UNIVERSITY OF COPENHAGEN

A thesis presented to the Faculty of Humanities in partial fulfillment of the requirements for the degree

Master of Science in IT & Cognition

Computer Generated Picturebooks

A cross-disciplinary human evaluation studying multimodal fusion in computer generated picturebooks

Martin Daniel Bare (MVK351)

MVK351@alumni.ku.dk

Supervisor: Manex Agirrezabal

May 2021

ABSTRACT

Stories are an important part of human history and remain as popular today as they have ever been. Although there are many ways to communicate stories, images and text have always been two of the most popular modalities to do so. When they are combined in picturebooks you can achieve many benefits, such as better understanding of the stories, teaching of visual literacy and creating content that is exponentially more complex than their unimodal counterparts. The idea of Multimodal Fusion has become increasingly popular in technical fields, and integrating research into picturebooks from disciplines like Psychology, Humanities and Education, can help to understand the intricacies that occur when combining computer generated images and computer generated textual stories. For this paper I have created several computer generated picturebooks using Bigsleep; BigGAN and CLIP for image generation, and Plan-and-write; high-level storyline architecture and GPT2, to generate text. To measure the results, I created a human evaluation form. The form focuses on several metrics pulled from a literature review, Specificity, Narrative Consistency, Relevancy and Multimodal Fusion. The three prior are intended to measure how the images represent the text. Specificity is a measure of how well the text fits individual images, Narrative Consistency measures how well images and text fit together as a sequence to create a story, and Relevancy measures how well the text and images mesh with one another to create picturebook text-image pairs. The Multimodal Fusion metric is intended to measure what type of product is created when text and images are combined and how the two modalities interact to create that product. 58 participants filled out the form and from that I conclude that the picturebooks are low in Specificity and Narrative Consistency, but high in Relevancy, and the Multimodal Fusion that occurs is perceived as enhancing or complementary and therefore beneficial in leading to a stronger product than their unimodal counterparts. This is significant, and

may indicate that computer generated picturebooks only need high Relevancy to replicate the benefits achieved by human tailored picturebooks. There are several limitations with the evaluation, such as no control over evaluations environments, subjectivity of human evaluations and the metrics, and using the word “story” to describe the picturebooks may all have biased results. From this I would argue that more work needs to be done in this area, as it is very sparse and the cross-disciplinary view allows one to draw conclusions that you would not normally find in Data Science papers. For future research, I believe that creating a holistic product with a UI to create stories could be beneficial. Using newer technology like Open-Ai’s DAll-E image generation and GPT3 would likely lead to more human-like results and strengthening and more closely measuring concreteness with the advent of “people” and “object” description generators may lead to higher Relevancy and Narrative Consistency. Finally, when working with and furthering this type of technology, it is important to consider how it may be used. It would be beneficial in improving human communication with and understanding of intelligent systems and creating stronger computer generated media. However, this type of technology could also be used for corporate greed in the wrong hands and may destroy jobs in industries such as Graphic Design. Keeping that in mind, the results seem to indicate that more human-centered and cross-disciplinary research into computer generated multimodal products could be beneficial for various fields in the future.

CONTENTS

Part I

INTRODUCTION

1

STORYTELLING

Storytelling is one of the most ancient human traditions and customs. You can find stories throughout all of human existence, from the remnants of cave paintings by our ancestors in Spain, to the most ancient written story in our possession, "*The Epic of Gilgamesh*" from Mesopotamia, to ancient Greek tragedies still influencing our stories today, Shakespeare taught in English classes all over the world, and now modern day movies, books and tiktok "story time" videos. The communication of stories has been an ever existing practice since our origins and is still as popular today as it ever was. From its constant popularity throughout human history, there is no doubt about how important and permeating stories are to human society as a mode of communication, recollection and imagination. Stories are often consumed for entertainment purposes, but are also useful tools. Therefore, it is not surprising that one of the most active fields within machine learning is *Natural Language Generation* (NLG), whose goal is to create computer-architecture that can emulate textual language and even create computer-generated stories. The sub-field specifically focusing on story generation is known as, *Automated Story Generation* (ASG), and is encompassed by story representation, story understanding, and story generation (Alabdulkarim et al., 2021). Currently, these fields of automatic-storytelling are evolving rapidly with the advent of new technology. This evolution is argued to be important as it allows us to learn about and communicate with intelligent systems (Alabdulkarim et al., 2021). As technology is an ever increasing presence in our lives, a deeper understanding of various technologies and their implications becomes very important. Recently NLG has advanced tremendously in scope

and proficiency, and is beginning to show increasing popularity in the public sphere. State-of-the-art NLG architectures, known as transformers, such as GPT2 and the recently released GPT3, have revolutionised language generation by raising the bar of increasingly human-like results produced by technology. GPT3 is now widely used for content creation for websites, copy-writing, script-generation and various other products. This technology brings us closer to the singularity where it will be impossible to distinguish between digital content created by humans and computer generated content.

There are many mediums used to communicate stories, but two of the most common are images and words. We often compare textual language to images. For example, most people will know the old saying, "A picture is worth a thousand words." There is much debate as to which form is better or worse for certain aspects of communication, but there is no doubt how important both mediums have been in the history of storytelling. They are both visual storytelling devices that are often paired together. It can even be argued that language itself evolved from images. Older written language systems such as the Egyptian hieroglyphs are collections of symbolic images like birds and Ankhs that when combined make up a textual language. Like ASG, image generation has taken a huge leap forward in the past two years. With the advent of Generative Adversarial Networks, otherwise known as GAN networks, and now Open-Ai's "Dall-E" image generation architecture, one can create impressive images from a textual input otherwise known as a prompt. From an entertainment perspective, we especially enjoy stories with accompanying images, like movies, since they are more expressive when compared to exclusively textual media. Images can also help comprehension and imagination of a textual story. Studying their interaction involves the study of multimodality. *Multimodality* is the study of the interaction that occurs in the combination of multiple literacies into a single medium. The product of this interaction is studied through the concept of *Multimodal Fusion*, or the study of the entity created through multimodal interaction (Shareha et al., 2009). Despite the concept of multimodality being nothing new, the use of the concept within technical fields like Data-science and Cognitive Science is growing and now being researched more than ever. As we begin to understand more benefits of

multimodal fusion, the integration of text with pictures will likely become increasingly important as these technologies evolve and their capabilities can be further used in the public sphere for content creation or teaching. Combining multimodal computer-generation to create picturebooks is now possible and will potentially provide many benefits that unimodal generation cannot.

2

MOTIVATION

This section is intended to provide the motivation behind why this work is important. It will first briefly highlight some research in various disciplines supporting the importance of picturebooks and the benefits of image generation over image retrieval. Finally, the section will address what I did for the paper.

The concept of multimodality is now widely taught in schools in most of Scandinavia, Australia, Singapore, British Columbia in Canada, and New Zealand (Callow, 2020). One of the most applied forms of images and text combined in a multimodal format are picturebooks. When combining images and text in a picturebook, this union can lead to a product whose interaction is limitless and unique because of the high degree of *Interanimation*, or fluid interaction between the two modalities (Lewis, 2001). An example of this fluid interaction, is how the exact same textual story like "*The Three Little Pigs,*" can be accompanied by different images leading to an entirely different story (Frederico, 2013). Even changing the colours, tone, or expression of a character in a single image within an image sequence can completely change a story by refocusing the context of the plot. In this way, the union of the images and text in a picturebook context increases the final product's complexity exponentially and is often argued to lead to a product that is greater than the sum of their parts. Most of the studies of how images affect narratives is conducted by psychologists on children. This makes sense as picturebooks are one of the most widely studied multimodal media in Education and their primary target are children for their teaching benefits during language acquisition

and development of literacy skills (Bus et al., 1995; Mol & Bus, 2011; Takacs & Bus, 2018). This is because children do not have as much experience with reading narratives and the pictures help to provide nonverbal representations of the plot-line, characters and setting (Bus et al., 1995; Mol & Bus, 2011). Additionally, there is evidence from case studies in a cross-cultural setting that children from other countries reading English picturebooks find the pictures important to understanding the narrative and learning the language (Birketveit & Rimmereide, 2017). There is further evidence from Gambrell and Jawitz in an experiment using 4th graders, that these benefits lead picturebooks to create impressive enhancements in comprehension and memory of stories as opposed to their unimodal counterparts (Gambrell & Jawitz, 1993). However, images also serve a larger purpose than language acquisition in picturebooks, as their inclusion can deepen the meaning of stories to make them more fulfilling and engaging. Research supports that children favour pictures that illustrate parts of the story, but also enhance the story or evolve the story with their inclusion (Birketveit & Rimmereide, 2017). The reading of picturebooks in a teaching context can also strengthen Visual Literacy, a concept that is becoming increasingly important in school-curriculum (Callow, 2020; O'Neil, 2011). *Visual literacy* is the study of how meaning is made in static or moving images (O'Neil, 2011). Visual literacy functions as a way to draw meaning out of pictures and is therefore the first skill learnt in reading, but is also important in adults for appreciation of art and other visual media. Some educators argue that visual literacy is important to foster at an early age as it will help learning acquisition and develop important skills in understanding all image focused media (Callow, 2020; O'Neil, 2011). Unfortunately, the same researchers argue that visual literacy has been neglected in former generations. This is a shame, as visual literacy helps people to not only understand the individual elements in a text-image interaction, but also understand the interaction between them and draw potentially endless meaning from multimodal fusion of text and images. This is an important skill in media comprehension and picturebooks function as an excellent way to help teach the skills necessary. Therefore, the increased relevance of teaching visual literacy is now making picturebooks more important than ever.

As of yet, the dominant form of illustration for generated narratives is currently *Image Retrieval*, where images are pulled from a human tailored dataset of images that closely fit text. Zakraoui et al., point out that although there are many multimedia databases created, their creation take massive amounts of human effort, require huge storage spaces and are still not capable of showing the complexity of relationships needed for generalised story illustration (Zakraoui et al., 2020). Another benefit for the field may be how image generation in a picturebook context could be a way to sidestep this need for curation, as a text prompt alone would create relevant images for the text. To my knowledge, there has been no attempt at creating entirely computer-generated picturebooks and evaluating the multimodal integration of images and text from a cross-disciplinary perspective. Therefore, I would argue that creating such books and evaluating whether they have the same benefits as human created picturebooks is a hole in the literature that needs to be filled.

After reviewing some of the possible benefits from the combination of images and text in computer generated media, one may be wondering why it is so sparsely researched. There are many instances of image and story generation, but the combination of the two are rare. Also, the papers that do combine image and text generation, often do not attempt to study the effects the two modalities have on one another during multimodal fusion. I believe one of the reasons that automatic picturebook generation using computers has yet to be widely studied, is because the currently available and publicly accessible image-generation technology has evolved so rapidly in the past year that it is the first time it is possible. Therefore, I believe that I am in a unique position to study this combination at the current juncture. For this paper, I will be using some of the best publicly available story generation architecture in combination with image generation architecture to generate picturebooks. To study the effect, I will be creating a human evaluation form and analysis that combines literature across various disciplines, the Humanities, Education, Psychology and Data-science, to understand if the combination of computer-generated images and text leads to a product that has the same benefits as human created picturebooks. I believe this is important work and hopefully one of the first of many future research projects into this rapidly evolving domain.

3

HYPOTHESIS

Images and text are both important on their own terms, but the literature argues that their combination adds more value than the components do on their own. Combining computer generated images and text is valuable as it leads to increased human-computer communication, is applicable in various learning scenarios and more engaging from an entertainment perspective. I will study if the product created through generated images and text in a picturebook context leads to a product that is greater than the sum of their parts.

This paper intends to contribute by filling a hole in the literature and highlighting that combining even current publicly available technology can lead to human-tested results that supports how stories, in the form of picturebooks, add something to textual stories that is not often explored in machine learning communities. The paper will, 1. Use current technologies available to the public to generate images and text and combine them to create picturebooks, 2. Highlight possible metrics for evaluating computer generated picturebooks, and 3. Create, administer and analyse a human evaluation form to show the possibilities of deeper understanding through the integrating of various disciplines when understanding multimodal fusion in machine generated picturebooks.

The paper will take the following format, 1. Literature Review 4, 2. Materials 10, 3. Method 14, 4. Results 18; Discussion 19; Limitations 21, 5. Future work 22 and 6. Conclusion 23.

Part II

LITERATURE REVIEW

4

INTRODUCTION TO LITERATURE REVIEW

Since there is so little research in the field of picturebook generation, the related works will briefly cover, 1. Natural language generation as a field, 2. Image generation as a field, 3. Story generation as a field, 4. Illustrated story generation in data science and their evaluations, and 5. Picturebooks and multimodal integration of image and text spanning various disciplines. I believe that this scope is necessary to get the proper breadth of what academic research is occurring and understand what the paper intends to accomplish.

5

NATURAL LANGUAGE GENERATION

This section is intended to give an overview of how and why NLG has recently evolved and the current state-of-the-art. It will first briefly cover the last few years of natural language generation, then the current state-of-the-art models, and finally give some examples of how it is used in the public sphere. The section will provide a brief overview and highlight some of the problems with the architecture that is necessary knowledge for the "Story generation" section below. As this is intended as an overview of the literature, details on the architecture used in this paper (GPT2) will be covered in more detail in the "Materials" section 10.

Natural language generation (NLG) is a subfield of Natural Language Processing (NLP). The overarching task in NLG is to create digital architecture that can reliably produce coherent, cohesive and readable text (Celikyilmaz et al., 2020). Because of its complex nature, NLG tasks have long been considered to be one of the most difficult computational tasks. The difficulties largely arise from the grammatical complexity of natural language and the extraction, simplification and transformation of the input information (Lu et al., 2018). In its application, there are many variations of this type of task, such as generating responses to questions, language translation, text summaries and story writing. Some of the previous solutions to these types of problems are in the form of Sequence-to-Sequence models (Seq2Seq), Reinforcement Learning models (RL), General Adversarial Networks (GAN) (Goodfellow et al., 2014) and Recurrent Neural Networks (RNNs) (Rumelhart et al., 1985); with LSTMs (Hochreiter & Schmidhuber, 1997) *Long Short-Term Memory* models, which are forms of

RNNs, achieving the most success until recently (Lu et al., 2018; Topal et al., 2021). One of the major problems with the previous architectures is that they lead to a *vanishing gradient problem*, or the problem that as sentences increase in length the probability of maintaining context with words currently being processed exponentially decreases the further a words is from others in the text sequence (Topal et al., 2021).

Recently, English language text-generation has seen a major boost in quality associated with the addition of more processing power and Attention Mechanisms. The current state-of-the-art in NLG are various types of Transformer Architectures which introduced novel architecture to Seq2Seq modelling by adding attention and eliminating recurrence and convolutions (Topal et al., 2021). *Transformers* are encoder-decoder based models that allow for self-attention which in turn allows for parallelism and pointwise ranking with fully connected layers improving a model's ability to extract long-range context without much additional computational impact (Fan et al., 2018). These types of models are great as they are faster to train because of parallelization and allow for positional encoding that improves scalability and transfer learning (Fotedar et al., 2020). Additionally, parallelization allows the models to sidestep the vanishing gradient problem and other long short-term memory problems of the previous state-of-the-art LSTM architectures (Topal et al., 2021; Vaswani et al., 2017). Transformer based NLG models can fine-tune downstream tasks, or re-train weights for specific types of text, and have improved the state-of-the-art in textual entailment, reading comprehension and commonsense reasoning, among other things (Fotedar et al., 2020). The newer models are currently being propelled by the application of the knowledge that adding more hyperparameters when training transformer-based architecture can lead to more human-like results (Radford et al., 2019).

The current state-of-the-art in text generation are transformer-based large scale neural models such as BERT (Devlin et al., 2018), ELMo (Peters et al., 2018) and OpenAI's GPT models (Brown et al., 2020; Radford et al., 2018, 2019). The transformer models are newer forms of Seq2Seq models, where they use text prior in a text-sequence to try and probabilistically predict what text is most likely next in a sequence. Currently the new iteration of OpenAi's transformer-based models, GPT3, is by

far the largest of these models with a massive 175 billion parameters (Brown et al., 2020). As of writing this paper, Open-AI's GPT3 API is in beta access and not readily accessible to the public. However, previous state-of-the-art models like GPT2 (The largest model containing 1775 million parameters) and BERT (The largest model containing 340 million parameters), are available to the public and have been used for a wide range of experimentation. These types of NLG models are achieving impressive results on summarization and machine translation tasks (Fotedar et al., 2020). However, in longer generation tasks there have been problems such as, "reference errors, intersentence incoherence, and a lack of fidelity to the source material" (Fotedar et al., 2020). Also, there are still some well known problems such as lexical diversity, repetition and word redundancy that still plague this type of generation.

Currently, these state-of-the-art *Application Programming Interfaces* (API's), or software intermediaries that allow multiple applications to communicate with one another, are being popularised in the public sphere through copywriting businesses that create content for websites, and videogames like *Ai-Dungeon*, that utilise this technology for profit [1]. In the academic setting, this technology is advancing rapidly with the addition of increasingly larger training sets. Now one of the biggest challenges of increasingly complex language generation, is the evaluation of the text because of the open-ended nature of many of the tasks (Celikyilmaz et al., 2020).

1 https://play.aidungeon.io

6

STORY GENERATION

This section is intended to provide an overview of how story generation has been tackled in the NLP communities and cover the state-of-the-art models. It will first relate Automatic Story Generation back to Natural Language Generation (NLG). Then briefly highlight how various researchers have tackled the problem of story generation. The section will then cover the state-of-the-art architecture, Plan-and-write, Plot Machines and Narrative Interpolation. Plan-and-write is the architecture used in this paper, and therefore only briefly referred to in this section, but can be found covered in more detail in the "Materials" section 10.

Automatic Story Generation or Controllable storytelling (Fan et al., 2018; Rashkin et al., 2020; See et al., 2019; S. Wang et al., 2020; Yao et al., 2018) is a widely studied and challenging problem within NLG. One of the tasks NLG is still struggling in, is attaining coherent narratives and remaining consistent with the input prompt. This can be referred to as *Narrative intelligence*, or the ability of human or computer entities to illustrate experience in the form of narrative (Riedl & Young, 2006). There are many types of narrative, but the most common and accepted is *Linear narrative*, which is composed of linear cause and effect events and the dominant format in traditional storytelling media such as novels and movies (Riedl & Young, 2006); this is the type of narrative we will be focusing on in this paper. It is important for linear narratives to retain thematic consistency and highlevel-plotting, which necessitates planning ahead and creativity. Neither is present in word-by-word generation, as the popular encoder-decoder systems do not inherently follow an underlying structure outside

of word-context probability. One of the ways that specifying NLG context has been tackled, and is now commonplace in NLG generation, is through a user input of a prompt that the story is then conditioned on while generating language (Fan et al., 2018). Conditioning on the prompt can make its stories more consistently grounded within the overall plot and reduce standard sequence models from drifting off-topic (Fan et al., 2018). Also, the ability to fine-tune models has allowed the user to "focus" pre-trained models for generation in particular subjects as it is redistributing weights to better emulate specific language, rather than training them from scratch.

In the past, narrative has been pre-scripted by human agents where the NLG is only used within the human tailored story points (Riedl & Young, 2006). This has been relatively successful in creating narratives, but it abuses the human agent's narrative intelligence in the interaction to craft the stories and limits the control of the NLG in creativity and scope. To solve this problem, some papers like, *Hierarchical Neural Story Generation*, have used a fusion model; first introduced by (Sriram et al., 2017), where a second model trained on and focusing on the link between the prompt and story keeps the standard language model on track (Fan et al., 2018). The advent of models such as fusion models aim to help increase modelling of long-range context within the story and the relevance of the overall story to the given prompt (See et al., 2019). *Fusion models* consist of two separate neural networks that communicate together, where one model generates overall structure and the other writes the content of the story (See et al., 2019). Several papers have taken the approach of having two models, where the first generates the plot structure in order to generate a high-level structure from a prompt, whereafter the Seq2Seq systems follow the generated premise (Fan et al., 2018). This type of generation has taken two forms in the state-of-the-art models, controlling the plot points or controlling the ending (Alabdulkarim et al., 2021). There are 3 state-of-the-art models that utilise the two model structures, Plan-and-Write (Yao et al., 2018), Plotmachines (Rashkin et al., 2020) and Narrative Interpolation (S. Wang et al., 2020). I will briefly cover Narrative Interpolation and Plotmachines below. Since Plan-and-write is used as the automatic storytelling architecture in this paper, it will be covered in further detail in the "Materials" section 10 below.

Plotmachines is a storytelling transformer architecture, formulating story generation as a story-guideline through outline-conditioned generation (Rashkin et al., 2020). This idea is based on the rough sketches of stories created by writers before writing the words of their stories. The plot outline is made up of rough un-ordered bullet points extracted with phrase extraction using *Rapid Automatic Keyword Extraction* (RAKE), which is a widely used keyword extraction algorithm (Rose et al., 2010). The un-ordered bullet points are then intertwined to make a story using parallel attention modules, Plotmachines and GPT2 (Rashkin et al., 2020). According to their human evaluation metrics, *Outline Utilization*, or comparing how paragraphs utilise the outline, *Narrative Flow*, how repetitive paragraphs are, and *Ordering*, deciphering order of textual paragraphs, the participants preferred Plotmachines' stories to their baseline. Also, their quantitative analysis found that compared to baseline rankings using traditional frameworks of GPT2, Fusion, and Plan-and-write, Plotmachines produced tighter narratives (Rashkin et al., 2020).

Narrative Interpolation is a form of ending-guiding story generation, which creates a beginning sentence and an ending guide, and then forms sentences in between in an attempt to form a coherent story (S. Wang et al., 2020). The wrapped model incrementally takes sentence-pairs and generates a sentence in-between two sentences. This is done through conditioning on the beginning and ending sentences and generating the text using GPT2 (S. Wang et al., 2020). Their methodology is that the ending is the most important part of the story, so first generating definite endings and continuously conditioning and generating towards the definite ending will make for good stories. Additionally, they use a pre-trained RoBERTa coherence ranking module (Liu et al., 2019) to rank generations and reject bad generations (S. Wang et al., 2020). They used a human evaluation with 3 metrics for textual story quality, *Coherence*, whether a story is logically consistent; *Faithfulness*, whether a story is faithful to the beginning and ending, and *Overall preference*, whether the human judge interprets and prefers the Narrative Interpolation sentence structure as a story (S. Wang et al., 2020). According to their metrics, the users preferred their generations as "stories" over the current state-of-the-art, GPT2 and Plan-and-Write (S. Wang et al., 2020).

7

IMAGE GENERATION

This section is intended to give a brief overview of image generation. It will begin by very briefly covering past architecture and then introduce and briefly describe how the currently used General Adversarial Networks (GAN) work as they are important for understanding the image generation architecture in the paper. Then it will briefly cover text-to-image generation and finally cover the current state-of-the-art in *zero-short image generation*, referring to generation of images outside of the training set. BigGAN and CLIP are used in the image generation within this paper and will therefore be referred to briefly but covered in more detail in their corresponding sections in "Materials" 10 below.

Image generation is a very difficult machine learning task. Images are complex and need to be of high quality to be distinguishable by humans. It is only recently that image generation is good enough to represent a wide variety of scenes. Over the past few years image generation has taken a huge step forward in quality with the advent of and popularisation of General Adversarial Networks (GAN). GANs were first proposed by Goodfellow., et al, in 2014 (Goodfellow et al., 2014). These types of models took a huge step forward from the previous state-of-the-art models, FVBNs (Fully Visible Belief Networks) first proposed by (B. Frey, 1998; B. J. Frey et al., 1996) and VAE (Variational Autoencoder) first proposed by (Kingma & Welling, 2013; Rezende et al., 2014), as they tend to be more realistic and less blurry than their competitors (Esfahani & Latifi, 2019). *GAN* architecture is based on the idea that there is a competition between two systems where a generative model tries to

"fool" an adversarial network (Goodfellow et al., 2014). The adversarial network, or discriminator network's task is to determine if the sample presented by the generative model is from the model or data distribution (Goodfellow et al., 2014). These GANs were then engineered to both be multilayer networks made up of two multilayer perceptrons, or feedforward neural networks, where neural nodes do not form a cycle. With this method, they could be trained through *backpropagation*, or backward propagation of errors that uses gradient descent for teaching feedforward neural networks, and dropout algorithms (Goodfellow et al., 2014).

