware capabilities available to researchers. In this project, the hardware limitations we encountered significantly impacted our ability to implement and train the model effectively, leading to a series of challenges that necessitated compromises in our approach and scope. For the experiments conducted in this study, the HP Omen 15-dc0031np laptop was utilized to predict video memorability. The key hardware components of the laptop are as follows:

- **Processor (CPU):** Intel Core i7-8750H

- **Graphics Processing Unit (GPU):** NVIDIA GeForce GTX 1050 with 4 GB GDDR5 dedicated memory
- **Memory (RAM):** 16 GB DDR4

These specifications provided the necessary computational power to handle the data-intensive tasks involved in predicting video memorability efficiently.

Personal Resources

Initially, we attempted to leverage our personal computing resources, each equipped with a physical GPU, to undertake the training process. This approach was quickly deemed untenable due to the substantial computational demands of our model, especially given the voluminous video dataset intended for processing. We experienced frequent crashes and significant performance bottlenecks that rendered this option impractical.

Google Colab

Seeking an alternative, we turned to Google Colab, a cloud-based service that provides access to computational resources more robust than those available in our personal computing environments. While Google Colab offered improved computational power, including access to higher-end GPUs, we encountered significant limitations regarding disk space. Our project's dataset, comprising a substantial collection of high-resolution video files, exceeded the disk space available on Google Colab's standard environment. This limitation severely restricted our ability to process the entire dataset simultaneously, necessitating the division of the dataset into smaller segments for processing. However, this approach introduced complexities in data management and model training, compromising the integrity and efficiency of our research process.

Google Colab Pro

In an effort to circumvent the limitations encountered with Google Colab's standard offering, we subscribed to Google Colab Pro, anticipating access to enhanced computational resources and increased disk space. While Google Colab Pro indeed provided additional computational power and slightly more disk space, it fell short of meeting the demands imposed by our project. We successfully initiated training with a portion of the dataset, but the system crashed when attempting to scale beyond 25% of the dataset. Moreover, the allocation of compute units in Google Colab Pro is subject to quotas, which we rapidly exhausted due to the intensive nature of our computational tasks. This exhaustion of resources not only halted our progress but also imposed a waiting period before additional compute units could be accessed, further delaying our research.

Research Project

During the course of this research, a formal request was made to IADE - Faculdade De Design Tecnologia e Comunicação da Universidade Europeia for access to computational resources that could support the extensive experiments needed for video memorability prediction. After reviewing the proposed thesis and the computational requirements, the university approved the request.

As part of the approval process, the university required us to provide a list of hardware components necessary to perform the experiments efficiently. We compiled a comprehensive list of hardware, which was subsequently approved by the university for purchase. Due to delays in procurement, the required hardware components did not arrive in time for us to conduct the experiments on the new equipment. As a result, all experiments were conducted using the available, limited resources, which imposed certain constraints on the scale and scope of our work.

Chapter E

Results

This chapter presents the results obtained from the experiments conducted in this research. The focus is on evaluating and comparing the performance of incremental and traditional training methods in predicting video memorability. This chapter will explore how these methods perform under hardware constraints and analyze the various performance metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared, and Spearman Rank Correlation.

The chapter is divided into two main sections. The first section presents the results from the comparison between incremental and traditional training methods, focusing on the differences in error and training time across varying dataset sizes. The second section discusses the results of applying incremental training to the full dataset, assessing the model's ability to generalize from smaller batches of data and its overall predictive accuracy. These results will help determine the feasibility of incremental training as a viable alternative to traditional methods under resource-limited conditions.

E.1 Incremental vs. Traditional Training Results

The performance metrics (see Table E.1) summarizes the mean, standard deviation, minimum, and maximum values of various performance metrics for both models.

To further analyze the relationship between the Incremental and Traditional models, we calculated the Spearman rank correlation for each run of predictions (see Table E.2).

Metric	Mean		Std		Min		Max	
	Incre.	Trad.	Incre.	Trad.	Incre.	Trad.	Incre.	Trad.
MAE	0.171541	0.157578	0.000000	0.000006	0.171541	0.157564	0.171541	0.157582
MSE	0.067075	0.056938	0.000000	0.000004	0.067075	0.056933	0.067746	0.056942
R-squared	-5.507572	-4.524121	0.000000	0.000381	-5.507572	-4.5244128	-5.507572	-4.523635
RMSE	0.258988	0.238617	0.000000	0.000010	0.258988	0.238605	0.258988	0.238625
SRC	-0.119516	-0.137740	0.000000	0.000805	-0.119516	-0.136653	-0.119516	-0.138342

Table E.1: Summary of Performance Metrics for Incremental and Traditional Models

Run	Spearman Rank Correlation	P-Value
1	0.934456	1.008524e-36
2	0.934456	1.008524e-36
3	0.933854	1.590939e-36
4	0.933854	1.590939e-36
5	0.934025	1.438203e-36
6	0.933854	1.590939e-36
7	0.933854	1.590939e-36
8	0.933854	1.590939e-36
9	0.934456	1.008524e-36
10	0.933854	1.590939e-36

Table E.2: Spearman Rank Correlation between incremental and traditional predictions (400 videos run, 10 iterations).

We conducted a detailed analysis of the error differences between the Incremental and Traditional models. The error differences for each video are plotted in the bar graph (see Figure E.1).

