A Preliminary Look at Generative AI for the Creation of Abstract Verbal-to-visual Analogies

Kara Combs
Air Force Research Laboratory
Kara.Combs.1@us.af.mil

Trevor J. Bihl
Air Force Research Laboratory
Trevor.Bihl.2@us.af.mil

Abstract

Generative artificial intelligence (GAI) appears useful in the creation of new data, which assists in the expansion of small, limited datasets in fields such as analogical reasoning (AR). This multidisciplinary study expands the number of AR visual datasets within the field of visual question answering. We introduce the first visual analogy dataset that includes abstract concepts by leveraging three text-to-image GAI generators, Text2Img, Craiyon, and Midjourney, to produce images for antonym and synonym analogies. Our visual dataset achieves up to 70% accuracy and performs better 84.6% of the time compared to the same evaluation on only textual information. Interestingly, results also imply that paid GAI services produce higher accuracy. This work shows the potential for GAI to aid in the development of abstract visual analogy datasets, which allows for a better understanding and incorporation of AR into cognitive-inspired AI models capable of analogy-based information fusion.

1. Introduction

Most artificial intelligence/machine learning (AI/ML) applications are brittle in the sense they can complete limited tasks and lack robustness for learning tasks outside of their initial training. From an AI user perspective, the goal for “strong” or “general” AI is to gain the ability to learn beyond initial capacities, which has continued to be a difficulty for AI/ML systems to overcome (IBM, 2023; Ray, 2019). However, within the past few years, there has been exceptional progress made in more general AI for natural language processing (NLP) (e.g., the generative pre-trained transformer (GPT) family of algorithms (OpenAI, 2023) (OpenAI, 2023), Google PaLM (Chowdhery, et al., 2022)); computer vision (e.g., the DALL-E family (OpenAI, 2022; Dayma, et al., 2022)); speech-to-text (e.g., Whisper (Radford, et al., 2022)); code generation (e.g., Codex (Zaremba & Brockman, 2021)); and many more fields. Recently, generative AI has been in the spotlight for its performance on par with or surpassing human performance for select tasks (Gozalo-Brizuela & Garrido-Merchan, 2023; Stokel-Walker & Van Noorden, 2023). However, these models are very large, with millions of parameters, see Table 1 compiled from (Griffith, 2023; Alston, 2023; Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022; Dayma, et al., 2022). The size of these models necessitates large amounts of data and storage. Algorithm learning and training results in typically long computation times depending upon the number and type of tasks. Given these hefty requirements, many avenues are being explored that look to balance an algorithm’s ability to learn and perform well as a function of computational needs. (Jovanovic & Campbell, 2022)

Table 1. Generative AI Models’ Number of Parameters

Model Name	Number of Parameters	Times Larger than GPT-1
GPT-1	117M	1x
GPT-2	1.5B	13x
GPT-3	175B	1496x
GPT-3.5	1.3B	44x
	6B	51x
	20B*	171x*
	175B	1496x
GPT-4	(est.) 1T	8547x
DALL-E	12B	103x
DALL-E 2	3.5B	30x
Craiyon	(est.) 400M	3x

*GPT-3.5 optimized for ChatGPT

In general, one proposal to assist in how machines learn is through analogical reasoning, which utilizes knowledge already known about a base domain and relates it to inferred ideas about a target domain, which is usually unfamiliar to the subject (Combs, Lu, & Bihl, 2023; Mitchell, 2021; Gentner & Smith, 2013; Gentner & Maravila, 2017). Initial research within analogical reasoning looked at human performance on solving verbal analogies such as “Man is to king as woman is to ____?” where “queen” completes the problem. Analogies are often stylized in the A:B::C:D form, Man:King::Woman:Queen, (Spearman, 1923;

URI: https://hdl.handle.net/10125/106520
978-0-9981331-7-1

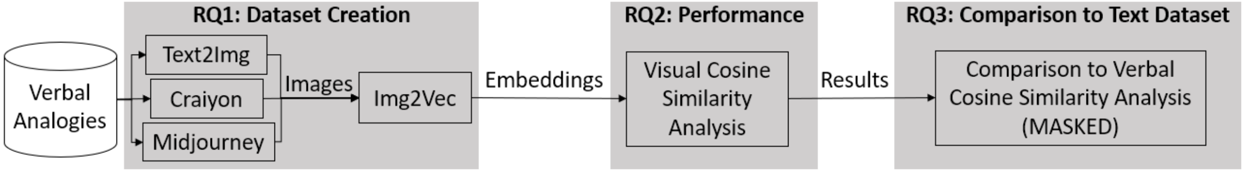

RQ1: Dataset Creation
RQ2: Performance
RQ3: Comparison to Text Dataset
Verbal Analogies
Text2Img
Craiyon
Midjourney
Images
Img2Vec
Embeddings
Visual Cosine Similarity Analysis
Results
Comparison to Verbal Cosine Similarity Analysis (MASKED)

