UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES
INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dept. of Computer Systems and Computation

Synthetic data generation and data augmentation techniques for image captioning with Stable Diffusion and large language models.

Master's Thesis

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

AUTHOR: Prieto Medina, Daniel Alejandro

Tutor: Domingo Ballester, Miguel

ACADEMIC YEAR: 2023/2024

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advent of deep learning models has changed the landscape for many different fields and opened many new lines of research. After some years, the power and complexity of these models have scaled rapidly, but their need of vast amounts of good quality data has grown too. However, obtaining it on those quantities is by no means an easy or cheap work, specially when working with multimodal models, where you need to process and annotate different types of data.

One of the many task restricted by what is mentioned before is Image Captioning, an important vision-language task, which often requires a tremendous number of finely labeled image-captions pairs for learning the underlying alignment between images and texts [1].

On that sense, this thesis focuses on the proposal of a multimodal data augmentation method that leverages the combination of recent text-to-image model Stable Diffusion and Large Language Models, to not only add new samples of data from preexisting annotations, but for creating new samples from scratch, generating new coherent pairs of image-captions.

1.1 Motivation

Image captioning is an important task at the intersection of computer vision and natural language processing (NLP). For the longest time, researchers have focused their efforts on optimizing models, refining extraction of informative features or developing better training techniques. However, data is a critical dimension regarding the performance of the models, but not as explored in comparison to those other axis, despite its recent surge in interest.

In their work, [2] show us how in different vision tasks, notable improvements in model performance can be achieved just by increasing the number of data seen in training. Specifically, [1] show us that having a large amount of high-quality pairs of image and captions is often the desired for supervised image captioning tasks. This requires to invest large amount of time and human resources in annotating images with descriptive sentences, and doing so has resulted in some quite large datasets like COCO [3], which [4] argues that it has a problem of containing an important amount of low quality images, simplistic annotations and repeated contexts, limiting the generalization capabilities of a trained model on the aforementioned dataset.

Historically, as with many other vision task, classic data augmentation techniques have been used to address the data issue and increase the number of training examples, including both augmenting the images [5, 6] and textual captions [7–9]. These classic techniques, although still effective, do not address the problem of having an incredible limited amount of data or only having one type of the of the data, specifically the textual one. In that sense, seeing the success of works like [1, 10], where they leverage the power of generative models like Stable Diffusion to generate synthetic data, we apply similar methods for the image captioning task and explore what results can be achieved with the usage of a synthetic dataset, so going a step further, we use Large Language Models (LLMs) to increase the amount of textual data we have at our disposition and feed it to a Stable Diffusion model to generate the pairs of image and caption data. With that said, we hope to show how these new generative models could be a more efficient and cheap alternative to increase the diversity and quality of datasets for image captioning, or even a solution for those scenarios where the lack of real sets of data is highly limited and constrict the possible solutions.

1.2 Objectives

The primary focus of this academic study is to explore the usage of generative models to create synthetic datasets for the training of image captioning models and compare their performance against the usage of real and human annotated pairs of image-caption. To achieve this, we will address the following secondary objectives:

- Train an image captioning model with a popular dataset for the task, so we have a point of comparison for the models trained according to the following points.
- Develop a pipeline for generating new high quality images using different Stable Diffusion.
- Develop a pipeline for generating new consistent descriptive sentences using Large Language Models.
- Combine the previously mentioned pipeline and develop a multimodal modal pipeline for generating pairs of image-captions data points for image captioning.
- Create a new datasets using the aforementioned pipelines and test their performance against real data.

1.3 Structure

The structure of this thesis has been designed to build upon a knowledge base that allows us to understand the working of all the components that make up this academic work. In that same matter, and understanding there is no real state-of-the-art metric or evaluation we can use to measure our efforts, it's been included a review of related works so we can better understand the context of this work and what it achieves.

- **Chapter 2** presents a recap of the concepts and technologies used for the development of this academic work, explaining in more detail the task of Image Captioning, delving into what are the core generative models used during this work, Stable Diffusion and Large Language Models, and the usage of Data Augmentation techniques in Image Captioning.
- **Chapter 3** offers a review of relevant works related to our case of study, so we can better understand the context of what we are doing and the goals presented in this particular work.
- **Chapter 4** will detail our case of study, going through the selected models, how the synthetic data was generated and the function of our multimodal pipeline works.
- **Chapter 5** present the experimentation done and the results obtained when working with the with the synthetic datasets and the pipeline as a whole.
- **Chapter 6** finalizes this thesis by synthesizing it and discussing whether we have managed to address the proposed objectives. Likewise, we will also discuss future work and what open lines of research this academic work leaves.

Chapter 2

Technological context

2.1 Image Captioning

As an everyday occurrence, individuals are exposed to vast amounts of images originating from diverse sources, such as the internet, television, newspapers, publicity or promotional materials, etc. While humans possess the innate ability to comprehend these visuals without the help of detailed captions, translating this capability to machines has presented a difficult task throughout the years.

First and foremost, image captioning is essential for advancing the development of artificial intelligence systems that can comprehensively understand visual content. Unlike tasks such as image classification, where the goal is to assign a single label to an image, captioning requires a deeper, more nuanced understanding of the scene, as it involves identifying objects, actions, spatial relationships, and context [11]. This capability to generate coherent and semantically rich descriptions from images a requires a higher level of understanding of the language from the machines.

Now then, when talking about specific applications of this task, image captioning significantly enhances content retrieval and organization. In the context of image search engines and multimedia databases, textual descriptions generated by captioning systems enable more accurate indexing and retrieval of images [11,12]. Rather than relying solely on metadata or manual tagging, which can be incomplete or inaccurate, automatically generated captions provide a more detailed and contextually appropriate description of images, improving the relevance and precision of search results. This has significant implications not only for general web-based image searches but also for more specialized domains, such as medical imaging, where accurate description and retrieval of visual data are essential for diagnostic and research purposes.

Moreover, the task is critical for improving accessibility, particularly for individuals with visual impairments. Automated image captioning technologies provide an invaluable tool for translating visual information into text, allowing people with disabilities to access and engage with online content, from social media to educational materials [13].

2.1.1 Encoder-Decoder architecture for Image Captioning

As for any other vision task, understanding the contents of an image largely depends on the quality of the features extracted and how informative they are. The methodologies employed for the extraction of these features can be classified into two main categories: traditional machine learning techniques and deep machine learning techniques, with the latter being the primary focus of this study.

As explained by Herdade et al. [14], the most conventional approach used for deep machine learning techniques is the encoder-decoder architecture, wherein an input image is transformed into an intermediate representation that encapsulates the information present in the image, which is then translated into a sequence of descriptive text. This encoding process may involve a singular feature vector produced by a Convolutional Neural Network (CNN) [11, 12], or it may incorporate multiple visual features extracted from various segments of the image, which are then fed to a Recurrent Neural Networks (RNNs) in order to generate captions. In instances where multiple features are utilized, these segments can either be sampled uniformly or determined using an object detection mechanism, the latter of which has been evidenced to enhance overall performance [14]. In the Figure 2.1 we can see a diagram showing in a simpler way the described architecture.