Within computer vision, *Text-to-Image Synthesis* is an important aspect that is currently dominated by GAN networks, "the task of text to image generation usually means translating text in the form of single-sentence descriptions directly into prediction of image pixels" (Esfahani & Latifi, 2019). This problem is however multimodal, as there can be many different types of pictures generated from a single sentence. There are many different possible combinations of pixels that will show the same thing equally well. For example, searching "a star in the sky" on Google-Images will come up with many results that all equally well show the prompt. S Reed et al., were the first to properly show a *GAN-CLS* model, a method of using GANs that are designed to generate images from input text descriptions, on a bird dataset that could produce relatively realistic 64x64 pixel images of birds and flowers (Reed et al., 2016). This type of task is based on differentiating 3 types of images, real images with real text, synthetic images with arbitrary text and real images with fake text (Reed et al., 2016). This trains the discriminator to distinguish between images with matching text, images that will not match any text and images the discriminator must learn to score as fake (Reed et al., 2016). This technique is now widely used in training GAN networks for realistic image generation. Generated images are usually evaluated on the *Inception score* or IS (Salimans et al., 2016), which is a score devised as an image classification neural network, and *Fréchet Inception Distance* or FID (Heusel et al., 2017), which calculates the difference between a real image and a generated image in the same domain.

There are too many state-of-the-art GANs trained for specific purposes, like generating human faces, cars or style-transfer, to cover in this paper. We will focus on GANs with large training sets for their ability to better generalise image output. One of the best publicly available pre-trained models for more generalised image generation using a text prompt is the BigGAN architecture (Brock et al., 2019). The advent of these networks saw a huge spike in the addition of large amounts of training data and the BigGAN architecture is trained on *Imagenet*, which is a massive sized human-annotated training set with over 14 million images with more than 100,000 synsets, where 80,000+ are nouns (Deng et al., 2009b) [1]. BigGAN is used in the generation of images for this paper and will therefore be covered in more detail in the "Materials" section 10 of the paper below.

The current state-of-the-art in zero-shot image generation is currently OpenAi's *Dall-E* architecture (Ramesh, Pavlov, Goh, et al., 2021), based on Image GPT (Chen et al., 2020), that runs on CLIP for image ranking and an image trained GPT3 type architecture for image generation (Radford, Kim, et al., 2021). However, Dall-E is currently not publicly available. It is worth mentioning its existence, but this paper will not go into further detail on the specifics of its performance or architecture as it is yet to be released publicly. However, OpenAI did release a smaller version of the full CLIP ranking algorithm which is used in various image generation architecture and in this paper. CLIP will be discussed further in the "Materials" section 10 later in the paper.

1 https://www.image-net.org/index.php

8

ILLUSTRATED STORY GENERATION AND EVALUATIONS IN DATA SCIENCE

This section will highlight some papers that have attempted generation of stories with accompanying images. The section will briefly highlight how they tackled the problem and also how they evaluated their results. As an important part of this paper is evaluating the multimodal fusion of generated text and image media, to understand how other people in the field evaluate their results is important. The metrics used in this paper, Specificity, Relevancy, Narrative consistency and Multimodal fusion, will all be discussed in more detail and defined more clearly within the paper in the "Methods" section 14. Retrieval papers often evaluate picture integration in stories as needing Coherence, Narrative Consistency/Relevance, and Concreteness.

There is a relatively large amount of research on human-created images and NLG interaction, such as caption generation from an image, story generation from an image, image-word substitution, and a wide range of other interactions. To my knowledge, there are very few attempts at creating a high-structure story with both generated image and text integration. The majority of research of this kind focuses on image retrieval, for example in papers like: (Gonzalez-Rico & Fuentes-Pineda, 2018; Huang et al., 2016; Nag Chowdhury et al., 2020) or story generation for retrieved images, for example in papers like: (Shin et al., 2018).

However, combining automatic story generation with image generation is rare. Some researchers have tried something similar before in an academic setting, like the paper, *Storytelling AI: A Generative*

Approach to Story Narration, where Fotedar, et al., explored generating a story using BigGAN and GPT2 with relative success (Fotedar et al., 2020). Their goal was the creation of the architecture and not its evaluation, but they do conclude that although the technology is there to supplement automatically generated stories with generated images, there is still human intervention required for filtering in its current state (Fotedar et al., 2020). Another paper that attempts this is, *Generating Images from Arabic Story-Text using Scene Graph* (Zakraoui et al., 2020). Their goal with the paper was to imitate the imaginary images visualised by children when engaging with a story to promote learning (Zakraoui et al., 2020). They do this by using a graph-based semantic representation system where relationships are represented as networks between nodes to simplify stories as graphs and using simple extraction of objects, their attributes and relationships to overcome subjectivity and diversity issues of complex expressions (Zakraoui et al., 2020). In other words, they reduce complex paragraphs to simple sentences and create a, "flexible and dynamic image repository based on scene graph" (Zakraoui et al., 2020). Their idea behind this type of node graph through the use of scene graphs is to use relationships of objects within stories as a way to retain relevancy and concreteness in their story generation. Their main concern is the creation of a system that is both flexible and dynamic. Their results were able to express clearly amounts of objects and their relationship to one another on various backgrounds. However, their ability to generalise their current model is difficult as it is based only on 96 categories of animals and natural backgrounds. This of course being an important part of varied and creative stories, needs to be expressed in further research if the breadth of the types of stories that can be generated is to be considered an important aspect. The paper was later expanded and evaluated according to how the characters mentioned in the input sentence were displayed in the generated images (Zakraoui et al., 2021). Here they create a metric called the *Story Character Classification Accuracy*, which is measured globally, and measures whether the characters in the input are represented in the output (Zakraoui et al., 2021). They found that with simple sentences, the images were able to represent the text well, but as sentences got more complex, images showed a fair but worse representation of the text (Zakraoui et al., 2021). This again reflects the complexity of

language and the generalisation problem. The paper, *Visual Storytelling*, introduced the dataset for sequential vision-to-language storytelling (Huang et al., 2016). They were focusing on object and concrete scenes based on cause and effect relationships that visualise events as they occur and as they change (Huang et al., 2016). This would lead each image to be related to the last in some way, and in turn create a product where each image is contextualised in its sequence of surrounding images. The paper argues that moving from pictures in isolation to images in sequence with a context, allows for more concrete visualisation of stories (Huang et al., 2016).

There are two components that are said to be important, Narrative Consistency between clauses in the generated story and Relevancy between generated story and imagery (Sapkale & Lukin, 2020). In this way, it is important for the generated story to be grounded in all the images presented with the story (Sapkale & Lukin, 2020). *Narrative consistency* is the degree a story is coherent and motivations of characters is justified within the logic of the storyworld (Sapkale & Lukin, 2020). The integration of pictures and images has been attempted through the creation of captions from images to fall in line with story pictures (Sapkale & Lukin, 2020). This is a way to retain the Relevancy of the pictures to the story. Stories can be seen as containing a Fabula and Sujet (Sapkale & Lukin, 2020). The *Fabula* functions as the "raw materials", or the objective events of the story, while the *Sujet* functions as how the story is shown to the audience (Sapkale & Lukin, 2020). In this way, the Fabula needs some consistency and logic behind it, and the Sujet can be seen as the motivation or focus of why the story should be told. In their work, (Sapkale & Lukin, 2020), they used the visual elements as the Fabula from which to extract the story, and the Sujet as the writing or what they want to communicate from the pictures. They used image sequences from the VIST and Visual Genome datasets constructed from Flickr for images, and the *Atlas of Machine Commonsense* (ATOMIC), a logical cause and effect based NLP dataset for creating plot points in the narrative. Their evaluation of the stories relied on *Relevancy* between images and story, *Concreteness*, or what the images convey as opposed to general description, and *Consistency*, which focuses on story coherency (Sapkale & Lukin, 2020). In their paper, *Analyzing Image-Text Relations for Semantic Media Adaptation and Personalization*, Hughes

et al., discuss a metric of understanding image-text relations through images being interpreted on a scale from General to Specific (Hughes et al., 2007). The distinction comes from how the images are interpreted in regards to pictures, as *Generalized* or abstract images can be interpreted in a wide variety of ways, *Specific* images are more representative of the accompanying textual information. The focus of this metric is the extraction of low-level information in the reader (Hughes et al., 2007). In their analysis when using this metric, Hughes et al., conclude that humans can predict some of the meaning of text through low-level image features (Hughes et al., 2007). From their report they also conclude that when generating multimodal media, text-image relations and low-level feature need to be taken into account and more specific imagery, for example sharper and more in focus characters will improve this type of generation (Hughes et al., 2007). They further argue that taking a multimodal approach into generation is a key strategy in attempting to close the semantic gap of text-image relations in future research (Hughes et al., 2007). Although these papers create metrics to measure how images relate to text on a surface level, they do very little in attempting to understand why the image and text work together and how they work together.

9

PICTUREBOOKS AND MULTIMODALITY

This section is intended to provide a cross-disciplinary look at picturebooks and their views of multimodal integration between images and text. This section will cover a scope of disciplines, Psychology, Humanities and Education. First, the section will discuss the definition of and spelling of picturebooks. Then this section is intended to provide literature to justify the use of the Multimodal Fusion metric in the evaluation; further detail can be found in the "Methods" section 14. The section is also intended to provide adequate vocabulary to discuss different types of multimodal fusion and what might be occurring in the reader when interacting with picturebooks. This will be further discussed in the "Discussion" section 19.

One of the most applied forms of multimodal fusion are the still widely used picturebooks. However, there is much debate in the community about what to call various books with image and text interactions (Nikolajeva & Scott, 2013). As the purpose of this paper is to generate image and text narratives and examine their worth as multimodal entities when perceived by humans, the purpose is not to engage in this debate and for brevity's sake I will accept one of the most prevalent definitions of various image-text books. For this paper, I will use the categorisation by Torben Gregersen, who makes the distinction between, 1. *Exhibit book*, a picture dictionary that has no narrative, 2. *The picture narrative*, which has few or no words but follows a narrative, 3. *Picturebook*, where pictures and words are seen as equally important in the narrative, and 4. *The Illustrated book*, where the images and text are independent from one another (Gregersen, 1974; Nikolajeva & Scott, 2013). Of these

categorisations, for the multimodal purpose of integration of image and text generated by computers, I will focus on picturebooks. Picturebooks can be spelled in many different ways (picture book, picture-book, picturebook, etc), but for this paper I will spell it with the words combined to signal the importance of the two elements together in their unity.

Picturebooks often use pictures as a way to carry a large amount of the narrative responsibility, and often the meaning emerges from the combination and interplay between words and images (Birketveit & Rimmereide, 2017). The interplay between words and pictures is complex and has been studied and condensed by Nikojavera and Scott to have 4 different possible interactions (Nikolajeva & Scott, 2013). They list them as, 1. *Counterpointing*, where the pictures tell a different story than the text, 2. *Complementary*, where the pictures and text fill in gaps in the narrative, 3. *Symmetrical*, where the text and pictures tell the same story in parallel, and 4. *Enhancing*, where the interaction between pictures and text together make up something that is greater than the sum of their parts (Nikolajeva & Scott, 2013). This type of relationship can be referred to as Multimodal fusion. *Multimodal fusion* is the integration of various modalities to combine various types of information and enhance an outcome of a task as the integration of various modalities can lead to a better outcome than unimodal information (Shareha et al., 2009).

Picturebook scholars imply that there is a synthesis that occurs between images and text when combined as they enhance one another, leading to an entity which is greater than the sum of their parts (Wolfenbarger & Sipe, 2007). This relationship is brought out through the integration of separate elements into a new whole and multimodal experience. The images and text work tightly together to convey temporal and spatial information that is present in both or one of the two modalities (Wolfenbarger & Sipe, 2007). From this, scholars argue that even if information is representative of the same thing, the two modalities are so different that they never represent the exact same information (Wolfenbarger & Sipe, 2007). This is argued to be one of the strengths of multimodal media, and one of the reasons that their combination improves both text and image. This comes from the play that occurs in the reader's exploration of relationships between words and pictures (Wolfenbarger &

Sipe, 2007). The interaction that occurs in these types of picturebooks has been described by Lewis as *Interanimation*, where the words are "pulled" through the text of the book and where the verbal text draws the attention of the reader to particular parts of images, and vice versa, where images draw attention to particular words, changes perception of words or creates perceived words in the mind of the reader not in the book (Lewis, 2001; Wolfenbarger & Sipe, 2007). An example of this could be visual description of a character being gathered from a visual representation of that character within an image that would not otherwise have been described in the text. This pulling words through pictures, shows the ways that pictures can affect the holistic interpretation of a story in the absence of text. This direct engagement helps readers to become active participants in picturebooks, by making sense of a story and a world through the scope of their own interpretation of the images as opposed to a passive reader that is described the whole story entirely through textual means (Wolfenbarger & Sipe, 2007). As the words are pulled through the images, the reader will encompass their own experiences and interpretations onto what meaning they are pulling from the picturebooks and what they are focusing on when synthesising the modalities into a single entity during multimodal fusion. This process is a type of self-assessment as well, as people will integrate their own analysis which is specific to their world view and experiences, and form a product more relevant to themselves. In this way, this process can also function as a mode of self-analysis and contemplation as meaning is formed (Wolfenbarger & Sipe, 2007). This type of visual literacy has been shown to be improved through more engagement with picturebooks, and its cousins like comics, and can almost be described as a bilingual ability of reading literacies (Wolfenbarger & Sipe, 2007). When engaging with the media in this way, there are several simultaneous inputs of stimuli that create a perception in the reader. In this way, because of their ability of integration with one's own experience in the multimodal combination of images and text, picturebooks can lead to self-assessment and understanding. The multimodal fusion creates layered meaning that in contrast creates a whole greater than the sum of its parts, that is important in application of imagination, self-reflection and understanding of art.

Part of this cohesion is created in the engagement of the form of the picturebooks with the overall context that it is represented in. A furthering of this context is the *Peritext*, or the physical features that encompass the story, such as the back and front cover, the jacket or even the dust that gathers on a book that has not been read in a while (Wolfenbarger & Sipe, 2007). This contextualises the experience even further in the engagement with the media, as the changing in colour throughout a book can for example indicate the passage of time. This shows that the context in which picturebooks are presented can further engage the imagination and engagement of the reader to pull meaning from picturebooks.

Part III

MATERIALS

10

INTRODUCTION TO MATERIALS

Within the “Materials” sections I will first cover the architecture that I used for the paper, the text-generation architecture using higher-order planning, Plan-and-write with GPT2, and the image generation architecture Bigsleep, which is a synthesis of BigGAN and CLIP. I will also outline the evaluation form. For replication purposes I will link where to find all the corresponding materials used and briefly describe how the more technical materials work that have not been covered in the ”Literature Review” section 4 above.

11

PLAN-AND-WRITE

Plan-and-write is a textual story generation architecture that uses two transformers akin to the fusion models. First a storyline of 5 linear "events" is generated as keywords, and then NLG is applied using the storyline as generation prompts to generate a 5 sentence story; an example can be seen in Fig. 1 below. For the NLG, I used GPT2. The Plan-and-write with GPT2 github page can be found in a link in the footnotes [1]. Plan-and-write has several variants, the one used for this project was a GPT2 variant not yet released. It was an exploratory extension created while exploring the BART-variant proposed in the, *Content Planning for Neural Story Generation with Aristotelian Rescoring*, paper (Goldfarb-Tarrant et al., 2020).

Title (Given)	The Bike Accident
Storyline (Extracted)	Carrie → bike → sneak → nervous → leg
Story (Human Written)	Carrie had just learned how to ride a bike. She didn't have a bike of her own. Carrie would sneak rides on her sister's bike. She got nervous on a hill and crashed into a wall. The bike frame bent and Carrie got a deep gash on her leg.

Figure 1: This shows how the "Title" is used to extract an ordered storyline of a story written by a human. Retrieved from the Plan-and-Write journal (Yao et al., 2018).

1 https://github.com/PlusLabNLP/StoryGenerationDemo

The higher order planning of Plan-and-write borrows heavily from poetry generation and conversational systems by using a sequence of words to approximate story plot outlines (Yao et al., 2018). The results show that this system creates more coherent and on topic stories than its predecessors (Yao et al., 2018). The problem formulation retrieved from the Plan-and-write journal can be found in Fig. 2 below. The title is first pulled and a storyline is generated from it. The storyline-points are used as input for the story generation system. For this project we use the Static Schema, the distinction between the two schemas can be found in further detail below.

Problem Formulation

Input: A title $\mathbf{t} = \{t_1, t_2, ..., t_n\}$ is given to the system to constrain writing, where t_i is the i-th word in the title.

Output: The system generates a story $\mathbf{s} = \{s_1, s_2, ..., s_m\}$ based on a title, where s_i denotes a sentence in the story.

Storyline: The system plans a storyline $\mathbf{l} = \{l_1, l_2, ..., l_m\}$ as an intermediate step to represent the plot of a story. We use a sequence of words to represent a storyline, therefore, l_i denotes a word in a storyline.

Given a title, the plan-and-write framework always plans a storyline. We explore two variations of this framework: the *dynamic* and the static schema.

Figure 2: The "Problem Formulation" screenshot is retrieved from the Plan-and-write journal (Yao et al., 2018)

To extract the storylines, the *Rapid Automatic Keyword Extraction* or RAKE algorithms (first presented in (Rose et al., 2010)) are used to find the weights of how important words are in the word sequence and extract the words with the highest weights to form the outline (Yao et al., 2018). The storyline is generated via context of the previous word in the story and the title (Yao et al., 2018). It is introduced as a content-introducing generation problem, where the title is used along with the last word as additional context (Yao et al., 2018). The context is retrieved through the use of the content-introducing method where the context is encoded into a vector as a bidirectional gated recurrent unit (BiGRU), incorporating auxiliary information (first introduced in (Yao et al., 2017)) (Yao et al., 2018). The generation has been proposed with two variants. The *Dynamic system* draws

from the generation into the planning and the *Static system* plans the storyline to completion first before the story generation takes place (Yao et al., 2018). An illustration of the two schemas can be found in Fig.3 below. For the Static generation (used in this paper), which outperformed the Dynamic variant in their metrics of Coherence and Relevancy, the storyline functions similarly to an outline created by writers before starting their story (Yao et al., 2018). This can limit some flexibility in the writing but according to the paper boosts coherence and relevance to the provided title as it plans, "what happens next" as a contextualisation of a story. It also writes a proper beginning, middle and end. This can be seen as an illustration in Fig. 3 below.

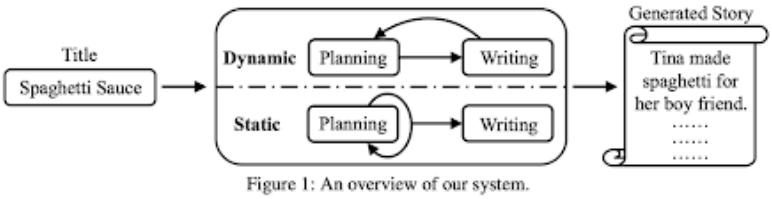


Figure 3: The screenshot above is an overview of how the system generates stories from the outline (Yao et al., 2018)

The storyline planner has been trained on the *ROCStories* corpus (Mostafazadeh et al., 2016). The corpus contains 98,162 "commonsense stories" with a mean of 50 words per stories and a total vocabulary size of 33,215 words (Mostafazadeh et al., 2016). Each story is 5 sentences long with a title that leads directly into each of them (Yao et al., 2018).

11.1 GPT2

Despite being about 2 years old, GPT2 or, *Generative Pre-trained Transformer 2,* is still one of the best publicly available and free to use generative language models. The version of GPT2 I used for the story generation was the pre-trained medium or 355M hyperparameter model from Huggingface library [2]. The model used was not fine-tuned on any additional data. The GPT2 model, put simply,

2 https://huggingface.co

was trained on a 40GB dataset called *WebText*, made up of data from 8 million web pages of wild data across the internet and is intended on predicting the next word given parts of the previously generated words (Radford et al., 2018). To ensure quality, the web scrape of the training data emphasised websites that were curated/filtered by humans (Radford et al., 2018). It utilises *Byte Pair Encodings* (BPE), which often operates on Unicode points and functions as a middle ground between character level and word level tokens, meaning some tokens will be whole words while others only parts of words (Radford et al., 2018). GPT2 uses interconnected stacks of transformer decoder blocks, which are used to convert vectors into intelligible language, with feed-forward or non cyclical neural networks and masked self-attention (Radford et al., 2018); this can be seen illustrated in Fig. 4 below. *Attention* is a mechanism based on human attention that enhances important parts of sequence text data, *Masked Self-Attention* means that it can only "pay-attention" to present and past BPEs whilst generating and will not reinterpret prior ones already generated (Radford et al., 2018).

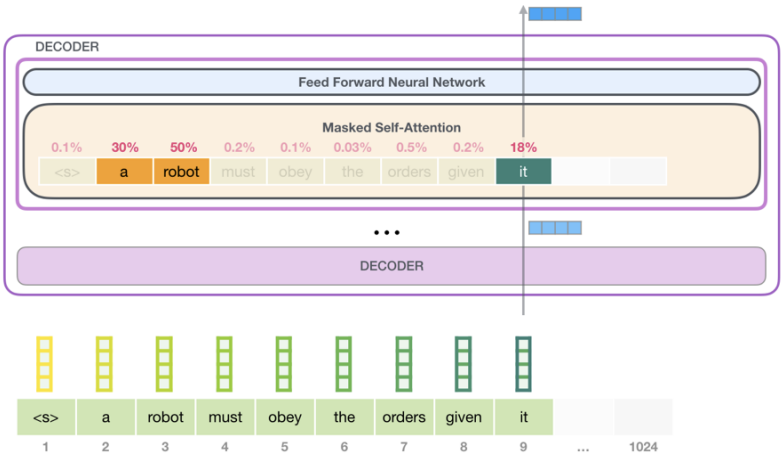


Figure 4: Retrieved from, (Alammar, August 12, 2019)

In regards to input encoding, trained GPT2 models utilise token embedding and positional encoding matrices that are intended to recognise vocabulary and where to positionally use it (Yao et al., 2018). The self-attention mechanisms of the decoders each help to understand the context of a word, like the word "he," to try and match with the most certainty which character it could be referring to

(“Tom”). Massive amounts of research has been conducted on GPT2 after its release, and it has been successfully applied in similar projects to this one by, (See et al., 2019; S. Wang et al., 2020).

12

BIGSLEEP

Bigsleep is an image generation model created by *Advanoun*. The author has permitted the architecture to be used for any purposes, even for financial gain, as long as their twitter is cited [1]. Bigsleep has been used in several online projects, and can be found on this github page [2]. I used the original notebook as opposed to the simple version as it was the only version available at the time and it allows for more control when generating images. The architecture is running on Google Colab using GPUs. I would restart the kernel until I got either T4s or P100's as the increased RAM made the generations much faster and smoother. The architecture runs on BigGAN and the ViT-B/32V CLIP model, which is covered in more detail below. Bigsleep generates pictures in 512x512 pixel resolution. The BigGAN takes a noise vector as input and generates images as output, CLIP takes in text as its input and outputs the image's features as text; this similarity can be represented as the cosine similarly of the learnt feature vector. In this way, BigGAN functions as the generator and CLIP the "guide." The generation stops when CLIP rates that it is the same match as the input description. Biglseep begins with an image generated by BigGAN without a prompt as its "template." This serves instead of a blank canvas for generation in an attempt to kickstart and speed up the generation with an image resembling something tangible instead of random noise. You can choose a seed to decide what type of image you want to start with, the amount of possible generations is not affected and in my experience they are mostly dogs and make little difference in the final result. This makes sense as dogs have one

1 https://twitter.com/advadnoun

2 https://github.com/lucidrains/big-sleep

of the largest representation in the training set of BigGAN, as is discussed in the journal (Brock et al., 2019). The entire story sentences from the Plan-and-write GPT2 model were individually and directly written into the BigSleep model for accompanying image generation. At times several pictures were generated from the same sentence according to a "Image Generation Checklist" which is discussed further in the "Methods" section 14 and can be found as a screenshot in the "Appendix" 24.

12.1 BIGGAN

As the theory behind GAN networks is already covered in the "Literature Review" section 4, I will only briefly discuss the details of BigGAN; for summarising purposes a simplistic graphic of the GAN architecture can be found in Fig. 5 below. BigGAN is a general adversarial network that improved on the prior state-of-the-art by scaling up the amount of data massively and improving performance both in the Inception score (from 52.52 to 166.5) and Fréchet Inception Distance (from 18.65 to 7.4), which are two image classification scores commonly used in evaluating image generation, this was done using ImageNet at 128x128 resolution (Brock et al., 2019). BigGAN improved the past GAN image generation architecture by adding self-attention to their model, called *SAGAN*, which made it possible to introduce an attention-map and allow the discriminator and generator to focus on different parts of the image (Brock et al., 2019). This in combination with more model parameters and larger batch sizes when training, increased size and quality of the generated images dramatically (Brock et al., 2019). The BigGAN model used for generating the pictures in BigSleep is the Huggingface Pytorch pretrained 512x512 resolution model with 12-layer, 768-hidden, 12-heads, and 110M parameters [3]. This version of BigGAN is trained on ImageNet and is intended to generate realistic images through the use of an adversarial system architecture. However, it can generate 1000 types of images better

3 https://github.com/huggingface/pytorch-pretrained-BigGAN/bl ob/master/README.md

than other images on average as there are more instances of these contained in the training set. The List of the 1000 things can be found in the footnotes: [4].

