A VISUAL-BASED METHOD FOR TEACHING CHILDREN USING TEXT TO IMAGE AND IMAGE TO TEXT TRANSLATIONS

Professional Practice/Seminar (IT890)

MASTER OF TECHNOLOGY

in INFORMATION TECHNOLOGY

by

N Sanjana Shree (222IT021)

under the guidance of

Prof. Ram Mohana Reddy Guddeti

DEPARTMENT OF INFORMATION TECHNOLOGY
NATIONAL INSTITUTE OFTECHNOLOGY
KARNATAKA SURATHKAL
MANGALORE - 575025

APRIL, 2023

DECLARATION

I hereby declare that the Professional Practice/Seminar (IT890) Work Report of the MTech (IT) entitled "A Visual-based Method For Teaching Children Using Text to Image and Image to Text Translations" which is being submitted to the National Institute of Technology Karnataka Surathkal, in partial fulfillment of the requirements for the award of the Degree of Master of Technology in the Department of Information Technology, is a bonafide report of the work carried out by me. The material contained in this professional practice/seminar report has not been submitted to any University or Institution for the award of any degree.

N Sanjana Shree
222IT021

Department of Information Technology

Place: NITK Surathkal
Date: 20-04-2023

CERTIFICATE

This is to certify that the Professional Practice/Seminar (IT890) Work Report Entitled "A Visual-based Method For Teaching Children That Uses Text To Image and Image To Text Translations" submitted by N Sanjana Shree (222IT021) as the record of the work carried out by her under my guidance, is accepted as the Professional Practice/Seminar (IT890) Work Report submission in partial fulfillment of the requirements for the award of the degree of Master of Technology in the Department of Information Technology.

Prof. Ram Mohana Reddy Guddeti
Dept. of Information Technology
NITK Surathkal, Mangalore

ACKNOWLEDGEMENT

I would sincerely like to thank Prof. Ram Mohana Reddy Guddeti (Department of Information Technology) for the unconditional and unbiased support during the whole session of study and development and for guiding me throughout the whole semester. He provided me with a favorable environment, without them, I would not have achieved my goal. They had always been there for me despite their busy schedule and were always a great source of inspiration. They had been easily approachable during and even after college hours. I sincerely thank them for that.

A blend of gratitude, pleasure, and great satisfaction is what I feel to convey my indebtedness to all those who have directly and indirectly contributed to the successful completion of the project.

N Sanjana Shree
222IT021

Department of Information Technology

ABSTRACT

For children to learn alphabets first step is to learn letters. Based on the letter given by the user a set of images is displayed. Image captioning and text-to-image translation are two related tasks in the field of computer vision and natural language processing. Image captioning involves generating textual descriptions of images automatically, using machine learning techniques. This task has many practical applications, including helping visually impaired individuals understand the content of images, and assisting in content-based image retrieval. Text-to-image translation, on the other hand, involves generating images from textual descriptions. This task is particularly challenging due to the complexity of mapping natural language descriptions to visual features in away that is both semantically meaningful and visually realistic. Both tasks require the use of deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), and the integration of multiple modalities of information, such as visual and textual data. Advances in these areas have led to the development of sophisticated models for image captioning and text-to-image translation, which are now being used in a wide range of practical applications, including e-commerce, digital art, and virtual reality environments.

Keywords: Deep Learning, Computer vision, Convolutional Neural Networks (CNN)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Image to text and text to image translation are two interrelated fields in the domain of computer vision and natural language processing. Image to text translation involves the conversion of visual information present in an image into meaningful textual representations. This task is typically achieved using deep learning-based models that analyze the visual features of the image and map them to corresponding words orsentences. On the other hand, text to image translation is the process of generating imagesfrom textual descriptions. This task is generally accomplished using generative models thatlearn to synthesize images that match the given textual input. Both of these tasks have a wide range of applications, including image captioning, scene understanding, and contentcreation, among others.

1.1 Text-to-Image Translation

Text-to-image translation involves converting written language into visual representations, such as pictures, diagrams, or other types of graphics. This process can be helpful for children who struggle with reading or have difficulty understanding written language. By creating images that represent the meaning of written words or sentences, children can develop a deeper understanding of language and improve their comprehension skills.

One way to introduce text-to-image translation to children is to have them create their own illustrations for stories or poems. For example, a parent or teacher could read a story to a group of children and then ask them to draw pictures of the characters orscenes from the story. This activity can help children visualize the story and better understand its meaning. Another way to introduce text-to-image translation is to have children create their own visual aids for presentations or reports. For example, if a child is giving a report on a particular animal, they could create a poster or a slideshow with pictures of the animal, along with captions or descriptions that explain its characteristics or habitat. This activity can help children develop their research and writing skills, as well as their ability to translate between written and visual modes of communication.

