UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

COS700 Research Proposal

Using AI to generate descriptions for data structure images for visually impaired persons

Student number: u20583703

Supervisor(s):
Dr. Linda Marshall
Mr Werner Hauger

Jano Esterhuizen

October 2024

Abstract

The goal of the present research is to create an AI-based solution for generating informative, accurate, and understandable descriptions for images and diagrams used in computer science, targeted to visually impaired students. Although the field of assistive technologies has been rapidly advancing, visually impaired students remain unable to access computer science content easily. Considering that computer science programs in schools often use images and diagrams to present information, it becomes imperative to enhance the accessibility of such materials for the target student group. In this way, the research will involve the development of a software prototype that produces narrations for images. The obtained descriptions will be added to the images, enabling screen readers to provide the students with the required clarifications.

The solution will be created using a deep learning model that perceives and describes images. When generating textual descriptions, the solution will rely on the best existing methods of natural language generation. Evaluation of the developed system and potentially adjusted methods will involve tests with visually impaired participants and the adoption of evaluation metrics, such as BLEU and ROUGE scores. Overall, the current analysis will contribute to the field of computer science by advancing the accessibility of education in this domain for an important part of the student population.

Keywords:

AI-generated descriptions, visually impaired, computer science education, natural language processing, NLP, image recognition, assistive technology

1 Introduction

One of the biggest challenges in computer science education in numerous academic institutions, including the University of Pretoria, is ensuring accessibility for visually impaired students. Diagrams and pictures used in education are often unavailable in formats that can be translated into descriptive text specifically those relating to data structures [1][2]. This lack of accessibility can make the learning process challenging for computer science students with visual impairments. This is an example of one of the struggles several students may face in higher education.

An instance of the invocation of an artificial intelligence solution lies in the description of visual data. Still, the descriptions cannot accurately determine if the need at the hand of the visually impaired learners is achieved, mostly when given complex content in computer science. The research has shown that while some momentum may have been gained in understanding context with the description of data structures in computer science, further research is still required.

The goal of this research is to overcome those issues by focusing on a better generation for image description targeting computer science images and diagram based, with well contextually relevant descriptions which adds more value in the education aspects. The project is tailored to generating human-readable but extremely precise professional descriptions that can easily slot in as alt text, and be read aloud with a screen reader. This will in turn allow visually impaired students to interact more meaningfully with computer science content, thereby leading them through better understanding of complex data structures and algorithms.Leveraging deep learning models for image processing and NLP, the system will generate descriptive narratives that are then rated based on educational benchmarks regarding accuracy, readability and relevance using metrics like BLEU [bilingual evaluation understudy], ROUGE [recall-oriented underscoring for gisting] or CIDEr [3][4].

Not only does the research provide a new perspective on computer science education, but it has broader ramifications for assistive technology as a whole—introducing an adaptable and scalable solution that can be applied to other visually demanding fields of study in principle. The aim of this study was to investigate and enhance the employment of AI in educational accessibility so as assist all students have access to learning regardless their visual capacity. This approach is set within current trends in edutech using

inclusivity, adaptive learning and the promise of AI as a tool to fill existing education gaps among marginalised groups.

2 Problem Statement

There is an urgent industry problem in the field of computer science education about students who are visually impaired or have low vision due to difficulties accessing educational diagrams and images [1]. The other is a first year computer science student who is visually impaired and has approached us at the University of Pretoria on campus to transform images & diagrams into descriptive text along with possible speech support, in her learning process. The above case irony reflects the larger problem of lack-of appropriation and access to greater tertiary education, which is experienced by students with disabilities in South Africa [2].

Although using A.I to automatically turn visual data into descriptions is a very promising solution, but still it has many challenges. The latter is in line with the existing literature on using AI for visually impaired people, but a specific context of educational content (especially the one from complex disciplines such as computer science) has been raised to be something that warrants more exploration [5]. Recent research has shown the generated descriptions by state-of-the-art models may not consistently serve what visually impaired users are accustomed to in their special field (e.g., context-aware data structuring within computer science) [6].

The purpose of the proposed research is to address this issue by designing a software tool that processes documents and interprets images in them (using AI), so it becomes easier for visually impaired people. The realisation of AI in support to the visually impaired with respect towards acquiring knowledge becomes centered on how meaning is conveyed. For example, one study conducted by Hoppe et al. evaluated automatic image descriptions and showed that they can provide visually impaired students with complete, correct and easy-to-read information [7]. These would need to be crafted specifically for complex academic content, e.g., computer science illustration. As an example, the description generation model introduced by Muscat & Belz is based on capturing spatial relations, which also gives insights into data structures [8].

In this study, we will use machine a learning approach especially making use of the models that effectively generate detailed descriptions for data structure diagrams and images. Thus improving the learning experience and academic achievements of visually impaired students in computer science.

3 Background

3.1 Introduction

This chapter discusses the related literature of artificial intelligence (AI) based description for images and diagrams especially related to data structures for the visually impaired. It starts by defining basic concepts regarding visual impairment and provides several definitions along with classifications to better comprehend some of the issues related to visually impaired individuals. The conversation transitions into looking at Computer Science learning, what has progressed, and setbacks in the space including how to increase accessibility via inclusivity in education.

The chapter then moves to natural language processing (NLP), focusing on the use of NLP in human-computer interaction by way of understanding and generating natural language. In it, we go over the evolution of NLP technologies and reveal some interesting methods in how these can be applied to generate meaningful image captions. We also detail how both quantitative metrics (e.g. BLEU, ROUGE, METEOR, CIDEr, and BERTScore) as well qualitative methods such user feedback collected through surveys or interviews have been used to evaluate the AI-generated image descriptions.

The chapter introduces the image recognition tasks addressed by AI systems, and how deep learning enabled greater accuracy and speed on each. This often deals with issues like data bias and also more ethically contentious topics, particularly in the fields of surveillance and privacy. This includes AI assistive software, an examination of some specific technologies that already exist to help the visually impaired and listing their pros and cons. Such examples as Vivid, AI-WEAR and more advance NLP models like GPT-4 are then used to demonstrate the current capabilities alongside ongoing challenges such as cost, personalisation & privacy concerns.

Finally this chapter also introduces a conceptual model that blends AI and assistive technologies together to automatically convert diagrams or images

into descriptions for visually impaired. We propose a framework that leverages machine learning vision and NLP to parse visual data — mimicking the way biological systems process this information — in order to generate rich, interpretable linguistic descriptions of images. This model is intended to make computer science education more accessible through assistance for the visually impaired while also contributing to support a very less inclusive learning environment.

3.2 Constructs

3.2.1 Visual impairments

It is important to understand and define a visually impaired person since scholars have made various definitions of people with visual impairments. According to some researchers, visual impairment is divided into four types: low vision; partially sighted; blind (functionally); and either near of completely [9][10]. According to Agesa a person that is visually impaired means they have poor eyesight as a result of illnesses, injuries, or birth defects [11].

Those who have poor vision or partially sighted have lower visual acuities, which makes it harder for them to interpret visual information as much as someone who is sighted [12].According to Pascolini and Mariotti, people who are partially sighted still have a visual acuity of less than 6/18 to light perception . This implies that a person with limited sight must travel 6 meters closer to an object that a person with normal vision can see at a distance of 18 meters [13]. In contrast, the visually impaired have either no light perception at all or very little light perception in the dominant or better eye, measuring less than 3/60 . These do not represent fractions; rather, they show the distance (3 meters) from the eye chart that both a normal-sighted person (60 meters) and a VI person must stand to view and distinguish the same image [13]. Smith states that the phrase "visual impairment" refers to a collective term that describes an individual who has any degree of vision loss that makes it difficult for them to do any work that requires light [14].

