Image Captioning of Art-Historical Photographs

Seminar
Tagging and Captioning Art-Historical Photographs
Winter Semester 2022/23

Smilla Fox, Elena Gensch, Valeria Tisch

Supervisor:
Alejandro Sierra Múnera, Jona Otholt, Hendrik Rätz

March 7, 2023

Contents

List of Figures

Acronyms

BLEU Bilingual Evaluation Understudy. 2, 10, 11, 16, 18

BLIP Bootstrapping Language-Image Pre-training. 3, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 19, 22, 23, 24, 25

CIDEr Consensus-based Image Description Evaluation. 2, 11, 16

MED Multi-modal Mixture of Encoder-Decoder. 8, 9

RCNN Region Based Convolutional Neural Networks. 7

SPICE Semantic Propositional Image Caption Evaluation. 2, 3, 11, 12, 16

ViT Vision Transformer. 7, 8

VLP Unified Vision Language Pre-Training. 7

WPI Wildenstein Plattner Institute. 2, 3, 6, 12, 14, 15, 16, 17, 18, 20, 21, 24, 25

To preserve historical artworks and make them easily accessible to more people they are often displayed in digital archives. Adding captions to the stored photographs and scans of artworks helps gain more information about them, especially for visually impaired people. The Wildenstein Plattner Institute possesses a large collection of art-historical photographs. Because manually creating captions for them is a time-consuming process, an automated solution is required. Unlike our approach, existing solutions trying to tackle the captioning task mainly focus on natural photographs. We specifically adapt to art images using the Bootstrapping Language-Image Pre-training framework. We finetune large pre-trained models with the art image dataset Artpedia and investigate methods to improve the training dataset and caption generation. This includes filtering the training dataset, using grayscale images, and enforcing certain words in the captions. In all approaches, the finetuned models show better results than the pre-trained one. The captions become more detailed and sometimes contain art-specific content like painting styles. However, they also have more mistakes.

1 Introduction

Archiving serves to gain insights into the artist's creative process, providing historical and cultural contexts, and ultimately preserving the works for future generations. In the field of art history, archiving has evolved in the past: recently, there has been a shift towards the digitization of archives.

The Wildenstein Plattner Institute (WPI) is an organization dedicated to the study of art-historical knowledge. The WPI's primary mission is to promote scholarly research, preservation, and accessibility of art-historical materials, including digital archives. The institute possesses a substantial collection of artworks and research materials. These include over 100,000 digital media files, among others photographs of artworks, studied over the years by the institute's scholarly affiliates. Furthermore, the WPI publishes the information online and thus makes it accessible to a wide audience.

Image captioning is a method that involves generating a natural language description of an image's visual contents. The process of captioning enhances the accessibility of the artwork for people with visual impairments. Moreover, captioning facilitates the use of digital archives by enabling keyword searches and faster identification of paintings depicting specific themes or subjects. Annotating the artwork data manually is time-consuming and costly. Existing solutions for automatically captioning image data focus on natural photographs, not artwork. Consequently, we aim to develop an automatic captioning system tailored explicitly to artwork images.

To create captions automatically, we work with the language-image pre-training framework BLIP [LXH22]. The goal is to generate detailed descriptions of the images' visible content. Our main contributions to achieve this include:

1. Finetuning the pre-trained captioning model on the Artpedia dataset [SCB+19] to adapt it to artwork images.
2. Filtering the dataset to remove low-quality captions and further finetuning the model on the remaining higher-quality captions.
3. Forcing the model to include certain words in the captions.
4. Evaluating the model's ability to generate captions with artist information, building on our finding that models finetuned on the Artpedia dataset also generate context information.

2 Related Work

With an increasing number of artworks becoming available digitally, there is a surge of interest in utilizing computer vision techniques, such as image classification and object

detection, to analyze and understand images from the art-historic field. Recently, there is a desire to combine visual and natural language representations of artworks.

Baraldi et al. [BCGC18] demonstrates the feasibility of cross-domain transfer between images and textual descriptions. They achieve this by using historical document illustrations along with textual commentaries and creating a shared embedding space for the images and texts. This mechanism, known as multi-modal retrieval, allows matching images with sentences. However, unlike our approach, this work focuses on matching an image to the correct part of a given text instead of generating text.

Similarly, the Text2Art approach by Garcia et al. [GV18] is a multi-modal retrieval task. It involves finding a relevant image or text from a closed set of images or texts given an artwork's text or image. In the process, they have published a collection of fine-art images along with textual comments, known as SemArt. The comments in the dataset also aim to describe techniques or art-historical context, making it unsuitable for our task to caption visual elements.

With the rise of transformers [VSP+17], transformer-based vision-language models are implemented, pre-trained on vast amounts of data: the Vision Transformer (ViT) [DBK+20] learns to attend and interlink information from both visual and linguistic modalities. Instead of text input ViT allows processing input images as sequences of patches. Similarly, the method of Unified Vision Language Pre-Training (VLP) is based on a transformer architecture, which generates captions for input images [ZPZ+20]. The models are pre-trained on datasets of natural photographs with ground-truth text descriptions.

Cetini [Cet21] applies this pre-trained model by finetuning it on artwork images. To extract important regional image features, the image input is preprocessed using the Faster RCNN method. Unlike previous research, this method aims to generate captions that describe only the visual elements of a painting, rather than incorporating textual descriptions that resemble keywords and not full sentences. However, the proposed pre-trained and finetuned models may be unsuitable for our task to generate natural language descriptions, as the used textual descriptions tend to resemble an accumulation of keywords. Additionally, the VLP model architecture is not robust against an imbalanced training dataset: the finetuned model identifies over-represented subjects of the training dataset incorrectly during validation. It is important to work with a model that can predict correct content with a high degree of certainty.

3 Methods

To solve the task of captioning art-historical photographs, we utilize a state-of-the-art framework called BLIP, a bootstrapping language-image pre-training for unified vision-language understanding and generation [LXH22]. First, an overview of BLIP's model

architecture, the Multi-modal Mixture of Encoder-Decoder (MED), and its pre-training objectives is provided. Next, the Captioning and Filtering (CapFilt) method, which is used for dataset bootstrapping, is described. To assess the generated captions, several evaluation metrics are applied. These are explained in subsequent sections.