1.2 Image-to-Text Translation

Image-to-text translation involves converting visual images into written language. This process can be helpful for children who are learning to write or who have difficulty expressing themselves in words. By looking at an image and describing what they see in words, children can practice using descriptive language and develop their vocabulary.

One way to introduce image-to-text translation to children is to have descriptions of pictures or scenes. For example, a parent or teacher could show a group of children a picture and ask them to write a paragraph or a few sentences describing what they see in the picture. This activity can help children develop their writing skills and improve their ability to express themselves in words.

Another way to introduce image-to-text translation is to have children create their own stories or poems based on pictures or other visual media. For example, a child could be shown a picture of a castle and asked to write a story about a princess who lives there. This activity can help children develop their creativity and imagination, as well as their ability to translate between visual and written modes of communication.

1.3 Motivation

Text-to-image and image-to-text translation are valuable tools for teaching children how to communicate effectively using both words and images. By introducing these concepts to children at a young age, parents and educators can help them develop essential language and communication skills that will serve them well throughout their lives. In addition to the activities mentioned above, there are many other ways to incorporate text-to-image and image-to-text translation into children's learning experiences. For example, children could be asked to create their own comic strips or storyboards, or to create their own visual diagrams or flowcharts to explain complex concepts. Overall, text- to-image and image-to-text translation are powerful tools for helping children develop their language and communication skills, and can be an engaging and fun way for children to learn about language and express their creativity.

Learning English and Kannada alphabets along with images can be a great way to help small children understand and memorize the letters. Images can make the learning process more engaging and help children associate each letter with a visual representation.

One way to incorporate images into the learning process is through the use of cards. Another way to use images in learning alphabets is through books. By making the learning process more engaging and fun, children are more likely to develop an interest in alphabets and enjoy the learning process. Small children can benefit from visual materials like pictures and essays because they can use them to improve their language, creativity, and imagination. It is quite effective to include young children in learning through visual pictures. Young children frequently learn best through images and other visual media since they are visual learners. Parents and teachers may improve their students' understanding and retention of key ideas by using visual aids in their teachings. Making visual aids like posters, diagrams, or charts is one approach to include visual pictures into your lessons. A parent or teacher may design a poster that illustrates the butterfly's life cycle or a diagram that breaks down a plant's component elements. Children's comprehension of the natural world may be expanded via this exercise, which can also help them become better at learning through visual media. Encourage kids to produce their own artwork or illustrations as another approach to employ visuals in the classroom. For instance, a parent or teacher could instruct kids to sketch their favorite animal or make a collage that illustrates a certain idea. Children's creativity can grow as a result of this practice.

Essays may be a useful educational tool for showing young kids how to communicate in writing. Parents and educators may aid children in honing their writing abilities, expanding their vocabulary, and learning how to organize their thoughts by encouraging them to write essays on a range of subjects. Having young children write little paragraphs or stories about subjects that interest them is one method to introduce essays to them. A youngster who enjoys playing outside can write a tale about a fantastic adventure they had in nature, or an animal lover might write a brief paragraph about their favorite animal. Children's imagination and creativity, as well as their capacity for written expression, can all be developed via this exercise.

Having young kids write about their experiences is another method to expose them to essays. A youngster may, for instance write an essay on a vacation they went with their family or about a noteworthy occasion or celebration they took part in. Children's abilities to observe, describe, and write in a narrative manner can all be improved via this practice. Text-to-image and image-to-text translation are two concepts that can help children develop their language and communication skills. In today's digital age, children are exposed to a variety of forms of communication, including written text and visual media. By teaching children how to translate between these two modes of communication, parents and educators can help them become more effective communicators and develop a deeper understanding of language.

2. LITERATURE SURVEY

Over the years, researchers have proposed various methods to tackle these tasks, ranging from rule-based approaches to deep learning-based models. Recently, deep learning-based models have shown promising results in both image captioning and text- to-image translation. In this literature survey, we will explore the various approaches proposed in the literature for image captioning and text-to-image translation. We will examine the strengths and weaknesses of each approach and discuss the future directions in this field.

2.1 Background

Image captioning and text-to-image translation are two challenging tasks in the field of computer vision and natural language processing. Image captioning involves generating a natural language description of an image, while text-to-image translation involves generating a visual representation from a textual description. These tasks have several practical applications, including image retrieval, content-based image search, and image summarization. Moreover, text-to-image translation can be used to create realistic images from textual descriptions, which has many potential applications in the fields of art, design, and entertainment.

2.2 Outcome of Literature Survey

The following listed papers listed are related to image captioning, text-to-image synthesis, and generative adversarial networks. Here is a brief summary of each paper:

In [1], authors proposed a deep learning model for generating captions for images based on the visual features of the images. In [2], authors presented a review of various deep learning models for image captioning, including both encoder-decoder and attention-based approaches. In [3], authors proposed a text-to- image synthesis approach to generate realistic images that can be used to improve image captioning. In [4], authors proposed a text-to-image synthesis method that uses a multiattention depth residual generation adversarial network (MADRGAN) to generate images.