3.2.2 Computer science learning

The evolution of the computer science education field has changed significantly over time as technology advances and undergo changes. In this construct we will look at key advancements and innovations in computer science education, referencing what has been found by the research community and their observations.

This article highlights the need for teachers and principals to encourage technology use in classrooms - calling out educators' implementations & pushing tech everywhere, not solely during special subject areas [15]. Modern computer science curricula have been shifting toward developing students' skills in both the foundations and actual practice of computing. This would include incorporating data science training across the entire curriculum and requiring sustainable, ethical focused coursework that prepares students for the analytical demands of the future[16].

Although there has been improvements, updating educational material and methods based on the pace of technological advancements can be a challenge. The attainment of this goal is, however, complicated by technical challenges such as limited resources and the pace at which technology advances linked with persistent needs for ongoing faculty development in education aimed at computer science students [17]. Looking forward, there is a trend towards more inclusive education strategies that not only teach people technical skills but also creative problem-solving and ethical considerations in computing. This focus will allow students to develop a flexible and versatile skillset that is relevant across many dynamic industries [18].

3.2.3 Natural Language processing

Artificial intelligence that deals with the interaction between computers and humans through natural language. The ultimate objective of NLP is to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful. NLP combines computational linguistics and artificial intelligence to process and understand human language. Early NLP research focused on linguistic structure analysis and fundamental technologies like machine translation and speech synthesis. Today, NLP technologies are integral to real-world applications such as spoken dialogue systems, social media mining, and sentiment analysis [19] . The integration of deep learning has propelled NLP forward, enhancing capabilities in language modelling, machine translation, and dependency parsing. This advancement is evident in the architectural improvements and the increasing accuracy of applications that manage complex language tasks [20].

Two of the examples that can be found in the real world concerning the use of NLP applications are IBM's Jeopardy and Wolfram Alpha. Both of them

show how, on the user side, a certain query can be processed with the help of various computational strategies [21]. On the one hand, those examples represent one of the goals that has been established for AI in general as a means of passing the Turing test by facilitating the way in which machines understand the natural language. However, on the other hand, there is still a number of challenges that have to be overcome despite the considerable progress in the field. For instance, these problems include adversarial attacks on models, bias, and the need for language understanding [22].

NLP is one of the leading fields where progress does not stand still. The branch of science that is closely connected to artificial intelligence and computational linguistics continues to develop and reach new horizons, encompassing other new applications along its way. The development of the techniques which make possible to understand and generate human language has truly bright perspectives.

3.2.4 Evaluation Methods

Evaluation metrics are quantitative measures that are used to evaluate certain aspects of a system, process, model, or reality. They are widely used in many fields, including, but not limited to, research, machine learning, information retrieval, and business. These metrics enable a standardised way to measure outcomes, compare measures, and make decisions. For example, in the realm of machine learning, several widely used valuation metrics such as accuracy, precision and recall measure how well a classifier performs in classification and predicting data, thus providing an idea for scholars on their predictive models [23].

Evaluating AI-generated image descriptions quantitatively is crucial for determining both their quality and how well they align with the image content. The most common metrics are BLEU (Bilingual Evaluation Understudy), ROUGE, METEOR, CIDEr and BERTScore, they are used regularly to evaluate the accuracy of AI generated descriptions in comparison to a source of truth.

BLEU (Bilingual Evaluation Understudy) is a popular metric for evaluating the accuracy of n-grams in generated text compared to a reference, mainly an exact word match [24]. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** measures the text generation on n-gram recall, telling how much of reference text has been copied into the

output [24]. **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** combines synonyms and stemming, considering both precision and recall so that it can compare how similar the generated text is to the reference text. [24]. **CIDEr (Consensus-based Image Description Evaluation)** is the consensus of image descriptions) compares generated captions to multiple human-provided reference sentences to determine how often a term appears in a reference set and then providing rewards or penalties, as appropriate [25]. **BERTScore** is a semantic similarity metric that matches each word in a generated text with the corresponding word in reference text using pre-trained BERT embeddings. It measures how similar generated and referenced words are depending on the context, rather than exact match of words [26].

In addition information can be collected with the help of surveys and interviews with visually impaired users. In particular, Popavic noted this qualitative data helps to identify areas of improvement and make sure that the AI-generated descriptions meet the needs of users. Thus, such information could be used to ensure that AI-generated descriptions evaluable and correspond to people's expectations [27]. It should be Note that combining both quantitative and qualitative metrics provide an opportunity to evaluate the performance and acceptance of AI-generated descriptions properly. At the same time, such a mixed approach also guarantees that the developers can enhance this technology effectively [4].

3.2.5 AI Image recognition

The use of Artificial Intelligence in image recognition, both images, and videos, has allowed many sectors to automate the process of identifying and classifying objects. This construct provides an overview of the improvements in AI image recognition, underpinned by the already conducted research and technological advancements. It should be stressed that AI technology, in particular, deep learning, has managed to transform computer vision, a part of which is image recognition. The convolutional neural networks and generative adversarial networks have enabled the improvement of imaging tools. Some tools are used in medical diagnosis while others are used in autonomous vehicles. There has been a tremendous improvement in the accuracy, speed, and efficiency of automatic image recognition systems [28].

Deep learning is the critical element of AI image recognition nowadays. Implemented by CNNs and GANs, these algorithms are trained on large

datasets to recognise recurring patterns and details. One of the most prominent uses of the technology is facial recognition, while it is also applied in automated surveillance and scene interpretation, among other things [29]. However, while there have been many beneficial advances in this field, they are not without challenges. One of the main ones is the issue of data bias: when the patterns recognised by a neural network are unjust, the results can also be unjust. It sometimes may result in the data's inaccuracy, pursuing nonexistent or very minor statistical trends. The other issue is the ethics of this nationwide level of tracking and surveillance in all, or nearly all, aspects of human life. Luckily, the experts are making progress in establishing fair, transparent, and accountable systems [30].

3.3 AI Assistive software

AI assistive technologies are resources designed to aid individuals with disabilities by enhancing their ability to perform tasks either impossible or very difficult for them. For the visually impaired, AI technologies have been particularly beneficial as they offer new ways to interact with the world through enhanced sensory inputs such as auditory or tactile feedback. Apps like Vivid allow users to gain information from the environment with added descriptions of how things work. It uses machine learning to process sensory information and provide auditory on what is present in the surrounding environment [31].

Current AI assistive solutions have limitations despite the progress. Most systems are prohibitively costly, not personalised for the end user and dependent on a constant connectivity supply. The design of AI-enhanced assistive adaptive technologies has to account for protecting privacy and ensuring inclusivity without infringing on the personal autonomy of users [32].

There are many AI methodologies that would offer tailored support for visually impaired students in computer science learning. For example, some techniques include real-time object detection, spatial audio feedback, and interactive learning tools that adjust to the speed and style of the learner. Real-time object detection as a part of this approach has already been used in the development of the smart text reader AI-WEAR, which was designed to support visually impaired students in reading by using Raspberry Pi and other innovative technologies. Therefore, it is clear that such methods are needed and can be beneficial [33].

Advanced NLP models such as GPT-4, developed by OpenAI, have made a notable contribution to a wide range of applications, and assistive technologies for visually impaired people are among them. The model provides more coherent and, what is more important, contextually correct text in comparison to the previous versions and earlier applications. For example, according to the study by Kozachek, GPT-3 could generate semantic errors in up to one quarter of long descriptions, while for GPT-4, this figure is two to four times lower. Additionally, the model generates accurate and detailed descriptions of relatively complex situations, which is of primary importance when it comes to providing educational content that is accessible for VI individuals [34]. As noted by Rahaman et al., GPT-4 is more performant in at least three aspects, i.e. improved training data, increased computation, and better responses, which implies that the newer model is advantageous for the application in question [35].