3.1 Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP)

BLIP is a recently developed framework, that has shown promising results in image captioning and visual question-answering tasks. BLIP's language-image pre-training uses image-text pairs to teach a model to "understand" visual and textual information and generate text. This pre-training phase allows the model to learn robust representations, which is of benefit in the second phase, where the model is finetuned on a downstream task. Finetuning is the process of further training the previous pre-trained model on a smaller set of data relevant to the task.

BLIP addresses pre-training from two different perspectives, the model and the data perspective. From the model view, the framework adapts flexibly to various vision-language tasks such as image-text retrieval, image captioning, and visual question answering. Multi-modal Mixture of Encoder-Decoder (MED) is proposed to pre-train a unified vision-language model with both understanding and generation capabilities. Contrary to uni-modal learning, which means that the model learns only from text or only from images, multi-modal learning refers to learning representations from various types of data while using the same model. For example, a speech recognition model might transcribe language by incorporating audio recordings and visual data like lip movements. The architecture of the multi-task model consists of four parts. Figure 1 illustrates how the different components work together. It enables the model to learn representations from both modalities simultaneously and combine them meaningfully.

The uni-modal image encoder encodes images with a Vision Transformer (ViT) [DBK+20]. The input image is divided into patches which are encoded as a sequence of embeddings.

Separately from the image encoder, the uni-modal text encoder uses a BERT-based encoder [DCLT18] to convert text to an embedding vector.

Both the uni-modal image and text encoder are optimized by the Image-Text Contrastive Loss (ITC). It aims to align text and pictures in the semantic space. After the encoding, the positive image-text pairs representations are supposed to be close to each other, as opposed to the negative pairs. An image-text pair is positive when the caption describes the image accurately. A negative one means the caption does not fit the image.

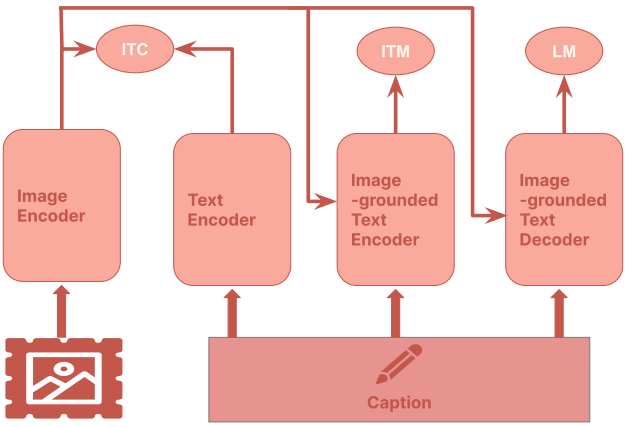


Figure 1: BLIP MED Architecture
Based on [LXH22]

The image-grounded text encoder operates similarly to the uni-modal text encoder but in a multi-modal way. The text and the image are taken jointly as inputs. In this manner, the text is encoded into an embedding vector while the visual transformer information is incorporated via cross-attention. It works by computing attention weights between each word in the caption and the corresponding visual elements in the images. Cross-attention enables the model to focus on more relevant parts of the two input modalities [GRM21].

The image-text matching loss (ITM) optimizes the image-grounded text encoder. Guided by ITM, the encoder aims to learn the multi-modal representation of image-text pairs that capture a more granular alignment between vision and language. A binary classifier is built on top of the text encoder. Based on the multi-modal feature, the linear layer predicts whether an image-text pair is positive or negative. It is more likely that negative pairings in a batch with higher contrastive similarity will be chosen to compute the loss. By applying that hard negative mining technique the model is forced to differentiate between minute but relevant differences in the instances and becomes more robust in the end.

The image-grounded text decoder also receives image and text information. Instead of encoding it, a text is produced while the visual information is incorporated via cross-attention. This part of the model is the most relevant for creating a caption. The image-grounded text decoder is optimized by the Language Modeling Loss (LM). That loss applies an auto-regressive model which employs probability calculations and predicts words based on a previous word when generating text. Cross-entropy measures how well the predicted caption matches the ground truth to help minimize the loss. Label smoothing is utilized to make the model less certain about its predictions and increase generalization. Therefore, LM endows the model with the ability to generalize visual data into meaningful captions.

BLIP not only views the problem from the model perspective but also from the data perspective. Commonly, image-text pairs collected from the internet are of low quality and contain numerous noisy alt-texts (alternative texts). These are textual descriptions of images, usually to support vision-impaired people. To address the issue of low-quality data, BLIP introduces two modules: a captioner and a filter. Both are initialized from the pre-trained MED model and finetuned individually on the high-quality human-annotated COCO dataset [LMB+14]. The captioner is an image-grounded text decoder that produces synthetic captions as additional training samples. The filter is an image-grounded text encoder that assesses how well an image-text pair matches. The filter removes bad captions from the synthetic dataset created by the captioner and from the web captions. A description is considered noisy when the filter predicts it does not match the image. A new high-quality dataset is formed by combining the filtered synthetic captions with the human-annotated ones. It is then used to pre-train a new model with the MED architecture.

3.2 Constrained Caption Generation

Captions can be forced to include certain predefined words: BLIP's image-grounded text decoder selects the output tokens that build the final caption based on the calculated token's probabilities at different time steps. Decoding strategies offer the opportunity to inject the tokens, representing the force words, at each time step during the generation. Once a predetermined token is included, the token will not be injected a second time, because the path already satisfies the constraint of including the token. This mechanism is known as Constrained Generation.

3.3 Evaluation Metrics

Natural language sentences can be expressed in a variety of ways. Given that diversity, determining whether a caption is accurate can be demanding. This challenge is not unique to our work but is shared by other related work. Even with appropriate labels, image captioning models make automatic evaluation a difficult task since captions are not binary classes. It is a common practice in the research community to adopt different metrics to evaluate the model's performance. We utilize three well-known metrics and one that we propose ourselves.

