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

# Detecting Hidden Meaning in Stock Images

## Master's Thesis

Pia Sülzle

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

Submission date: November 30, 2023

# Declaration

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

Leipzig, November 30, 2023

. . . . . .

Pia Sülzle

**Abstract**

Stock images are images that each of us encounters in our daily lives. They are intended to convey a feeling and are used as symbolic representations in many contexts. In order to maximize the sales of licenses and allow stock images to be used in a variety of contexts, the design process focuses on keeping the images generic. Due to their frequent use in different contexts, they are ideal for the previously under-studied area of quantitatively finding meaning in images. The automated detection of hidden meaning in images could enable Text2Image models such as DALL-E to better understand and represent implicit meanings. This thesis therefore addresses the question of whether it is possible to find hidden meaning in images using existing models such as CLIP. For this purpose, a stock photo dataset is created, which contains a set of captions for some of the images. This allows investigating different contexts in which a single image appears. Additionally, various experiements were conducted to determine the feasibility of extracting meaning from a set of captions. Analysis of the captions using bag of words and topic modeling led to the discovery of hidden meaning within the examined images. This demonstrated the presence of a hidden meaning in the images that can be differentiated from their description. Further experiments showed that there is no possibility of using CLIP to distinguish between meaning-bearing and descriptive captions. Therefore, it is not yet suitable for recognizing meaning in images on its own. Furthermore, we developed a model to generate a hidden meaning embedding. Due to the limitations of CLIP regarding the distinction between meaning and description and the few possibilities to transfer CLIP embeddings back to interpretable forms such as text, this approach has not yet been able to achieve satisfactory results. However, we outline a possibility to extract hidden meaning with future models and further work in this area. This thesis establishes a groundwork for future research through the creation of a novel stock photo dataset and the execution of diverse experiments, aimed at uncovering the underlying meanings within images.

# Contents

**1 Introduction** **1**

**2 Background** **5**
2.1 Foundations of Image Interpretation . . . . . 5
2.2 Stock Images . . . . . 10
2.3 CLIP - a multimodal Model . . . . . 12

**3 Related Work** **15**
3.1 Meaning in Texts and Images . . . . . 15
3.2 CLIP retrieval and LAION-5B . . . . . 21

**4 Experiments & Results** **24**
4.1 Creation of a Stock Image Dataset . . . . . 24
4.1.1 Scraping Stock Images . . . . . 25
4.1.2 Caption Crawling . . . . . 26
4.2 Caption Analysis . . . . . 30
4.2.1 Bag-of-Word-Analysis of Captions . . . . . 31
4.2.2 Topic Modeling of Captions . . . . . 34
4.2.3 Similarity Analysis of Embeddings . . . . . 38
4.3 Extract Hidden Meaning from Images . . . . . 46
4.3.1 Creation of a Hidden Meaning Embedding . . . . . 47
4.3.2 Translation of Hidden Meaning Embedding . . . . . 50
4.3.3 Conclusion of the Hidden Meaning Extraction . . . . . 53

**5 Discussion and Conclusion** **55**
5.1 Future Work . . . . . 60

**Bibliography** **61**

# Chapter 1

# Introduction

Every one of us is familiar with such images: an article about stress at work, supplemented with a with a picture of a group of focused, young people sitting around a desk and discussing ideas, a happy family on a walk in the autumnal forest on the website of an insurance company. These are images that we encounter everywhere every day and think nothing more of, images that hardly stick in our minds after we have read the article or left the website. These high resolution images convey some kind of feeling, even if they seem somehow staged. The people that are depicted are usually stereotypically attractive and the situations they are in are vague: An office, a doctor's practice, a park or forest somewhere in the world. These images are often so called *stock photos*: Images from some image database held in stock that *stand for* something. The stock photo agencies sell licenses for the use of these images. A good stock photo is an image that can be used particularly well in various contexts and can therefore be sold frequently. Therefore, stock photos are designed to convey more than one meaning, because the goal is to be able to sell licenses for such photos multiple times. We frequently encounter such images and often do not question them further, and without wasting any thought we can understand why certain images are used as symbolic images e.g. for articles. In general, images often have a meaning that is immediately clear to us without understanding exactly why. Images contain codes that we have learned to decode. This code is necessary to find meanings in pictures. Many researchers have already dealt with the finding of meaning in images. Often the ability of images to carry more meanings than texts is pointed out. The meanings that a picture can carry always depend on the people looking at it and what they associate with certain symbols in pictures. And although finding meaning in images is a field of research often considered theoretically, it is rarely quantitatively analyzed in practice. Therefore, an analysis of meanings in images is an interesting field of research. Stock photos are ideal for such analyzes

as they are created without a specific intention of meaning. They only have meaning ascribed to them in the process of use by various parties. Often stock photos are intended primarily to evoke feelings and not to stand too concretely for anything. The way people interpret stock photos and the meanings they ascribe to them often only becomes clear in the context in which these images are used. One question that arises in the context of this work is whether many people associate similar things with images even though they are not directly visible on the image, i.e. whether they decode the code of an image in a similar way. Stock photos are a good source for this question due to their multiple use. Since licenses for stock photos are not always sold to just one party, but multiple times, the same stock photos sometimes appear several times on the internet, so that it is possible to find several contexts for a single image. This makes it possible to get to the bottom of the question of whether different recipients associate the same things with one and the same image.

The present work is primarily concerned with the description of a picture and in contrast to it, its hidden meaning. What can actually be seen in a picture, which objects and activities, is regarded as a *literal description*. However, everything that is not directly visible on the image and that puts the image into context, giving it meaning, is referred to as *hidden meaning* in the following work. Object detection and classification of images is already a major topic in the field of machine learning. Large image datasets such as MS-COCO[1] or ImageNet[2] are designed to train models that are supposed to recognize what can be seen in an image or classify images into certain categories. Object detection has been a much-researched field since the early days of deep learning. Finding meanings in images, on the other hand, has received little attention. Yet it would be interesting if computers could be taught to understand images the way we do. Models such as DALL·E and Stable Diffusion excel at representing tangible objects and actions, but understanding and representing implicit meanings remains a challenge (Chakrabarty et al., 2023). Therefore, this work is intended to lay a kind of foundation and investigate whether it is possible with existing tools to recognize not only the obvious in an image, but also the hidden. To find this hidden meaning and to understand what people see in images, texts that appear together with an image are a good starting point. Such texts could be captions as these can show the context in which an image was used. In order to be able to analyze the images and the associated texts together, a model is needed that makes both comparable. With CLIP a model was published, which makes it possible to represent images and texts in the same vector space (Radford et al., 2021). Thus, images and texts can be made comparable for the computer. CLIP converts images and texts

[1] https://cocodataset.org/
[2] https://image-net.org

into embeddings, for which a similarity value can be calculated, thus it is now possible to make a statement about how similar an image is to a text. One question that arises is essential in the context of the present work: Does the similarity of CLIP embeddings represent descriptions or meanings? One of the possibilities would be that a high CLIP embedding similarity of an image and a text, indicates that the image actually shows what the text describes. Another possibility would be that a high embedding similarity makes a statement that the text captures one of the meanings of an image. Both possibilities could help to reveal the hidden meaning of an image by examining images and their accompanying texts. Thus, the use of CLIP could be a good way to investigate images for their meanings.

With the subject of the thesis of detecting hidden meaning in stock images, the following research questions arise:

- **Q1** To what extent is there a difference between what is seen in a picture and what the picture means?
- **Q2** Can this difference be quantified with the help of the multimodal model CLIP?
- **Q3** Can one of the meanings of an image be extracted with the help of the multimodal model CLIP?

Since a computational analysis of stock images has not been done before and thus no dataset is available that is composed of stock photos only, a stock photo dataset will be created as part of this work. This dataset will encompass stock images as well as the captions that can be found in association with them. This not only lays the groundwork for the experiments that will be conducted in this thesis but poses an important contribution to future work on the topic.

The research questions will be tackled with a combination of different approaches. Methods used include a bag of word (BoW) analysis on the image captions to find clues to the level of meaning in images through the frequency of words. Topic modeling on image captions will be applied to find meaning through topics and possibly reveal a difference between levels of meaning and description. The latter will also be analyzed by looking at similarities between CLIP embeddings. Furthermore, generated image descriptions and image captions will be compared using the text-only model SBERT. The results of these similarity analyzes will contribute to the creation of a Siamese Network (SN), with which further work in this direction should be able to generate a hidden meaning embedding (HME) from an image.

Chapter 2 lays the groundwork for this thesis. The first part of the chapter deals with the question of how meaning is revealed in images and how the relationship of images and texts can contribute to meaning making. The second

part of the chapter deals with stock photos in general, what stock photos are, what makes them different from other images, and why it is interesting to deal with their meaning. The third part introduces CLIP, a model that is very relevant to this thesis and to answering the research questions. Chapter 3 examines the interplay between text and images in multimodal data, exploring various methodologies. Besides, the LAION-5B dataset and the CLIP-retrieval based on it are also presented. In Chapter 4 the different experiments are presented and evaluated individually. The first part deals with the creation of the dataset necessary for the later analyzes. The second part of the chapter deals with the largest topic of caption analysis, analyzing word frequencies, similarities, and topic that occur in captions. The individual analyzes are described in more detail and evaluated using example images. In the following part, the experimental idea of creating a HME for images is cited and a possible approach for it is presented. Finally, Chapter 5 summarizes and discusses the findings of the thesis. After, the possible outlooks and further work that emerge from the results and limitations of the thesis are presented.

# Chapter 2

# Background

In order to understand why it is important and useful to study images and their meaning, and why stock images are an interesting field of research for these analyzes, this chapter provides essential context and background. The initial section of the chapter examines the differentiation between the visible component of images and their underlying meanings, demonstrating that images frequently convey meaning exceeding their apparent content. The succeeding section delves into the nature of stock images, emphasizing their distinct attributes in comparison to other image types. At the end of this chapter, we provide a detailed overview of the CLIP, a model that can represent images and text in the same vector space. Thus, CLIP is well suited for the analyzes in this thesis, because a linkage of images and associated text can provide information about their meaning as will be shown in Section 2.1.

## 2.1 Foundations of Image Interpretation

The first question that arises from the title of this thesis is: What is the hidden meaning of an image? To answer this question, it is reasonable to look at how people see and understand images. In the present work we distinguish between the hidden meaning of an image and the literal description. In order to first of all bring closer what is meant by this, we can have a look at Figure 2.1. The photograph depicts two women playing with a beach ball in a lake on a sunny day. At first sight this picture appears to convey any further meaning. However, it was used in this example with the caption 'Cooling down is needed soon: In Germany, the first heat wave of the year is about to start'(English translation of the German caption: 'Eine Abkühlung ist bald vonnöten: In Deutschland steht die erste Hitzewelle des Jahres in den Startlöchern'). Thus, the terms *heat wave* and *cooling* are presented in relation to the image. This creates a connotation that is not visually represented. It is unclear if the


**Figure 2.1:** Image taken from an German Online Newspaper. The picture shows two women in a lake playing with a beach ball. The caption is: 'Cooling down is needed soon: In Germany, the first heat wave of the year is about to start'(English translation of the German caption: 'Eine Abkühlung ist bald vonnöten: In Deutschland steht die erste Hitzewelle des Jahres in den Startlöchern'). Source: https://www.tz.de/welt/italien-unwetter-hochwasser-ueberschwemmung-hitzewelle-deutschland-sommerwetter-spanien-zr-92304148.html

image was taken during a heatwave or if the lake is cooling. Nonetheless, this association exists within the image and is referred to as *hidden meaning* in this thesis.

**Images** According to Barthes, an image contains three different kinds of messages (Barthes, 1977). To grasp the three messages, Barthes presents the image in Figure 2.2. The image belonged to an advertisement of a company named Panzani. The first message an image may contain is the *linguistic message*. This is any word or linguistic material that can be found on an image. The linguistic message in this image is both the writing that can be seen on the various items and the writing on the image that can be seen at the bottom right. All that is needed to understand this message is an understanding of the French language. Barthes refers to the description of what is visibly present in an image as the *uncoded iconic message*, which he describes as the 'first degree of intelligibility'. Not only shapes and colors are perceived, but the actual objects shown. According to Barthes, this message is uncoded because it is supposed to say exactly what can actually be seen. In this type of message, the elements of the image simply represent what they are. To come back to Figure 2.2: A tomato is just a tomato, without any underlying symbolisn or additional meaning.


**Figure 2.2:** An image from a Panzani advertising campaign. Roland Barthes uses the image to explain his different levels of meaning in *Rhetoric of the Image*.

The third type of message an image can contain is the *coded iconic message*. This type of message is not very direct and becomes clear mainly through connotations. This message is the most relevant for this work, as it holds the meaning of an image apart from what is shown, and thus can be representative of *hidden meaning*. In Figure 2.2, the use of fresh ingredients and an open bag conveys the feeling of returning from a market. In addition, the color scheme of the picture evokes a sense of Italy. Analyzing this message is challenging due to several factors. Firstly, an image can have several meanings, not all of which can be recognized or even perceived by every viewer. The meanings that an image can have therefore depend not only on the creator of the image, but also on the recipient and whether both parties share a common knowledge. The sense of Italy or the return from the market, e.g., might not have been conveyed through Figure 2.2 at all in another culture or a less Eurocentric society.

The three messages that an image can convey show its ability to communicate through varying perspectives. Much of the work that deals with the meaning of images relates to Barthes' considerations. Another terminology

introduced by Barthes is that of two levels of meaning in images: *connotation* and *denotation*(Barthes, 1977). Denotation is equivalent to the literal, explicit meaning of an image (Sturken and Cartwright, 2009). Connotation, in contrast, is contextual and is more likely to be the symbolic meaning or range of possible meanings conveyed by cultural codes. According to Leeuwen (2020), all visual communication is coded. Codes have a major impact on how we see images and how we read meaning from images (Sturken and Cartwright, 2009). Nevertheless, codes are difficult to decode automatically because they can also have different meanings in different contexts. In our work, the connotative meaning conveyed by code is equivalent to the hidden meaning. Humans have developed the skill to interpret codes instinctively (Sturken and Cartwright, 2009). There are many more approaches to find meaning in images. In the present work, we distinguish between the hidden meaning and the literal description of an image. To draw comparisons to the two levels of meaning described by Barthes, one could equate the literal description with the denotation and the hidden meaning with the connotation. Another important idea related to image analysis is *semiotics*. As stated by Aiello and Parry (2020), semiotics explores the creation and interpretation of meanings through signs. A sign can be anything, but what is crucial is that it can be understood by a community of individuals who share the same cultural or social background. Although signs can be subjectively interpreted and their understanding can differ across cultures, semiotics offers valuable insights into how images communicate meaning. There are other approaches and works on the meaning and 'reading' images, yet Aiello and Parry say that visual communication is still an understudied field of research. Often these approaches are based on the work of Barthes and draw on his views on finding visual meaning. Many of these approaches are therefore also similar, and overall speak of the ambiguous determination of a particular meaning. The meaning of an image is always ambiguous and depends on both the cultural and social circumstances in which the image was created and the cultural and social circumstances in which the recipient of the image finds himself. Finding meaning requires a close analysis of the signs and codes present in an image. This suggests that it is difficult to find meaning in images using computational methods. However, many of the approaches to meaning-finding suggest that text that occurs with an image can make an important contribution to the meaning of the image.

**Image-Text-Relationship** The relationship between images and texts can contribute to the meaning and especially the interpretability of an image. Often the term *meaning multiplicatio*n is used, which was characterized by the semiotician Jay Lemke (1998). He states the combination of different modes, such as image and text, can carry more meaning than any of these modalities

alone. Thus, an image considered alone can carry less meaning than an image considered in conjunction with a text. This can also be seen in Figure 2.3. The image alone does not say much more than that the person in the image is smoking. With the two different captions, however, the image has two different meanings.