Sodbeans is an example of an auditory programming environment created for blind and visually impaired learners. This environment can provide students with information about the code through sound as well as syntax errors or program output. This gives students the opportunity to engage with programming tasks more independently [36].

Exploring the realm of AI assistive software, it becomes clear that current state is marked by important accomplishments, as well as many obstacles that are yet to be resolved. Although the daily functioning of visually impaired individuals has become much easier with the help of AI, its many problems, such as significant cost, privacy concerns, and the necessity for further personalisation, show that there are many ways in which AI can still be improved in order to facilitate the needs of visually impaired users, especially concerning their retrospective and prospective use of information in educational contexts.

3.4 Conceptual model

The provided conceptual model incorporates artificial intelligence with assistive technologies for producing original solutions related to visually impaired people. Thus, the described model involves the mutual work of computer vision and NLP to analyse the visual data, generating a comprehensive and understandable acoustic description of others' interaction with the world. The most important part of the model is relying on biological vision to connect the 'where' pathway with the 'what' one. This conceptualisation

supports an understanding of physiology-oriented low-level functional vision pathways and the high-level trails linked to object identification, which are to be united for overall image understanding. The authors state the following "We present a conceptual framework inspired by biological vision which integrates low-level vision functionalities oriented to actions with identification and recognition capabilities" [37].

Making the model even more complex, the past few years have seen rapid advances in AI that have facilitated the development of more advanced assistive tools that are intuitive and interactive at the same time. For example, Vasireddy et al. discuss the development of an Image Caption Generator based on Vision Transformers and GPT-2. After the description of the image is generated by GPT-2, its key features are delivered to the user in a more intuitive manner, "By combining the strengths of computer vision and NLP, our paper aims to extract significant image features using ViT and generate contextual, human-like descriptions through GPT-2" [38]. It is possible to judge of the efficiency of this model because it allows the visually impaired individuals to increase their independence in a variety of context. For example, regarding the 'e-Vision', Migkotzidis et al. state, "The proposed system is a context-aware solution and builds upon three important day-to-day activities" [39].

4 Methodology

4.1 Introduction

This study aims at developing, implementing, and evaluating an AI-based system that would generate descriptive text for images and diagrams presented in computer science educational materials for visually impaired students. The methodology consists of the following steps: system development, initial testing, usability testing, and performance evaluation. More specifically, an AI system will be developed that would be based on modern machine learning models specialising in image recognition and natural language processing. This system will be subject to initial testing to ensure its operability. Next, the performance of the system will be evaluated: it would be possible to estimate the BLEU and ROUGE scores and compare the AI-generated descriptions to the manually annotated ones.

4.2 Experimental Methodology

4.2.1 Experimental Setup

1. Hypothesis: The AI-driven program using OpenAI's models will generate accurate and contextually relevant descriptions for diagrams and images in educational materials.

2. Participants: Visually impaired individuals who will evaluate the generated descriptions.

3. Materials: PowerPoint and Word files with embedded images.

4.2.2 Procedure

Phase 1: System Development and Initial Testing:

1. Develop the AI-driven program as outlined in the system design.

2. Conduct initial testing to ensure the system functions as expected and generates descriptions.

Phase 3: Performance Evaluation:

1. Compare AI-generated descriptions with manually annotated ground truth using quantitative metrics (CIDEr).

4.3 System Design

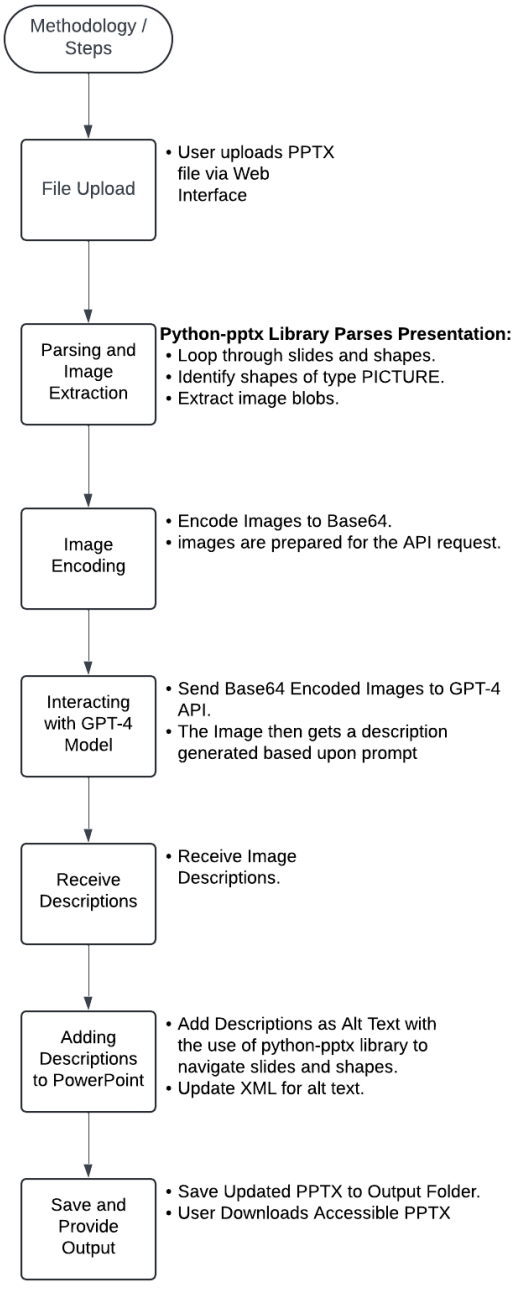

Figure 1: Design

This proposed system is intended to automatically generate textual descriptions for images within PowerPoint files in order to improve their accessibility. This process will start when any user upload the power point file through web interface. The system will use the python-pptx library to parse the presentation and retrieve images from each slide In detail, it will loop through each slide and shape to find the shape type PICTURE and save the image blobs to an output directory.

The system will encode each of the images to base64 so they can be sent back to OpenAI's GPT-4 model in a format it is expecting. The image data should be encoded because an API request that contains the image data in a form for the model to process must be provided. The system will provide descriptions of several prompt types. For all the classes of prompts, there will be different System and User prompt types that would guide GPT-4 to yield rich descriptions suiting to the context.

Once the descriptions are fetched, they will be placed in the original PowerPoint file as alt texts for corresponding images. To do so will require navigation to the specific slide and shape within the presentation as well in the XML representation of the shape. The python-pptx library allows this by giving access to the underlying XML elements directly.

When all the images have been processed (each time detecting an image and modifying it with correct alt text), the system will then save that newly updated PowerPoint file to a specified output folder. This will save the file and send a link to the user that will allow them to download their updated presentations. The result will be a PowerPoint file that can be used by screen readers, which will make the file accessible to visually impaired users.

4.4 Evaluation Metrics

Evaluation will be done by a quantitative and qualitative approach to enclose all the assessment type that help in model performance enhancement. The significance of CIDEr, a consensus-based metric, is well documented for the quantitative evaluation regarding how close machine-generated descriptions would be to human-annotated ones [25]. On the other hand, a more qualitative approach will include a survey and interviews with users who will be making use of the program and gathering feedback based on usability and experience.

4.4.1 Accuracy of Descriptions

The evaluation of how accurate the generated descriptions are will be conducted by comparison with the manually annotated source of truth descriptions against CIDERr evaluation metric.

