Leipzig University
Institute of Computer Science
Degree Programme Computer Science, M.Sc.

Classification of Multimodal Social Media Posts

Master's Thesis

Pieer Achkar

1. Referee: Prof. Dr. Martin Potthast
2. Referee: Dr. Harry Scells

Submission date: April 15, 2024

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, April 15, 2024

. . . .
Pieer Achkar

Abstract

Navigating the complex landscape of digital media culture, this thesis presents an innovative approach to classify social media posts using state-of-the-art deep learning models for comprehensive analysis of both image and text. Various vision models, including ResNet, SEER, Vision Transformer (ViT), Shifted Window Transformer (Swin), and CLIP, along with BERT for text analysis, are used to classify Instagram content into "norm-beauty" and "divers" categories. The evaluation demonstrates the individual capabilities and synergistic potential of combining visual and text analysis. With SEER excelling in image classification and BERT in text analysis, especially with context-rich data. The thesis emphasizes the importance of the Large Language and Vision Assistant (LLaVA) as a zero-shot classifier, using advanced prompt engineering to significantly refine its classification accuracy. Additionally, it proposes a late fusion model that fuses ViT with BERT for multimodal analysis, further improving classification accuracy. Overall, this thesis reveals the robust capabilities of deep learning approaches, with a particular focus on multimodal models, in the field of data science to solve complex tasks. It underscores the transformative power of prompt engineering in improving the effectiveness of zero-shot classifiers and confirms the strength of custom-trained models on a given dataset.

Contents

Chapter 1

Introduction

In an increasingly interconnected world, the digital realm has become an essential aspect of daily life. Social media, once considered a communication platform, now stands at the intersection of culture and information. It reflects and influences societal beliefs, values, and norms. Its role in shaping perspectives on topics ranging from global politics to deeply personal questions of identity and self-worth is critical. Among these topics, the perception of beauty has undergone significant evolution due to the widespread influence of these platforms, with research indicating that social media has a major impact on the perceptual, emotional, cognitive and behavioural aspects of body image, encouraging narrow definitions of beauty and contributing to body dissatisfaction and a range of psychological well-being issues, as highlighted by Levine and Santos [2021].

The complexity of social media content, particularly when dealing with subjective themes such as beauty, poses a challenge that is well-suited for technical innovation. This thesis takes on this challenge by leveraging advanced deep learning algorithms, with a particular focus on the multimodal nature of social media content. The main objective is to classify social media posts, specifically from platforms such as Instagram, into two distinct categories: "norm-beauty" and "divers". Through this process, the thesis will investigate whether these modern algorithms can effectively distinguish and measure subjective concepts present in real-world data. The data used in this thesis reflects the perspectives of a small group of people who share the same cultural background. Therefore, it is important to mention that this thesis offers a narrow perspective through which beauty and diversity can be explored.

Research on image aesthetics and beauty classification has been conducted using various methodologies, indicating a rich field of study. For example, previous studies have used convolutional neural networks (CNNs) to categorize images from the web as visually pleasing or not, often without a specific focus on human subjects Phatak and Borkar [2020]. Others have applied traditional machine learning techniques, including support vector machines (SVMs) and neural networks, to assess facial beauty in naturalistic photographs Yan et al. [2016]. Such efforts highlight the diversity of approaches to understanding aesthetic values. However, they also indicate an opportunity for a more refined exploration of beauty standards in the specific context of multimodal social media content, a gap that this thesis aims to fill. A detailed examination of related work in Chapter 2 will further contextualize these early findings within the broader landscape of beauty classification research.

This thesis adopts a comprehensive approach to classify Instagram posts by exploring both their visual and textual elements. First, the investigation focuses on the visual content of the posts. Several vision models, such as Vision Transformer (ViT), Swin Transformer, ResNet, and SEER, were fine-tuned and evaluated to determine how much the visual content can influence the classification into "norm-beauty" or "divers" categories. Additionally, the ability of CLIP to perform zero-shot classification is also evaluated, providing a sense of how the model can recognize and categorize images without the need for additional, specific training data. After the image analysis, the focus shifts to the textual content of the posts. BERT is then fine-tuned to classify the captions, exploring the linguistic expressions associated with beauty and diversity. The work then proceeds to the multimodal aspect by utilizing the Large Language and Vision Assistant (LLaVA) model, which is a state-of-the-art multimodal large language model (MLLM). The model's ability to perform as a zero-shot classifier is extensively evaluated through various experiments. Finally, the development of a Late Fusion model combining ViT and BERT contributes further to the multimodal classification approach, which involves processing text and images together, thereby capturing the full spectrum of complexity of social media content. This approach highlights the strengths of both visual and textual analysis and demonstrates the power of integrating them through advanced multimodal techniques.

The structure of this work is constructed to build upon this perspective:

- **Chapter 2, Related Work:** Reviews existing literature to highlight the novel intersection this thesis explores between beauty, aesthetics and computational methods. It contrasts this work with previous studies, setting the stage for the methodologies employed.

- **Chapter 3, Background:** Provides a foundation in the key concepts, including multimodality, the advancement of large language models (LLMs) and Prompt Engineering, which are crucial for better understanding the approach of the thesis to classify social media content.

- **Chapter 4, Methodology:** Details the technical framework, describing the architecture of each deep learning model used to analyze visual and textual data. It also explains the concept of fine-tuning and integrating models for multimodal classification.

- **Chapter 5, Evaluation:** Offers a thorough evaluation of the performance of the models on different tasks, discussing the setup for each experiment, data splitting, and metrics used. The analysis highlights the effectiveness and limitations of the proposed models.

- **Chapter 6, Conclusion:** Concludes the study by summarizing key findings and contributions to the field of data science. It also acknowledges the limitations of the research and suggests avenues for future research on the sensitive classification of beauty standards.

Chapter 2

Related Work

In recent years, the field of computational aesthetics has seen a growing interest, with researchers increasingly applying image classification models to evaluate or categorise the aesthetic value of images. This wave of research seeks to capture the complex human perception of beauty in visual content, across a wide range of compositional elements and subjective interpretations. While the challenge is immense, the progress made reflects the emerging ability of artificial intelligence (AI) to address and quantify the underlying difficulties of aesthetic evaluation. With each advancement in the field, AI's capacity to address and resolve the challenges of aesthetic evaluation has grown remarkably, marking a significant step forward in the understanding and application of computational methods in the realm of aesthetics.

Computational Aesthetics in Digital Photography In the digital photography context, Suchecki and Trzcinski [2017] approached aesthetic evaluation using a CNN, analyzing a massive dataset of 1.7 million Flickr photos. By fine tuning the AlexNet neural network for binary classification, their study was able to classify images as aesthetically pleasing or not, with an accuracy of 70.9%. This method, based purely on visual information, demonstrates the potential of deep learning to identify aesthetic values in photographs. The results provide important insights into the features that contribute to the aesthetic appeal of a photograph, such as colour saturation, sharpness and contrast. This study establishes a basic framework for the thesis, which investigates the evaluation of social media images against conventional ideals of beauty or diversity.

The study "Deep learning for assessing the aesthetics of professional photographs" by Chambe et al. [2022] evaluates the performance of aesthetic assessment models in the context of professional photography. The models were initially trained on competitive photographs from the AVA dataset, which is known for its wide variety of aesthetic scores and semantic labels. However, the models encountered new challenges when applied to other types of photography, such as fashion, architecture, and sports. After fine-tuning the models on data with various photographic categories, the results improved significantly. The study's findings emphasize the importance of fine-tuning models on domain-specific data to enhance accuracy and reliability. This is central to this thesis as it reflects the challenges of applying advanced deep learning models to the rich domain of social media.

Exploring Alternative Approaches in Aesthetic Evaluation A comprehensive review of machine learning techniques for automatic aesthetic evaluation in images is presented in the work of Bodini [2019], which explores various methodologies, including deep learning approaches, and how these have evolved from philosophical and neuroaesthetic perspectives. A critical analysis of various datasets is performed, such as the AVA dataset, and their impact on computational aesthetics is assessed. The paper also discusses the challenges and limitations of aesthetic evaluation, such as the binary criteria of 'ugly vs. beautiful' and the need for continuous ranking in aesthetic evaluation. The research insights enhance the understanding of the complex nature of beauty and its evaluation. Additionally, it supports the thesis objectives by demonstrating the importance of quantifying beauty and the necessity of considering various factors, such as cultural and socio-educational contexts, when discussing beauty.

Chandakkar et al. [2017] work introduces a different approach to computational aesthetics, referred to as 'relative aesthetics'. The study focuses on selecting the more aesthetically pleasing image from a pair, deviating from traditional binary classification models. The authors utilized a customized dataset derived from the AVA dataset to focus on comparing images within the same category, while avoiding pairs with significant differences in their ratings. A deep neural network model, specifically a Siamese model, is trained using relative comparisons. The model performs significantly better in aesthetic evaluation than the binary method. The findings offer a broader understanding of aesthetic evaluation and highlight an alternative approach that goes beyond traditional binary classifications in the field. It inspires future work that may follow this thesis to approach the evaluation of beauty as a non-binary concept.

Facial Beauty Evaluation in Machine Learning Moving to another domain within the field of aesthetics, the study by Choudhary and Gandhi [2016] focuses on facial beauty evaluation. The study utilises a range of machine learning models, including SVM, k-Nearest Neighbour (KNN), Decision Tree and Artificial Neural Network (ANN), to classify levels of facial attractiveness using the SCUT-FBP dataset, which contains images of Asian female faces. The classification was conducted in both binary (attractive or not attractive) and multi-class formats (five levels of attractiveness). While the SVM model showed lower performance, the KNN and ANN models proved to be more effective, achieving accuracies as high as 88% and 87% for binary class classification, respectively. This study highlights the potential of machine learning in assessing subjective features like facial attractiveness and the importance of considering various classification schemes, including binary and multi-class, for a comprehensive understanding of beauty standards. It demonstrates promising results that motivate further investigation of ANN in the problem of classifying images based on beauty and diversity standards.

In their study, Bougourzi et al. [2022] apply an advanced deep learning method for facial beauty prediction (FBP) using the SCUT-FBP5500 dataset, which consists of 5500 frontal facial images with different attributes such as age, gender and ethnicity. Each of these images has been rated on a beauty scale of one to five by 60 diverse volunteers, adding complexity to the FBP task. The REX-INCEP framework is the core element of their approach, combining the capabilities of ResneXt-50 and Inception-v3 models to optimise feature extraction for FBP. Furthermore, dynamic robust loss functions that adjust parameters adaptively during training, and an ensemble regression model that consolidates predictions from multiple models, further enhance the approach. By outperforming several CNN architectures, this study not only sets a new benchmark in FBP, but also demonstrates the potential of applying sophisticated machine learning techniques to the subjective domain of aesthetics. This again highlights the ability of machines to learn about such subjective tasks, which fits well with the focus of the thesis on using advanced computational methods to navigate the subjective and complex domain of aesthetics and beauty standards.

Concluding observations Reflecting on the aforementioned studies in the field of computational aesthetics, it is clear that this landscape is experiencing rapid progress, with machine learning and deep learning emerging as key innovation factors. This progress aligns with the objectives of this thesis, which aims to explore and push the boundaries of AI's ability to capture the diverse and subjective essence of beauty.

This thesis differs from the previously mentioned works by taking a multimodal approach that integrates and analyzes both visual and textual data specifically from social media. It examines whether the textual narratives shared by social media users alongside their posts influence or correlate with the visual content, thereby enhancing the ability of AI models to better capture the nuanced concept of beauty. It suggests that understanding beauty in the digital realm from a machine perspective goes beyond purely visual analysis; requiring an integrated examination of both images and accompanying text to fully capture the breadth and depth of beauty as perceived and expressed by individuals. Investigating this interplay between textual and visual elements provides a pathway to more sophisticated AI-driven analysis, with the goal of significantly improving models' understanding of beauty.

Chapter 3

Background

3.1 Multimodality

3.1.1 Definition

Multimodal Machine Learning (MMML) is an emerging field that combines data from different sources, such as text, images, and audio, to develop sophisticated systems with enhanced understanding and interaction capabilities. In recent years, there has been a notable shift towards utilizing multimodal data sources in ML-based classification models. These methods, originally developed for unimodal data, are now being adapted to handle the complexity and variety of representations found in real-world data. IV et al. [2021] highlight the transition from unimodal to multimodal approaches in MMML, emphasizing the inherent representational challenges and the innovative solutions that have emerged. This evolution is particularly relevant to the classification of Instagram posts, where the combination of visual and textual data forms a complex dataset for analysis.

However, defining 'multimodality' remains a challenge. In their discussion, Parcalabescu et al. [2021] argue that human-centered and machine-centered definitions are limited in their ability to capture the full range of multimodal interactions. Instead, a task-relative definition is proposed, which suggests that the nature and requirements of the task at hand should determine the modality of the inputs and out:

> *"A machine learning task is multimodal when inputs or outputs are represented differently or are composed of distinct types of atomic units of information."*

This approach provides a more refined understanding of multimodality and aligns it closely with the specific goals and contexts of machine learning tasks.

3.1.2 Challenges

Baltrusaitis et al. [2019] have identified five major challenges in MMML, each of which addresses a different aspect of integrating and interpreting different types of data:

- **Representation:** This involves creating feature vectors that integrate heterogeneous data types, including text, images, and audio. Two approaches are available to address this challenge. Joint representations combine features from multiple modalities into a unified vector, while coordinated representations align separate feature vectors for each modality in a common space. The choice between these methods depends on the task's requirements and the nature of the data involved.

- **Translation:** This is the conversion of data from one modality to another, such as transforming video into textual descriptions. It is addressed either by example-based methods, which rely on dictionary lookups or k-nearest neighbor searches to match values across modalities, but are limited by the available training data, or by generative methods, which are more creative as they generate new outputs rather than simply retrieving information. This approach includes grammar-based methods that generate text within predefined rules, as well as encoder-decoder networks that convert data from the source modality to another target modality. Due to advances in deep learning and the availability of large multimodal datasets, generative methods are becoming increasingly popular.

- **Alignment:** Focuses on mapping corresponding sub-elements between different modalities, which is critical for tasks such as synchronizing video with captions. The process includes explicit alignment, which aligns modalities based on related components, and implicit alignment, which is used in tasks such as speech recognition where specific alignments are not pre-defined.

- **Fusion:** Involves integrating information from various modalities to enhance prediction and analysis. There are two main categories of fusion strategies: model-agnostic and model-based approaches. Model-agnostic methods, which are of primary interest in this thesis, include early, late and hybrid fusion techniques, differing in the degree to which modalities are combined. These techniques are discussed in more detail later in this section. In contrast, model-based approaches utilize algorithms such as multiple kernel learning and neural networks, which consider inter-modality relationships.

- **Co-learning:** This includes knowledge sharing between modalities, which is particularly beneficial when there's an imbalance in information richness between them. This approach is categorized into three co-learning approaches. Parallel co-learning leverages data shared across modalities, enhancing learning in one modality with well-labeled data from another. Non-parallel co-learning works on shared concepts or categories, supporting tasks such as zero-shot learning to recognize unseen concepts in one modality based on their presence in another. The hybrid method connects two non-parallel modalities via a common dataset or modality, making it suitable for tasks such as multilingual image captioning. Each method uniquely addresses the challenges of data diversity in multimodal learning.

3.1.3 Classification Framework

The field of multimodal classification contains a variety of terminologies, each of which is subject to different interpretations. Specifically, the terms 'early', 'late', and 'hybrid' have inconsistent definitions across different studies, contributing to inconsistencies within the fusion process. To address this, IV et al. [2021] introduced a structured framework, offering clarity in the design and execution of multimodal classification models. This framework consists of five principal stages: Preprocessing, Feature Selection, Data Fusion, Primary Learner, and Final Classifier, each playing a pivotal role in the multimodal classification system as outlined in Table 3.1.

IV et al. [2021] also provide a detailed description of each of the aforementioned stages. Since the data fusion stage plays a central role in the development of a multimodal classification model within this thesis, it is necessary to examine it in more detail. As previously indicated in Table 3.1, this stage can be divided into Fusion Architecture and Data Fusion Technique.