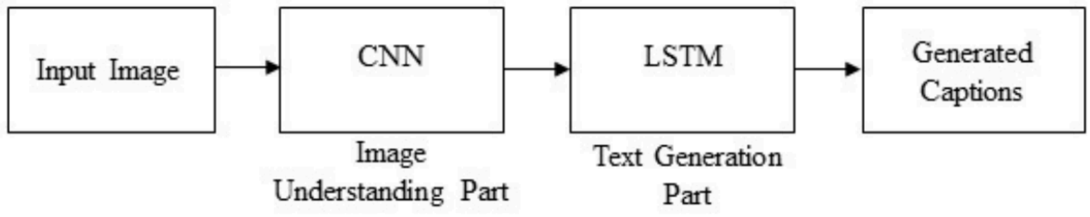


Figure 2.1. Diagram of a common encoder-decoder image captioning model.

This has been the common approach for many years, however, with the advent of Transformers and encoder-decoder attention based models, the majority of the recent efforts on the task leverages this new technologies, replacing common use of CNNs and Long Short-Term Memory (LSTMs) as the models for the encoder and decoder modules [12].

2.2 Data Augmentation

As stated in previous sections, deep learning models and any field of artificial intelligence in general, have been able to go through many breakthroughs in the last couple of years and the new advances seems to only keep coming faster. Shorten and Khoshgoftaar [15] mention that the fast pacing of all these advances has been fueled by the combination of the latest deep network architectures, more powerful computation, and access to big data.

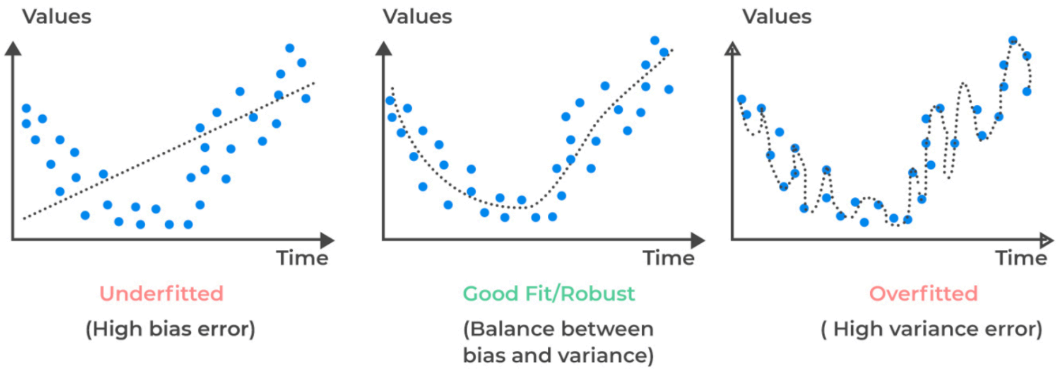


Figure 2.2. Examples of different fitting states of a network.

Improving the generalization ability of these models is one of the most difficult challenges and the proper way to do it varies depending on the type of data and task at hand. Generalizability in machine learning refers to a model's ability to perform well on new, unseen data that was not part of the training dataset. It is a key measure of a model's effectiveness, as the goal of most machine learning systems is not merely to perform well on the training data, but to accurately predict or classify new instances that the model encounters in real-world scenarios [15].

Generalizability hinges on the concept of the model capturing the underlying patterns and structures in the data rather than memorizing the specific details or noise present in the training set [16]. When a model generalizes well, it is able to infer from the learned data and apply those insights to different, but related, data. Poor generalization typically indicates overfitting, where a model has become too specialized to the training data, capturing noise or irrelevant features that do not extend to new data points [17]. Underfitting, on the other hand, occurs when a model is too simplistic and fails to capture the complexity of the data, leading to poor performance on both training and unseen data.

When constructing a model, the desired tendency during training is that the validation error must continue to decrease with the training error. The application of one or more data augmenting techniques helps into achieving this decrease through the generation of new examples that hopefully can represent a more comprehensive set of possible data points, thus minimizing the distance between the training and validation set, and as well as any future testing sets [15]. Even though there exists other types of successful regularization techniques, like dropout, batch normalization, transfer learning and some more, data augmentation approaches overfitting from the root of the problem, the training dataset [16]. However, it must be noted that this is done under the premise that the original dataset can yield additional insights through various augmentation techniques. Such augmentations serve to artificially increase the size of the training dataset, and this is normally achieved either through methods of data warping or by employing oversampling strategies [15].

In the context of supervised learning, data augmentation is commonly applied in domains such as image processing, natural language processing, and speech recognition. For image-based tasks, augmentation techniques might include transformations like rotations, flips, cropping, scaling, and color adjustments [17, 18]. In natural language processing, augmentation can involve strategies such as synonym replacement, word swapping, or paraphrasing. The idea is to create new samples that retain the essential features of the original data while introducing variability to improve the model's robustness [7].

2.3 Diffusion Models

Diffusion models in machine learning are a class of generative models that learn to generate data by gradually denoising a random noise signal, effectively reversing a diffusion process. These models are based on a process where data is progressively corrupted by noise over a series of steps, and the model is trained to reverse this process, recovering the original data from the noisy input. This approach is grounded in stochastic differential equations and draws inspiration from thermodynamics, where a system evolves from order to disorder [19]. In diffusion models, the reverse process allows the system to move from disorder (noise) to a structured output (data), such as images or text.

In practice, the training of diffusion models involves learning the parameters of a neural network that can predict how to reverse each step of the noise corruption process. During inference, the model starts with a sample of pure noise and iteratively denoises it to generate new samples that resemble the distribution of the training data [20, 21]. This iterative refinement process allows diffusion models to produce high-quality, realistic samples across various domains, such as image synthesis and speech generation [19].

2.3.1 Stable Diffusion

Stable Diffusion is a deep learning model designed for generating high-quality images by applying diffusion processes, specifically focusing on generating images from textual descriptions. It belongs to a class of diffusion models that incrementally denoise a random noise distribution to produce structured outputs, following a reverse process of a learned diffusion model. In contrast to other types of Diffusion models, Stable Diffusion employs a latent diffusion process, which means that the denoising is performed in a latent space (a compressed representation of data) rather than directly on pixel-level data [18, 19, 21], as shown in the Figure 2.3. This reduces computational costs and increases the efficiency of the model while still producing highly detailed images.

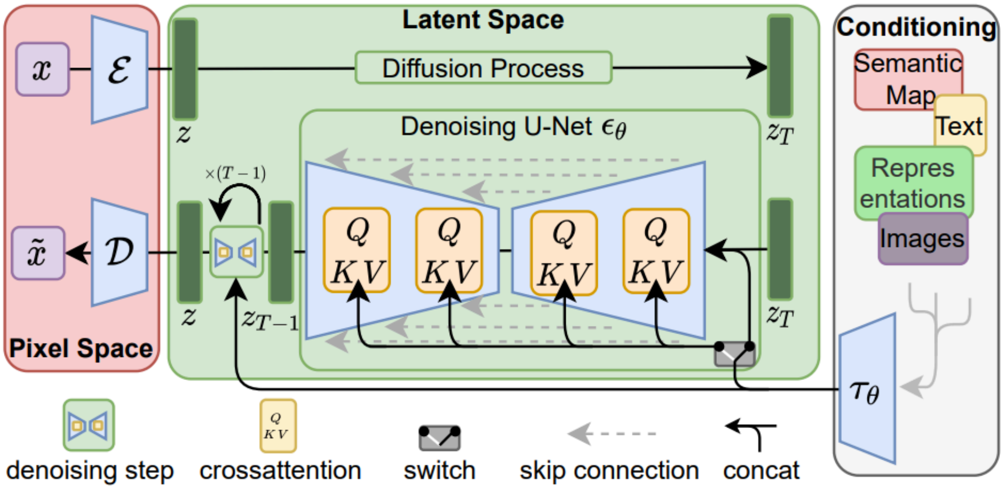


Figure 2.3. Overview of Latent Stable Diffusion by Rombach et al.