Figure 1. Analysis Flowchart

Sternberg & Nigro, 1980). Many of these analogies were constructed around abstract ideas, which had definite semantic meanings but were ambiguous in terms of how they might be visualized. For example, take the analogy Up:Down::Rich:Poor (from (Sternberg & Nigro, 1980)). This has limited analogical reasoning research in the visual realm to concrete examples such as Cheese:Cow::Fried Egg:Chicken (from (Krawczyk, et al., 2008)). The lack of visual analogy datasets for abstract concepts is the primary motivation for this paper. Given generative AI's ability to take a text prompt and create a high-quality and accurate visual version, and widespread availability, we propose to leverage generative AI in the creation of visual analogies.

This paper discusses the following research questions:

- RQ1: How can generative AI be leveraged to create semantically abstract visual analogy datasets?
- RQ2: How do abstract visual analogy datasets created by different generative AI algorithms compare to one another as measured by accuracy on analogy problems?
- RQ3: How does accuracy performance on visual analogy datasets compare to previous performance on their verbal (textual) dataset counterparts?

The process implemented and its correspondence to the research questions are shown in Figure 1. We address these research questions throughout the paper. In Section 2, we present background on analogical reasoning, visual question answering (VQA), and generative AI. Section 3 details the creation of the visual datasets and Section 4 details the evaluation method and metrics thereof. The results are presented along with a discussion on the datasets, analysis, and future work in Section 5. Finally, the paper provides conclusions in Section 6. Our contributions to the literature are the following:

1. Method for the conversion of verbal analogies into visual analogies;
2. Development of 3 new visual analogy datasets considering abstract concepts;
3. Performance comparison of the newly created generative-AI-produced visual analogy datasets to one another and their verbal-only counterpart. 2. Background

This research is grounded in three research fields: analogical reasoning, visual question answering, and generative AI. This multidisciplinary study looks to expand the current state of analogical reasoning visual datasets through text-to-image generative AI. The idea of proposing a visual question prompt (such as a visual analogy) to yield an answer is the focus of visual question answering (VQA) problems.

2.1. Analogical Reasoning

There is a wide variety of analogy problems from natural language processing and cognitive science beyond those in word-based form (i.e. *A:B::C:*D), including sentence-based (also known as similes or metaphors) and story-based (i.e., parables) analogies (Combs, Lu, & Bihl, 2023; Ichien, Lu, & Holyoak, 2020). As mentioned earlier, there has been a significant amount of research regarding human performance (Goswami, 1991; Vendetti, Matlen, Richland, & Bunge, 2015; Christie, Gao, & Ma, 2020; Morsanyi, Stamenkovic, & Holyoak, 2020; Guerin, Wade, & Mano, 2021) and algorithm/machine development & performance (Combs, Bihl, Ganapathy, & Staples, 2022; Lu, Wu, & Holyoak, 2019; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014) on verbal *A:B::C:D* analogies. This has specifically sparked interest in natural language processing with the rise of vector space models (e.g., word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), FastText (Joulin, Grave, Bojanowski, & Mikolov, 2017)) and transformer models (e.g. Bidirectional Encoder Representations from Transformers (BERT) family (Devlin, Chang, Lee, & Toutanova, 2019; Liu, et al., 2019; Sanh, Debut, Chaumond, & Wolf, 2019). Though NLP algorithms possess a broad range of abilities, which makes them more versatile for general text evaluation, there is a plethora of verbal analogy datasets from both cognitive science and the natural language processing domains (Ichien, Lu, & Holyoak, 2020). For a verbal analogy, the respondent is presented with a text-based prompt; whereas, in visual analogy, the prompt is image-based, which leads us to VQA.

2.2. Visual Question Answering (VQA)

In VQA, the questions are posed as images and can cover target ID, classification, spatial relations, and many more (Patadia, Kejriwal, Shah, & Katre, 2021). The literature has several reviews that cover the most popular VQA methods and datasets (Zou & Xie, 2020; Patadia, Kejriwal, Shah, & Katre, 2021; Wu, et al., 2017). An example of a VQA problem would be the presentation of Figure 2 along with the external question: "How many leftover donuts are in the box on the red bicycle?" (Ren, Kiros, & Zemel, 2015).