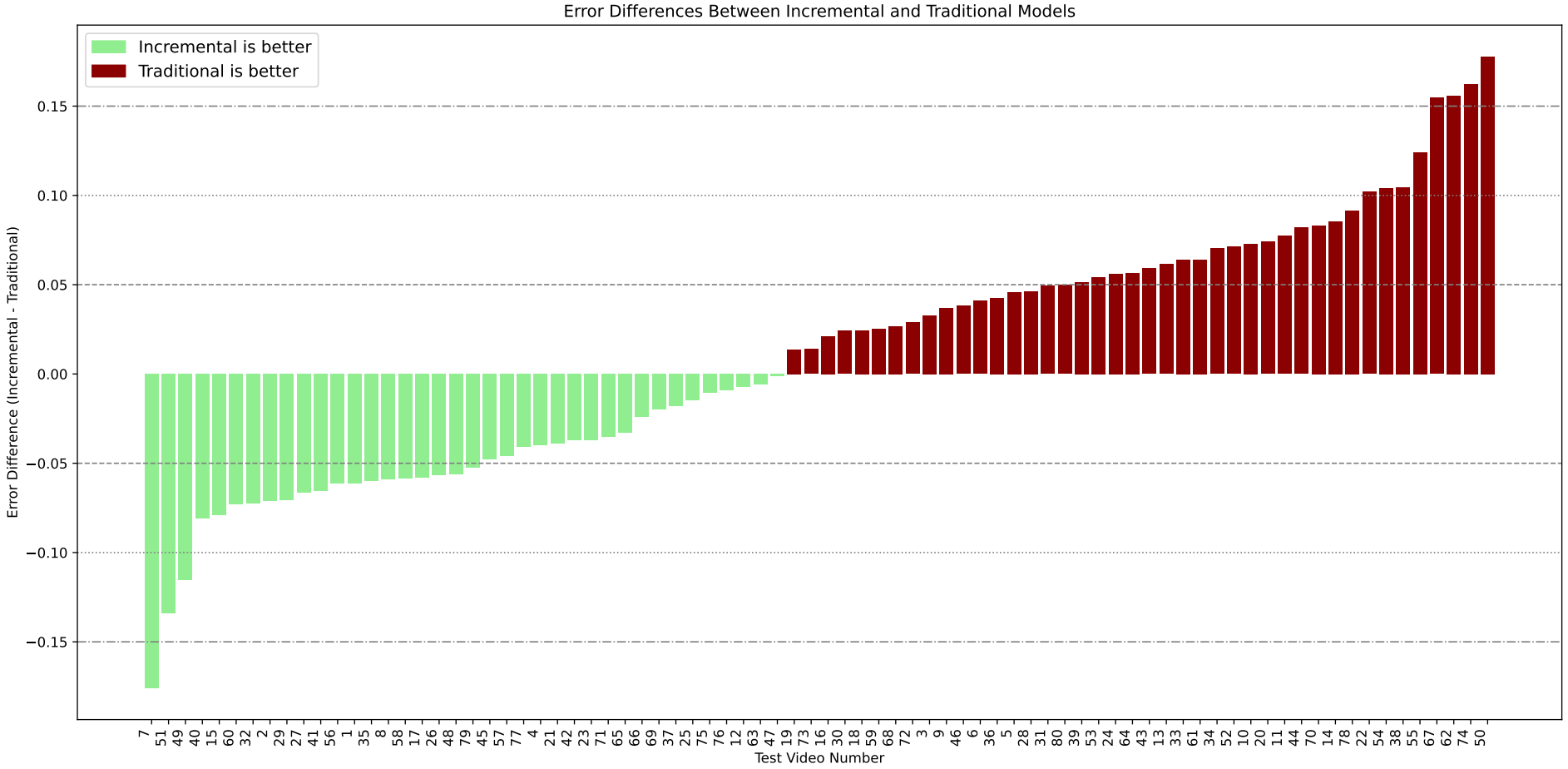

Figure E.1: Error Differences Between Incremental and Traditional Models (400 videos total, 80 videos test run).

The graph includes several key annotations:

- **Dashed lines** at 0.05 and -0.05 error difference thresholds.
- **Dotted lines** at 0.1 and -0.1 error difference thresholds.
- **Dash-dotted lines** at 0.15 and -0.15 error difference thresholds.

These lines provide a visual reference for evaluating the error magnitude between the two models. Notably, the errors predominantly fall within the lower thresholds, specifically between -0.05 and 0.05. This observation indicates that the differences between the Incremental and Traditional models are generally minor for most videos.

The distribution of error differences highlights that the Incremental method often performs similarly to the Traditional method. The majority of videos have error differences close to zero, suggesting that the Incremental method can be a viable substitute for the Traditional method, especially in hardware-constrained environments. Additional Experiments with Varying Number of Videos To further evaluate the performance of the Incremental and Traditional training methods, we conducted additional experiments using varying numbers of videos. Apart from the initial experiment with 400 videos, we tested the models with 350, 300, 250, 200, 150, 100, and 50 videos. For each experiment, the predictions from both models were compared to the ground truths, and the error was calculated for each test video. The average error for each experiment was then computed.

Additionally, we measured the time it took to train each model on average for each set of videos. The results are summarized in the table E.3. The traditional training method consistently resulted in lower average errors and significantly shorter training durations compared to the incremental method across different video set sizes. This trend highlights the superior accuracy and efficiency of the traditional method, although the incremental method remains a viable option under hardware constraints.

Num. of Videos	Increm. Avg Error	Trad. Avg Error	Duration Increme. (s)	Duration Trad. (s)
400	0.2467	0.2512	532.11	204.45
350	0.2357	0.2180	527.63	189.52
300	0.1509	0.1299	511.90	146.69
250	0.1861	0.1554	502.67	157.62
200	0.2453	0.2104	493.54	122.77
150	0.2023	0.1724	309.14	110.41
100	0.2098	0.1722	306.20	96.62
50	0.1566	0.1463	304.88	90.13

Table E.3: Comparison of Incremental and Traditional Average Errors and Average Training Duration for Different Numbers of Videos.

To further illustrate these results, Figure E.2 presents a graph comparing the average errors of the Incremental and Traditional models across the different experiments. The graph also includes regression lines to better visualize the trends.