In [5], authors proposed a novel deep neural network architecture called ON-AFN (Ordered-attentive Fusion Network) for generating image captions. The proposed architecture combines residual attention and ordered memory module to improve the performance of image caption generation. The residual attention mechanism helps the model to focus on the most relevant parts of the image, while the ordered memory module

preserves the temporal order of the generated words.. In [6], authors proposed an attention-based LSTM model for image captioning, which used both visual attention and textual attention mechanisms. In [7], authors proposed a cycle-consistent adversarial network (CycleGAN) for unpaired image-to-image translation, which could be used for tasks such as style transfer and image synthesis.

In [8] , authors proposed a text-to-image synthesis method which used a residual generative adversarial network (RGAN), which generated high- resolution images from textual descriptions. In [9], authors proposed a bidirectional generative adversarial network (BiGAN) for text-to-image synthesis. The BiGAN consists of a generator and a discriminator network that work in a bi-directional manner to generate realistic images from textual descriptions. The generator network takes as input the textual description and generates a low-resolution image, which is then refined by a second generator network to produce a high-resolution image. The discriminator network evaluates the realism of the generated images and provides feedback to the generator networks.

In [10], authors proposed a deep architecture for image caption generation. The proposed architecture consists of an encoder-decoder network with attention mechanism. The encoder network processes the input image and extracts a feature vector that is used by the decoder network to generate the caption. The attention mechanism allows the model to focus on different parts of the image during the caption generation process.

In summary, literature survey on image captioning and text-to-image translation has revealed that both tasks are active research areas in computer vision and natural language processing. Image captioning typically involves using an encoder-decoder framework with attention mechanisms to generate natural language descriptions for images.

Recent approaches have focused on using pre-trained models and reinforcement learning to improve the quality of generated captions. On the other hand, text-to-image translation typically uses a generative adversarial network (GAN) to generate images from textual descriptions. Recent research has explored using pre-trained language models and visual attention mechanisms to improve the quality of the generated images. However, there are still significant challenges in generating both high-quality captions and realistic images that are semantically coherent with the given textual descriptions, leaving room for further research and development in these fields. Table 2.1 Summarizes the salient features of some key existing works.

2.1 Summary of Literature Survey

SNO.	Paper Name	Authors	Methodology
1	Image Captioning Using Deep Learning	C.S.Kanimozhiselvi et al. [1]	The methodology proposed in the IEEE research paper is a standard approach to image captioning using deep learning. It involved collecting a large dataset, preprocessing the data, designing a neural network architecture, training the model, evaluating its performance, and generating captions for new images.
2	Facilitated Deep Learning Models for Image Captioning	Imtinan Azhar et al. [2]	This approach utilized the encoder-decoder framework, where the encoder network extracted image features, and the decoder network generated captions. The authors proposed a novel mechanism that combined attention and gating mechanisms to improved the performance of the decoder network.
3	Text to Image Synthesis for Improved Image Captioning	Md. Zakir Hossain et al. [3]	This method aimed to improve the image captioning task by generating realistic images from textual descriptions using a conditional generative adversarial network (cGAN). The generated images are then used to train an image captioning model.
4	A Text-to-Image Generation Method Based on Multiattention	Shuo Yang, Xiaojun Bi et al. [4]	This approach combined the advantages of attention mechanisms and residual networks to generate high-quality images from textual descriptions. The proposed model consists of a generator network that generates images and a discriminator network that discriminates between real and generated images.
5	ON-AFN: Generating Image Caption based on the Fusion of Residual Attention and Ordered Memory Module	C. Lu et al. [5]	They proposed a novel deep neural network architecture called ON-AFN (Ordered-attentive Fusion Network) for generating image captions.

6	Unpaired Image-to-Image Translation usingCycle- Consistent Adversarial Networks	Jun-Yan et al. [6]	This approach utilized a cycle-consistent adversarial network (CycleGAN) to perform unpaired image-to-image translation between two domains. The generated images are used to train an image captioning model.
7	Text to Image Synthesis using Residual GAN	Priyanka, Mishra, et al. [7]	This method used a residual GAN to generate high-quality images from textual descriptions. The authors proposed a residual learning approach to improve the quality of the generated images.
8	StackGAN++:Realistic Image Synthesis withStacked Generative Adversarial Networks	Han, Zhang et al. [8]	This model utilized a stacked GAN to generate high-resolution and realistic images from textual descriptions. The authors proposed a two-stage training process that generated low-resolution images in the first stage and high-resolution images in the second stage.
9	Text to Image Synthesis With Bidirectional Generative Adversarial Network	Z. Wang et al. [9]	They proposed a bidirectional generative adversarial network (BiGAN) for text-to-image synthesis. The BiGAN consisted of a generator and a discriminator network that work in a bi-directional manner to generated realistic images from textual descriptions.
10	Image Caption Generation Using A Deep Architecture	A. Hani et al. [10]	They proposed a deep architecture for image caption generation. The proposed architecture consists of an encoder-decoder network with attention mechanism. The encoder network processes the input image and extracts a feature vector that is used by the decoder network to generate the caption. The attention mechanism allows the model to focus on different parts of the image during the caption generation process.