CIDEr uses a vector edition of both the generated and reference sentences. The TF-IDF weights for a sentence are obtained by calculating the frequency and then weighing the words compared to all the reference sentences. It will prevent the model from punishing frequent words and it helps in rewarding uncommon words (that still might be meaningful). The cosine similarity of the generated description with respect to the human reference descriptions was then sequentially computed after weighing these candidate and ground truth, i.e., machine-generated and human annotated, integer lows. This comes down to a score that aligns with how closely the human annotations matched the machine output. It is designed to handle multiple reference caption, and thus particularly appropriate for such tasks where there can be many correct ways in describing an image. This feature is designed to increase robustness and reliability in comparison to simpler metrics such as BLEU or ROUGE, which might only rely on a single reference sentences or may lack the sophistication necessary to diversely account for correct answers human language [25].

This report will use CIDEr as the main quantitative evaluation metric for scoring the alignment between machine-generated text and human-annotated references. The implementation will include the following:

1. Preparation of Reference Data: There will be multiple human-annotated descriptions created and collected for each image in the dataset. The references are necessary for CIDEr to calculate the consensus score. Having more references gives a better base for any generated description to be evaluated as CIDEr is especially known to work well when multiple valid descriptions are provided [25].

2. Description Generation: The model will generate text description given an input image. Generated description will be compared against a set of reference descriptions.

3. TF-IDF Weighting: Every word in each of the reference descriptions will be given a TF-IDF weight based on how much it appeared across all other reference descriptions. The algorithm down-weights words that come up in a reference frequently, and gives the ones that are far

less frequent (but also necessary to describe well) much more weight [25].

4. Cosine Similarity Calculation: Calculate the cosine similarity between the TF-IDF vectors for the generated and reference sentences to count how well a machine-generated description captures the richness of human-generated references.

5. Scoring and Analysis: the ultimate CIDEr score is going to be determined as to whether or not the human annotated descriptions and the machine generated captions agree. A higher CIDEr score means the generated description is more consistent with human references, so the model works better.

4.4.2 User Satisfaction

Using semi-structured interview will be beneficial for this paper to collect the qualitative data and learn more about participants' and their experience of using AI generated descriptions. However, in this paper such method of collecting data will not be used. This qualitative data will make it clearer, in what areas the descriptions are adequate and in what areas they could be improved as stated by Graham [4].

This paper will not provide any user satisfaction data. As a recommendation for future work, investigating user satisfaction through analysing the statistics of participants' ratings and the detailed data of their insights from the interviews will be a warranted and beneficial approach. Such approach will contribute to understanding user satisfaction and the practical effectiveness of the descriptions generated by the AI.

5 Tool creation

The VisionAI System development went through several stages, combining theoretical background of previous research and practical realisation knowledge about cutting-edge artificial intelligence and assistive software technologies. Our target domain is Computer Science, focusing on educational materials because detailed and easy to understand descriptions for images & diagrams in text-based learning resources are required by visually impaired students. In the following section, we explain our process of building this system, what tools and methods we used and review how it reflects or contradicts with existing literature from the background

5.1 Architecture Overview

The VisionAI system is structured as a front-to-back multi-layered architecture. In addition to backend & frontend technologies it includes external AI services such as deep learning and natural language generation.

5.1.1 Frontend

The frontend of the system was created with React, which is a widely used JavaScript library for building UIs. We selected React because it provides an efficient way to create dynamic web applications, gives flexibility and has a good ecosystem around manipulated DOM. There is a simple web interface of the front end which allow users to upload their PowerPoint files including images and diagrams. The frontend is linked with the backend and communicates with an API that provides the status of a users upload and progress the processed files. This is very important because enables both visually impaired users and educators and others to efficiently manage and navigate the system.

5.1.2 Backend

Flask is a Python based web framework used for development of the backend side in an effective way. It is essentially the thing between what the user faces and sends to, processing it in order to create image descriptions. HTTP handling for incoming requests, file uploading management, AI model invocation and resulting in modified PowerPoint files.

The backend performs several essential tasks:

- **File handling**: It receives uploaded PowerPoint files, processes them securely, and stores them in a specified directory on the server.
- **AI model integration**: The backend communicates with external services (i.e., OpenAI's API) to generate descriptions for the images extracted from the uploaded PowerPoint files.
- **Response generation**: Once the files are processed, the backend returns the modified file to the user, now equipped with alt text descriptions for images.

5.1.3 AI Integration

The AI uses OpenAI's GPT-4 model to perform its main function: describing what is in the images it pulls out of PowerPoint slides. GPT-4 is a best-in-

class natural language processing (NLP) model that can produce human-like text in response to demanding contexts. We were able to feed this model in a way where we can make sure that generated descriptions were not only grammatically correct but also contextually appropriate for educational content.

5.2 Technologies Used

5.2.1 Flask

Flask version 2.0.1 was selected as the back-end framework due to its basic, yet high-level approach. It is simple to operate on web request that suits it as a best for building REST API which deals with file uploads and using AI services. Since the tool logic relates to file handling, image processing and AI integration — it was important that flask supports modular development of flask applications.

5.2.2 React

Using React version 18.3.1 for the frontend enables us to create an interactive user interface with immediate feedback. Due to React's Componentate Architecture, state management of the app (e.g. tracking upload progress and displaying error messages), can be easily done as well as making these ones interact with backend asynchronously through API requests.

5.2.3 OpenAI GPT-4

The OpeniAI version 1.37.1 made use of the GPT-4 model which is the key component allowing it to produce detailed descriptions. GPT-4, which was chosen for its ability to generate contextually coherent (and human-like) text crucial in creating descriptions that visually impaired students could depend on when trying to understand an intricate diagram or image in computer science.

5.2.4 python-pptx

The **python-pptx** library version 0.6.21 is used to manipulate PowerPoint files programmatically. It allows us to:

- Extract images from slides.
- Modify the slides to include generated descriptions as alt text.

- Save the modified file for the user to download.

5.3 Detailed Workflow

The VisionAI system uses a structured workflow approach that starts with the user interacting with the system and gives output of a fully accessible PowerPoint file:

5.3.1 File Upload and Image Extraction

If the user uploads a PowerPoint file through the React front-end, this file is sent to Flask back end and saved in its secure area. This uses the `python-pptx` library to open a file, iterate through each slide and identify any shapes that are images. It then extracts and saves the images in a temporary folder. This is essential because the images are what will be entered into an AI model, which will be generating the descriptive text.

5.3.2 Image Encoding and AI-Based Description Generation

The images are extracted and base64 encoded before sending to OpenAI GPT-4 via an API call. The model then has a prompt for each image telling it what to do in order to generate the description. The prompt is generated keeping in mind the image type and desired output:

- **General prompt**: Provides a broad description suitable for general images.
- **Exam prompt**: Focuses strictly on the visual elements without giving away context that could help in an exam setting.
- **Data structure prompt**: Focuses on the technical aspects of diagrams related to computer science, particularly data structures.

The GPT-4 model processes these inputs and generates descriptions, which are then returned to the backend.

5.3.3 Reintegration of Descriptions

The `insert_alt_text` function re-inserts the descriptions into their respective places in powerpoint as alt texts. It helps to make the PowerPoint accessible for assistive technology like screen readers, which will read these descriptions out loud to people with vision-related disabilities. Not only

does this step ensure accessibility compliance, it also enhances the learning experience for students needing craft feedback audibly.

5.3.4 File Saving and Response

After the processed PowerPoint file is ready, it is then saved to an output directory and return as a downloadable link for user. The user can directly download the updated file with these descriptions to make it more accessible.