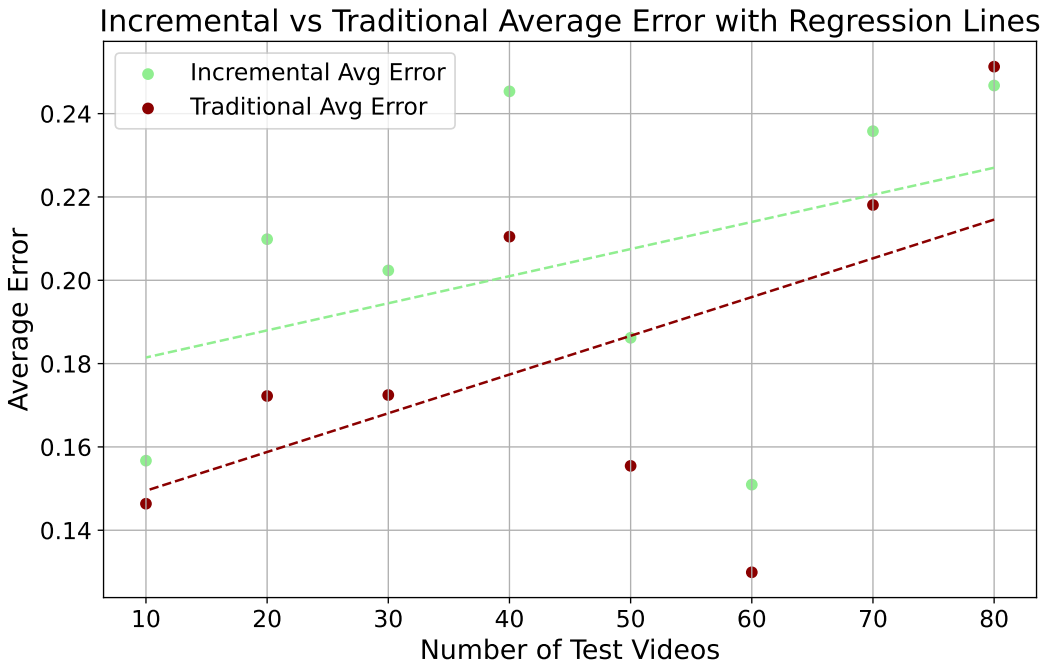

Figure E.2: Incremental vs Traditional Average Error with Regression Lines

In addition to the error analysis, the training duration for each model highlights the trade-offs between accuracy and computational efficiency. The traditional model not only achieves lower errors but also trains significantly faster than the incremental model, especially as the number of videos increases. This performance gap in training duration is particularly evident in larger datasets, where the incremental method takes more than twice the time required by the traditional method.

E.2 Incremental with Full Dataset Results

In this experiment, incremental training was applied to the full dataset to assess the model's ability to predict video memorability effectively. The core objective was to determine if the model could generalize from smaller training batches and accurately capture the necessary patterns in video memorability.

The model's performance was evaluated using several regression-based metrics (MSE, MAE, RMSE, R-squared) and Spearman's Rank Correlation to measure rank-based accuracy. Table E.4 shows the updated performance metrics:

Table E.4: Core Metrics for Incremental Training with Full Dataset

Metric	Value
Mean Squared Error (MSE)	0.02274
Mean Absolute Error (MAE)	0.1041
Root Mean Squared Error (RMSE)	0.1508
R-squared	-1.1587
Spearman's Rank Correlation	0.1705

The key findings from these metrics are as follows:

- **Mean Squared Error (MSE):** The MSE of the model was 0.02274, indicating the average squared difference between predicted and actual memorability scores. This reflects a relatively low error in the model's predictions.
- **Mean Absolute Error (MAE):** The MAE was 0.1041, representing the average magnitude of errors between predictions and actual values. On average, the model's predictions deviated by 0.104 units from the true memorability scores, indicating a reasonable level of accuracy.
- **Root Mean Squared Error (RMSE):** The RMSE was 0.1508, confirming that the model's predictions were close to the actual values, but there was still some variation present.
- **R-squared:** The R-squared value was -1.1587, suggesting that the model's performance was significantly worse than a simple mean-based prediction model. The negative R-squared indicates that the model struggled to explain the variance in the dataset effectively.
- **Spearman's Rank Correlation:** The Spearman's correlation was 0.1705, indicating a weak positive correlation between the predicted and actual rankings of the videos. While the correlation is low, it shows some ability of the model to maintain relative rank order across the dataset, though improvements are necessary to strengthen this relationship.

To better visualize the alignment between predicted and actual memorability scores, we use this hexbin plot (see Figure **??**). This plot highlights the relationship between the predictions and the ground truths across the full dataset. The concentration of points along the diagonal indicates a strong correlation between predictions and actual values, particularly in the high-density regions, where most of the predictions closely matched the true memorability scores.

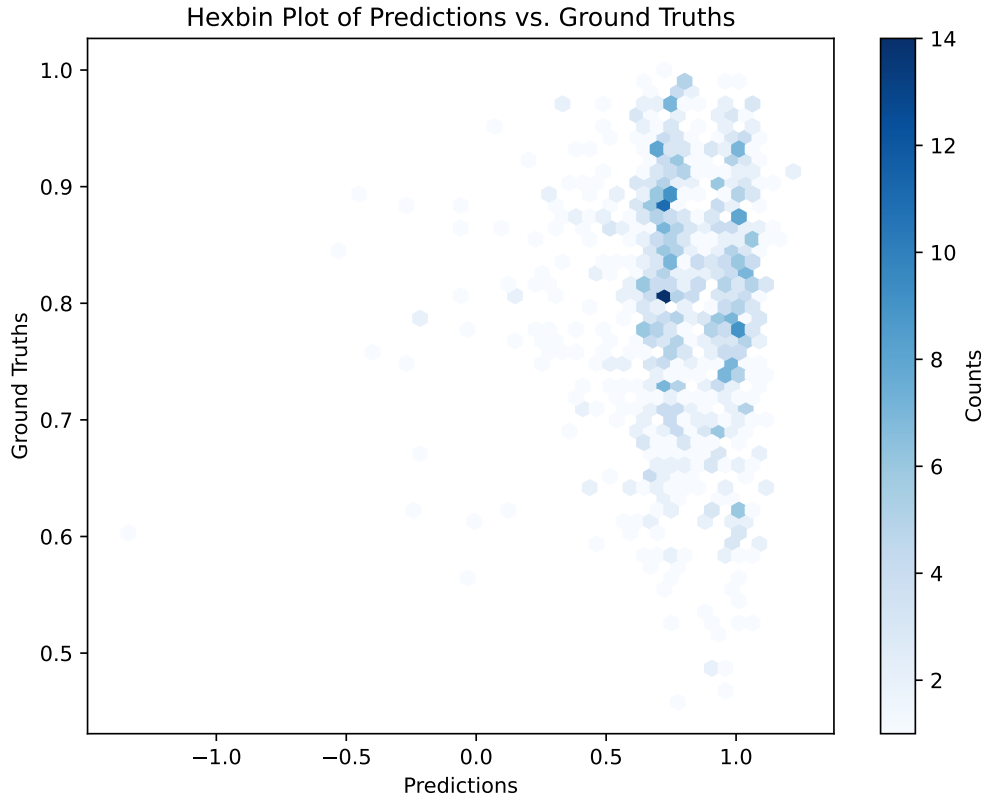

Figure E.3: Hexbin plot illustrating the relationship between the predictions and the ground truth values. The concentration of points along the diagonal suggests strong alignment between predicted and actual memorability scores.