**Figure 2.3:** An example of meaning multiplication. Two different captions for the same image allow the image-text pair to convey two different meanings. Source: Kruk et al. (2019)

According to Barthes there are three different relationships in which texts can stand to images (Barthes, 1977). Figure 2.4 shows an overview of the three kinds of relationship, which are divided into equal and unequal. Barthes uses the term *relay* for the only equal relationship. Here, the importance of one modality does not outweigh the importance of the other, and the relationship between the two tends to be balanced. The meaning of the whole emerges primarily from the combination of both, image and text. The unbalanced relationships of texts and images are divided into *anchorage* and *illustration*. By anchorage, Barthes refers to the reinforcement of an image through the text. The text can help to classify the image by identifying the elements of the scene or the scene of the image itself. In Barthes' words, the text answers the question *'What is it?'*. An anchorage is thus, roughly speaking, a description of the image with the purpose of conveying a clear understanding of its intended meaning. Illustration is distinguished from anchorage by Barthes as the relationship between text and image in which the image is serves to explain the meaning of the text. In contrast to anchorage, where the primary meaning is obtained from the image and the text is simply supportive, the text defines the meaning in this case and the image serves a supporting role. Barthes' classification into the three types of text-image relationships is the foundation for much further research in this area. Bateman (2014) proposes the following five text-image relationships as a summary of different classification approaches:

- illustration, which means the image is subordinate to the text,

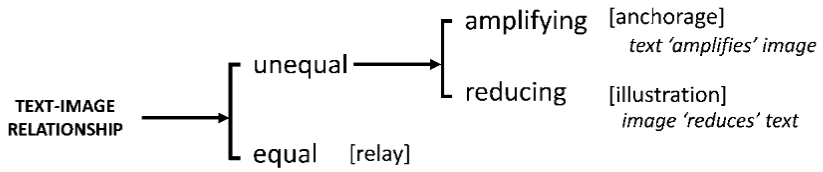



**Figure 2.4:** Barthes' classification of text-image relationships represented graphically as a systemic network by Bateman (2014)

- pictoral exemplification, where the image serves as an example of what the text describes and thus provides new information,
- labelling, where the image is identified by a textual element, such as the name of a work of art,
- mutual determination, which combines relay and anchorage
- contradiction, where the meaning of the image and the text are opposing

Nevertheless, in Bateman opinion, there can be no conclusive and correct systematization of these relationships, since the different types can overlap or be mutually exclusive. Furthermore, it is possible to find definite instances for every relationship type of relationship, as well as cases that do not neatly fit into one or overlap into multiple categories. In summary, researchers have proposed typologies of relationships between text and images. However, there have been few systematic studies that demonstrate how recipients perceive and interpret these relationships (Bateman, 2014).

## 2.2 Stock Images

In order to understand why it is interesting to use stock images for the analysis of meanings in images, it is crucial to understand the concept of stock images.

Even if not everyone is familiar with the term, most of us have seen such images in many different contexts, often used as symbolic images for newspaper articles or in advertising. One of the largest image and photography agency that sells licenses for stock photos is *Getty Images*[1] (Frosh, 2020). In Figure 2.5, it is evident that stock photos from Getty Images are used as symbol images in articles in leading newspapers, such as the New York Times. Basically, stock images are images, usually photographs, that are held in 'stock' by agencies in order to sell licenses for those images (Frosh, 2002). In order to sell stock

[1] https://www.gettyimages.com

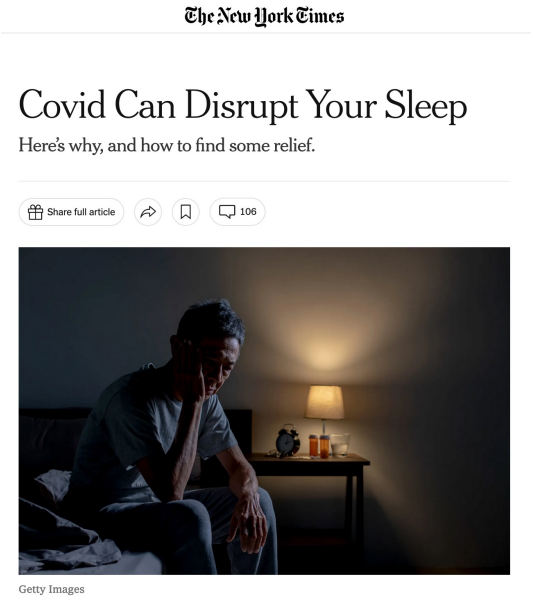



**Figure 2.5:** Getty Image in New York Times Article with the title 'Covid Can Disrupt Your Sleep'. Source: https://www.nytimes.com/2023/09/20/well/live/sleep-covid-symptoms.html

photo licenses multiple times, they must be suitable for a variety of contexts especially in advertising and marketing (Frosh, 2002; Ward, 2007). Researchers therefore refer to stock photos as *generic images* (Frosh, 2002; Machin, 2004; Ward, 2007). Stock photos are specially designed so that they are inherently ambiguous can hold more a wide range of possible meanings (Ward, 2007). The final use of a stock photo is uncertain during the design process, so they are kept generic to enable recipients to interpret them according to their preferences. Neither the photographers nor the stock photo agencies know in advance for which products, media platforms or contexts the images will ultimately be used (Frosh, 2020). One design decision that is commonly made to enable multiple uses of stock photos is *decontextualization* (Machin, 2004). This means that in stock photos, the background is often out of focus, or there is just a color as background. This ensures that an image can be more easily placed in different contexts. Furthermore, the people shown in the pictures become 'typical examples', where the exact place and time are insignificant (Leeuwen, 2020).

According to Machin (2004), an important point for images that sell particularly well is that an image should not capture or document a particular situation, but rather evoke a feeling in the recipient. In Figure 2.5, it is unclear what kind of scene the person is in. It is possible that the time of day

is morning and the room is still dark, or it could be evening. The presence of medicine cans on the nightstand suggest that the person may be unwell, although it is also possible that the medicine belongs to someone else. The scene is ambiguous, but suggests a feeling of melancholy and discomfort. It cannot be determined whether the person is experiencing sleep issues due to Covid-19, however, the image captures the essence of the accompanying article. In his research on stock photos, Frosh concludes that the inherent ambiguity in these images, makes them a intriguing basis for analyzes.

## 2.3 CLIP - a multimodal Model

Contrastive Language-Image Pre-training (CLIP) is a multimodal neural network that learns visual concepts using natural language monitoring (Radford et al., 2021). Using CLIP, visual and linguistic concepts can be linked, mapped into the same vector space, and thus made comparable.

The idea behind CLIP is to learn visual concepts through natural language supervision. Learning via natural language has several advantages over other training methods. In conventional approaches, especially in computer vision, models are trained using annotated datasets, i.e. images and their class labels. Instead of class labels, for example, captions that are occurring with the images are used in natural language learning approaches. Therefore, data used for natural language learning does not need to be annotated and is available in large quantities on the internet. This means that such models are much more scalable, as there is no need to label a large dataset using, for example, crowdsourcing approaches. In addition to this great advantage, models based on natural language do not only learn the representation of an image, but can also combine the representation with natural language. The natural language supervision approach is therefore the focus of Radford et al.'s work.

As the main motivation for CLIP is the ease of scalability, one concern of Radford et al. was to train CLIP on the largest dataset possible. Typical image datasets, such as MS-COCO (Lin et al., 2015), are therefore too small at about 100,000 images. Larger datasets, such as YFCC100M (Thomee et al., 2016), contain a very large amount of images, but the descriptions or titles are not usable to train the model on natural language. Therefore, it was necessary to create a new dataset to train CLIP. For this dataset, 400 million image-text pairs were crawled from a variety of publicly available sources on the internet. Since the dataset should cover a wide range of visual concepts, a list of queries was created that included words that occur at least 100 times in the English language Wikipedia[2]. The text of an image-text pair included in the dataset

[2]https://en.wikipedia.org/

had to contain one of these queries. To balance the dataset, up to 20,000 pairs per query were included in the dataset.

After initial trials, Radford et al. determined that training CLIP to predict every word within a text linked to the image required excessive effort. Therefore, Radford et al. decided to take a different approach. CLIP was then trained on the task of predicting an entire text that belongs to an image. To do this, CLIP was given a batch of image-text pairs, and CLIP was asked to predict which of the possible image-text pairs actually occur in the dataset. To accomplish this task, CLIP learns a multimodal embedding space by training image and text encoders simultaneously. In doing so, CLIP tries to maximize the cosine similarities of the image and text embeddings of the true pairs and minimize the cosine similarities of the false pairs. CLIP trains both the text encoder and the image encoder from scratch without initialization with pre-trained weights. The structure of the pre-training can be seen in Figure 2.6. For the image encoder, the ResNet-50 architecture from He et al. (2015) was

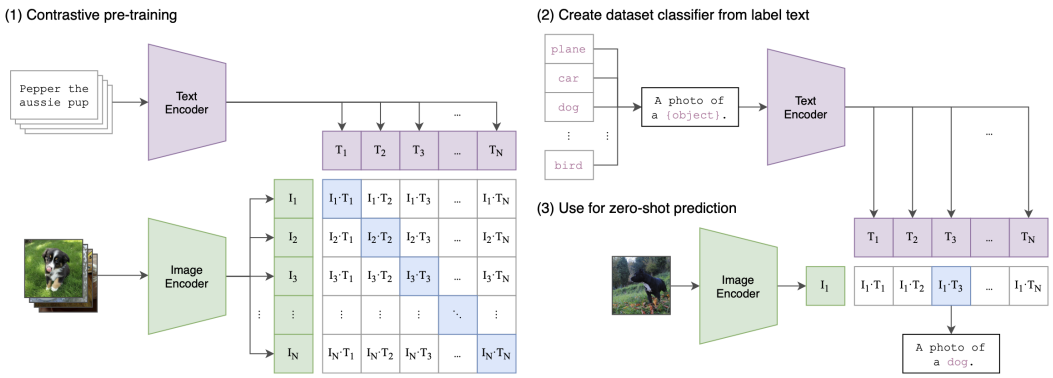



**Figure 2.6:** (1) Summary of the CLIP-approach. For CLIP, both a text encoder and an image encoder are trained to predict a matching image-text pair from a set of images and texts. (2)/(3) CLIP for zero-shot prediction. By extending labels to natural language, CLIP can also be used for zero-shot prediction on conventional image datasets. Source: Radford et al. (2021)

used as the base architecture with a few modifications. The transformer architecture of Vaswani et al. (2017) with architectural modifications proposed by Radford et al. (2019) was used as the basis for the text encoder.

CLIP, unlike other image models, was trained on natural language and not on class labels. So, in order to test CLIP for generalizability on other datasets, the data had to be prepared. To do this, the labels were first converted into a natural language text snippet, as shown in Figure 2.6, where, e.g., the label *dog* was converted into *A photo of a dog.* Thus, the ability of CLIP to compute

the most likely image-text pair from a batch could be used for other datasets as well. With such experiments, it was shown that CLIP is able to recognize a variety of visual concepts in images and associate them with their names. Radford et al. assume that as a result, CLIP models can be applied to almost any visual classification task.

# Chapter 3

# Related Work

In this chapter, we provide a overview of related work relevant to the scope of this thesis, focusing on multimodal data and their various applications. The first section deals with research focusing on the interaction between textual and visual elements. The work presented in this subsection clarifies the connections and relationships between images and texts, forming a crucial foundation for this thesis. The last section presents the LAION-5B dataset and the CLIP retrieval based on it, which has played an important role in the creation process of the dataset presented in this work.

## 3.1 Meaning in Texts and Images

Müller-Budack et al. (2021) present two multimodal approaches. The first approach focuses on quantifying entity coherences between images and texts. The second approach is about quantifying contextual image-text relationships. Müller-Budack et al.'s work is primarily concerned with the meaning of text and the extent to which images can support or even change the meaning of the associated text. Müller-Budack et al. (2021) distinguish the relationship between images and text in three different modes. Accordingly, in relation to the text, an image can be (a) decorative, adding little or no information to the meaning of the text, (b) information-enriching, giving important or additional details to the text, or (c) misleading, adding conflicting visual information to the text. Müller-Budack et al. present an approach for verifying cross-modal entity consistency. The goal is to automatically check semantic relations in the form of common entities between image and text pairs. This is done by measuring the cross-modal similarity for these entities. For this purpose, first of all named entities are extracted from a text with spaCy[1]. The named entities

[1]https://spacy.io

are then categorized and cross-checked against knowledge databases. The image belonging to the text is analyzed to extract various entities such as people using face recognition or places using geolocation models. These extracted entities are then compared to those identified in the text. Therefore, images from image search engines such as Google are crawled for the entities mentioned in the text. For the recognized faces from the source image and the crawled images of persons found in the text, the cosine similarity of the feature vectors is computed and thus a *cross-modal person similarity* is computed. The cross-modal similarity of other entities is computed similarly. The process can be seen in Figure 3.1. The other approach presented by Müller-Budack et al.

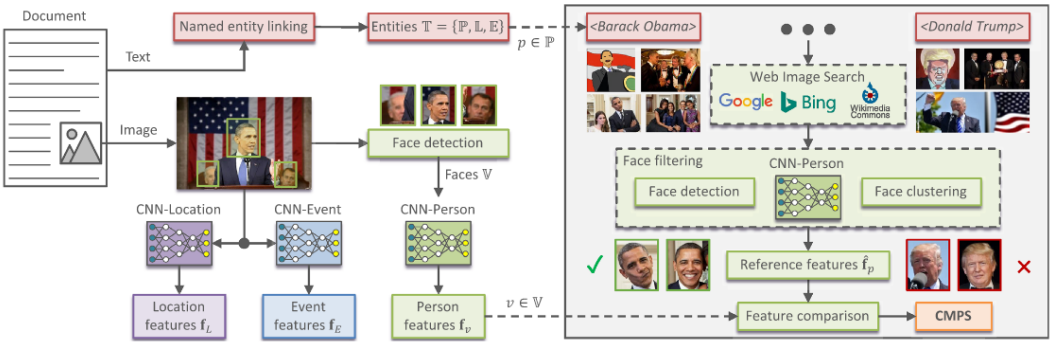



**Figure 3.1:** Overview of the Cross-modal Entity Consistency Workflow from Müller-Budack et al. (2021). With the help of Named Entity Linking and different CNNs for different entities, location, events and names are extracted from text and images. Example using person entities: Web Image Search is used to search for images for the person names found in the text and compare them with the persons found in the reference image.

deals with quantifying the contextual semantic relationship between texts and images. Using part-of-speech tagging, nouns are extracted from articles as they are considered appropriate words to represent general concepts of the text. Then, embeddings are computed for the extracted nouns. Scenes are extracted and classified from the images using a model trained on the Places365 dataset[2]. In addition, the probability of the scene is calculated. Embeddings are also calculated for the extracted scene labels. The cosine similarity is then calculated for the embeddings of each noun and the embeddings of the scene class labels and weighted based on the probability of the scene. Finally, the maximum similarity of all these calculations is the *Cross-modal Context Similarity*. In Figure 3.2 the process can be seen. Alikhani et al. (2020) investigate in their work the goals and information needs of captions. They are partic-

[2]http://places2.csail.mit.edu

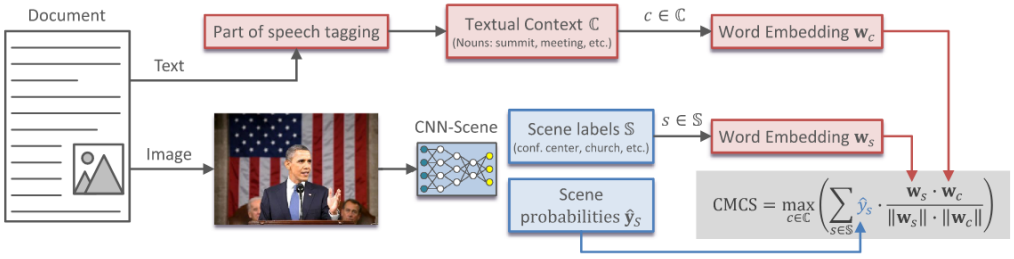



**Figure 3.2:** Overview of the Cross-modal Context Consistency Workflow from Müller-Budack et al. (2021). Extracting the textual scene context from the document. Additionally extracting the visual scene and scene probability from the image using a special CNN. Calculation of the cross-modal context consistency of the similarity of both scenes.

ularly concerned with the coherence relationships between images and text. When generating image captions using conventional models, Alikhani et al. repeatedly noticed that models 'made up' information that was not present in the images or requested in the captions. For their work, they consider five image-text coherence relationships. The *visible* relationship is about the text presenting information that characterizes what is shown in the image. *Subjective* refers to a coherence relation in which the text is a reaction or judgment of the speaker to the image. When the text describes a longer dynamic process of which the image captures only a representative snapshot, Alikhani et al. refer to the relationship as *action*. In the *story* relationship, the text is an independent description, of the circumstances presented in the image, such as instructions or background information. *Meta*, the final relationship examined by Alikhani et al., describes a relationship in which the text provides details regarding the scene depicted in the image and, e.g., the creation process of the image. These relationships are often non-exclusive, meaning that an image-text pair can have multiple of these relationships. Alikhani et al. trained their model using a dataset of 10,000 publicly available image-caption pairs. These pairs are manually labeled with the introduced relationship types. During annotation, it was also observed that image-caption pairs from various publication sources exhibit different distributions of coherence relations. Compared to image-text pairs from the Daily Mail domain, which were primarily labeled as *story*, pairs from the Getty Images domain mainly exhibited *meta* and *visible* relationships. A model was trained on the annotated dataset to generate coherence-dependent captions. Along with the image, the model was provided with the desired coherence label. Consequently, the authors observed a reduction in 'hallucinations'. In Figure 3.3 it is evident that the coherence-agnostic model partially 'made up' things like for image (a) *beautiful girl* or in (b) *best*

(a) coherence-aware *Meta*: A girl in the winter forest. coherence–agnostic: beautiful girl in a red dress.