The model operates by being conditioned on text, often using a pre-trained language model like CLIP [22] (Contrastive Language-Image Pretraining) to interpret textual descriptions. By coupling text and image representations in this manner, Stable Diffusion can generate images that align with the semantic content of the input text [18]. It is designed to produce images iteratively, starting with a random noise input and gradually refining the image across several steps based on learned denoising processes, resulting in a coherent visual representation of the provided textual input [19, 21].

Stable Diffusion has become a prominent model in the domain of generative AI due to its ability to produce diverse, high-quality outputs with relatively lower computational demands compared to other models [18]. It has broad applications in art generation, content creation, and other domains where synthesizing realistic or creative images from textual prompts is required. The model's architecture and training methodology also emphasize scalability and versatility, making it adaptable to various use cases in computer vision and natural language processing.

There are, in general, three main components in latent diffusion as shown in Figure 2.4:

- A **Variational Auto-encoder (VAE)** [23].
- A **UNet**.
- A **text encoder**.

The function of the VAE is twofold: encoding complex visual data into a compact, continuous latent representation and then decoding the latent representations back into the original image space [19, 24]. This allows the Stable Diffusion model to operate efficiently in the latent space, significantly reducing the computational burden that would be required if diffusion were applied directly to pixel-level data.

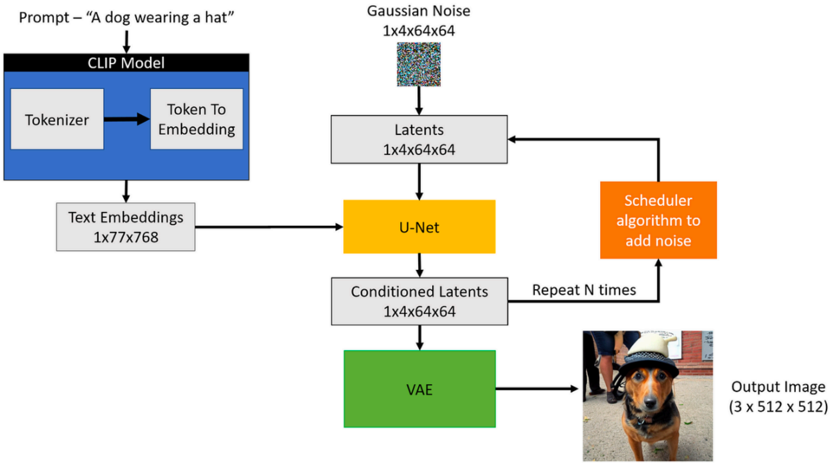


Figure 2.4. Diagram with a more simple look on Stable Diffusion.

The VAE in Stable Diffusion consists of two primary components: the encoder and the decoder. The encoder maps the high-dimensional input image into a lower-dimensional latent code, capturing essential features while discarding redundant information [25]. This latent space is where the diffusion process, which involves iteratively denoising random noise, takes place. Once the diffusion process is complete, the decoder reconstructs the latent representation back into the image space, producing a high-quality image from the refined latent representation [24, 25].

By incorporating the VAE, the Stable Diffusion model gains the ability to work in a more computationally tractable space, which allows for faster and more resource-efficient image generation without sacrificing quality [20, 25].

The U-Net [26] architecture serves as the core neural network responsible for learning the denoising process during the diffusion phase. Specifically, U-Net is used to progressively remove noise from a latent representation of the image, transforming noisy inputs into clean, coherent outputs. This process is essential in the reverse diffusion process, where the model begins with random noise and iteratively refines it to generate an image that corresponds to the input text prompt [25].

The U-Net architecture is well-suited for this task due to its ability to capture both local and global features through its symmetric structure, consisting of an encoder (downsampling) and a decoder (upsampling) [25]. The encoder progressively reduces the resolution of the input, extracting high-level, abstract features, while the decoder reconstructs the input by gradually increasing its resolution, this behavior. Skip connections between corresponding layers in the encoder and decoder allow for the preservation of fine-grained details, facilitating a more accurate reconstruction during the denoising process [26]. These connections are crucial for ensuring that both low-level details and high-level contextual information are retained throughout the generation process.

Lastly, the text encoder transforms the input prompt into an embedding space, which then acts as input for the U-Net. This embedding offers crucial guidance for the noisy latents during U-Net's denoising training. Generally the text encoder is a simple transformer-based model that converts a sequence of input tokens into a corresponding sequence of latent text embeddings. Importantly, Mishra [25] comments, Stable Diffusion does not create a new text encoder; instead, it employs an existing pre-trained text encoder, specifically CLIP, as introduced in the original research.

2.4 Large Language Models

LLMs are advanced artificial intelligence systems designed to process and generate human language. These models are built using deep learning techniques, particularly the Transformer architecture, and are trained on vast amounts of text data from diverse sources, that includes text from books, websites, articles, and other sources. LLMs are capable of performing a variety of natural language processing tasks, including translation, summarization, and text generation, by learning patterns in language structure and semantics from the training data [27–29]. Their large-scale architecture allows them to capture complex linguistic nuances, making them highly effective for a wide range of language-based applications via text generation [28, 29].

To make a quick distinction, current LLMs utilize comparable Transformer architectures and pre-training goals, such as language modeling —that are also employed by smaller language models. Nevertheless, LLMs markedly increase the dimensions of the model, the volume of data, and the overall computational resources, often by several orders of magnitude, leading to a substantial enchantment of the models capabilities, like many researchers have concluded [27].

2.4.1 GPT-3 model

GPT-3 [30], or Generative Pre-trained Transformer 3, is a state-of-the-art language model developed by OpenAI. It is based on the Transformer architecture and represents a significant advancement in natural language processing due to its unprecedented scale and performance. GPT-3 is pre-trained on massive datasets containing diverse text from a wide range of sources, enabling it to learn complex language patterns, semantics, and contextual understanding [31, 32].

With 175 billion parameters, GPT-3 is capable of generating highly coherent and contextually relevant text across various tasks. It can perform language-related tasks such as translation, summarization, text completion, and question answering without task-specific fine-tuning, relying on its general understanding of language learned during pre-training [30,31]. One of the distinguishing features of GPT-3 is its ability to produce human-like text in response to prompts, making it highly versatile for applications like content generation, dialogue systems, and automated writing.

Despite its impressive capabilities, GPT-3 has limitations, such as occasional factual inaccuracies and an inability to understand beyond the patterns it has learned during training. Nevertheless, it marks a significant step forward in the development of large-scale language models and serves as a foundation for ongoing research in natural language processing, artificial intelligence, and multimodal understanding [32, 33].

2.4.2 LLaMA model

The LLaMA (Large Language Model Meta AI) language model, developed by Meta (formerly Facebook), is a state-of-the-art generative language model designed to advance the field of NLP. LLaMA builds upon the Transformer architecture, similar to models like GPT, but distinguishes itself through its focus on efficiency and performance in low-resource settings [33, 34].