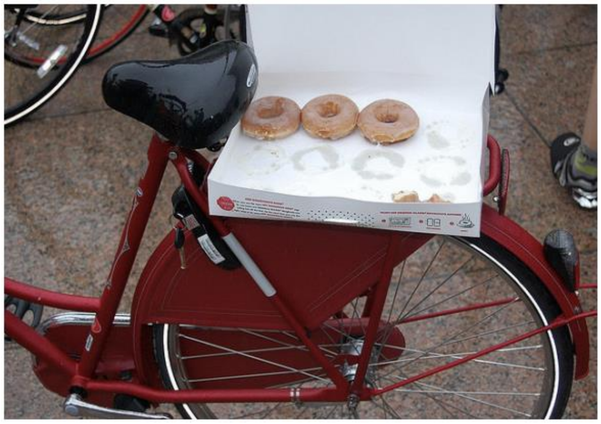


Figure 2. COCO-VA Image 5078

2.3. VQA for Analogical Reasoning

As mentioned earlier, analogies come in many different forms such as word-, sentence-, and story-based; however, our focus is on word-based (*A:B::C:D*) analogies due to being the simplest form to represent visually. Much of the work done in visual analogies concerns abstract geometrical problems such as Bonyard Problems (Bongard, 1967; Yun, Bohn, & Ling, 2020), Procedurally Generated Matrices (PGM) (Barrett, Hill, Santoro, Morcos, & Lillicrap, 2018), and more recently, Raven's Progressive Matrices (RPM) (Raven & Court, 1938; Zhang, Gao, Baoxiong, Zhu, & Song-Chun, 2019). It is worth noting that multi-modal research involving combining textual and visual analogies has been performed by (Ota, Shirai, Miyao, & Maruyama, 2022; Lu, Liu, Ichien, & Holyoak, 2019) and different prospectives for 3D objects have also been explored by (Reed, Zhang, Yuting, & Lee, 2015). However, it is beyond the scope of this paper.

The first analogy reasoning algorithm created, ANALOGY, was constructed to solve abstract geometrical problems (an example of which is shown in Figure 3) (Evans, 1964). Presented with the images labeled *A*, *B*, and *C* and options *1-5* in Figure 3, respondents are tasked with selecting the best option to fill the *D* slot. Similar to verbal *A:B::C:D* analogies, discovering the relationship between *A* and *B* and utilizing that to select *D* given *C* is key to solving these problems. However, these have been expanded into 3x3 matrices as seen with RPM problems as shown in Figure 4 (Zhang, Gao, Baoxiong, Zhu, & Song-Chun, 2019). When presented with the 3x3 prompt on the left side of Figure 4, respondents are tasked with selecting from options 1-8 on the right side. Despite being useful, these problems are far from real-world images (such as Figure 2) that computer vision algorithms are expected to accurately analyze, evaluate, and select the more appropriate response.

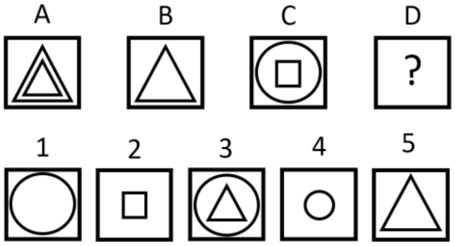


Figure 3. Example ANALOGY Problem (originally from MASKED)

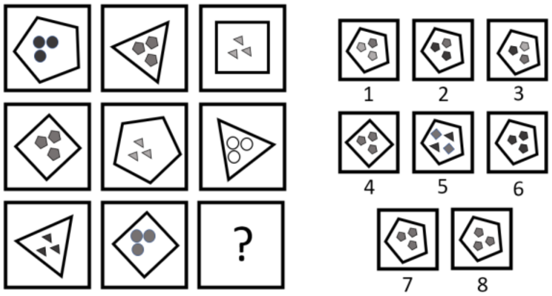
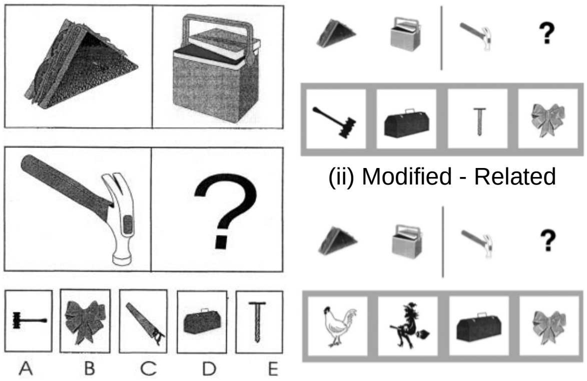