(b) coherence-aware *Visible*: the pizza at restaurant is seen. coherence–agnostic: the best pizza in the world.

**Figure 3.3:** Images with their corresponding captions generated by a coherence-aware model and the coherence-agnostic model by Alikhani et al. (2020).

*pizza* that may not have been intended. The model that includes the coherence relation in the generation can thus generate more accurate captions that match the intended relationship.

In their work, Kruk et al. (2019) examine the complex relationship between image and text in the context of Instagram posts. For this purpose, they developed three taxonomies that deal with different areas of text-image relationships. The first taxonomy is about categorizing an author's intentions. In doing so, they distinguish between eight different intentions that an author may have when composing an image and a caption. These include but are not limited to the *expressive* intention, which involves expressing feelings, admiration or attachment towards e.g. a group, or the *entertainment* intention, which aims to entertain people through e.g. art or humor. Kruk et al. introduce two taxonomies in addition to the author's intention. One of these taxonomies is the contextual relationship between the literal meaning of the image and the caption. There are three different kinds of relationship for this type. If the literary meaning of the caption and the image barely overlap, it is a minimal contextual relationship. The opposite is true in case of a close contextual relationship, where there is a significant overlap between the literary meaning of the caption and the image. This is the case, e.g., when the caption describes what is shown on the image. The last type of contextual relation is the transcendental relation. This type is given when the meanings overlap, but additional information is given in the caption. The final taxonomy presented by Kruk et al. is the *semiotic* taxonomy. This taxonomy is concerned with the relationship between what is represented by modalities. This is to analyze the meaning of the signs depicted in both the image and caption objectively. The semiotic relationship, as classified by Kruk et al. is divided into *divergent*, *parallel* or *additive*. Divergent in this context refers to the relationship, where there is a difference between the meanings suggested by the image and by the text. That means the image and text semiotics are contradictory. A

relation is called parallel when the meaning of the text and image are identical, independent of one another. The last type of semiotic relationships is the additive one, which exists when the meaning of the image and the meaning of the text reinforce or change each other. Kruk et al. have also found in their experiments that the semiotic relationship are not always homologous. The contextual relationship between an image and its caption can be minimal and at the same time the semiotic relationship can be additive. For the analysis based on the taxonomies, Kruk et al. created their own dataset, called MDID (Multimodal Document Intent Dataset). It consists of 1299 Instagram posts annotated according to the three taxonomies. Meaningful examples were selected for various taxonomies and the respective categories. The dataset was utilized to train a neural network that computes the probabilities for the different categories within the taxonomies for new image-text pairs. Using their model, Kruk et al. demonstrated that the combination of both image and text produces improved categorization results compared to either image or text alone. This proved the role of meaning multiplication.

Chakrabarty et al. (2023) deal with visual metaphors in their work. Metaphors are a linguistic device of evoking images in the mind of the recipient by linking certain expressions, thus making it easier to convey ideas. In their work, Chakrabarty et al. examined how to visually depict these metaphors in using diffusion-based text-image models like DALL·E 2. This is a challenging task for such models, because metaphors convey meanings only implicitly through symbols. Models like DALL·E or Stable Diffusion are unable to capture implicit meaning. Therefore, it is necessary to *translate* the linguistic metaphors using an LLM before feeding them into DALL·E to create visual metaphors. Chakrabarty et al. used GPT-3 to identify the implied meaning, and thus translated linguistic methapers such as *'My bedroom is a pig sty.'* Using Chain of Thought prompting introduced by Wei et al. (2022) before passing the text to DALL·E to generate an image. Therefore, Chakrabarty et al. input the prompt *Your task will be to elaborate a metaphor with rich visual details along with the provided objects to be included and implicit meaning. Make sure to include the implicit meaning and the objects to be included in the explanation* together with the image into the GPT-3 model. The visual elaboration generated by GPT-3 was subsequently used as a prompt for DALL·E to generate a visual metaphor. In Figure 3.4, the impact of translating the linguistic metaphor on the quality of the visual metaphor is evident. On the left, DALL·E generates images solely from the metaphor, while the right, the metaphor was translated beforehand using GPT-3, so DALL·E no longer needs to understand and implement the rhetorical concept and implied meaning itself. As a result of their work, Chakrabarty et al. have created a high quality dataset containing 6476 visual metaphors for 1540 linguistic metaphors and their corresponding visual

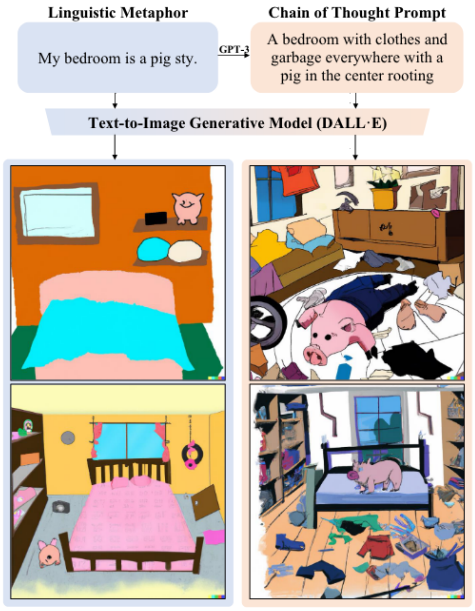



**Figure 3.4:** Visual metaphors generated by DALL·E 2. Left: Generating image by prompting DALL·E 2 the actual linguistic metapher. Right: Generating images by prompting DALL·E 2 the Chain of Thought 'translated' metapher using GPT-3. Source: Chakrabarty et al. (2023)

elaborations.

Another work that deals to some extent with the meaning of images is *Embedding-based Stance and Persuasiveness Classification* by Torky et al. (2023). The paper is a submission of the Webis group to the subtasks of ImageArg 2023[3]. The focus is on the analysis of argumentative stances in images, which has received increasing attention in recent years. The tasks are performed on the multimodal dataset called ImageArg by Liu et al. (2022). The dataset consists of tweets containing image and text pairs that deal with two controversial topics, namely gun control and abortion. The first of the two tasks of ImageArg 2023 deals with the classification of argumentative stances in tweets. An image-text pair is to be classified according to whether it supports or opposes one of the controversial topics. For this purpose, Torky et al. worked with the text and employed a BERT model, with which they were able to achieve an F1 score of 0.84 for the classification. They were thus able to train a classifier that can reliably determine the stance of a tweet. The second task deals with the classification of image persuasiveness. The aim was to show whether the image of the tweet makes the associated text more convincing. To do this, the researchers used concatenated CLIP embeddings. They used 512-dimensional CLIP embeddings for the image and the text and then concatenated these into a uniform 1024-dimensional CLIP embedding per tweet. They trained a binary classifier on these composite CLIP embeddings. This classifier has achieved an F1 score of 0.56. Torky et al. conclude that their model cannot satisfactorily recognize whether an image supports a text or not. Nevertheless, it was found that the model was better able to identify combinations in which the image does not make the text more convincing than in the opposite case.

## 3.2 CLIP retrieval and LAION-5B

As previously outlined in Section 2.3 CLIP is a multimodal model, which can represent images and texts in one and the same vector space and thus makes it possible to compare both modalities. Thus, CLIP is a relevant model for the performed analyzes and experiments in the present work. Furthermore, CLIP is crucial for generating the dataset used for these experiments. Here, the CLIP retrieval of Beaumont (2022) based on the LAION-5B[4] dataset is used. Both the LAION dataset and the CLIP retrieval are described in this section.

[3]https://imagearg.github.io

[4]https://laion.ai/blog/laion-5b/

**LAION-5B** The LAION-5B dataset is a large multimodal dataset consisting of 5.85 billion image-text pairs. This dataset is the first of its kind to made public and allows research on multimodal language-vision models within a wide community. The dataset is based on Common Crawl's WAT files[5]. These WAT files contain raw HTML data from numerous web pages. From the provided HTMLs, the initial step was to identify all IMG tags that possess an alt text attribute. Subsequent to this, language detection was conducted on these descriptions, categorizing them into three segments: texts definitively identified as English, texts identified as another language with certainty, and texts where no specific language could be confidently determined. The last category often included descriptions that were very short or e.g. contained proper names. With the help of CLIP, the over 50 million candidates were trimmed, allowing subsequent filtering to be applied to only the appropriate image-text pairs. Therefore, CLIP embeddings were calculated for both the image and the alt text. The cosine similarity was calculated between the two embeddings. Samples with a similarity below a threshold (0.28 for English language and 0.26 with MCLIP for multilingual samples) were eliminated. The image-text pairs were then filtered and only pairs meeting the filter criteria were added to the dataset. E.g. samples containing alt texts that are only 5 characters long or an image size of less than 5 KB are not included in the dataset. Image files that are too large are also not included, as they could potentially be DOS bombs. In addition, duplicates were also removed. This resulted in a dataset of 5.85 billion image-text pairs in the dataset after preprocessing and filtering based on the above criteria.

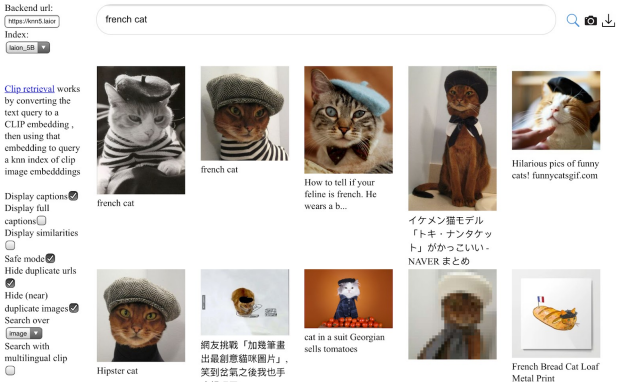



**Figure 3.5:** Screenshot of the CLIP frontend by Beaumont. CLIP-retrieval on the LAION-5B dataset can thus be carried out easily.

[5]https://commoncrawl.org

**CLIP-retrieval** Beaumont has developed CLIP retrieval[6] to easily search on LAION datasets using CLIP embeddings. For this purpose, a large KNN index was created using `autofaiss`[7] for the LAION-5B and the LAION-400M dataset. This makes it possible to query images or texts and find images with similar CLIP embeddings in the large datasets. CLIP retrieval enables effortless calculation of CLIP embeddings and retrieval of their nearest neighbors from the index, i.e. the CLIP embeddings that are most similar to the calculated one. The input for quering the index can be either a text or an image. Then, a CLIP embedding is created from the input. This embedding is used to search the index and find similar CLIP image embeddings. To make it easier to use Beaumont has also created a frontend, which can be seen in Figure 3.5.

[6]https://github.com/rom1504/CLIP retrieval
[7]https://github.com/criteo/autofaiss

# Chapter 4

# Experiments & Results

In the following chapter the experiments and results of this work are presented. Section 4.1 deals with the dataset on which the following experiments were performed. We created this dataset specifically for this work because there were no usable datasets for the use case at hand. The process of creation and further information about the dataset are thus described in the first part of this chapter.

In Section 4.2, we describe various analyzes on captions of the dataset. Specifically, the analyzes include a Bag of Words analysis, a Topic Modeling, and a similarity analysis, all performed on a selection of image captions from the dataset. Each of these analyzes is explained and the workflow is described. Since meaning of images always depends on the individual image itself and cannot be generalized, some examples and the corresponding results are presented for each analysis. Each of the analyzes is then evaluated.

Section 4.3 discusses an approach for extracting a hidden meanings from an image. Initially, we present the development of a neural network intended to produce a hidden meaning embedding for an image. The second part deals with the possible translation of this embedding. For this experiment the results are explained directly afterwards.

Figure 4.1 illustrates an overview of the procedure and the experiments conducted throughout chapter.

## 4.1 Creation of a Stock Image Dataset

This section outlines the compilation of the data that serves as the basis for the experiments carried out as part of this thesis. The first subsection describes the creation of the stock photo dataset itself. The second part deals with the crawling of captions that appear together with images from the dataset.

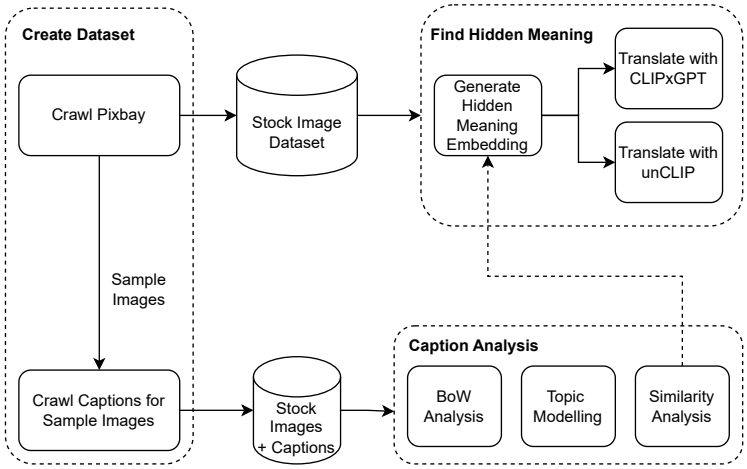



**Figure 4.1:** Shown is our workflow for the experimental setup of this work. First, we create a stock image and caption dataset by crawling Pixabay and retrieve captions of the stock images in the CLIP retrieval model. Next, we analyze the captions using various linguistic analyses. Finally, we train a model for the generation and extraction of hidden meanings based on the stock images and the captions.

These captions form an important part of the data, as the majority of the experiments deal with the analysis of these captions.

### 4.1.1 Scraping Stock Images

There are many large image datasets that can be used for all sorts of tasks, such as MS-COCO[1], for object recognition, or the Flickr image dataset[2] for image captioning. However, none of these image datasets are usable for the present work because this thesis exclusively focuses on stock photos with stock photos. As described in Section 2.2, stock images can be found on various websites, in newspaper articles, advertisements, and in many other mediums. Typically, image datasets are collected from the internet by crawling, e.g., Google image search results for specific categories, such as in the MS-COCO dataset, or by collecting images from Flickr (Lin et al., 2015). Therefore, it is very likely that available image datasets contain stock images. However, it is difficult to distinguish between stock photos and other photo types, making it impossible to filter out stock photos from a dataset that contains both stock and non-stock photos. In order to perform analyzes based on stock photos, it is necessary to use a dataset that consists only of stock photos. Such a dataset did not exist before, so we created a new dataset for this work. To create the

[1]https://cocodataset.org
[2]https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

```
{
"total": 4692,
"totalHits": 500,
"hits": [
    {
        "id": 195893,
        "pageURL": "https://pixabay.com/en/blossom-bloom-flower-195893/",
        "type": "photo",
        "tags": "blossom, bloom, flower",
        "previewURL": "https://cdn.pixabay.com/photo/2013/10/15/09/12/flower-195893_150.jpg"
        "previewWidth": 150,
        "previewHeight": 84,
        "webformatURL": "https://pixabay.com/get/35bbf209e13e39d2_640.jpg",
        "webformatWidth": 640,
        "webformatHeight": 360,
        "largeImageURL": "https://pixabay.com/get/ed6a99fd0a76647_1280.jpg",
        "fullHDURL": "https://pixabay.com/get/ed6a9369fd0a76647_1920.jpg",
        "imageURL": "https://pixabay.com/get/ed6a9364a9fd0a76647.jpg",
        "imageWidth": 4000,
        "imageHeight": 2250,
        "imageSize": 4731420,
        "views": 7671,
        "downloads": 6439,
        "likes": 5,
        "comments": 2,
        "user_id": 48777,
        "user": "Josch13",
        "userImageURL": "https://cdn.pixabay.com/user/2013/11/05/02-10-23-764_250x250.jpg",
    },
    {
        "id": 73424,
        ...
    },
    ...
]
}
```


**Figure 4.2:** Sample response for a Pixabay API request. Source: https://pixabay.com/api/docs/

dataset, we crawled the stock photo site Pixabay[3]. We chose Pixabay because they provide free stock photos and an API for image retrieval.