3.3.1 Bilingual Evaluation Understudy (BLEU)

Bilingual Evaluation Understudy (BLEU) citebleu is a widely applied metric for evaluating the quality of machine-generated natural language sentences, such as captions. It calculates the n-gram overlap between the generated captions and the reference captions. An n-gram is a sequence of n-words. N can have different values. For example, BLEU4

uses 4-grams. The BLEU score ranges from 0 to 1. A score of 1 indicates a perfect overlap between the generated and reference caption. As opposed, a score of 0 means there are no common words. It should be noted that the BLEU metric does not take word order or importance into account.

3.3.2 Consensus-based Image Description Evaluation (CIDEr)

Consensus-based Image Description Evaluation (CIDEr) [VZP14], similar to BLEU, is a metric that assesses the degree of n-gram overlap between generated and reference captions. However, CIDEr weights the importance of a word based on the term frequency-inverse document frequency (tf-idf). Tf-idf is a statistical measurement technique that assumes that important words are rare, whereas irrelevant ones occur very often across the corpus. That leads to a more sophisticated evaluation. CIDEr calculates the cosine similarity between the generated caption and the reference captions. The resulting score ranges from 0 to 10 because the cosine similarity is multiplied by 10. With a rising score, the similarity between the predicted and ground truth sequence increases. Nevertheless, the score rarely exceeds 1, which is already considered a very high score indicating the high quality of the generated captions.

3.3.3 Semantic Propositional Image Caption Evaluation (SPICE)

Semantic Propositional Image Caption Evaluation (SPICE) [AFJG16] addresses the limitations of the previously described metrics. N-gram overlap might be insufficient to detect the same meaning. The sentences could have a similar meaning but still a low n-gram overlap or the other way around. SPICE parses the captions into a set of propositions and compares them. Specifically, SPICE recognizes relations, objects, and attributes, and builds scene graphs for the generated captions as well as for the references. Figure 2 displays an example of a scene graph. Tuples of objects, relations, and attributes are extracted from these scene graphs, establishing a structured comparison between the reference and produced captions. The metric score ranges from 0 to 1, but the score decreases when the number of reference captions increases. For instance, given five reference captions, scores lie typically in the range of 0.15-0.20, while with 40 reference captions, scores lie in the range of approximately 0.03-0.07 [AFJG23]. This expected result is due to the impact of the recall component of the metric. Compared to other metrics, SPICE captures human judgments more accurately than other metrics.

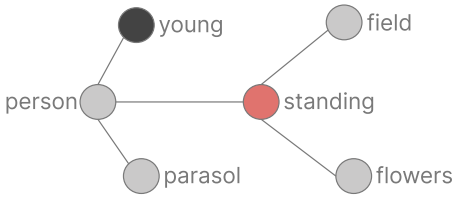


Figure 2: Exemplary scene graph of the caption *A young person with a parasol standing on a field with flowers* parsed by SPICE

3.3.4 Tag Accuracy

The mentioned evaluation metrics provide valuable insights but they all require actual captions or natural language sentences to serve as references. However, in some scenarios, only tags are available. Tags are keywords describing or categorizing an image, usually containing one to three terms. In such cases, these metrics cannot be used to assess the model's performance in a reasonable time and effort. This limitation highlights the need for a metric that can perform automatic evaluation given only tags or labels. We introduce a metric that measures tag accuracy. It determines to which extent the reference labels occur in the generated captions. To achieve this, synonyms for the existing tags are first identified and grouped. Next, all tags and synonyms are stemmed, and stop words are removed. The accuracy is calculated by dividing the number of matching tag groups, including the synonyms, by the total number of tags, as seen in Equation 1. The resulting score ranges from 0 to 1. A score of 1 indicates that all tag groups were present in the generated caption. A score of 0 means that none were present.

$$\text{Tag Accuracy} = \frac{\#\text{Matched Groups of Tags}}{\#\text{Tags}} \tag{1}$$

4 Results

To assess the effectiveness of the BLIP model for captioning art-historical photographs, we conducted a series of experiments on two distinct datasets. The following section presents our findings. The first dataset, Artpedia, provides image-text pairs, whereas the second dataset, WPI, consists of partially labeled images. Both datasets contain art-historical photographs and are presented in detail in subsequent sections. Our first experiment is finetuning the BLIP model on the Artpedia dataset, which is followed by finetuning the model on grayscaled Artpedia images. Additionally, we filter the Artpedia dataset and finetune the model on a smaller, higher-quality subset. All finetuned models are compared to the base model of BLIP. Each approach is evaluated on an Artpedia test subset and on the WPI dataset and quantitative and qualitative analysis was conducted.

Furthermore, we explore the potential of constrained caption generation, where words are forced to be included in the predictions. Finally, we investigate the model's ability to recognize the artists of the artworks.

4.1 Datasets

4.1.1 Artpedia Dataset

To finetune the BLIP base model, we utilize the manually annotated Artpedia dataset by [SCB+19]. The dataset consists of 2,930 paintings and 28,212 corresponding descriptions retrieved from Wikipedia. Given the URLs, we download the images from which 2,875 were accessible. As of now, 55 links are broken.

Diversity in historical periods might be important for finetuning as it allows for the model to learn a broad range of painting styles and visual attributes. The artworks date from the 13th through the 21st centuries, according to Figure 3, ensuring a diverse representation of historical art periods.

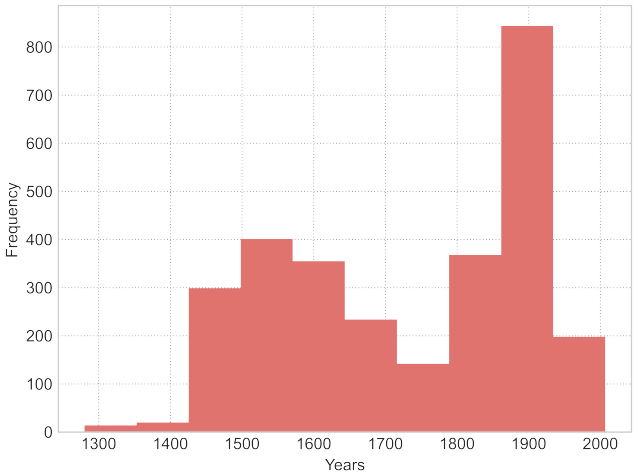


Figure 3: Distribution of Artpedia Paintings Periods

The descriptions are distinguished into 9,173 visual sentences and 19,039 contextual sentences. However, we end up with 9,014 visual sentences in total due to the unavailable images. The visual sentences intend to capture the visual content of the works whereas the contextual sentences describe their historical context. Only visual sentences are of relevance to us because we are captioning the images based on their appearance. However, the visual sentences also often contain contextual information.