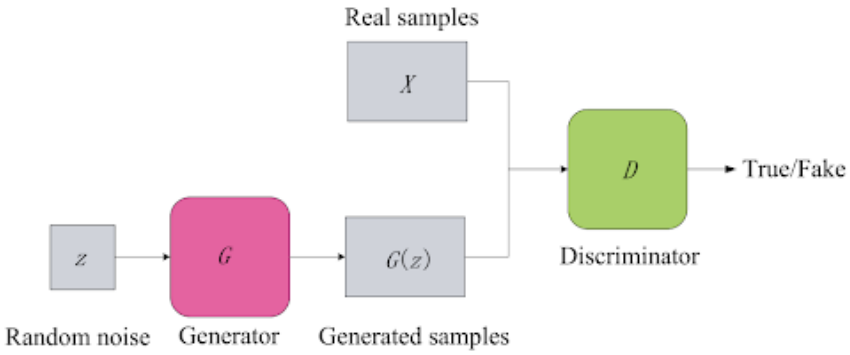


Figure 5: Briefly summarises how GAN architecture works. Retrieved from (L. Wang et al., 2020)

12.2 CLIP

One of the newest additions to publicly available image generation systems, and one of the biggest improvements, is CLIP. CLIP or *Contrastive Language–Image Pre-training*, is a neural network that is intended to perform zero-shot transfer, natural language supervision, and multimodal learning [5] (Radford, Kim, et al., 2021). CLIP therefore functions as a state-of-the-art zero-shot image-text pair classifier that distinguishes images from each other (Radford, Kim, et al., 2021). CLIP was trained on image-text pairs pulled from the internet at large and therefore less expensive and more varied in its training than the previously used ImageNet which is limited to 1000 categories (Radford, Kim, et al., 2021). CLIP will be given the characteristics of a visual concept and using its text-encoder create its own classifier that according to their research is often competitive with most supervised model variants (Radford, Kim, et al., 2021). In other words, CLIP trained on raw-text associated with images in the wild created from internet users around the web, and therefore had a much larger and varied training set than prior state-of-the-art. It is impressive because it can be generalised in its recognition

4 https://gist.github.com/yrevar/942d3a0ac09ec9e5eb3a

5 https://openai.com/blog/clip/

of images by not being optimised for a single task (“Zero-shot”). It reaches this performance by training a text encoder and image encoder and then computing the scaled cosine similarity matrix of the text features and image features together by minimising the scale of the diagonal values to match corresponding features; as illustrated in Fig. 6 below (Radford, Kim, et al., 2021). CLIP can be used to predict either specific text-sequences or utilise *Bag-of-words Contrastive*, where the order of the words does not matter in prediction. The model used for this paper is the ViT-B/32V, but the best version is currently the unreleased ViT-L/14-336px model (Radford, Kim, et al., 2021).

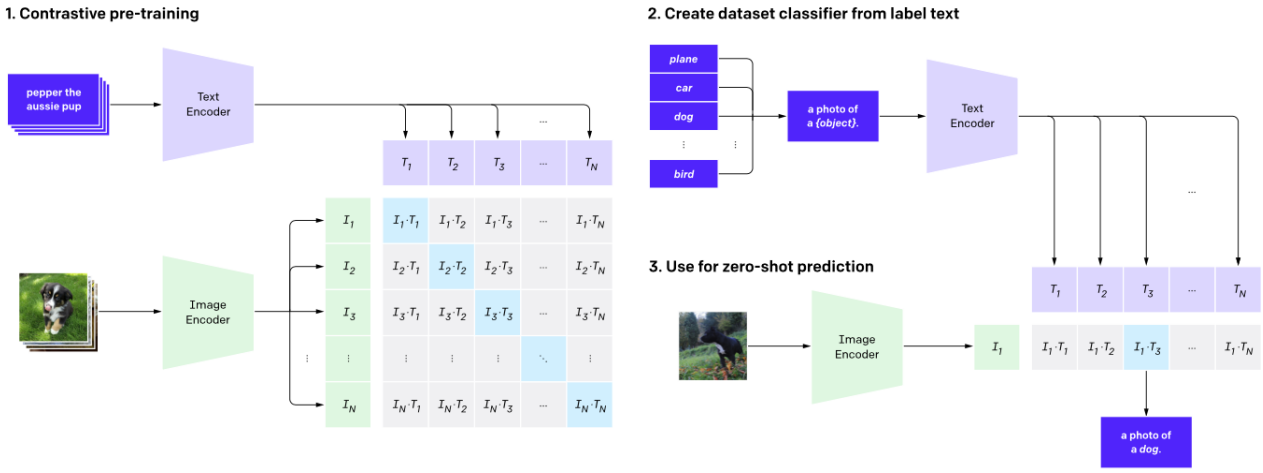


Figure 6: Fig. 6, briefly shows how the text and image encoding works to calculate the cosine similarity matrix of the text and image features. Retrieved from (Radford, Sutskever, et al., 2021)

13

PICTUREBOOK EVALUATION FROM

The evaluation was created using Jotform[1], which is a free-to-use online form creation tool with many customisation options and cross-device support. The evaluation is called, "*Picturebook Story Evaluation Form,*" and a clone of the original form can be found as a footnote [2]. After completion, the evaluation contains 12 pages in all, excluding the "Thank You" page after submission. The first page is a summary page of what the evaluation is intended to accomplish, an indicator of how long it will take (10-15 minutes) and example of an image-sentence pair, and words of encouragement intended to reduce pressure in participants (the form explicitly states that the evaluation is of the computer and not of the participants themselves). It also contains a progress-bar at the top to give a sense of how long participants are through the evaluation. The form itself is entirely anonymous and no information is gathered besides participant's answers. The evaluation form contains a total of 4 Tasks, each with various questions. The evaluation contains a total of 23 questions, with 13 unique computer generated images excluding the example image, and 2 complete image generated stories. Before each task there is a short description of what the participants will have to do. The questions are varied and involve multiple-choice, image-picker, short-answer, scale-rating and optional long-answer questions. The "Maybe" category was created to provide and in-between answer if the participants did not find the "Yes" or "No" categories representative of their answers, and an additional "Other"

1 https://eu.jotform.com/myforms/?

2 https://form.jotform.com/211254776600351

option could be filled out with an accompanying short-answer question if participants wanted to write their own answer. Most questions, aside from the ones that are optional, have a red asterisk to indicate they are required when filling out the evaluation so one cannot skip them. This was a measure to avoid incomplete evaluation data. For the purpose of easier navigation I have named the various tasks and sub-tasks for easier reference. They will be listed on the next page in the order they appear in the evaluation with a short description (Each question can also be found in the "Appendix" 24 in larger size and better quality):

Task1, *Non-Narrative Task*, this task has 3 questions for 3 images that are in a non-narrative context;

Figure 7: This shows Task 1 (Q1.1,Q1.2 and Q1.3) without the header explaining the task. This will also be referred to as, “The non-narrative task” in the paper.

Task 2, *Matching-task*, this is a task of 5 questions where images from the generated picturebook "The new job," are scrambled and the participants need to match a textual story to their matching image;

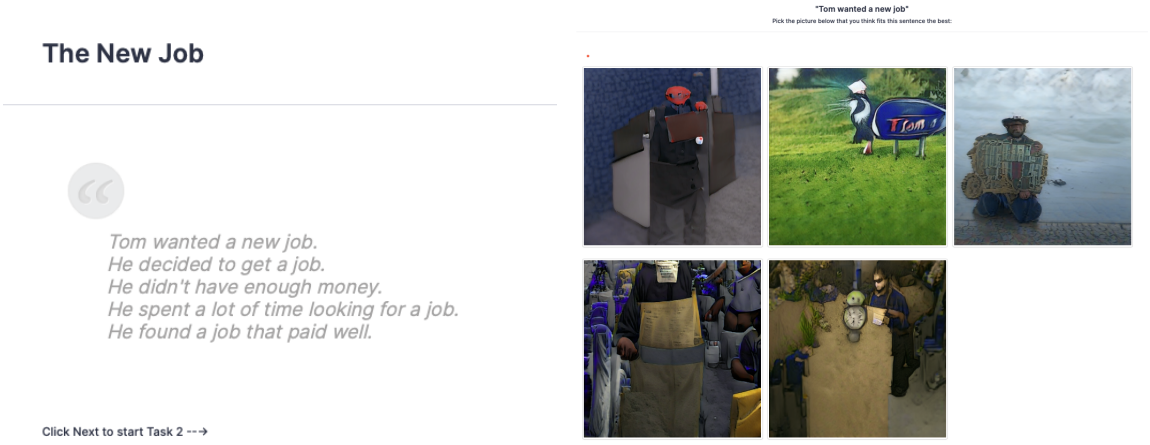


Figure 8: This shows the textual story presented to the participants (left), along with the first of the 5 slides that the participants match to each story-sentence (right). This is Task 2 in the evaluation and will also be referred to as, "The matching task" in the paper.

Task 2 Questions, *"The new job" evaluation task*, this task involves answering questions about "The new job" story and the previous matching task;

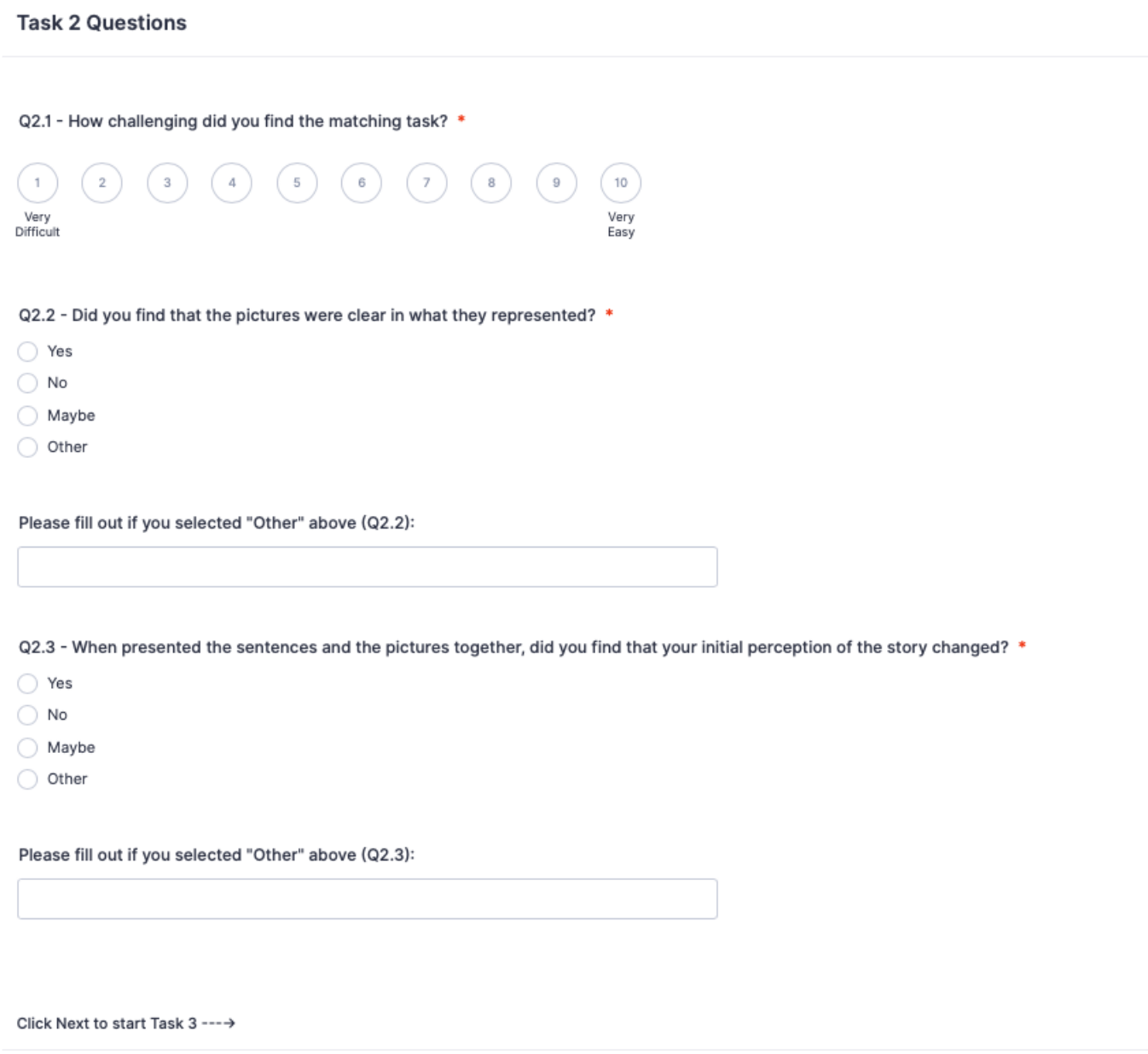


Figure 9: This shows Task 2 questions which are asked after the matching task. It will also be referred to in this paper as, " 'The new job' evaluation task".

Task 3, *The story writing task*, for this task participants are given the picturebook "The scary ghost" without the text and need to fill in 5 short answer questions to write their own story from the images;

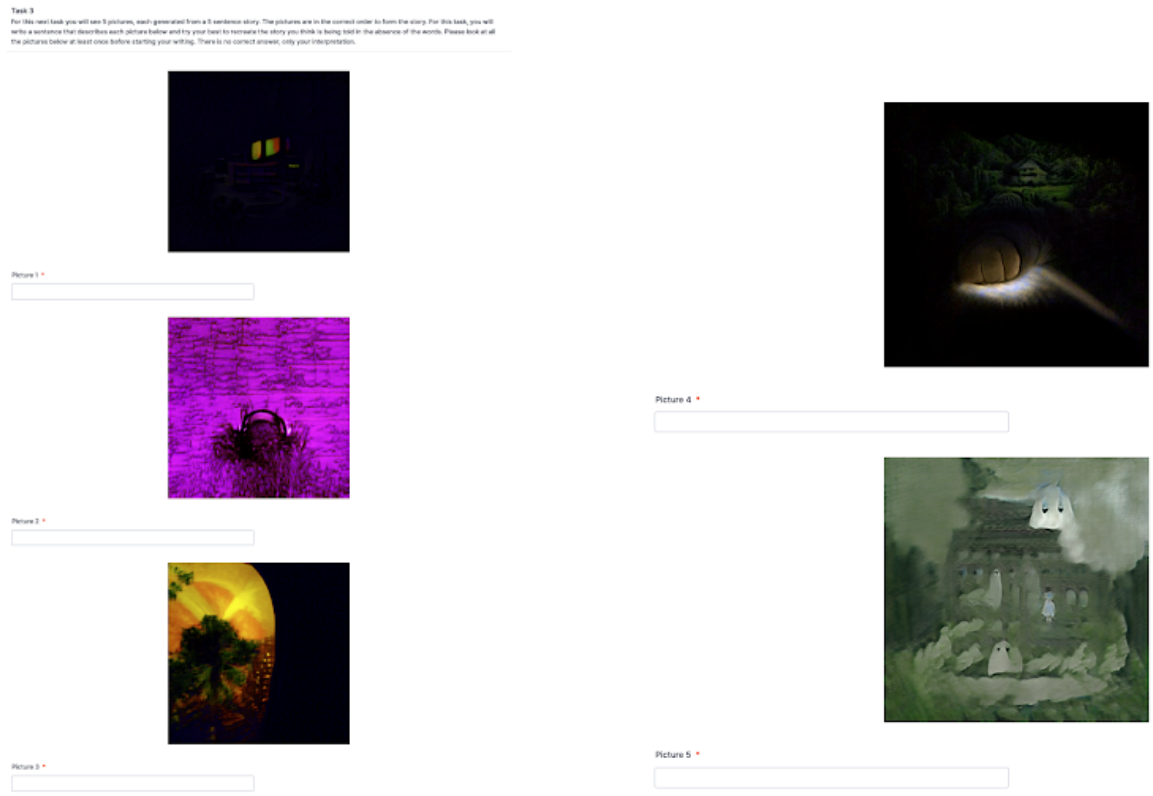


Figure 10: This shows Task 3, the left are the first 3 pictures and on the right are the last 2 pictures, where participants are writing their own stories with image prompts. This will also be referred to as, "The story writing task" in the paper.

Task 3 Questions, *The participant story evaluation task*, this involves various questions to do with the writing of the story in the story writing task;

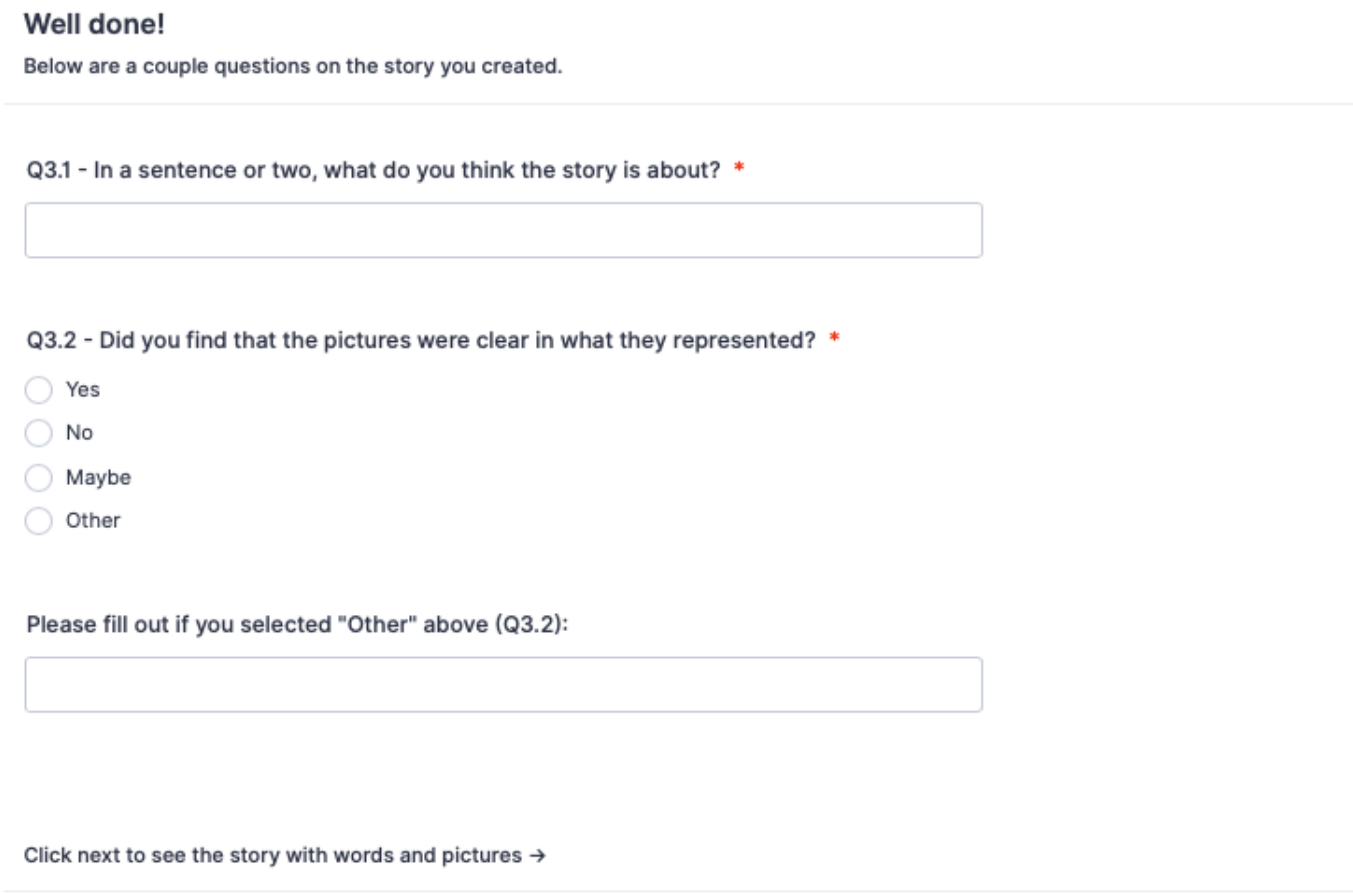


Figure 11: These questions are directly after the 5 images known as The story writing task. This section includes Q3.1 and Q3.2 and will also be referred to as, “The participant story evaluation task.”

This is not a task in itself. Before being presented Task 4, participants were shown the computer generated story “The scary ghost” using the same images they had used to write their own story, but now accompanied with the computer-generated textual prompts. Then the participants had to answer questions for it on the next page in the, “A scary ghost” comparison task.

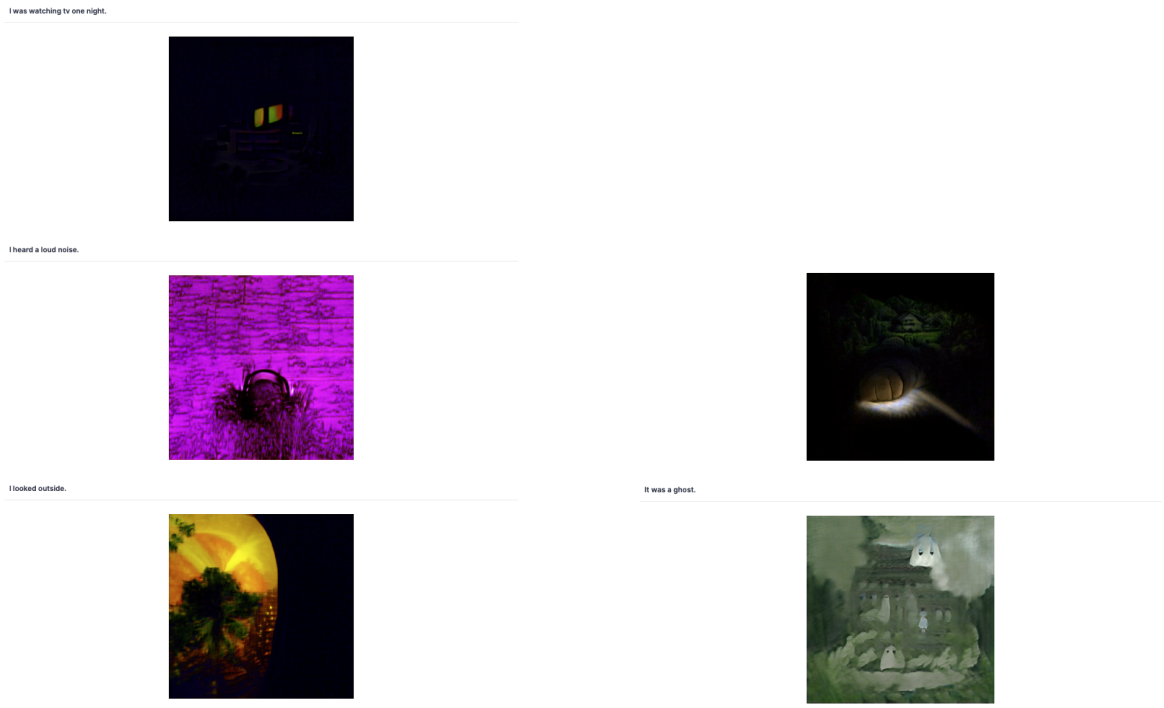


Figure 12: The 5 story image-text pairs goes from top to bottom, and left to right.

Task 4 Questions, *"A scary ghost" comparison task*, this is the final task in the evaluation form where participants compare their answers from the story writing task to the computer generated story with the same accompanying pictures. There are also other questions regarding integration of images and text that covers the entirety of the evaluation.