Fusion Architecture This can be divided based on the stage at which fusion occurs during the associated procedures:

1. *Early Fusion:* In this approach, data from multiple modalities are separately pre-processed and their features extracted before being integrated (see Figure 3.1). The integration is achieved by methods such as concatenation, where the features from each modality, whether traditional

Stage	**Description**
Preprocessing	This is a preliminary step in model building to refine and enhance the raw data. It can involve removing irrelevant or corrupted data, ensuring a balanced representation of different classes, and extending the dataset using augmentation techniques.
Feature Selection	Transforming raw data into a form suitable for further processing by the model requires the extraction of high-level features that encapsulate the essence of the data. This process covers a variety of techniques such as manual feature engineering, text encoding and CNN-based feature extraction.
Data Fusion	The integration of data from different modalities into a coherent representation occurs at this stage. It allows the multimodal model to make effective use of complementary information. Additionally, multimodal models can be distinguished by their architectural design and data fusion methods.
Primary Learner	During this stage, the training process is performed, where the models learn from the combined data representations. This can be done independently for each modality or integrated with the feature extraction and final classification stages.
Final Classifier	As the final step in the learning process, this stage provides the final classification output, which may include predicted labels or class probability distributions. There are a range of models available for this purpose, ranging from simple neural networks or decision trees to advanced ensemble models.

Table 3.1: Overview of Multimodal Classification Model Stages. (Adapted from IV et al. [2021])

feature vectors or outputs from pre-trained neural networks, are combined into a single representation. The resulting combined data is then fed to the primary learning model. Early Fusion can be useful when the modalities are highly correlated or have a direct one-to-one relationship IV et al. [2021]. However, challenges may arise when modalities have different sampling rates or when continuous and discrete data types need to be aligned. To address these issues, various techniques are employed, such as PCA for dimensionality reduction or strategies for aligning heterogeneous data Joshi et al. [2021]. Despite these challenges, early fusion remains a fundamental technique in multimodal learning, providing the basis for complex model training and subsequently influencing the effectiveness of the final classifier.

2. *Late Fusion:* In comparison to early fusion, this method is distinguished

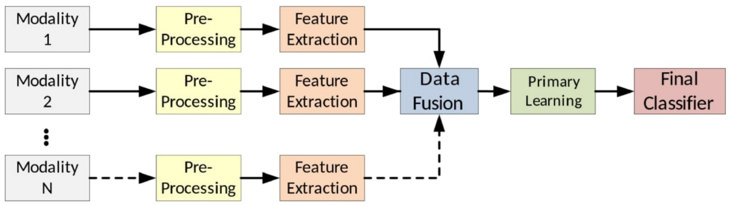

Figure 3.1: Early Fusion in Multimodal Classification Framework. (Source: IV et al. [2021])

by the independent extraction of features from each modality prior to the final classification stage. The results of this stage usually consist of a combination of low-level features learned by a deep network, or the probabilities derived by each modality's respective classifier. These distinct outputs are then combined to generate the final classification decision. One of the main benefits of this architecture is its flexibility, which enables the customization of each modality with specialized algorithms. However, there is a notable trade-off in this method, as it may ignore cross-modality learning opportunities, which may limit the model's ability to distinguish interdependencies between different types of data IV et al. [2021]. Moreover, it requires extensive learning, as each modality needs a separate supervised learning phase, followed by another learning phase to fuse the representations. This multistage process may lead to a loss of correlation within the mixed feature space, which is another drawback of this approach Snoek et al. [2005]. Figure 3.2 demonstrates the general approach for late fusion.

3. *Cross-modality Fusion:* This allows a dynamic exchange of information between modalities, either before or during the primary learning phase, which differentiates it from the static early and late fusion approaches. It enables modalities to use each other's context to improve the overall predictive power of the model, allowing flexible and interactive data sharing that can vary in scope and timing throughout the learning process. The results of such strategic collaboration have been shown to potentially outperform traditional fusion methods, offering a promising opportunity for advancing multimodal problem solving IV et al. [2021]. Figures 3.3 and 3.4 show a general approach to cross-modality architectures, where modalities engage in single or multiple data-sharing operations during the learning process.

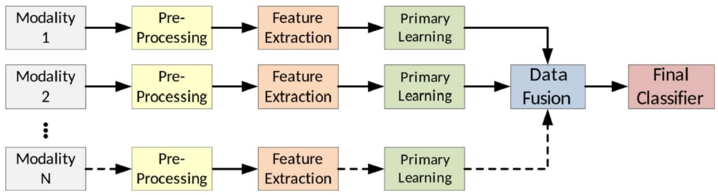

Figure 3.2: Late Fusion in Multimodal Classification Framework. (Source: IV et al. [2021])

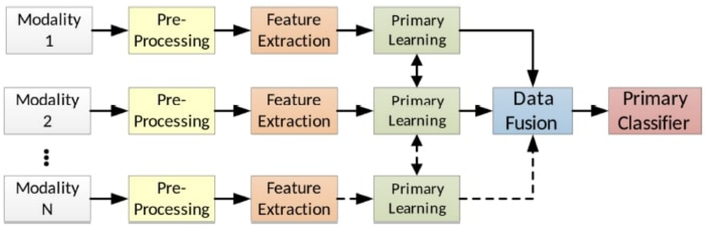

Figure 3.3: Single data-sharing operation in cross-modality fusion architecture. (Source: IV et al. [2021])

Data Fusion Technique The integration of information from different modalities to form a unified feature representation is an integral part of multimodal learning, with techniques such as concatenation and merging being widely used. *Concatenation* is a straightforward method that combines different features into a single comprehensive vector that is suitable for both raw data and processed neural network outputs. On the other hand, *merging* takes a more nuanced approach by using arithmetic operations or network layers to combine features and create a feature set that captures the complex dynamics between modalities IV et al. [2021]. These methodologies are central to creating a coherent representation, which is necessary to handle the complexity inherent in tasks that require synergistic fusion of multiple data types. The objective of fusion techniques is to enhance the capabilities of the combined modalities within a unified semantic representation Joshi et al. [2021].

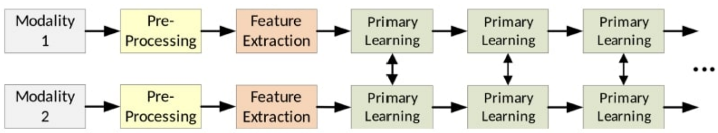

Figure 3.4: Multiple data-sharing interactions across learning stages in cross-modality fusion. (Source: IV et al. [2021])

An additional valuable consideration, IV et al. [2021] points out, is that in real-world scenarios, it's common for different stages within the multimodal classification framework to use identical models for processing. For example, in a late fusion setup, each modality may independently perform feature extraction and primary learning using the same model type. In contrast, an early fusion architecture might consolidate learning and classification within a single model, a practice typical of many traditional machine learning models, where extracted features are concatenated and fed directly into a classifier.

3.2 Multimodal Large Language Models (MLLMs)

3.2.1 Evolution

Large Language Models (LLMs) Language models have progressed from simple rule-based frameworks to the advanced LLMs of today. While basic models based on linguistic rules established the foundation, they struggled to capture the finer details of complex language usage. Advancements in statistical modeling toward the end of the 20th century improved data handling, yet deep linguistic understanding remained difficult to achieve Douglas [2023].

In the 2010s, language models made a major step forward with the introduction of neural network-based models. The Recurrent Neural Network Language Model (RNNLM) improved the generation of coherent and contextually relevant natural text by processing sequences of words. It helped overcome the limitations of statistical models by enabling complex pattern recognition and sequential data processing, which are essential for understanding the flow and structure of language Mikolov et al. [2010].

With the introduction of the Transformer model in 2017, a paradigm shift was achieved in the field of Natural Language Processing (NLP). Its innovative architecture, which utilizes self-attention mechanisms, enabled parallel processing of sequences, making it significantly more efficient than previous models at learning long-range dependencies within a piece of text Vaswani et al. [2017]. This architectural progress provided the foundation for subsequent developments in language models, including BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. [2018]. This model utilizes bidirectional training, enabling it to understand the context of a given word based on all the words around it, as opposed to previous models that processed text in one direction.

GPT-3, a pre-trained Transformer developed by OpenAI in 2020, continued the evolution of previous versions of Transformers, scaling the architecture up to 175 billion parameters. Together with advanced training techniques, The model achieved remarkable performance on a variety of different NLP tasks, often with minimum or no task-specific training Brown et al. [2020]. GPT-3 is considered a milestone in the development of LLMs due to its deep and rich understanding of language. It can generate text that closely resembles human language, solve complex problems, and detect nuanced details in text. Following the success of GPT-3, OpenAI introduced GPT-4 in 2023, further enhancing the capabilities of language modeling. GPT-4's enhancements in text understanding and generation made it more capable of mimicking human-like conversation and reasoning across a wider range of languages and domains. Its enhanced performance sets new benchmarks within LLMs, achieving human-level performance across a range of benchmarks OpenAI [2023].

The emergence of LLMs from large institutions has been accompanied by a notable trend toward the democratization of AI technology, highlighted by the release of open-source models such as Meta AI's LLaMA[1] and Mistral's Mistral 7B[2]. These open-source initiatives are changing the AI landscape by making cutting-edge technology available to a broader community of researchers, developers, and enthusiasts. The adoption of open source models has enabled broader access to modern AI technologies, as well as encouraged a culture of collaboration and innovation within the field.

Beyond language-centric models The transition from Large Language Models (LLMs) to Multimodal Large Language Models (MLLMs) that began in the 2020s is an important conceptual change toward a richer, more complex understanding of intelligence, similar to human sensory perception. While traditional LLMs perform well in NLP tasks, they have a fundamental limitation, which is their inability to process information beyond text Yin et al. [2023]. This constraint emphasizes the necessity for a comprehensive and integrated approach to comprehend the diverse nature of communication and information, thereby opening the path for the emergence of MLLMs.

Recent models such as GPT4V[3], Gemini[4] and LLaVA (to be discussed further in this thesis) are examples of this transition, representing a move from purely

[1] `https://ai.meta.com/blog/large-language-model-llama-meta-ai/`
[2] `https://mistral.ai/news/announcing-mistral-7b/`
[3] `https://openai.com/research/gpt-4v-system-card`
[4] `https://deepmind.google/technologies/gemini/#introduction`

textual analysis to a multimodal approach that incorporates visual, auditory, and textual inputs. This evolution mirrors the complexity of human cognition, which does not rely on a single mode of perception, but integrates multiple sensory inputs to interact with and understand the world. The ability to aggregate and interpret information across multiple different modalities is key to enabling both humans and AI systems to effectively navigate and make sense of the rich variety of information in the natural world.

Integrating multimodal input has extended the application of LLMs, improving their usability in domains such as human-computer interaction, image recognition, and speech generation. Through extensive experimentation, it has been shown that the cross-modal knowledge transfer provided by MLLMs can significantly enhance reasoning capabilities, often surpassing the performance of models limited to a single modality Wu et al. [2023]. This proliferation highlights the immense potential of AI to surpass conventional boundaries, indicating a promising future for innovation and application in a wide range of fields.

3.2.2 Applications

The broad applicability of MLLMs has led to a variety of use cases, each of which demonstrates the potential of integrating multimodal data for complex problem solving and decision making. Some of the widely reported use cases are:

- **Visual Question Answering (VQA):** It is a multipurpose application of MLLMs that combines computer vision with NLP to answer questions based on image content. It requires a deep understanding of both visual and textual data, going beyond simple object recognition or scene description. Due to its dynamic and real-time nature, VQA is closely related to practical AI applications Yuan [2021]. In the fashion industry, VQA improves the efficiency and accuracy of product labeling by automatically answering attribute-based questions from images Wang et al. [2022a]. In the medical field, it has recently been adopted for a variety of applications, including assisting clinicians in decision making, enhancing medical education through image interpretation, automating disease diagnosis, and providing answers to patient questions that do not require a doctor's visit Al-Sadi et al. [2021].

- **Image Captioning:** The objective of this task is to create natural language descriptions based on visual content, connecting visual perception and linguistic expression. This is crucial for aiding blind people and improving human-robot interaction, as it requires not only the recognition of visual elements, but also the translation of these perceptions into coherent, contextually relevant sentences Laina et al. [2019]. In the past, this task has typically relied on human annotators, which can sometimes result in certain limitations, such as descriptions that may be overly simplistic or repetitive. However, MLLMs offer a solution to these challenges by learning a joint embedding space of language and image features. This allows them to generate more descriptive and accurate annotations from images without necessarily relying on paired datasets.

- **Emotion Recogniton:** This task focuses on the interpretation of human emotions by integrating various types of multimodal data, such as images, video, audio, and text. It has gained attention for its potential to improve user interactions and mental health applications. The use of MLLMs allows researchers to systematically evaluate the the capabilities of these models in tasks such as recognition of facial expressions, analysis of visual sentiment, and detection of micro-expressions. Lian et al. [2023] have made significant progress in this field by demonstrating that GPT-4V can analyze both image and text input and can partially capture temporal information from video frames. However, they have addressed some challenges. For instance, the interpretation of audio data remains limited, and there are inconsistencies in the security of the evaluation. The authors' quantitative evaluation of GPT-4V sets a new benchmark in the field, demonstrating its potential and identifying areas for future development in multimodal emotion understanding.

3.3 Prompt Engineering

In the field of advanced language models, prompt engineering is considered a key and powerful technique. It involves carefully designing and constructing input queries, or "prompts", which play a critical role in guiding the behavior of LLMs to produce specific, desired outcomes Kaddour et al. [2023]. The core principle of prompt engineering is that the way prompts are structured and presented to the model has a profound impact on their responses. By effectively engineering these prompts, it is possible to leverage the capabilities of the model to produce a variety of meaningful and accurate outputs.

As language models have evolved, several innovative prompting techniques have been introduced to improve the model's performance for specific tasks. Some of these methods include:

1. **Zero-shot Prompting:** Refers to the ability of the model to perform tasks without being presented with examples or being trained in that particular domain of knowledge. This approach takes advantage of the model's pre-training on a wide range of data, enabling it to perform well and provide accurate responses to questions it has never encountered during training. For example, a model could be given a prompt such as *"Translate this sentence into French: The boy goes to school"* and provide an accurate response even without explicit training on translation tasks. However, it is important to note that Zero-shot prompting may have certain limitations. One main limitation is the possibility of inconsistent performance across tasks and languages, particularly those that are not well represented in the training data. In addition, as mentioned by Sanh et al. [2021], creating effective prompts that yield the desired response can be a challenging process that may require significant trial and error. Moreover, while Zero-shot Prompting is capable of producing grammatically correct results, it may not always have the necessary domain-specific accuracy or contextual understanding, especially in highly specialized domains such as law or medicine Martínek et al. [2022].

2. **Few-shot Prompting:** Unlike Zero-shot prompting, this method provides the model with a small amount of examples to guide its reasoning. For example, in Few-shot video language learning, a model might be shown only a few examples of video frames paired with descriptions, and then learn to generate accurate descriptions for new, unseen video frames Wang et al. [2022b]. This method has the advantage of quickly adapting to new tasks with limited data, making it particularly useful in situations where large-scale datasets are unavailable or not feasible. However, the quality of this method can be highly dependent on the representativeness and robustness of the examples provided. In some cases, Few-shot prompting can lead to inconsistencies or biases if the sample prompts are not carefully selected Patel et al. [2022]. This approach has been applied to a variety of complex reasoning and problem-solving scenarios. For example, it has been used to improve the performance of small language models, transforming them into efficient Few-shot learners without the need for fine-tuning Zhang et al. [2021]). Evaluation of Few-shot prompting often focuses on its flexibility, efficiency in learning from limited data, and ability to generalize to new tasks or domains.

3. **Chain of Thought (CoT) Prompting:** Introduced by Wei et al. [2022], this technique has significantly improved the reasoning abilities of LLMs by guiding them through intermediate reasoning steps, especially

in complex tasks that require multi-step logical reasoning. For example, the addition of phrases such as *"Let's think step by step"* in Zero-shot CoT helps models decompose problems and arrive at more accurate, interpretable answers Kojima et al. [2022]. To automate and refine this process process, innovations such as the Automatic CoT by Zhang et al. [2022b] aim to generate diverse chains of reasoning, minimize manual intervention, and maximize the representativeness of examples. Despite potential challenges such as dependence on example quality and model size, CoT prompting represents a strong advance in the ability of LLMs to tackle sophisticated reasoning tasks with improved accuracy and depth of understanding.

The exploration of the aforementioned prompting techniques covers important developments in the field, but only briefly delves into this rapidly evolving area. As AI research advances, new and innovative prompting approaches are being developed, enriching the tools available for interacting with modern advanced models. This overview provides a sufficient foundation for the scope and purpose of this thesis, recognizing that prompt engineering is an expansive and dynamic field that offers a wide avenue for future investigation.