LLaMA is designed with the goal of achieving strong performance in natural language tasks while utilizing fewer parameters than many contemporary large-scale models. This approach allows it to maintain competitive accuracy and versatility across various NLP tasks, such as text generation, summarization, question answering, and translation, while being more computationally efficient. The model is pre-trained on a diverse dataset consisting of text from books, articles, and websites, allowing it to learn linguistic patterns and context for a broad range of applications [34].

The model's smaller size, relative to other large language models like GPT-3, makes it more accessible for researchers and developers who may not have access to extensive computational resources [33]. Despite its reduced parameter count, LLaMA achieves high performance on benchmarks and demonstrates that smaller models can still deliver strong results, provided that they are trained effectively and optimized for specific tasks [35].

Chapter 3

Related work

This section synthesizes prior works directly relevant to our topic, highlighting methodologies, findings, and theoretical frameworks that have shaped the field. By critically analyzing existing literature, we identify gaps and limitations that our research aims to address, positioning our work within the broader academic conversation. This review not only validates the significance of our study but also demonstrates how it extends or challenges established knowledge

3.1 Data issues and limitations in Image Captioning

In recent years, there has been significant progress in image captioning models, primarily driven by advancements in deep learning techniques. A widely adopted and effective framework is the Encoder-Decoder architecture, which employs CNNs to extract features from images and utilizes RNNs to translate these features into coherent natural language descriptions [11]. The integration of attention mechanisms [36] and Transformers [37] has further enhanced the performance of these models. Nevertheless, the training of image captioning systems through fully supervised methods necessitates extensive datasets comprising paired images and captions, leading to suboptimal performance of state-of-the-art models when data is scarce [1]. To mitigate this issue, researchers have proposed unpaired image captioning (UIC) strategies alongside data augmentation methods.

UIC aims to generate captions using models trained on unpaired images and captions, which can be obtained independently from various online sources, thereby alleviating the financial burden associated with collecting paired datasets. This innovative approach has garnered significant interest within the academic community. For example, Gu et al. [38] implemented language pivoting by incorporating additional caption data in Chinese, while Feng et al. [39] created a UIC framework that leverages a visual concept detector. Other notable contributions, such as those by Laina et al. [40], have focused on utilizing extensive text corpora to develop shared multi-modal embeddings for images and textual descriptions. Recent advancements, including the work of Zhu et al. [41], have sought to enhance UIC by tackling visual concept recognition using only image-level labels. Furthermore, semi-supervised learning techniques have been employed to augment these methodologies, with Chen et al. [42] generating absent visual information from textual inputs and Kim et al [43]. utilizing Generative Adversarial Networks (GANs) to assign pseudo-labels to unlabeled images. A majority of UIC research relies on external datasets beyond traditional image captioning and frequently necessitates additional information or annotations, which can incur substantial costs.

In the realm of data augmentation for image captioning, the main goal is to create supplementary training data that exhibits increased diversity. Although augmentation strategies have been widely utilized for images and text independently across various machine learning applications, their application in vision-language tasks, such as image captioning, is relatively rare [44]. Most investigations into data augmentation within this domain tend to concentrate on either the visual component or the textual description, rather than concurrently adjusting both elements. For example, researchers like Katiyar and Borgohain [5], along with Wang et al. [6], have utilized conventional image manipulation techniques, including cropping, flipping, and mirroring. Nonetheless, these transformations can inadvertently introduce noise by altering the inherent meaning of the images. On the textual front, Atliha and Šešok [44] implemented methods such as synonym replacement and paraphrasing using BERT, while other studies have examined techniques like word permutation, back translation, and various natural language processing (NLP) augmentation strategies.

Nevertheless, the field of multi-modal augmentation, which involves the simultaneous modification of both images and their corresponding captions, remains largely underexplored [1], and even less trying to fabricate a dataset without parting from some type of real refenrence. Feng et al. [39] introduced an innovative approach that integrates CutMix [45] with caption editing, wherein segments from a different image are incorporated, and the caption is revised to accurately reflect the modified image. Despite the proposal of this method, it has yet to be practically applied, and no empirical results have been presented to validate its effectiveness. In that sense, this work hopes to introduce a valuable pipeline that not only can simultaneously augment both types of data, but create new datasets from scratch, without having any reference from real world examples.

3.2 Image and caption synthesis

Text-to-image synthesis represents a sophisticated multimodal challenge that entails the creation of high-quality images derived from descriptive textual prompts. Traditionally, this domain has been primarily influenced by generative models, notably Generative Adversarial Networks (GANs) [46] and Variational Autoencoders (VAEs). Nevertheless, these models exhibit specific drawbacks: GANs are often challenging to optimize and are limited to datasets with restricted variability, while VAEs, despite their ability to generate high-resolution images, frequently do not meet expectations regarding sample quality [1]. Recently, a novel category of deep generative models, referred to as diffusion models, has emerged, showcasing enhanced performance and outpacing GANs in the realm of text-to-image generation. Although cutting-edge models such as DALLE-2 [47] and Imagen [48] can produce exceptionally high-quality images, their inference processes tend to be resource-intensive and protracted. In this research, we utilize an advanced Latent Diffusion Model, specifically Stable Diffusion, for the purpose of text-to-image synthesis [1, 10, 28, 49]. Stable Diffusion is capable of generating high-resolution images that are on par with those produced by leading models, while simultaneously demanding considerably fewer computational resources and time [2].

Conversely, the application of generative models in image captioning tasks has been relatively rare. For instance, Kim et al. [43] employed CycleGAN as a foundational model for unpaired image captioning, although its effectiveness was limited. More recently, Li et al [50]. illustrated that the optimal caption for an image is the one that facilitates the most precise reconstruction of the original image, utilizing Stable Diffusion for text-to-image generation alongside Flamingo [51] for the reverse process. However, despite the notable achievements of diffusion models in text-to-image generation, their potential for enhancing the quality of generated image captions remains largely unexplored.

Chapter 4

Case of Study

This section will present the pipeline constructed for creating the synthetic data developed in this academic work and the set up created for its testing, in that sense, we will go through the generation of the images and captions from real data, the assemble of the pipeline for generating fully synthetic data and the models throughout all the experiments.

4.1 Dataset selection

We use the Dataset Flickr30K [52] as the base of our experiments and the construction of the synthetic datasets. This a common dataset used for different vision task and consist of 31,783 annotated images, sourced from the Flickr platform, with five different descriptive sentences each, so having 158,915 captions.

The captions, written by human annotators, provide rich and varied descriptions of the visual content. This makes the dataset valuable not only for generating captions but also for understanding the relationship between visual features and natural language descriptions [52, 53]. The diversity of captions reflects the different ways in which people can describe the same image, helping to train models that are robust to linguistic variations.

Flickr30k has been a key dataset for training and evaluating machine learning models aimed at generating coherent and contextually accurate captions for images [53]. It is used extensively in image captioning tasks where the goal is to generate human-like descriptions, as well as in broader vision-language research such as visual question answering (VQA) and visual grounding.