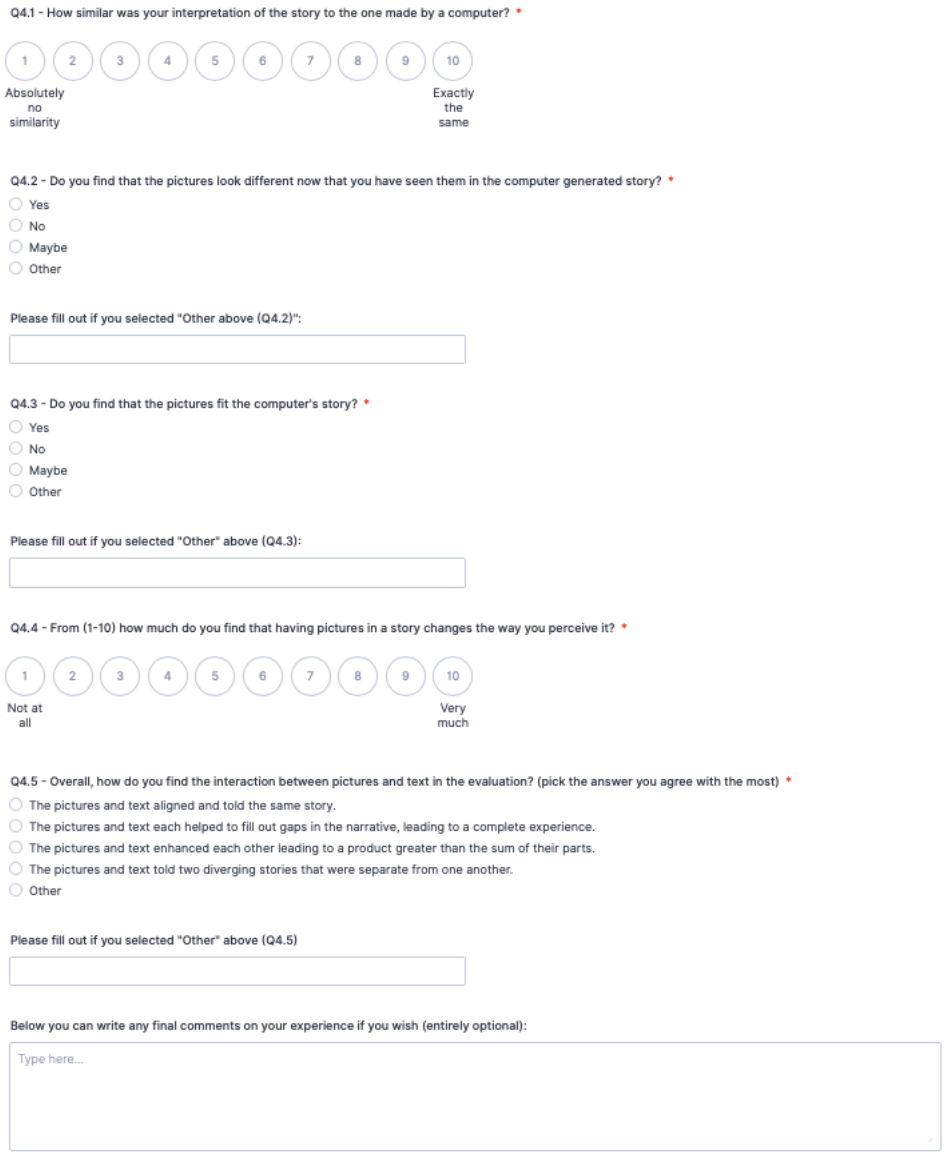


Figure 13: This shows Task 4. Right before these question is the full "The scary ghost," computer generated story with text which can be found in the "Appendix" 24. These question will also be referred to as "The scary ghost comparison task," in the paper.

The various tasks and their corresponding name can be found individually, larger and in higher quality for visual referral in the “Appendix” 24. The metrics used and their distribution throughout the evaluation will be discussed further in the “Methods” section 14 below.

Part IV

METHODS

14

INTRODUCTION TO METHOD

The methods section will cover the methodology behind the narrative generation process, image generation process and the evaluation. The main goal of this section is to discuss why I chose the various architectures and evaluation metrics and how I used them to produce the final results. It will begin by illustrating why I chose Plan-and-write for narrative generation and the specifics of how the pictures were generated. I will then discuss why I chose Bigsleep and how I created the images. Finally, I will discuss why I chose the evaluation form, followed by each metric in the evaluation; Specificity, Relevancy, Narrative Consistency, and Multimodal Fusion, what they measure and why I chose them.

15

NARRATIVE GENERATION PROCESS

The original idea when proposing the thesis was to use GPT3 (Radford et al., 2019) to generate long-form stories and then retrieve pictures from the labelled image dataset, Image-net (Deng et al., 2009a). However, I could not get access to GPT3 as it is in Beta access and you need permission from Open-AI, so I opted to use GPT2 (Radford et al., 2018).

When training GPT2 to generate stories I found that I could not achieve consistent linear narratives. For automatic story generation, I tried different models. I used a fine-tuned GPT2 117M and a 774M model trained on a 28mb .txt dataset of children's stories manually pulled and combined from *Project Gutenberg* (n.d); a free to use database of books with expired copyright licences. However, even with the bigger model and a loss-rate of less than 1.0, meaning that the fine-tuning was thoroughly trained on the dataset, the results did not lead to consistent narratives. Another GPT2 model was trained on stories in the same dataset, split into beginning, middle and endings. To remain consistent in extraction, the data was split by the first 5% for the beginning, final 5% for the ending and the rest for the middle. This was to create fixed story-points rather than manually determining a story's beginning, middle and endings, which would be very time consuming and could lead to possible errors and replication problems. This model was also ineffective in consistent narrative generation. However, the Plan-and-write (Yao et al., 2018) model with higher-order architecture and GPT2 exclusively for NLG did in my opinion generate consistent narratives. As mentioned in the "Materials section," Plan-and-write is a story generation architecture that first generates an outline which serves as prompts

when generating a story. I chose Plan-and-write because I liked that the generated stories were only 5 sentences long and was impressed with the results of the BART variant in their paper (Yao et al., 2018). The sentences generated by Plan-and-write are simple in their language, which was indicated by (Zakraoui et al., 2020) to boost Relevancy in image generation, and the stories were closely related to the prompt and had a red-thread throughout, with what I subjectively determined to be definite endings. Also, having stories that span 5 sentences was advantageous when producing picturebooks as I wanted an image for each sentence and, 1. It was less time consuming to generate pictures to accompany fewer sentences, and 2. The stories were short and therefore easier to evaluate. The Plan-and-write architecture with GPT2 was retrieved by contacting the creators who sent some preliminary code on a GPT2 version that I converted to a Google Colab format and ended up using for the final product (More information and a link can be found in the "Materials" section 14 above). The reason I decided to use GPT2 with the architecture is because I had familiarised myself with the architecture through previous experimentation and found that it had been applied successfully in other recent papers trying to accomplish something similar (See et al., 2019; S. Wang et al., 2020). I chose to use the Open-AI pre-trained GPT2 version and not my own fine-tuned GPT2 model as my own model was more specified in its application and would likely have clashed with the ROCstories storyline generator.

With the longer stories, I experimented with various forms of extracting prompts from sentences, such as noun-extraction. However, with the 5 sentence stories with Plan-and-write and GPT2, I found that the information provided within the sentences was often concise enough that when experimenting with the image generation I could use the whole sentences as prompts. Using the entirety of the story-sentences as prompts also made the generation of stories more representative of what the current architect was able to accomplish and was more in-line with the definition of picturebooks as an entity where images and text are equally important (Gregersen, 1974). I came up with titles for generated stories randomly on the fly. Also, I made a list of reasons I would regenerate stories that aligned with common problems in narrative generation (The "Textual story generation checklist" can also

be found in the “Appendix” 24). This list was made to prevent the various problems with image generation discussed in the “Literature Review” section 4. The list of criteria for deletion are as follow: if the story had repetition of sentences, if the story contained repetition of words within sentences (this was addressed in the Plan-and-write paper as something they eliminated, but I still put it in as a failsafe), if the main character of the story did not carry throughout the entire story, if the story had individual sentences that were nonsensical, or if the story did not make any references whatsoever to the input title. Most of the stories were only generated once, with only a few exceptions of stories being discarded because they did not meet the outlined requirements. The Plan-and-write model would first create a static storyline from a prompt such as the one illustrated in Fig. 14, and then generate a 5 sentence story using the storyline as prompts.

Title/prompt: “The Scary Ghost”

Storyline: night -> heard -> looked -> turned flashlight -> ghost,

Story generated: “I was watching tv one night. I heard a loud noise. I looked outside. I turned on the flashlight. It was a ghost.”

Figure 14: This is an example of how the story, “The scary ghost” was generated using Plan-and-write and used in the evaluation. The three steps involve a title, a storyline and finally a story. Each of the 5 sentences in the story were then used to generate an image in Bigsleep.

16

IMAGE GENERATION PROCESS

The first idea I had was to use image retrieval, but because of recent advances in image generation since the proposal, and the advent of CLIP, I decided to go the route of generating the pictures from scratch using a GAN architecture. I settled on Deepdaze or Bigsleep. After running preliminary tests, Bigsleep was chosen because I found the pictures subjectively more appealing and accurate in their representations than Deepdaze. Also, the Bigsleep architecture had successfully been used in visualization projects such as "*This is not a real band*" [1].

Figure 15: This shows pictures generated for comparison using the same prompt, "a whale flying over a mountain" by Bigsleep (left) and Deepdaze (right). The first generated picture was selected for both architectures.

1 https://twitter.com/ai_metal_bot

For the image generation with Bigsleep, I experimented with adding “in the style of cartoon” to the end of the prompt and other forms of augmentation. This was an attempt for more consistent thematic appearance and to promote the perception of a continuous and single product. I eventually found that the sentences by themselves without any modification were more in-line with the non-assisted computer generation that I was aiming for and were just as effective as my other attempts in thematic tone. After experimenting with having different pictures as starting points and changing the learning rate, I settled on the default starting image and a 5e-2 learning rate, as I found the starting image did not change the results meaningfully and subjectively I found that it was consistently the best learning rate for producing quality pictures.

To incorporate what was learnt in prior work, stating that some human intervention was necessary when using image generation (Fotedar et al., 2020), I created a checklist to accompany my picking of the images. I wanted the generation to accurately represent the architecture’s capabilities without too much human “cherry picking”, or human intervention in selecting the best results, but the literature review deemed it necessary to have at least some filtration of results. I made the checklist as brief as possible and only eliminated pictures that had definite errors or did not have resemblance of mentioned characters. As mentioned in the “Materials” section 10 above, the starting image often resembled dogs and sometimes birds and this occasionally bled into the generated image and affected the generation to not match the prompt, you can see examples of this in Fig. 16 below. The checklist would help eliminate the pictures that were: too similar to the original starting image, did not contain at least the resemblance of a character referred to in the sentence (Inspired by the “Story character classification” from, (Zakraoui et al., 2021)), images that were too blurry, and images that were just the input sentence written in words (The “Image generation checklist” can be found in the “Appendix” 24). If all the criteria were met, I would pick the first image that was created.

The images took anywhere between 20-25 minutes to generate. If the image was not relatively sharp after 25 minutes it would be discarded and the Google Colab notebook would be restarted with the same prompt. If the images checked all the boxes on the list, they were downloaded from the

notebook and put into a folder with the story title name. During story generation, no more than 3 pictures were generated for any one prompt and several sentences were paired with the first image generated. I manually input the sentences from stories generated from the Plan-and-write architect. I generated 6 complete stories in all.

Figure 16: “Ryan had dreams of becoming a singer” resembles a bird (left), and prompt: “He decided to get a job” resembles a dog (right).

17

EVALUATION

For the evaluation form, I decided on 4 metrics that I wanted to evaluate: 1. Specificity vs Generalization, 2. Relevancy, 3. Narrative Consistency and 4. Multimodal Fusion. The evaluation form was designed to take about 10-15 minutes to complete as it requires a lot of creativity and attention from participants and I did not want to make it so long that it would be mentally taxing. The main focus of the evaluation was to gather information and try to understand how human participants perceive the integration of computer generated text and images when paired together. However, it needs to be considered that when engaging with computer generated media, a classical approach of picturebook analysis, for example outlined in (Callow, 2020), would be too specifically tailored towards human created picturebooks and may not accurately represent the computer generated media. Therefore, I created an evaluation that is a synthesis of image-text integration from various fields, including Educational research, Psychology, Humanities and Data Science. Participants were gathered on social media in various pre-established groups and through direct messaging.

I thought to make adequate conclusions in integrating computer generated imagery, I first had to measure how the generated images are representative of the text, and Specificity vs Generalization is argued by Hughes et al., to do just that (Hughes et al., 2007). I also thought it necessary to find how cohesive the images and text were in a narrative context; if the images and text together form a cohesive story. For this I chose the metric *Narrative Consistency* or interchangeably Narrative Cohesion, which is measuring how interchangeable images would be in a textual story. The concepts

of image-text Relevancy and Narrative Consistency were pulled from various literature, but the version used in the evaluation was specifically defined by the metrics used by Sapkale Lukin (Sapkale & Lukin, 2020). Finally, I wanted to measure the product of the integration of images and text as Multimodal Fusion. Questions were created to measure how the text and images were integrated through the systems proposed by Nikolajeva Scott, Enhancing, Counterpointing, Complementary or Symmetrical (Nikolajeva & Scott, 2013). I decided against using the *Inception score* (Salimans et al., 2016), which is a score devised as an image classification neural network, and *Fréchet Inception Distance* (FID) (Heusel et al., 2017), which calculates the difference between a real image and a generated image in the same domain, because these metrics are specific for evaluating images and require testing images in the same domain. The FID score require images similar to the ones generated, but images generated for stories would maybe not represent real images as they are imaginary images for fictional stories. The scores are better in ranking image generation systems, and BigGAN which I used in the generation reached high scores in both, so I did not find it necessary.

From these metrics, the evaluation was split up to measure individual images and sentences in isolation and images and sentences in a story context. The integration between the images is largely done through the writing of text by participants. Further information will be stated about the various metrics and how they were defined and measured within the evaluation in the corresponding sections below.

17.1 SPECIFICITY VS GENERALIZATION METRIC

A story that is represented through images alone can be interpreted in various ways. The images as they relate to text, will be on a spectrum of *Specificity*, where the image can represent a large number of possible sentences (a General image) or a singular sentence (Very Specific). I wanted a way to measure the initial textual story that is “pulled” through the images when observed by participants. I picked this metric as it was discussed by Hughes et al., as a way to measure generated imagery and its

synthesis with text (Hughes et al., 2007). If the images are more Generalized, it will indicate the text is not closely related to the details presented in the image. This type of metric is important as it will help identify how images and text synthesise in terms of the similar details present in both instances.

Q1.3 from the non-narrative task and Task 3, and as an extension Q3.1, from the story writing task and participant story evaluation task respectively, are designed to measure Specificity vs Generalization. In each case, the participants are given images and need to write what they believe the prompt is. If the human written prompts are all similar, then the images would be more specific to the prompt as it would be accurately representing a single possible narrative. If the human written prompts are all very different from the images, the images are more *Generalized*, which means that they lack Specificity to the textual story and represent a wide variety, or no possible text accompaniments. The "A scary ghost" comparison task, specifically Q4.1, is also designed to measure this through asking the participants how close they think their own written story from images alone is when compared to a computer generated story with the same images.

Another way this is measured is through Q2.2 from "The new job" evaluation task, and Q3.2 from the story writing task, which are intended to measure Specificity by asking how "clear" pictures presented in various tasks were. This is a more integrated measure of this task, as it is analysing how clear images are in being representative in a narrative context. If images are not clear, they would be more Generalized and if they are clear they would have more Specificity.

17.2 RELEVANCY METRIC

This evaluation metric is tailored for image-text pairs. It is important to make a distinction between Relevancy and Specificity vs Generalization. *Relevancy* is measured in the evaluation when both text and image are present and perceived as one entity; in this way Relevancy is presented as a multimodal product to be evaluated, and Specificity vs Generalization is a unimodal product that is presented and it is the participant's task to integrate the other modality. In this evaluation, the Relevancy metric is

obtained with an image, and integrating text. This is important as a baseline to see if the images in a non-story context are recognisable as what they are proposed to represent. If there is low Relevancy between images and text, it would mean that when the two are represented together as one product, there is a disassociation between the two. For example, if an image of a whale is accompanied by a caption "The squirrel climbed the tree," the image and text are disassociated and in turn have low Relevancy.

Q1.1, Q1.2, Q4.3 are questions designed to measure Relevancy. I will describe each task briefly: Q1.1 is from the non-narrative task and is intended to measure the amount an image expresses the given prompt on a scale of 1-10. When the two are paired, the participants will indicate how high Relevancy the images and text prompts have with one another.

Q1.2 is a multiple choice question from the non-narrative task, where given an image and 5 prompts, 4 of them are random prompts from other stories and 1 of them the correct prompt, the participants will be tasked in matching the image to the correct prompt. When presented the image, if the correct prompt is picked, it will show high Relevancy between image and text.

Q4.3 is a 1-10 scale evaluation of how images fit the text in a complete story from the "A scary ghost" comparison task. After reading a story, the participants are asked how the images fit the text prompts. The closer the score is to 10, the higher the Relevancy of the images to the text sentences.

17.3 NARRATIVE CONSISTENCY METRIC

Narrative Consistency is defined as the measure of whether the combination of words and images leads to a cohesive story that is internally consistent. In the evaluation it is the measure of how the images can be drawn from text to fit into a narrative-whole. The methodology behind it is whether, when images are scrambled, the textual story can help align the sequence of images to consistently create a narrative. This metric focuses on text first and then images.

Task 2 is from the matching task where a textual 5 sentence story is first given within its proper sequence in its entirety. Then participants are presented a scrambled version of the images and need to pick the appropriate image to fit each textual story segment. There are 5 images that each match a single sentence, and each need to be matched in the proper sequence to form the computer generated narrative. If the pictures are matched in the proper sequence, then the Narrative Consistency and internal coherence of the textual story when combined with the images will be high. This metric is focusing on integration, and how coherent the sequences of computer generated images are when combined with text.

17.4 MULTIMODAL FUSION METRIC

The *Multimodal fusion metric* is designed to measure how the participants perceive the entity created when combining images and text in the stories and how they perceived this integration; the multimodal product that is created when the images and text are combined as a single narrative.

Q4.5 is from the "A scary ghost" comparison task and is measuring the Multimodal fusion metric directly by asking participants how the combination of the images and text pairs presented as stories, Task2, Task3 and Task4, fit into one of the 4 types of Multimodal Fusion: Enhancing, Counterpointing, Complementary or Symmetrical, or an option for "Other" if they do not find any of them are appropriate.

Additionally, throughout the evaluation in Q2.3 from the matching task and Q4.2 from the "A scary ghost" comparison task, participants are asked if they believe that the addition of pictures in the stories changed the way they perceived the product. This is intended to measure if the multimodal products affect each other. Finally, Q4.4 from the "A scary ghost" comparison task asks whether participants believe that the addition of pictures in stories changes their perception of the product as a more general question to help gauge how the participants find the images affect textual stories as a whole. If the perception is largely changed when combining text and images together in narrative form, we can

infer that the product created from intertwining image and text leads to something that is different from their unimodal counterparts.

Part V

RESULTS

18

RESULTS

To better organise the results, the appropriate questions are split up into each corresponding metric for an easier overview of each measurement. Within each metric, the questions will be briefly summarised to add context to the results and within each metric section they will be displayed in the order that they appear in the evaluation itself. The same clone of the evaluation found in the "Materials" section 10 can be found in full from the link in the footnote [1] and each story and question will be found in the "Appendix" 24. All the results are rounded to 2 decimal places. For data processing and graph creation, the evaluation form data was downloaded as a .csv spreadsheet from the Jotform website. All the results were processed via Python 3 in an Anaconda Jupyter notebook. All graphs were made in Python using Plotly and Matplotlib, tables were made in GoogleDocs.

18.1 SPECIFICITY VS GENERALIZATION OF IMAGES

All the Questions in this Results sub-section were intended to measure "Specificity VS Generalization" and are ordered by appearance in the evaluation.

1 https://form.jotform.com/211254776600351

18.1.1 *Task 1, Q1.3*

Task 1, Question 1.3 was a short-answer question from the non-narrative task. The participants were provided a computer-generated image from a story. The images were presented in isolation with the prompt, "She won the football game." Participants were asked, "What do you think the picture above shows? (In one sentence)." Since participant answers were so varied within the short-answer questions, I manually parsed the responses and added them to relevant lists according to matching words between their written sentence and the computer-generated prompt. None of the answers matched the text prompt and no participants mentioned the word "Won." However, slightly less than half of participants mentioned the word "Football" (28/58 or 48%). It is also worth noting that other participants mentioned words related to other sports activities (25/58 or 43%) as opposed to being entirely unrelated to sports (5/58 or 9%). However, according to the metrics in this paper, this would not be closely matching the text prompt and still show a higher level of Generalization as opposed to Specificity.

18.1.2 *Task 2, Q2.2.*

Task 2, Question 2.2 was a multiple-choice question with 4 possible options ("Yes," "Maybe," "No," and "Other"), pulled from the "The new job" evaluation task. The question was phrased as, "Did you find that the pictures were clear in what they represented?," and was related to the matching task (Task 2) generated from the title "The New Job." When asked if people thought that the pictures in Task 2 were clear in what they presented, 32/58 or 55% said "No," 14/58 or 24% said "Maybe," 9/58 or 16% said "Other," and only 3/58 or 5% said "Yes." The results can be found in a bar-graph in Fig. 17 below for visual comparison. This shows a higher degree of Generalization as opposed to Specificity.

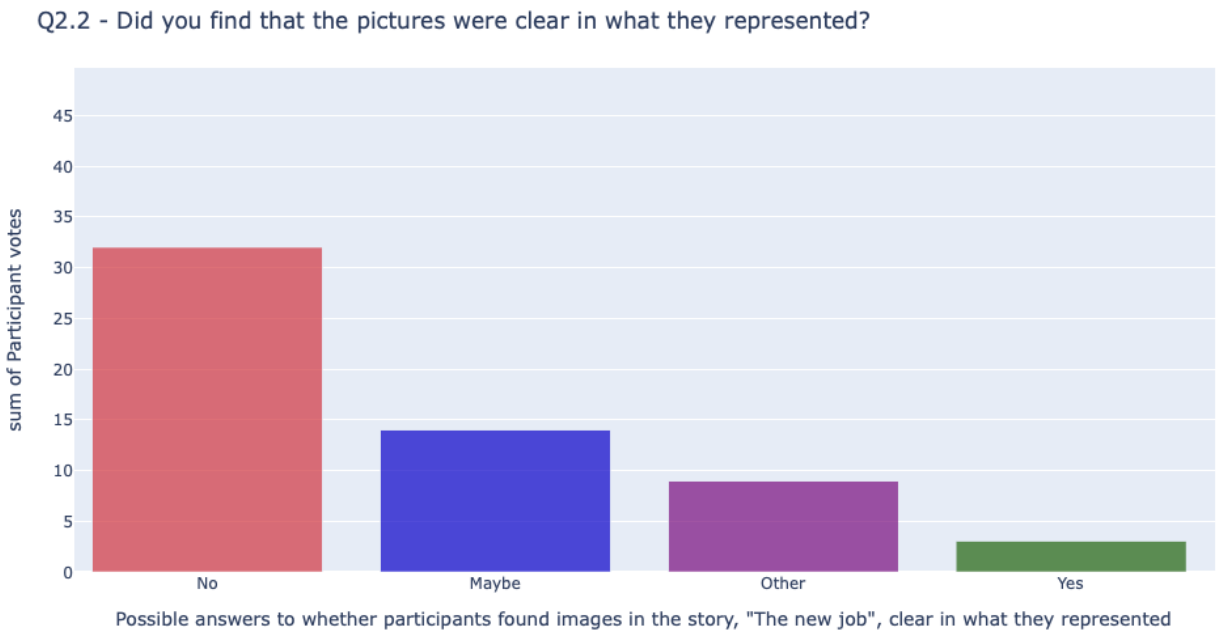


Figure 17: This shows a bar-graph representing the distribution of participant answers from Q2.2: pulled from the, “The new job” evaluation task.

18.1.3 *Task 3 and Task 3, Q3.1.*

Task 3 was a sequence of 5 short-answer questions for 5 image and text pairs in the story generated from the prompt, “The Scary Ghost.” The question was phrased as,

“For this next task you will see 5 pictures, each generated from a 5 sentence story. The pictures are in the correct order to form the story. For this task, you will write a sentence that describes each picture below and try your best to recreate the story you think is being told in the absence of the words. Please look at all the pictures below at least once before starting your writing. There is no correct answer, only your interpretation,”

and is also referred to as the story writing task. As it was a sequence of short-answers with 58 unique participant answers for each question, I will only broadly discuss the results. All in all, I will mention that most stories did not contain the words used in the generation of the images. The only exceptions were image 1 and 5 in the sequence. Image 1 was generated from the prompt, “I was

watching tv one night." Here 18/58 or 31% mentioned televisions, 33/58 or 57% mentioned that it was dark or in the night. In image 5, which was generated from the prompt, "It was a ghost," 47/58 or 81% mentioned the word ghost or haunted. However, despite the medium score in 1 and high score in 5 of Specificity, from reading the prompts I would conclude that the 3 remaining short-answer results (2,3 and 4) had very little resemblance to the prompts and therefore the short-answer questions as a whole lean more towards a higher Generalization than Specificity. Task 3.1 was a short-answer question pulled from the participant story evaluation task, and was phrased as, "In a sentence or two, what do you think the story is about?" This question is an extension of Task 3, and also showed high levels of Generalization. Of the 58 responses only 5/58 or 9% mentioned the word "ghost" or "ghosts," when asked what they thought the story was about. Task 4, Q4.1 below, is pulled from the same section and has each participant rank the similarity between their story and the computer generated story and will therefore be used to draw further conclusions from this task.