Chapter 4

Methodology

4.1 Data

4.1.1 Data Collection & Labeling

The dataset utilized in this thesis consists of Instagram posts, which include images and their corresponding captions. The collection was limited to public posts only, in accordance with ethical standards and Instagram's privacy policy.

The acquisition of the data occurred in the summer of 2023, at the Ludwig Uhland Institute for Empirical Cultural Studies at the University of Tübingen. The objective of the seminar was the examination of social media feeds, specifically Instagram. The students used ethnographic methods to categorize Instagram posts into two categories: "norm-beauty" and "divers" representations. They were divided into groups and tasked with developing and refining criteria to guide their post selection. Initially, 50 posts were obtained for each category, with an emphasis on overall thematic representation rather than strict compliance with the initial criteria.

Throughout the seminar, the students conducted a critical evaluation and adjustment of their selection criteria, particularly concerning the "norm-beauty" category where gender bias was detected. This reflective process resulted in a more inclusive and balanced dataset curation. An important component of the seminar was the establishment of a subgroup dedicated to identifying subtle forms of racism and discrimination present in the posts. However, the subgroup was dissolved shortly after its formation due to the emotional challenges of reviewing such sensitive content.

Figure 4.1: Representative samples of 'norm-beauty' (left) and 'divers' (right) images.

The resulting dataset, consisting of 472 'norm-beauty' and 456 'divers' posts, was organized into an Excel file containing links to the original Instagram content. As part of the work in this thesis, web crawling techniques were used to obtain both images and captions from the links.

Due to ethical concerns and the desire to preserve the privacy and integrity of the original Instagram users, this thesis does not include actual images from the dataset. Instead, representative examples were generated using the *Stable Diffusion XL* image generation model. This approach was chosen to demonstrate the types of visual content categorised as "norm-beauty" and "divers", without directly displaying personal or sensitive information.

To ensure the reliability of the labels and their reproducibility, the dataset was subjected to an additional verification step. A sample of 10% of the dataset was independently labeled by seven annotators with similar cultural backgrounds. The accuracy rates of the annotators ranged from 73% to 96%. Furthermore, a majority voting process was used to determine the most frequently assigned label for each post. This resulted in an overall accuracy of 91%, demonstrating the coherence and reliability of the labels. Overall, these observations suggest that the decision-making process behind the labeling of posts into "norm-beauty" and "divers" representations is based on a collective understanding that is to some extent reproducible.

4.1.2 Data Preprocessing

The images were processed to standardise their colour representation, ensuring consistency across the dataset. If an image was not in the RGB standard, it was converted to this format to prevent variations from affecting the analysis.

For the textual data associated with each Instagram post, significant variability was observed, reflecting the diverse nature of social media interactions. The content of the posts ranged widely in format and content; some contained only emojis, others were structured as commercial or personal stories, and a few were entirely textless. Given that the focus of further analysis is to classify this data, the main preprocessing step was to remove entries with empty captions.

4.1.3 Data Augmentation

Data augmentation is an important element of image-based machine learning, especially when dealing with limited datasets. It enables the expansion of training data diversity and quantity whilst creating a reliable dataset from a limited number of source images. This is very useful in scenarios where it is either impractical or impossible to acquire a large number of unique images. By augmenting images with different techniques, the machine learning models are exposed to a wider range of scenarios and variations. Such diversity is valuable because, from a machine's perspective, even small changes to an image can significantly alter how it is processed and understood.

Several augmentation techniques were employed to enrich the dataset, following established best practices in the field of machine learning, as highlighted by Yang et al. [2022]. These techniques included

1. **Rotation:** Each image was rotated by 15 degrees, introducing a new perspective while preserving the original content of the image.

2. **Horizontal Flip:** This method created a mirror image of each original, thereby adding variation to the dataset while maintaining the structural integrity of the subjects.

3. **Brightness Adjustments:** The brightness of each image was modified to mimic different lighting conditions. This included both increasing and decreasing brightness levels to prepare the models for real-world lighting variations.

4. **Gaussian Noise Addition:** To simulate common photographic challenges such as graininess, Gaussian noise was added to the images. This

adds a challenge to the model so that it can process and analyse images under less than ideal conditions.

5. **Background Removal:** A custom background removal tool was employed to isolate foreground subjects, focusing on the main subject and reducing potential background distraction.

6. **Foreground Removal:** In contrast, this method removed the main subject, leaving only the background. This could help to understand the context and environment of the subjects.

4.1.4 Data Transformation

Transforming raw data to ensure consistency for machine learning applications is another important step for conducting multiple experiments to address the problem at hand. Both the textual and visual data underwent some transformation to achieve uniformity and improve the effectiveness of the analysis.

An initial transformation stage focused on textual content, primarily image captions. *GPT-4*, OpenAI's most advanced generative language model, was utilized to standardize and reformulate these captions into a coherent structure. This process included:

- **Language Standardization:** All non-English captions were translated to English.
- **Conversion of Emojis and Hashtags:** Emojis and hashtags were replaced with corresponding natural language descriptions, preserving the original posts' sentiment and context.

Guiding Prompt for Textual Transformation:
"Reformulate the following Instagram caption in one sentence to capture the essence of it. Focus on the meaning of the caption, the sentiment, and the context.
- If the caption has emojis, remove them and explain them in natural language.
- If the caption has hashtags, remove them and explain them in natural language.
- If the caption is not in English, translate it to English and reformulate."

In addition to transforming the textual data, the images were also transformed into textual descriptions. This was achieved using the LLaVA model. LLaVA analyzed each image to provide a detailed description that included various aspects such as pose, body prominence, and skin appearance.

Guiding Prompt for Visual Transformation:
Analyze the person in the image. Provide a JSON response with the following fields:

```
{
  "pose_and_posture": "Description of their stance",
  "body_prominence": "How their body is displayed",
  "skin_appearance": "Details of skin texture and features",
  "body_features": "Information on weight, slimness, muscularity, and facial characteristics",
  "disability_or_syndrome": "Indicators of any disabilities or syndromes"
}
```

Following the presentation and discussion of the dataset, the remaining sections of this chapter will provide a detailed technical explanation of the models and methods used in this thesis.

4.2 Text Models

4.2.1 Bidirectional Encoder Representations from Transformers (BERT)

In the field of NLP, the introduction of BERT by Google AI was a significant advancement. With its bidirectional processing capabilities, it enabled a rich understanding of linguistic context, providing a considerable improvement over the unidirectional approach of previous models Devlin et al. [2018]. The underlying architecture is built on the transformer principles developed by Vaswani et al. [2017], which is recognized for its ability to handle sequential data and its implementation of advanced attention mechanisms. BERT uses only the encoder part of the Transformer architecture, which is well suited to understanding the context and relationships between words in a text.

The preprocessing of the input text starts by converting tokens into embeddings that the model can understand. This process includes three main components:

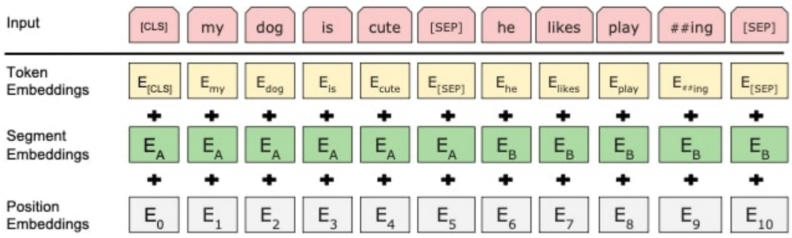

Figure 4.2: Illustration of BERT's Input Processing. The diagram shows the addition of token, segment, and positional embeddings to the input sequence, preparing it for encoding by BERT. (Source: Devlin et al. [2018])

1. **Token Embeddings:** Each token in the input sequence is converted into a vector. Additionally, special tokens are appended: a [CLS] token is placed in front of the input to serve as an aggregated representation for classification tasks, and [SEP] tokens are added to mark the end of sentences.

2. **Segment Embeddings:** To enable the model to distinguish between sentences in tasks that involve multiple sentences, such as question answering, BERT adds segment embeddings. These are binary flags associated with each token indicating whether it belongs to the first sentence (Sentence A) or the second (Sentence B).

3. **Positional Embeddings:** These are used to specify the position of each token within the sequence, preserving word order information.

The embeddings from these three sources are combined to create a comprehensive representation for each token which is fed into the model. This process, as shown in Figure 4.2, is essential for BERT to contextualize individual tokens and understand the entire input sequence.

BERT is available in two primary variants, the base model with 12 encoders and 12 attention heads, totaling approximately 110 million parameters, and the large model with 24 encoders and 16 attention heads, totaling approximately 340 million parameters, designed for more computationally intensive tasks. Two main strategies are used to train the model:

- **Masked Language Modeling (MLM):** The aim of this method is to improve the model's ability to predict words that are masked in the context of the text. For instance, considering a modified sentence such as "Rain is essential for the growth of [MASK]." BERT is tasked to infer

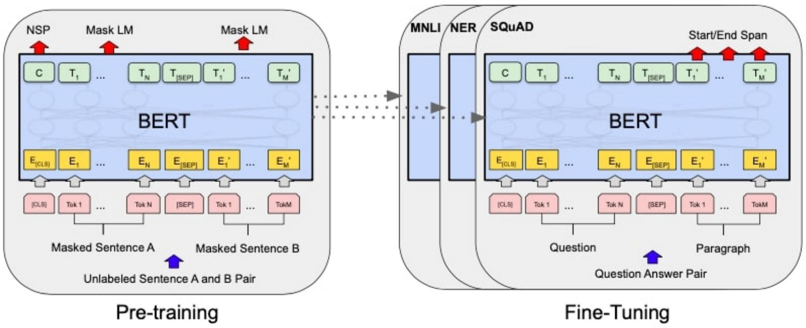

Figure 4.3: BERT's Training Process. The left side shows pre-training with NSP and MLM on unlabeled data, while the right side shows fine-tuning BERT for specific tasks such as question answering. (Source: Devlin et al. [2018])

the masked word - in this case, "plant" - based only on the surrounding context. To prevent the model from overfitting to the [MASK] token and ensure generalizability, the training procedure is diversified: out of the tokens intended for masking, 80% are in fact replaced by [MASK], while 10% are replaced by random tokens and the remaining 10% are left unchanged. This approach forces the model to predict the masked word, but also to improve its predictive ability, regardless of the token's appearance. The model's ability to acquire knowledge is therefore directed towards the reconstruction of the masked tokens, leading to an incremental but profound learning of bidirectional language representation, which is central to a wide range of NLP applications.

- **Next Sentence Prediction (NSP):** This teaches the model to understand the connection between pairs of sentences, which is essential for tasks that require a deeper understanding of textual relationships, such as question answering. In this task, BERT evaluates whether the second sentence in a pair logically follows the first. During training, the model is presented with two sentences separated by a [SEP] token. It must predict whether the second sentence is a true continuation or a random insertion. This binary decision is based on the output of the [CLS] token and is achieved by a classification layer within the model.

BERT's pre-training involves learning from large text corpora to understand sentence context and word relationships, which is a fundamental building block. Then, through fine-tuning, the model adapts to a given task by adjusting its input-output structure and efficiently learning the necessary domain-specific knowledge. This two-stage approach, as shown in Figure 4.3, ensures that the model is applicable to various NLP applications and thus achieves a high degree of maturity.

4.3 Vision Models

4.3.1 Residual Neural Network (ResNet)

Introduced in "Deep Residual Learning for Image Recognition" by He et al. [2015], ResNet represents a significant advancement in deep neural networks in the field of computer vision, as it overcomes the vanishing gradient problem that hinders effective updating of network parameters. This problem was a critical challenge in the era post AlexNet [1], a CNN model that won the ImageNet [2] 2012 competition. As architectures such as AlexNet became deeper, they encountered training difficulties caused by the vanishing gradient problem. ResNet, with its skip connections, introduced a new approach to deep network training, improving depth and efficiency, leading to its success in the 2015 ImageNet Challenge and establishing its role in the advancement of computer vision and deep learning.

ResNet's architecture focuses on residual learning. This approach, where the network layers approximate residual functions, is defined as $F(x) = H(x) - x$. It assumes that if nonlinear layers are capable of approximating complex functions, they should also be capable of approximating the residual functions. According to this, the layers do not learn the direct mapping of $H(x)$, but rather the difference from the input, which has been shown to simplify the learning process, especially in deeper networks. The residual learning approach is complemented by the concept of skip connections, as shown in Figure 4.4. These connections allow gradients to bypass specific layers, a mechanism crucial for addressing the training challenges often encountered in deep neural networks. Skip connections facilitate identity mapping, ensuring that inputs can be carried forward with minimal change, which is essential for optimising the training of deep neural networks. By focusing on refining a smaller subset of features, it enables the training of deeper architectures, effectively preventing the degradation of training performance that is typically observed in very deep networks He et al. [2015].

The various architectures of ResNet, ranging from ResNet-34 to ResNet-152, where the number refers to the number of layers in the neural network archi-

[1] `https://en.wikipedia.org/wiki/AlexNet`
[2] `https://www.image-net.org/challenges/LSVRC`

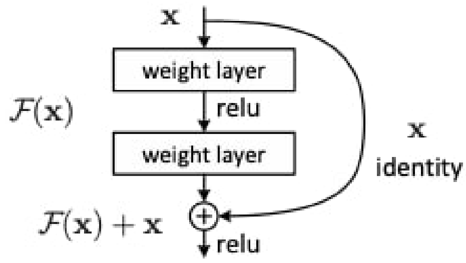

Figure 4.4: A standard residual block in ResNet, demonstrating the skip connection (Source: He et al. [2015]).

tecture, demonstrate the ability to scale depth while maintaining performance, significantly reducing image classification error rates on datasets such as ImageNet. The influence of ResNet on the field of computer vision is evident in its widespread application for solving complex tasks He et al. [2015].

4.3.2 Self-Supervised Model (SEER)

Developed by Meta AI, SEER (Self-supERvised) departs from traditional supervised learning models by leveraging self-supervised learning. It was trained on a large and diverse dataset of more than one billion public Instagram images, intentionally excluding EU-sourced content due to privacy issues. The dataset, while randomly selected and unfiltered, was analyzed for geographic and gender diversity, resulting in a comprehensive representation from 192 countries.

Model Architecture Goyal et al. [2022] sought to scale up to a highly dense network with 10 billion parameters, based on the RegNet architecture — a CNN variant known as "Regularised Network". RegNet was chosen for its promising scalability, documented in Radosavovic et al. [2020], offering a flexible "design space" as compared to fixed architectures. This flexibility allows for custom modifications necessary to support SEER's large number of parameters.

The RegNet architecture is built upon an initial design space called AnyNet, which is a flexible interpretation of the ResNet structure, shown in Figure 4.5 that represents a segmented approach: the initial processing Stem, the computationally intensive Body with four stages of X-blocks, and the concluding Head. The X-blocks, foundational to the model's adaptability, are illustrated in Figure 4.6 and are parameterized by their width, bottleneck ratio, and

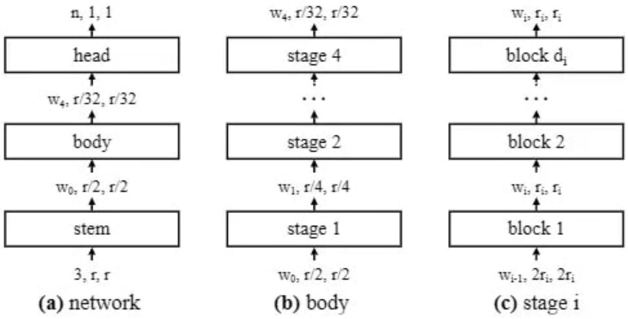

Figure 4.5: Overview of the network architecture: (a) shows the high-level division into Stem, Body, and Head; (b) the four stages within the Body, each comprising multiple computational blocks; (c) zooms into Stage 1, illustrating the block structure with its associated parameters. (Source: Radosavovic et al. [2020])

group width at a fixed resolution. This design affords the model 16 degrees of freedom, facilitating extensive exploration within the architectural landscape. This exploration is marked by a series of iterations from AnyNetXA through to AnyNetXE, each progressively refining the design space:

- **AnyNetXA:** The baseline model with an unconstrained ResNet-like architecture, allowing for approximately 10^{18} possible structures, accounting for all permutations of its four parameters.
- **AnyNetXB:** Constrains the bottleneck ratio, reducing the design space to around 10^{16} possibilities.
- **AnyNetXC:** Standardizes the group width across layers, further narrowing down the possibilities to 10^{14}.
- **AnyNetXD:** Implements non-decreasing layer width, limiting structural variations to 10^{13}.
- **AnyNetXE:** Applies non-decreasing layer depth, yielding a pragmatic 10^{11} design variations.