For example, Figure 4 is described with the following visual and contextual sentences.

Visual sentence
It is also stylistically earlier to that work, being painted without pseudo-perspective, and having the angels around the Virgin simply placed one above the other, rather than being spatially arranged.
Contextual sentence
The Maest is a painting by the Italian artist Cimabue, painted around 1280 and housed in Mus du Louvre of Paris, France.

Figure 4: Example of an Artpedia image with visual and contextual sentences

Each painting is associated with three visual sentences. We take pairs of single-sentence captions and paintings as input for finetuning and evaluating the model. The visual sentences contain 28,541 different terms, of which 216 can be considered stop words. Excluding the stop words, the most frequent words include for instance “paintings”, “depicts”, “left”, and “shows”. More are shown in Figure 5. It is expected to see these words in the generated captions by the finetuned models as well.

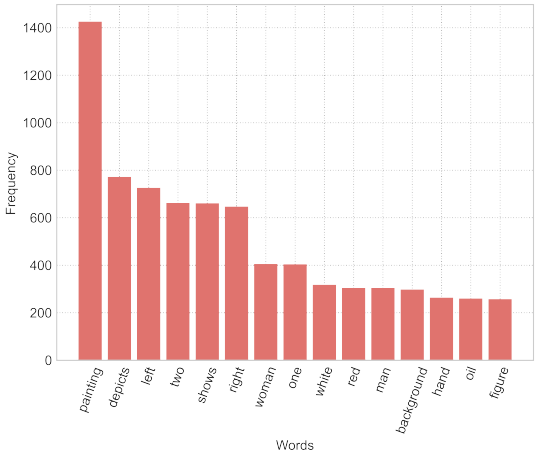


Figure 5: Frequent terms in the visual sentences in Artpedia dataset

4.1.2 Wildenstein Plattner Institute Dataset

The Wildenstein Plattner Institute (WPI) dataset consists of 19,130 entries. An entry contains the image URLs and title and might include genres, topics, names, and places.

We consider an image as labeled if there is at least one element in topics or genres. That results in 10,433 labeled entries, which is about 55% of the whole dataset. There are 193 distinct genres and 1,414 topics.

The most prevalent genres are for instance notes, letters, and photocopies. These might look significantly different from the Artpedia samples. The 15 most frequent genres are shown in Figure 6.

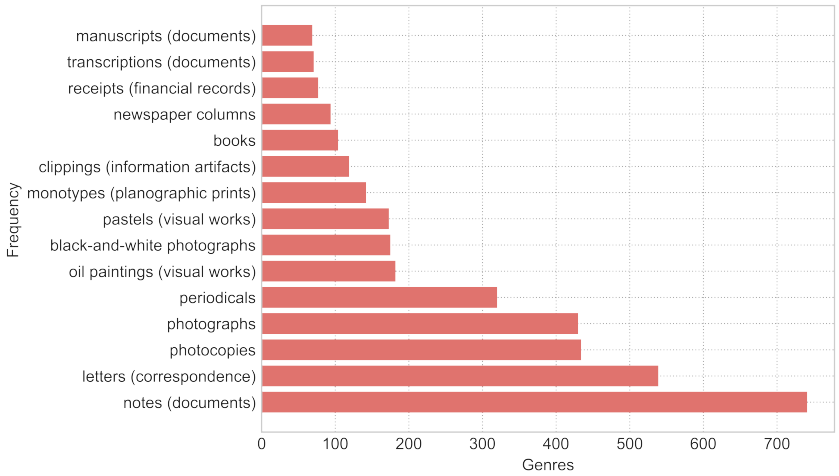


Figure 6: Frequent genres in the WPI dataset

Our objective is to create captions for artwork images. Therefore, we have excluded all samples from the WPI test dataset that contain images of documents (such as letters or prints). This involves two steps. Firstly, we eliminate all images that are tagged with a genre related to a document term. Secondly, we utilize the large BLIP model pre-trained on COCO to generate captions and subsequently remove all samples with captions containing the term ”document” or its synonyms. The resulting test dataset of WPI contains 7,463 image-tags-pair samples.

The two datasets are still similar to each other. For instance, both contain a higher frequency of the word “women” compared to “men”, and “oil” and “painting” are among the most frequent labels. More frequent topics of the WPI dataset are seen in Figure 7.

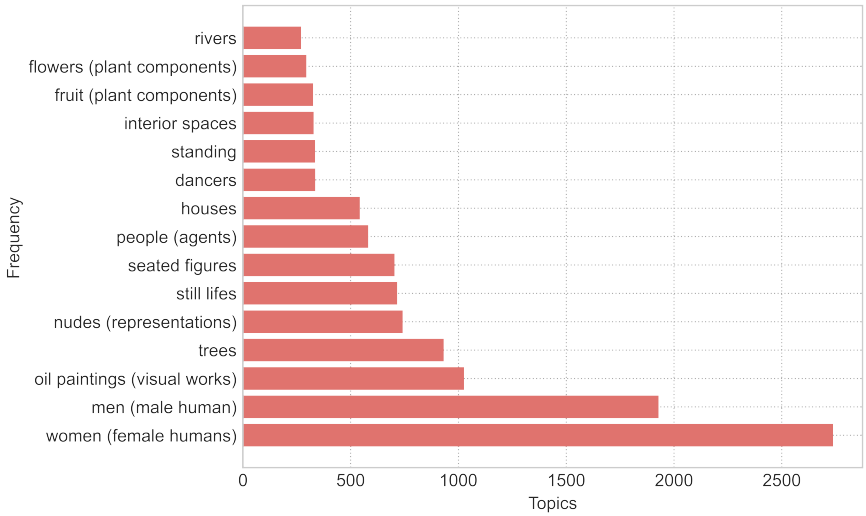


Figure 7: Frequent topics in the WPI dataset

4.2 Finetuning

Metric	Pretrained	Finetuned	Grayscale	Filtered
BLEU4	0.7	3.0	3.1	**3.6**
CIDEr	3.0	8.6	7.8	**10.6**
SPICE	4.6	5.8	4.9	**6.6**
Tag Accuracy	3.3	**16.7**	13.3	14.5