18.1.4 *Task 3, Q3.2.*

Task 3, Question 3.2 was a multiple-choice question with 4 possible options ("Yes," "Maybe," "No," and "Other"). This was pulled from the participant story evaluation task, and was to be answered after finishing the story writing task. Question 3.2 was presented as, "Did you find that the pictures were clear in what they presented?," referring to the images in the story writing task where participants had to write their own story-sentences for the computer generated images from the "A scary ghost" story. From this, 21/58 or 36% answered "No," 12/58 or 21% answered "Yes," and 20/58 or 34% answered "Maybe," 5/58 answered "Other." This shows that although some of the answers were high in Specificity, the results as a whole lean more towards Generalized as opposed to Specificity in terms of image-text interaction. A visual representation can be seen in a bar-graph below, in Fig. 18.

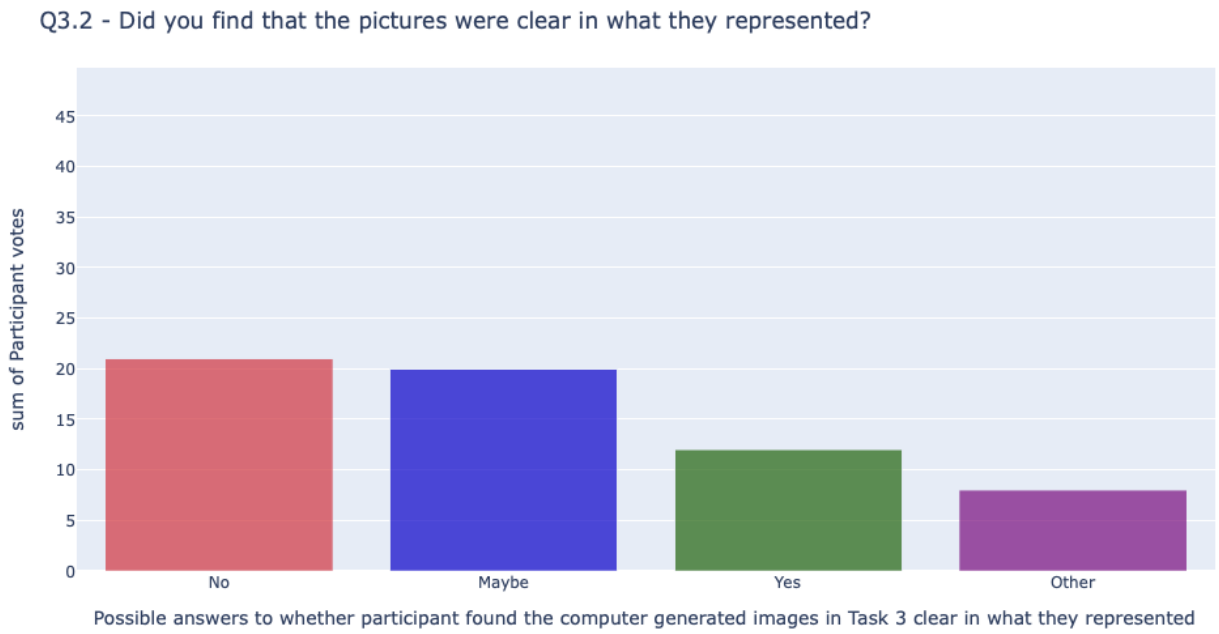


Figure 18: This shows a bar-graph representing the distribution of answers on Task 3: pulled from the story writing evaluation task

18.1.5 *Task 4, Q4.1*

Task 4, Question 4.1 was a scale-rating question from 1-10, that involved rating the computer generated story from Task 3 (the story writing task). The question was phrased as, "How similar was your interpretation of the story to the one made by a computer?," and was pulled from the "A scary ghost" comparison task. The possible answers ranged from 1 (Absolutely no similarity) to 10 (Exactly the same). The results from the question were quite mixed. However, no participants rated it 10 (exactly the same). The mean was 4.81 showing quite an even spread across the rating scale, with two peaks at 2 to 3 and 6 to 7. From this I would conclude that it shows a mixed level of Specificity. Fig. 19 below shows a bar-graph visually representing the data.

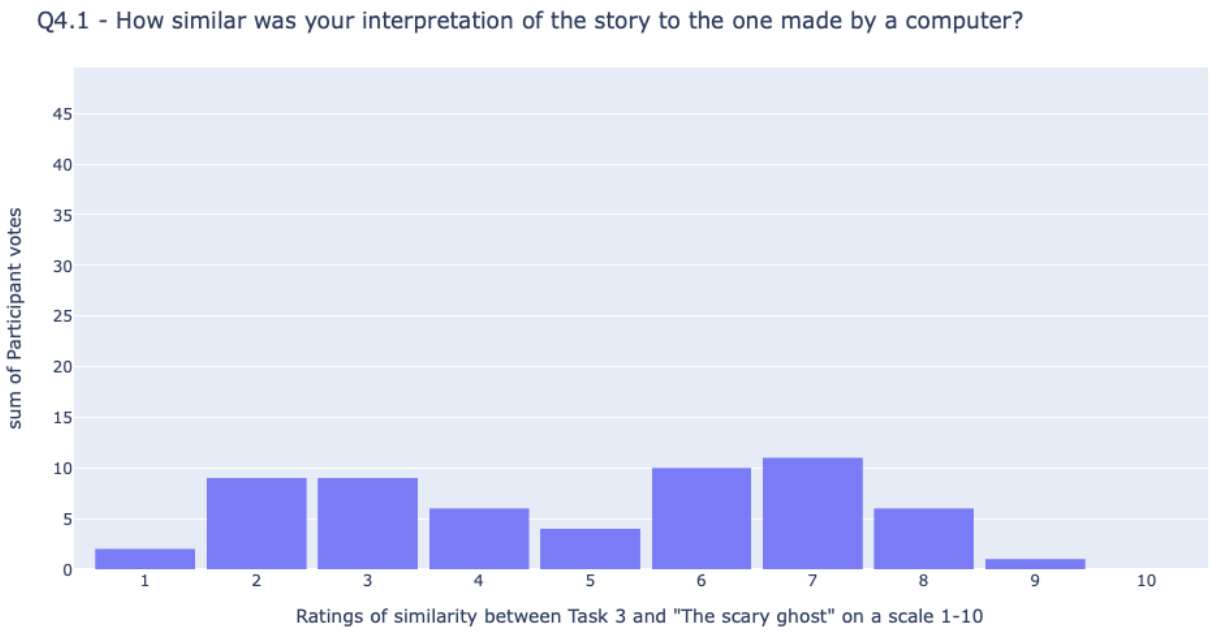


Figure 19: This shows a bar-graph representing the distribution of participant answers from Q4.1: pulled from the, “A scary ghost” comparison task.

18.2 RELEVANCY OF TEXT AND IMAGES

All the Questions in this Results sub-section were intended to measure “Relevancy” and are ordered by appearance in the evaluation.

18.2.1 *Task 1, Q1.1*

Task 1 Question 1.1 was a scale-rating question ranging from 1-10 for a single image generated from the prompt, “She was very scared,” pulled from the non-narrative task. The rating scale ranged from 1 (“it does not look like that at all”) to 10 (“It looks exactly like it”) and the question was phrased as, “On a scale from (1-10), how much does the picture above express ‘She was very scared.’” The most picked number was 8, with 14/58 or 24%, the least picked was 1 with 0/58 or 0%. The mean was 7.59, which means that the rating was on average quite high and people found that the picture

expressed the prompt quite clearly and in turn that it had high Relevancy. A bar-graph representing the vote distribution can be found in Fig. 20 below. The bar-graph shows a skewed distribution, with the participants favouring the higher numbers ranging from 6-10.

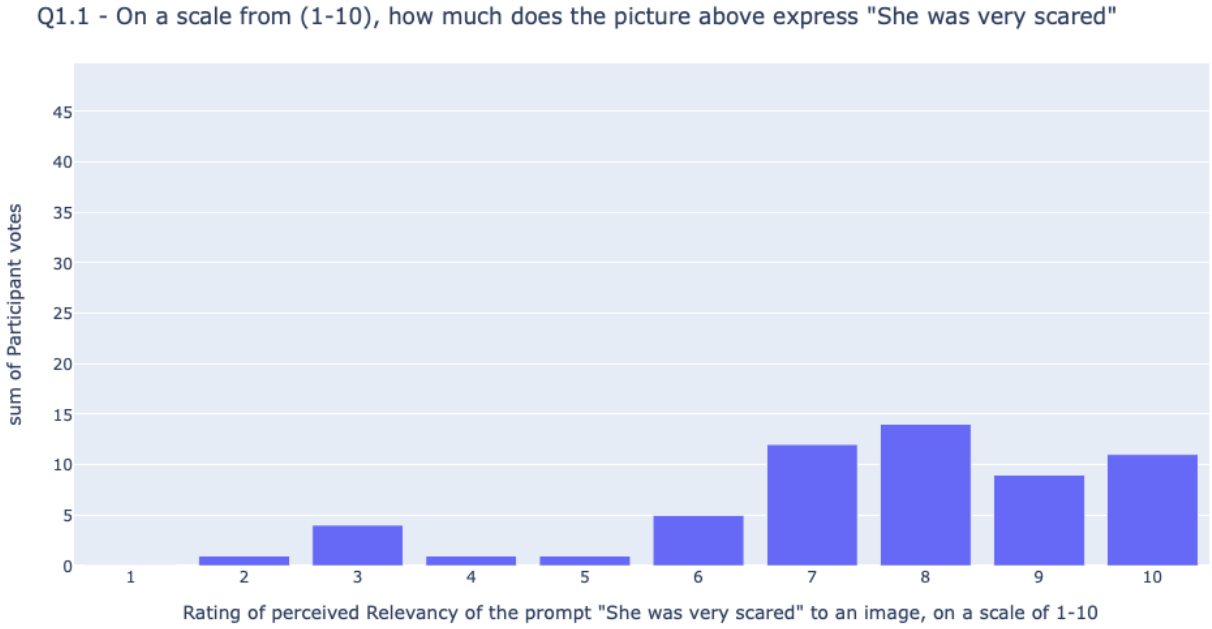


Figure 20: This shows a bar-graph representing the distribution of participant answers from Q1.1: pulled from the non-narrative task.

18.2.2 *Task 1 Q1.2*

Task 1, Question 1.2 is a multiple-choice question with 5 possible choices, containing the correct prompt and 4 random prompts from other stories. This question was pulled from the non-narrative task, and phrased as, “Which sentence best describes what is happening in the picture above?” Participants had to try and match the correct prompt to the image generated from, “Verney was a good student.” The correct answer was matched the majority of the time with, 45/58 or 78% of participants matching the prompt correctly to the picture. The rest of the options chosen in descending order were, 5/58 or 9% chose, “It was a great camping trip,” 4/58 or 7% chose, “He would greet people,”3/58 or 5% chose, “Ryan had dreams of becoming a singer,” and finally, 1/58 or 2% chose, “He eventually made a

lot of people's ears." These results would indicate a high degree of Relevancy between image and text. The results can be found in Fig. 21 as a bar-graph below for visual interpretation of the data.

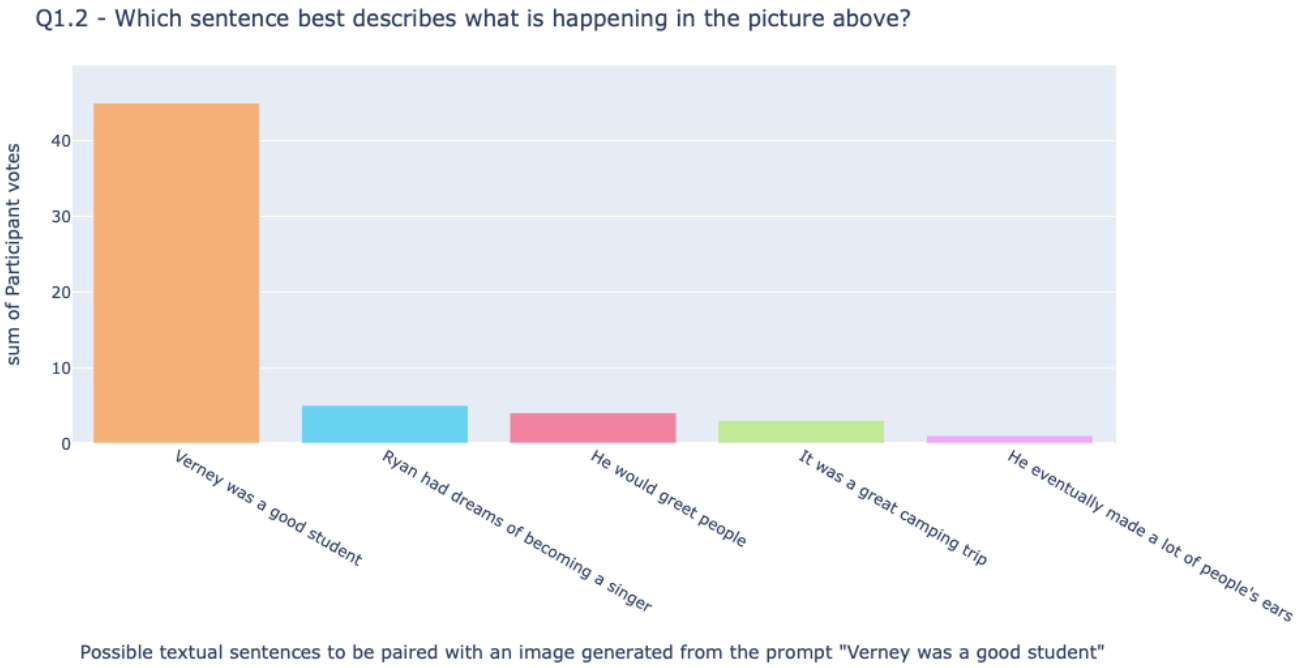


Figure 21: This shows a bar-graph representing the distribution of participant answers from Q1.2: pulled from the non-narrative task.

18.2.3 *Task 4, Q4.3*

Task 4, Question 4.3 was a multiple-choice question with 4 possible options ("Yes," "Maybe," "No," and "Other"), pulled from the "A scary ghost" comparison task. When asked, "Do you find that the pictures fit the computer's story?," referring to the complete "The scary ghost" computer generated story, the majority of people responded "Yes," with 35/58 or 60%. The second highest option was "Maybe," with 18/58 or 30%, 3/58 or 5% answered "Other," and only 2/58 or 3% said "No." The high numbers would indicate that the image and text have quite a high Relevancy to one another. A bar-graph representing the results visually can be found below in Fig. 22.

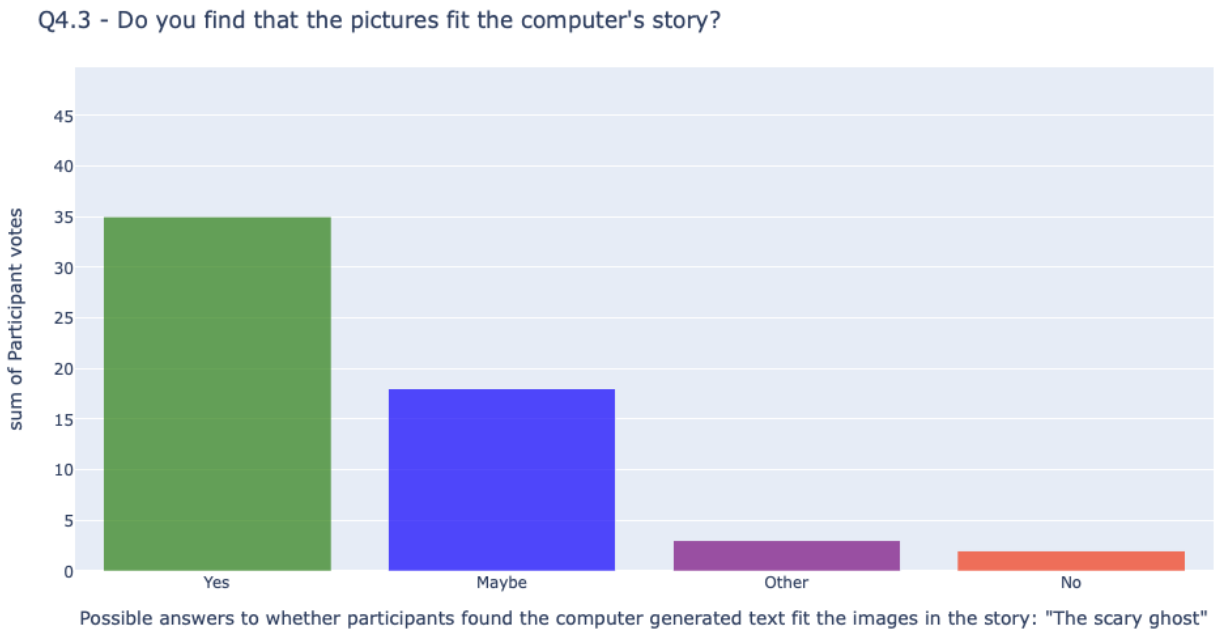


Figure 22: This shows a bar-graph representing the distribution of participant answers from Q4.3: pulled from the, “A scary ghost” comparison task.

18.3 NARRATIVE CONSISTENCY OF TEXT AND IMAGES

The question below in this Results sub-section was intended to measure “Narrative Consistency”.

18.3.1 *Task2*

Task 2 was a sequence of 5 matching tasks where the participants were presented the textual story “The new job,” and were then presented each sentence in the correct story order with the 5 scrambled generated images. This task was phrased as,

“On the next slide you will be presented with a complete 5 sentence story in its entirety. After you read the story you will be presented with one sentence at a time from that story, and a choice of 5 pictures for that one sentence. Your task is to select the picture that you think best fits the presented sentence. Each sentence pairs with one picture,”

and is referred to throughout the paper as the matching task. For each sentence the participants were asked to, “Pick the picture below that you think fits this sentence the best.” The participants were informed that each sentence had 1 matching image.

The task was matched largely incorrectly. Only 1/58 of participants arranged the story in the correct order. Out of the 5 images, only 1 image was matched correctly a majority of the time. This was image number 4 out of 5, where 33/58 or 57% of participants matched the image to the story-sentence correctly. The second (2/5) text-image pair was matched correctly the least, with only 3/58 or 5% correct matches. Fig. 23 below is a table of the data. Within the table, the columns represent the images as labelled by their correct order and the rows represent the sequences of text in the correct order. To get a visual representation of the matching task, you can find the full task in the “Appendix” section 24.

Chart of Image sequences vs Textual story sequences

Order picked	First	Second	Third	Fourth	Fifth
First	12	14	**20**	9	*3*
Second	12	*3*	12	**18**	13
Third	*3*	**40**	5	*3*	7
Fourth	9	*1*	11	**33**	4
Fifth	10	*3*	**26**	5	14

Bolded is the most picked option, italics is the least picked. The green boxes are the correct matches.

Figure 23: This is a table of the data from Task 2 otherwise known as the matching task. The columns are the images in correct order, the rows are the text in correct order.

18.4 MULTIMODAL FUSION OF TEXT AND IMAGES

The question below in this Results sub-section was intended to measure the "Multimodal Fusion" metric.

18.4.1 *Task 2, Q2.3*

Task 2, Question 2.3 was a multiple-choice question with 4 possible options ("Yes," "Maybe," "No," and "Other"). The question was phrased as, "When presented the sentences and the pictures together, did you find that your initial perception of the story changed?," referring to the matching task for "The new job" story in the matching task. This question was pulled from "The new job" evaluation task. Of the answers, 22/58 or 38% said "Yes," 12/58 or 21% said "Maybe, " and 24/58 or 41% said "No." A bar-graph representing the answers visually can be found below in Fig. 24.

Figure 24: This shows a bar-graph representing the distribution of participant answers from Q2.3: pulled from the, "the new job" evaluation task.

18.4.2 *Task 4, Q4.2*

Task 4, Question 4.2 is a multiple-choice question pulled from the “A scary ghost” comparison task, with 4 possible options (“Yes,” “Maybe,” “No,” and “Other”). The question was referring to the participant written story when compared to the computer generated story and was phrased as, “Do you find that the pictures look different now that you have seen them in the computer generated story?” After being presented with the computer generated story in Task 4, 26/58 or 45% of participants responded “Yes,” that the pictures did look different after seeing them in a different context, 13/58 or 22% said “Maybe,” 17/58 or 29% said “No,” and 2/58 or 4% responded “Other.” A bar-graph visually representing the distribution of answers can be found in Fig. 25 below.

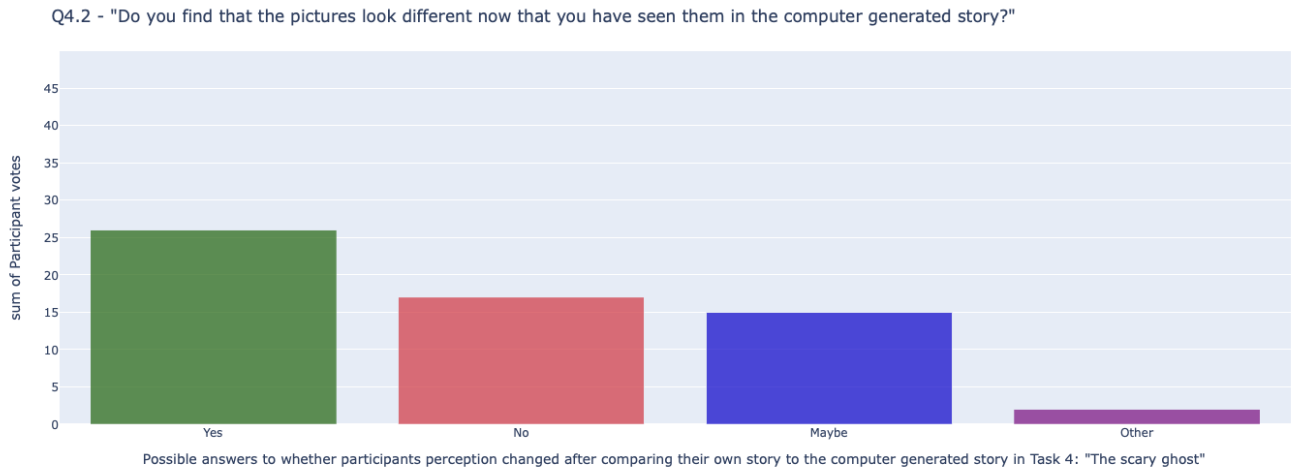


Figure 25: This shows a bar-graph representing the distribution of participant answers from Q4.2: pulled from the, “A scary ghost” comparison task.

18.4.3 *Task 4, Q4.4*

Task 4, Question 4.4 was a scale-rating question from 1-10, asking, “From (1-10) how much do you find that having pictures in a story changes the way you perceive it?” This was pulled from the “A scary ghost” comparison task. The rating ranged from 1 (Not at all) to 10 (Very much). This question

was proposed as a general question, but after experiencing the whole evaluation was designed to reflect on how the images changed the stories above. For this task, 42/58 or 72% answered 8 or above, 0/58 people responded 2 or below and the mean was 8.12. For visual referral, the data is represented in a bar-graph in Fig 26. The bar-graph is skewed with two peaks at 8 and 10, with participants favouring the higher ratings, 7-10.

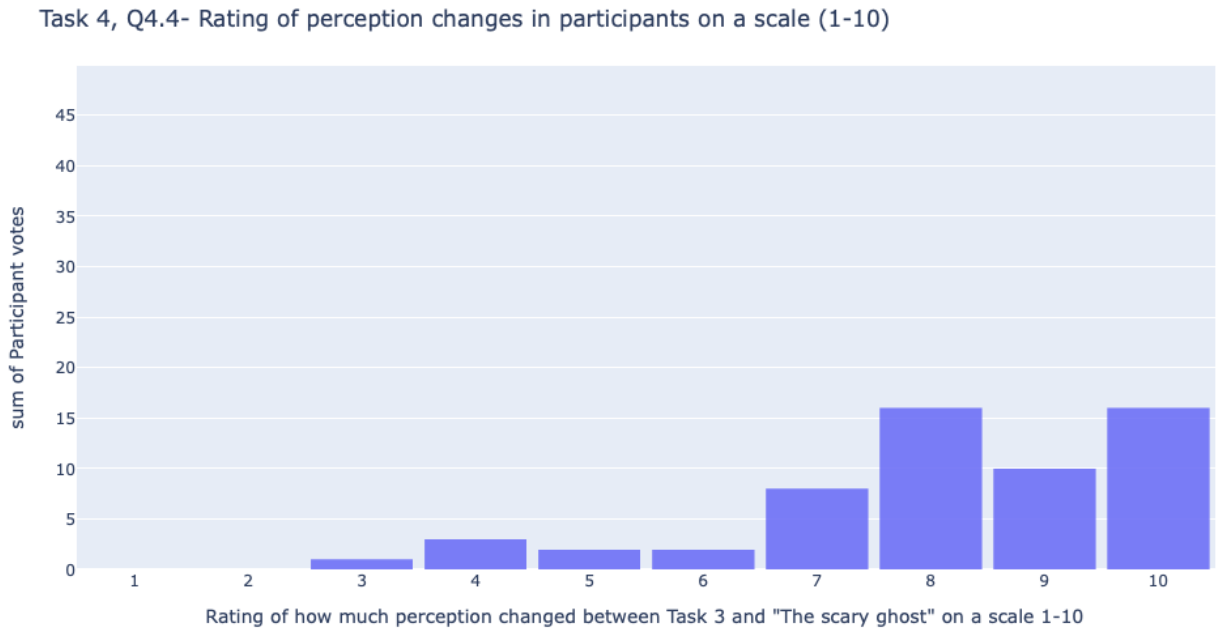


Figure 26: This shows a bar-graph representing the distribution of participant answers from Q4.4: pulled from the, “A scary ghost” comparison task.