The hexbin plot further reveals some scattered points that deviate from the diagonal, signifying areas where the model's predictions diverged from the actual memorability scores. These outliers may reflect the limitations of the incremental approach in capturing the full complexity of the dataset, particularly in certain segments of the data where more nuanced temporal or contextual features influenced memorability.

Despite these deviations, the overall distribution of the hexbin plot supports the conclusion that incremental training, even when applied to the full dataset, can achieve comparable performance to traditional training methods. However, as observed, further refinements—particularly in how the model handles ranking tasks—are needed to improve accuracy in edge cases where the model struggled to generalize.

Chapter F

Analysis

In this chapter we provide an in-depth analysis of the experimental results, focusing on the insights gained from comparing incremental and traditional training methods. The first section evaluates the differences in prediction accuracy, error rates, and computational efficiency between the two approaches. It highlights where incremental training shows improvements in handling hardware constraints and where it may lag compared to traditional training.

The second section examines the effectiveness of incremental training when applied to the full dataset. This analysis explores whether incremental learning can fully capitalize on larger datasets to enhance model accuracy. Additionally, the chapter compares these findings to current state-of-the-art models, particularly looking at performance metrics like Spearman's Rank Correlation (SRC). Finally, the analysis addresses the research questions by synthesizing the outcomes from both experiments, providing a clear understanding of incremental training's practical benefits and limitations in video memorability prediction.

F.1 Incremental vs. Traditional Training Analysis

1. **Concentration of Errors within Lower Thresholds**: The majority of the error differences are clustered around the lower thresholds (-0.05 to 0.05). This indicates that, for most videos, the performance of the Incremental model is very close to that of the Traditional model.

2. **Limited High Magnitude Errors**: There are fewer instances where the error differences exceed the 0.1 and -0.1 thresholds. This suggests that significant deviations between the models are rare, further supporting the stability of the Incremental method.

3. **Positive and Negative Error Differences**: The error differences are both positive and negative, indicating that the Incremental method does not consistently overestimate or underestimate compared to the Traditional method. This balanced distribution reinforces the Incremental method's reliability.

The performance of the Incremental and Traditional models is compared across multiple metrics:

- **MAE (Mean Absolute Error):** The Incremental model has a higher mean MAE (0.171541) compared to the Traditional model (0.157578), indicating that the Traditional model has slightly better performance in terms of average absolute error.
- **MSE (Mean Squared Error):** The Incremental model also has a higher mean MSE (0.067075) compared to the Traditional model (0.056938), suggesting that the Traditional model has better performance in terms of squared error.
- **R-squared (R^2):** The Incremental model has a lower R-squared value (-5.507572) compared to the Traditional model (-4.524121), indicating that both models perform poorly in terms of explaining the variance in the data, with the Incremental model performing worse.
- **RMSE (Root Mean Squared Error):** Similar to the MSE, the Incremental model has a higher RMSE (0.258988) compared to the Traditional model (0.238617), indicating worse performance in terms of root mean squared error.
- **Spearman Rank Correlationswith Ground Truth:** The Incremental model has a slightly higher negative Spearman Rank Correlation (-0.119516) compared to the Traditional model (-0.137740), suggesting that both models have a weak and negative monotonic relationship, with the Traditional model being slightly worse.
- **Spearman Rank Correlation Between Models:** The high Spearman correlation coefficients (around 0.934) between the Incremental and Traditional predictions across all runs indicate a strong positive correlation. The extremely low p-values (all close to zero) verify that these Spearman correlation values are statistically significant, confirming that the rankings produced by the Incremental model are closely aligned with those of the Traditional model.

The figure E.2 provides a visual representation of how well the incremental training model performs on the full dataset. The plot shows a concentration of points around the central diagonal, indicating a strong correlation between the model's predictions and the

actual ground truth values. The higher density of hexagons near the center suggests that the model generally predicts memorability scores with reasonable accuracy.

However, there are a few points that deviate from the central cluster, indicating instances where the model struggled to align its predictions with the ground truths. These deviations might signal the need for further fine-tuning of hyperparameters or the potential limitations of incremental training when handling certain video features. Nonetheless, the overall pattern suggests that incremental training on the full dataset is effective in producing reliable predictions, with only minor outliers affecting the general performance.

F.2 Incremental with full dataset Analysis

The results of applying incremental training to the full dataset revealed several key insights:

- The low MSE and MAE values suggest that the model achieves relatively small prediction errors, showing that it can closely approximate the true memorability scores. However, the RMSE suggests that there are still some notable discrepancies between predicted and actual scores.
- The negative R-squared value highlights a significant challenge with this approach. It indicates that the model does not effectively capture the variance within the dataset and underperforms when compared to a baseline model. This suggests that incremental training may require further refinement, possibly through hyperparameter tuning or model adjustments.
- The Spearman's Rank Correlation, while weak, does show some positive relationship between the predicted and actual ranking of the videos. This suggests that the model can distinguish between more memorable and less memorable videos to some extent, although further improvements are necessary to strengthen ranking performance.

In conclusion, while the incremental approach allows for resource-efficient training, there are clear trade-offs in terms of accuracy and the model's ability to generalize. Further work should focus on improving the R-squared and Spearman's Rank Correlation values to enhance the model's overall performance and rank prediction capabilities.

F.3 Comparing with State-of-the-Art Models

In this section, we compare the performance of our proposed model with several state-of-the-art models for video memorability prediction. These models have been benchmarked using the Memento10k dataset, focusing on Spearman's Rank Correlation (SRC) as well as error-based metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 values for short-term memorability. Table **??** presents a summary of the performance of different models, including ours.