Using Pixabay's API, for one search term up to 500 images can be crawled at once. The API's response to a request contains various information per image, such as the URLs for the image in various sizes, the image's Pixabay-internal ID, the number of likes an image has on the Pixabay page, and a list of tags associated with the image. An example response to an API request can be seen in Figure 4.2. However, since Pixabay expects a search term as a request and as mentioned above a maximum of 500 images are returned per search term, it is necessary to create a list of search terms in order to create a large dataset. The dataset is based on a list of search terms that we created using ChatGPT 3.5[4]. To do this, ChatGPT was queried for typical topics for stock photos and a list of 133 topics was created. Terms included in this list are e.g. *office*, *people* and *family*. For each of these topics, a request was sent to the API and the 500 images per search term were downloaded in `webFormat`. The dataset consists of 66,277 stock at the end.

### 4.1.2 Caption Crawling

The meaning of images cannot always be gleaned from the image alone. Often the text-image relationship has an impact on the meaning of an image, as described in Section 2.1. Also, photographers and stock image agencies themselves cannot always foresee the context in which images will ultimately

[3]https://pixabay.com/de/
[4]https://chat.openai.com

be used and how recipients will actually interpret the images, as described in 2.2. Therefore, it is not enough to use a dataset consisting only of the images. In order to learn for each image the context in which it is actually used in the real world, a set of captions that occurred with the image in the internet. The various captions used with a single image can offer insight into how the image is perceived and comprehended by those who use ist. Therefore, to analyze the context of the image, we tried to gather a large number of captions per image. For this purpose, the CLIP-retrieval on the LAION-5B dataset described in Section 3.2 was used. The CLIP client from Beaumont's CLIP-retrieval was used to crawl for duplicate images and their corresponding captions on the index *laion5B-L-14* with the CLIP model *ViT-L/14* from openAI[5]. In a first exemplary analysis, a few sample images from the stock photo dataset were searched using the CLIP-retrieval frontend[6]. Using the *laion5B-L-14* index, the most appropriate images were identified for a sample of reference images. Therefore, we chose to utilize this index in conjunction with the corresponding CLIP model. To find captions for the images from the dataset, one image at a time is used as a reference image. To crawl the captions, an embedding is calculated for the reference image using the CLIP model. The reference embedding is then used as input for the CLIP client for querying the index. After initial investigations we have discovered that there were not much more than 200 duplicates in the LAION-5B dataset for many of the examined images. This may be due to the relatively small image database Pixabay, which is free but much less used than e.g. Getty Images. Therefore, and to make the dataset a bit more balanced, we limited the expected number of results per image to 200. We specified 1000 as the number of results to be returned by the CLIP client, even though we only want to collect 200 captions. Since we are only looking for duplicates, i.e. images that are exactly the source image, a duplicate score is calculated for each of the returned result images. The duplicate score is the inner product of the embedding of the source image and the embedding of the currently viewed result image returned by the CLIP client. Not every image with a high duplicate score is actually a duplicate. Therefore, we conducted a previous test to determine the approximate value at which the resulting images are no longer considered duplicates. We set the duplicate detection threshold to *0.975* based on an analysis of a limited number of test images from the LAION-5B dataset. Our preliminary tests showed that the majority of the analyzed images had no more than 200 duplicates, which led us to set this specific threshold to make the search for duplicates more efficient. Each result image that falls below this threshold is not used, since it is assumed that it

[5]https://huggingface.co/openai/clip-vit-large-patch14. During the course of the thesis, this particular model was used whenever CLIP was used.

[6]https://rom1504.github.io/clip-retrieval/

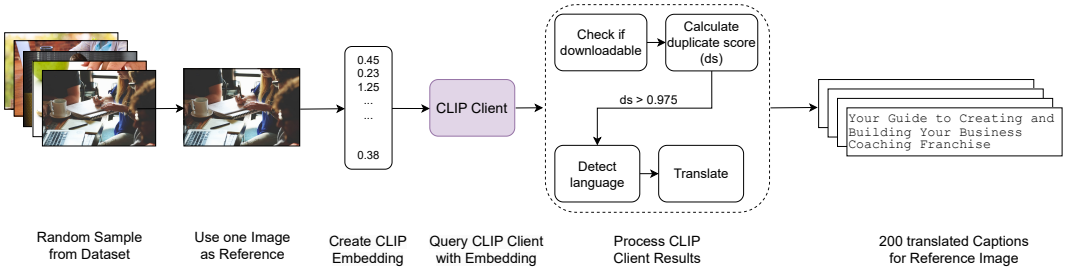



**Figure 4.3:** Overview of the caption crawling process. One image from a set of images is used as a reference image. For this image an embedding is computed using CLIP. This embedding is used to query the LAION-5B dataset using the CLIP client. The results obtained by the CLIP client are processed and stored after the downloadability test and reaching the duplicate score. To obtain the final caption dataset, the language of the captions is recognized and a translation is performed. The process is repeated for the other images in the sample.

shows another merely similar image. The corresponding caption is saved for each result image that is marked as a duplicate. Once 200 captions were found for an image, the search is terminated and the same procedure is followed with the next image in the dataset. In addition, another condition that causes the search to be aborted was introduced. This was done because searching for images and the calculating the duplicate score is time-consuming. Examining a first test results, we found that for some images, despite retrieving 1000 results, not enough images exceeded the duplicate threshold. Therefore, once 50 of the analyzed images have a low duplicate score, the search is terminated because it is presumed that insufficient images can be found. This value resulted also from the exemplary examination of the first results. Since the results returned by the CLIP client are sorted by a similarity score, it is unlikely that after the first 50 images below the duplicate threshold there would be enough duplicates at all. For the calculation of the duplicate value, an attempt is made to download the images via the URL returned by the CLIP client. If this was not possible, no value could be calculated and the next result image was checked, therefore it was necessary to request 1000 result images from the CLIP client despite both abort conditions. The URL was also used to check if there is already an caption with the same image URL in the result set. For each image from the dataset for which 200 results could be found, an entry was then added to a caption file in json format. Each image key has a list of occurences as a value. The entry for each occurence contains an URL, a caption, an ID returned by the CLIP client, and the calculated duplication score. Since the computation is very time consuming, as mentioned above, despite introduced parallelization of crawling and abort conditions, captions were only crawled for

a randomly selected part of the dataset of 1000 images. This results in overall 200,000 captions.

Our analysis of the captions included text similarity and word frequency. To ensure consistent results, it was necessary to analyze all captions in the same language. However, limiting the analysis to English text would have further reduced the already small number of available captions per image. Therefore, we decided to translate non-English captions in order to include them in the analyzes. This decision allowed us to consider a wider range of data, although it posed the challenge that translations can affect the original word meanings and contexts.

Therefore, we detected the language of the crawled captions with the help of `langdetect`[7]. Captions for which the language could be detected were translated with the help of *NLLB-200*[8]. The model NLLB-200 was chosen because it is easy to use. The recognized language and the translation were additionally included in the result json. The recognized language was included to enable possible following analyzes on language.

| Language | Absolut | Relative (%) |
|---|---|---|
| English (en) | 61792 | 30.90 |
| French (fr) | 14413 | 7.21 |
| Spanish (es) | 13402 | 6.70 |
| German (de) | 13215 | 6.61 |
| Russian (ru) | 10247 | 5.12 |
| Polish (pl) | 10008 | 5.00 |
| Italian (it) | 9774 | 4.89 |
| Portuguese (pt) | 7189 | 3.59 |
| Japanese (ja) | 7096 | 3.55 |
| Dutch (nl) | 6158 | 3.08 |
| Miscellaneous (misc) | 46706 | 23.35 |

**Table 4.1:** The languages and the corresponding absolute and relative frequencies are shown. In addition to the 10 most frequent languages and the total number of the 44 other languages.

As can be seen in 4.1, 30.9% and therefore the majority of the captions are already in English. However, almost 70% of the crawled captions could not be used without translation. In total, there are captions in 54 languages. For the 44 languages grouped under misc, there is an average of 1061.5 captions in the dataset. The number of different languages recognized also suggests a

[7]https://pypi.org/project/langdetect/

[8]https://huggingface.co/facebook/nllb-200-distilled-600M

future analysis at language level, as the meaning in images also depends on the social and cultural background of the recipients, as described in Section 2.1. Different language could be an indicator of different cultural backgrounds, so that an analysis in this direction would be possible.

## 4.2 Caption Analysis

The captions of the individual images were used for the following analyzes. The following section describes the procedure for each of the analyzes carried out and shows the results using a number of examples. To get a first impression of the contexts in which an image is used, a simple BoW analysis was performed, which is described in Section 4.2.1. For the same purpose, a topic analysis using BERTopic was applied to the captions, which is presented in Section 4.2.2. In the next experiment, described in Section 4.2.3, the similarities between the CLIP embeddings of an image and the corresponding captions are examined. This was to test the possibility to distinguish the meaning and description of an image with the help of CLIP. Based on the results of this experiment, an analysis of the similarities of embeddings created with a text-only model, namely SBERT, was performed.

**Generating Description for an Image** Since some of the analyzes are based on text only, an Image2Text model was used to create a descriptions for all of the stock photos in the dataset. In order to perform the text analyzes and to work out a difference between what can be seen on an image and what the image means, the assumption was made that the generated description of an image can stand in for the image. Thus, this description denotes what can be seen in the image and should thus be distinguished from the *hidden meaning* of the image. To ensure that the generated caption describes satisfactorily what is shown on the image, two of the most popular Image2Text models from Huggingface[9] were tested and evaluated. The two models are the ViT-GPT2 Image Captioning model from ydshieh[10] and the BLIP Image Captioning model from Salesforce[11]. For the test, descriptions were generated for the first 100 images in the dataset with both models. These 100 images and the corresponding descriptions were then manually reviewed to determine which of the generated captions better described the image. The descriptions generated by BLIP were better for 90 of the 100 images. To help you make an unbiased decision, we have developed a simple decision tool that allows you

[9] https://huggingface.co
[10] https://huggingface.co/ydshieh/vit-gpt2-coco-en
[11] https://huggingface.co/Salesforce/blip-image-captioning-large


**Figure 4.4:** Decision tool for evaluating which Image2Text model generates the best results for our use case: ViT-GPT2 or BLIP.

to view the image and its description without knowing which model produced which description. The structure can be seen in Figure 4.4. The decision was made to use the BLIP model for generating descriptions.

### 4.2.1 Bag-of-Word-Analysis of Captions

The initial analysis of the captions of stock photos focuses on word frequencies. The idea of this analysis is based on the assumption that captions for a particular image can convey meaning through the words used. Therefore, this analysis identifies the most common terms in the captions to provide insights into the image's contexts and potential interpretations, regardless of what is shown on the image itself. The first step of this analysis is the preprocessing of the captions. This involves transferring the captions into cleaned word lists. Therefore, special characters (!,?.) and numbers are removed. For every cleaned caption a word list is created. Afterwards, stop words are removed from the word lists. The list of stop words is sourced from the `nltk` library[12]. The final step of preprocessing the captions is lemmatizing the words to be able to join words with the same root. This preprocessing is done with each of the captions crawled for one image. The description of that image generated by the

[12]https://www.nltk.org

Image2Text model is preprocessed in the same way. After the preprocessing the word lists of all captions together are added together. This creates a list that contains all the words that occurred in any caption with the image to be viewed. From this list every word is removed, which the preprocessed description list contains. The assumption is that words contained in the description list are objects or actions, which depicted in the image. Since the description is representative of the image, this assumption is made. Not all image captions convey meanings, sometimes they simply describe them. Therefore, it is necessary to eliminate those descriptive words from the caption word list. If all words from the description word list are removed from the caption word list, the remaining words should refer to the imags's meaning. To find out which words are most frequently associated with the image without actually being on it, the final step of this analysis is to count the word frequencies of the filtered caption word list. The 10 most frequently occurring words are an indication of the *hidden meaning*. These are words that were frequently associated with the image even though they do not appear in the description of the image, thus it can be assumed that recipients most frequently associated these top 10 words with the images. An overview of the analysis and its steps can be seen

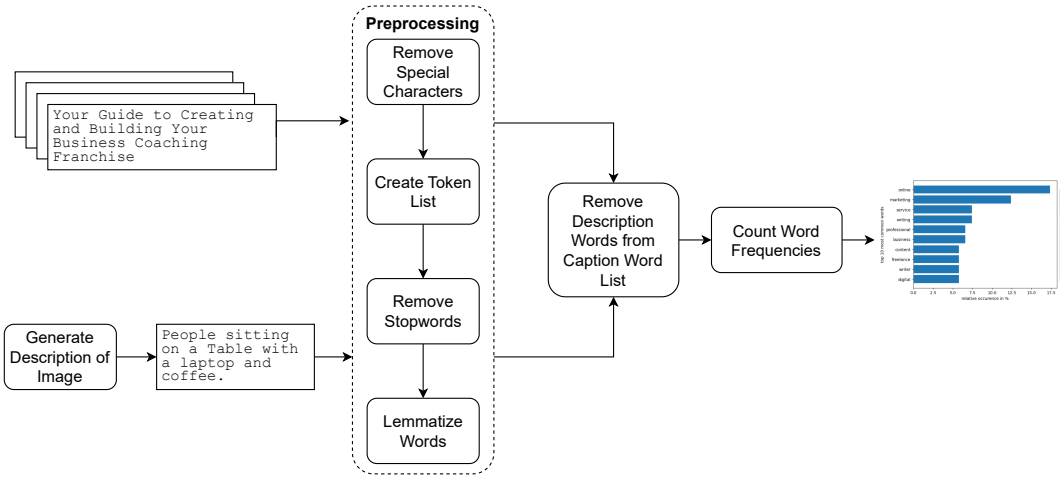
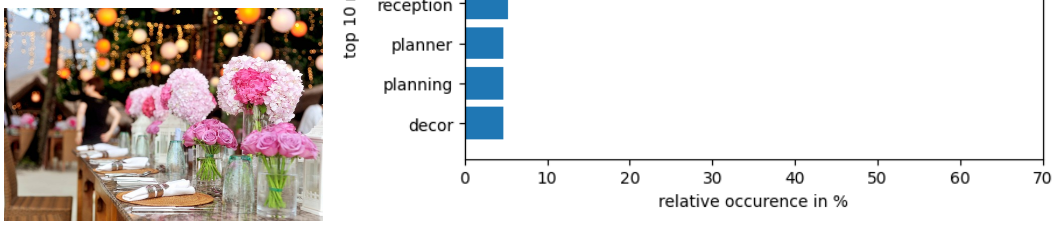



**Figure 4.5:** Overview of the steps of the BoW analysis. An image description is generated for the image to be analyzed. Both the captions belonging to the image and the caption are preprocessed. Special characters and stopwords are removed and the texts are tokenized and lemmatized. This results in a caption list with words that occur in all captions and a list with the words from the description. The words from the description are removed from the entire caption word list. The word frequencies are then counted and the top 10 most frequent words are shown in a bar chart.

in Figure 4.5.

**Results** Since this analysis is limited to one image at a time, examples with corresponding analysis results are presented here.

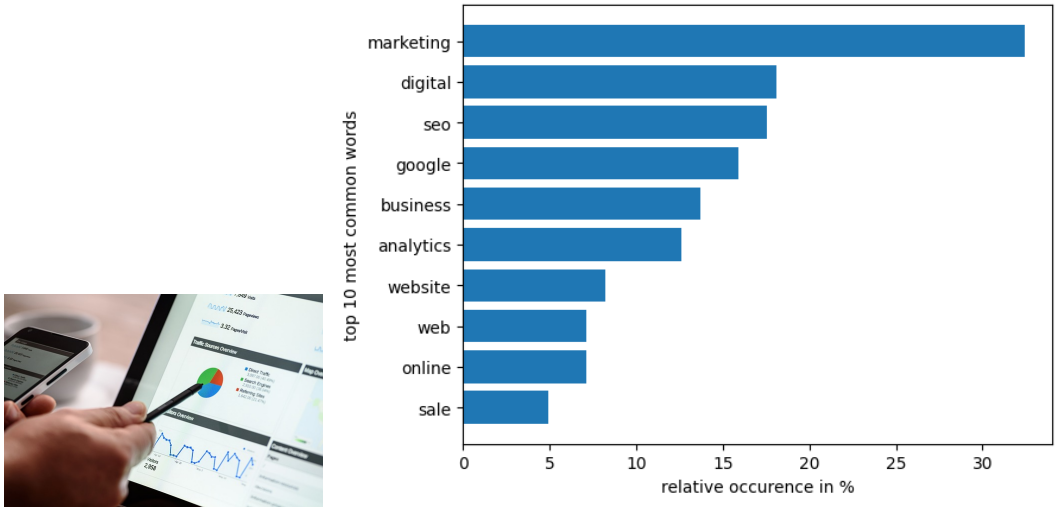
Figure 4.6 displays the result of the BoW analysis for the dataset image *img_ 015577* and the image itself. The image is shown on the left (Figure 4.6a). In the foreground of the image, there is a table that has been arranged and decorated with pink and rose flowers. The background is unfocused but displays a location decorated with light chains and lanterns in which another table can be seen. The description generated by BLIP is *there are many vases of flowers on the table with place settings*. The results of the BoW analysis can be seen in Figure 4.6b. There the relative frequencies of the top 10 words from the crawled captions are shown. The value thus tells in what percentage of the captions the individual words occurred. The word that occurred by far the most in the captions is *wedding*. Although the location could be decorated for any event, wedding seems to be the word most people who used this image associate with it. Words that are otherwise commonly associated with the image are more obvious at first glance, such as *decoration* or *event*.