The decision to use Flickr30k instead of COCO, which is probably the most used dataset for benchmarking and comparison for image caption stems from the fact that we are going to recreate a scenario where we have a limited amount of data, so having more than a 100 thousands images for our experiments wans't necessary and working with a smaller dataset facilitated working with it. On the other hand, COCO pairs of image-captions have an irregular amount of descriptive sentences per image, contrary to Flickr in which all pairs have 5 captions, in these was a desirable property for the experiments we wanted to conduct.

4.2 Generating synthetic images

So, first of all, when working Flickr, we used the Karpathy [54] split set as it is the most common way of working with these datasets, so that is how we defined our train, test and validation partitions. On the other hand, we reduced our training dataset to 12 thousand examples, and this was mainly for two reasons, first is what was already mentioned, we want to work parting from the fact we have a limited amount of data, and secondly, to ease the amount of computation need for the execution of all the experiments.

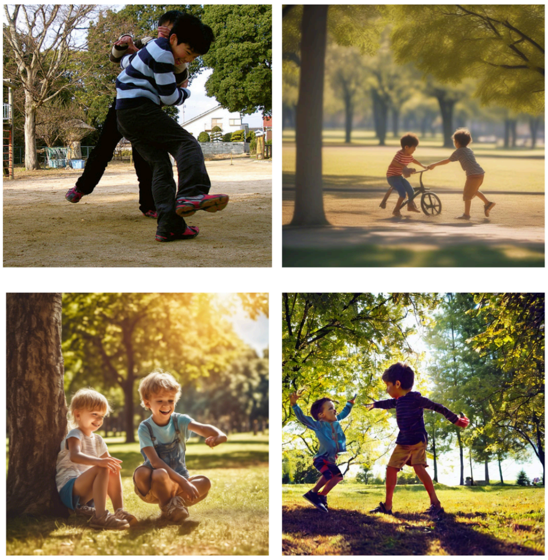


Figure 4.1. Examples of images generated with different diffusion models, the upper left one being the real image, the upper right generated with Stable Diffusion XL-Base-1.0, the lower left with Stable Diffusion 3, and the lower right with Stable Diffusion 2.1.

First of all, we have Flickr, which consist of all the pairs of image-captions as they come from the dataset and will be used for comparison. For generating our new pairs of data we used 3 different models of Stable Diffusion, all of them available through Hugginface or the Stability Ai website, these were Stable Diffusion 2.1, Stable Diffusion 3 and Stable Diffusion XL-Base-1.0 [20]. As mentioned before, we apply our different Stable Diffusion models using the captions found on Flickr for generating our synthetic images, with this, we form our first synthetic datasets called SFlickr *image* , were we have pairs of real captions and generated images. For the generation of the image we gave the longest description, going with the intuition that these were the more detailed ones regarding the description of the image. The predefined settings were used for all models except for Stable Diffusion 2.1, were we increased the number of denoising steps for better quality images, but only a small amount was added taking into consideration the increase on computation time.

Regarding the prompting, for each of the models a very simple prompt was used, asking to generate an image in accordance to the caption, the only relevant detail here was the specification of making the generated images realistic.

On another note, we relied on the paid API of Stability Ai for the generation of the images, the other two models were used on machines at our disposal. As a side note, asking the images to be realistic or not was one of the possible options at our disposition using the Stability Ai API, but nonetheless, we kept the specification of realism on the prompt, assuring results like in Figure 4.1.

4.3 Generating synthetic captions

As suggested by prior research on Image Captioning, the diversity of the descriptive sentences for the images plays a relevant role on increasing the performance of the models [1]. Thus, simulating a lack of annotations for the images, we randomly discarded 4 out of the five captions we had for each pair and rephrased the remaining one, again, opting for using the longest description out of the five. With these, we have obtained our second and third dataset denoted as SFlickr *caps* and SFlickr *pairs*. This was done using GPT-3.5-turbo [30] and Llama 3.1-8B-Instruct [35] for paraphrasing the captions like it has been done in related works [2, 44, 49]. This lead to create a new dataset Flickr *limitcaps* where the number of captions was randomly reduced up to only 2 captions per image.

This was the part if the experiment that required the most tweaking with the prompts, but we will delve more into this issue in the next chapter. An parameters, we kept the temperature of the model low, at around 0.5, this was enough for the models to have the ”creativity” to paraphrase the captions with sufficient variations. As for the final prompt. we had the most succes using the following one:

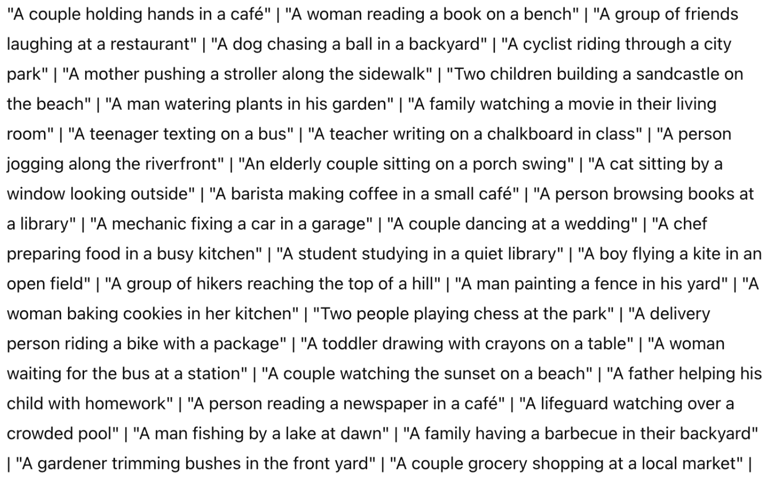
“Your job is to paraphrase a sentence, keep the main adjectives and verbs unchanged, the main subjects of the phrase should be consistent. For example: ”Two kids playing in a park” could be paraphrased to ”A couple of kids playing outdoors at the playground”.“

This prompt was sufficient to generate good and different enough captions for most cases.

4.4 Generating synthetic image-captions pairs

Next, we discarded the idea that we had the rest of the examples as only captions and created totally synthetic pairs of image and captions. For this, we used GPT-3.5-turbo to randomly generate descriptions of common day scenarios (Figure 4.2), then use those with Stable Diffusion XL-Base-1.0 to generate the images, and rely again on GPT-3.5-turbo to paraphrase the captions, creating the SFlickr *fully* dataset. However, thanks to the limitations on the usage of the language model and the increasing difficulty of keep the descriptions diversified enough, we could only successfully generate 7000 fully synthetic pairs.

Like this, combining all the other modules developed previously, we have a pipeline for creating pairs of image-captions from scratch, without relying in any human annotation or previously existing data.



"A couple holding hands in a café" | "A woman reading a book on a bench" | "A group of friends laughing at a restaurant" | "A dog chasing a ball in a backyard" | "A cyclist riding through a city park" | "A mother pushing a stroller along the sidewalk" | "Two children building a sandcastle on the beach" | "A man watering plants in his garden" | "A family watching a movie in their living room" | "A teenager texting on a bus" | "A teacher writing on a chalkboard in class" | "A person jogging along the riverfront" | "An elderly couple sitting on a porch swing" | "A cat sitting by a window looking outside" | "A barista making coffee in a small café" | "A person browsing books at a library" | "A mechanic fixing a car in a garage" | "A couple dancing at a wedding" | "A chef preparing food in a busy kitchen" | "A student studying in a quiet library" | "A boy flying a kite in an open field" | "A group of hikers reaching the top of a hill" | "A man painting a fence in his yard" | "A woman baking cookies in her kitchen" | "Two people playing chess at the park" | "A delivery person riding a bike with a package" | "A toddler drawing with crayons on a table" | "A woman waiting for the bus at a station" | "A couple watching the sunset on a beach" | "A father helping his child with homework" | "A person reading a newspaper in a café" | "A lifeguard watching over a crowded pool" | "A man fishing by a lake at dawn" | "A family having a barbecue in their backyard" | "A gardener trimming bushes in the front yard" | "A couple grocery shopping at a local market" |

Figure 4.2. Extract of random descriptions generated using GPT-3.5-turbo.