5.4 Key Features from the Background Section

The development of the VisionAI system makes use from the concepts that was discussed in the background section from this research report, more specifically in areas like AI, NLP, and assistive technologies.

5.4.1 Natural Language Processing (NLP)

Clearly, GPT-4 is already following the NLP principles discussed in paper Background. This goes on to solve the problem faced by visually impaired students in reading complex academic content — specifically, domains like computer science where description of context matters.

5.4.2 AI Image Recognition

This is backed up by the discussion of CNNs & GANs resulting in image recognition and description generation processes. Our system does not implement these models directly, but the basic idea of recognising patterns in images and translating them into useful descriptions are fundamental to how our app is based.

5.4.3 Assistive AI Technologies

The System extends on prior work using assistive technologies such as Vivid and Sodbeans to increase the accessibility for visually impaired users. The VisionAI system takes this idea a step further, and is applied to complex educational content and establishing an entering point for people from traditionally under-stress fields of study.

5.5 Conclusion

VisionAI is an AI-based system developed to provide accessibility and accessibilities for visually impaired students in education. Using state-of-the-art knowledge of a NLP model and other web technologies, VisionAI provides an innovative means to increase educational access for all students but especially those within the technical domain like Computer Science. This is in response to the goals introduced before and prepares for future advances of assistive technology.

6 Evaluation Model

A detailed procedure was carried out for the examination of AI-generated image descriptions in VisionAI tool. The goal of this process was to assess as fairly and carefully as possible how the model performed for various prompts by comparing AI-generated descriptions with ground-truth references provided from expert annotators in our field. The main objective was to evaluate the fidelity of descriptions given different receiving prompt inputs and for testing which types of prompts are better suited to be used.

6.1 Evaluation Setup

The evaluation was done with the custom framework that we developed to work on image description dataset. The AI model generated these descriptions in response to given prompts, and their precision was evaluated against the provided ground truth descriptions from experts. This method helped to make sure evaluations were based on human intelligence, more so in the case of technical content such as computer science visualisation.

6.1.1 Dataset and Ground Truth

The dataset contains eight images of data structure trees content that were selected as common examples of visual content found in the study of computer science (e.g., data structure diagrams and flowcharts). Domain experts annotated these images to derive ground-truth descriptions that were semantically meaningful and relevant from a pedagogical point of view. These descriptions assembled by experts are meant to serve as reference points or grounds of truth for comparison against the output of an AI.

A comparison is made for each generated description against the ground-truth descriptions according to the CIDEr metric. Using expert crafted

rubrics as reference points helps in ensuring that the assessments are stringent and quality is maintained as per academic standards.

6.1.2 Prompts for AI Generation

A few prompts were created to analyze the flexibility and correctness of VisionAI model related to relevant descriptions. The following prompts were all designed to stress different aspects of the input and evaluate the model response towards different specificity and context information.

- General visual inventory prompts (e.g., "Conduct a comprehensive visual inventory of this image.")
- Analytical prompts focusing on explaining the image's visual structure (e.g., "Provide a thorough visual analysis of this image.")
- Contextual prompts simulating a teaching scenario (e.g., "Imagine you're explaining this image to someone learning data structures.")
- Technical prompts tailored for specific content, such as data structures (e.g., "Please provide a structured analysis of each tree diagram element.")

Thus, the evaluation was not just about how accurate the AI-generated descriptions were, but rather which prompts produced a certain construction of what could be considered an acceptable description. The prompts that were used to compared over eight images and CIDEr scores showed which prompt returned descriptions that are closest to the ground of truth references. The goal was to see what types of prompts produced the most contextually accurate, rich descriptions for each image presented in the dataset.

6.2 Framework for Evaluation

This custom evaluation framework also helped us compare human-provided ground truths with the descriptions generated by an AI model. This framework enabled the loading of datasets, multiple prompts for each image and made it ultra-efficient to generate descriptions.

6.2.1 Description Generation and Comparison

The AI model would provide descriptions for each image based on input prompts included with the dataset provided. Then the expert provided

ground-truth descriptions are used to compare them with the AI Generated ones. We evaluate the performance of various prompts on effectiveness for helping model to generate correct descriptions. This comparison was done by taking a description and comparing it to the ground truth, looking at whether or not that prompt gave descriptions with good performance in terms of human-generated references.

This part of evaluation actually refers to some ideas belonging to AI Assistive Software in the back ground. When it comes to assistive AI technologies, the description accuracy is essential for accessibility. Here as well, the evaluation process went into consolidating AI-generated descriptions on multiple prompts to make it accurate and useful for visually impaired users. The framework allowed for a detailed comparison of how different prompts influenced the quality of the AI's output.

6.2.2 Flexibility for Dataset Expansion

This framework was inherently scalable, so the dataset could grow should more images and ground-truth descriptions be added. As more data gets added, the evaluation process could be able to improve and offer a better identification of how well an AI model works. Especially, flexibility of the framework is crucial here due to different AI model developments.

6.3 Use of Background Concepts in Evaluation

Several key concepts from the **Background** were applied in the design and execution of the evaluation process:

- **Expert-Provided Ground Truth**: The background of the Evaluation Methods section pointed out that human-provided references are crucial when assessing AI descriptions. In our case, expert-provided ground truths were used to guarantee that the descriptions adhered to an standard of accuracy and relevance, especially for computer science education.
- **NLP for Prompt-Based Generation**: This use of different prompts corresponds with the ideas we learnt in Natural Language Processing section. Through testing the AI at describing images from various prompts, evaluation investigated how well such a model was able to read and understand natural language inputs.
- **Assistive AI Tools**: The evaluation process is also influenced by AI Assistive Software principles with the aim to create a description that

gives accessibility as well as utility for visually impaired users. Measure of comparing prompts for consistent and meaningful descriptions are an essential component to ensure that the instructions could directly support its learning mission in educational settings.

6.4 Evaluation of Prompts

CIDEr scores were logged for each prompt for eight different images in order to evaluate the performance of each prompt for generating correct image descriptions. The CIDEr scores, both at the individual image score level (i.e., each image should only have one score), and an average score across all images for each prompt are tabulated in the following Figure 3 . Higher scores indicate greater correspondence to descriptions made by experts.