**(a)** img_015577.jpg **(b)** Result of BoW Analysis for img_015577.jpg

**Figure 4.6:** Image 015577 and corresponding result of BoW Analysis

Figure 4.7 shows the result of the BoW analysis for the image *img_ 000025*. In the image itself (Figure 4.7a), a hand can be seen rather blurredly in the foreground, pointing at a screen with a pen. On the left side of the image, half of a cell phone can be seen with some numbers on it. On the screen itself, which takes up most of the image, there are numbers, statistics, and charts. The generated description of the image is *someone is using a cell phone and a laptop computer to display graphs*. The result of the BoW analysis (Figure




**(a)** img_000025.jpg **(b)** Result of BoW Analysis for img_000025.jpg

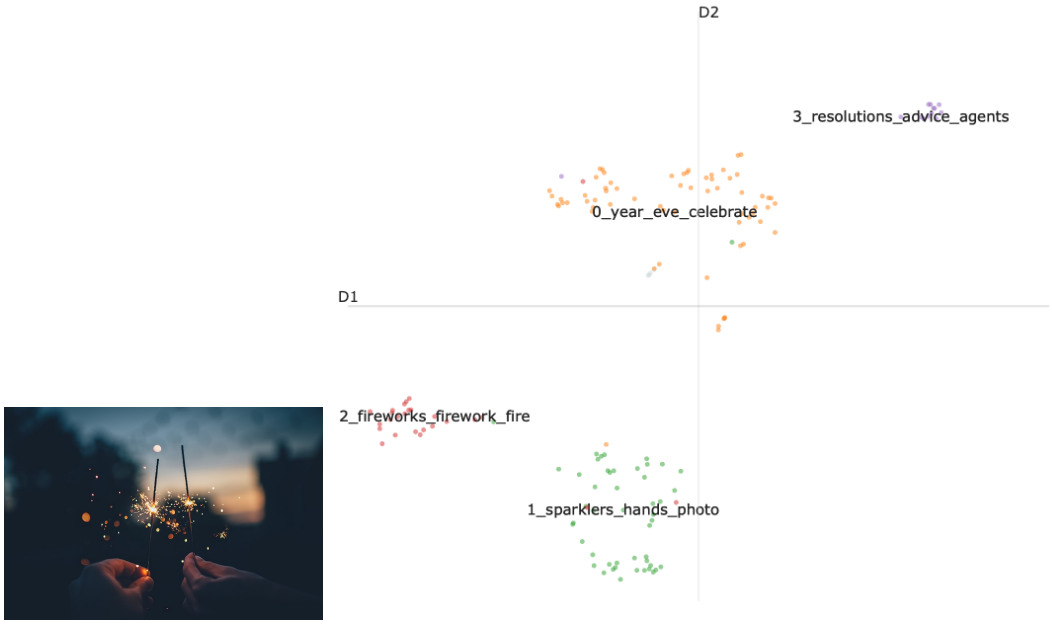
**Figure 4.7:** Image 000025 and corresponding result of BoW Analysis

4.7b) again shows the relative occurrences of the top 10 words from the crawled captions. The word that occurs most frequently with the image is *marketing*. The image is often used in the context of selling things, business analytics and search engine optimization (SEO). A closer look at the image reveals that words like *visitors* and *search engines* can be seen on the screen. Therefore, it is understandable why words that emerge from the analysis as the most common words are so often associated with the image. Nevertheless, marketing or Google is not something that comes directly from the image, it is rather a symbolic image that can indicate such issues and therefore is often used in this context.

**Conclusion** The analysis based on BoW indicates a discrepancy between the what is shown on an image and its meaning. The manual evaluation of the images and their most common terms provides some explanation as to why the image is utilized in these contexts. The findings from this analysis offer valuable insights into the contexts in which an image is frequently used. Despite its simplicity, the analysis has already provided useful insights.

### 4.2.2 Topic Modeling of Captions

The present work is about capturing meanings from images and differentiating them from what is shown. In the previous analysis, we connexted certain words with the corresponding picture. Therefore, the meaning of an image should be




**(a)** img_015522.jpg **(b)** Result of Topic Modeling for img_000223.jpg

**Figure 4.8:** Image 015522 and corresponding result of Topic Modeling

extracted through words that commonly co-occur with the picture. It is also assumed in the present thesis that the meaning of an image is reflected in the topics that are related to the image. Therefore, topic modeling was applied to the captions in the following analysis. As in the previous analysis, the contexts in which an image is used were to be elaborated in order to gain insight into the meaning of an image. For topic modeling, the crawled captions for an image were preprocessed in the first step. For this purpose, as in the BoW analysis, the special characters, numbers and stop words were removed. Then, the topic model *BERTopic*[13] from Grootendorst (2022) was used. To limit the number of words per topic that mean the same thing, we used *MaximalMarginalRelevance* like suggested by Grootendorst with a diversity score of 0.2.

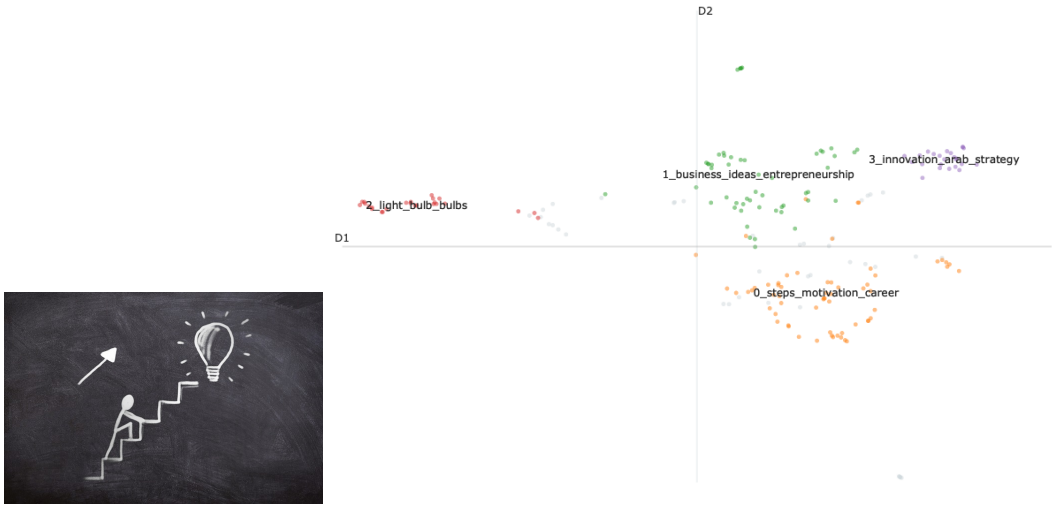
**Results** In order to show to what extent topic modeling could be used to find underlying concepts in captions, the following part presents some images and their respective results from topic modeling with BERTopic.

Figure 4.8 displays an image and the results of applying topic modeling to its captions. In the stock photo (Figure 4.8a), you can see two hands in the foreground, each holding a lit sparkler. The background is out of focus, but the sky is probably just after sunset and silhouettes of trees and a house

[13]https://maartengr.github.io/BERTopic/index.html

can be glimpsed. Figure 4.8b shows the results of the topic modeling. The diagram displays dots that represent individual documents, in this case the captions. Four topics were retrieved from BERTopic. The different topics can be identified by the colors of each document item shown in the diagram. At first glance, the four topics can easily be divided into two groups, since two of the topics are very close to each other, but at the same time far away from the other two topics. The topics, which are shown on the diagram in the lower left-hand corner, are labeled *1_sparklers_hands_photo* and *2_fireworks_firework_fire*. If we compare these topics with the image, it becomes apparent that the names primarily include items that are related in some way to the visual content. As described earlier, there are hands and sparklers in the picture, as mentioned in topic 1. An additional clue that topic 1 is descriptive is the prediction of the description *someone holding sparklers in their hands with a dark background* to this topic. In a way, sparklers can also be counted as fireworks, so the terms that appear frequently in topic 2 are also related to what is seen in the image. Thus, the captions from topics 1 and 2 are primarily concerned with describing the image. However, when examining the remaining topics, it becomes clear that none of the objects depicted in the picture can be found in the names. The larger of the two topics, *0_year_eve_celebrate*, appears to deal with New Year's Eve, and is therefore related to the meaning of the image. Captions that fall into this topic include 'Where to celebrate new years eve?' or 'Happy New 2017 with the right New Year's Promises'. The sparklers in the picture make many people think directly of New Year's Eve and this is reflected in the corresponding captions. Captions such as 'How to Live Up to New Year's Resolutions' or 'New Year: The most popular resolutions made every year' can be found in *3_resolutions_advice_agents*. New Year is a common topic of discussion, primarily regarding resolutions. These resolutions often relate to New Year's Eve and the conclusion of a year. So, it can be said that hidden meanings for the image 4.8a include New Year and resolutions. Neither is in any way depicted in the image, and yet apparently many people who have used this image associate these things with it. So, topic modeling here has given a good indication of contexts related to the meaning of this image.

Another example where topic modeling worked quite well can be seen in Figure 4.9. In the picture 4.9a there is a drawing presumably with chalk on a black board. The drawing contains a stick figure climbing up a staircase on hands and feet. Above it is an arrow pointing to the upper right, where the staircase ends. At the top of the stairs is a presumably glowing light bulb depicted. Figure 4.9b visualizes the results of topic modeling for this image. For this image four topics were found by BERTopic. At first glance, it is apparent that the documents for one of the topics are mapped further away from all other documents. The name of that topic is *2_light_bulb_bulbs*. This name




**(a)** img_000223.jpg **(b)** Result of Topic Modeling for img_000223.jpg

**Figure 4.9:** Image 000223 and corresponding result of Topic Modeling

alone seems to indicate that the topic is based primarily on descriptive captions because a light bulb can be seen in the image. Captions belonging to this topic include *Drawing lightbulb with blackboard* or *Graphic drawing of a person climbing a ladder to a lit lamp in the blackboard*. Thus, this topic does indeed seem to contain mostly descriptive captions. An assumption could be made that the other three topics are primarily focused on the context in which the image is used, therefore, tend to emphasize the meaning of the image. Another interesting topic is *0_ steps_ motivation_ success*. At first glance, *steps*, is a word that can also be related to what is seen in the image, however, steps does not necessarily have to be related to stairs, as there are other meanings, as seen in captions *10 Steps to Successful Starting as a Motivator* or *Five simple steps to an inspiring speech*. Nevertheless, the steps that can be seen in the image are a symbol of, e.g., a *career ladder*. Other meanings that the image can convey, can be found in the two topics of *1_ business_ ideas_ entrepreneurship* and *3_ innovation_ arab_ strategy*. The light bulb is a frequently used symbol for ideas and is therefore likely to be associated with words such as *entrepreneurship* and *innovation*. A caption that can be representative of topic 1 is *How to Get an Idea for a New Business*.

**Conclusion** With the help of topic modeling, trends could be found in the captions. The results presented here show that it is possible to summarize concepts in the captions using topic modeling and thus to detect certain levels of meaning from the captions. In some cases it is even possible to distinguish

the descriptive topics from the meaningful topics.

### 4.2.3 Similarity Analysis of Embeddings

This experiment deals with the computation of similarities of embeddings. One of the first questions that arose regarding the use of CLIP for the present work was whether CLIP is more likely to capture descriptions or meanings. The question here is whether a high similarity of CLIP embeddings indicates literary similarity or similarity at the level of meaning. This was examined by computing similarities between the CLIP embedding of an image and the CLIP embeddings the crawled captions associated with the image. Following this experiment, we applied a similar analysis using the SBERT language model to the image's texts, i.e. the captions and the description. In the following section, these experiments are considered in more detail and evaluated on two examples.

**Analysis of CLIP Embedding Similarities** The analysis of CLIP embeddings' similarities was based on the assumption that there is a difference between the similarity of embeddings for descriptive texts and the image, and the similarity of embeddings for meaningful texts and the image. If the CLIP embeddings of texts holding meanings related to the image and the reference image have a significantly higher similarity compared to that of descriptive texts and the image, or vice versa, the similarity of CLIP embeddings can serve as a clue to detect the meaning. As described in Section 2.3, CLIP was trained primarily on image-text pairs crawled from the internet. Large image-text datasets crawled from the internet, e.g. the LAION-5B dataset described in Section3.2, often consist of images with their associated alt text attributes. Unfortunately, there is no detailed information about the dataset CLIP was trained on and the dataset is not publicly available. However, it is possible that the dataset used to train CLIP contains alt attributes associated with images. According to W3C guidelines[14], the alt text of an image should not necessarily describe the visual characteristics of an image. Instead, the alt text should rather convey the same meaning as the image. This guideline for using the alt attribute suggests that CLIP is more likely to capture an image's meaning level than its description, assuming alt attributes were present in the training dataset. For the analysis of similarities, the list of captions per image that was expanded with the description generated by the BLIP model. Since it could not be confirmed whether the caption list included a depiction at all, the inclusion of the generated description was required as a point of reference

[14]https://www.w3.org/TR/2016/NOTE-WCAG20-TECHS-20161007/H37

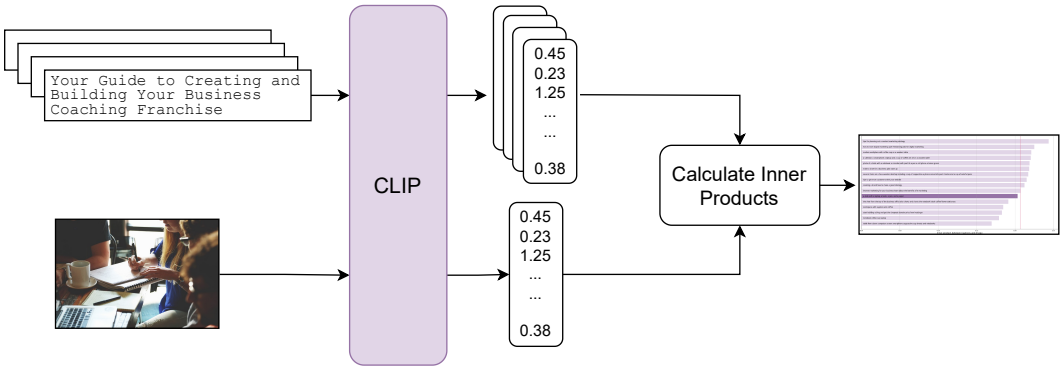



**Figure 4.10:** Overview of the individual steps of similarity analysis with CLIP. CLIP is used to calculate embeddings for the image to be analyzed and the associated crawled captions for the image. These embeddings are used to calculate the similarity between a caption and the image. The similarity values are then plotted on a bar chart.

because the goal was to find out whether CLIP similarity captures meanings or descriptions. CLIP embeddings were calculated for every text, including the description and all captions, as well as for the reference image. Then, the similarities of the CLIP embeddings were calculated. Since only the similarity of the texts to the image was relevant, the cosine similarity of the CLIP embeddings was calculated for each combination of one of the texts with the image. The similarities between the image and texts were sorted. An overview of the analysis and its steps can be seen in Figure 4.10.