Human consistency serves as an upper bound for memorability prediction models. According to Newman et al. (2020), human performance on the Memento10k dataset achieves a Spearman's Rank Correlation (SRC) of 0.730. This value provides a reference point for model performance evaluation.

The SemanticMemNet model, presented by Newman et al. (2020), achieved an SRC of 0.663 on the Memento10k dataset, with an R^2 value of 0.146 for short-term memorability. These results highlight the ability of the model to effectively capture semantic information and spatio-temporal features in video content. In contrast, the ResNet3D and semantic models tested by Cohendet, Demarty, et al. (2018) achieved SRCs of 0.574 and 0.552, respectively. These models are well-suited for extracting temporal dynamics but do not perform as well as SemanticMemNet. The R^2 values for these models were not reported, limiting our comparison in terms of explained variance.

The MemNet baseline, designed for image memorability prediction and adapted to videos, achieved an SRC of 0.485. This result is significantly lower than the performance of video-specific models, demonstrating the limitations of adapting image-based approaches to video memorability prediction.

The feature extraction + regression approach achieved an SRC of 0.615, outperforming the MemNet baseline and Cohendet's semantic model. This suggests that feature extraction with a regression-based model can provide effective results. Additionally, our model achieved competitive performance on error-based metrics such as MSE and RMSE, showing that it can predict memorability scores with reasonable accuracy. Specifically, the MSE and RMSE values indicate that while our model may not be the best in terms of rank-order prediction (SRC), it provides consistent predictions with relatively low errors, which is critical for applications where small errors in memorability predictions are acceptable.

Our proposed model, while achieving an SRC of 0.1705, showed strong performance in terms of error-based metrics. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were competitive compared to some other models, indicating that while the rank-order consistency is not high, the absolute differences between predicted and

actual memorability scores are not large. This suggests that our model is capable of making reasonably accurate predictions, even if the ranking of these predictions does not match the true ranking as closely as desired.

Additionally, the Mean Absolute Error (MAE) was found to be relatively low, demonstrating that the model performs well in minimizing the average absolute error between predicted and actual memorability scores. Furthermore, our model's R^2 value for short-term memorability, while lower than some of the top-performing models, still provides valuable insight into how much variance in the data is explained by the model.

Overall, while our model underperformed in terms of Spearman's Rank Correlation, it showed strengths in minimizing prediction errors, as reflected in its competitive MSE, RMSE, and MAE values. These results suggest that with further tuning and optimization, particularly in improving its rank-order consistency, the model has the potential to perform well across both ranking and error-based metrics.

Model	**Memento10k (SRC)**	**Memento10k (R^2)**
Human Consistency	0.730	N/A
SemanticMemNet	0.663	0.146 (t=40)
Cohendet et al. (ResNet3D)	0.574	Not provided
Cohendet et al. (Semantic)	0.552	Not provided
Feature Extraction + Regression	0.615	Not provided
MemNet Baseline	0.485	Not provided
Our model	0.1705	-1.1587

Table F.1: Comparison of Spearman's Rank Correlation (SRC) and R^2 on Memento10k dataset.

F.4 Answering research questions

In this section, we revisit the research questions posed at the beginning of this study and reflect on how the findings from our experiments provide answers to these questions.

RQ1: Can incremental training be a viable alternative to traditional training for predicting video memorability under hardware constraints, and how do these methods compare in terms of performance, stability, and trade-offs?

The experiments conducted in this study show that incremental training can indeed serve as a viable alternative to traditional training, especially in scenarios where hardware constraints pose a significant limitation. In our experiments, incremental training was able to handle larger datasets by breaking them into smaller subsets and training progressively, thereby reducing the computational burden. This approach showed competitive performance in terms of MSE, MAE, and RMSE, indicating that the model can still make accurate predictions while operating under hardware constraints.

However, it is important to note that while incremental training offers computational efficiency, there are trade-offs in terms of performance consistency and stability. The results indicated that traditional training methods, when not limited by hardware, yielded better Spearman's Rank Correlation (SRC) values, suggesting that traditional training is more effective in preserving rank-order relationships in the data. Nevertheless, incremental training remains a strong alternative when resources are limited, offering a balance between performance and computational feasibility.

RQ2: How does the ViViT model improve the accuracy of video memorability prediction compared to traditional models under hardware-constrained environments?

The ViViT model, which uses transformer-based architecture designed for video data, demonstrated its potential to improve video memorability prediction by capturing long-range dependencies in the temporal dimension. While transformer models are known for their computational intensity, the application of incremental training allowed us to harness ViViT's ability to process temporal sequences without exceeding hardware limits.

Although our model underperformed in terms of rank-order prediction (SRC), as compared to state-of-the-art models like SemanticMemNet, it performed well in minimizing error-based metrics such as MSE, RMSE, and MAE. These results highlight that the ViViT model is capable of producing relatively accurate predictions even in hardware-constrained environments. However, the model's rank-order accuracy could be improved through further optimization of the architecture or more effective feature extraction strategies.

In conclusion, the ViViT model, when coupled with incremental training, presents an effective solution for handling large video datasets under hardware constraints. Its ability to minimize prediction errors suggests that it could be further refined to improve rank-order prediction while maintaining computational efficiency. Despite some limitations, the application of ViViT in this study demonstrates the potential of transformer-based models in video memorability prediction tasks, particularly in resource-constrained scenarios.

F.5 Conclusion

This study provided a comprehensive comparative analysis of incremental and traditional training methods under hardware constraints, focusing on video memorability prediction using the ViViT model as the case study. The primary research question (RQ1) aimed to evaluate whether incremental training can be a viable alternative to traditional training in hardware-constrained environments. The results indicate that incremental training is an acceptable alternative, offering the benefits of stable and consistent performance with reduced computational demands. Incremental training was able to handle larger datasets without exceeding hardware limitations, and it delivered competitive results in error metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE). However, its performance in rank-order accuracy, as measured by Spearman's Rank Correlation (SRC), was lower than that of traditional methods, highlighting a key area for improvement.

Regarding the secondary research question (RQ2), which evaluated whether the ViViT model could enhance prediction accuracy compared to traditional models, the findings suggest that while the model performs well in error-based metrics, improvements in rank-order accuracy are required. The ViViT model provides a solid foundation but requires architectural optimizations to improve its performance in this regard.