Doing a recap of all the synthetic datasets created, we have the following list:

- SFlickr *image* consist of 12 thousand pairs of image and 5 captions where the images where generated using three different Stable Diffusion models that we are denoting with a SD2.1, SD3 and SDXL. For the models running on our machines, we encountered that sometimes, especially for the Stable Diffusion 2.1 model, we had some occurrences of images being only a black background, so we had to discard and repeat the process when we had average pixel values of 0 for the generated images.

- SFlickr *caps* consist of 12 thousand pairs of image and 1+n captions, where those n captions are the ones generated through paraphrasing, noting that some of the captions n captions are discard for being the same using a simple character based approach, as we don't wanted to compare them in a semantic way. Knowing that n is less or equal to 4, each image would always have one of the original captions and up to four synthetic ones, meaning that all the pairs had less or exactly 5 captions.

- SFlickr *pairs* is a combination of the previous two datasets, where the pairs consiste of synthetically made images and paraphrased captions deriving from a subset of the original captions in Flickr.

- SFlickr *fully* consisted of 7 thousand pairs of image and n captions. The base captions were discarded the same way as in the previous dataset, and the process of generating the images and additional captions per example was done the same way as described for the other first two datasets.

4.5 Image captioning model

The model used for testing our new synthetic datasets is based on the work of Xu et al. [36]. The resulting model keeps the same architecture as described in their work, however, seeing the recent increase on the usage of Vision Transformers as the backbone for the image encoder mentioned by Croitoru et al. [55], the model used a Swim Transformer [56], specifically one of the checkpoints available through the models API of PyTorch.

For the most part, we used a standard architecture for Image Captioning having the Swim model as the image encoder, a simple attention mechanism described on the referenced paper and a Long Short-Term Memory (LSTM) module to generate captions, just as described in the Section 2.1. The selection of a simpler model and not trying to recreate the latest state-of-art at image captioning stems from the fact the we are only trying to address the issues related to data used for training these types of models, so the knowledge produced during this experiments should translate to more complex models and, in such case, benefit at some degree too.

4.6 Metrics selection

For the evaluation of Image Captioning models there exists multiple metrics for the measuring of performance, in general, we can use metrics found in some NLP tasks such as machine translation thanks to the fact that, at bottom line, we are evaluating how pertinent a generated text compares to a set of annotations. On the execution of this academic work, the trained models were evaluated using the following two metrics.

4.6.1 BLEU

BLEU [57] (Bilingual Evaluation Understudy) is a metric commonly used in the domain of machine learning to evaluate the quality of text generated by models, particularly in tasks such as machine translation. It measures the similarity between a machine-generated output and one or more human reference texts by comparing n-grams (contiguous sequences of n words) within the two. The core idea behind BLEU is to quantify how much overlap exists between the n-grams of the generated text and the reference text, thus assessing the accuracy and fluency of the model's output [16].

BLEU computes scores for various n-gram sizes, such as unigrams, bigrams, trigrams, and so on, and assigns a precision-based score to the output. However, instead of focusing purely on exact word matches, BLEU also incorporates a brevity penalty to penalize outputs that are too short, thereby encouraging models to generate outputs that are not only precise but also complete [16, 57, 58]. The final BLEU score is a weighted geometric mean of the n-gram precision values, providing a single score that reflects the overall quality of the generated text in terms of both its correctness and completeness.

While BLEU is widely adopted for its simplicity and ability to provide a quantitative measure of text generation quality, it has limitations. One significant issue is that it tends to reward exact word matches, which can lead to penalizing valid paraphrases or linguistic variations [16, 58]. As a result, BLEU is often complemented by other metrics that account for semantic similarity, as it primarily focuses on surface-level lexical overlap rather than capturing the full meaning or context of the generated text. Despite this, BLEU remains a foundational metric in machine learning, especially in the evaluation of machine translation and other language generation tasks [58].

4.6.2 CIDEr

CIDEr [59] (Consensus-based Image Description Evaluation) is a metric designed for evaluating the quality of captions generated by machine learning models, particularly in image captioning tasks. Unlike simpler metrics that focus purely on word overlap, CIDEr places emphasis on capturing both syntactic and semantic similarity between the generated captions and a set of human-provided reference captions [16]. By doing so, it aims to more closely reflect human judgments of caption quality.

As Vedantam et al. [59] exposes in their work, the metric operates by comparing the n-grams (sequences of n words) in the generated caption to those in the reference captions, weighting these comparisons using a variant of term frequency-inverse document frequency (TF-IDF). This weighting helps ensure that more important and informative words contribute more to the final score, while common or less significant words are down-weighted. Additionally, CIDEr takes into account the diversity of human annotations by rewarding captions that reflect consensus among multiple human references, which helps address the natural variability in how different people describe the same image.

In academic research, CIDEr is recognized for its ability to better capture the nuances of natural language compared to earlier metrics like BLEU or ROUGE. By focusing on meaningful n-grams and considering consensus across human captions, CIDEr provides a more nuanced assessment of caption quality, balancing linguistic precision with flexibility in expression [16]. This makes it a preferred evaluation metric in tasks where capturing the richness and variability of human language is essential, particularly in image captioning and other multimodal generation tasks.

Chapter 5

Experiments and Results

As stated before, we generated our datasets using a combination of Stable Diffusion models and LLMs. For the generation of the images only for Stable Diffusion 2.1 the number of denoising steps was slightly increased from 50 to 65, all the parameters are frozen for the during the generation process and we initially discarded some of the images that resulted in a black background only. For the textual data, we used LLMs to paraphrase some of the original captions founded on Flickr30K, doing a similar approach to the ones commented by Koshla and Saini [49].

For the implementation of the model, as discussed before, we used the work of Xu et al. [36] for the implementation and used the Karpathy split during the evaluation. For the training in question, the weights of the Swim model were frozen, and an early stopping policy based on the BLEU was used, always training the model up to a max of 30 epochs. After those first steps of training, the models were trained for a couple more steps, unfreezing some of the last layers of the image encoder. All the metrics reported where obtained from the test set coming from original from Flickr30K.

5.1 Synthetic Data vs real available data

For the first experiment, we trained our models as described before so we can compare how does the model perform when completing their data with synthetic ones. Here we are assuming that we have always the same 12 thousand examples and for each model varies the origin of the images and the number of captions and their origin. The experiments results are shown on Table 4.1.