Building on the foundations introduced by AnyNetXE, the RegNet architecture introduces two distinct series: RegNetX and RegNetY. RegNetX is built through an empirical optimization of parameters, including the initial layer width w_0, slope w_a, and quantization w_m, which collectively define the network's width and depth. RegNetY improves the design by adding a squeeze

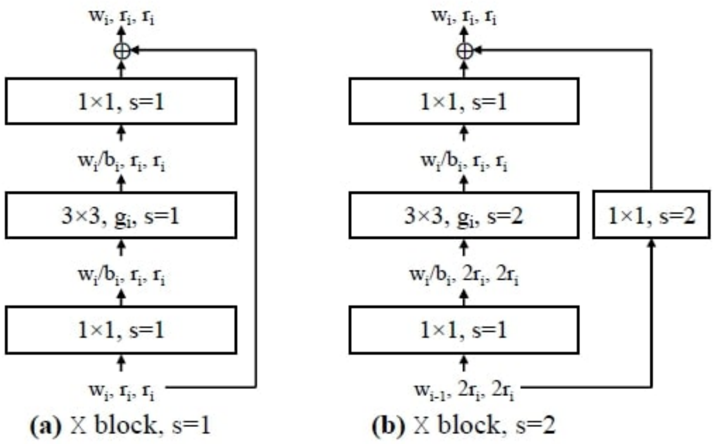

Figure 4.6: (a) The X block, a basic building block of the AnyNetX architecture with stride $s = 1$. (b) Illustrates the downsampling variant of the X block with stride $s = 2$. (Source: Radosavovic et al. [2020])

and excitation (SE) layer which uses an SE ratio q to adjust the filter responses, thereby enhancing the network's feature discrimination capabilities. These models undergo a detailed parameterization process to ensure hardware compatibility, such as the widths being divisible by 8 and conforming to the group size g, determined by the bottleneck ratio b.

The SEER model is the result of strategic scaling of the RegNet architecture, with a focus on increasing width over depth to efficiently increase model size. The choice was based on experiments that showed that models with increased width and depth improved performance without the inefficiencies of higher resolution or the use of complex models. As a result, SEER maintains the base resolution while significantly increasing the width to provide a good balance between scale and training efficiency.

Self-Supervised Learning (SSL) SEER was trained using the Swapping Assignments between multiple Views (SwAV) technique introduced by Caron et al. [2020]. SwAV uses a sophisticated method to learn visual features through self-supervision. It has a multi-stage process that starts by generating different image crops, including both large and small sizes. These crops are processed by a neural network, such as ResNet-50, which outputs feature vectors. These vectors are then associated with prototype vectors that define distinct visual categories. By maximizing the similarity between image features and these prototypes, cluster assignments are created. The model is trained to predict clusters of one image view based on another, effectively teaching it to recognize the same object across different image views. This self-predictive mechanism, reinforced by swapped vector predictions, enables SwAV to develop consistent and transferable representations for visual recognition tasks.

According to Goyal et al. [2022], SEER is a more robust, fair, less biased, and less harmful model compared to models trained on curated datasets such as ImageNet. By scaling the model's capacity, they demonstrate that it can capture a wide range of concepts, including semantic information, artistic style, and geographic diversity. The performance has been validated on over 50 benchmarks, demonstrating improved generalisation capabilities and strong performance across a range of computer vision tasks.

4.3.3 Vision Transformer (ViT)

To leverage the power of attention-based mechanisms in computer vision, Dosovitskiy et al. [2021] propose an architecture that approaches image processing from the perspective of sequence transformation. Inspired by the transformer models, ViT represents images as a series of distinct elements.

The core of ViT's architecture consists of partitioning an input image into fixed-size patches, typically 16×16. These patches are then flattened and linearly projected to a high-dimensional space. Each patch is represented by a D-dimensional vector. This process is similar to tokenization in NLP, where words are converted into tokens before being embedded. To preserve positional information lost during patch flattening, a learnable 1D position embedding is added to each patch embedding. Following a similar technique to BERT, ViT introduces a "class token", an additional learnable embedding that is appended to the sequence of patch embeddings. This token serves as an aggregate representation of the entire image and is used for classification after being processed by the Transformer encoder.

The encoder in ViT consists of multiple identical layers, each of which includes a multi-headed self-attention (MSA) and a multi-layer perceptron (MLP) network. Layer normalization (LN) is applied to each MSA and MLP, and residual connections are integrated around both components to improve gradient flow during training. The self-attention mechanism is essential as it allows the model to assess and prioritize different image patches, thereby focusing on the most informative patches for specific tasks. The MLP, which has two layers with GELU non-linearity, further processes features derived from the attention

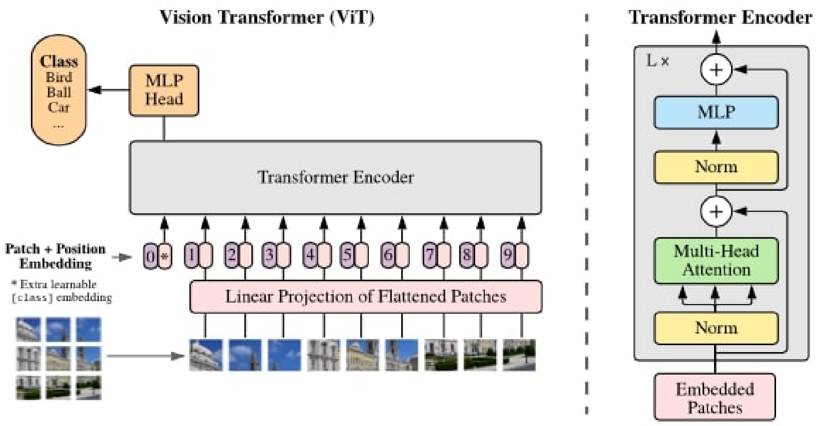

Figure 4.7: ViT architecture, showing the flow from input image patches to final classification output. (Source: Dosovitskiy et al. [2021])

mechanism. Figure 4.7 illustrates the ViT architecture, outlining the transformation from input image patches to the final classification output. One of the key features of the model is the ability to apply global attention across all image patches, which differs from the local attention of CNNs. As a result, the model is able to capture extensive dependencies between patches, providing a more comprehensive understanding of the image.

There are different sizes of ViTs, such as ViT-Base, ViT-Large, and ViT-Huge. The authors have also developed a hybrid model that incorporates CNNs to form input sequences, thereby improving the feature extraction process. These models offer flexibility in terms of patch size, thereby addressing different computational needs and tasks. When trained on extensive datasets such as ImageNet-21k and JFT-300M, ViTs demonstrate remarkable capabilities. They often equal or exceed the performance of traditional CNN models such as ResNet, while requiring fewer computational resources. Notably, ViT-H, the largest model in the ViT series, exhibits a faster training process compared to its CNN counterparts. However, the performance of these models on smaller datasets does not achieve the same level of effectiveness as CNNs. This is due to inherent design differences between the two architectures. CNNs have a convolutional structure that naturally has an inductive bias that favors local connectivity, which aids in pattern recognition within images. In contrast, ViTs lack these built-in biases and instead heavily rely on rich training datasets to learn the underlying visual features Dosovitskiy et al. [2021]. While this reliance provides greater flexibility, it also creates a dependency on massive datasets. It reflects a broader trend in deep learning, where transformers are increasingly recognised as more general computational frameworks compared to specialised networks such as CNNs.

4.3.4 Shifted Window Transformer (Swin)

The Swin Transformer builds on the fundamental concepts of the ViT to address its computational inefficiencies, in particular when processing high-resolution images. By design, ViT's computational complexity increases quadratically with image size, making it poorly suited for high density prediction tasks such as object detection and semantic segmentation. By introducing hierarchical feature maps and shifted window attention mechanisms, the Swin Transformer achieves a significant reduction in computational complexity and improves the processing of images at different resolutions Liu et al. [2021].

Unlike the fixed scale approach of ViT, the Swin Transformer employs a flexible patch merging strategy that produces adaptive and scalable feature maps. This strategy combines features from a 2×2 grid of adjacent patches into a single patch, enriching the channel depth and reducing the spatial dimensions by half. This preserves essential information while creating a compressed yet detailed representation of the input. Specifically, the merging strategy groups each $n \times n$ neighboring patches and concatenates them depth-wise, effectively downsampling the input by a factor of n. As a result, the input's dimensions are transformed from $H \times W \times C$ to $\frac{H}{n} \times \frac{W}{n} \times (n^2 \times C)$, where H, W, and C represent the height, width, and channel depth, respectively. Through recursive application across the network layers, this technique constructs a hierarchical feature maps, capturing details at diverse scales essential for the model's effectiveness in complex vision tasks. This architecture balances computational efficiency with enhanced predictive performance for various image analysis applications. Figure 4.8 illustrates the hierarchical architecture of the Swin Transformer, showcasing the adaptable feature maps in contrast to the single-resolution feature maps typical of ViT.

Another key component enhancing the capabilities of this model is the Swin Transformer Block, which refines the processing of complex visual information. It replaces the standard MSA of the ViT with two specialized modules designed to optimize computational efficiency and accuracy:

- **Window-based Multi-Head Self-Attention (W-MSA)**: Instead of considering the entire image at once, which is computationally intensive, W-MSA divides the image into smaller, more manageable sections called

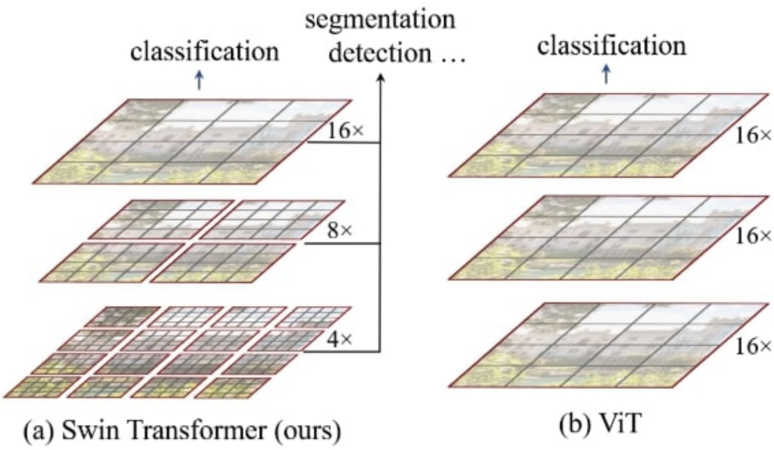

Figure 4.8: A comparison of the hierarchical feature maps in Swin Transformer (a) and the uniform single-level feature maps in Vision Transformer (ViT) (b) (Source: Liu et al. [2021])

.

windows. Within each window, the self-attention mechanism focuses on understanding the relationships between patches, which are small parts of the image, thereby simplifying the process and reducing the computational workload. This localized attention is comparable to understanding a scene not by looking at everything in sight, but by focusing on one part at a time.

- **Shifted Window Multi-Head Self-Attention (SW-MSA)**: While W-MSA is efficient, it works in isolation, which means it might overlook the complete context of the image. SW-MSA solves this problem by slightly shifting the positions of the windows, allowing each window to consider its neighbors. This shift enables the model to assemble a more complete understanding of the entire image. The effect is like shifting the perspective slightly to see what's happening at the edge, providing a broader view.

The Swin Transformer Block consists of a repeating pattern of two types of layers. The first type includes the W-MSA module for concentrated attention within windows. This is followed by a layer that normalizes the data, ensuring consistency and stability in the model's predictions. Finally, an MLP layer is included, which acts as a complex filtering layer that further refines the data. The second type is similar, but it utilizes the SW-MSA module to expand the attention's scope to different areas of the image. The design of the Swin Transformer Block is presented in Figure 4.9. It shows the flow of information

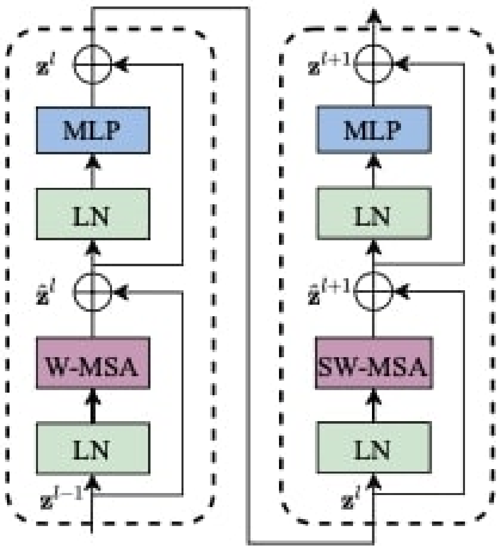

Figure 4.9: Illustration of the Swin Transformer Block, detailing the sequential flow from input to output (Source: Liu et al. [2021])

.

as it's processed by these layers, starting with the input features, moving through the specialised W-MSA and SW-MSA modules, and ending with the output features ready for the next steps in image analysis.

4.4 Multimodal Models

4.4.1 Contrastive Language-Image Pre-training (CLIP)

Introduced by OpenAI, CLIP is an open-source multimodal model that unifies the capabilities of NLP and computer vision. It is designed for self-supervised learning of image representations, utilizing a dual-encoder framework to interpret a combination of textual and visual information.

The pre-training process, shown in Figure 4.10, is critical for CLIP's learning capabilities. Jointly training an Image Encoder and a Text Encoder, CLIP adopts a contrastive learning approach to generate image embeddings $[I_1, I_2, \ldots, I_N]$ using either a ResNet or ViT, and text embeddings $[T_1, T_2, \ldots, T_N]$ using a Transformer model with GPT2-style modifications. Each encoder outputs an $N \times d_e$ matrix, where d_e is the size of the latent dimension. In this contrastive learning framework, the model's goal is to maximize the diagonal elements of the resulting $N \times N$ similarity matrix, corresponding to correct

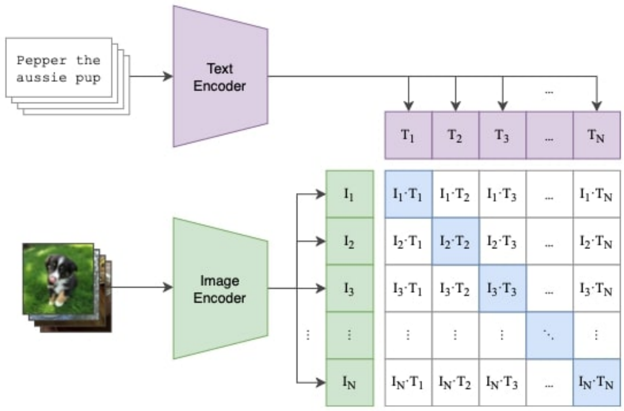

Figure 4.10: The Contrastive Pre-training step of CLIP. (Source: Radford et al. [2021])

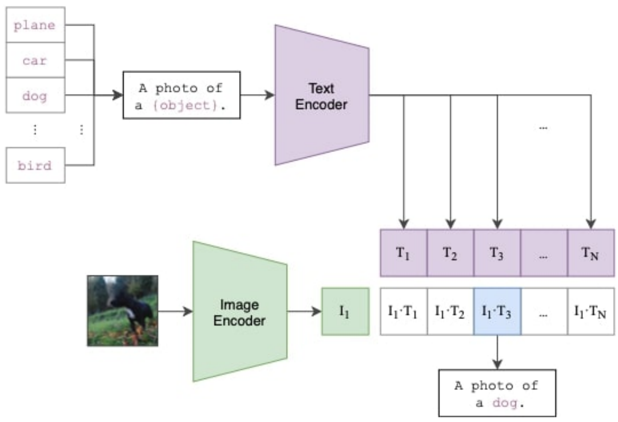

Figure 4.11: Zero-shot classification using CLIP. (Source: Radford et al. [2021])

image-text pairs, and minimize the off-diagonal elements, indicating dissimilar pairs.

After this pre-training, CLIP can perform Zero-shot Classification, as shown in Figure 4.11. Unlike traditional methods, it does not require further fine-tuning for classification tasks. Descriptive text prompts and images are encoded into embeddings, and the highest cosine similarity between them determines the image's label.