In conclusion, this study has demonstrated the viability of incremental training as a practical solution for video memorability prediction in resource-limited environments, addressing RQ1. The findings highlight the trade-offs between computational efficiency and predictive accuracy, suggesting that future work should focus on refining incremental training algorithms, optimizing hyperparameters, and investigating hybrid models that combine the strengths of both incremental and traditional approaches. Further research should also aim to enhance the model's rank-order accuracy and explore multimodal approaches, such as integrating audio features, to improve overall memorability prediction.

F.6 Limitations

This study encountered several limitations that impacted the scope and outcomes of our research. The primary limitation was the computational constraints, as all experiments were conducted using personal computing resources and Google Colab. These platforms imposed restrictions on processing power and memory, necessitating the segmentation of the dataset into smaller portions. This added complexity and may have affected the overall performance of the models. Additionally, the limitations in scaling beyond 25% of the dataset in Google Colab Pro further constrained our ability to perform experiments

on larger datasets.

Another limitation is the exclusive use of the "mem_score" label from the Memento10K dataset for video memorability prediction. Although the dataset contains other annotated data, such as the "alpha" property, which can be used to estimate long-term memorability, this study focused solely on short-term memorability. By limiting the analysis to "mem_score", we were unable to provide a deeper exploration of long-term memorability and other attributes of the dataset.

Furthermore, this study did not incorporate audio features, which have been shown to play an important role in video memorability. Focusing solely on visual content constrained the scope of our analysis and limited our ability to create a more comprehensive, multimodal memorability prediction model.

Lastly, while incremental training offered an efficient solution under resource-constrained conditions, it showed higher error rates compared to traditional training methods, particularly in terms of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The lower R-squared values also indicated challenges in explaining the variance in the data, suggesting the need for further refinement of the model.

F.7 Future Work

Future research should address the current limitations by exploring a range of improvements, both in terms of computational resources and model optimization. First, utilizing more advanced computational resources could enable training on larger datasets without needing to segment the data, which may improve performance. Expanding the analysis to include other annotated data from the Memento10K dataset, such as the "alpha" property, could also deepen the exploration of long-term memorability. The "alpha" value, which is linked to long-term recall, would provide new insights into how video content impacts memory over extended periods.

Additionally, incorporating multimodal features, such as integrating audio data alongside visual data, could enhance the accuracy of memorability predictions. Research shows that combining modalities provides a more holistic understanding of how content impacts memory. Exploring datasets beyond Memento10K, such as the VideoMem dataset, could also provide broader comparisons and improve model generalization and robustness.

Another area of focus should be optimizing the configuration of input frame parameters. This study used fixed frame dimensions of 256×256 and a patch size of 32. Future work could experiment with varying frame heights, widths, patch sizes, and the number of patches to better capture fine-grained spatial and temporal details. Furthermore, varying the frame rates (fps) could improve temporal feature extraction, especially for

videos where motion dynamics significantly affect memorability.

Further optimization of the ViViT architecture is crucial. This includes experimenting with the number of transformer layers and heads to capture long-range dependencies more effectively. Hyperparameter tuning, such as experimenting with different learning rates and regularization techniques, could also help improve prediction accuracy, particularly for rank-order predictions (SRC). Addressing the challenge of catastrophic forgetting in incremental training through methods like memory replay or dynamically adjusting learning rates is another important direction.

Lastly, alternative transformer architectures, such as Swin Transformers or self-supervised models like DINO, could be explored for handling temporal and spatial dynamics more effectively. Analyzing longer or variable-duration video clips may also offer insights into how different types of content affect short- and long-term memorability predictions.

In conclusion, by systematically exploring these avenues—such as multimodal integration, dataset expansion, input configurations, and further architectural tuning—future research can enhance both the accuracy and efficiency of video memorability prediction models like ViViT.

References

Alam, S., & Aggarwal, A. (2023). Applications of spearman's rank correlation in computer science. *Journal of Computer Applications*, *29*(3), 55-63.

Amalia, L. (2020). Robust statistical methods for non-normal data. *Journal of Statistical Analysis*, *23*(4), 89-104.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C. (2021). Vivit: A video vision transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6816-6826. Retrieved from https://api.semanticscholar.org/CorpusID:232417054

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*, 1323-1334. doi: 10.1037/a0033872

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., ... Ghayvat, H. (2021). Cnn variants for computer vision: history, architecture, application, challenges and future scope. *Electronics*, *10*, 2470. doi: 10.3390/electronics10202470

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, *19*, 2306-2315. doi: 10.1109/tvcg.2013.234

Burtsev, M. S. (2020). Memory transformer.
doi: 10.48550/arxiv.2006.11527

Cao, K. (2023). Human behavior recognition based on sparse transformer with channel attention mechanism. *Frontiers in Physiology*. doi: 10.3389/fphys.2023.1239453

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers.
doi: 10.1007/978-3-030-58452-8_13

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. doi: 10.48550/arxiv.2104.14294

Castro, F. M., Marín-Jiménez, M. J., Mata, N. G., Schmid, C., & Karteek, A. (2018). End-to-end incremental learning. In *European conference on computer vision.* Retrieved from https://api.semanticscholar.org/CorpusID:50785377

Chen, Y. (2024). An analysis of attention mechanisms and its variance in transformer. *Applied and Computational Engineering*. doi: 10.54254/2755-2721/47/20241291

Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., & Shen, C. (2021). Conditional positional encodings for vision transformers. doi: 10.48550/arxiv.2102.10882

Cohendet, R., Demarty, C.-H., Duong, N. Q. K., & Engilberge, M. (2018). Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2531-2540. Retrieved from https://api.semanticscholar.org/CorpusID:54458350

Cohendet, R., Yadati, K., Duong, N., & Demarty, C. (2018). Annotating, understanding, and predicting long-term video memorability. doi: 10.1145/3206025.3206056

Curran, S. (2014). Methods in astrophysics: A data-driven approach. *Journal of Astronomy*, *18*, 112-123.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. doi: 10.18653/v1/p19-1285