Table 1: Evaluation results on Artpedia (BLEU4, CIDEr, SPICE) and WPI (Tag accuracy)

4.2.1 Finetuning on Artpedia

For finetuning on the Artpedia dataset we use the training, validation, and test split provided by the Artpedia dataset [SCB+19]. The visual sentences serve as ground-truth captions. Due to the limited available amount of data, we want to use all visual sentences the dataset contains. During training, every image can only be associated with a single caption. Therefore, we include images multiple times with different captions. For our experiments, we work with the LAVIS library [LLL+22]. It implements the BLIP architecture in Pytorch and provides pre-trained models. The pre-trained model we use has been trained on the human-annotated datasets COCO [LMB+14], Visual Genome [KZG+16], and three web datasets.
We set the batch size to 16 because more images do not fit into the GPUs memory and

train for 5 epochs. The other hyperparameters are set to the default values which can be found in the BLIP paper. On a single NVIDIA Ampere GPU, this takes approximately 5 hours.

4.2.1.1 Quantitative Results

We additionally evaluated the models on the WPI dataset using tag accuracy. The results can be found in Table 1. Here, there is also a significant difference between the pre-trained and finetuned model. Hence, the BLIP model fine-tuned on Artpedia identifies topics and subjects of a painting in the WPI test dataset more accurately than the pre-trained model.

4.2.1.2 Qualitative Results

We additionally evaluated the outcomes of the experiment by comparing the captions created by the pre-trained and finetuned models for example images. This gives us more insights into how captions' quality changes.

In Figure 8 we can see some captions created by the pre-trained model and the model finetuned on the Artpedia dataset. Both models are not always correct. Yet, all captions mention the main content, two men/soldiers/sailors on a boat or at sea. The pre-trained model generates more general captions, whereas the finetuned model's captions are more descriptive. They include expressive adjectives like "gloom" and "stormy". The second caption from the finetuned model contains context information (French, German) that we cannot verify easily. We can see these differences in several examples.

Pretrained model:
1. *two men holding a machine gun while standing on a ship*
2. *two soldiers stand on a boat deck while holding their arms around each other*

Finetuned model:
1. *two sailors on the deck, on a gloom stormy day, drinking a beer*
2. *the french soldiers on the pier looking at a german battleship gun and barrel*

Figure 8: Example image with captions created by two different models

In Figure 9 we can find captions generated for an example from the WPI dataset. Here the caption from the finetuned model is again more detailed. Moreover, it contains art-specific content ("pen and ink drawing") and the painter of the drawings. In this case,

the painter is correct. Due to the context information in many Artpedia captions the model finetuned on Artpedia also often creates captions with context information. The pre-trained model describes the drawing as photos and does not include any context information.

Pretrained model:
several pictures are shown with words and images an image of a black and white set of photos
Finetuned model:
scenes of French countryside is seen here in a pen and ink drawing by vincent van gogh

Figure 9: Example image from the WPI dataset with captions created by two different models

4.2.2 Grayscale

To adjust better to the grayscale images in the WPI dataset we transform the images from the Artpedia dataset to grayscale. We then train and evaluate the model on these grayscale images. All parameters are the same as in the experiment using the original Artpedia dataset.

4.2.2.1 Quantitative Results For this experiment, we compare the results of the model finetuned on the grayscaled Artpedia dataset to those of the model trained on the original Artpedia dataset. The results can be found in Table 1. Only the BLEU4 metric is a little higher for the grayscale version when evaluating the model on the Artpedia test split. The other two metrics decrease. This might be because some ground-truth captions contain colors. These cannot be learned when using grayscale images. However, for us, the evaluation of the model on the WPI dataset is more interesting as the intent of this experiment was to prepare the model better for this dataset. The tag accuracy, found in Table 1, does not increase compared to the model finetuned on the original Artpedia dataset. Hence, the training on grayscale images did not have the desired effect of improving the results on the WPI dataset.

4.2.2.2 Qualitative Results

When looking at example captions for the models trained on normal and grayscale Artpedia there is no significant difference.

4.2.3 Filtering

The Artpedia dataset contains a significant number of poorly composed visual sentences including plenty of contextual information. To improve its quality, we use the Image-grounded Text Encoder from the BLIP framework to filter out inadequate captions. The filter computes matching scores for the images and their ground-truth captions. We use a pre-trained model from the LAVIS library. This model was trained on the same datasets as the captioning model from the previous experiments. We cannot finetune the filter model because we do not have access to labels that describe if the captions match the image or not. Figure 10 shows the distribution of the matching scores. The filter tends to assign extreme scores, with the majority of sentence-image pairs scoring close to 0 (indicating a very poor match) or close to 100 (indicating a high match).

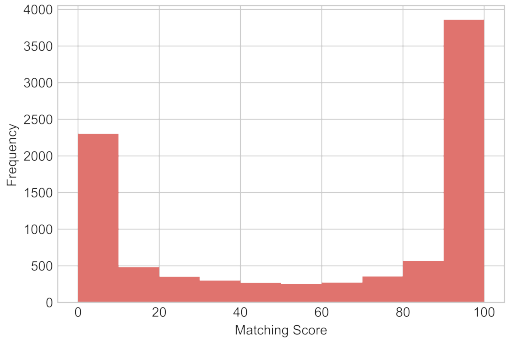


Figure 10: Matching scores distribution of the Artpedia dataset

We chose 80 as the threshold because the median lies at approximately 78. Therefore, only sentence-image pairs scoring at least 80 will be included in the training samples to finetune another BLIP model. It leaves us with about half of the samples in our training dataset, which is still enough to finetune the model. With a higher threshold, the dataset will become too small.

We finetune the same pre-trained model as in the previous experiments and use the same parameters. The evaluation also takes place on the original unfiltered test and validation sets to ensure comparability with the other experiments.

4.2.3.1 Quantitative Results

As it can be seen in Table 1 the model trained on the filtered dataset achieves the best results in all three metrics on the Artpedia test split. Even though the difference in the values is not as big as the difference between the pre-trained and finetuned model this is still a promising result.