CLIP utilizes the WebImageText (WIT) dataset for training, which consists of 400 million image-text pairs. This was created by OpenAI to overcome the limitations of smaller datasets. It serves as the seed for the model's learning approach, allowing it to learn from a variety of visual concepts combined with natural language descriptions.

As reported by Radford et al. [2021], CLIP demonstrates remarkable Zero-shot transfer performance on various vision datasets, competing with fully supervised baselines such as ResNet. The model performs well in image recognition, as well as in learning features useful for a variety of downstream tasks. However, the authors acknowledge that achieving optimal results with Zero-shot learning in CLIP requires massive computational power. Furthermore, while the rich dataset WIT contributes to the model's diverse learning experience, it does not inherently improve the model's efficiency with respect to data. Therefore, CLIP's accuracy is supported by the volume of data rather than an enhanced ability to learn from limited information.

4.4.2 Large Language and Vision Assistant (LLaVA)

LLaVA is an open-source multimodal model that combines an LLM with state-of-the-art vision capabilities. It represents a significant advancement in multimodal conversational AI, integrating a vision encoder with a language model to achieve new levels of visual and linguistic comprehension. Parallel to GPT-4V's capabilities, LLaVA demonstrates remarkable flexibility in handling multimodal chat interactions. It responds to a variety of images and instructions, including those never encountered before.

Data Liu et al. [2023b] have identified a significant gap in multimodal AI research, which is the lack of high-quality datasets where text specifically serves as an "instruction" for an image. To address this issue, they created their own dataset by taking advantage of GPT-4's ability to convert existing image-text pairs from the COCO [3] dataset into instruction-based data, without presenting the image directly to GPT-4. This process begins with an image X_v and its corresponding caption X_c, from which a series of questions X_q are formulated. These questions are designed to guide an AI assistant in effectively describing the image's content. Initially, this approach produced data that was functional but lacked diversity and depth in its instructional scope. To enhance the dataset, the methodology was further refined to take full advantage of GPT-4's advanced capabilities to create more complex instruction sets based on the image. To do this, the image is represented in two ways: cap-

[3] `https://cocodataset.org/`

tions, which provide different textual interpretations of the visual scene, and bounding boxes, which precisely identify and locate objects within the image, an example is shown in Figure 4.12. This dual-representation improves the quality of the data and allows for more complex interactions between the visual and textual elements. The enriched dataset, which consists of 158,000 unique sets, contains three types of instruction-following data:

- **Conversation**: This involves question-and-answer dialogues about the visual content, focusing on object identification, quantification, and spatial location.
- **Detailed Description**: In-depth descriptions are generated by asking detailed, tailored questions to GPT-4.
- **Complex Reasoning**: These are questions requiring logical, step-by-step reasoning for accurate answers.

Architecture The architecture of LLaVA is designed to combine the functionalities of a pre-trained LLM and a visual model so that they can work together. The language processing unit is based on Llama, which serves as the LLM. For visual processing, when an image X_v is received, the model utilizes a pre-trained version of CLIP (ViT-L/14) to extract the visual features, which are denoted as $Z_v = g(X_v)$.

The main idea of the architecture is to connect the visual and linguistic domains. A linear layer, conceivable as a trainable projection, is employed to transform Z_v into a series of visual tokens H_v that are compatible with the language model's understanding. This is achieved by applying a projection matrix W to the visual features, generating language embedding tokens according to $H_v = W \cdot Z_v$. These visual tokens have the same dimensionality as the word embedding space within the Llama model, facilitating an effective integration of visual information into the language processing workflow. As a result of this transformation, the language model is provided with a sequence of visual tokens H_v, which are combined with regular language instructions H_q to generate responses that are informed by both the textual context and the visual input. This integrated process is visually represented in Figure 4.13, which illustrates the flow from image input to language output, encapsulating the integration between the LLM and visual data within LLaVA's architecture.

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Figure 4.12: Generation of LLaVA instruction-following data using GPT-4 by presenting the caption and bounding boxes. (Source: Liu et al. [2023b])

Training The objective of training LLaVA was to enable the model to engage in conversations about images in the same way as chatting. The training procedure involved exposing the model to simulated chats, with a clear end-of-chat signal `<STOP>` to indicate the end of each message. This effectively taught the model the flow of dialogue. The model's learning process was centered around customized sets of questions and answers for each image. These sets were presented sequentially, starting with the first question after the image, followed by the remaining questions one by one. The training unfolded in two distinct stages:

- **Feature Alignment Pre-training**: Aiming to balance concept coverage and training efficiency, this stage used 595K image-text pairs from the

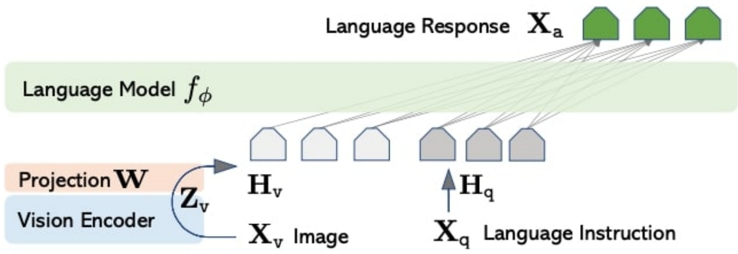

Figure 4.13: The LLaVA Architecture. (Source: Liu et al. [2023b])

CC3M [4] dataset, which is a collection of image and caption pairs. Each data point was adapted for instruction-following tasks and treated as a single conversational turn. The weights of the visual encoder and LLM remained fixed, and the focus was on adjusting the projection matrix to align image features with the LLM's pre-trained word embeddings.

- **End-to-End Fine-tuning**: This stage was designed to improve the model's capabilities through careful training, while retaining the visual encoder weights and optimizing both the projection layer and the LLM weights within the LLaVA framework. The model was fine-tuned using the custom dataset generated, which included response scenarios such as a multimodal chatbot and analysis of scientific QA [5] dataset, thereby fostering the model's ability to process and reason with both text and visuals to derive responses.

Liu et al. [2023b] demonstrate the capabilities of LLaVA in understanding and responding to images within a chatbot context. They achieved comparable results to multimodal GPT-4 despite using a smaller training dataset. The COCO validation split was used for a quantitative evaluation, which demonstrated LLaVA's enhanced instruction-following capabilities.

LLaVA 1.5 Building on the original LLaVA model, Liu et al. [2023a] introduced an enhanced version called LLaVA 1.5. They implemented strategic enhancements to refine the system's visual reasoning and multimodal interaction capabilities. These developments were applied systematically across multiple domains:

1. **Response Formatting**: The initial version encountered difficulties generating excessively verbose responses due to unspecific prompts. LLaVA

[4] `https://github.com/google-research-datasets/conceptual-captions/`
[5] `https://scienceqa.github.io/`

1.5 resolved this issue by utilizing more specific prompts that instruct the model to restrict its responses to a single word or phrase, resulting in a significant improvement in conciseness.

2. **Architectural Advancement**: A key improvement was the replacement of simple linear projection with a two-layer MLP as the vision-language connector. This architectural advancement has enabled LLaVA 1.5 to process multimodal data more effectively, resulting in enhanced performance across various tasks Liu et al. [2023a].

3. **Scaling Improvements**:

   - *Enhanced Image Resolution*: By increasing the resolution of visual inputs to 336, LLaVA 1.5 enables a clearer and more detailed understanding of images, capturing finer details with greater accuracy.
   - *Expanded Visual Dataset*: The GQA [6] dataset has expanded the model's training with a wider range of visual questions, scene graphs, and relational data, enhancing its visual comprehension.
   - *Incorporation of Conversational Data*: The ShareGPT dataset has added a rich set of conversational contexts to the model, diversifying its linguistic training and enhancing its natural language processing capabilities.
   - *Enlarged Model Capacity*: By replacing the LLM they used with the Vicuna-13B (which is also a version of Llama), the model has improved its processing and text generation capabilities.

A comprehensive overview of the models performance in comparison to its competitors is provided in Figure 4.14, which shows its capabilities on key benchmarks in the multimodal domain. It shows the progress made in the development of LLaVA 1.5, highlighting its efficiency and quality of results, even when compared to larger, more data-intensive models.

4.4.3 Late Fusion Model

Building upon the foundational principles outlined in the background section on multimodality in 3.1, this thesis proposes a late fusion model designed for the joint classification of images and text. The model aims to utilize the strengths of both visual and textual elements to potentially improve classification accuracy. The decision to use a late fusion approach was primarily driven by its flexibility, which allows for independent and specialized processing of

[6] `https://cs.stanford.edu/people/dorarad/gqa/`

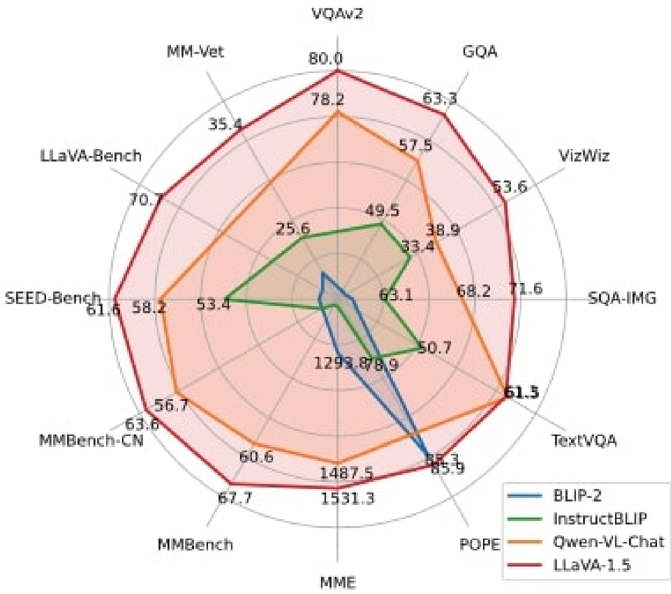

Figure 4.14: Comparative analysis of the performance of LLaVA 1.5 on various multimodal evaluation tasks, alongside modern multimodal models. (Source: Liu et al. [2023a])

each data modality prior to integration. This methodological choice is beneficial when dealing with complex and diverse data types such as social media content, where the interplay between visual elements and textual narratives is both rich and variable.

The model employs a two-step process to extract the learned representation of each data type. A ViT model extracts a 768-length visual feature vector from its last hidden layer for visual content. Similarly, a BERT model is utilized to extract a 768-length feature vector from its final hidden layer for textual content. This parallel processing of each modality ensures that the model effectively captures the complex details and specificities inherent in both visual and textual data. The feature vectors from both ViT and BERT are then concatenated to form a unified feature vector of length 1536.

Finally, a linear layer acts as the final classifier, processing this combined feature vector, capturing the rich, cross-dimensional insights it provides, and using it to predict the class of each data point. Figure 4.15 demonstrates the architecture of this model.

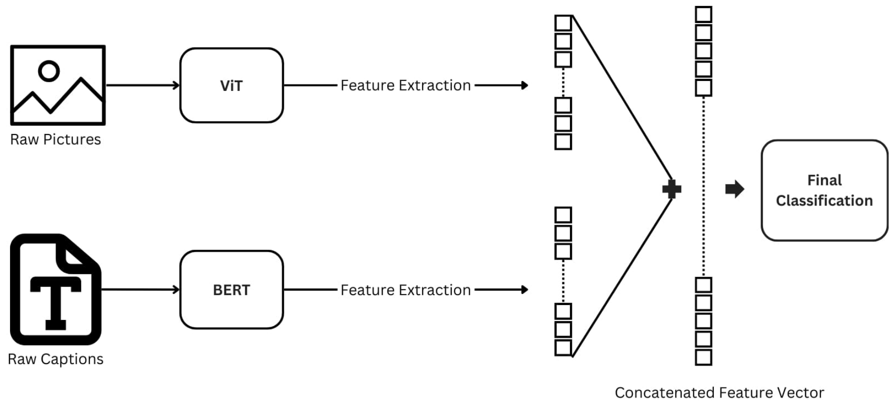

Figure 4.15: The late fusion model architecture.

4.5 Transfer Learning

In the traditional machine learning paradigm, as illustrated in Figure 4.16 (a), each task requires a separate learning system that is independently trained from scratch on task-specific data. Although this approach is straightforward, it often requires a large amount of labelled data and significant computational resources for each new task. In contrast, transfer learning, depicted in Figure 4.16 (b), seeks to optimise this process by transferring the knowledge gained from learning from one source task to improve the learning system for a new target task. This approach dynamically enriches a single learning system with insights from previous, related tasks, enabling it to adapt more efficiently to the new task with potentially limited data.

For example, in the case of a neural network model trained to recognise human faces for security purposes, the model has learned to recognise different facial features, including eyes, noses and mouths, as well as the correlations between them. If the goal is to create a system that identifies specific facial expressions, transfer learning can be used instead of training a model from scratch. By using a smaller dataset that focuses specifically on expressions, the model's pre-existing knowledge of facial features can be fine-tuned. This approach avoids the need for the model to relearn the fundamentals of facial structure and instead allows it to concentrate on the finer details of expressions. As a result, the training process for the new task is significantly accelerated.

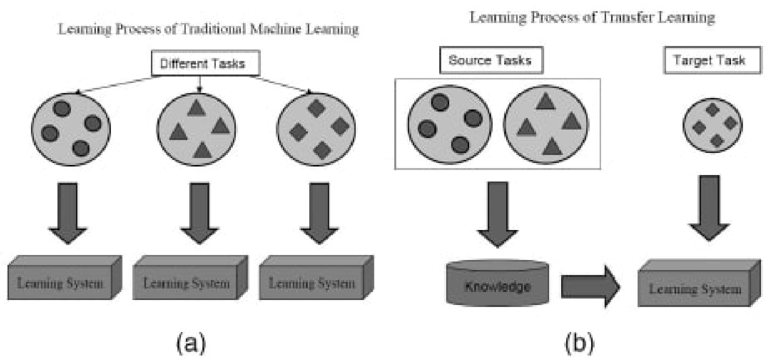

Figure 4.16: A visual comparison between the traditional machine learning and transfer learning methodologies. (Source: Pan and Yang [2010])

A formal definition of transfer learning is presented by Pan and Yang [2010], who describe it as the process in which knowledge from a source domain D_S and task T_S is applied to a different, yet related, target domain D_T and task T_T. Specifically, they define a domain as consisting of a feature space X and a marginal probability distribution $P(X)$, where X includes feature vectors $x_1, x_2, \ldots, x_n$. Correspondingly, a task is characterized by a label space Y and a conditional probability distribution $P(Y|X)$. The goal of transfer learning is then to improve the predictive performance in the target task by leveraging the information learned in the source domain and task, with the condition that $D_S \neq D_T$ or $T_S \neq T_T$. This process enables the use of a pre-trained model on the target task, potentially reducing the need for a large labeled dataset in the target domain.

One commonly used technique in transfer learning is *Fine-Tuning*. This process involves modifying the parameters of a pre-trained model to adapt it to a new, related task. As a result, the model's existing knowledge is customized to the particular characteristics of the new task, making it more effective and efficient in meeting the requirements of the target task. In this thesis, the fine-tuning approach has been used for both the vision and text models previously discussed in this chapter. The vision models, which were originally trained on extensive datasets such as ImageNet, have been further refined to classify images into categories of "norm-beauty" or "divers". Similarly, in text analysis, BERT, which has been pre-trained on large corpora of text, is fine-tuned using the captions associated with these images to assist the classification process.

Chapter 5

Evaluation

5.1 Overview

This chapter provides a comprehensive evaluation of the classification results, starting with images, then text, and finally multimodal classification. In order to ensure a broad understanding of the key elements underlying this evaluation, it is essential to outline the computational environment, the data split used to train the models, and the metrics used to assess their performance.

5.1.1 Computational Environment

The main programming language used in this thesis was *Python*, favored for its robust libraries and strong community support in data science. All computational tasks were performed using *Google Colab Pro*, leveraging the processing power of the *NVIDIA Tesla V100 16GB GPU* for tasks requiring intensive graphical processing. This was essential for the efficient computation required to train and evaluate advanced deep learning models. For the storage and management of datasets, models, and results, *Google Drive* was utilized, ensuring accessible data management during development.

5.1.2 Data Split

The dataset was divided into training, validation, and test sets with proportions of *80%*, *10%*, and *10%*, respectively. These splits were utilized for all classification tasks, ensuring a consistent and unbiased evaluation framework for model performance in image, text, and multimodal classification.