2.3 Problem Statement

This project focus on teaching the young children alphabets, sentences and visual analysis of a particular image using deep learning models such as stable diffusion model and Encoder-Decoder model.

2.4 Objectives

Study the different approaches for text to image and image to text translation to teach the young children: Text-to-image and image-to-text translation are advanced topics that may not be suitable for young children to learn directly. However, some basic concepts related to these topics can be introduced to children in a simplified way, depending on their age and cognitive development level.

To implement text to image translation to teach young children: Text to image translation involves using natural language processing and computer vision to generate images based on text descriptions. This technology can be used to create visual aids for teaching young children by providing them with engaging, interactive, and informative visuals.

To implement image to text translation to teach young children: Image to text translation involves using computer vision and natural language processing to generate text descriptions based on images. This technology can be used to create informative and engaging captions for images, which can be helpful in teaching young children.

3. METHODOLOGY

Text-to-image and image-to-text translation are complex tasks that require collecting large and diverse datasets, preprocessing the data, and training deep learning models such as GANs. For text-to-image translation, the text is converted into a vector space representation, and the GAN generates an image that matches the input text. For image-to-text translation, the image is fed into a convolutional neural network to extract visual features, which are then used to generate a textual description using a recurrent neural network or transformer-based model. These methodologies require careful tuning and evaluation to produce high-quality results.

3.1 System Modules and Components

3.1.1 Learn Kannada and English alphabets with their words

English Alphabets: For each English alphabet there is an English word starting with that alphabet which is mapped. The program asks for console input to give an alphabet. The words starting with that alphabet is scraped from google and displayed on the screen.

3.1.2 Image-to-Text Translation

Kannada Alphabets: For each kannada alphabet there is a kannada word starting with that alphabet which is mapped. The program asks for console input to give an alphabet. The words starting with that alphabet is scraped from google and displayed on the screen.

3.1.3 Text-to-Image Translation

Diffusion Models have recently showed a remarkable performance in Image Generation tasks and have superseded the performance of GANs on several tasks such as Image Synthesis. These models have also been able to produce more diverse images and proved to not suffer from Mode Collapse. This is due to the ability of the *Diffusion Models* to preserve the semantic structure of the data.

3.2 Tools for Building Models

3.2.1 Stable Diffusion Model

These models are highly computationally demanding, and training requires a very large memory and carbon footprint which makes it impossible for most researchers to even attempt the method. This is due the fact that all Markovian states need to be in memory for prediction all the time which means multiple instances of large Deep-Nets being present inmemory all the time. Furthermore, training time for such methods also becomes too high (e.g., days to months) because these models tend to get stuck in the fine-grained

imperceptible intricacies in the image data. However, it is to be noted that thisfine-grained image generation is also one of the main strengths of *Diffusion Models* so, it isa kind of paradoxical to use them.

A text-to-image model using deep learning called Stable Diffusion was launched in 2022. Although it may be used for various tasks including inpainting, out painting, and creating image-to-image translations directed by text prompts, its primary usage is to produce detailed visuals conditioned on text descriptions. In cooperation with a variety of academic researchers and nonprofit groups, the start-up Stability AI created it. Stable Diffusion Models is just a rebranding of the LDMs with application to high resolution images while using *CLIP* as text encoder. The paradise cosmic beach is taken as input. Then, by diffusing information over the grid via a method known as heat diffusion, the image is turned into a higher-dimensional representation.

The image gets blurrier and less detailed as the diffusion process goes on, but the high-dimensional representation that is produced captures key aspects of the image in a form that can be used to subsequent tasks like classification or grouping. The visualization displays the final high-dimensional representation as a sequence of colored dots, where each colour denotes a distinct group of pixels with a comparable set of properties. Fig 3.1 shows working of Stable Diffusion model. Fig 3.2 shows Images of Stable Diffusion model. Fig 3.3 shows Seed value output.