18.4.4 *Task 4, Q4.5*

Task 4, Question 4.5 is a multiple-choice question with 5 possible choices. The question was phrased as, “Overall, how do you find the interaction between pictures and text in the evaluation? (pick the answer you agree with the most),” and pulled from the “A scary ghost” comparison task. The 5 possible choices are,

1.“The pictures and text aligned and told the same story” (Symmetrical), 2. “The pictures and text each helped to fill out gaps in the narrative, leading to a complete experience” (Complementary), 3.

"The pictures and text enhanced each other leading to a product greater than the sum of their parts" (Enhancing), 4. "The pictures and text told two diverging stories that were separate from one another" (Counterpointing), and 5. "Other."

In their responses, 33/58 people or 57% responded that they believe the pictures and images together create something greater than the sum of their parts. This would support that the multimodal fusion was enhancing. Additionally, 18/58 or 31% believe that the pictures and images together fill in parts of the narrative leading to the full experiences, showing that some some participants found that the mulitmodal fusion was complementary. Only 7/58 or 12% of people responded with another option or "Other." A visual representation can be found in the pie-graph in Fig. 27 below.

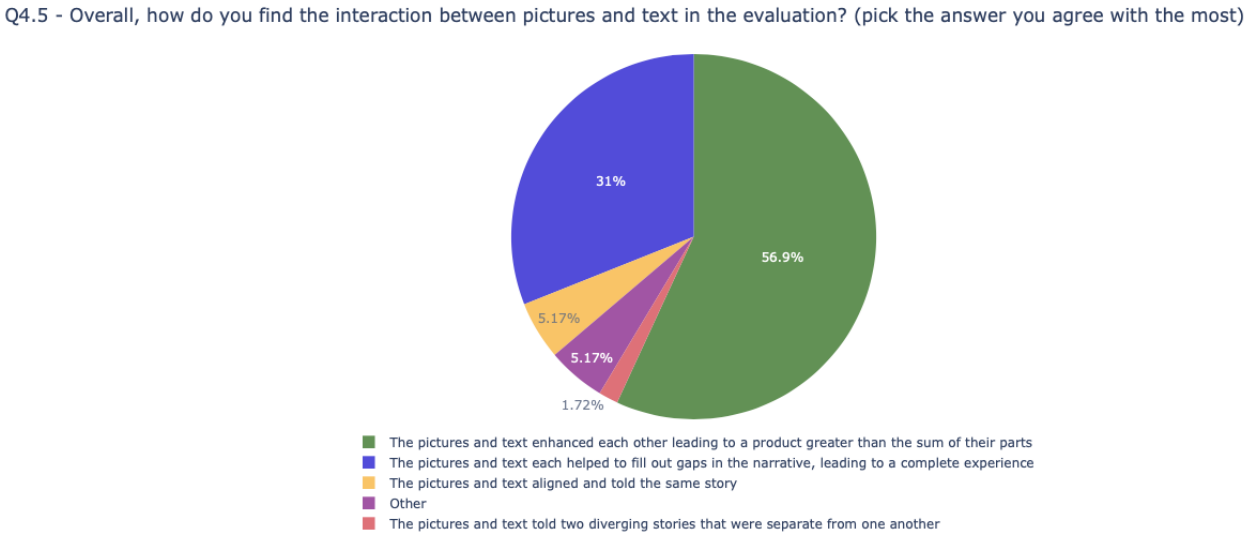


Figure 27: This shows a pie-graph representing the distribution of participants answers from Q4.5: pulled from the, "A scary ghost" comparison task.

Part VI

DISCUSSION

19

INTRODUCTION TO DISCUSSION

The discussion section is split up into Analysis of Results 20, Limitations 21 and Future work 22. The Analysis of Results section will cover each metric in their corresponding section and how the results relate to the literature outlined in the “Literature Review” section 4 above, followed by a “Further Results Analysis” subsection that will be discussing all metrics together and discuss the results from the evaluation as a whole. This is followed by the “Limitations” section, which will cover possible limitations with the evaluation and generation. Lastly, the ”Future Work” section will discuss what can be done in the future to improve on and further the results from this paper. Also, it is worth noting that the results from the “Other” category did not provide additional information and were therefore discarded.

20

ANALYSIS FROM RESULTS

20.1 SPECIFICITY V.S GENERALIZATION RESULTS ANALYSIS

In regards to what the images portrayed, participants reported consistently that they were not very clear. This would indicate a high level of Generalization as opposed to Specificity. This can be seen in various answers from the evaluation, but is especially well represented in Q4.1 from the "A scary ghost" comparison task, where an average score of participants was 4.8/10 of how similar the human created stories were in comparison to the one generated by the computer. Considering the scale varies from "Absolutely no similarity" (1) to "Exactly the same" (10), it would indicate that participant's stories were likely not very similar to the computer generated story. From this, one could argue that the computer generated images were interpreted in many different ways, and would mean that the images are leaning towards communicating a variety of meanings as opposed to specifically representing the text-prompts. From viewing the images, I believe this makes sense as the pictures are quite abstract in their representation. Many images that Bigsleep generates are "dreamlike", abstract, blurry (even when filtering out the very blurry images) and do not have something highly-recognizable in the center of the image. In the literature, these are discussed as important factors in achieving high Specificity, indicating that images could be strengthened by considering image-text relations and improving sharpness and centering points of interest (Hughes et al., 2007). Since many of the images do not match these criteria, it makes sense that the computer generated pictures in isolation lean more towards

Generalized rather than Specific images. This also supports prior research that emphasised that human intervention is still necessary when using current image generation technology for storytelling (Fotedar et al., 2020). Potentially, the Specificity of the images could have been strengthened by increased human "cherry picking" when linking the images to the text during image-generation.

However, one could argue that there are some exceptions to this conclusion. In Q1.3 from the non-narrative task, which presented an image generated from the sentence, "she won the football game," a large number of people mentioned the words "football" 28/58 (48%) or other sports related activities 25/58 (42%), as opposed to answers that were entirely unrelated 5/58 or (9%). Also, in Task 3, otherwise known as the story writing task, participants did have higher accuracy in mentioning words in the textual prompt used to generate the images. The first image generated with the prompt "I was watching tv one night," included some similar wording from participants to the prompt, as 18/58 (31%) mentioned televisions and 33/58 (57%) mentioned the words "dark" or referencing it being night. Although these words were mentioned, the scenes themselves that they described were not in line with the textual prompt. Many of the answers were referring to individual words in the prompt, but not the prompt as a whole. Since these individual pictures had participants guessing similar words to the text-prompt, it could maybe be explained that having one instance of an image lowers the rate of error in visual literacy, while longer sequences of images caused participants to compose a narrative that diverged from the images and in turn the semantic meaning behind the story. Potentially, having longer sequences of images without textual accompaniment increases error as the narrative increases in complexity and readers perceive it taking diverging paths. However, this is only theorising and no conclusions can be drawn on this hypothesis with the current data.

Furthermore, the aforementioned level of image abstraction could potentially be solidified further with the use of more concrete vocabulary and image-generation prompts. *Concreteness* refers to how much a word, or the degree that the concept of a word refers to a perceptible entity (Brysbaert et al., 2014; Charbonnier & Wartena, 2019). A more in depth account of concreteness and its possible benefits will be discussed in the "Further Results Analysis" portion of the discussion below.

I would argue that this metric should be considered in moderation when increasing Specificity, as part of the strength of what picturebooks can accomplish is a difference between information communicated in modalities when combined (Wolfenbarger & Sipe, 2007). But, this metric should not be entirely disregarded as having images and text that are completely misaligned could potentially be disruptive to the storytelling and confuse the reader. To conclude, I do believe that steps should be taken to raise the low-level of Specificity in the current state of image generation used in the evaluation, but as the technology gets better and Specificity of computer images to text gets higher, I think it is worth studying its value in more human-centered evaluations.

20.2 RELEVANCY RESULTS ANALYSIS

The results seem to indicate that images with accompanying text had higher Relevancy with one another. When evaluating Relevancy of images to text, in Q1.1 and Q1.2 from the non-narrative task, the image and text prompts had a high degree of Relevancy. Q1.1 I asked participants to rate the expression of fear in a picture generated from the prompt, “She was very scared.” The responses had a mean of 7.58/10 in expressing the prompt. Additionally, 45/58 (83%) participants chose the prompt, “Verney was a good student,” correctly from the other 4 prompts when matching the text to the generated image. This shows a high Relevancy indicated by the images expressing the accompanying text correctly. This was again shown in the question, “Do you think the pictures fit the computer’s story?” In Q4.2, from the “A scary ghost” comparison task, where 35/58 (60%) said “Yes,” 18/58 (36%) said “Maybe,” and only 2/58 (3%) said “No.” This supports that the majority of participants believed that when the computer’s stories were presented with text and images together, the images had a high degree of Relevancy with the text prompt.

The indication of Relevancy being reasonably high in regards to text-image pairs in the story-sequences is a good indication that when paired together the images are solidified in their appearance and relation to the text from their abstract form when perceived by the participants. This again ties

back to the idea that the meaning emerges with the interaction of images and text (Birketveit & Rimmereide, 2017). In isolation the pictures seem to be more abstract / Generalized as discussed above in "Specificity vs Generalisation Results Analysis," but with the integration of the textual storylines, the results indicate that the images are interpreted to have high Relevancy to the sentences and represent the text. This supports other literature that concludes that images in a context convey meaning more easily than in isolation (Huang et al., 2016). This observation is important as it could help indicate the importance of text-image pairs when considering image generation at its current stages and has several implications that are discussed in more depth below in the "Further Results Analysis" subsection.

20.3 NARRATIVE CONSISTENCY RESULTS ANALYSIS

Within Task 2, also referred to as the matching task, the pictures did not seem to be representative of the textual story when scrambled and presented to participants. When presented with a textual story and having to match the correct images to them in a narrative context, only one image (4) that was generated from the matching text-prompt was mostly picked correctly. For the rest of the sentences, the coupled pictures were not matched correctly most of the time. This would indicate that all the images did not support the story and the story therefore contained a low Narrative Consistency (Sapkale & Lukin, 2020). Within this evaluation, the textual story was presented first and therefore functions as the Fabula, or raw ingredients, and the image as the Sujet, or the interpretation that the participants had (Sapkale & Lukin, 2020). The Sujet was relatively consistent with participants picking similar pictures for conditions, but it was misaligned with the computer-generated story. This misalignment may indicate that the participants interpreted a story-arch adjacent to the one told by the computer-story. This is interesting as it could show a difference between how computers and humans communicate and pull semantic meaning from text in other modalities. This could potentially warrant future research that will be discussed further in the "Future Work" section 22.

Narrative Consistency is a metric designed to analyse the internal consistency of picturebooks. This metric was adapted in an attempt to understand computer generated images specifically, but it may be flawed in that part of the cohesion comes from image-text pairing. The metric itself relies on the image-sequence to be perceived as a story, but in the act of separating and scrambling the text and image components, this internal coherence and overall Narrative Consistency may have been dis-tangled. I still believe that this metric provides important insights, but it would be worth considering the reliance of images and text as a *Hermeneutic circle*, or a reactive and ever increasing circle of meaning creation through the engagement of unimodal modalities in a multimodal whole (Nikolajeva & Scott, 2013), in picturebooks when drawing conclusions from the results in the evaluation. Potentially, separating the elements in order to understand their individual components may have disrupted this circle as the individual parts are not seen as a whole. This possible disruption of meaning creation is important to consider when further analysing the results.

Despite its possible flaws, one can still make the argument from the evaluation that that Narrative Consistency was lacking. In this case, having low Narrative Consistency seems to indicate that the images and text in isolation do not match up cohesively when not directly associated together in a narrative context. This could potentially be explained by the abstraction of the pictures and their leaning towards Generalization in terms of image-text relations.

20.4 MULTIMODAL FUSION OF IMAGES AND TEXT RESULTS ANALYSIS

From the evaluation, it seems as though people largely believe that the combination of pictures and images did have an effect on the final product. When measuring to what extent, people reported that images affected their perception of the written stories, there was a mean score of 8.12 /10; as seen in Q4.4 from “Not at all” (1) to “Very much” (10), pulled from the “A scary ghost” comparison task. When asked how the images and story interacted in Q4.5 from the “A scary ghost” comparison task, 33/58 (57%) replied that, “the interaction led to a product greater than the sum of their parts,”

and 18/58 (31%) responded that, “the images and text each helped to fill out gaps in the narrative leading to complete experience.” This supports that the multimodal fusion created from the image and text integration was mostly perceived as enhancing or complementary. If combining these possible answers as the two that show the greatest positive attribution of multimodal fusion, then 88% of participants answered that each modality helped the other in creating a full story. This could indicate that the interplay between the two modalities created meaning (Birketveit & Rimmereide, 2017). This would support my hypothesis that computer generated picturebooks are able to fill in the same niche that conventional human picturebooks do by creating meaning “between” the modalities through their complex interanimation, or continuous acting upon one another in the perception of the reader. In this interanimation, one could regard the Sujet as the perception of the participants or “extra” non-physical elements filled in by the minds of the participants when perceiving the Fabula or raw ingredients. The literature argues that when this multimodal fusion or interplay between modalities occurs, its through a hermeneutic circle of meaning creation through participants “pulling” information through each medium in combination with their personal beliefs, experiences and understanding of the world (Nikolajeva & Scott, 2013). If these elements are present, then it would potentially be possible to use computer-generated picturebooks to improve visual literacy in a teaching setting or at the very least creating a better entertainment product than a unimodal story variant.

To further engage with these results, we should also examine how much participants believed their perception of text changed when combined with images. In Q4.2 from the “A scary ghost” comparison task where participants had to write their own story before seeing the computer version, many of the answers just under 50% of answers indicated that it changed their perception. For this question, 26/58 (45% said “Yes”), that the images changed their perception, 13/58 (22%) said “Maybe,” and 17/58 (29%) said “No.” Most people responded with “Yes” or “Maybe”, as opposed to “No”, which would be a weaker argument than Q4.5, but again indicate the images and text were seen as a story and affected each other’s meaning. This evidence is supporting of how multimodal media, specifically images when acting on text in picturebooks, are able to present narratives as a greater whole by each

acting on one another and enhancing their unimodal counterparts. This supports the hypothesis and the literature that this instance of multimodal fusion leads to a better outcome than the unimodal information (Shareha et al., 2009).

However, it is important to consider the contrast of answers from Q2.3, where many participants reported that their initial perception did not change when seeing the combination of images and text in a story. Q2.3 from the "The new job" evaluation task, involved first seeing a story as text in isolation and then introducing pictures in the matching task. When asked if seeing the images changed their perception, 22/58 (38%) said "Yes," (12/58) 21% said "Maybe," (24/58) 41% said "No." This could potentially be an indication that the pictures did not change their initial visualisation in their mind's eye. Potentially, this could be because they had already established a visualisation of the story that was already ingrained and functioned as a filter for perception. These results could be an indication that the initial perception gathered by the participants was withheld to some degree after the integration of another modality. Another explanation that is worth mentioning, was that the complete story with text and images was never shown to the participants and it remained fragmented and aligned with their perceived image-sequence after analysis. Without showing the text and images together as the computer generated them, it may not lead to association through Relevancy and potentially they would retain their initial perception. To conclude, although the results from the evaluation are slightly inconsistent, there is still more evidence in support of multimodality enhancing or complementing narratives, than evidence contradicting it.

20.5 PUTTING IT ALL TOGETHER: FURTHER RESULTS ANALYSIS

The results seem to indicate that the images are abstract and Generalized in their representation and images lose Narrative Consistency with textual stories when scrambled. However, when computer-generated images and text are presented together as matching pairs, they seem to show a high degree of Relevancy and there is support that the integration of images and text as a multimodal narrative

enhances or complements both modalities. From the analysis of the results, I would argue that generated images and text do influence and add to each other when integrated as a multimodal entity in line with the literature (Nikolajeva & Scott, 2013; Wolfenbarger & Sipe, 2007). As mentioned in the section, "Specificity vs Generalization Results Analysis" above, this could potentially be due to the perception of individuals helping to focus the abstraction of the computer generated images. A possible explanation backed by consistent results in the psychological community is that *Priming*, or how one stimuli affects perception of another subsequent stimuli, can affect ambiguous images (Goolkasian & Woodberry, 2010). Potentially, the text is a more concrete representation than the more ambiguous imagery and through this relation when combined in image-text pairs helps to solidify the ambiguity of the images as a specific picture of the text in the narrative context. Research by Goolkasian and Woodberry, found that primes that were loosely tied, indirect and focused on the semantic relationship between the two stimuli were more successful in priming (Goolkasian & Woodberry, 2010). In opposition, when focusing on the physical characteristics of the ambiguous figure this effect was lessened and the priming was less effective (Goolkasian & Woodberry, 2010). This is consistent with other priming literature and could indicate that the textual-story functions as priming for the perception of the imagery to fit the narrative better. This would help to explain the images out of sequence leading to poor Narrative Consistency when the images are presented without their priming-textual pairs, but there in turn being a high Relevancy when picture-image pairs are presented as a complete narrative or story-segment. If not primed with text, participants could be more inclined to view the images more closely leading to different interpretations of the images than when accompanied with a primer-sentence. Potentially the interplay between ambiguous elements in the images combined with the priming can lead to a deeper story in the interpretation of participants and may be an alternative way to explain the image and text interaction.

It does seem as though pictures were able to express emotions quite vividly, for example in the representation of a scared woman. This is in line with some of the strengths of images as a way to express emotions and setting. Additionally, even though participants were not able to use the images

to write a similar textual story to the computer, the general theme was captured quite well. Many people in Q3.1 from the story writing evaluation task, indicated that they thought the story was about something horror related, using words like: “ghosts”, “monsters” and/or “horror.” The fact that results indicate that the images were quite Generalized in what they displayed but were still effective in strengthening the fusion when integrated, may be evidence to support that computer generated images do not need high levels of Specificity to still provide a benefit when a textual narrative is present. This is an important consideration when engaging with the technology. This could potentially be explained by the idea that pictures and images are never entirely representative of the same thing, as the two modalities are intrinsically different in how they represent story elements (Wolfenbarger & Sipe, 2007). Perhaps, this combined with the type of Multimodal Fusion metric results would indicate that having high Relevancy between images and text is sufficient in the creation of computer-generated picturebooks where images and text work together to strengthen the narrative.

My research accounted for concreteness by creating a requirement for a visual representation of characters in images mentioned in the textual prompt during image generation; as seen in the “Image Generation Checklist.” This was to filter for and align with the character classification accuracy metric outlined in the literature review (Zakraoui et al., 2021). However, this could potentially be strengthened further by having more objects related to the environments and prompts. It seemed as though images that contained objects mentioned in the text prompts had higher degrees of Narrative Consistency and Relevancy as opposed to non-object images-text pairs. As seen throughout the evaluation, but most specifically in the image-text pair where the word, “time” was visually represented as a clock (the majoritively correctly picked image in the matching task, Task 2) and, “student” by a school uniform and books (in Q1.2 where 78% of participants picked the proper prompt from the non-narrative task). These were matched correctly with a high accuracy and this may be due to the concrete visualised objects being represented to express words in the story-sentences. This could mean that concreteness of sentences and representation thereof leads to a stronger Relevancy of images and text together in a story context. Concrete expression is a common metric seen in the literature, and

it could potentially strengthen Relevancy of text and image interaction. Concreteness could also be another possibility for the lack of Narrative Consistency, since the character was not distinct in their appearance throughout the story. Having more concrete character descriptions could potentially help to solidify the concreteness and heighten the score of these metrics (This is discussed further in "Future Work" 22 below).

The literature also argues that one of the most important things to consider when engaging with picturebooks is the message of the author when creating the work. The intrigue of this author message is that computers themselves are integrating huge amounts of human created data into the creation of these narratives. If there are messages that are to be communicated by authors in narrative, and particularly picturebooks, it is interesting to consider "who" the author becomes. Can the computer be considered the author? Is it the congregation of the data that the models were trained on? Or is it the person who wrote the prompts and combined the images with the text? Or potentially the lack of an author with a direct message means that the stories allow for further personal engagement and self-reflection when engaging with computer-generated media. If this is the case, it could be argued that the differences in perception are important self-reflective practices that serve as introspective exercises while simultaneously strengthening visual literacy. These types of questions are philosophical and potentially semantic in nature, but they are important considerations if one considers the author's message an important component in media communication. Engaging further in the relation between a message and computer authorship may lead to a better understanding of why this technology is important and what its future implications could be. Additionally, this may help us to understand human-computer communication further, which is an important avenue to consider as they become a larger part of our lives.

All in all, the implications of this evaluation seems to be that computer-generated textual stories could benefit from the integration of images into their narrative. Even though the current technology is in its infancy and the evaluation did not contain the yet to be released state-of-the-art in image or language generation, the integration was still reported by participants to benefit both modalities. I

believe from these results and analysis, I can reasonably conclude that the integration of computer generated images and text stories could be beneficial in further application and requires more attention and research.

21

LIMITATIONS

Upon concluding the analysis, it is important to discuss several limitations that could potentially have influenced the results. First of all, as discussed above in the "Narrative Consistency" discussion section 19, the metric may be flawed in how it was evaluated. This metric is difficult to distinguish as the idea of internal cohesion and Narrative Consistency are quite subjective concepts. Having a condition that measures these metrics within a narrative sequence of image-text pairs, and not just a scrambled text and image condition, may be beneficial in strengthening the conclusions drawn from the metric. Also, having more stories and conditions would allow for stronger conclusions to be drawn in further research. Two stories may not be enough as it varies so greatly between generated images as they are computer generated and each prompt has an almost limitless amount of possible results. Also, as the pictures and images are a small sub-sample of the amount of possible images, this may impede on future replications of results. Potentially implementing a seed option in future iterations of the image generation would allow for easier replication of results. Also, the picking of the images with some manual help does mean that there was some bias in the pictures used for the evaluation task. The checklist was filled out individually by me and was not consulted by a third party before entering it into the evaluation.

Although the evaluation was a measure of integration of images and text, the participant tasks were to write prompts for images or match images to prompts. This could potentially cause problems as the tasks were never focused on the creation of images; for example, having participants draw

images. I avoided this task as it would be difficult to implement and might make some participants feel inadequate when presented with drawing tasks, but this could potentially cause problems as it may influence how participants engage with the media. This could potentially be combated through visualising text as images in their mind and comparing the visualised image to the one displayed in an evaluation or writing a description of it. However, this could be difficult to implement as it may require prior training for visualisation and accurate reporting.

In the case of the evaluation, the combination of pictures and images were assumed to already be considered fully fledged stories when consumed by the participants. This assumption could potentially cause problems as it might contain a pre-asserted bias to view the images and sentences as a whole product as opposed to separate elements. This may have influenced results of the integration between the two elements to follow this bias. Additionally, the use of the word "picture" instead of the word "image" may denote that the images are representative of the text. Using the word "image" in the evaluation may have helped eliminate this bias. Another possible problem is that the environment that the participants were doing the evaluation is not controlled. This means that some may have done it quicker and not paid as much attention to the evaluation. Having the reading and evaluation done in more controlled conditions could potentially help the participants engage with the materials in a more controlled way.

In regards to the participants, it would have been beneficial to evaluate whether people have visual problems such as the need for corrective lenses. This could potentially cause differences in evaluation, especially since the computer-generated images tend to be blurry. Also visual literacy was not accounted for and could be important as this metric may have shown how individuals are able to view images in a specific context of understanding that may have an effect on their narrative perception when engaging with the image-text pairs. I decided not to ask for personal data in the evaluation form, but its inclusion could potentially have allowed for further analysis if included in the future.

22

FUTURE WORK

Even considering the limitations, the results are promising and warrant future research. Potentially having a concreteness score for each story and a ranking system would help further understanding in how concreteness in textually generated stories could help to strengthen associated images. There is research into concreteness ranking that could be applied, for example Charbonnier Wartena, in their paper, "*Predicting word concreteness and imagery*", have created a metric that they report can predict word concreteness using word embeddings with a strong linear correlation with human evaluation results (Charbonnier & Wartena, 2019). Generating many stories, ranking them by a concreteness score and picking the top ranked stories could integrate more perceivable text artefacts that may translate into images and increase metrics such as Specificity and Relevancy.