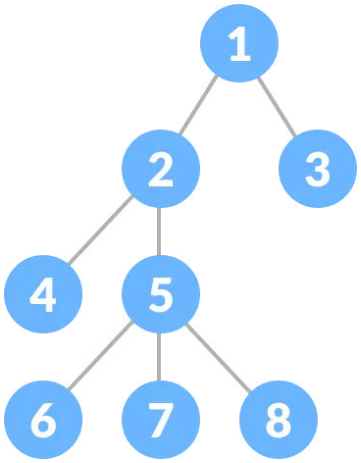

Image 1

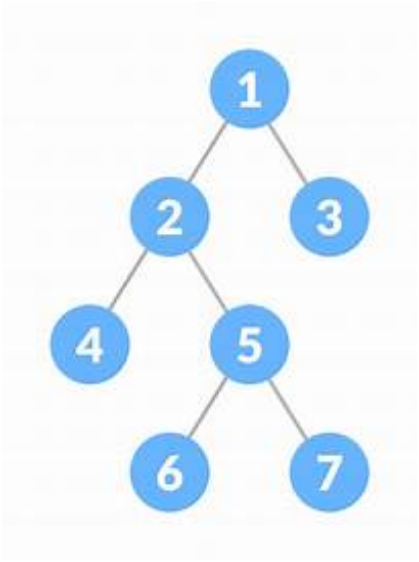

Image 2

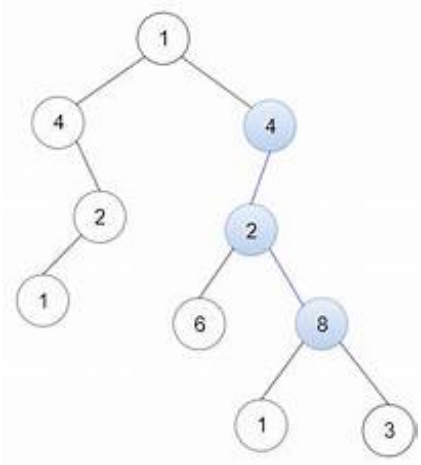

Image 3

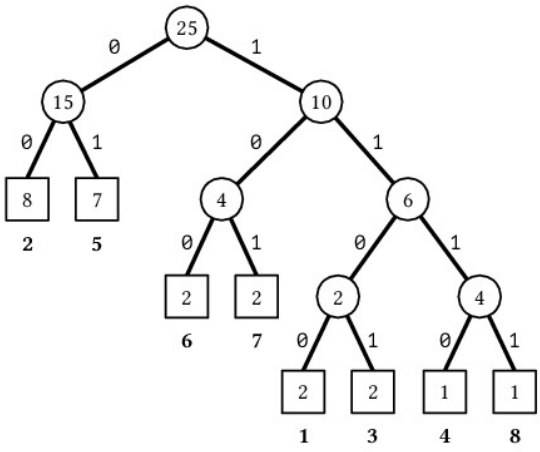

Image 4

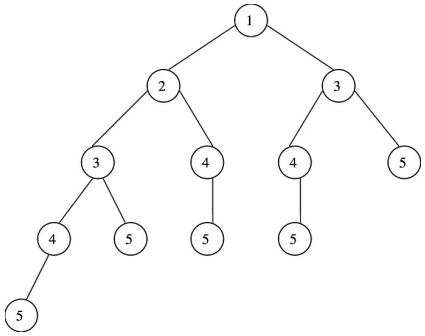

Image 5

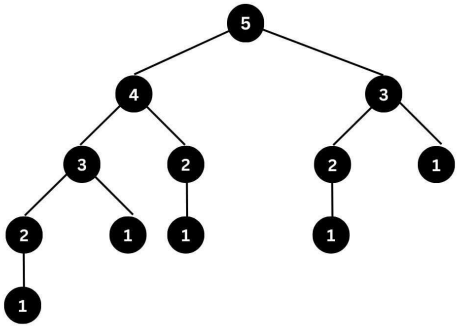

Image 6

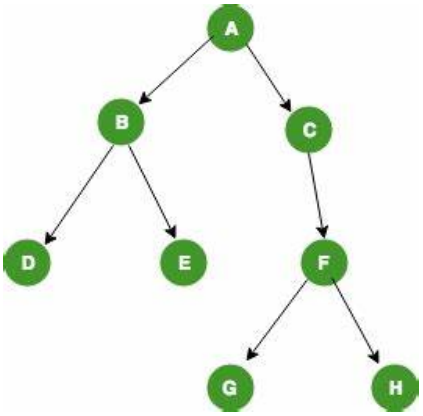

Image 7

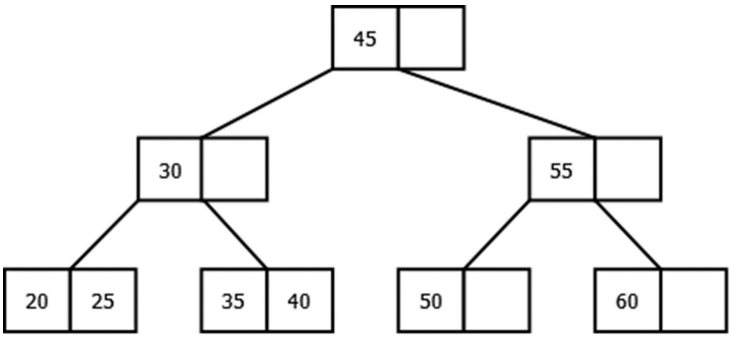

Image 8

Table 1: Images arranged from 1 to 8

Prompt 1	Conduct a comprehensive visual inventory of this image. List and describe all observable elements, including their size, position, color, and any apparent relationships between them.
Prompt 2	Provide a thorough visual analysis of this image, detailing its composition, colors, shapes, and any text or numbers present. Describe the overall structure and how individual elements relate to each other.
Prompt 3	Imagine you're explaining this image to someone who cannot see it. Offer a detailed, step-by-step description of all visual elements, starting from the most prominent features and progressing to the smallest details.
Prompt 4	For each tree image, please describe its overall type (such as binary, directed, or search tree) and its visual style including node colors, shapes, and how labels are presented. Starting from the root node at the top, explain its value and position, then describe how the main branches extend downward, including each node's children and their values, plus any special characteristics like empty squares or directional arrows. Finally, note any notable structural features like the deepest paths or how values are ordered, keeping each description to 3-4 sentences while maintaining clear positioning and relationships between nodes, and ensure each description stands independently without referencing other images.
Prompt 5	Please provide a structured analysis of each tree diagram by first describing its fundamental type and visual appearance including shapes and colors. Then walk through the structure beginning at the root node, detailing how each level connects to the next and specifying the values and positions of all nodes, making sure to note any unique features like arrows, empty spaces, or special formatting. End by highlighting any notable patterns or characteristics of the overall structure, keeping the description focused and self-contained in 3-4 clear sentences.
Prompt 6	Examine each tree diagram and describe its core structure by identifying its type and visual elements, then trace the connections from the root node through each branch, noting all values and their positions. Document the relationships between parent and child nodes, including any distinctive visual elements or patterns in the structure, and conclude with any significant features of the overall organization, maintaining brevity while capturing all essential details in a self-contained description.

Figure 2: Prompt used in evaluation

Prompt	Image 1 CIDEr Score	Image 2 CIDEr Score	Image 3 CIDEr Score	Image 4 CIDEr Score	Image 5 CIDEr Score	Image 6 CIDEr Score	Image 7 CIDEr Score	Image 8 CIDEr Score	Avarage CIDEr Score
Prompt 1	0.2052	0.4679	0.2971	0.2971	0.5486	0.3056	0.4185	0.2091	0.3547
Prompt 2	0.3326	0.3194	0.3574	0.2639	0.5079	0.3958	0.3933	0.2292	0.3499
Prompt 3	0.3886	0.3235	0.3169	0.2869	0.346	0.3892	0.2934	0.1503	0.3118
Prompt 4	0.3938	0.3515	0.3274	0.2915	0.5087	0.4681	0.3912	0.236	0.371
Prompt 5	0.3427	0.3824	0.3394	0.3226	0.4766	0.4859	0.5111	0.2668	0.3909
Prompt 6	0.5576	0.481	0.3289	0.241	0.4318	0.3354	0.3821	0.2268	0.3731

Figure 3: CIDEr Scores for Prompts Across Images

- **Average CIDEr Scores:** The performance of a prompt was quantified as the average CIDEr score over the eight images for every prompt. An higher average score indicates that a prompt matched the expert-generated descriptions more closely. For instance, better performing prompts were evident when written in structured instructive format describing specific aspects of data structures, indicating that technical prompts performed better in this educational setting.

- **Prompt Variability and Effectiveness:** CIDEr scores showed that more unambiguous and concrete prompt wording led to higher consistency and relevance in generated descriptions. Average scores were lower for general prompts or ones without specific technical cues, signaling that contextualized prompts relating to the content type and educational context helped the model.