Figure 4. Example RPM Problem (originally from MASKED)

Beyond abstract geometrical, other forms of visual *A:B::C:D* analogies have also been investigated; however, they are tricker than their verbal counterpart due to the difficulty in representing abstract concepts. The literature shows that individuals likely interpret verbal analogies differently based on their own experiences with the subject matter and multiple definitions (Khatena, 1972). For example, consider the homonym, "bat." Immediately upon reading "bat" an image of the animal or a baseball bat likely popped into your head, or perhaps you imagined the verb version of the bat. Similarly, consider synonyms such as "father" and "dad." There might be a slight distinction semantically, but the mental image is likely very similar if not the same. There has been very limited research involving a mix of visual and textual aspects of an analogy. Multiple definitions and images are why analogies can be difficult to evaluate and construct in a visual sense. Thus, only two visual analogy datasets have been created: the Goranson Analogy Test (GAT) (Goranson, 2001; Krawczyk, et al., 2008) and the Visual Analogy Question Answer (VAQA) dataset (Sadeghi, Zitnick, & Farhadi, 2015).

2.3.1. GAT Dataset. The original GAT dataset consisted of 24 word-based analogies in Version A and 24 clip-art analogies in Version B (see Figure 5) (Goranson, 2001). Each analogy presented *A, B,* and *C* with five options for *D.* Later, the original GAT Version B was reduced to only include 4 options for 18 analogies (2 in the practice set and 16 in the test set) (Krawczyk, et al., 2008). Additionally, this dataset was split based on options that were purposely picked as distractors to the correct answer (directly derived from the original GAT Version B analogies) (see Figure 5(ii)) and unrelated options (derived from other unrelated GAT analogy choices) (see Figure 5(iii)) The GAT Version B with distractor options (those shown in Figure 5(ii)) simply referred to as “GAT” from this point forward, has been used in two studies. One was a typical cognitive science study interested in human performance and response rates when answering the analogies (Wong, Schauer, Gordon, & Holyoak, 2019). In the computer vision realm, the GAT dataset and its corresponding textual labels were used as inputs to ResNet50-A (ResNet50 adapted to make analogy predictions) and the Bayesian Analogy with Relational Transformations, respectively, in one study interested in a computer’s ability to solve analogies (Lu, Wu, & Holyoak, 2019). The limitations of this dataset are its size (the number of analogies) and the concreteness of the images pictured within the analogies. None are abstract ideas, which is a foundational pillar of analogical reasoning research.



(i) Original (iii) Modified - Unrelated

Figure 5. GAT Version B Analogy Problems

2.3.2. VAQA Dataset. The VAQA dataset consisted of real-world images of animals and vehicles (motorcycles, cars, bicycles, etc.) (Sadeghi, Zitnick, & Farhadi, 2015). The two primary relations were attribute (which was color) and action (swimming, sitting, standing, eating, lying, etc.). The latter of which was reserved for only animal-to-animal analogy. Given images for *A, B,* and *C,* selections for *D* are retrieved by the algorithm to best complete the analogy. An example of an attribute analogy is shown in Figure 6(i) where the white motorcycle is to the yellow motorcycle as a white bird is to a yellow bird. In Figure 6(ii), an example action analogy is shown where a dog swimming is to a dog standing as horses swimming are to horses standing. The primary limitation of the VAQA dataset is that relationships captured by the analogies are semantically obvious and uninteresting (the “relationship” between a white motorcycle and a yellow motorcycle is that they are both motorcycles in the case of Figure 6(i)).

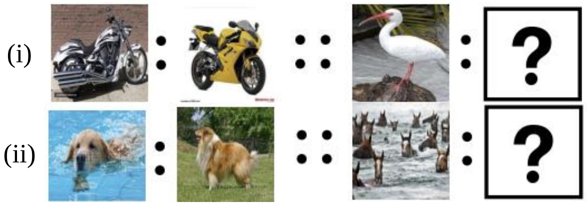


Figure 6. VAQA Attribute and Action Analogy Examples (Sadeghi, Zitnick, & Farhadi, 2015)