On the WPI dataset, the tag accuracy is lower than for the model trained on the original Artpedia dataset. This shows us that filtering does not improve the model's performance on art images in general.

4.2.3.2 Qualitative Results In Figure 11 we can see an example of a matching score created by the filter for an image-text pair. This example demonstrates a common problem the filter has. For our purposes, captions with only visual information are needed. This caption contains a lot of context information, like when the painting was painted and who the person in the painting is. Still, the caption gets a good score. This shows us that the filter is not able to detect context information in the caption. It seems to pay more attention to a good sentence structure and words like "it depicts" or "the painting depicts". It is difficult to quantitatively evaluate the filter's performance. However, when looking through examples, we can find many occurrences of good caption scores for bad captions or bad matching scores for good captions. Therefore, we can assume that there is still room for improvement in the filter's accuracy.

Caption:
Painted in 1889, it depicts actress Ellen Terry in a famous performance of William Shakespeare's tragedy Macbeth, wearing a green dress decorated with iridescent beetle wings.

Matching score: 0.999

Figure 11: Example score for image-text matching

The difference between the quality of captions from the models trained on the unfiltered and filtered dataset cannot be seen by looking at examples.

4.3 Constrained Caption Generation

The WPI images are tagged with words or word groups (but not sentences or larger groups of words, that are linked together). We evaluate the impact of forcing tags (or

other words) in some images with constrained generation as described in subsection 3.2. We utilize a model finetuned on Artpedia.

Artpedia only provides full sentences as reference captions while WPI only offers tags as references. Both datasets lack supplementary information on visual features that can be used as input for constrained generation. Consequently, we cannot quantitatively compare the impact of constrained generation with previous experiments, as we cannot incorporate information from the reference as input.

The constrained generation approach guarantees the inclusion of certain words and produces syntactically correct captions, even when injecting wrong visual information (see Figure 12.3). The caption is further semantically correct if we force words, that are known as correct visual elements. Without forcing the WPI tag, the tree is not mentioned for the painting in Figure 12.1 in contrast to the caption Figure 12.2 although it is a relevant part. Furthermore, the constrained decoding mechanism enables you to target specific characteristics of the painting when you force certain words in the output: In Figure 12.4 you can see that the model correctly outputs information on the image's background.

1. Caption without constrainted generation
a painting of people sitting and standing around a campfire
2. Force a WPI tag: tree
The painting of a firepit and a dark tree in the painting's middle
3. Force false visual information: red roses
The painting shows a group of people sitting around a bonfire with a woman standing next to it and red roses
4. Force position: In the background
A painting of the woods at nighttime with a group of people dancing around a campfire in the background

Figure 12: WPI Example for constrained captioning with different force words

4.4 Artists - Context in Captions

Some samples in the Artpedia dataset have visual descriptions, that include the artist. As seen in Figure 9, the finetuned model is capable of generating artist information. However, we do not have insight into the model's overall performance: is the artist's

information correct, and how often does it occur? Additionally, the experiments focus on the question, of whether the finetuning process impacts the artist generation's ability. Thus we evaluate the pre-trained and the finetuned models on the original and filtered Artpedia test dataset: We examine whether BLIP is able to recognize artists correctly when seeing an image. Artist names for the Artpedia paintings are extracted from the Wikidata [VK14] knowledge database through the painting's title. The resulting test dataset consists of the image-artists pairs (instead of the visual descriptions as reference captions). A painting can have multiple or no painters, which is why the artists are stored in lists. The amount of empty artist lists in the test dataset due to unknown origins or outdated titles in Artpedia is negligible (28 of 1,025 samples). As the artist set in Artpedia is fixed, the artists can be treated as multi-labels. Therefore, each generated caption is checked for the presence of artists from the Artpedia paintings. A caption that meets the artist recognition task correctly identifies all ground truth artists without mentioning any other artists. For the purpose of comparing the artists in the artist recognition task, a precision, recall, and f1-score are assigned to each artist. We use the overall f1-score, precision, and recall for quantitative evaluation.

4.4.1 Evaluation of one Artist

First, we focus on an artist that is mentioned often in the generated captions by the finetuned model, Van Gogh. The training dataset consists of 105 samples painted by Van Gogh, and 63 reference captions utilized for finetuning including the artist's name. This allowed the model to learn to recognize patterns in the paintings and associate them with the artist, as it was exposed to many image-caption pairs of Van Gogh during training. Despite a significant increase in the score, it is worth noting that 28% of the captions

Metric	Pretrained	Finetuned
Precision	0.00	0.72
Recall	0.00	0.28
F1-score	0.00	0.40

Table 2: Scores for Van Gogh mentions in the generated captions, see more in Table 4

mentioning Van Gogh actually refer to paintings created by other artists, revealing the continued uncertainty in generating accurate artist recognition (Table 2). Nevertheless, the majority of generated captions miss the artist information of Van Gogh completely or use other designations for the artist: e.g. Van Gogh can be referred to as *Vincent Van Gogh*, *Van Gogh, Arles* or just *Van Gogh* etc. in the Wikidata database, as the names used for one artist are not unified. This explains the lower recall.

4.4.2 Quantitative Results

The generated captions from the pre-trained model have received consistently low metric scores, with all scores measuring at 0.0: These captions do not contain a reference to the painting's artists. Hence, the pre-trained models are not suitable for recognizing artists from the visual content.

Overall accuracy and recall are higher for the finetuned models (see Table 3). This observation suggests that the finetuning process enables the model to acquire knowledge of the painting's artist when seeing an image with reference captions including such information during training. As the model is not finetuned on the task of generating artist information, the recall is lower compared to the precision which means that the majority of generated captions do not output artist information. However, the model still outperforms the pre-trained ones in this regard.

The evaluation of generated captions that were finetuned on the filtered dataset reached slightly lower scores compared to the original Artpedia dataset that was trained on a larger dataset. The effectiveness of the pre-trained filter is limited for the artist recognition task. Nevertheless, the increased size of the unfiltered training dataset has only a small beneficial impact on generating artist information. Thus, using a filtered dataset customized for the task might pay off.