- The fact that all technical prompts(i.e. those prompts targeting specific components of data structures) scored consistently higher, confirmed that domain specific prompts lead to more accurate descriptions of the educational topics. Conversely, general visual analysis prompts received a much lower score, indicating that the tool may fail to capture important aspects of the scene without context.

The results obtained from this evaluation highlight the need for customised prompt design for quality AI descriptions of educational images especially in fields like computer sciences. That is a huge insight for prompt optimization to unlock accessibility tools for students who identify as having visual impairments.

6.5 Evaluation Results

CIDEr scores for eight images and 6 types of prompts, together with evaluation results for the descriptions generated by VisionAI tool, provide insight into performance across different types of instructions. Key findings that were discovered includes:

6.5.1 Average CIDEr Scores by Prompt Type

The average CIDEr scores were higher for prompts that contained more structured, technical directions related to data structures. Prompts that contained specific references to elements of computer science concepts (for example, "tree diagrams") averaged 0.39 over images, indicating better model alignment to domain descriptions with specificity. Less specific but more general prompts like "Provide a thorough visual analysis of this image" scored about 0.35 suggesting a general but less accurate coverage of details which may be relevant in educational contexts.

6.5.2 CIDEr Score Range Across Images

For the most discrete images like simple flow-charts, or diagram components, some prompts even reached scores of as high as 0.55, but scores varied significantly between images. Scores were generally lower for complex images using generic prompts, highlighting again how much the structured prompt facilitates the model in producing accurate, contextually relevant descriptions.

6.5.3 Impact of Prompt Specificity

Structured prompts generated higher scores than unstructured prompts in every image tested, with a consistent average difference of 0.05 points for the CIDEr score. Overall, detailed, context-driven prompts led to higher scores with structured prompts. This implication indicates that when working with complex visual content, AI models such as VisionAI benefit from explicit, unambiguous prompts. For users with vision impairments, this means less general and more detailed descriptions and overall better usability for tools like VisionAI in educational use cases.

6.5.4 Average Score Comparison

When compared to general prompts that scored lower, the top-scoring prompt was able to align more closely to expert references, earning an average CIDEr

score of 0.39. This highlights the importance for prompt tailoring of VisionAI's descriptions in order to align with the educational thresholds of accessibility of computer science learning content.

6.5.5 Overall Evaluation

Our CIDEr score analysis indicates that the tool excels when driven by comprehensive and domain-specific prompts, as we have seen in the case of computer science oriented educational images. Our results validate that context-enhanced AI via customized prompts greatly improves the wordings of the descriptions generated from the AI which can increase the assistive nature of tools in educational settings.

The implications of these findings is that future versions of VisionAI should build support for prompt engineering as a first class feature where educators can input structured prompts resulting in a more on target description when describing for visually impaired learners. The significance of the CIDEr scores show that there is room in tuning accessibility tools via prompt engineering, which is beneficial for CS blind student in complex subjects.

6.6 Scalability and Future Work

This framework of evaluation is promising and can serve as a good basis for assessing how accurate AI-generated descriptions are given different cues in the form of prompts. This allows the framework to evaluate how well that model performs over an increasingly larger sample size of images and prompts types. The paper also suggest that future work involve qualitative feedback from visual impaired users, in order to put these descriptive qualities into context with respect to how effectively they satisfy the needs of people in real-world settings.

6.7 Conclusion

Evaluation of the VisionAI model involved comparing AI descriptions with expert-curated ground truths. The evaluation helped to know how capable the model is in handling a wide range of prompts and provide meaningful context-specific descriptions. The characteristics of the evaluation framework ensures that we can leverage its scalability to continue use it on more tests sets as our dataset gets larger, contributing towards our broader goal for broadening participation in computing education.

7 Discussion

Results obtained from this research ascertain that an AI-driven model is much more effective in creating fuller and most context proximate descriptions of diagrams and images in the computer science domain for visually challenged students. This model, utilising the latest natural language generation (NLG) and image recognition technologies is designed to describe complex data structures with a high degree of accuracy.

The main benefit of the system as developed is that it can describe complicated images and diagrams such as graphs, trees etc., which are typically integral parts in computer science education. Using GPT-4 for NLP, the system was able to produce accurate descriptions that also conveyed this information effectively. The tool made it possible for visually impaired students to access materials that are otherwise difficult or impossible without a visual counterpart.

The evaluation metrics with CIDEr also supported the results that AI descriptions are closely related to those of human-annotated captions. Naturally, the model was capturing relevant portions of images — it made holistic sense to learn associations among different elements in structures as diverse as nodes and edges in graphs or hierarchy contained within tree structures. Since this model is capable of handling complex text well, it may be suited for educational applications that require both accuracy and clarity.

While this model worked effectively, we occasionally observed minor details were missed or omitted in the descriptions generated by it. These are not crucial parts of the description but lack thereof just leaves descriptive gaps in the image. For instance, the model could get bindings wrong in some cases of a graph or might miss labeling some less important nodes/edges not central to the image, or it may provide an ambiguous description for elements which are even somewhat a peripheral part of UML. Although these inaccuracies are negligible in the context of educational understanding, they do not detract from its clarity and represent a potential subject for model refinement to produce more nuanced subtle details.

There is also the possibility that these minor omissions are due to properties natural language processing models, despite being advanced ML systems do not always exhibit robustness with respect or precision across all components of complex diagrams. Refinements to the model's attention mechanisms or additional layers of image processing can be used in an attempt to fill these minor gaps. Nonetheless, the system can deal with challenging images fairly well (in a few cases just omitting small pieces), which is an important result to show that it would most likely be workable in practice for learning processes.

The prompts which were used to steer the natural language generation process, played a key role in making this model work. Further, the model used a tailored prompt with respect to what kind of image was being described (technical diagram/formula or flowchart like in case above and data structures otherwise), generating captions confined only to relevant contextual details within an image. The ability to easily adapt the system is essential in that it allows us to handle all of the different types of images that appear frequently in computer science content.

Although not the focus in this study we will also consider adding qualitative feedback from visually impaired users to future Totalitics versions. And user feedback could help make the model better by letting us know how well these descriptions land and whether they are in fact meeting any needs of our audience. It could also lead to more personalisation, such as tweaking down the type or level of detail/complexity in descriptions depending on what students have separately indicated they understand well.

Another promising area for further research is personalisation. While the current system produces good descriptions for a majority of users, one possible weakness is that individual preferences can be somewhat polarised — especially in terms of what information will an image need to have more details about. The system could be further advanced in the future to include adaptive learning techniques where the AI changes its descriptions right on the fly, dependent upon feedback from users or predefined preferences. In turn, this would provide a personalised learning experience through the audience with impaired vision when studying computer science topics.

Furthermore, whereas this study had a strong emphasis on screen reader feedback through auditory channels one could imagine sytem extensions that would take in comibined multi-sensory feedback such as tactical interfaces etc. This has the potential to make an educational experience even better, in particular when it comes down spatially oriented data structures and tactile exploration combined with audio descriptions can provide more immersive understanding of the material.

In conclusion, it is established that the AI-driven system shows a high accuracy in understanding and captioning challenging images or complex diagrams related to computer science. Sometimes details of little significance are left out in the process but while this exists, hardly did it take away from them being taught meaningfully to visually challenged learners. Future developments should focus on further improving the detail recognition of such a system and incorporating user feedback to make this an effective tool for increasing accessibility in technical education.

8 Conclusion

References

[1] N. C. Sanderson, S. Kessel, and W. Chen, "What do faculty members know about universal design and digital accessibility? A qualitative study in computer science and engineering disciplines," *Universal access in the information society*, vol. 21, pp. 351–365, 3 2022.