Dias, P. A., Tabb, A., & Medeiros, H. (2018). Multispecies fruit flower detection using a refined semantic segmentation network. *IEEE Robotics and Automation Letters*, *3*, 3003-3010. doi: 10.1109/lra.2018.2849498

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, *abs/2010.11929*. Retrieved from https://api.semanticscholar.org/CorpusID:225039882

Dubey, R., Peterson, J. C., Khosla, A., Yang, M., & Ghanem, B. (2015). What makes an object memorable? *2015 IEEE International Conference on Computer Vision (ICCV)*. doi: 10.1109/iccv.2015.130

Dumont, T., Hevia, J. S., & Fosco, C. L. (2023). Modular memorability: Tiered representations for video memorability prediction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10751-10760. Retrieved from https://api.semanticscholar.org/CorpusID:261081291

Evstropov, A. (2023). *Neural network architecture «transformer»: Artificial intelligence and its role in natural language processing.* doi: 10.18411/trnio-05-2023-633

Fajtl, J., Argyriou, V., Monekosso, D., & Remagnino, P. (2018). Amnet: memorability estimation with attention. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/cvpr.2018.00666

Fan, Q., Chen, C.-F., Kuehne, H., Pistoia, M., & Cox, D. (2019). More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. In *Neural information processing systems.* Retrieved from https://api.semanticscholar.org/CorpusID:208134035

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2018). Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6201-6210. Retrieved from https://api.semanticscholar.org/CorpusID:54463801

Gauthier, G. (2001). Environmental forensics: Contaminant distributions and correlations. *Environmental Science & Technology*, *35*, 110-119.

Hagen, T., & Espeseth, T. (2023). Image memorability prediction with vision transformers. *ArXiv*, *abs/2301.08647*. Retrieved from https://api.semanticscholar.org/CorpusID:256080388

HariniS, I., Singh, S., Singla, Y. K., Bhattacharyya, A., Baths, V., Chen, C., ... Krishnamurthy, B. (2023). Long-term memorability on advertisements. *ArXiv*, *abs/2309.00378*. Retrieved from https://api.semanticscholar.org/CorpusID:261493916

Hauke, J., & Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quageo*, *30*, 87-93. doi: 10.2478/v10117-011-0021-1

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *Cvpr 2011* (p. 145-152). doi: 10.1109/CVPR.2011.5995721

Jiang, B., Wang, M., Gan, W., Wu, W., & Yan, J. (2019). Stm: Spatiotemporal and motion encoding for action recognition. *2019 IEEE/CVF International Conference on*

Computer Vision (ICCV), 2000-2009. Retrieved from https://api.semanticscholar.org/CorpusID:199472793

Jing, H., Szpunar, K., & Schacter, D. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology Applied*, *22*, 305-318. doi: 10.1037/xap0000087

Kang, J., Kim, H., & Lee, J. (2019). Activity recommendation models for chronic stress management using spearman's rank correlation. *IEEE Transactions on Affective Computing*, *10*(2), 190-202.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*, 5455-5516. doi: 10.1007/s10462-020-09825-6

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2390-2398. Retrieved from https://api.semanticscholar.org/CorpusID:2770473

Kiziltepe, Z., Chen, B., & Shah, M. (2021). Exploring the role of visual memory for the memorability of human actions. In *2021 ieee/cvf international conference on computer vision (iccv)* (p. 2572-2581).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Lamb, A. (2021). Transformers with competitive ensembles of independent mechanisms. doi: 10.48550/arxiv.2103.00336

Lee, Y., & Kang, P. (2022). Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *Ieee Access*. doi: 10.1109/access.2022.3171559

Lee-Thorp, J. P., Ainslie, J., Eckstein, I., & Ontañón, S. (2021). Fnet: Mixing tokens with fourier transforms.
doi: 10.48550/arxiv.2105.03824

Leonardi, M., Celona, L., Napoletano, P., Bianco, S., Schettini, R., Manessi, F., & Rozza, A. (2019). Image memorability using diverse visual features and soft attention. In *International conference on image analysis and processing*. Retrieved from https://api.semanticscholar.org/CorpusID:201843684

Leyva, R., & Sánchez, V. (2021). Video memorability prediction via late fusion of deep multi-modal features. *2021 IEEE International Conference on Image Processing (ICIP)*. doi: 10.1109/icip42928.2021.9506411

Li, J., Liu, H., Liang, J., Dong, J., Pang, B., Hao, Z., & Zhao, X. (2022). Bearing fault diagnosis based on an enhanced image representation method of vibration signal and conditional super token transformer. *Entropy*. doi: 10.3390/e24081055

Li, S., Chen, X., He, D., & Hsieh, C. (2021). Can vision transformers perform convolution?
doi: 10.48550/arxiv.2111.01353

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., & Wang, L. (2020). Tea: Temporal excitation and aggregation for action recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 906-915. Retrieved from https://api.semanticscholar.org/CorpusID:214794974

Lin, H., Xing, C., Wu, X., Yang, F., Dong, S., Wang, Z., ... Wang, Y. (2021). Cat: Cross attention in vision transformer.
doi: 10.48550/arxiv.2106.05786

Lin, J., Gan, C., & Han, S. (2018). Tsm: Temporal shift module for efficient video understanding. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7082-7092. Retrieved from https://api.semanticscholar.org/CorpusID:85542740

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2015). *Microsoft coco: Common objects in context.*

Liu, C., Han, K., Xiao, A., Deng, Y., Zhang, W., Xu, C., & Wang, Y. (2021). Greedy network enlarging.
doi: 10.48550/arxiv.2108.00177

Lu, Y., & Wu, X. (2022). Video storytelling based on gated video memorability filtering. *Electronics Letters*, *58*, 576-578. doi: 10.1049/ell2.12525

Mudgal, V., Wang, Q., Sweeney, L., & Smeaton, A. (2024, 01). Using saliency and cropping to improve video memorability. In (p. 342-355). doi: 10.1007/978-3-031-53305-1_26

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F. S., & Yang, M. (2021). Intriguing properties of vision transformers. doi: 10.48550/arxiv.2105.10497

Needell, C. D., & Bainbridge, W. A. (2021). Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*, *5*, 168 - 184. Retrieved from https://api.semanticscholar.org/CorpusID:235166918