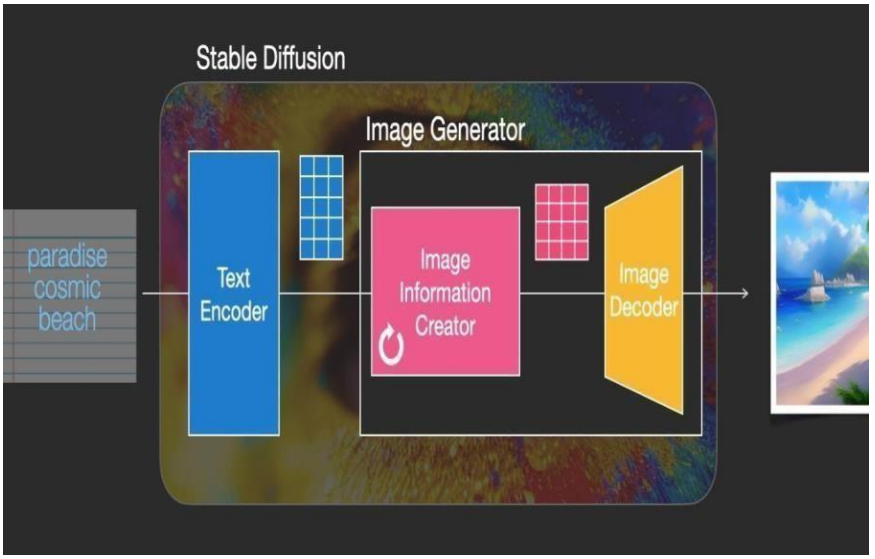

Figure 3.1: Working of Stable Diffusion model

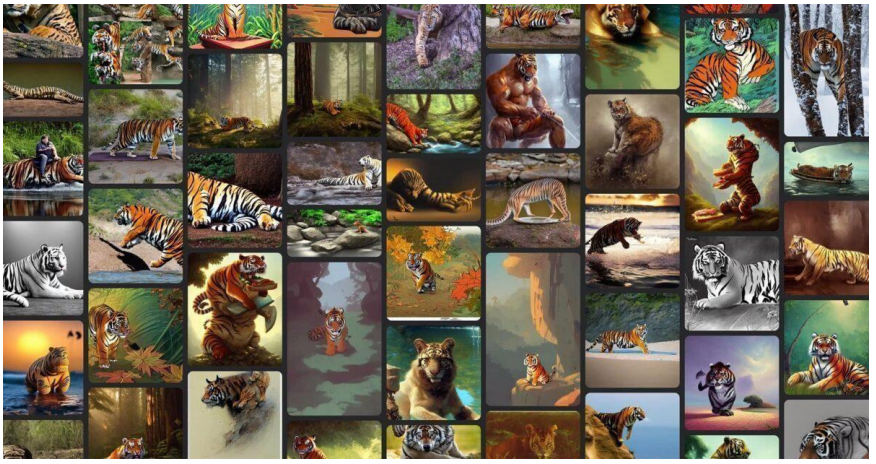

Figure 3.2: Images of Stable Diffusion model

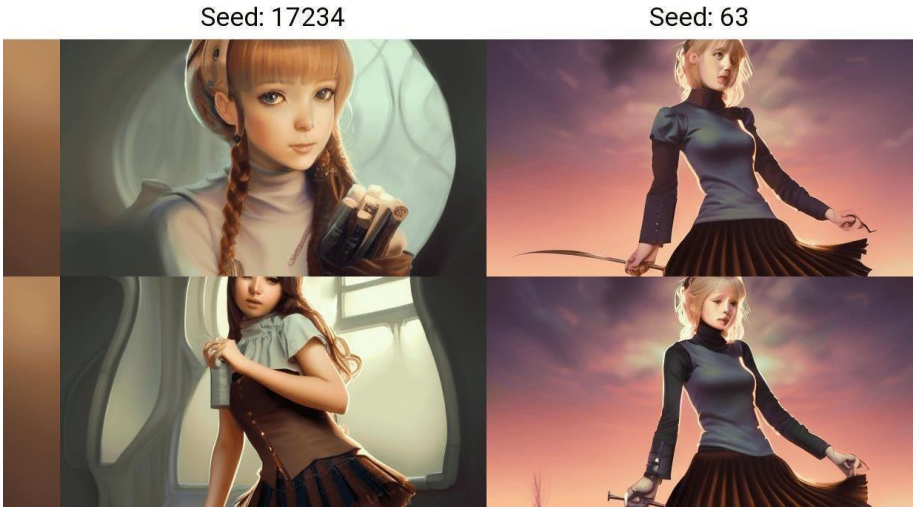

Figure 3.3: Seed value output

The Stable Diffusion "txt2img" text to image sampling script takes a text prompt as well as many option parameters for sample kinds, output picture dimensions, and seed values. In accordance with the model's interpretation of the prompt, the script generates an image file. Invisible digital watermarks are placed on created images to help viewers recognize them as being from Stable Diffusion, however these watermarks lose their effectiveness when the image is scaled down or rotated.

Each txt2img generation will use a unique seed value that will have an impact on the final picture. Users can choose to use a new seed to experiment with various created outputs or stick with the same seed to get the same picture output as a prior image generation.

The number of inference steps for the sampler may also be changed by the user; a larger value requires more time, whilst a lesser value could lead to aesthetic flaws. The user may also control how closely the output image resembles the prompt by adjusting the classifier-free guidance scale value. While use cases looking for more specified outputs may choose to use a higher number, more exploratory use cases may choose to use a lower scale value.