**Results of the Similarity Analysis with CLIP** Because including all captions in the resulting diagrams would result in lost overview, a total of 10 descriptive and 10 more meaningful captions were manually selected from the 200 captions per image for each of the two sample images. This ensures that both descriptive and meaningful captions are included in the analysis and that the results remain clear. The images used to evaluate the CLIP embedding similarity analysis are shown in Figure 4.11. The image 4.11a shows a desk from above with a calendar, a cup of coffee, a cell phone, sheets of paper with some diagrams on them, a computer screen, and a cup with pens. The result of the CLIP embedding similarity analysis for this image can be seen in Figure 4.12. In the diagram, the value plotted on the X-axis is the inner product, i.e. the cosine similarity between the embedding of the image under consideration and embedding of the respective caption or generated description. The generated description is highlighted with a darker color to distinguish it from the crawled captions. The caption with the highest similarity (0.24) to the image is *'tips for planning out a content marketing*

**(a)** img_000384.jpg **(b)** img_019768.jpg

**Figure 4.11:** Images for Similarity Analysis

*strategy'* and the caption with the lowest similarity (0.17) is *'table from above computer screen smartphone cappuccino cup sheets and notebooks'*. This first look at the result suggests that image captions that have high CLIP embedding similarity are more likely to be meaning-bearing. This could indicate that CLIP embeddings capture meanings rather than descriptions. However, upon a closer examination of the results, it is evident that there is no identifiable threshold at which the captions become more likely to convey meaning or serve as a description. Instead, we have not observed a trend that would support our assumption that CLIP is better suited to reflect one of the two categories. The generated description has the 3rd highest similarity with 0.23. The captions ranked 6th, 7th and 8th in terms of similarity are also descriptive captions. Furthermore, the caption *start building a blog and get the cheapest domain price from hostinger* has one of the lowest similarities and does also not fit among the descriptive captions.

The second image presented to display the results of the CLIP similarity analysis is depicted in Figure 4.11b. It shows a meadow or a field in the foreground and a few wind turbines in the background during sunrise or sunset. The results of the CLIP similarity analysis for this image are shown in Figure 4.13. Here, the caption that has the most similar CLIP embedding to the image is *'photo of a wind farm at sunrise'* with a similarity value of 0.27. The caption with the lowest similarity (0.19) is *'bid signs agreement with government to develop energy projects'*. So initially, this appears to be the opposite of the first example at first glance. The caption with the highest similarity is more descriptive and the one with the lowest similarity is more focused on conveying meaning. But again, a closer examination of the results shows that no trend emerges. Among the 5 captions with the highest similarity, 3 are descriptive and two are about energy production and technology, so they are more meaning-bearing.


**Figure 4.12:** The results of the CLIP similarity analysis for image img_000384.jpg are displayed. Captions are represented on the y-axis. To enhance visibility, the bar for the generated description of the image is darker compared to the other bars. The x-axis indicates the inner product between the caption and the image. The average similarity values are marked by the horizontal line.

The captions shown here are of course only a small subset for both examples. To still get an idea of how similar the similarity values of all the crawled captions for those two images are, we calculated the standard deviation of the total of similarity values for both images. The standard deviation of all the similarity values of image 4.11a is 0.022 and the standard deviation of all the similarity values of image 4.11b is 0.022. The values are both very low, indicating that all the inner products of the respective image embeddings with the embeddings of the crawled captions are all very similar. This suggests that it is difficult to divide the set of crawled captions into two groups, meaningful

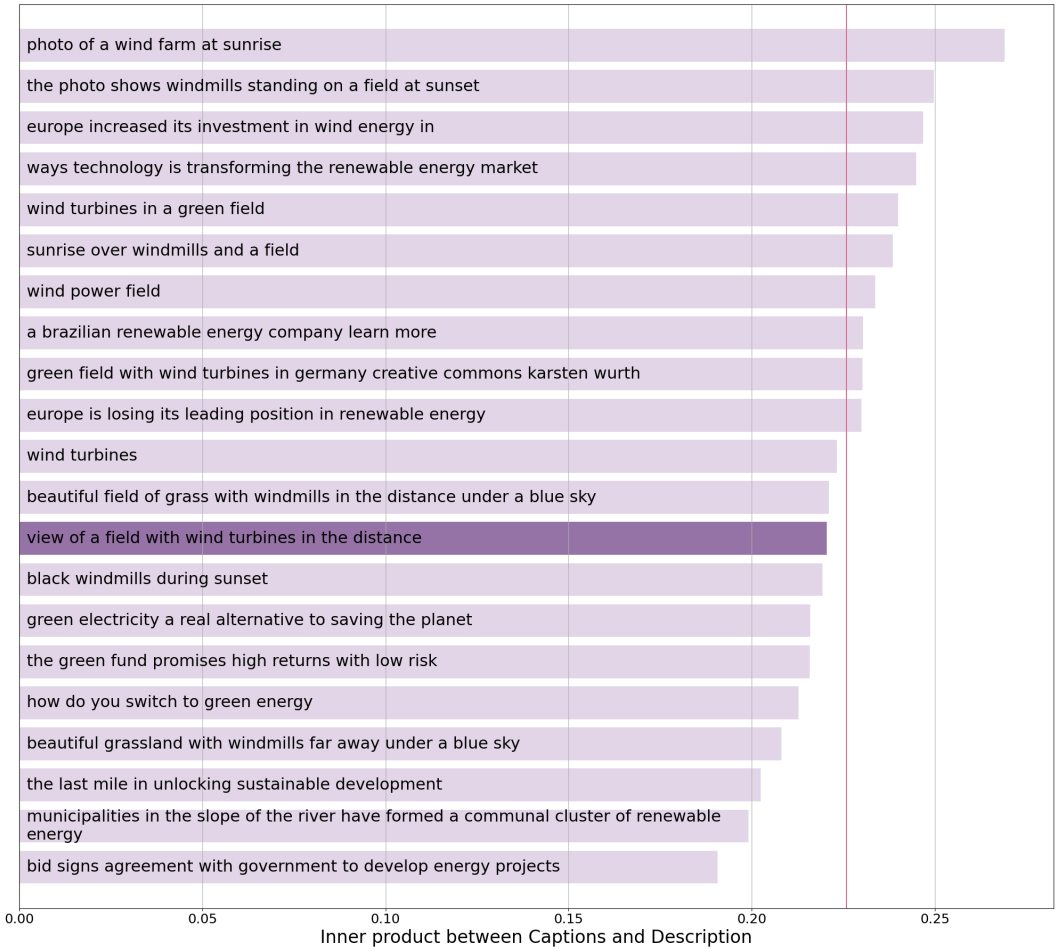



**Figure 4.13:** The results of the CLIP similarity analysis for image img_019768.jpg are displayed. Captions are represented on the y-axis. To enhance visibility, the bar for the generated description of the image is darker compared to the other bars. The x-axis indicates the inner product between the caption and the image. The average similarity values are marked by the horizontal line.

and descriptive.

To verify our assumption that CLIP embedding similarities could be useful for distinguishing between descriptive and meaningful, results were obtained on a larger scale in the course of the work. However, unfortunately no significant difference between the CLIP embedding similarities could be found. The similarities were always very close in the analyzed images. This only allowed the conclusion that CLIP captures descriptions as well as meanings. This can also be concluded from the fact that CLIP was trained on exactly such image-caption pairs as were used here in the analysis. Image captions cannot

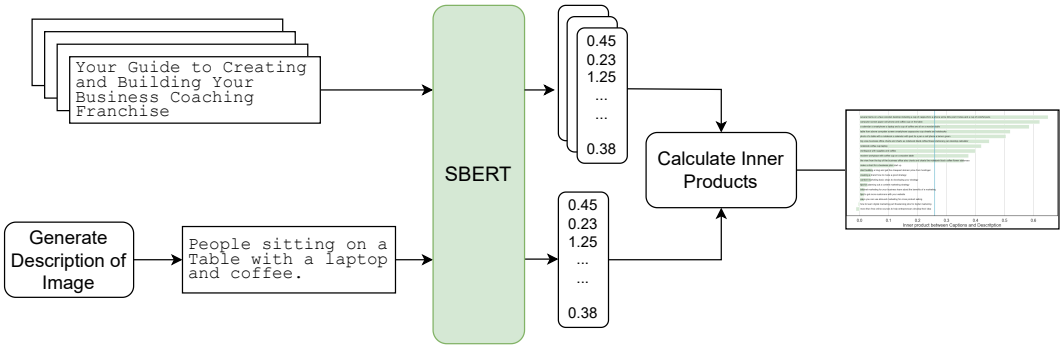



**Figure 4.14:** Overview of the individual steps of the similarity analysis with SBERT. A description is generated for the image to be analyzed. Embeddings are calculated for the generated image description and the corresponding crawled captions with the help of SBERT. The similarity between each caption embedding and the description embedding is calculated. These similarity values are then plotted on a bar chart.

generally be classified as only descriptive or just meaning-bearing. Additionally, it cannot be determined whether the quality of CLIP's training data was good since neither our data nor CLIP's training data underwent any cleaning or annotation. Thus, the assumption that a particular CLIP similarity could be an indication of meaning of an image could unfortunately not be substantiated. Thus, CLIP was not usable for the present work in the way it was intended.

**Analysis of SBERT Embedding Similarities** Since the previous analysis did not yield satisfactory results in the experiments, the question arose whether it would be possible to find differences between descriptive captions and meaningful captions by using a LM namely SBERT[15]. However, since a multimodal model like CLIP, is not utilized, the corresponding image can no longer be considered as a reference. So now that only text was used again, the description of the image created by BLIP was used as reference, as in Section 4.2.1 and 4.2.2. In order to be able to perform the experiment in this way, the assumption was again made that the generated description can be representative of the image. The list of captions crawled per image was utilized in this analysis. This time, SBERT was used to compute the embeddings for the captions. In addition, an embedding was also calculated for the generated description of the image. Since in this analysis assumes that the description represents the image, similarities were calculated between a caption and the description respectively. Again, cosine similarity was used. As in the previous

[15]https://www.sbert.net

analysis, the similarities were sorted and analyzed in more detail. An overview of the analysis and its steps can be seen in Figure 4.14.

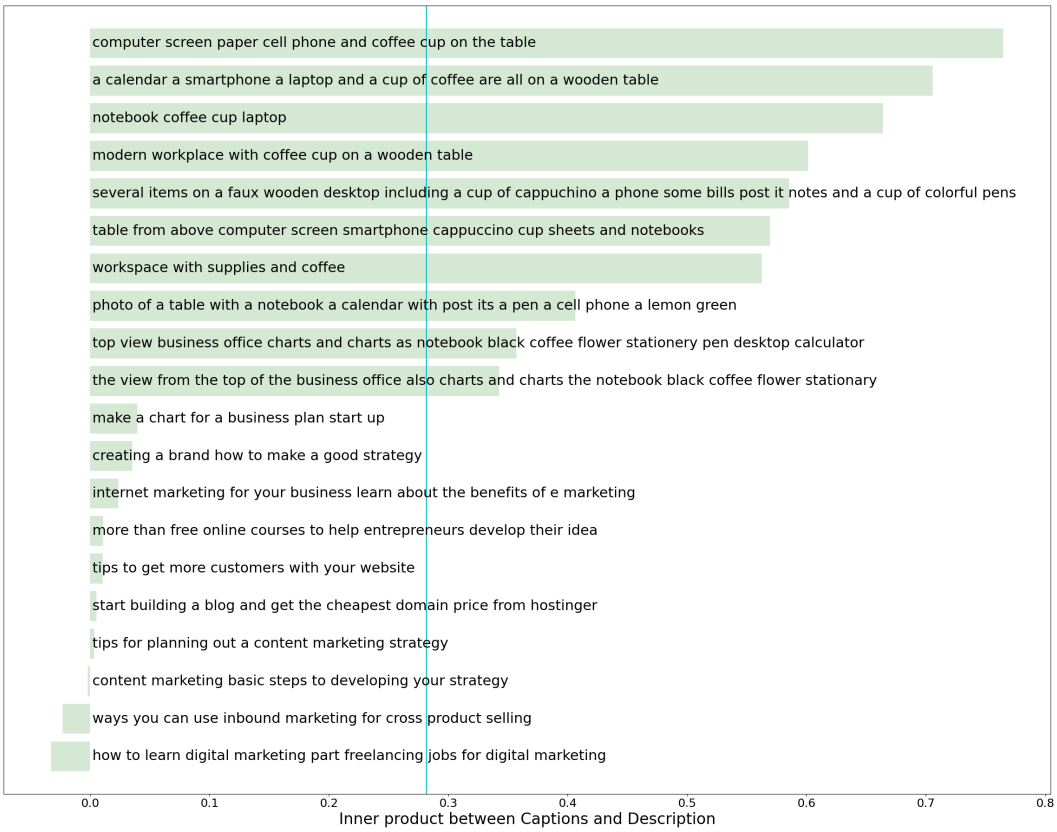



**Figure 4.15:** The results of the CLIP similarity analysis for image img_000384.jpg are displayed. Captions are represented on the y-axis. The graph displays the inner product value between the caption and the generated description of the image on the x-axis. The horizontal line illustrates the average similarity values.

**Results** To make the analysis of similarities with SBERT embeddings comparable to the analysis with CLIP embeddings, the results are evaluated using the same images (Figure 4.11) and also the manually selected captions. In the SBERT similarity analysis graphs, the value plotted on the X-axis is also the similarity between embeddings. But this time the similarity between the embeddings of the captions and the embedding of the generated description is considered. Figure 4.15 shows the results of the SBERT similarity analysis for Figure 4.11a. Even at first glance, it is noticeable that there is a much larger variance in similarity values. A closer look shows that exactly 10 of the analyzed captions have a similarity value above the mean value of 0.26 and 10

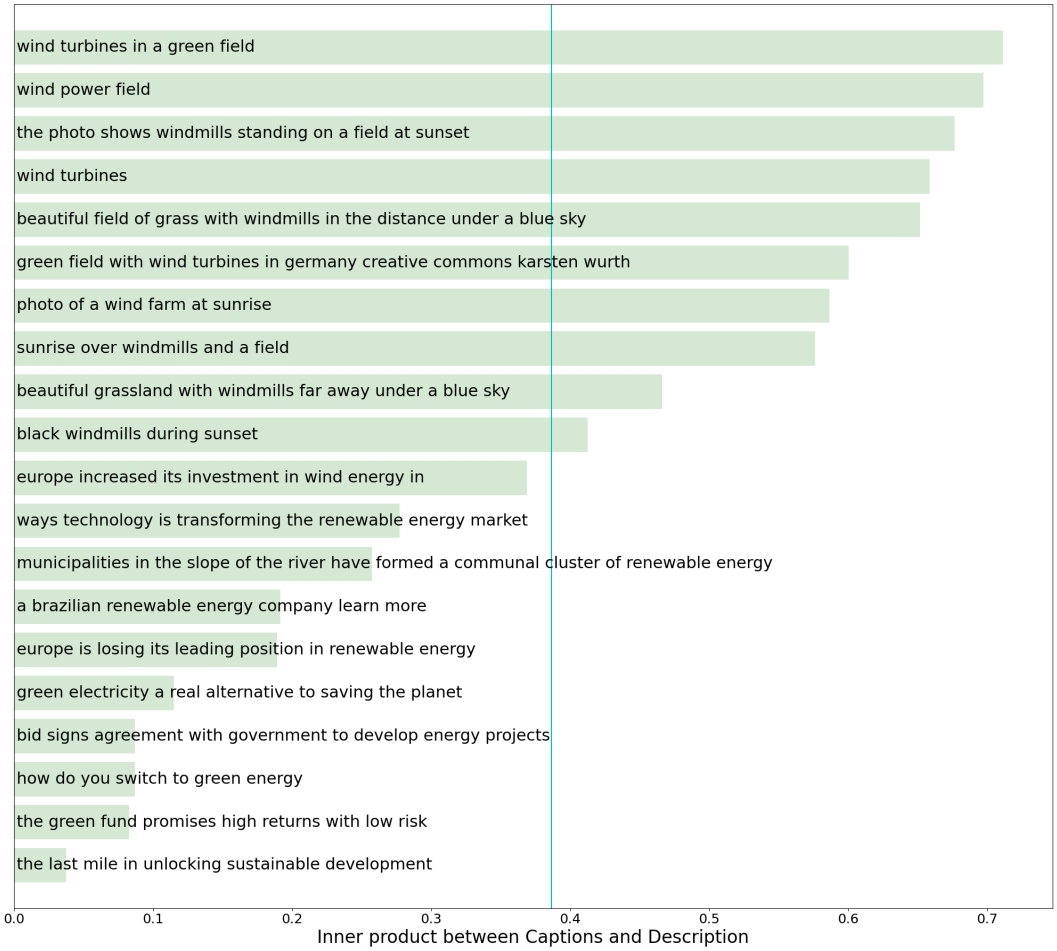



**Figure 4.16:** The results of the CLIP similarity analysis for image img_019768.jpg are displayed. Captions are represented on the y-axis. The graph displays the inner product value between the caption and the generated description of the image on the x-axis. The horizontal line illustrates the average similarity values.

have a similarity value below the mean value. Now looking at both groups, it can be seen that the captions from the group with high similarity values are all descriptive captions. The captions with the low similarity values are therefore those that are more likely to hold the meaning of the image. In this case, for example, this includes the caption *'make a chart for a business plan start up'* or *'tips to get more customers with your website'*. Both are captions that are more concerned with the meaning of the image, what we associate with it, or what the image 4.11a might be a symbolic representation for. None of these meaning-bearing captions are depicted the image itself, and yet the images were used in this business context and with these captions. The results

for the similarity analysis with SBERT for image 4.11b can be seen in Figure 4.16. Again, at first glance, it is noticeable that there is a larger difference in similarity values and that 10 of the similarity values are above the mean of 0.25 and 10 are below. Again, the captions that contain what is seen in the image have a higher similarity to the description than those that contain a level of meaning of the image. The captions, which have a low similarity value, refer to the meaning level of the image. Examples for those captions include *'ways technology is transforming the renewable energy market'* or *'how do you switch to green energy'*. Sustainability and renewable energy is a topic the image is often associated with, even if these things are not shown in the image. This is probably because wind turbines are a symbol of renewable energy, and so this seems to be an underlying meaning of the image.