For future work, it would be interesting to use newer versions of image generation. OpenAi's Dall-E (Ramesh, Pavlov, Goh, et al., 2021) for example, has produced impressive results that would likely make sharper and clearer representations of pictures than Bigsleep. Additionally, Latent2vision and Latent3Vision are currently the newest version of open-source image generation architecture by the creator Advanoun (the creator of Bigsleep), and was just released a week ago at the time of writing this paper [1]. An example of both of these technologies can be found in Fig. 28 below. From a subjective perspective, I believe these newer generation architectures seem to carry Specificity of prompts better than Bigsleep and may lead to better results in metrics such as Specificity. As stated

1 https://twitter.com/advadnoun

above in the “Specificity vs Generalization Analysis” 20, if the Specificity turns out to be very high in better technology, it may be beneficial to study how a varying gradient of Specificity vs Generalized images changes Multimodal Fusion of narratives in human participants.

Figure 28: ABOVE: Generated by OpenAi’s DALL-E with accompanying prompt. Retrieved from (Ramesh, Pavlov, Gabriel, et al., 2021). BELOW: Generated by Latent2Vision from prompt “A soliloquy to the past”. Retrieved from, (“A Soliloquy to the Past [Latent2Visions]”, 2021)

Furthermore, one of the limitations of the technology used in the paper, is that the sequence of pictures does not carry over the likeness of characters established earlier in the sequence. Having a system to hold onto context in some way and carry it throughout the image generation may lead to a more recognisable Narrative Consistency, as the characters would be more consistent in their appearance throughout the entirety of the story. This could potentially be done through a textual character "look" generation system that would then carry adjectives into the prompt when generating characters. This system could potentially scrape the story for adjectives about the character and then substitute the character's name and/or pronoun for a generated description. For example, this could replace the name "Tom" with, "a tall, skinny, postman with brown hair, blue eyes and a letter-bag hung over the shoulder." Another possibility of using this technique, could be pulling concrete and associated words from more abstract words; such as associated words to School, like book, school-uniform and pencil. This could have a substitution method where the associated and more concrete words are used as input/prompts when generating, potentially leading to more concrete representations when generating pictures. Potentially, another possibility would be to combine the architectures, such as Bigsleep and Plan-and-write, into one single architecture that both generates the images and text.

Another potential way to strengthen narrative cohesion could be to create an overarching filter or similarity in pictures that could help associate sequences as one single entity. Potentially applying a filter or a UI when interacting with the pictures would maybe help to increase the cohesive experience between images and text. This would again be drawing from the idea of *Peritext*, the elements that surround the story and give it contextual meaning (Wolfenbarger & Sipe, 2007), or the integration of the context by which the story elements are integrated as discussed in the "Literature Review" 4. Since part of the intrigue gathered through the picturebooks is the communication and engagement of human and computer intelligence through stories, providing a context to view the picturebooks could be an interesting way to study the relationship. This could potentially involve printing and binding the picturebooks to separate them from a computer screen or maybe integrating the machine generated

pictures with human drawings to further study the influence that the computer generated pictures have as a product.

As this is a preliminary test in researching multimodal story integration of computer generated media, future research could focus on creating more engaging and varied stories. Using newer and larger NLG transformers like GPT3 would likely lead to more engaging stories. Since the GPT3 architecture is trained on more data and seems to be better in every metric that it has been tested on when compared to GPT2, using the better architecture may help strengthen stories. This could also potentially allow for longer range stories, potentially using GPT3 combined with Plot Interpolation to control definite beginnings and endings, would allow for more machine creativity in the creation process and in turn a more varied possibility of generated stories. One of the major hindrances of the current story-planning software is the pre-planning limiting creativity and interestingness of the stories. Having a larger amount of creativity in the textual stories could potentially lead to a more varied and interesting scope of subjects, characters, images generated and in turn picturebooks. Finally, since "the pictures and stories together create something greater than the sum of their parts" option was the most actively picked selection of type of Mulitmodal Fusion, finding a way to measure this relationship further would be very interesting. For example, this could be exploring how the same pictures with different text, or vice-versa, would change story comprehension and fusion. Also, to see what type of meaningful experience would be had when engaging with similar elements scrambled together as image-text pairs. This would also be an interesting way to test the priming theory and test whether abstractness of images are solidified in a narrative context through text. Using this as a way to analyse visualisations of images as compared to the computer images of the same prompt may help to lead to an understanding of how computer and human participants differ in their visualisations of multimodal integration of images and texts in narrative stories.

Part VII

CONCLUSION

23

CONCLUSION

From what was gathered in this paper, I would conclude that further research into computer generated picturebooks would be beneficial for various fields of research. Even considering that there were several limitations, the results still indicate that although Specificity and Narrative Consistency were low, the Relevancy was quite high and the type of text and image integration was helpful in improving aspects of the storytelling. The evaluation supported that computer-generated picturebooks were also effective in Multimodal Fusion that aided both modalities. This shows that even though this technology is in its infancy, it can still create multimodal media that leads to a product worth studying. The integration of various disciplines helped to highlight some of the strengths and weaknesses of the few ways that data-scientists have been evaluating these types of interactions. I would emphasise that although more costly and subjective, integrating a cross-disciplinary human evaluation of these types of technologies helps to highlight the human aspect and application of this type of media in its current state and in the future. As far as I understand, this is the first type of cross-disciplinary evaluation of computer-generated picturebooks, but hopefully not the last.

The application of this type of technology would potentially be possible in the classroom. As the integration of multimodal media in many countries is gaining traction as a way to teach many aspects of reflection, such as visual literacy, the use of computer-generated picturebooks may create more instances of multimodal media to study and also help to engage youth in human-computer communication. Other possible applications are creative media. The combination of image and text is

used in many avenues that we see everyday like advertising and art. These creative media have been dominated by human content, but may soon see a rise in computer-generated content with the growth of NLG and image generation.

To reiterate, image generation is in its infancy but has recently seen a huge evolution with the advent of publicly available technology like CLIP and the currently studied Dall-E. These technologies make it possible to generate incredibly accurate representations of text. Also, its worth considering that Dall-E is in its first iteration, if Dall-E 2 is as big a leap forward as GPT2 to GPT3, the already impressive results will likely be a huge leap forward in terms of quality. There are so many ways to engage in future research of this kind, through increasing Specificity of image generation in a narrative context, to generating physical character traits to keep consistency across narratives, or integrating new technology. While researching for this paper there were new growths and developments and improvements on the technology almost every week, so there may be entirely new technologies available by the time I am done writing that we are not even aware of yet.

Perhaps one of the most interesting aspects of computer-generated media are the relations that it develops between intelligent systems and people. As stories are such an old art-form that helps define us as humans, what will it mean when/if computers are able to do it as well or better. Would the advent and popularisation of this type of media change the way we understand and interact with computers? Will computers be able to express the human experience as well as humans themselves can? If computers are able to express emotion as well as us within a storytelling context, it may begin to break down barriers of how computers interact with us socially. The idea of a voice assistant like Siri describing its day of being in my pocket would make it difficult to distinguish between my smartphone and talking to another human on the phone. Creating these types of stories and embracing them in a classroom setting would probably increase understanding and relationships of the next generation with this type of technology. The negative possibilities are also immense. This technology could potentially put many graphic designers and copywriters out of business if the technology gets so good that it can create human-like media. Computers only need to be invested in once and thereafter

do not need a wage. Therefore, in the wrong hands this technology could be detrimental to many people's professional careers. Some people may be reduced from creatives to filtration systems/quality assurance technicians for powerful machine learning architecture.

It is also interesting to witness all the research going on into Dall-E and how multimodal integration can expand how we think about computational systems as "understanding" what they are writing. Current research seems to be indicating that Dall-E has developed something similar to human multimodal neurons which may be one of the strongest arguments yet for a computer system "understanding" language and images literally, symbolically and conceptually (Goh et al., 2021). This is a huge finding and may be the first step towards general intelligence, as one could argue that these intelligent systems are beginning to resemble the way we humans think more and more. I would predict that the next iterations of GPT-like architecture will begin to add multimodal components, as they are already studying these types of systems and it seems to be adding benefits to machines as much as it does to people. However, as these systems become increasingly intelligent it does put into question whether this type of technology should be applied in the classroom. Many qualified academics are arguing for the dangers of future AI as an existential threat. This used to feel like a Science fiction novel or at most a far off future, but the field of AI is currently moving so quickly that researchers barely have time to study what is currently available before the next big iteration is released. Multimodal fusion may be one of the keys that finally brings this technology into the next definite era. Will it be beneficial to encourage computer generated multimodal fusion based products to increase our communication and understanding of intelligent systems for future generations, or will this serve as a way to desensitise youth to the dangers of AI and make the apocalypse more approachable through playful applications of potentially dangerous technology.

Part VIII

APPENDICES

24

APPENDICES

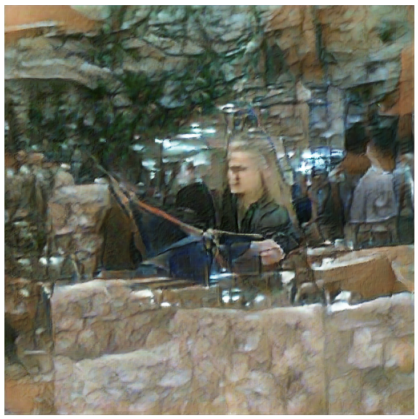
The Appendix will contain screenshots of the full evaluation form, with each picture taken individually and the text and image checklists. A full clone version of the evaluation form can also be found in this link: (https://form.jotform.com/211254776600351).

Can a computer make a picturebook?: Story evaluation form

Thank you so much for taking the time to fill out this form. This is designed as a way to evaluate how computers generate picturebooks. In other words, this is intended to evaluate computer-capabilities and not your capabilities. This evaluation is intended to be a fun and creative way to give us valuable data on how computer created images and sentences work together to make stories.

So what does a computer picture-book look like?

Each story created by the computer is 5 sentences with a beginning, middle and end. From each sentence a picture is "generated" or created by the computer to accompany the sentence like that in a children's picturebook. Below is an example of a single picture and its corresponding sentence from one of these stories.



"Legolas went shopping at the mall"

Your task

For this evaluation you will be given various tasks associated with identifying story elements in pictures and/or text. This evaluation requires some creativity and writing. You can spend as much or as little time on each task as you are willing. There is a progress bar at the top of the form that functions as a rough guide of how far you are through the evaluation. It will on average take around 10 minutes to complete. Have fun!

Your information

All the information you submit is entirely anonymous and will not be traced back to you. All information gathered will be entirely for educational purposes and deleted upon completion of the project.

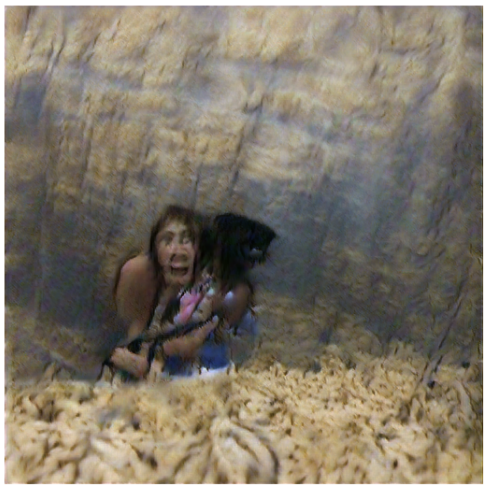
Figure 29: Opening Page

Task 1

Task 1

Within this task you will be presented with 3 different ways to evaluate stories. Each picture is entirely independent from the others. All the pictures and sentences were created by a compute stories.

Figure 30: Task 1 Description



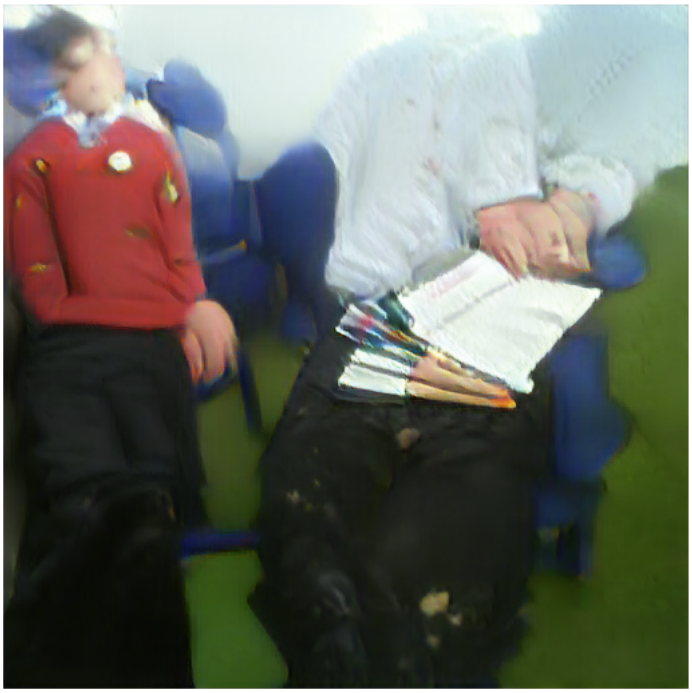
Q1.1 - On a scale from (1-10), how much does the picture above express "She was very scared". *

1 2 3 4 5 6 7 8 9 10

It does not look like that at all

It looks exactly like it

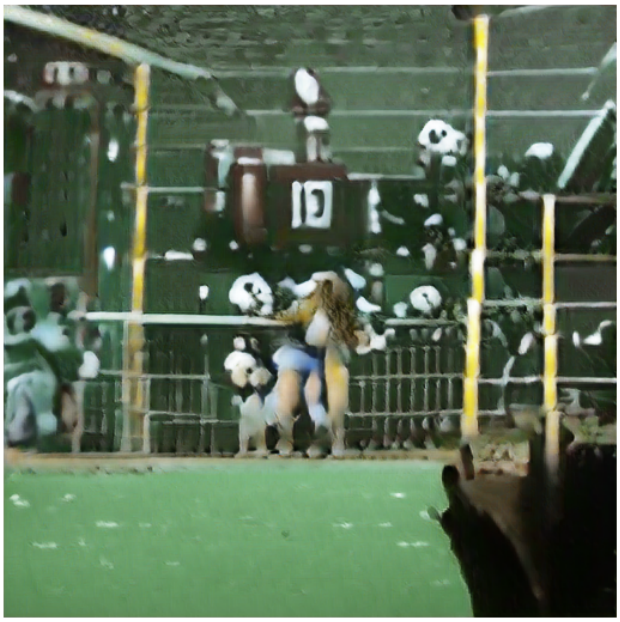
Figure 31: Q1.1



Q1.2 - Which sentence best describes what is happening in the picture above? *

- It was a great camping trip.
- Ryan had dreams of becoming a singer.
- Verney was a good student.
- He would greet people.
- He eventually made a lot of people's ears.

Figure 32: Q1.2



Q1.3 - What do you think the picture above shows? (In one sentence) *

Figure 33: Q1.3

Task 2

Task 2

On the next slide you will be presented with a complete 5 sentence story in its entirety. After you read the story you will be presented with one sentence at a time from that story, and a choic
that one sentence. Your task is to select the picture that you think best fits the presented sentence. Each sentence pairs with one picture.

Figure 34: Task 2 Description

The New Job

Tom wanted a new job.
He decided to get a job.
He didn't have enough money.
He spent a lot of time looking for a job.
He found a job that paid well.