Another sample script called "img2img" that takes a text prompt, the path to an existent picture, and a strength value from 0 and 1 is also included in Stable Diffusion. The script generates a new picture using the initial picture that includes the text prompt's specified parts as well. The output image's extra noise is indicated by the strength value. More variance within the image is produced with a greater strength value, but it may also result in an image which is not conceptually compatible with the given challenge.

3.2.2 Encoder-Decoder Model

The process of creating a natural language description of a picture is known as image captioning. An encoder network to extract the image's features and a encoder- decoder network to create the caption are the two primary parts of the task. Fig 3.4 shows working of Encoder-Decoder.

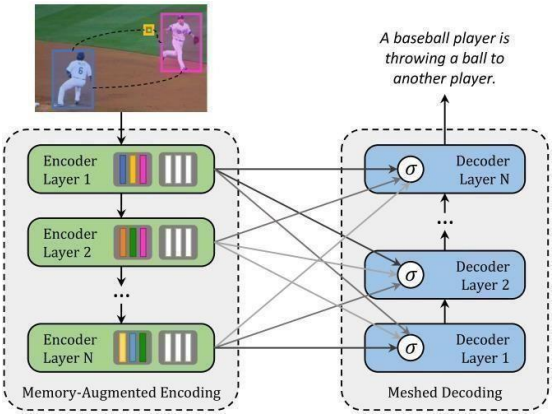

Figure 3.4: Encoder-Decoder working

Figure 3.5: Captioning pixelwise

Fig 3.5 shows the Image captioning pixel-wise. Using a feedforward neural network is one method of implementing the encoder decoder network. The feedforward network creates the caption one word at a time using the input of the encoded picture attributes. The feedforward network predicts the probability distribution across the full vocabulary at each time step and samples the following word based on this distribution.

1. The embedded preceding word and the encoded picture characteristics are sent into the decoder network at each time step. This input is processed by the feedforward network, which then generates the probability distribution across the vocabulary from which the following word is sampled.
2. Here is a more detailed explanation of the decoder network for image captioning using a feedforward neural network.
3. Embedding Layer: The input to the decoder network is the previously generated word, which is first converted into a fixed-dimensional vector using an embedding layer. The embedding layer maps each word in the vocabulary to a dense vector of a specified size.
4. Concatenation with Image Features: The embedded word vector is then concatenated with the encoded image features to form a combined input vector.

The image features are usually extracted using a convolutional neural network (CNN) and can be thought of as a fixed-length vector representation of the image.

5. Feedforward Layers: The combined input vector is then passed through one or more feedforward layers, each consisting of a dense layer with a non-linear activation function, such as ReLU. The output of each dense layer is fed as input to the next dense layer in the network. Dropout layers can also be included between the dense layers to regularize the model and prevent overfitting.

6. Output Layer: The final feedforward layer in the network is a dense layer with softmax activation that outputs the probability distribution over the entire vocabulary. Each element in the output vector represents the probability of generating a particular word in the vocabulary given the input image and the previously generated words in the caption.

7. Sampling: To generate the next word in the caption, the word with the highest probability is sampled from the output probability distribution. This word is then embedded and concatenated with the image features to produce the input for the next time step. This process is repeated until an end-of-sentence token is generated, indicating the completion of the caption.

The loss function for the model is typically cross-entropy, which measures the difference between the predicted probability distribution and the true probability distribution of the target word. The model is trained using backpropagation and stochastic gradient descent (SGD) to minimize the loss.

In summary, the decoder network for image captioning using a feedforward neuralnetwork is a sequence of layers that takes as input the encoded image features and the previously generated word, and generates the next word in the caption by predicting the probability distribution over the vocabulary.

3.3 Implementation Diagram

The implementation of my project consists of Learning alphabets and words where the child gives a letter and the corresponding words and images are obtained. Study different deep learning models. Used stable diffusion model for text to image translation and encoder decoder for image to text translation. Fig 3.6 shows Implementation Diagram.