We can see that the results obtained from the training of the model is consistent through the different datasets (Table 5.1). First thing to notice is that dramatically reducing the number of captions per pair represents a considerable drop on the performance of the model. On the other hand, we can see that the synthetic images produced with Stable Diffusion 3 and

Training Data	BLEU	CIDEr
Flickr	24.6	46.7
Flickr limitcaps	16.7	24.8
SFlickr caps GPT3	25.3	47.5
SFlickr caps Llama	24.4	45.8
SFlickr image SD2.1	19.3	36.8
SFlickr image SDXL	22.4	43.1
SFlickr image SD3	24.3	45.8
SFlickr pairs SD2.1 GPT3	20.1	38.1
SFlickr pairs SDXL GPT3	22.7	42.7
SFlickr pairs SD3 GPT3	25.1	47.2
SFlickr pairs SD2.1 Llama	19.1	35.5
SFlickr pairs SDXL Llama	21.8	41.8
SFlickr pairs SD3 Llama	24.1	45.1

Table 5.1. Results of only real data against synthetically augmented sets. Flickr: Original dataset | Flickr limitcaps: Original dataset with most of their captions discarded per sample | SFlickr: Dataset augmented with synthetic data | caps: Contains synthetic captions | image: Contains synthetic images | pairs: Contains both synthetic images and captions | SD: Stable Diffusion.

XL-Base achieved nearly the same results as if training with only the original data. When exploring the images produced with Stable Diffusion 2.1 we can notice a real degradation on the quality of the images when there are faces or more complicated objects present, this is probably due to the fact that human or animal faces require a lot more details to be adequately represented, which the model is no able to produce in the amount of inference steps used during the generation.

Finally, and what is more interesting, removing some of the original captions and paraphrasing them with the GPT model achieves a higher score on BLEU and CIDEr (regardless of the images used) than when training with all the five original captions. This could relate to the observation made by Suprabhanu et al. [4] and Atliha and Šešok [44], where the diversity of the captions plays an important role on the performance of the models and, in contrast, there are many examples of simplistic or vague annotations in relation of what it's seem on the image. However, being the increase only of a few tenths, is difficult to conclude how big of an impact this could actually be.

5.2 Usage of fully synthetic data

For the second half of the experiments, we are now working with the fully synthetic pairs of images-captions. In this case, we will be training with our original pairs, but adding a certain amount of fully synthetic data, demoted as Flickr + XK SFlickr *fully*.

Training Data	BLEU	CIDEr
Flickr	24.6	46.7
Flickr + 1K SFlickr fully	24.8	47.2
Flickr + 3K SFlickr fully	25.4	48.4
Flickr + 5K SFlickr fully	25.9	49.7
Flickr + 7K SFlickr fully	27.1	55.2

Table 5.2. Performance when using different amounts of synthetic data, including model trained on only synthetic data. Flickr: Original dataset | SFlickr fully: Dataset created synthetically from scratch | NK: Number of data points taken as a sample.

As shown in Table 5.2, the addition of fully synthetic data does show a noticeable increase in the performance of the model. These results were expected, adding more good quality examples helps the models to generalize better and this aligns with what is shown in the bibliography. That's why it became of more interest to test this fully synthetic dataset in other ways, specifically, we are comparing how the models perform using only synthetic data and when only using real one.

To achieve this, a random split of the original was made data, and it was discarded 5 thousand pairs, doing this we end up with two datasets of 7 thousands of image-captions pairs. Of course, the original pairs remained with 5 captions per image and the synthetic one had between 3 to 5 captions.

As we can see in Table 5.3, our crafted synthetic achieves basically the same performance as the original one, so our methods of generating the captions and images seems capable of already imitating the same quality as real, human annotated examples. This results follows the results of Akrout et al. [10], where they were able to construct a dataset for skin diseases detection which obtain almost the same results as working with the real data. This hint us that modern generative models, at some point, will reach a point in which they will solve many of the current data issues that we find not only in image captioning but across many other tasks on the field of computer vision, an even beyond that.

Training Data	BLEU	CIDEr
Flickr reduced	16.3	23.7
SFlickr fully	16.1	23.3

Table 5.3. Performance of the fully synthetic dataset against one with real, human annotated pairs. Flickr reduced: Original dataset reduced to 7k data points | SFlickr fully: Dataset of 7K pairs of data points created synthetically from scratch.

5.3 Qualitative analysis of the pipeline

As mentioned before, some prompting engineering and reviewing of the generated content was necessary for seeking good quality of the images and captions. For the image generation the issues where more concrete, the main example being that there is a clear problem when generating faces the amount of distortion was problematic, as shown in Figure 5.1. Even though this happened with all models, this was especially noticeable for the model of Stable Diffusion 2.1.

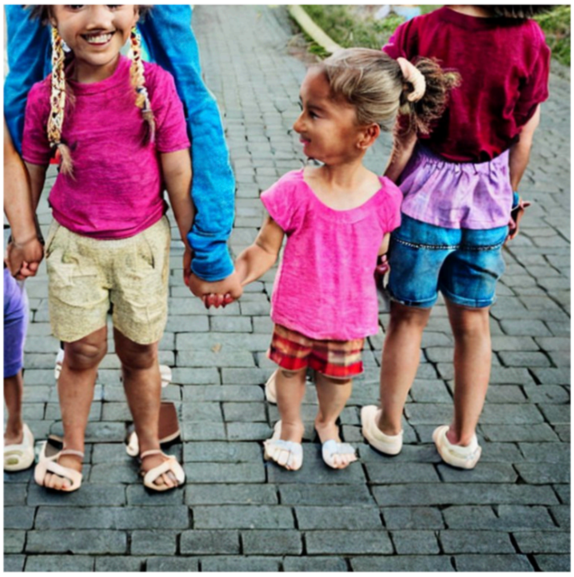


Figure 5.1. An example of distortion on faces generated with Stabel Diffusion.

In the Figure 5.2, we see that sometimes the images can be generated with a cartoonish or not realistic style when not specified on the prompt, this was a common issue found specially for Stable Diffusion XL-Base-1.0. However, in contrast to the one mentioned before, this issue was solved rather easily and consistently by specifying on the prompt the style we wanted to use during the generation.

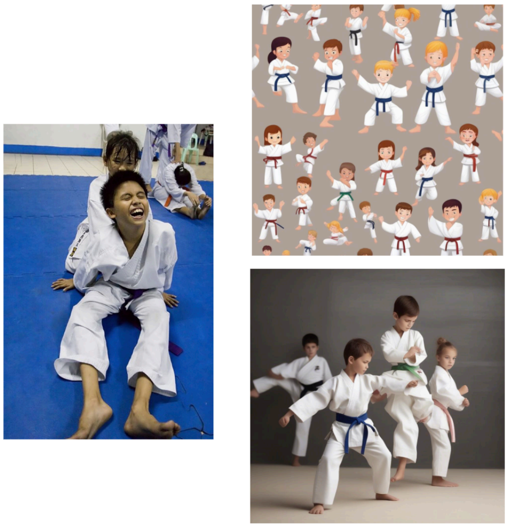


Figure 5.2. At the left we have the original image, at the upper right there is the image without asking for realism, the lower right one is the result when specifying a realistic style.

As mentioned in the previous chapter, going for too simple prompts when doing the paraphrasing of the captions for the augmented data resulted in to vague description in comparison to the original one, as shown in Figure 5.3, when asking the models to just paraphrase the caption, for the case of a black and white dog, we ended up having the dog described as a ”spotted dog”, which may be true as a matter of fact, but a dog can have spots of many different colors, so this types of losses in the original description creates inadequate captions for some of the images.