5.1.3 Metrics

This thesis primarily focuses on accuracy, F1 score, and the balance between training and validation loss, each of which provides unique insights into the behaviour of the models.

Accuracy This is the main indicator of model performance in data classification. It is defined as the ratio of correctly predicted observations to the total observations:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

In this thesis, high accuracy is emphasized as the main goal, particularly in the context of classifying Instagram posts into 'norm-beauty' and 'divers' categories. It is crucial as it reflects the model's effectiveness in accurately discerning the nuanced differences between these categories. A model with high accuracy ensures accurate categorisation, which is essential for applications such as content filtering or cultural trend analysis, where it is vital to accurately identify subtle differences in the data.

Precision, Recall, and F1 Score While accuracy provides an overall measure of effectiveness, precision and recall offer more insights. Precision is the ratio of true positive predictions to all positive predictions made, while recall measures the ratio of true positive predictions to the total number of actual positives in the dataset:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1 Score is the harmonic mean of Precision and Recall, providing a balance between the two by considering both false positives and false negatives:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

It ensures a more balanced evaluation of model performance, especially in contexts where precision and recall are both crucial.

Balance between Training and Validation Loss Monitoring the balance between training and validation loss during training is important for evaluating a model's ability to generalize. A small gap indicates good generalisation, suggesting the model is learning patterns relevant to unseen data. A large gap, however, may indicate overfitting. This balance serves as a vital indicator of model robustness and predictive power.

5.2 Image classification

The exploration of multimodal social media posts begins with the analysis of the visual content. Given the inherently rich, contextual information that images can provide at a glance, image classification is the first cornerstone of this investigation. The goal here is to distinguish between "norm-beauty" and "divers" posts based solely on the visual narratives presented in the images.

5.2.1 Model Selection

To address the complex nature of visual data, five advanced models were used: ViT_Base, Swin_Base, ResNet-50, SEER, which will be referred to as the vision models, and CLIP. The first four models were selected for their proven performance in image classification tasks across a wide range of domains, and were *fine-tuned* to meet the specific requirements of the given dataset.

The CLIP model has demonstrated promising capabilities in Zero-shot classification, using large image-text pairs to enable robust learning of visual representations. As pointed out by Zhang et al. [2022a], this innovative approach can effectively transfer knowledge across different tasks without additional training. Therefore, this model was chosen to classify images in a Zero-shot way. This is done by embedding images and their corresponding labels in a shared space, and then by assessing the similarity between these embeddings, CLIP can classify images without the need for model fine-tuning. The incorporation of CLIP into the model ensemble brings a new perspective to the analysis, allowing for the assessment of Zero-shot learning efficacy within the context of the given image classification task.

5.2.2 Experimental Setup

The experimental setup was designed to evaluate and compare the performance of the selected vision models under different conditions. Specifically, image resolutions were varied (224×224 and 384×384) to examine how increased resolution affects model performance and the models' ability to capture and utilize detailed visual information. Data augmentation was also employed as a variable in the experiment to evaluate its impact on the models' performance. By analyzing the results with and without data augmentation, it is possible to gain a better understanding of the potential benefits of augmentation in enhancing the models' generalization capabilities.

Training Configuration Initial manual experimentation was conducted with a range of hyperparameters, including the following epochs $[2, 3, 4, 5]$, and learning rates $[1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}]$. After preliminary investigations, it was found that the following set of hyperparameters could serve as a good starting point for all vision models:

- **Epochs:** 4
- **Batch Size:** 16
- **Learning Rate:** 5×10^{-5}

For the CLIP model, the image resolution was preset to 336×336, in line with the model's default configuration.

5.2.3 Results and Analysis

The performance of the image classification models is comprehensively presented in Table 5.1, where bold values indicate the best performance within each model family.

The SEER models demonstrates enhanced performance, particularly at higher resolutions, signifying its capability to effectively process increased image detail. However, a noticeable difference between training and validation loss when training on augmented data suggests a tendency to overfit. Despite this, the SEER models achieve high accuracy and F1 scores on the test data, indicating a degree of generalizability. SEER_384 trained on raw data emerges as the most effective model in this family.

Within the ResNet family, the dynamics between resolution, data augmentation, and loss metrics present a complex scenario. Higher resolutions do not consistently improve performance; in particular, ResNet_384 trained on raw data does not outperform its lower-resolution counterpart, ResNet_224, in terms of test accuracy or F1 scores. Data augmentation has been shown to improve performance metrics. This indicates their ability to extract more diverse and generalizable features from enriched datasets. While the narrow loss margins in the augmented version of ResNet_384 suggest good generalization capabilities, it is possible that these models may not be fully exploiting their learning capabilities. The findings suggest that while ResNet architectures inherently benefit from enriched data, unlocking their true potential may require further fine-tuning of the model or varying the training data to ensure that they are challenged enough to learn more nuanced and discriminative features. The most suitable performing model in this case would be the ResNet_384 trained on augmented data.

The ViT models demonstrate a good ability to classify the images, benefiting from the higher resolution, as evidenced by the better performance of ViT_384 over ViT_224 in terms of test accuracy and F1 scores. While the data augmentation improves validation accuracy, it also increases the gap between training and validation loss, suggesting a potential for overfitting. This could potentially be improved with refined regularization or learning rate adjustments to ensure that these models capture generalizable patterns and maintain robust performance across different datasets. However, the ViT models maintain strong performance, particularly in higher resolution setups, with the Vit_384 trained on raw data being the strongest model.

The Swin Transformer models exhibit similar performance trends to the ViT family, taking advantage of higher image resolutions to improve their classification capabilities. In particular, the Swin_384 model, especially when trained on raw data, exhibits a very good balance between training and validation loss, underscoring its robust ability to generalize. Similar to the pattern seen with ViT models, data augmentation results in a larger gap between training and validation loss, suggesting a potential overfitting problem. Overall, the Swin family is comparable to the ViT models in its ability to handle complex visual data.

The CLIP model, despite its innovative Zero-shot learning approach, demonstrates modest performance on the complex task of classifying social media posts into "norm-beauty" and "divers" categories, as evidenced by a test accuracy of **0.49** and an F1 score of **0.62**. This result underscores the complex nature of the task, which goes beyond simple image classification into the realm of subjective interpretation and contextual understanding. CLIP's method, which is primarily designed to correlate images with text, may not fully capture the subtleties required for this particular challenge. It suggests that the difficulty of the task may exceed the model's Zero-shot capabilities,

Model	DS	T Loss	V Loss	V Acc	Test Acc	Test F1
SEER_224	raw	0.20	0.47	0.80	0.78	0.81
SEER_224	aug	0.10	0.48	0.87	0.84	0.84
SEER_384	raw	0.18	0.29	0.87	**0.87**	**0.87**
SEER_384	aug	0.09	0.47	0.86	0.86	0.86
ResNet_224	raw	0.67	0.69	0.58	0.65	0.73
ResNet_224	aug	0.46	0.50	0.75	0.77	0.78
ResNet_384	raw	0.68	0.68	0.65	0.59	0.58
ResNet_384	aug	0.41	0.46	0.81	**0.78**	**0.79**
ViT_224	raw	0.20	0.58	0.85	0.75	0.78
ViT_224	aug	0.09	0.62	0.78	0.73	0.76
ViT_384	raw	0.14	0.24	0.89	**0.84**	**0.84**
ViT_384	aug	0.05	0.31	0.91	0.84	0.84
Swin_224	raw	0.20	0.23	0.92	0.80	0.83
Swin_224	aug	0.08	0.48	0.87	0.84	0.85
Swin_384	raw	0.18	0.14	0.94	**0.84**	**0.85**
Swin_384	aug	0.05	0.31	0.91	0.84	0.84
Clip_336	raw	-	-	-	0.49	0.62

Table 5.1: Performance metrics of Image Classification models on different datasets.

suggesting the need for more refined approaches, possibly incorporating richer contextual insights or task-specific fine-tuning, to adeptly navigate the complicated landscape of social media post classification.

5.2.4 Discussion and Implications

The exploration of classifying Instagram posts based on visual input alone has yielded promising results. The performance of fine-tuning the vision models confirms the viability of this visual-centric approach. In particular, models such as SEER and Swin Transformers at higher resolutions confirm that visual content is a powerful medium for distinguishing and categorizing images within the realm of aesthetics and beauty. The impact of image resolution on classification results has been shown to be significant in research. For example, Wollek et al. [2023] in their study of chest X-rays, and Cerit et al. [2016] in their study of automatic gender and age classification of faces, both show that higher resolution improves the model's ability to detect fine-grained details, as well as positively affects the accuracy of the model. Overall, these results represent the depth and richness that visual narratives can provide. They highlight the potential of images, when interpreted by sophisticated models, to reveal layers of context and meaning.

However, challenges and complications have been encountered in this process. The tendency of models to overfit when trained with augmented data is a reminder of the sensitivity required in model training. Although data augmentation introduces diversity, it also requires careful balancing to prevent models from learning features that are too specific to the training set. Cao et al. [2022] highlighted the significance of dataset curation and diversity in data augmentation to address unfair behavior and overfitting in deep learning models. Their research underscores the importance of using a broader and more varied datasets, and highlights how balanced feature development and distribution-aware augmentation can improve fairness and increase diversity, potentially mitigating the problem of overfitting and leading to models that truly understand underlying patterns rather than dataset-specific anomalies.

5.3 Text classification

This section discusses the text classification part of this thesis, which is important for understanding the textual narratives within the social media posts. By analyzing the text in addition to the images, a better insight into the data can be gained.

5.3.1 Model Selection

The choice to utilize BERT as the underlying model for text classification was motivated by its widespread availability, open source nature, and simplicity of fine-tuning. The accessibility and adaptability of the model make it an appropriate candidate for advanced text analysis tasks, allowing it to be efficiently adapted to specific research needs.

To address the challenges posed by the dataset, a multilingual and uncased variant of BERT was selected. This variant effectively handles the diverse linguistic content of the dataset, ensuring comprehensive coverage and accurate analysis of captions in multiple languages without the need for language-specific pre-processing. Additionally, the 'uncased' option allows for case-insensitive text processing. This feature can be particularly helpful when analyzing social media text, where non-traditional capitalization is common. This allows the model to focus on semantic understanding rather than stylistic differences.

DS	T Loss	V Loss	V Acc	Test Acc	Test F1
captions	0.42	0.44	0.76	0.74	0.74
reformulated_captions	0.41	0.38	0.82	0.76	0.75
descriptions	0.40	0.34	0.86	0.79	0.82

Table 5.2: Text Classification Performance Metrics

5.3.2 Experimental Setup

The aim was to assess the model's ability for classifying the text presented in the captions, reformulated captions, and descriptions into 'norm-beauty' and 'divers'.

- **Captions:** The original text that accompanies Instagram posts, often informal, in various languages, and including emojis or hashtags.
- **Reformulated Captions:** As detailed in Section 4.1.4, the captions were transformed using GPT-4 to ensure language standardization, translating non-English captions to English and converting emoji and hashtags to natural language descriptions. The aim was to achieve coherent structures that preserve the original sentiment and context.
- **Descriptions:** These are detailed descriptions of the images, which have been transformed into text format using the LLaVA model. The descriptions include information on pose, body prominence, and skin appearance, among other features. The purpose of this transformation is to provide a complete and objective description of the visual data, which could be used to support the classification task.

Training Configuration The model was fine-tuned over 2 epochs with a learning rate of 1×10^{-5}, a configuration chosen to balance precision and generalization, minimizing overfitting while adapting to the classification task's complexities.

5.3.3 Results and Analysis

The results of the experiments conducted are summarized in Table 5.2, which compares the performance metrics across different text types.

The model's performance on original captions indicates its competence in dealing with raw, unmodified social media text, achieving a test accuracy and F1 score of 74%. The results indicate effective, but not optimal, classification capabilities in dealing with the informal and diverse nature of original captions.

The increase in accuracy and F1 score to 76% and 75% for the reformulated captions highlights the value of preprocessing the captions for clarity and coherence. This preprocessing seems to aid the model in navigating the text more effectively, leading to improved classification outcomes.

The descriptions deliver the best model performance, with a test accuracy of 79% and an F1 score of 82%, demonstrating the benefit of detailed, contextual text in improving classification accuracy. Additionally, the lower training and validation losses for descriptions indicate good model fit, proving the effectiveness of using enriched textual data to support the learning process of the model.

5.3.4 Discussion and Implications

The results of the text classification analysis reveal the ability of the BERT model to effectively process and classify the diverse textual content of social media posts. By examining the model's performance on different types of text, from original captions to reformulated captions and detailed descriptions, several observations can be derived. Interestingly, the model's relatively good performance on raw captions suggests an inherent ability to distinguish between "norm-beauty" and "divers" posts, even within the informal and heterogeneous nature of social media language. This indicates that BERT has a robust classification ability that goes beyond shallow textual features and exploits the underlying semantic contexts of the captions.

Further analysis of the results shows the benefits of preprocessing and contextual enrichment on model performance. By translating non-English captions, converting emojis and hashtags into descriptive language, and standardizing text across posts, the data becomes more consistent and accessible to the model. These steps make it more efficient to uncover the hidden meanings and contextual information embedded in the text, which is crucial for accurate classification. Moreover, the description dataset, which provides the most detailed and objective view of the visual content and yields the highest accuracy and F1 scores, suggests that providing BERT with rich, descriptive text allows for a more comprehensive understanding and classification of the posts.

Overall, as indicated by Sun et al. [2019], fine-tuning BERT on task-specific data supports the premise that such domain-centric approaches can significantly enhance the ability of the model to more closely match the specific characteristics of the dataset at hand, thereby optimizing the accuracy of the model's performance on the given task.

5.4 Multimodal classification

After examining the classification capabilities of models using only visual and textual data, the investigation moves to a more holistic approach using multimodal classification. This approach combines visual and textual knowledge to address the complexity of social media posts, with the goal of leveraging the combined power of these modalities. This section explores the performance of combining images and text by using LLaVA for Zero-shot classification, as well as the late fusion approach that combines ViT with BERT. In addition, to better evaluate the capabilities of LLaVA and discuss the effect of adding textual information, this section also includes an examination of LLaVA's ability to classify images independently.

5.4.1 Model Selection

LLaVA The *LLaVA 1.5* was selected due to its outstanding performance on various benchmarks, as demonstrated in Figure 4.14. It can be used as a Zero-shot classifier to process and integrate multimodal data, combining both visual and textual inputs, without the need for task-specific training. Furthermore, the model can be prompted using natural language. This allows the model's behaviour and focus to be refined and directed through prompt engineering, thereby enhancing the model's ability to recognise the subtle thematic and aesthetic differences that define the categories of interest, improving the model's classification capabilities and tailoring it to the specific needs of social media content analysis. The LLaVA version utilized in this thesis is a quantized variant of the 13 billion parameter model. Quantization is the process of compressing a model's weights, activations, and other parameters into a more computationally efficient format without significantly degrading its performance Liu et al. [2023c].

Late Fusion This model strategically employs ViT with 384 resolution for image processing and BERT for text processing of raw captions, taking advantage of their shared transformer architecture for seamless integration. This choice is supported by the similarity of their configurations, including the identical size of their last hidden output layer (768 dimensions), which simplifies the fusion process. This approach combines the strengths of both models to analyze the contributions of visual and textual data and address the complex interaction between these modalities in social media content.

5.4.2 Experimental Setup

LLaVA The experimentation with LLaVA involved a systematic process of prompt engineering, input variation, and the use of system prompts to refine the overall behavior of the model. Initially, the experimentation started with simple prompts such as *"classify the following image into one of the following categories: 'divers' or 'norm-beauty.'"*. This basic prompt aimed to assess the model's inherent ability to distinguish between the two broad categories with minimal input. This experiment resulted in an accuracy of less than 50%, highlighting the necessity for more advanced prompt design to effectively guide the model's classification process.