Metric	Pretrained	Finetuned	Filtered
Precision	0.00	0.20	0.17
Recall	0.00	0.40	0.39
F1-score	0.00	0.14	0.12

Table 3: Overall results for artist evaluation on Artpedia

5 Conclusion

We found that the BLIP captioning model transfers well to artwork images. Even a pre-trained version that was not trained on art images yet can already create general descriptions of artwork. When finetuning the model on the Artpedia dataset the resulting generated captions become more detailed and diverse. In comparison to the pre-trained model, they contain more art-related descriptions. However, we can also find more mistakes and often context information in the captions. We would like to avoid context information and instead focus on visual information. It is difficult to verify if context information is correct, so this only makes the task harder. The finetuned model's problems are most likely due to the caption quality in the Artpedia dataset. As the sentences are not originally intended to be captions, they are often not suited well. Our quantitative

evaluations on the Artpedia and the WPI dataset show that the metrics improve significantly with the finetuned model. Hence, BLIP models finetuned on datasets consisting of artwork image-text pairs are more suitable to be transferred to art-historical datasets, such as the WPI dataset, than pre-trained models.

A promising approach for improving the dataset is filtering it. Despite the filter not being completely accurate it already enables better finetuning results on the Artpedia dataset. Nevertheless, the model finetuned on the original Artpedia dataset transfers best to the WPI dataset. This shows the limitations of using only the general pre-trained filter for art-historical images.

Finetuning the model on grayscale images does not improve results compared to finetuning on original or filtered image datasets. This applies to the WPI dataset too. Hence, removing color information during training does not simplify caption generation with BLIP, even for historical datasets that provide only grayscale images.

To further improve the results for the WPI dataset, we can take advantage of the tags it provides. By using these to force words in the captions important content of the image is mentioned. This is a way to improve the captions by including additional sources of information beyond the image.

Even though our focus lies on visual content images, we also investigated the BLIP model's ability to predict context information, e.g. the artist. For artists that often appear in the Artpedia dataset, this is already possible, even though the mentioned artists are not always correct. Nevertheless, it shows predicting more context information about the art images is possible, especially with more training data.

In conclusion, we can say that we can already generate sensible captions for many artwork images: the results of all finetuning experiments indicate a significant improvement in the recognition of artwork-related topics for the WPI compared to pre-trained models. It is worthwhile to further investigate how the results can be improved even more.

6 Future Work

The problems we identified in our experiments led us to more ideas on how our captioning approach can be improved. Here, we present some directions that could be further investigated.

6.1 CapFilt

So far, we have only used a filter to improve our dataset. However, one could make use of the CapFilt approach described in the BLIP paper. This approach does not only include filtering out bad captions but also creating synthetic captions to create more

training data. Synthetic captions could be generated for the images from the Artpedia dataset that are filtered out, on WPI images or other art images collected from the web. Moreover, the filtering process can be improved by finetuning the filter. One approach to do this is to create labels for image-caption pairs that classify them as either positive or negative pairs.

Instead of classifying image-text-pairs, one could make use of the information already at hand to create a "context filter". This is a classifier that gets a caption as input and decides if it contains contextual information or not. It can be trained on the Artpedia dataset with the contextual sentences as one class and the visual sentences as another. After training, the filter can be used on the visual sentences to filter out those that contain context information.

6.2 Context Evaluation

In addition to identifying artists, captions may also contain references to entities from various contextual categories, such as geographic locations, human nationalities, or historical events. Thus, performing an in-depth analysis of pre-defined contextual categories is needed to assess the extent to which BLIP can learn context information for specific categories from both image content and associated textual data. This knowledge can be utilized to filter sentences containing context information BLIP is able to learn.

References

[AFJG16] Anderson, Peter ; Fernando, Basura ; Johnson, Mark ; Gould, Stephen: SPICE: Semantic Propositional Image Caption Evaluation, 2016. – ISBN 978–3–319–46453–4, S. 382–398

[AFJG23] Anderson, Peter ; Fernando, Basura ; Johnson, Mark ; Gould, Stephen: *SPICE: Semantic Propositional Image Caption Evaluation - Peter Anderson*. `https://panderson.me/spice/`, 2023. – (Accessed on 02/21/2023)

[BCGC18] Baraldi, Lorenzo ; Cornia, Marcella ; Grana, Costantino ; Cucchiara, Rita: Aligning text and document illustrations: towards visually explainable digital humanities. In: *2018 24th International Conference on Pattern Recognition (ICPR)* IEEE, 2018, S. 1097–1102

[Cet21] Cetinic, Eva: Iconographic image captioning for artworks. In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III* Springer, 2021, S. 502–516

[DBK+20] Dosovitskiy, Alexey ; Beyer, Lucas ; Kolesnikov, Alexander ; Weissenborn, Dirk ; Zhai, Xiaohua ; Unterthiner, Thomas ; Dehghani, Mostafa ; Minderer, Matthias ; Heigold, Georg ; Gelly, Sylvain ; Uszkoreit, Jakob ; Houlsby, Neil: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020

[DCLT18] Devlin, Jacob ; Chang, Ming-Wei ; Lee, Kenton ; Toutanova, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018

[GRM21] Gheini, Mozhdeh ; Ren, Xiang ; May, Jonathan: Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation, 2021, S. 1754–1765

[GV18] Garcia, Noa ; Vogiatzis, George: How to read paintings: semantic art understanding with multi-modal retrieval. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, S. 0–0

[KZG+16] Krishna, Ranjay ; Zhu, Yuke ; Groth, Oliver ; Johnson, Justin ; Hata, Kenji ; Kravitz, Joshua ; Chen, Stephanie ; Kalantidis, Yannis ; Li, Li-Jia ; Shamma, David A. ; Bernstein, Michael S. ; Fei-Fei, Li: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. In: *International Journal of Computer Vision* 123 (2016), S. 32–73

[LLL+22] Li, Dongxu ; Li, Junnan ; Le, Hung ; Wang, Guangsen ; Savarese, Silvio

; Hoi, Steven C. H.: *LAVIS: A Library for Language-Vision Intelligence*. 2022

[LMB+14] Lin, Tsung-Yi ; Maire, Michael ; Belongie, Serge ; Hays, James ; Perona, Pietro ; Ramanan, Deva ; Dollár, Piotr ; Zitnick, C.: Microsoft COCO: Common Objects in Context, 2014. – ISBN 978–3–319–10601–4