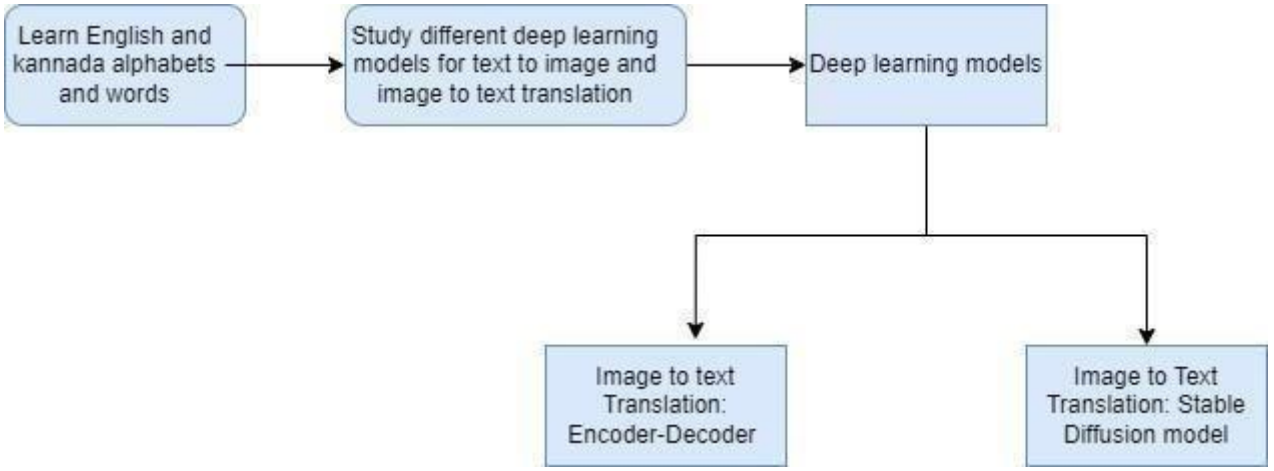

Figure 3.6: Implementation diagram

4. RESULTS AND DISCUSSION

4.1 Learning Kannada and English alphabets

Figure 4.1: English Alphabets with words

Figure 4.2: Kannada Alphabets with words

This section includes understanding English and Kannada alphabets with their words. The user gave the input alphabet as “D”. Dog was the word mapped to that alphabet. So the dog images are displayed along with the google images link of dog images for better understanding of the children. Figure 4.1 and 4.2 show the English and Kannada letters/words, respectively.

4.2 Text-to-Image Translation

Figure 4.3: Text to image translation

Machine Learning was the invention of GANs (Generative Adversarial Networks) a method that introduced the possibility to think beyond what was already present in the data, a steppingstone for a whole new field which is now called, Generative Modelling. However, after going through a booming phase, GANs started to face a plateau where most of the methods were struggling to solve some of the bottlenecks faced by the adversarial methods. It is not the problem with the individual methods but the adversarial nature of the problem itself. Some of the major bottlenecks of the GANs are:

- Lack of diversity in Image Generation
- Mode Collapse
- Problem learning Multimodal distribution
- High Training Time
- Not Easy to Train due to the Adversarial Nature of the problem formulation

Due to these disadvantages of GANs stable diffusion models can be used. The above figure asks for prompt, height, width of the image. On the basis of entry values it generates an images. Figure 4.3 shows text-to-image translation.

4.3 Image-to-Text Translation

Figure 4.4: Image input for Image captioning

```
100/100 [==============================] - 77s 766ms/step - loss: 2.7406 - masked_acc: 0.4176 - val_loss: 2.8421 - val_masked_acc: 0.3956
Epoch 14/100
100/100 [==============================] - ETA: 0s - loss: 2.7465 - masked_acc: 0.4196

a man in a red wetsuit is surfing
a man wearing a yellow kayak is surfing
three opposing light men swinging on wave

100/100 [==============================] - 72s 722ms/step - loss: 2.7465 - masked_acc: 0.4196 - val_loss: 2.9335 - val_masked_acc: 0.3916
Epoch 15/100
100/100 [==============================] - ETA: 0s - loss: 2.7667 - masked_acc: 0.4168

a man in a red wetsuit is surfing
a man in a wetsuit is surfing
a surfer rides a surfboard dressed in a wave

100/100 [==============================] - 75s 748ms/step - loss: 2.7667 - masked_acc: 0.4168 - val_loss: 2.9226 - val_masked_acc: 0.3947
Epoch 16/100
100/100 [==============================] - ETA: 0s - loss: 2.7232 - masked_acc: 0.4230

a man in a red wetsuit is surfing
a surfer rides a wave
a young skateboard splashes through the surf

100/100 [==============================] - 76s 766ms/step - loss: 2.7232 - masked_acc: 0.4230 - val_loss: 2.9111 - val_masked_acc: 0.3919
Epoch 17/100
100/100 [==============================] - ETA: 0s - loss: 2.7256 - masked_acc: 0.4237

a man in a red wetsuit is surfing
a surfer in a red wetsuit is surfing
a man in the swimming trunks of a surfboard

100/100 [==============================] - 73s 730ms/step - loss: 2.7256 - masked_acc: 0.4237 - val_loss: 2.8854 - val_masked_acc: 0.3976
Epoch 18/100
100/100 [==============================] - ETA: 0s - loss: 2.7326 - masked_acc: 0.4196

a man in a red wetsuit is surfing on a wave
a surfer rides a wave
a surfer slides through the air inside a swimming ocean

100/100 [==============================] - 73s 729ms/step - loss: 2.7326 - masked_acc: 0.4196 - val_loss: 2.8786 - val_masked_acc: 0.3938
```