To improve the model's performance, the focus was narrowed to specifically identify whether an image conformed to "norm-beauty" standards. This simplification allowed the model to better recognize features associated with conventional beauty standards. As the experimentation progressed, the prompts were progressively enriched with detailed instructions targeting key image features such as the subject's pose, skin texture, clothing, and environmental context. This incremental refinement aimed to guide the model's analytical focus towards important visual elements that influence the perception of beauty within an image. Throughout the development process, performance was monitored using the test dataset to track improvements and adjustments over time. This continuous evaluation allowed for dynamic refinement of prompts, ensuring each iteration contributed meaningfully to enhancing the model's accuracy. The result of this process was the development of 2 comprehensive prompts, which are:

Prompt 1:
Analyze the person in the image. Provide a JSON response with the following fields:

```
{
  "pose_and_posture": "Description of their stance",
  "body_prominence": "How their body is displayed",
  "skin_appearance": "Details of skin texture and features
```

```
",
  "body_features": "Information on weight, slimness,
   muscularity, and facial characteristics",
  "disability_or_syndrome": "Indicators of any disabilities
    or syndromes",
  "aligns_with_beauty_standards": "boolean value (0 or 1)
   for whether the individual's appearance aligns with
   traditional beauty standards on social media",
  "explanation": "brief explanation for
   aligns_with_beauty_standards"
}
```

Prompt 2:
"Analyze the person in the image. Beyond the immediate visual elements, reflect on the cultural, social, or personal significance conveyed in the image and caption. Does the image challenge or conform to traditional beauty norms? How does the caption complement or contrast with the visual message? Provide a JSON response with the fields outlined below:"

```
{
  "pose_and_posture": "Describe the physical stance and any
    implied emotions or attitudes",
  "body_prominence": "Detail how the body is displayed,
   including context such as clothing and setting",
  "skin_appearance": "Note skin texture, features, makeup,
   tattoos, or other adornments",
  "body_features": "Assess weight, slimness, muscularity,
   and facial characteristics, focusing on conformity or
   divergence from beauty standards",
  "disability_or_syndrome": "Identify any disabilities or
   syndromes, focusing on representation",
  "aligns_with_beauty_standards": "Boolean (0 or 1)
   indicating if the appearance aligns with traditional
   social media beauty norms",
  "explanation": "A brief explanation for the '
   aligns_with_beauty_standards' decision, linking observed
    elements to beauty norms"
}
```

The experiments also varied the type of input, from purely visual to combinations of text and image, to observe the effects on classification results. By adding *"considering the Instagram caption written by the person who posted it: 'caption'"* to the prompts, the corresponding caption for each image was easy to incorporate into the prompt.

In addition to adjusting the content-specific prompt, a considerable amount of experimentation has been dedicated to customizing the system prompt. This prompt is designed to provide guidance on the expected behavior of the model in a general sense, influencing its approach to the task at hand. The prompt was customized for the specific use case of social media posts and included instructions for the model to prioritize certain aspects over others or to adopt a certain analytical perspective when processing the input. The default system prompt of the LLaVA model was:

Default LLaVA System Prompt:
"A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions."

In contrast, the system prompt designed for the specific use case at hand:

Desgined LLaVA System Prompt:
"In the role of a social scientist, you are tasked with classifying social media images into 'norm-beauty' and 'divers' categories. 'norm-beauty' images typically feature conventional beauty standards and poses, while 'divers' images represent a wider spectrum of human diversity, including unique physical traits and unconventional styles. Analyze these images critically, focusing on their representation of societal norms and diversity."

For all experiments, the configuration parameters were set with a *max_length* of 1024 and a *temperature* of 0.2. The *max_length* parameter defines the maximum number of tokens the model processes in a single prompt. The *temperature* parameter controls the randomness of the model's responses, with a lower value such as 0.2 resulting in more deterministic and consistent results.

Late Fusion For this model, the experiments focused on finding the optimal training parameters for the linear layer, which serves as the final classifier. A series of experiments were conducted to find the optimal combination of epoch and learning rate, exploring setups with 2, 3, and 4 epochs, along with learning rates of $[1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}]$. The goal of this exploration was to identify a training configuration that would allow the linear layer to effectively leverage the value of the combined data from ViT and BERT, thereby enhancing its ability to perform multimodal classification.

Two main approaches were explored, one using the fine-tuned versions of ViT and BERT, which had already learned a refined representation for the dataset, and the other using the base models of ViT and BERT. The optimal configuration was found to be:

- **Epochs:** 4
- **Batch Size:** 16
- **Learning Rate:** 5×10^{-5}

5.4.3 Results and Analysis

LLaVA Before discussing the results, an example will illustrate the response behavior of the LLaVA model. For demonstration purposes, the images with ID 4 and ID 6 from the group of representative generated images presented earlier in Figure 4.1 will be considered. The responses presented below are generated using the Prompt 2 and the default LLaVA system prompt.

ID 4:

```
{
  "pose_and_posture": "The person in the image is standing with a relaxed and confident posture. They are smiling and looking directly at the camera, which suggests a positive and approachable demeanor.",
  "body_prominence": "The person is wearing a bikini, which accentuates their body shape. The setting is a beach, which is a common location for showcasing swimwear and summer fashion.",
  "skin_appearance": ""The skin appears to be well-maintained and tanned, which is often associated with beauty standards in many cultures.",
  "body_features": "The person has a slim and toned physique, which aligns with traditional beauty norms that often prioritize thinness and muscularity.",
  "disability_or_syndrome":  "There are no visible disabilities or syndromes in the image.",
  "aligns_with_beauty_standards": 1,
  "explanation": "The image aligns with traditional beauty norms because the person is slim, toned, and has a well-maintained appearance. The confident pose and the beach setting further contribute to the perception of beauty
```

```
    and attractiveness."
}
```

ID 6:


```
{
  "pose_and_posture": The person in the image is standing
    with a straight posture, looking directly at the camera.
     The facial expression is neutral, and there is no overt
     emotion or attitude conveyed.",
  "body_prominence": "The body is displayed prominently,
    with the top half of the suit jacket open, revealing the
     chest area. The setting is a plain background, which
    puts the focus on the person.",
  "skin_appearance": "The skin texture appears to be smooth
    , and there are no visible tattoos or adornments. The
    person has a full beard and mustache, which are well-
    groomed.",
  "body_features": "The person appears to be of average
    weight, with a slim build. There are no visible muscular
     features, and the facial characteristics are typical of
     a middle-aged adult.",
  "disability_or_syndrome": "There are no indications of
    any disabilities or syndromes in the image.",
  "aligns_with_beauty_standards": 0,
  "explanation": "The image challenges traditional beauty
    norms in several ways. The person is shirtless, which is
     not a typical pose for a professional portrait. The
    beard and mustache are well-groomed, but the lack of
    visible muscularity and the open suit jacket suggest a
    departure from the idealized male physique often
    portrayed in media. The neutral expression and the
    absence of overt emotion or attitude also contribute to
    a non-traditional portrayal of beauty."
}
```


The performance of the LLaVA model on the dataset is summarized in two tables. The first table 5.3 presents the overall results of LLaVA's classification on the entire dataset. The second table 5.4 is intended to provide a more comparative perspective, detailing LLaVA's performance specifically on the test dataset.

In evaluating the performance of the LLaVA model across different experimental configurations, a systematic analysis reveals the differing effects of prompt detail, the integration of multimodal data (images and text), and the application of custom system prompts. This structured investigation aims to clarify the factors that contribute to the model's ability to correctly classify the data.

The initial set of experiments, focused on image-only analysis, illustrates the importance of prompt complexity. The use of a more general prompt (Prompt 1) resulted in the model achieving a moderate level of accuracy (0.72) and an F1 score (0.78). In contrast, the use of a more detailed prompt (Prompt 2), even in the absence of textual data, significantly improved the model's performance, as evidenced by the experiment "llava_2_img," which showed an increase in accuracy (0.79) and F1 score (0.81). This improvement underscores the assumption that a comprehensive prompt facilitates a more accurate interpretation by the model, thereby refining its classification capability.

Next, the analysis was extended to include textual data in addition to visual input, which had a noticeable impact on the model's performance. When captions were added to Prompt 1, the model showed a slight improvement in both accuracy (0.74) and F1 score (0.80). The addition of captions to Prompt 2 also resulted in a considerable increase in performance, achieving an accuracy of 0.81 and the highest F1 score of 0.86 in the "llava_2_img_caption" experiment. This demonstrates the positive effect of adding textual context as an extra input, which improves the model's ability to accurately recognize and classify content.

Further exploration of the role of custom system prompts alongside these configurations provides additional insight into the complex dynamics of model performance optimization. The use of custom prompts alongside the more general Prompt 1, particularly in image-only scenarios, showed an improve-

Experiment	Model Type	Total Acc	Total F1
llava_1_img	Vision	0.72	0.78
llava_1_img_cust	Vision	0.80	0.83
llava_1_img_caption	Vision + Text	0.74	0.80
llava_1_img_caption_cust	Vision + Text	0.82	0.84
llava_2_img	Vision	0.79	0.81
llava_2_img_cust	Vision	0.56	0.65
llava_2_img_caption	Vision + Text	0.81	0.86
llava_2_img_caption_cust	Vision + Text	0.78	0.78

Table 5.3: LLaVA model performance metrics on the full dataset

Experiment	**Model Type**	**Test Acc**	**Test F1**
llava_1_img	Vision	0.67	0.76
llava_1_img_cust	Vision	0.79	0.83
llava_1_img_caption	Vision + Text	0.73	0.79
llava_1_img_caption_cust	Vision + Text	0.83	0.85
llava_2_img	Vision	0.75	0.76
llava_2_img_cust	Vision	0.48	0.58
llava_2_img_caption	Vision + Text	0.82	0.82
llava_2_img_caption_cust	Vision + Text	0.85	0.85

Table 5.4: LLaVA model performance metrics on the test dataset

ment in model performance, with a notable increase in both accuracy and F1 score. This trend was similarly observed in multimodal scenarios involving images and text with Prompt 1, where the integration of custom prompts also led to performance improvements. Such observations underscore the potential of custom prompts to refine model focus and analytical precision across single and combined modalities.

However, extending this custom prompting approach to the more detailed Prompt 2 introduced unexpected complexities. Comparing custom prompts to both Prompt 1 and the more detailed Prompt 2 revealed a challenging balance between prompt specificity and the interpretive flexibility of the model. While it was expected that custom prompts would generally improve performance by sharpening the analytical focus of the model, the results of the "llava_2_img_cust" and "llava_2_img_caption_cust" experiments suggested otherwise. Specifically, in image-only scenarios using Prompt 2, the inclusion of custom system prompts unexpectedly led to a decrease in performance, suggesting that overly restrictive prompts may limit the model's ability to reason autonomously. Similarly, in multimodal scenarios where both images and captions were analyzed, the expected improvement from custom prompts did not appear consistently. This pattern highlights the importance of carefully balancing prompt specificity to preserve the model's inherent analytic capabilities, particularly when dealing with the complex interplay of visual and textual data.

This analysis emphasizes the significance of well-designed prompts in enhancing the classification accuracy of the LLaVA model. It also demonstrates that the integration of textual data alongside visual input, when carefully combined with thoughtfully designed prompts, can significantly enhance the model's ability to distinguish between complex social media content. The results support

Model	T Loss	V Loss	V Acc	T Acc	T F1
caption_vit384_bert_base	0.001	0.32	0.89	0.88	0.88
caption_vit384_bert_fine_tuned	0.02	0.36	0.86	0.92	0.92

Table 5.5: Comparative Performance of Late Fusion Model Using Base and Fine-Tuned Configurations

a balanced strategy in which custom system prompts are tailored to enhance the model's inherent strengths without overly restricting its analytic scope, thereby optimizing performance under different experimental conditions.

Late Fusion The results summarized in Table 5.5 aim to explore the effects of using base versus fine-tuned versions of the ViT and BERT models within the Late Fusion architecture. It seeks to demonstrate how these two configurations affect the model's effectiveness in classifying the data.

It is shown that both configurations have acceptable values for training and validation losses, indicating the ability to learn and generalize adequately. However, there is potential to optimize the difference between these losses to ensure a more balanced training-validation loss ratio, which could further improve model robustness and reduce the risk of overfitting.

Remarkably, the fine-tuned configuration shows a strong increase in test performance, with test accuracy and F1 score significantly better than the base configuration. This demonstrates the fine-tuned model's ability to better learn from and classify the data, and points to the benefits of using models that are optimized for the given data.

5.4.4 Discussion and Implications

The analysis of multimodal classification using the LLaVA and Late Fusion approaches demonstrates the benefits of integrating visual and textual data to solve complex challenges. The application of the LLaVA model, despite its general-purpose training, illustrates the potential of MLLM to adapt to complex classification tasks through innovative prompt engineering and Zero-shot learning capabilities. This adaptability is important because it reflects the model's ability to effectively handle data related to different topics, although refinements in prompt design are needed to fully exploit its capabilities.

An additional advantage of the LLaVA model is its ability to provide natural language explanations, which greatly enhances transparency and helps to interpret its reasoning process. This ability is instrumental in explaining the differences between the "norm-beauty" and "diverse" categories from it's point of view, providing insight into the model's decision criteria. Such explanations encourage deeper analytical understanding as well as increased confidence in the model's results. It is useful for researchers and practitioners as it clarifies the model's evaluation mechanisms and supports ongoing refinement of classification strategies to increase accuracy.

However, it is important to note that using the LLaVA model required additional data cleaning and processing steps. In particular, despite instructions for the model to maintain a JSON structure with specific fields such as "aligns-with-beauty-standards" filled with a boolean value of 0 or 1, there were instances where the model deviated from these expectations. Occasionally, the model would output a string or use TRUE/FALSE instead of the required numeric boolean values. In other cases, the model provided overly detailed descriptions of the image, resulting in nested dictionaries for what should have been simple fields within the JSON response. These inconsistencies, required a careful data cleaning approach to ensure the consistency and reliability of the model outputs. While this process was feasible, it emphasized the importance of closely monitoring and adjusting the dataflow to address the varying outputs generated by the model. The cases of non-compliance with the expected JSON structure were relatively few and could be efficiently resolved, ensuring that the integrity of the analysis remained unaffected.

In contrast, the Late Fusion model, which strategically combines the strengths of ViT and BERT, shows a significant improvement in classification performance over the use of either model alone. This improvement highlights the complementary nature of visual and textual information in providing a more complete understanding of the data. However, the analysis also identifies potential areas for improvement within the late fusion approach. For example, further experimentation could be conducted with different data fusion techniques instead of concatenation, such as weighted sum or applying PCA for dimensionality reduction prior to fusion. Another experimental approach could be to use MLPs rather than a liner layer within the model. These strategies could refine the model's ability to extract and interpret the critical features from each modality.

Chapter 6

Conclusion

This thesis has taken a step forward in exploring the digital realm of social media, with a particular focus on Instagram, to analyze and categorize representations related to concepts such as beauty and diversity. The quest has focused on exploring the capabilities of various computational approaches in identifying complex and subjective perspectives on beauty in social media content.

The investigation began with an in-depth analysis of images, recognizing them as the primary means of conveying beauty standards on social media platforms. Advanced image processing models, such as CLIP, ViT, Swin Transformer, ResNet, and the SEER model, were utilized to assess their capacity to differentiate and categorize visual representations related to beauty and diversity. This visual-centric approach laid the groundwork for a more comprehensive analysis, which was gradually expanded to consider the textual narratives accompanying the images, using the BERT model to capture the relationship between these rich textual expressions within social media posts.

Building on the insights gained from the visual and textual evaluations, the thesis moved into the realm of multimodal classification. In this advanced stage of the analysis, the LLaVA model proved to be a major success, demonstrating its ability to combine and interpret both visual and textual data. LLaVA's approach was innovative and represented a new step forward, with the results underscoring the model's ability to deal with challenging tasks and demonstrating the potential of MLLMs to handle complex classifications with high accuracy. The importance of prompt engineering was also highlighted, showing how the use of targeted prompts can significantly refine the model's output, ensuring more accurate and context-aware classifications.

Models such as LLaVA are being adapted as Zero-shot classifiers, as in the work of Islam et al. [2024]. In their study, they used the LLaVA model to evaluate its classification capabilities on different datasets: MNIST, cats vs. dogs, Hymenoptera (ants vs. bees), and an innovative set focused on pox vs. non-pox skin images. They also used customized prompts for zero-shot learning, demonstrating LLaVA's efficiency in accurately classifying images without the need for prior fine-tuning. The model achieved high accuracies of 85%, 100%, 77%, and 79% across these datasets, respectively. Furthermore, their research highlighted the adaptability of the model through fine-tuning on a specialized task involving the identification of autism in children based on facial images. Prior to fine-tuning, LLaVA's test accuracy was 55%, which significantly increased to 83% after fine-tuning. This work by Islam et al. [2024] emphasizes the significant potential of LLaVA and similar models to revolutionize various application areas within the field of AI, and identifies them as instrumental in advancing Zero-shot classification tasks.