In contrast to the first analysis of embedding similarities, a higher variance could be found here in the similarity values. For both images, the standard deviations in this experiment over the entire set of captions are significantly higher compared to the values obtained in the previous experiment, with a value of 0.21 for Figure 4.11a and 0.2 for Figure 4.11b. A closer analysis of the results of some examples revealed that below a certain similarity value (which is often the mean but varies per image), the texts are no longer descriptive, but convey a meaning of the image. If the similarity between a caption and the generated description of an image is low, it is more probable that the examined text conveys one of the meanings of the image.

**Conclusion of the Similarity Analysis** In conclusion, it is not feasible to differentiate between meaning and description using CLIP. The use of SBERT embeddings for the similarity calculation, on the other hand, makes it possible to distinguish between meaning and description. To determine whether a caption emphasizes description or meaning, utilizing similarity analysis is a noteworthy methodology. However, relying solely on CLIP similarities does not definitively determine whether the text is conveying meaning or providing a description. Therefore, it is necessary to additionally consider the similarity between the text and the description of the image. Therefore, it can be concluded that a text conveying a meaning should be as similar as possible to the reference image while differing as much as possible from the image description.

## 4.3 Extract Hidden Meaning from Images

After the captions have been examined in various analyzes, the results of the similarity analysis were used in this last experiment to possibly find one of the meanings of an image. The goal of this work was, among others, to find

a way to extract a meaning from an image using technical tools. The results of the similarity analyzes from Section 4.2.3 can be used as a clue for this. As described there, a high embedding similarity between a text and a reference image and a low embedding similarity between this text and the description of the reference image could indicate that the text conveys one meaning of an image. These findings were used in the following chapter to create a *hidden meaning embedding* (HME) for an image. The efforts for this are described in the first part of this section. However, since an embedding would not be interpretable or usable, the second part of the section is about the possible *translation* of the HME.

### 4.3.1 Creation of a Hidden Meaning Embedding

The objective was to create a neural network capable of receiving an image and its description as input and producing an embedding as output. This embedding should ideally convey one meaning of the input image, and is therefore referred to as a hidden meaning embedding. To construct the model, we utilized the results of the similarity analysis. This analysis showed that the similarity between an embedding of a meaning-bearing caption and the embedding of the image itself is close to the similarity between the embeddings of a descriptive caption and the image. However, to differentiate between description and meaning, the analysis has demonstrated that the similarity score between embeddings of an image's description and its meaning must be significantly low to classify a caption as conveying meaning. These findings have been incorporated into the creation of the network. Thus, a meaning embedding should have a high similarity to the image embedding on the one hand, but at the same time have a low similarity to the description embedding. Thus, for the creation of the hidden meaning embedding, both the image itself and the literal description of the image are necessary.

The HME to be created must be adjusted based on both inputs. On the one hand, the similarity of the HME and the input image embedding must be maximized and on the other hand, the similarity between the HME and the description embedding must be minimized. To meet these requirements, we chose Siamese network (SN) architecture. In a SN, two different loss values are adjusted simultaneously with the same weights to compare two input vectors with one output vector. To compute the input embeddings CLIP was used here , because as already described CLIP offers the possibility to map images and texts by embeddings in the same vector space and thus to make them comparable. So, the resulting HME must also be a CLIP embedding. The structure of this experiment can be seen in Figure 4.17. The images from the stock photo dataset described in 4.1 were used as training data for the

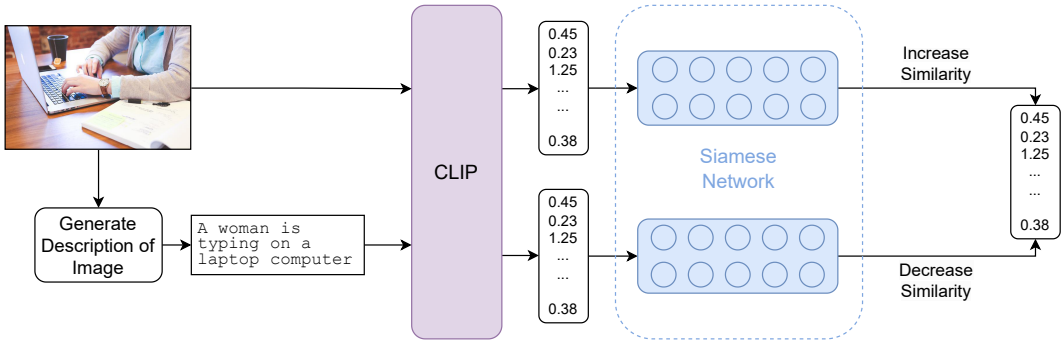



**Figure 4.17:** Conceptual overview of SN training for HME generation. CLIP embeddings are computed for an image and its generated description. Both embeddings are used as input for the SN. The SN produces an output CLIP embedding with a high similarity to the image embedding and a low similarity to the description embedding.

model, and in addition to each image, the generated description was used. In the context of this work, annotating the stock image dataset would have been very resource-intensive and time-consuming. Moreover, the meaning attribution by human annotators would be subjective. Therefore, an unsupervised approach was chosen to test the possibility of accomplishing the task of extracting hidden meaning solely through using CLIP embeddings of an image and its description.

In the first step, CLIP embeddings were calculated for both the images and the generated descriptions. The pre-trained ViT-L/14 model from the openai clip library is used to generate the CLIP embeddings from the images and descriptions in the dataset. It converts the image and the image description text into a 768-dimensional embedding. The input in the form of two CLIP embeddings is already a highly complex representation. Thus, we decided to build a basic neural network consisting of only one linear layers and batch normalization. We have made this choice because, based on the characteristics of CLIP embeddings, there is no need for significant transformations to alter them. However, we need to keep the basic structure and features of the original CLIP embeddings to some extent, as it is important that the HME remains an interpretable CLIP embedding. The SN has been implemented in PyTorch, using a shared embedding layer for processing two input vectors. The layer consists of a linear layer, ReLU activation, and batch normalization sequence, which is repeated twice. Stabilizing and enhancing the learning process is achieved by using batch normalization and ReLU after each linear transformation. The output layer of the network calculates the dissimilarity between the two preprocessed embeddings. While this network can handle various input

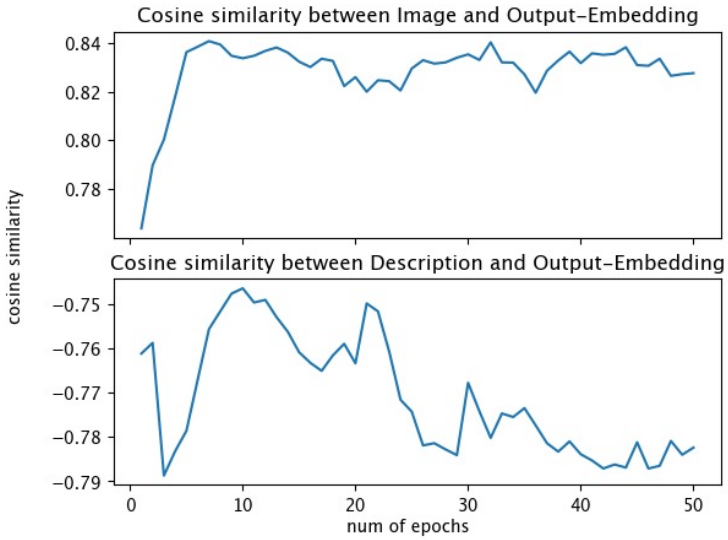



**Figure 4.18:** The loss curve of the two similarity calculations during the training of the SN over 100 epochs can be seen.

dimensions, it was specifically tested with 768-dimensional CLIP embeddings. The model was trained on the stock image dataset for 50 epochs using a batch size of 32. For testing purposes, one image was excluded from this dataset, resulting in the final training being conducted on 66,226 stock image and description embeddings. The Adam optimizer was utilized with a learning rate of 0.001. In the course of the training, an image embedding and the corresponding description embedding are passed to the SN as input. Thereupon the weights are adjusted in the individual epochs. The cosine similarity between one of the input embeddings and the output embedding is used as the loss in an alternating manner. In each 'even' epoch, the similarity between the image embedding and the output embedding is calculated. In these epochs, the goal is to maximize the cosine similarity, so here the negative cosine similarity is the loss and the weights are adjusted accordingly. In the 'odd' epochs, the similarity between the description embedding and the output embedding is calculated. Here the similarity should be minimized so that the loss is the cosine similarity and the weights are adjusted to minimize the similarity. As can be seen in Figure 4.18, there was no continuous loss curve. In trainings with more epochs, the loss developed partly as it should, and the similarity between output and image increased and the similarity between output and description decreased. However, the embeddings generated with SNs trained over more epochs could not be translated into correct images using unCLIP, so that the meaningful development of the loss curves did not prove to be a criterion for the quality of the final HME. The values of the loss show that the

similarity between the output and the image is much higher than the similarity between the output and the description. This shows that the general goal of generating output that is similar to the image and dissimilar to the description has already been achieved.

### 4.3.2 Translation of Hidden Meaning Embedding

In order to interpret the CLIP embedding generated by the SN described in Section 4.3.1, it must be 'translated' in some way. Therefore we tried to translate the embedding with two different approaches. In the first approach we used the CLIPxGPT Captioner[16] to translate the embedding with the help of GPT into a natural language text to find out if the generated embedding contains a meaning of the corresponding image. The second approach works with unCLIP[17] and tries to generate an image from the embedding instead of natural language to check if a meaning level can be found in the image.

**Figure 4.19:** The image img_00001.jpg from the stock image dataset is shown. The corresponding description generated by BLIP is: *a close up of a jar filled with coins and a plant growing out of it.* This image was not part of the training dataset.

**Translation with CLIPxGPT** The first idea to make the generated HME analyzable was to translate the embedding back into natural language. Since the choice of possible techniques and tools to translate CLIP embeddings back into natural language is limited, the CLIPxGPT captioner was chosen for this approach. This is based on CLIP and GPT-2 and is actually intended to create captions. The CLIPxGPT captioning model was trained on the Flickr30k[18] dataset. Beside a CLIP model, which was used without any fine-tuning, a mapping module and some layers of GPT-2 were trained for the creation of CLIPxGPT. To generate captions for an image, a CLIP embedding

[16]https://github.com/jmisilo/clip-gpt-captioning
[17]https://huggingface.co/docs/diffusers/api/pipelines/unclip
[18]https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

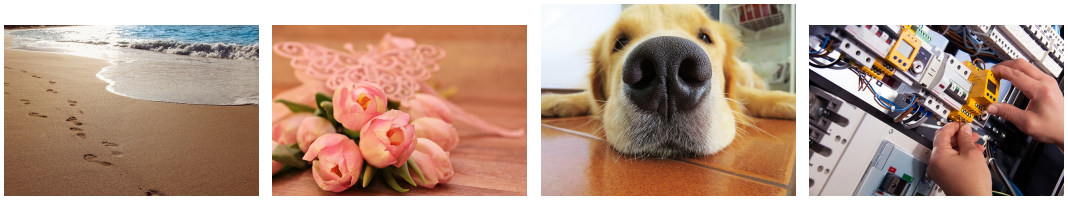
is initially calculated. This CLIP embedding is then mapped to a GPT-2 embedding by the mapping module and this GPT-2 embedding is transformed into text by the text encoder. Thus, a caption is created from the image by computing and mapping CLIP and GPT-2 embeddings which is the originated use case. To adapt this model for our specific use case, we removed the image encoder. Instead of converting an image to an embedding, we now input an embedding directly to the model. Thus, the HME generated by the SN should be 'translated' into a natural language text using the CLIPxGPT captioner.

| Epoch | Generated Text |
|---|---|
| 1 | Commuters are walking down the street. |
| 10 | Commuters are watching a man in a blue shirt and a woman in a white shirt. |
| 20 | Commuters are watching a band play. |
| 30 | Commuters are watching a hockey game. |
| 40 | Commuters are watching a group of people in a field. |

**Table 4.2:** Texts generated by CLIPxGPT for output embeddings of the SN during training are shown. Every 10 epochs the HME was generated for the image img_000001.jpg in Figure 4.19. The resulting HMEs were converted into texts using CLIPxGPT to verify the process.

The assumption was that in this way one of the different meanings that an image might have could be contained in the text. Unfortunatly, this approach does not work. The results of CLIPxGPT during the training of the SN can be seen in table 4.2. Every 10 epochs in the training of the SN, an HME was generated for the image img_000001.jpg, as shown in Figure 4.19. This image was not part of the training dataset and therefore serves as a test image. Texts were generated from these embeddings using CLIPxGPT. The texts do not say anything about the hidden meaning of the original image. However, the fact that those translations do not really carry the meaning level of an image does not necessarily have to be due to the fact that the SN does not work. Upon closer examination of the CLIPxGPT captioner, it becomes apparent even with the originally intended use that partially no satisfactory results can be created. Thus, even when inputing an image, the model partially 'made up' objects and people that were not visible in the image. Figure 4.20 shows some examples where images were passed to the CLIPxGPT captioner and still no reasonable captions were generated. On closer examination of the results in Figure 4.20, it is noticeable that in some way themes or objects that can be seen in the images also appear in the captions, such as *beach* in Figure 4.20a or *dog* in Figure 4.20c. But the captions only describe the images in some extent

and in many cases hallucinate things in addition. So the CLIPxGPT model and approach of using GPT-2 to translate CLIP embeddings into texts does not really seem to be possible so far. That means that, unfortunately, this approach can not give conclusive information about whether the HME gets a level of meaning of the image or not.



**(a)** A person in a black suit is walking on the beach.

**(b)** Woman with a white hat and a white veil is standing in front of a table with a tablecloth and a tablecloth with a flower on it.

**(c)** Dog with glasses on.

**(d)** A man in a blue shirt is working on a computer.

**Figure 4.20:** Images and their corresponding generated caption using CLIPxGPT.

**Translation with unCLIP** Another possible approach to enable the interpretation and evaluation of the SN possible is 'translating' the embedding using unCLIP[19]. UnCLIP is a model used by DALL·E for generating images from CLIP embeddings. A CLIP embedding can be passed to the model, and an image is generated from this CLIP embedding. We considered this approach due to the poor results of translating with the help of CLIPxGPT. The idea here was to convert the embedding generated by the SN into an image that may contain a hint to the meaning of the image originally processed by the SN. UnCLIP was used without any further adjustments or fine-tuning.

Figure 4.19 shows the image and the corresponding caption that were used during the training of the SN to generate an output embedding every 10 epochs. The output embeddings were converted into images with the help of unCLIP. These images can be seen in Figure 4.21. Almost all images contain coins, although these can also be seen in the original image. However, these images can hardly be used for our purpose. Usable images are difficult to achieve, as the hidden meaning of images are often concepts that cannot be visualized directly. So this approach, although it worked more or less, it only results in images generated from the HME having to be manually interpreted again to

[19]https://huggingface.co/docs/diffusers/api/pipelines/unclip


**Figure 4.21:** Images generated by unCLIP for output embeddings of the SN during training are shown. Every 10 epochs the HME was generated for the image img_000001.jpg in Figure 4.19. The resulting HMEs were converted into images with unCLIP to verify the process.

obtain the meaning of the HME and thus the original image. This approach is therefore also not expedient.