[2] L. Vincent and D. Chiwandire, "Funding and inclusion in higher education institutions for students with disabilities," *African journal of disability*, vol. 8, no. 1, pp. 1–12, 2019.

[3] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

[4] Y. Graham and Q. Liu, "Achieving accurate conclusions in evaluation of automatic machine translation metrics," pp. 1–10, 2016.

[5] D. L. Fernandes, M. H. F. Ribeiro, F. R. Cerqueira, and M. M. Silva, "Describing image focused in cognitive and visual details for visually impaired people: an approach to generating inclusive paragraphs," *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 1 2022.

[6] A. R. Crockett and G. Gannod, "Improving understanding of data structures for the blind with tactile media and a user-centered iterative approach," *2020 IEEE Frontiers in Education Conference (FIE)*, pp. 1–8, 2020.

[7] A. Hoppe, D. Morris, and R. Ewerth, "Evaluation of automated image descriptions for visually impaired students," in *International Conference on Artificial Intelligence in Education*, 2021.

[8] A. Muscat and A. Belz, "Learning to generate descriptions of visual data anchored in spatial relations," *IEEE computational intelligence magazine*, vol. 12, pp. 29–42, 8 2017.

[9] P. R. Cox and M. K. Dykes, "Effective classroom adaptations for students with visual impairments," *Teaching Exceptional Children*, vol. 33, no. 6, pp. 68–74, 2001.

[10] A. M. Mulloy, C. Gevarter, M. Hopkins, K. S. Sutherland, and S. T. Ramdoss, "Assistive technology for students with visual impairments and blindness," *Assistive technologies for people with diverse abilities*, pp. 113–156, 2014.

[11] L. Agesa, "Challenges faced by learners with visual impairments in inclusive setting in trans-nzoia county," *Journal of Education and Practice*, vol. 5, no. 29, pp. 185–192, 2014.

[12] C. L. La Voy, *Mathematics and the visually impaired child: An examination of standards-based mathematics teaching strategies with young visually impaired children*. PhD thesis, University of Kansas, 2009.

[13] D. Pascolini and S. P. Mariotti, "Global estimates of visual impairment: 2010," *British Journal of Ophthalmology*, vol. 96, no. 5, pp. 614–618, 2012.

[14] D. W. Smith, "Assistive technology competencies for teachers of students with visual impairments: A delphi study.," 2008.

[15] C. Raymond, "Technology integration in the classroom," vol. 2016, pp. 1–6, 2016.

[16] E. V. Dusen, A. Suen, A. Liang, and A. Bhatnagar, "Accelerating the advancement of data science education," pp. 1–4, 2019.

[17] A. N. Kumar, J. Beidler, Bhagyavati, H. Farian, M. Haas, Y. Kushleyeva, F. J. Lee, and I. Russell, "Innovation in undergraduate computer science education," *Journal of Computing Sciences in Colleges*, vol. 21, pp. 138–142, 2005.

[18] D. Passey, "Computer science (cs) in the compulsory education curriculum: Implications for future research," *Education and Information Technologies*, vol. 22, pp. 421–443, 2017.

[19] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, pp. 261 – 266, 2015.

[20] A. Raj, R. Jindal, A. K. Singh, and A. Pal, "A study of recent advancements in deep learning for natural language processing," *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pp. 300–306, 2023.

[21] N. Shah, "Recent technological advances in natural language processing and artificial intelligence," *ArXiv*, vol. abs/1208.4079, 2012.

[22] M. Omar, S. Choi, D. Nyang, and D. A. Mohaisen, "Robust natural language processing: Recent advances, challenges, and future directions," *IEEE Access*, vol. 10, pp. 86038–86056, 2022.

[23] V. Plevris, G. Solorzano, N. Bakas, M. El, and A. B. Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," *8th European Congress on Computational Methods in Applied Sciences and Engineering*, 2022.

[24] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *Annual Meeting of the Association for Computational Linguistics*, 2014.

[25] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2014.

[26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *ArXiv*, vol. abs/1904.09675, 2019.

[27] M. Popovic, "Informative manual evaluation of machine translation output," pp. 5059–5069, 2020.

[28] D. K. Prajapati, D. Upadhayay, A. Kumawat, and A. P. Shekhawat, "Application of ai computer vision and image recognition," *Industrial Engineering Journal*, 2022.

[29] C. Li, X. Li, M. Chen, and X. Sun, "Deep learning and image recognition," *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, pp. 557–562, 2023.

[30] A. Hosny, C. Parmar, J. Quackenbush, L. Schwartz, and H. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, pp. 500 – 510, 2018.

[31] S. Alhazmi, M. Kutbi, S. Alhelaly, U. Dawood, R. Felemban, and E. Alaslani, "Utilizing artificial intelligence techniques for assisting visually impaired people: A personal ai-based assistive application," *International Journal of Advanced Computer Science and Applications*, 2022.

[32] N. McDonald, A. K. Massey, and F. Hamidi, "Ai-enhanced adaptive assistive technologies: Methods for ai design justice," *IEEE Data Eng. Bull.*, vol. 44, pp. 3–13, 2021.

[33] A. Llorca, "Ai-wear: Smart text reader for blind/visually impaired students using raspberry pi with audio-visual call and google assistance," *International Journal of Advanced Research in Computer Science*, 2023.

[34] D. Kozachek, "Investigating the perception of the future in gpt-3, -3.5 and gpt-4," *Proceedings of the 15th Conference on Creativity and Cognition*, 2023.

[35] M. S. Rahaman, M. M. T. Ahsan, N. Anjum, H. J. Terano, and M. M. Rahman, "From chatgpt-3 to gpt-4: A significant advancement in ai-driven nlp tools," *Journal of Engineering and Emerging Technologies*, 2023.

[36] A. Stefik, C. Hundhausen, and D. W. Smith, "On the design of an educational infrastructure for the blind and visually impaired in computer science," *Proceedings of the 42nd ACM technical symposium on Computer science education*, 2011.

[37] M. Chessa, N. Noceti, F. Odone, F. Solari, J. Sosa-García, and L. Zini, "An integrated artificial vision framework for assisting visually impaired users," *Comput. Vis. Image Underst.*, vol. 149, pp. 209–228, 2016.

[38] I. Vasireddy, G. Bindu, and R. B, "Transformative fusion: Vision transformers and gpt-2 unleashing new frontiers in image captioning within image processing," *International Journal of Innovative Research in Engineering and Management*, 2023.

[39] P. Migkotzidis, F. P. Kalaganis, K. Georgiadis, E. Chatzilari, G. Pehlivanides, S. Tsafaras, K. Monastiridis, G. Martinidis, S. Nikolopoulos, and I. Kompatsiaris, "e-vision: An ai-powered system for promoting the autonomy of visually impaired," vol. 3, pp. 33–53, 2020.

Plagiarism and Generative AI Declaration

Full names	Jano Esterhuizen
Student number	u20583703
Topic of work	Using AI to describe images for the visually impaired

Declaration

1. I understand what plagiarism is and I am aware of the University's policy in this regard.
2. I declare that this dissertation (essay, report, project, assignment, dissertation, thesis, etc.) is my own original work. Where other people's work has been used (either from a printed source, internet or any other source), this has been properly acknowledged and referenced in accordance with the requirements as stated in the University's plagiarism prevention policy.
3. I certify that this assignment represents my own work. I have not used any free or commercial systems or services offered on the internet or text-generating systems embedded into software. e.g. ChatGPT, Gemini etc.
4. I have not used another student's past written work to hand in as my own.
5. I have not allowed, and will not allow, anyone, to copy my work with the intention of passing it off as his or her own work.

Signature __Jano_____

Date: __14/07/2014_____