Figure 35: Task 2 Textual Story

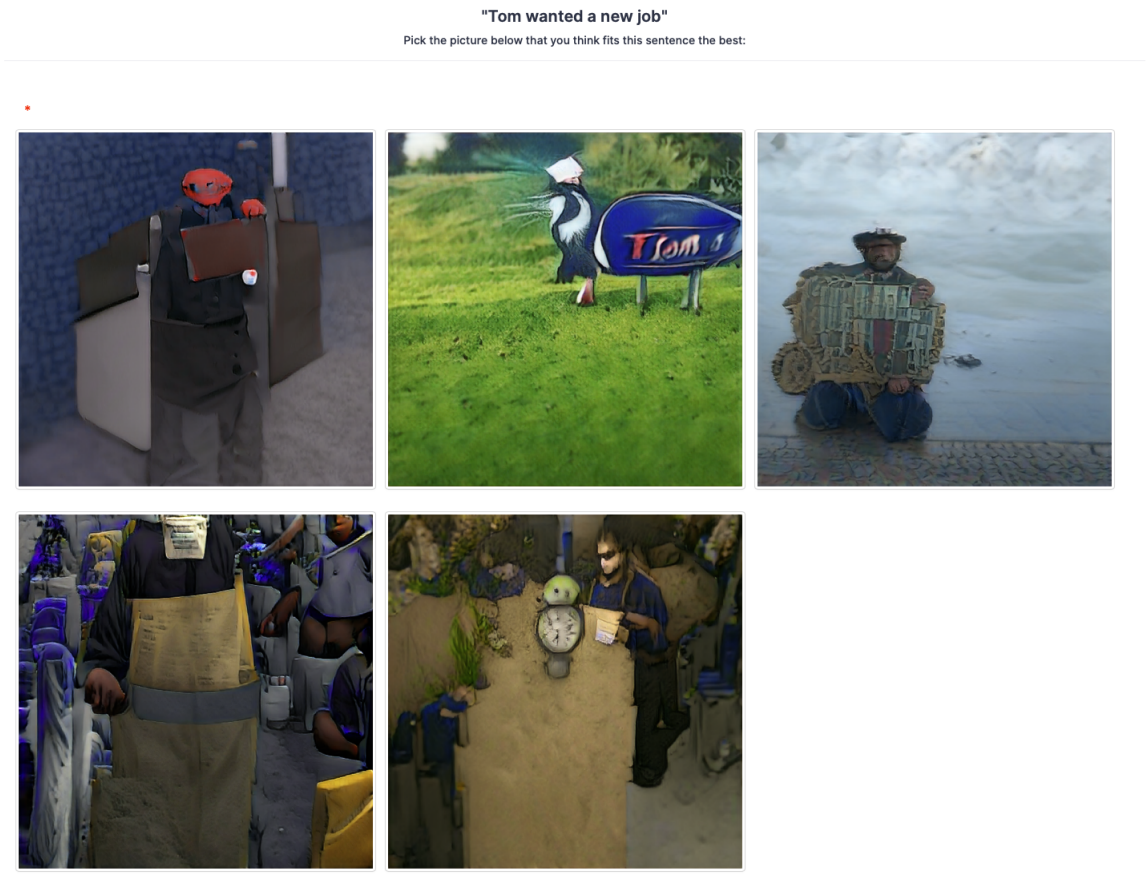


Figure 36: Task 2, Image sequence 1

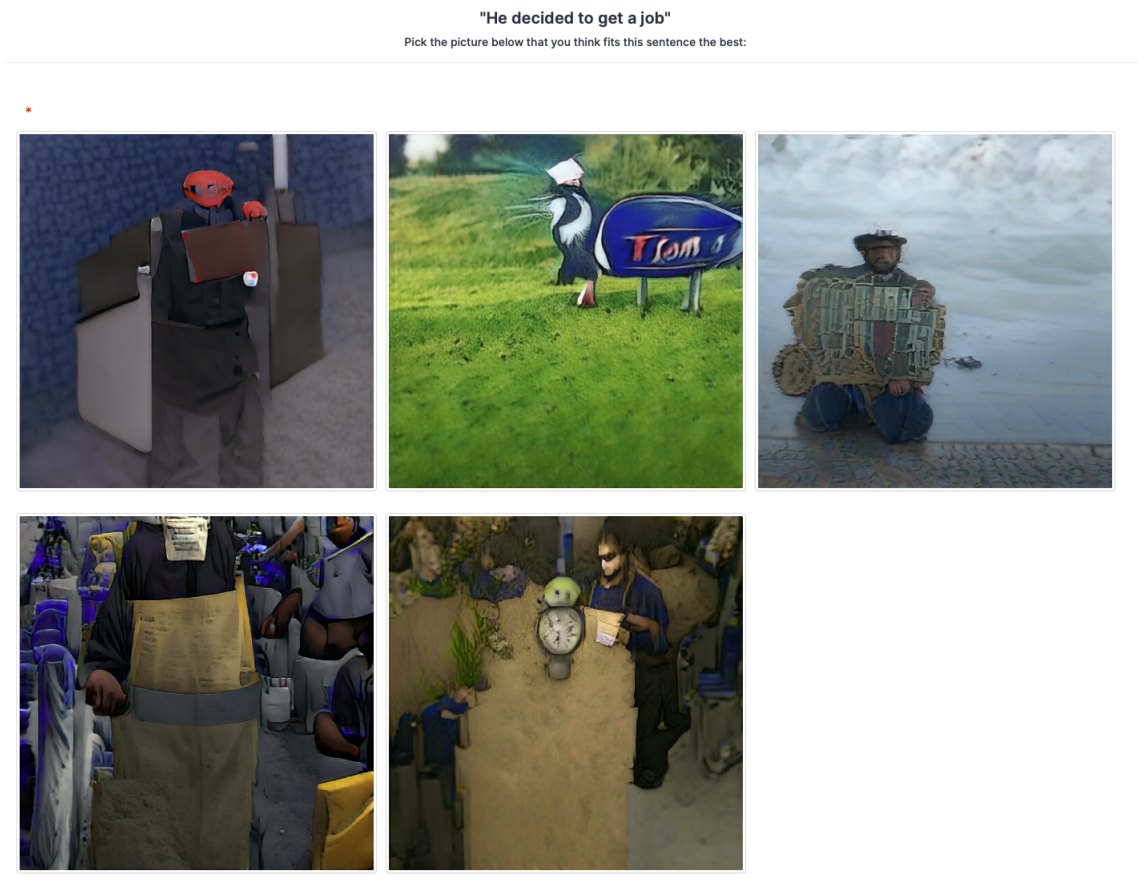


Figure 37: Task 2, Image sequence 2

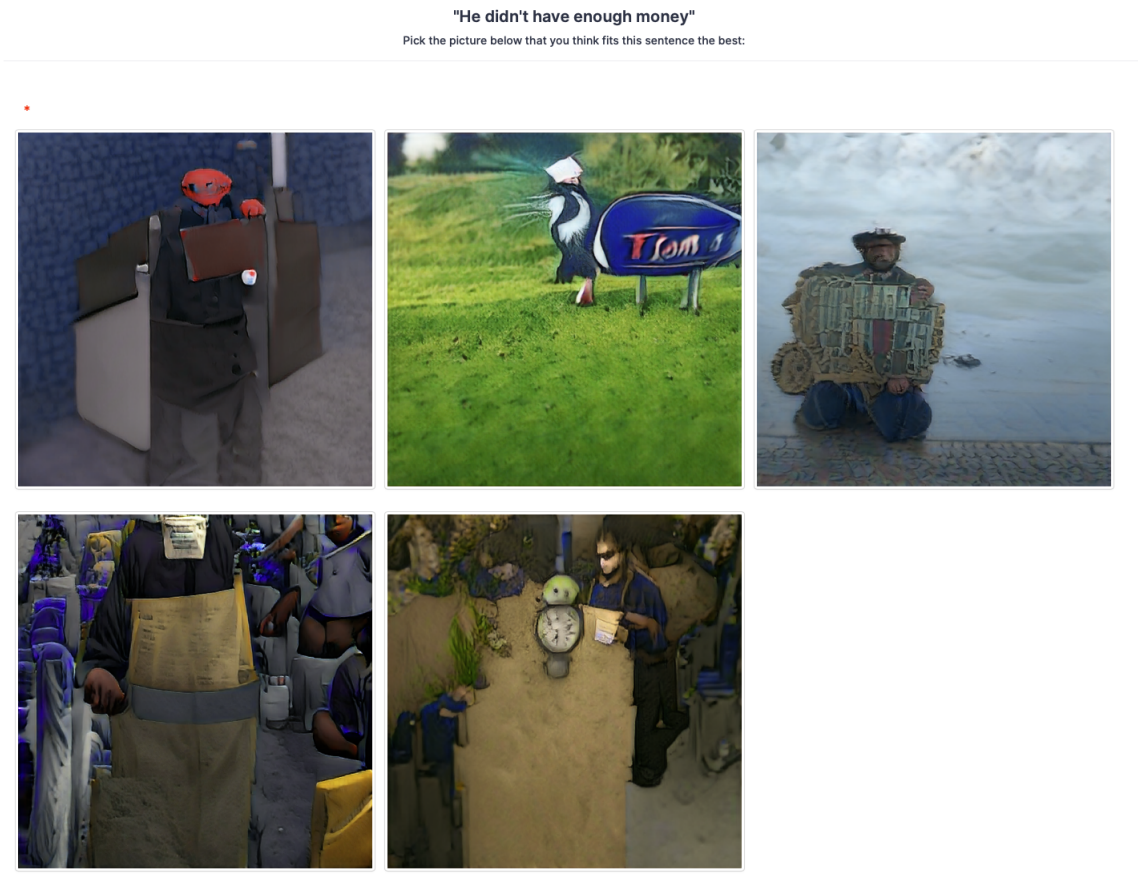


Figure 38: Task 2, Image sequence 3

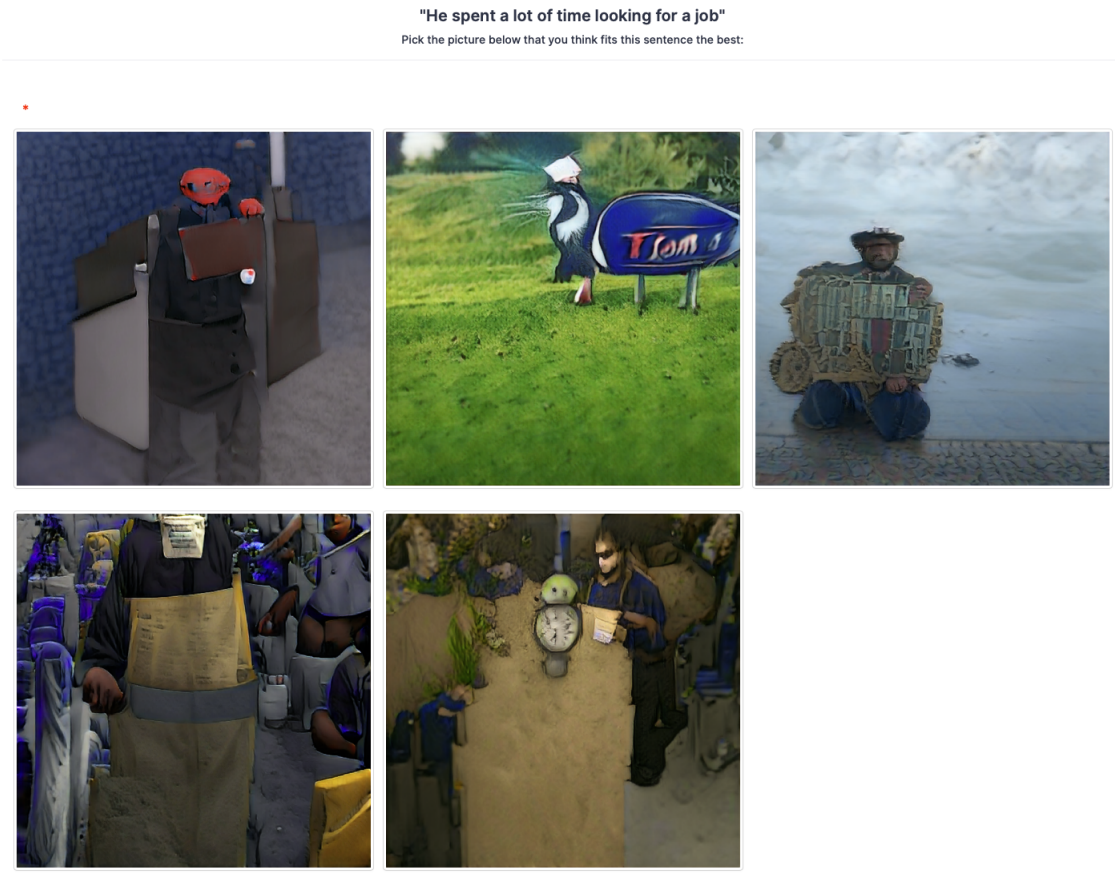


Figure 39: Task 2, Image sequence 4

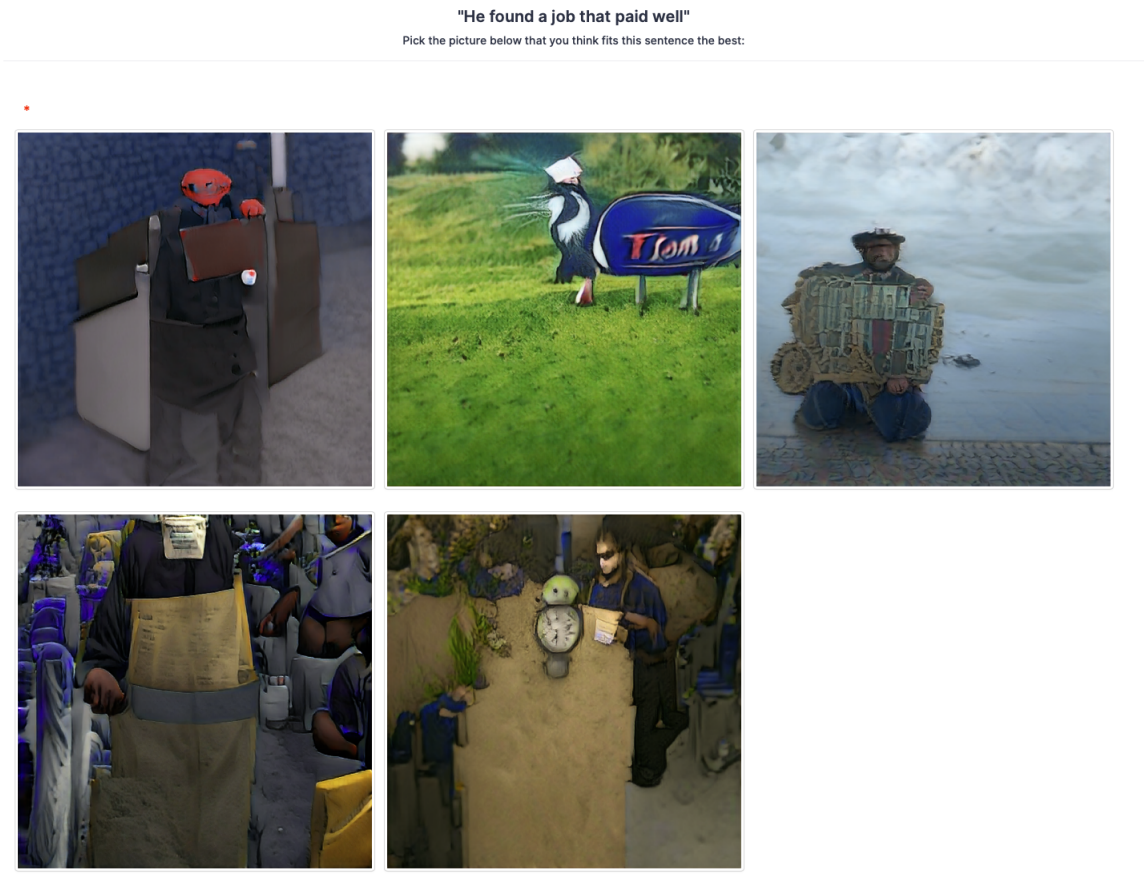


Figure 40: Task 2, Image sequence 5

Figure 41: Q2.1

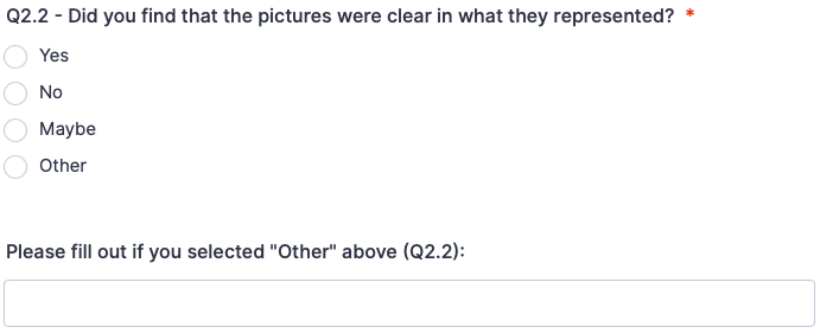


Figure 42: Q2.2

Task 3

Figure 43: Task 3 Description

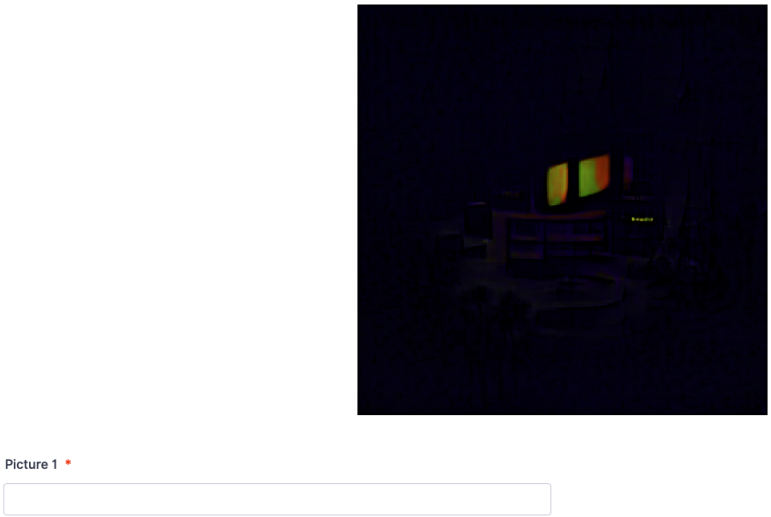
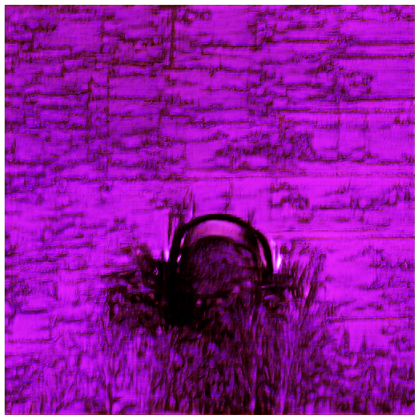
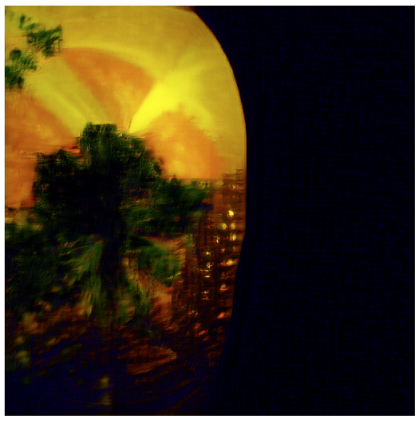


Figure 44: Task 3, Short-answer 1



Picture 2 *

Figure 45: Task 3, Short-answer 2



Picture 3 *

Figure 46: Task 3, Short-answer 3

Picture 4 *

Figure 47: Task 3, Short-answer 4

Picture 5 *

Figure 48: Task 3, Short-answer 5

Well done!

Below are a couple questions on the story you created.

Figure 49: Task 3 Evaluation Description

Q3.1 - In a sentence or two, what do you think the story is about? *

Figure 50: Q3.1

Q3.2 - Did you find that the pictures were clear in what they represented? *

- Yes
- No
- Maybe
- Other

Please fill out if you selected "Other" above (Q3.2):

Figure 51: Q3.2

Task 4

Now you will see the computer generated story

At the bottom you can compare it to your interpretation in Task 3!

Figure 52: Task 4 Description

I was watching tv one night.

Figure 53: “A scary ghost,” Computer Story, Text-Image pair 1

I heard a loud noise.

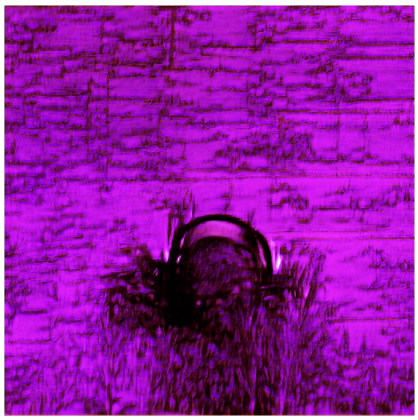


Figure 54: “A scary ghost,” Computer Story, Text-Image pair 2

I looked outside.

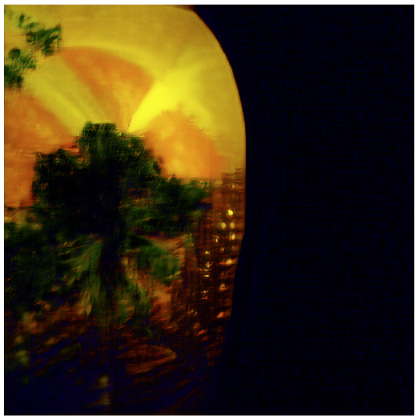


Figure 55: “A scary ghost,” Computer Story, Text-Image pair 3

I turned on the flashlight.

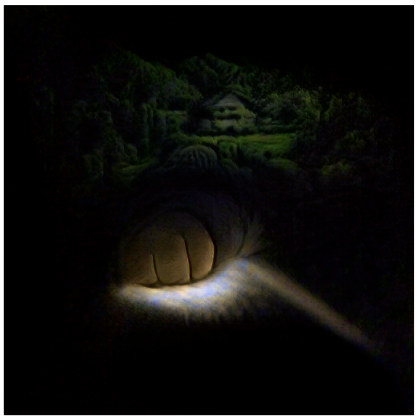


Figure 56: “A scary ghost,” Computer Story, Text-Image pair 4

It was a ghost.

Figure 57: “A scary ghost,” Computer Story, Text-Image pair 5

Q4.1 - How similar was your interpretation of the story to the one made by a computer? *

1 2 3 4 5 6 7 8 9 10

Absolutely no similarity

Exactly the same

Figure 58: Q4.1

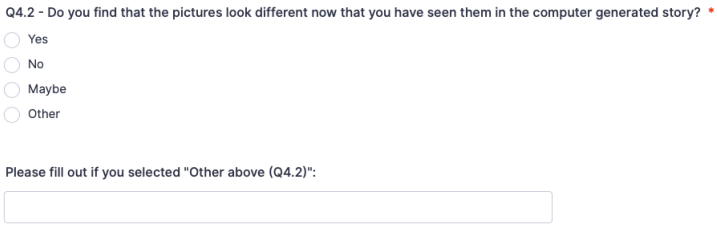


Figure 59: Q4.2

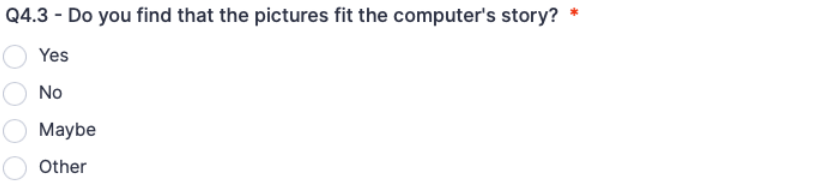


Figure 60: Q4.3

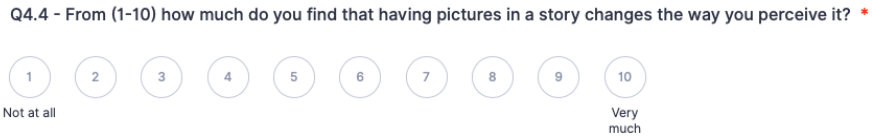


Figure 61: Q4.4

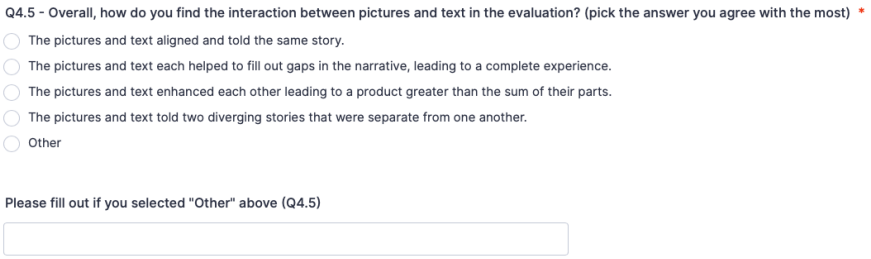


Figure 62: Q4.5

Figure 63: Optional Final Comments Section

Screenshots from word documents checklists made to filter content:

Text checklist:

Checklist for images:

1. It does not resemble the starting image drastically
2. If there is a character in the sentence, there must be a resemble of a character presented in the image
3. The image must be relatively sharp
4. The image must not be the sentence itself spelled out

Figure 64: Textual Story Generation Checklist

Image Checklist:

Textual story generation checklist:

1. Is the same sentence repeated more than once?
2. Is the same word used several times in a single sentence?
3. Is the story completely dissociated with the input title?
4. Does the story have sentences that do not make any sense?
5. If the story has a main character, are they present throughout the narrative?

Figure 65: ‘Image Generation Checklist

REFERENCES

Alabdulkarim, A., Li, S., & Peng, X. (2021). Automatic story generation: Challenges and attempts.

Alammar, J. (August 12, 2019). The illustrated gpt-2 (visualizing transformer language models) [Accessed: 2021-05-25].

Birketveit, A., & Rimmereide, H. E. (2017). Using authentic picture books and illustrated books to improve l2 writing among 11-year-olds. *The Language Learning Journal*, *45*(1), 100–116. https://doi.org/10.1080/09571736.2013.833280

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*. https://arxiv.org/abs/2005.14165

Brysbaert, M., Warriner, A., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *BEHAVIOR RESEARCH METHODS*, *46*(3), 904–911. http://dx.doi.org/10.3758/s13428-013-0403-5

Bus, A. G., van IJzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*, *65*(1), 1–21. https://doi.org/10.3102/00346543065001001

Callow, J. (2020). Visual and verbal intersections in picture books – multimodal assessment for middle years students. *Language and Education*, *34*(2), 115–134. https://doi.org/10.1080/09500782.2019.1689996

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *CoRR*, *abs/2006.14799*. https://arxiv.org/abs/2006.14799

Charbonnier, J., & Wartena, C. (2019). Predicting word concreteness and imagery. *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, 176–187. https://doi.org/10.18653/v1/W19-0415

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. *International Conference on Machine Learning*, 1691–1703.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009a). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009b). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Esfahani, S. N., & Latifi, S. (2019). A survey of state-of-the-art gan-based approaches to image synthesis. https://doi.org/http://dx.doi.org/10.5121/csit.2019.90906

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation.

Fotedar, S., Vannisselroij, K., Khalil, S., & Ploeg, B. (2020). Storytelling AI: A generative approach to story narration. *Proceedings of AI4Narratives - Workshop on Artificial Intelligence for Narratives in conjunction with the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI 2020), Yokohama, Japan, January 7th and 8th, 2021 (online event due to Covid-19 outbreak)*, *2794*, 19–22. http://ceur-ws.org/Vol-2794/paper4.pdf

Frederico, A. (2013). The construction of meaning in three fairy tale enhanced electronic picturebooks. *Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI*. https://doi.org/10.29173/cais713

Frey, B. (1998). Graphical models for machine learning and digital communication.

Frey, B. J., Hinton, G. E., Dayan, P., et al. (1996). Does the wake-sleep algorithm produce good density estimators? *Advances in neural information processing systems*, 661–670.

Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children's story comprehension and recall. *Reading Research Quarterly*, *28*, 264–276.

Goh, G., Voss, C., Amodei, D., Carter, S., Petrov, M., Wang, J. J., Cammarata, N., & Olah, C. (2021). Multimodal neurons in artificial neural networks [Accessed: 2021-05-25].

Goldfarb-Tarrant, S., Chakrabarty, T., Weischedel, R. M., & Peng, N. (2020). Content planning for neural story generation with aristotelian rescoring. *CoRR*, *abs/2009.09870*. https://arxiv.org/abs/2009.09870

Gonzalez-Rico, D., & Fuentes-Pineda, G. (2018). Contextualize, show and tell: A neural visual storyteller.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Goolkasian, P., & Woodberry, C. (2010). Priming effects with ambiguous figures. *Attention, perception amp; psychophysics*, *72*(1), 168–178. https://doi.org/10.3758/app.72.1.168

Gregersen, T. (1974). *småbørnsbogen," in børnel-og ungdomsbøger*. Copenhagen: Gyldendal.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, *abs/1706.08500*. http://arxiv.org/abs/1706.08500

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., & Mitchell, M. (2016). Visual storytelling. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1233–1239. https://doi.org/10.18653/v1/N16-1147

Hughes, M., Salway, A., Jones, G., & O'Connor, N. (2007). Analyzing image-text relations for semantic media adaptation and personalization. *Hughes, Mark and Salway, Andrew and Jones, Gareth J.F. and O'Connor, Noel E. (2007) Analyzing image-text relations for semantic media adaptation and personalization. In: SMAP 2007 - Second International Workshop on Semantic Media Adaptation and Personalization, 17-18 December 2007, London, UK.* https://doi.org/10.1109/SMAP.2007.43

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lewis, D. (2001). *Reading contemporary picturebooks: Picturing text*. Routledge.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Lu, S., Zhu, Y., Zhang, W., Wang, J., & Yu, Y. (2018). Neural text generation: Past, present and beyond.

Mol, S., & Bus, A. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological bulletin*, *137*, 267–96. https://doi.org/10.1037/a0021890

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. F. (2016). A corpus and evaluation framework for deeper understanding of commonsense stories. *CoRR*, *abs/1604.01696*. http://arxiv.org/abs/1604.01696

Nag Chowdhury, S., Cheng, W., de Melo, G., Razniewski, S., & Weikum, G. (2020). Illustrate your story: Enriching text with images. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 849–852. https://doi.org/10.1145/3336191.3371866

Nikolajeva, M., & Scott, C. (2013). *How picturebooks work*. Routledge.

O'Neil, K. E. (2011). Reading pictures: Developing visual literacy for greater comprehension. *The Reading Teacher*, *65*(3), 214–223. https://doi.org/https://doi.org/10.1002/TRTR.01026

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, *abs/1802.05365*. http://arxiv.org/abs/1802.05365

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Radford, A., Sutskever, I., Kim, J. W., Krueger, G., & Agarwal, S. (2021). Clip: Connecting text and images [Accessed: 2021-05-23].

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://openai.com/blog/better-language-models/

Ramesh, A., Pavlov, M., Gabriel, & Gray, G. S. (2021). Dall·e: Creating images from text [Accessed: 2021-05-20].

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation.

Rashkin, H., Celikyilmaz, A., Choi, Y., & Gao, J. (2020). Plotmachines: Outline-conditioned generation with dynamic plot state tracking.

Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. *ICML*, *48*, 1060–1069. http://dblp.uni-trier.de/db/conf/icml/icml2016.html#ReedAYLSL16

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *International conference on machine learning*, 1278–1286.

Riedl, M., & Young, R. (2006). From linear story generation to branching story graphs. *IEEE Computer Graphics and Applications*, *26*(3), 23–31. https://doi.org/10.1109/MCG.2006.56

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. https://doi.org/10.1002/9780470689646.ch1

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (tech. rep.). California Univ San Diego La Jolla Inst for Cognitive Science.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *CoRR*, *abs/1606.03498*. http://arxiv.org/abs/1606.03498

Sapkale, A., & Lukin, S. M. (2020). *Maintaining consistency and relevancy in multi-image visual storytelling* (tech. rep.). DEVCOM Army Research Laboratory Playa Vista.

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C. D. (2019). Do massively pretrained language models make better storytellers? *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. https://doi.org/10.18653/v1/K19-1079

Shareha, A. A., Rajeswari, M., & Ramachandram, D. (2009). Multimodal integration (image and text) using ontology alignment. *American Journal of Applied Sciences*, *6*(6), 1217.

Shin, A., Ushiku, Y., & Harada, T. (2018). Customized image narrative generation via interactive visual question generation and answering.

A soliloquy to the past [latent2visions] [Accessed: 2021-05-16]. (2021).

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2017). Cold fusion: Training seq2seq models together with language models.

Takacs, Z. K., & Bus, A. (2018). How pictures in picture storybooks support young children's story comprehension: An eye-tracking experiment. *Journal of experimental child psychology*, *174*, 1–12.

Topal, M. O., Bas, A., & van Heerden, I. (2021). Exploring transformers in natural language generation: Gpt, bert, and xlnet.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, L., Chen, W., Yang, W., Bi, F., & Yu, F. R. (2020). A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, *8*, 63514–63537. https://doi.org/10.1109/ACCESS.2020.2982224

Wang, S., Durrett, G., & Erk, K. (2020). Narrative interpolation for generating and understanding stories.

Wolfenbarger, C. D., & Sipe, L. R. (2007). A unique visual and literary art form: Recent research on picturebooks. *GSE Publications*. https://repository.upenn.edu/gse_pubs/32%22

Yao, L., Peng, N., Weischedel, R. M., Knight, K., Zhao, D., & Yan, R. (2018). Plan-and-write: Towards better automatic storytelling. *CoRR*, *abs/1811.05701*. http://arxiv.org/abs/1811.05701

Yao, L., Zhang, Y., Feng, Y., Zhao, D., & Yan, R. (2017). Towards implicit content-introducing for generative short-text conversation systems. *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2190–2199.

Zakraoui, J., Moutaz, S., Al-ma'adeed, S., Aljaam, J., & El-Seoud, S. (2021). Visualizing children stories with generated image sequences. https://doi.org/10.1007/978-3-030-67209-6_55

Zakraoui, J., Moutaz, S., Asghar, U., Aljaam, J., & Al-ma'adeed, S. (2020). Generating images from arabic story-text using scene graph, 469–475. https://doi.org/10.1109/ICIoT48696.2020.9089495