Alongside LLaVA, the Late Fusion model emerged as an auxiliary but essential component of the multimodal analysis framework. This model strategically combined the strengths of ViT and BERT through a late fusion approach and emphasized the potential synergy of merging visual and textual modalities. Although LLaVA remained the focus, the Late Fusion model provided a complementary perspective that enriched the multimodal classification task with its unique insights. This model provided the best results among the tested approaches, demonstrating that models specifically trained on the given dataset can still outperform large, general-purpose models. This finding indicates the value of custom model training and reinforces the benefit of customizing models to fit the specific characteristics and challenges of the data being analyzed.

The systematic methodology of this thesis, beginning with image analysis, moving to textual review, and ending with multimodal classification, successfully revealed the complexity of the dataset. Through this approach, the unique contributions of each data modality to the final analysis were highlighted, thereby indicating the vital role of both images and text in representing beauty and diversity on social media. Through this sequential investigation, the importance of integrating these modalities to fully understand social media narratives became apparent, emphasizing the efficacy of a multimodal strategy.

Contributions to the Field of Data Science This thesis makes a meaningful contribution to the field of data science, specifically in the domain of social media content analysis and the exploration of beauty concepts. Through an innovative application of advanced deep learning models, this work has successfully navigated the difficult dynamics of beauty and diversity as represented on social media platforms, offering new perspectives and methodologies for understanding these subjective constructs. The main contributions of this research are:

- **Innovative Application of Advanced Deep Learning Models:** The application of state-of-the-art models such as ViT, Swin Transformer, SEER, as well as the novel use of the LLaVA model for Zero-shot classification, underscores the potential of deep learning to analyze the complex nature of the visual and textual narratives in social media. Overall, this thesis demonstrates the feasibility of using these technologies to understand highly nuanced concepts, and sets a strong direction for future research in this area.

- **Development of a Multimodal Analysis Framework:** By integrating visual and textual data analysis and utilizing multimodal models such as LLaVA and the Late Fusion Framework, this thesis introduces a comprehensive workflow for analyzing social media content. The presented methodology expands the scope of beauty analysis and improves the understanding of social media content and provides a valuable resource for future studies in similar domains.

- **Advancement in Prompt Engineering:** The exploration of prompt engineering with the LLaVA model highlights the importance of carefully crafted prompts in refining the output of deep learning models. This thesis adds to the developing field of prompt engineering by showcasing its efficacy in directing the model's focus, thereby improving classification accuracy and providing more context-aware analysis.

Reflection on Limitations While considering the findings of the research, it is important to acknowledge the limitations encountered along the path. One of the primary challenges is the subjective nature of classifying beauty and diversity. From an anthropological angle, defining clear criteria for both categories is particularly challenging because cultural constructs of beauty and diversity are inherently fluid and vary significantly across different social and cultural environments. The dataset used in this thesis reflects a narrow perspective based on annotations provided by individuals, which, while valuable, does not accurately represent the full spectrum of societal or individual perceptions of beauty. The methodological approach, which uses machine learning techniques to replicate the perspectives of the participants, further challenges the subjective viewpoint. The results of the classification models are representations of representations, adding layers to the subjective nature of the task.

Another limitation of the dataset is related to its diversity, size, and annotation reliability. The 'norm-beauty' category was more straightforward to distinguish compared to the 'divers' category, which encompassed a wide range of diversity. This highlights the challenge of capturing the full spectrum of human diversity in a comprehensive and accurate way. Furthermore, ensuring annotation reliability poses significant challenges due to the subjective nature of beauty and diversity, which can result in varying interpretations among annotators. To address these limitations, a more comprehensive dataset is necessary, incorporating diverse perspectives and utilizing robust annotation methods to increase the reliability and representativeness of the data.

In reflecting on these limitations, it becomes clear that while this thesis has made progress in exploring the digital representation of beauty and diversity, there remains a broad landscape of complexity that has yet to be fully understood. The challenge of subjective classification, along with the constraints related to the dataset's diversity and annotation reliability, presents opportunities for further exploration and refinement in future research.

Directions for Future Work The research findings have identified new directions for future work in the field of analyzing social media content in relation to beauty. The following outlines potential research paths for further exploring the implications and complexities related to this domain:

1. **Extensive analysis of LLaVA model outputs:** The LLaVA model's explanations for correctly classified data points provide valuable insights. A more detailed analysis, possibly using advanced topic modeling techniques, could reveal the underlying themes or hidden narratives within these explanations. This approach would enhance the interpretation of the model's mechanisms, as well as provide a deeper look into the cultural and societal signals that define beauty and diversity on social media platforms.

2. **Experimentation with Various Prompting Techniques:** Prompt engineering is key to improving the performance of models such as LLaVA. Future research could explore various prompt engineering techniques, such as few-shot or chain of thought, to determine the most effective methods for guiding models in challenging tasks. By systematically testing and refining different prompting approaches, new levels of precision

and context awareness in model output could be achieved, leading to more accurate and sophisticated analysis of social media content.

3. **Exploration of LLaVA 1.6:** The release of LLaVA 1.6 introduces a number of improvements over its previous versions, including superior reasoning, optical character recognition (OCR), and world knowledge capabilities Liu et al. [2024]. These improvements could be beneficial in future research to address the challenges of classifying complex, multimodal social media content with greater accuracy and depth. By utilizing LLaVA's advanced features, it could be possible to make a significant improvement in the understanding of multifaceted representations of beauty and diversity, setting a new standard in the field.

4. **Creation of a More Reliable Dataset:** The foundation of any robust analysis lies in the quality of the dataset. It is recommended that future work prioritize the creation of larger and more diverse datasets that are carefully annotated through processes that ensure high reliability, such as inter-annotator agreement. Engaging annotators from a wide range of cultural backgrounds could enrich the dataset with diverse perspectives on beauty and diversity, thereby enhancing the depth and breadth of the analysis.

5. **Adoption of a Graded Scale for Annotation:** Transitioning to a graded scale for image annotation represents a significant methodological shift from binary classification. This approach acknowledges the subjective and spectrum-based nature of beauty and diversity, allowing for a more detailed and nuanced categorization. By capturing a wider range of perceptions, a graded scale could provide a richer, more dimensional understanding of social media aesthetics, facilitating more refined analyses.

Concluding Remarks This thesis represents an innovative advancement in integrating AI into the field of digital aesthetics, specifically addressing the challenge of uncovering subjective perspectives on beauty and diversity in social media. By comparatively evaluating the capabilities of various advanced AI models, it pushes the boundaries of traditional data science and AI, revealing new depths in the analysis of multimodal data.

Looking to the future, there is great potential for advanced AI models to evolve into powerful tools that provide users with insight into the content that populates their social media feeds. As research in this area progresses, AI could play a crucial role in promoting greater awareness among social media users, enabling them to critically evaluate and engage with the content they are exposed to. Such developments could significantly contribute to promoting a more inclusive, diverse, and conscious digital ecosystem. This could celebrate and understand the plurality of beauty standards and the richness of diversity in all its complexity.

Overall, this thesis stands as a demonstration of the power of AI and data science in uncovering the layers of social media aesthetics, marking a critical step toward a more accurate and comprehensive understanding of how beauty and diversity are curated, shared, and perceived in an increasingly interconnected digital world. Moving forward, the exploration and innovation at the intersection of AI, data science, and social media analysis will only continue to enrich our understanding of these essential aspects of the human experience, paving the way for a future where technology and human insight converge in the pursuit of a deeper understanding and appreciation of the diverse cultural landscape that defines us all.

Bibliography

Aisha Al-Sadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Fumie Costen. Visual question answering in the medical domain based on deep learning approaches: A comprehensive study. *Pattern Recognit. Lett.*, 150:57–75, 2021. doi: 10.1016/J.PATREC.2021.07.002. URL https://doi.org/10.1016/j.patrec.2021.07.002.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607. URL https://doi.org/10.1109/TPAMI.2018.2798607.

Matteo Bodini. Will the machine like your image? automatic assessment of beauty in images with machine learning techniques. *Inventions*, 4(3), 2019. doi: 10.3390/inventions4030034. URL https://www.mdpi.com/2411-5134/4/3/34.

Fares Bougourzi, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowl. Based Syst.*, 242:108246, 2022. doi: 10.1016/J.KNOSYS.2022.108246. URL https://doi.org/10.1016/j.knosys.2022.108246.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Yushi Cao, David Berend, Palina Tolmach, Guy Amit, Moshe Levy, Yang Liu, Asaf Shabtai, and Yuval Elovici. Fair and accurate age prediction using distribution aware data curation and augmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2867–2877. IEEE, 2022. doi: 10.1109/WACV51458.2022.00292. URL https://doi.org/10.1109/WACV51458.2022.00292.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020. URL https://arxiv.org/abs/2006.09882.

Betul Cerit, Salih Arda Boluk, and M. Fatih Demirci. Analysis of the effect of image resolution on automatic face gender and age classification. In *24th Signal Processing and Communication Application Conference, SIU 2016, Zonguldak, Turkey, May 16-19, 2016*, pages 853–856. IEEE, 2016. doi: 10.1109/SIU.2016.7495874. URL https://doi.org/10.1109/SIU.2016.7495874.

Mathieu Chambe, Rémi Cozot, and Olivier Le Meur. Deep learning for assessing the aesthetics of professional photographs. *Comput. Animat. Virtual Worlds*, 33(6), 2022. doi: 10.1002/CAV.2105. URL https://doi.org/10.1002/cav.2105.

Parag S. Chandakkar, Vijetha Gattupalli, and Baoxin Li. A computational approach to relative aesthetics. *CoRR*, abs/1704.01248, 2017. URL http://arxiv.org/abs/1704.01248.

Garima Choudhary and Tapan Kumar Gandhi. Indexing facial attractiveness and well beings using machine learning. In *2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 1–6, 2016. doi: 10.1109/R10-HTC.2016.7906813.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Michael R. Douglas. Large language models. *CoRR*, abs/2307.05782, 2023. doi: 10.48550/ARXIV.2307.05782. URL https://doi.org/10.48550/arXiv.2307.05782.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *CoRR*, abs/2202.08360, 2022. URL https://arxiv.org/abs/2202.08360.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Ashhadul Islam, Md. Rafiul Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Zubair Shah. Pushing boundaries: Exploring zero shot object classification with large multimodal models. *CoRR*, abs/2401.00127, 2024. doi: 10.48550/ARXIV.2401.00127. URL https://doi.org/10.48550/arXiv.2401.00127.

William C. Sleeman IV, Rishabh Kapoor, and Preetam Ghosh. Multimodal classification: Current landscape, taxonomy and future directions. *CoRR*, abs/2109.09020, 2021. URL https://arxiv.org/abs/2109.09020.

Gargi Joshi, Rahee Walambe, and Ketan Kotecha. A review on explainability in multimodal deep neural nets. *CoRR*, abs/2105.07878, 2021. URL https://arxiv.org/abs/2105.07878.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *CoRR*, abs/2307.10169, 2023. doi: 10.48550/ARXIV.2307.10169. URL https://doi.org/10.48550/arXiv.2307.10169.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *CoRR*, abs/2205.11916, 2022. doi: 10.48550/ARXIV.2205.11916. URL https://doi.org/10.48550/arXiv.2205.11916.

Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. *CoRR*, abs/1908.09317, 2019. URL http://arxiv.org/abs/1908.09317.

Martha Peaslee Levine and Júlia Scherer Santos. *Social Media and Its Effects on Beauty*. IntechOpen, London, United Kingdom, 2021.

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Shun Chen, Bin Liu, and Jianhua Tao. GPT-4V with emotion: A zero-shot benchmark for multimodal emotion understanding. *CoRR*, abs/2312.04293, 2023. doi: 10.48550/ARXIV.2312.04293. URL https://doi.org/10.48550/arXiv.2312.04293.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023a. doi: 10.48550/ARXIV.2310.03744. URL https://doi.org/10.48550/arXiv.2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023b. doi: 10.48550/ARXIV.2304.08485. URL https://doi.org/10.48550/arXiv.2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. URL https://arxiv.org/abs/2103.14030.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: data-free quantization aware training for large language models. *CoRR*, abs/2305.17888, 2023c. doi: 10.48550/ARXIV.2305.17888. URL https://doi.org/10.48550/arXiv.2305.17888.

Jirí Martínek, Christophe Cerisara, Pavel Král, Ladislav Lenc, and Josef Baloun. Weak supervision for question type detection with large language models. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3283–3287. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-345. URL https://doi.org/10.21437/Interspeech.2022-345.

Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA, 2010. doi: 10.21437/INTERSPEECH.2010-343. URL https://doi.org/10.21437/Interspeech.2010-343.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191. URL https://doi.org/10.1109/TKDE.2009.191.

Letitia Parcalabescu, Nils Trost, and Anette Frank. What is multimodality? *CoRR*, abs/2103.06304, 2021. URL https://arxiv.org/abs/2103.06304.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. Bidirectional language models are also few-shot learners. *CoRR*, abs/2209.14500, 2022. doi: 10.48550/ARXIV.2209.14500. URL https://doi.org/10.48550/arXiv.2209.14500.

Madhura Phatak and Prashant Borkar. Image aesthetics assessment using deep learning (based on high level image attributes). *International Journal of Engineering Research Technology (IJERT)*, 8(05), 2020. ICSITS – 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross B. Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. *CoRR*, abs/2003.13678, 2020. URL https://arxiv.org/abs/2003.13678.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207, 2021. URL https://arxiv.org/abs/2110.08207.

Cees Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In HongJiang Zhang, Tat-Seng Chua, Ralf Steinmetz, Mohan S. Kankanhalli, and Lynn Wilcox, editors, *Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005*, pages 399–402. ACM, 2005. doi: 10.1145/1101149.1101236. URL https://doi.org/10.1145/1101149.1101236.

Maciej Suchecki and Tomasz Trzcinski. Understanding aesthetics in photography using deep convolutional neural networks. *CoRR*, abs/1707.08985, 2017. URL http://arxiv.org/abs/1707.08985.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583, 2019. URL http://arxiv.org/abs/1905.05583.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Min Wang, Ata Mahjoubfar, and Anupama Joshi. Fashionvqa: A domain-specific visual question answering system. *CoRR*, abs/2208.11253, 2022a. doi: 10.48550/ARXIV.2208.11253. URL https://doi.org/10.48550/arXiv.2208.11253.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. Language models with image descriptors are strong few-shot video-language learners. *CoRR*, abs/2205.10747, 2022b. doi: 10.48550/ARXIV.2205.10747. URL https://doi.org/10.48550/arXiv.2205.10747.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL https://arxiv.org/abs/2201.11903.

Alessandro Wollek, Sardi Hyska, Bastian O. Sabel, Michael Ingrisch, and Tobias Lasser. Exploring the impact of image resolution on chest x-ray classification performance. *CoRR*, abs/2306.06051, 2023. doi: 10.48550/ARXIV.2306.06051. URL https://doi.org/10.48550/arXiv.2306.06051.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. *CoRR*, abs/2311.13165,

2023. doi: 10.48550/ARXIV.2311.13165. URL https://doi.org/10.48550/arXiv.2311.13165.

Mengjia Yan, Yurou Duan, Siqi Deng, Wenjia Zhu, and Xiaoyu Wu. Facial beauty assessment under unconstrained conditions. In *8th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2016, Ploiesti, Romania, June 30 - July 2, 2016*, pages 1–6. IEEE, 2016. doi: 10.1109/ECAI.2016.7861087. URL https://doi.org/10.1109/ECAI.2016.7861087.

Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey. *CoRR*, abs/2204.08610, 2022. doi: 10.48550/ARXIV.2204.08610. URL https://doi.org/10.48550/arXiv.2204.08610.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. doi: 10.48550/ARXIV.2306.13549. URL https://doi.org/10.48550/arXiv.2306.13549.

Desen Yuan. Language bias in visual question answering: A survey and taxonomy. *CoRR*, abs/2111.08531, 2021. URL https://arxiv.org/abs/2111.08531.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. Differentiable prompt makes pre-trained language models better few-shot learners. *CoRR*, abs/2108.13161, 2021. URL https://arxiv.org/abs/2108.13161.

Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. *CoRR*, abs/2207.09519, 2022a. doi: 10.48550/ARXIV.2207.09519. URL https://doi.org/10.48550/arXiv.2207.09519.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *CoRR*, abs/2210.03493, 2022b. doi: 10.48550/ARXIV.2210.03493. URL https://doi.org/10.48550/arXiv.2210.03493.