### 4.3.3 Conclusion of the Hidden Meaning Extraction

The purpose of developing a HME was to identify the meaning of an image without relying on numerous captions for the same image. Using the SN, it is possible to generate an embedding in the shape of a CLIP embedding that has both a high similarity to the input image embedding and a low similarity to the input description embedding. The assumption that a CLIP embedding could potentially integrate one of an image's meanings emerged from the findings of the similarity analysis outlined in Section 4.2.3. However, the initial findings of the similarity analysis suggest that relying solely on CLIP similarity might not be enough to detect meaning, at least not in CLIP's present state. Since the SN was intended to create a CLIP embedding, only CLIP was used to create both input embeddings, although the second similarity analysis was based on SBERT. The results of the similarity analysis were utilized in creating the SN. It is uncertain if this circumstance contributed to the ineffective outcome of the SN. Nonetheless, no definitive conclusions can be drawn.

To improve and adapt the SN, it would have been necessary to translate the embedding to be able to evaluate it. A translation would have made it possible to interpret the result of the SN. The translation using CLIPxGPT was able to create a natural language sentence from the HME, but this sentence did not contain any meaning of the input image in all tested cases. However, as described above, the CLIPxGPT captioner does not produce satisfactory results even in the original usecase, so it cannot be conclusively determined for this approach whether the translation with CLIPxGPT or the generated HME itself is the problem. Thus, this approach cannot be used to improve the SN. Due to the generally not clearly recognizable meaning of images, the unCLIP approach could not really be used any further. Images that are created in

this way would have to be examined again by humans in order to interpret a meaning into them. As this process of finding meaning in images is subjective, it cannot be assumed that the meaning found in the image generated for the HME also reflects the meaning of the original image. However, this approach was able to show that the generated HMEs can indeed be interpreted as CLIP embeddings and that images can be generated from them. This means that a properly working approach to textual translation of CLIP embeddings could provide better information in the future and thus be used to improve the approach. Although the approach did not really work in the present way, it cannot be said with certainty that it is impossible to generate one of the hidden meanings of an image in this way. However, the experiments we carried out enabled us to create a basic framework that can serve as a basis for future work.

# Chapter 5

# Discussion and Conclusion

In this chapter, the work and the findings of the experiments are summarized and discussed. In particular, possible sources of error and problems that have become apparent in the course of the work are examined.

The present work deals with detecting meaning in stock images with the help of technical means such as CLIP.

Due to the intentional ambiguity of stock photos, they were considered an interesting foundation for the topic of this work. The nature of stock images raises the question to what extent there is a difference between the visible content of an image and the recipient's associated meaning, as well as the context in which the image is used. In addition, the question arose whether it is possible to find this context, i.e. this meaning of an image, using computational methods.

Stock photos served as the data basis for this work, as they are an interesting field in the context of detecting meaning. Stock photos are intended to have several meanings and one and the same image can be used several times in different contexts on the internet. It is therefore particularly interesting to investigate whether the same image is always used in a similar context, i.e., whether it has a hidden meaning. As no stock photo dataset existed before, it was necessary to create one in the course of this thesis. The free stock photo site Pixabay was crawled for this purpose. The final dataset consists of 66,277 stock photos on approximately 133 topics from a list created with the help of ChatGPT. In order to have a basis for analysis from which subject areas and meanings for images could possibly be extracted, it was necessary to collect several texts or captions for each image, which were used together with the respective image. CLIP retrieval was conducted on the LAION-5B dataset for this purpose. Duplicates of the respective image were searched for, and the captions of the duplicates were saved. This made it possible to find 200 captions per image for a part of the stock photo dataset, on which

the further analyzes were carried out. The number of captions was limited to 200, as exemplary analyzes showed that the images from the dataset often did not occur much more than 200 times in the LAION-5B dataset. Searching for more duplicates would have considerably increased the time required for caption crawling. Despite exemplary checking and adjustment of the duplicate score so that only duplicates were actually found, it cannot be fully guaranteed that every image found and identified as a duplicate is actually a duplicate and not simply a very similar-looking image. Therefore, in order to be able to carry out and interpret the analyzes, it was assumed that all images found that exceed the duplicate score are indeed duplicates. One reason for the relatively rare occurrence of the respective images in the dataset could be the use of Pixabay as a stock photo base, as Pixabay is a relatively small and less frequently used stock photo site. The largest and most frequently used stock photo site is Getty Images. The images from this agency are also used by major newspapers such as the New York Times. Using photos from this site as a basis might have ensured that the images appear more frequently in the LAION-5B dataset and that the collected captions are of better quality. Many of the captions collected via the CLIP retrieval were of poor quality, contained the file names or the page names of the websites and did not necessarily have anything to do with the respective image. This was also noted by Betker et al. in *Improving Image Generation with Better Captions*, where they show that Text2Image models, e.g., can work much better with better captions and they note that captions crawled from the internet are often of very poor quality. This was also a problem in the present study: Even if many duplicates could be found, the corresponding captions were often of poor quality. In addition, many of the captions were not in English, so that a translation of the non-English captions was necessary in order to be able to carry out the analyzes on a large proportion of the captions at all. As the interpretation of images also depends on the social and cultural background of the viewer, people from different cultures may understand or interpret images differently. This means that the translation of captions may mean that the results are not entirely valid. In addition, the assumption must be made that the translations using the NLLB-200 model are correct. This could also be a possible source of error and limit the validity of the results.

For some of the analyzes it was not possible to use the image itself, as they are based solely on text. Therefore, a caption was generated for each image using the BLIP model from Salesforce. This generated caption should represent the image. This is of course a potential source of error, as there is no guarantee that the model actually generates a correct description of the image. Although a small part of the stock photo dataset was used to check whether the model generates descriptive captions, it can still be a source of error. In order

to carry out the analyzes anyway, the assumption is made that the generated description can be representative of the image itself. The first analysis of the captions is the BoW analysis. This extracted the top 10 most frequent words that appeared in the crawled captions but not in the generated description. The assumption for this analysis is that words that appear in the description have nothing to do with the hidden meaning of the image, as they describe objects or activities that can be seen in the image itself. Words that remain and are frequently used together with the picture, on the other hand, give an indication of the hidden meaning, or are possibly already the hidden meaning of the picture. For some of the pictures analyzed, this simple and naive approach provided good starting points for finding the meaning of a picture. Nouns in particular give a good indication of the contexts in which an image is used and the frequent occurrence of these can already be seen as a kind of meaning of the image. To a certain extent, this analysis could support the assumption that images often evoke certain connotations and connections in the recipients despite their 'open meaning'. This can partly be shown by the very frequently occurring words. In the subsequent analysis, topic modeling was applied to the crawled captions. Although topic modeling is not necessarily designed for the short and, above all, few texts, it was also possible to find some indications of the meanings that an image carries. For some of the images analyzed, it was possible to divide the topics found into descriptive and meaningful. The names of the themes are derived from frequent words that appear in the captions, so that in some cases the names and the spacing of the themes in the diagrams made it possible to see which themes dealt with what could be seen in the image and which ones tended to illuminate a level of meaning in the image. The topics found with BERTopic, like the BoW analysis, were thus able to provide indications of topics and contexts that were frequently used together with an image.

With the help of the first similarity analysis, the second research question, whether it is possible to find a difference between descriptive and meaningful captions using CLIP, was to be answered. The goal was to find out whether the value of the CLIP embedding similarity between an image and a corresponding caption can give any information about the relationship between the two modalities. The intention was that, for example, a high similarity of CLIP embeddings could indicate that the text encompasses a level of meaning of the image. The analysis of different images showed that there was no significant difference between the similarities of an image and its captions. The descriptive captions thus had a comparable high embedding similarity to the image itself as the meaningful captions. Thus, CLIP was not as advantageous as was hoped at the beginning of the work. Although the hope was that CLIP would work better for the use case of the work, it is not necessarily surprising

that this was not the case. CLIP was trained with unverified data crawled from the internet. It can therefore have been trained with both descriptive and meaningful captions or texts. Captions from the internet, which are often also based on the alt texts, are often of poor quality or are sometimes simply relatively independent of the image itself (Betker et al., 2023). For this reason, it is not necessarily possible to distinguish between descriptive and meaningful with the help of CLIP alone. The second similarity analysis was based on the results of the first. As it was not possible to distinguish between meaning and description using CLIP and the similarities between images and texts, only the texts were considered in this experiment. The generated description was used as a representation of the image and the similarity between the description and the crawled captions was calculated. It was found that there is indeed a difference in similarities with the description as a reference. The average is often the threshold. Captions that have an above-average similarity to the description are also descriptive. The captions that have a low similarity to the generated description are instead more meaningful.

A further aim of the work was to train a neural network with which it is possible to extract one of the levels of meaning from an image. The theoretical foundations of meaning in images suggest that finding meaning in images is a very complex analytical task. However, the last part of the thesis tested the feasibility of finding hidden meaning in CLIP embeddings. The results of the similarity analyzes served as the basis for the creation of the neural network. The neural network received both the CLIP embedding of an image and the CLIP embedding of the corresponding description as input and, in the best case, create a CLIP embedding that contains the hidden meaning of the image. To achieve this, the output embedding is supposed to be as similar as possible to the image embedding and at the same time as dissimilar as possible to the literal description. The model is a Siamese network that transforms and combines the two input embeddings and finally creates an output embedding with the help of the alternately adapted loss. In order to evaluate the model in a meaningful way, two approaches were tested to translate the generated embedding.

The first approach was to translate the embedding into a natural language text using a model called CLIPxGPT. A text is easier to interpret and to check whether one of the meanings of the image occurs in it. Unfortunately, translating CLIP embeddings back into natural language text is not possible yet. One of the few approaches that deals with this translation is the CLIPxGPT approach. This attempts to create a text from an embedding by mapping CLIP embeddings to GPT embeddings. However, CLIPxGPT does not deliver satisfactory results in the actual use case of generating captions for images. With the help of CLIPxGPT it was possible to create a text from the output em-

bedding of the SN. However, due to the inadequate results in the actual use case of CLIPxGPT, no meaningful interpretation could take place and thus the SN could not be improved by the translation with CLIPxGPT. With the help of unCLIP it was possible to create images from the generated HMEs. But, regardless of how well this translation approach ultimately worked, translating the HME into an image would only have been helpful in a few cases. This is because many of the hidden meanings that images can have cannot really be described with the help of objects and actions. Meanings such as *business*, *innovation*, or *seasons* are merely concepts that cannot be literally translated into an image. In order to extract these concepts from a hidden meaning embedding with the help of unCLIP, a concrete analysis of the result image would again be necessary. This would bring us back to the initial question and we would have to understand the meaning of an image again. It would have been a circular argument, so to speak. However, since the translation into natural language hardly promised any success, this approach should provide a way of interpreting the results of the Siamese network.

As neither of the two translation approaches really worked well, it was not possible to test and improve the SN in the course of this thesis. The approach could therefore neither be supported nor completely refuted with the available models and tools at this point in time.

In conclusion, it can be said that with the help of the analyzes, it was possible to find indications of hidden meaning for the images examined. The challenge in finding the hidden meaning of an image lies primarily in the interpretation of the results. Meanings in images are often subjective and ambiguous and are heavily dependent on human understanding of context, making it a very complex task to find the hidden meaning using solely computational methods. Nevertheless, some of the experiments, such as topic modeling and similarity analysis with SBERT, were able to find a difference between descriptive and meaningful captions. Unfortunately, CLIP was not able to bring the expected benefits. The creation of a HME also turned out to be unsuccessful for the reasons already explained.

Despite the challenges in interpreting image meanings computationally, this thesis makes significant contributions to the fields of image analysis and artificial intelligence. The creation of a stock image dataset and the application of methods like topic modeling and similarity analysis with CLIP and SBERT represent innovative steps in the exploration of image context and meaning. The experiments conducted offer insights into the subjective and multifaceted nature of image interpretation. Thus, this thesis also contributes to discussions about how meaning is constructed and perceived by people from different backgrounds, cultures, and socializations. The methodologies and findings of this thesis may influence future research in these research fields and may

enhance our understanding of the complex relationship between visual content and perceived context in images and other media.

## 5.1 Future Work

Despite some problems, this thesis provides a groundwork for further research into detecting meaning in images with the help of computational tools. Further, more in-depth analyzes could be carried out following this work. Therefore, the stock photo dataset with the corresponding captions can be used to further explore the topic. A manual annotation of the existing captions into descriptive and meaningful would be beneficial for future work, as it would allow for more in-depth experiments. Perhaps a better foundation for analyzing the level of meaning would be using longer texts and not just the alt attributes of the images, as available in the LAION-5B dataset. As has been shown, captions are insufficient and sometimes of poor quality, when crawled from the Web, as other researchers have noted, so an analysis of full article texts for an image, e.g., could provide more information about the meaning and context of the image itself. Due to the limitations of CLIP, many of the experiments carried out did not produce satisfactory results. A new version of CLIP trained on better captions could ensure that re-running experiments and re-training the SN produces more meaningful results.

Another interesting question is the clarification of cultural differences in relation to meaning in images and the decoding of codes. In research on meaning in images, it is often said that the decoding of codes and meanings in images can depend on the social and cultural background of the recipients. It would therefore also be interesting to investigate whether captions from different languages relate different meanings to an image. This would make it possible to investigate to what extent the cultural background of the users of the stock photos has an influence on the attribution of meaning or the context in which an image is used. The existing caption data set could be used for this analysis, as it already contains a large number of captions in various languages, as described above.

# Bibliography


Aiello, G. and Parry, K. (2020). *Visual communication: understanding images in media culture*. SAGE, Los Angeles London New Delhi Singapore Washington DC Melbourne.

Alikhani, M., Sharma, P., Li, S., Soricut, R., and Stone, M. (2020). Clue: Cross-modal Coherence Modeling for Caption Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535. arXiv:2005.00908 [cs].

Barthes, R. (1977). The Rhetoric of the image. In *Image-Music-Text*, pages 32–51. Fontana, London.

Bateman, J. A. (2014). *Text and image: a critical introduction to the visual-verbal divide*. Routledge, Taylor & Francis Group, London ; New York.

Beaumont, R. (2022). Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. `https://github.com/rom1504/clip-retrieval`.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., and Ramesh, A. (2023). Improving Image Generation with Better Captions.

Chakrabarty, T., Saakyan, A., Winn, O., Panagopoulou, A., Yang, Y., Apidianaki, M., and Muresan, S. (2023). I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors. arXiv:2305.14724 [cs].

Frosh, P. (2002). Rhetorics of the Overlooked: On the communicative modes of stock advertising images. *Journal of Consumer Culture*, 2(2):171–196. Publisher: SAGE Publications.

Frosh, P. (2020). Is Commercial Photography a Public Evil? Beyond the Critique of Stock Photography.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].

Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., and Divakaran, A. (2019). Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. arXiv:1904.09073 [cs].

Leeuwen, Theo van, G. K. (2020). *Reading Images: The Grammar of Visual Design*. Routledge, London, 3 edition.

Lemke, J. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs].

Liu, Z., Guo, M., Dai, Y., and Litman, D. (2022). Imagearg: A multimodal tweet dataset for image persuasiveness mining. *arXiv preprint arXiv:2209.06416*.

Machin, D. (2004). Building the World's Visual Language: The Increasing Global Importance of Image Banks in Corporate Media. *Visual Communication*, 3(3):316–336. Publisher: SAGE Publications.

Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Hakimov, S., and Ewerth, R. (2021). Multimodal news analytics using measures of cross-modal entity and context consistency. *International Journal of Multimedia Information Retrieval*, 10(2):111–125.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs].

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sturken, M. and Cartwright, L. (2009). *Practices of looking: an introduction to visual culture*. Oxford University Press, New York, 2nd ed edition. OCLC: 144224088.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59(2):64–73. arXiv:1503.01817 [cs].

Torky, I., Ruth, S., Sharma, S., Salama, M., Chaitanya, K., Gollub, T., Kiesel, J., and Stein, B. (2023). Webis @ ImageArg 2023: Embedding-based Stance and Persuasiveness Classification. In Alshomary, M., Chen, C.-C., Muresan, S., Park, J., and Romberg, J., editors, *10th Workshop on Argument Mining (ArgMining 2023) at EMNLP*. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*. arXiv: 1706.03762.

Ward, C. G. (2007). Stock Images, Filler Content and the Ambiguous Corporate Message. *M/C Journal*, 10(5). Number: 5.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.