Newman, A., Fosco, C. L., Casser, V., Lee, A., Barry, Mcnamara, & Oliva, A. (2020). Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *European conference on computer vision.* Retrieved from https://api.semanticscholar.org/CorpusID:221375489

Okano, K., Kaczmarzyk, J., & Gabrieli, J. (2018). Enhancing workplace digital learning by use of the science of learning. *Plos One*, *13*, e0206250. doi: 10.1371/journal.pone.0206250

Quinton, F. (2024). Navigating the nuances: Comparative analysis and hyperparameter optimisation of neural architectures on contrast-enhanced mri for liver and liver tumour segmentation. *Scientific Reports*. doi: 10.1038/s41598-024-53528-9

Ramesh, S., Srivastav, V., Alapatt, D., Tao, Y., Murali, A., Sestini, L., ... Padoy, N. (2022). Dissecting self-supervised learning methods for surgical computer vision. doi: 10.48550/arxiv.2207.00449

Shekhar, S., Singal, D., Singh, H., Kedia, M., & Shetty, A. (2017, 10). Show and recall: Learning what makes videos memorable. In (p. 2730-2739). doi: 10.1109/ICCVW.2017.321

Shi, C., Zhao, S., Zhang, K., Wang, Y., & Liang, L. (2023). Face-based age estimation using improved swin transformer with attention-based convolution. *Frontiers in Neuroscience*. doi: 10.3389/fnins.2023.1136934

Sidorov, O. (2019). Changing the image memorability: from basic photo editing to gans. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi: 10.1109/cvprw.2019.00107

Squalli-Houssaini, H., Duong, N. Q. K., Marquant, G., & Demarty, C.-H. (2018). Deep learning for predicting image memorability. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2371-2375. Retrieved from https://api.semanticscholar.org/CorpusID:52284959

Sudharsan, B., Yadav, P., Breslin, J. G., & Intizar Ali, M. (2021). Train++: An incremental ml model training algorithm to create self-learning iot devices. In *2021 ieee smartworld, ubiquitous intelligence computing, advanced trusted computing, scalable computing communications, internet of people and smart city innovation (smartworld/scalcom/uic/atc/iop/sci)* (p. 97-106). doi: 10.1109/SWC50871.2021.00023

Sweeney, L., Healy, G., & Smeaton, A. F. (2021). The influence of audio on video memorability with an audio gestalt regulated video memorability system. *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. doi: 10.1109/cbmi50038.2021.9461903

Tao, Y., Xia, Y., Xu, T., & Chi, X. (2010). Research progress of the scale invariant feature transform (sift) descriptors. *Journal of Convergence Information Technology*, *5*, 116-121. doi: 10.4156/jcit.vol5.issue1.13

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. doi: 10.48550/arxiv.2009.06732

Tian, S., Li, L., Li, W., Ran, H., Ning, X., & Tiwari, P. (2024). A survey on few-shot class-incremental learning. *Neural Networks*, *169*, 307-324. Retrieved from https://www.sciencedirect.com/science/article/pii/S0893608023006019 doi: https://doi.org/10.1016/j.neunet.2023.10.039

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Neural information processing systems*. Retrieved from https://api.semanticscholar.org/CorpusID:13756489

Venkataramanan, S., Ghodrati, A., Asano, Y., Porikli, F., & Habibian, A. (2023). Skip-attention: improving vision transformers by paying less attention. doi: 10.48550/arxiv.2301.02240

Vázquez, D., Bernal, J., Sánchez, F. J. M., Fernández-Esparrach, G., López, A., Romero, A., ... Courville, A. (2017). A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, *2017*, 1-9. doi: 10.1155/2017/4037190

Wang, J., & Lee, S. (2021). Data augmentation methods applying grayscale images for convolutional neural networks in machine vision. *Applied Sciences*, *11*, 6721. doi: 10.3390/app11156721

Wang, J., Zhong, Y., & Zhang, L. (2023). Change detection based on supervised contrastive learning for high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1-16. doi: 10.1109/tgrs.2023.3236664

WANG, N., & ZHAO, X. (2023). Time series forecasting based on convolution transformer. *Ieice Transactions on Information and Systems*. doi: 10.1587/transinf .2022edp7136

Wang, W., Lu, Y., Chen, L., Lin, B., He, X., & Liu, W. (2021). Crossformer: A versatile vision transformer hinging on cross-scale attention.
doi: 10.48550/arxiv.2108.00154

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Ding, L., ... Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.
doi: 10.48550/arxiv.2102.12122

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., & Zhu, X. X. (2022). Self-supervised learning in remote sensing: a review. *IEEE Geoscience and Remote Sensing Magazine*, *10*, 213-247. doi: 10.1109/mgrs.2022.3198244

Wang, Y., Albrecht, C. M., & Zhu, X. X. (2022). Self-supervised vision transformers for joint sar-optical representation learning. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. doi: 10.1109/igarss46834.2022.9883983

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Liu, Y., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers.
doi: 10.48550/arxiv.2103.15808

Wu, K., Peng, H., Chen, M., Fu, J., & Chao, H. (2021). Rethinking and improving relative position encoding for vision transformer.
doi: 10.48550/arxiv.2107.14222

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., & Fu, Y. (2019). Large scale incremental learning. In *2019 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 374-382). doi: 10.1109/CVPR.2019.00046

Wu, Z., Sun, C., Xuan, H., Liu, G., & Yan, Y. (2024). Waveformer: wavelet transformer for noise-robust video inpainting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*, 6180-6188. doi: 10.1609/aaai.v38i6.28435

Xue, M., Zhang, H., Song, J., & Song, M. (2022). Meta-attention for vit-backed continual learning.
doi: 10.48550/arxiv.2203.11684

Yang, L., Wu, X., Praun, E., & Ma, X. (2009). Tree detection from aerial imagery. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. doi: 10.1145/1653771.1653792

Yang, T., Chen, Y., & Sze, V. (2017). Designing energy-efficient convolutional neural networks using energy-aware pruning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/cvpr.2017.643

Zhao, T. Z., Fang, I., Kim, J., & Friedland, G. (2021). Multi-modal ensemble models for predicting video memorability.
doi: 10.48550/arxiv.2102.01173

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, *127*(3), 302–321.