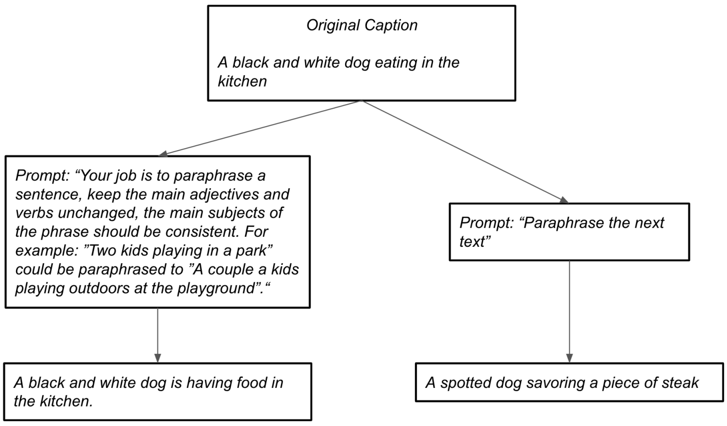

Original Caption
A black and white dog eating in the kitchen
Prompt: “Your job is to paraphrase a sentence, keep the main adjectives and verbs unchanged, the main subjects of the phrase should be consistent. For example: ”Two kids playing in a park” could be paraphrased to ”A couple a kids playing outdoors at the playground”.“
Prompt: “Paraphrase the next text”
A black and white dog is having food in the kitchen.
A spotted dog savoring a piece of steak

Figure 5.3. A comparison of the results of different prompts when generation new captions.

All of these issues mentioned during the generation process of the data is probably the biggest bottleneck, because they were manually spotted when reviewing the examples generated during the experiments, so this is an important limitation on the consideration for the usage of this pipeline on datasets of bigger dimensions.

Chapter 6

Conclusions

With the completion of this academic work, we have created a successful prototype of a pipeline for multi-modal data augmentation for the specific use of the image captioning task, allowing to a research to expand and introduce more variability to an already existing and limited dataset. On the other hand, we have explored the possibility of creating fully synthetic datasets. This has shown us to that, for a small amount of data points, synthetic data can achieve results comparable to using real one. Of course, more extensive experiments should be conducted to better understand the effect of these techniques when working on more extensive sets of data, or if the generative models have the capability of producing variable enough examples when trying to achieve amounts of data like COCO, who ranges in the hundred of thousands pairs of image and captions.

Another important aspect that limits the usage of a pipeline is the inability to measure in a quantitative way the pertinence of the augmented data is relation to the original information they come from and, in the case of the captions, how well they relate to the other synthetic data generated for the same data point. Having said this, the effectiveness of synthetic images are constrained by the capabilities of the text-to-image models utilized. For instance, if we where to try and build a dataset for tasks such as human face recognition, it is improbable to attain satisfactory results, as models like Stable Diffusion demonstrate limitations in accurately rendering human faces.

The other issue that arises is in regards to limitations on processing power and the cost of the generative models, especially Stable diffusion, as creating just one of these datasets takes days or even weeks when only working with one machine with standard hardware specs.

Having said this, we find that it could be opportune to work on the next proposals after concluding this work:

- Explore the possibility of creating a robust way of assessing the quality of the generated data. A generalized multimodal data quality assessment is an open line of work that should contribute greatly in the future.

- Study the effects of working generating bigger and more diverse datasets, so it can be properly measured the current applications of these techniques for a system tackling a real case.

Bibliography

[1] Changrong Xiao, Sean Xin Xu, and Kunpeng Zhang. Multimodal data augmentation for image captioning using diffusion models. In *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, LGM3A '23, page 23–33, New York, NY, USA, 2023. Association for Computing Machinery.

[2] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017.

[3] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[4] Sai Suprabhanu Nallapaneni and Subrahmanyam Konakanchi. A comprehensive analysis of real-world image captioning and scene identification, 2023.

[5] Sulabh Katiyar and Samir Kumar Borgohain. Image captioning using deep stacked lstms, contextual word embeddings and data augmentation. *CoRR*, abs/2102.11237, 2021.

[6] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. *CoRR*, abs/1604.00790, 2016.

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

[8] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. Learning to evaluate image captioning. *CoRR*, abs/1806.06422, 2018.

[9] Ingrid Ravn Turkerud and Ole Jakob Mengshoel. Image captioning using deep learning: Text augmentation by paraphrasing via backtranslation. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–10, 2021.

[10] Mohamed Akrout, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, Máté Kovács, and István Fazekas. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images, 2023.

[11] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning, 2018.

[12] Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. Image captioning: A comprehensive survey. In *2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)*, pages 325–328, 2020.

[13] Hiba Ahsan, Nikita Bhalla, Daivat Bhatt, and Kaivankumar Shah. Multi-modal image captioning for the visually impaired. *CoRR*, abs/2105.08106, 2021.

[14] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[15] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019.

[16] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

[17] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey, 2023.

[18] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2023.

[19] Tianyi Zhang, Zheng Wang, Jing Huang, Mohiuddin Muhammad Tasnim, and Wei Shi. A survey of diffusion based image generation models: Issues and their solutions, 2023.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[21] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), nov 2023.

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

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[24] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 36:2814–2830, 2022.

[25] Onkar Mishra. Stable Diffusion Explained — onkarmishra. `https://medium.com/@onkarmishra/stable-diffusion-explained-1f101284484d`, 2023. [Accessed 16-09-2024].

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[27] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.

[28] Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. A survey on data augmentation in large model era, 2024.

[29] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

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

[31] Rahib Imamguluyev. The rise of gpt-3: Implications for natural language processing and beyond. *International Journal of Research Publication and Reviews*, 4:4893–4903, 03 2023.

[32] Luciano Floridi and Massimo Chiriatti. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4):681–694, December 2020.

[33] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874, 2024.

[34] Meta. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`, 2024. [Accessed 16-09-2024].

[35] Abhimanyu Dubey et al. The llama 3 herd of models, 2024.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[37] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections, 2022.

[38] Jiuxiang Gu, Shafiq R. Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. *CoRR*, abs/1803.05526, 2018.

[39] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. *CoRR*, abs/1908.09317, 2019.

[41] Peipei Zhu, Xiao Wang, Yong Luo, Zhenglong Sun, Wei-Shi Zheng, Yaowei Wang, and Changwen Chen. Unpaired image captioning by image-level weakly-supervised visual concept recognition, 2022.

[42] Wenhu Chen, Aurélien Lucchi, and Thomas Hofmann. Bootstrap, review, decode: Using out-of-domain textual data to improve image captioning. *CoRR*, abs/1611.05321, 2016.

[43] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. *CoRR*, abs/1909.02201, 2019.

[44] Viktar Atliha and Dmitrij Šešok. Text augmentation using bert for image captioning. *Applied Sciences*, 10(17), 2020.

[45] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019.

[46] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[49] Cherry Khosla and Baljit Singh Saini. Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pages 79–85, 2020.

[50] Hang Li, Jindong Gu, Rajat Koner, Sahand Sharifzadeh, and Volker Tresp. Do dall-e and flamingo understand each other?, 2023.

[51] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[52] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event

descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[53] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015.

[54] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions, 2015.

[55] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

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

[57] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[58] Ehud Reiter. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401, 09 2018.

[59] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014.