Figure 4.5: Captions for image in each epoch

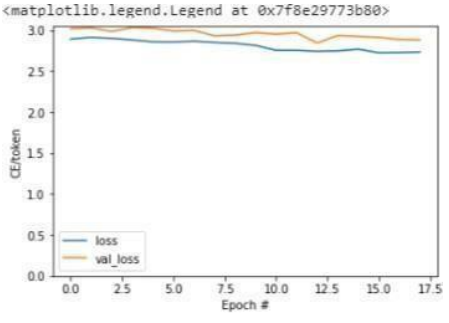

Figure 4.6: Loss Graph

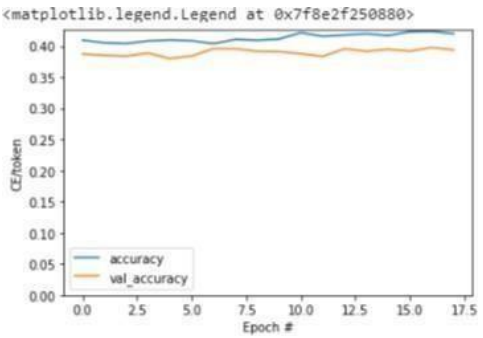

Figure 4.7: Accuracy graph

The URL of the image is given. The model learns for 100 epochs. After 100 epochs it gives the description of the image. Figure 4.4 gives the image which is given as input. Figure 4.5 tells the epoch wise training results. Finally the model gives the caption for the image.The validation and accuracy results are shown above. Figure 4.6 shows captioning in each epoch. Figure 4.6 shows loss graph. Figure 4.7 shows Accuracy graph.

5 CONCLUSIONS AND FUTURE WORK

5.1 Conclusion

Image captioning and text-to-image translation are two exciting and rapidly developing areas in the field of computer vision and natural language processing. Image captioning involves generating natural language descriptions of images, allowing machines to understand and describe visual content like humans. Meanwhile, text-to- image translation aims to generate realistic images from textual descriptions, which can have applications in areas such as virtual reality, gaming, and content creation. In recent years, significant progress has been made in both fields, with the development of deep learning techniques such as convolutional neural networks (CNNs) and generative adversarial networks (GANs) proving particularly effective. However, there are still many challenges to be addressed in these areas. For image captioning, there is a need to develop models that can generate more accurate and diverse captions, as well as handle complex scenes with multiple objects and actions. In text-to-image translation, generating high- resolution images that are both realistic and faithful to the textual input remains a challenge.

5.2 Future Work

Future work in image captioning and text-to-image translation will likely involve the integration of other modalities such as audio and video, as well as the development of more interpretable models that can provide insights into how they generate their output. Additionally, there is a need for research into ethical considerations around these technologies, such as issues of bias and privacy. Overall, these areas hold great potential for advancing our understanding of how machines can understand and generate visual content, and we can expect to see continued progress in the years to come.

REFERENCES

[1] C. S. Kanimozhiselvi, K. V, K. S. P and K. S, "Image Captioning Using Deep Learning,"2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-7

[2] I. Azhar, I. Afyouni and A. Elnagar, "Facilitated Deep Learning Models for Image Captioning," *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, 2021

[3] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in *IEEE Access*, vol. 9, pp. 64918-64928, 2021

[4] S. Yang, X. Bi, J. Xiao and J. Xia, "A Text-to-Image Generation Method Based on Multiattention Depth Residual Generation Adversarial Network," *2021 7th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2021, pp. 1817-1821

[5] C. Lu, Z. Qu, X. Wang and H. Zhang, "ON-AFN: Generating Image Caption based onthe Fusion of Residual Attention and Ordered Memory Module," *2021 IEEE Asia- Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Dalian,China, 2021, pp. 1328-1334

[6] Zhu, Jun-Yan & Park, Taesung & Isola, Phillip & Efros, Alexei. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2242-2251. 10.1109/ICCV.2017.244.

[7] Mishra, Priyanka & Rathore, Tribhuvan & Shivani, Shivani & Tendulkar, Sachin. (2020). Text to Image Synthesis using Residual GAN. 139-144. 10.1109/ICETCE48199.2020.9091779.

[8] Zhang, Han & Xu, Tao & Li, Hongsheng & Zhang, Shaoting & Wang, Xiaogang & Huang, Xiaolei & Metaxas, Dimitris. (2017). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. IEEE Transactions of Pattern Analysis and Machine Intelligence. PP. 10.1109/TPAMI.2018.2856256.

[9] Z. Wang, Z. Quan, Z. -J. Wang, X. Hu and Y. Chen, "Text to Image Synthesis With Bidirectional Generative Adversarial Network," *2020 IEEE International Conference on Multimedia and Expo (ICME)*, London, UK, 2020, pp. 1-6

[10] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," *2019 International Arab Conference on Information Technology (ACIT)*, Al Ain, United Arab Emirates, 2019, pp. 246-251