[LXH22] Li, Dongxu ; Xiong, Caiming ; Hoi, Steven: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022

[SCB+19] Stefanini, Matteo ; Cornia, Marcella ; Baraldi, Lorenzo ; Corsini, Massimiliano ; Cucchiara, Rita: Artpedia: A New Visual-Semantic Dataset with Visual and Contextual Sentences. In: *Proceedings of the International Conference on Image Analysis and Processing*, 2019

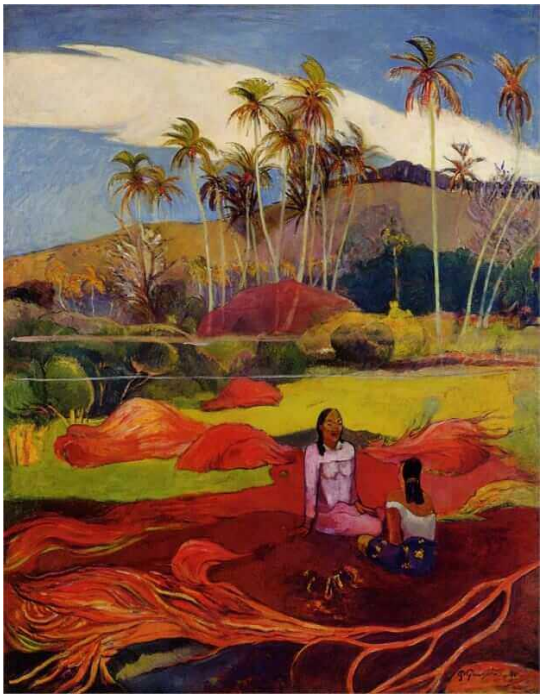
[VK14] Vrandečić, Denny ; Krötzsch, Markus: Wikidata: a free collaborative knowledgebase. In: *Communications of the ACM* 57 (2014), Nr. 10, S. 78–85

[VSP+17] Vaswani, Ashish ; Shazeer, Noam ; Parmar, Niki ; Uszkoreit, Jakob ; Jones, Llion ; Gomez, Aidan N. ; Kaiser, Łukasz ; Polosukhin, Illia: Attention is all you need. In: *Advances in neural information processing systems* 30 (2017)

[VZP14] Vedantam, Ramakrishna ; Zitnick, C. ; Parikh, Devi: CIDEr: Consensus-based Image Description Evaluation. (2014), 11

[ZPZ+20] Zhou, Luowei ; Palangi, Hamid ; Zhang, Lei ; Hu, Houdong ; Corso, Jason ; Gao, Jianfeng: Unified vision-language pre-training for image captioning and vqa. In: *Proceedings of the AAAI conference on artificial intelligence* Bd. 34, 2020, S. 13041–13049

7 Appendix



Pretrained model:
1. *A painting of a woman sitting in a field*
2. *an image of a painting with trees in the background*

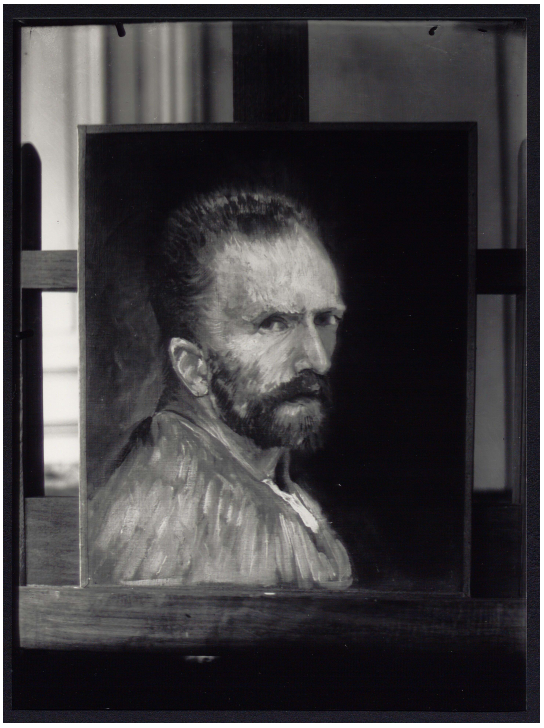
Finetuned model:
1. *a woman who is holding a baby, sitting on the ground while palm trees are in the background*
2. *a painting by paul gauguin that depicts a woman on a large red rug*

Figure 13: Example image from the Artpedia with captions created by the pre-trained and finetuned models

Pretrained model:
1. *a black and white photo of a woman sitting on a bench*
2. *this drawing depicts a woman with flowered hair, a square frame, and a picture of herself*

Finetuned model:
1. *an woman in traditional dress sitting on a cushion with flowers falling from her hat*
2. *frida is shown in the centre of the painting and has an elaborately decorated panel with black - and - white flowers*

Figure 14: Example image from the WPI dataset with captions created by the pretrained and finetuned models



Pretrained model:
1. *there is a portrait of an older man that has beard and long hair*
2. *this photo is black and white of a bearded man*

Finetuned model:
1. *the artist's face is an example of selfportraits, probably that in van gogh's paintings*
2. *his face is visible in the reflection, where he has placed an easel*

Figure 15: Example image from the WPI dataset with captions created by the pretrained and finetuned models

(a) Self-Portrait with Bandaged Ear (Vincent Van Gogh 1889)

Finetuned model
man with a bandaged face is a portrait of vincent van gogh, and one of his most known pieces of expressionism

(b) Caption correctly includes Vincent Van Gogh

(c) Paris Street; Rainy Day (Gustave Caillebotte 1877)

Finetuned model
a painting by the painter vincent van gogh

(d) Caption falsely includes Vincent Van Gogh

Figure 16: Captions with references to the artist Vincent Van Gogh

Metric	Vincent van Gogh	Rembrandt	Paul Gauguin	Caravaggio
Precision	0.72	0.58	0.5	0.67
Recall	0.28	0.21	1.0	0.16
F1-score	0.4	0.31	0.67	0.26

Table 4: Artist evaluation on a finetuned model (Artpedia), pre-trained model scores 0.0 for every artist and metric