2.4. Generative AI

Recently, text-to-image generative AI has taken the world by storm with the rise of Text2Img, DALL-E 2, and generative adversarial networks (GANs) (Cao, et al., 2023; Zhang, Zhang, Zhang, & Kweon, 2023). GANs rose to prominence due to their ability to create high-quality synthetic data (such as images) based on real data (Karras, Laine, & Aila, 2019; Karras, et al., 2020). This is especially beneficial to the expansion of current datasets and the curation of new datasets to provide more training and evaluation material for AI/ML algorithms (Karras, Laine, & Aila, 2019; Karras, et al., 2020; Cao, et al., 2023). However, one of the downfalls is that GANs require ground truth data in the same medium as its outputs such as images. Other forms of generative AI have allowed for the creation of chat-box-like querying as showcased by ChatGPT’s ability to create reasonably sensible content (textual or visual) given a prompt. Additionally, this allows models to be multimodal in the sense the prompt and the generated result are not necessarily of the same medium (Cao, et al., 2023). Some of the popular examples of this include DALL-E 2, Midjourney, and Craiyon (formerly DALL-E mini) where an image is generated from a textual description, but the text can also be used to generate auditory speech, music, knowledge graphs, computer programming code, and more (Cao, et al., 2023; Zhang, Zhang, Zhang, & Kweon, 2023). Since generative AI has had a promising history of image generation and requires little human input (as compared to searching for appropriate public images), it was leveraged to create the datasets in this work.

3. Dataset Curation

Given the limited number of visual analogy datasets and their clear limitations (such as size, inherent relationships captured, etc.) and the difficulties in their creation, one potential solution to this was to query generative AI with prompts from a verbal analogy dataset. For this study, a shortened version of the Sternberg & Nigro data was utilized as the prompts for the generative AI engines. The generative AI algorithms selected were Text2Img (utilizes the Stable Diffusion model), Craiyon (formerly DALL-E Mini), and Midjourney based on their popularity, ease of use, and cost (Borji, 2022).

3.1. Sternberg & Nigro Verbal Analogy Dataset

There are many verbal analogy datasets; however, Sternberg & Nigro dataset was selected due to its simplicity, size, and popularity among studies (see (Morrison, et al., 2004; Combs, Bihl, Ganapathy, & Staples, 2022). The original Sternberg & Nigro dataset was developed for the evaluation of the abstraction abilities of students and consisted of 197 analogies with four options for *D* to complete the analogy with the given *A*, *B*, and *C* words (Sternberg & Nigro, 1980). The 197 analogies span five relationships: antonyms (Yes:No::True:False), synonyms (Early:Almost::Car: Auto), categorical (John: Name::Dinner:Supper), functional (Bird:Fly::Rabbit: Hop), and linear ordering (Month:Year::Inch:Foot). In a modified version, that is utilized in this study, the number of options was reduced to two options, the correct answer (*D*) and one of the incorrect options (*D'* [D prime]) (Morrison, et al., 2004). For our preliminary study, only 40 analogies were considered with 20 being synonyms and 20 being antonyms. This was due to the time needed to generate the images. Individual words were used as the prompts for the generative AI engines.

3.2. Generative AI Datasets

Information related to the cost, interface, and daily limits (if any) of the generators are shown in Table 2. Image quality was independent of tiers except for watermarks and logos in some instances. The free versions of Text2Img (backend uses Stable Diffusion v4) and Craiyon and the most basic tier of Midjourney. To query the AI image generators, the individual words were passed without any additional content or formatting. This process was repeated for each *A, B, C, D*, and *D'* word in the selected analogies with some variation for the compilation of the datasets.

Table 2. Generative AI Algorithm Comparison

	Text2Img	Craiyon	Midjourney
Free?	Yes	Yes	No
Paid?	No	\$60-\$240/yr	\$96-576/yr
Interface	Chat-like Website	Queried Website	Discord Server
Limits	~450/mo.	Unlimited	200/mo.

The Text2Img generator creates one image for a given prompt, which was used for analysis. Default parameters were used: width = 512 px, height = 512 px, number of denoising steps = 20, and guidance scale = 7.5. In total, the Text2Img dataset had 5 images per analogy (one that corresponds to *A*, *B*, *C, D,* and *D'*) for a total of 200 images. Craiyon produces nine images given one prompt, all of which were used in the analysis (via an averaging method described in the next section). Craiyon did not have any customizable settings available at the time. The Craiyon dataset had 45 images per analogy (9 images for each *A*, *B*, *C*, *D*, and *D'* word) for a total of 1,800 images. Midjourney produces four images for a given prompt; however, at the discretion of one author (*n*=1), the most representative image was selected. The default parameters for Midjourney were also used and the Midjourney dataset had 5 images per analogy for a total of 200 images.

A small example of images in the Text2Img (i), Craiyon (ii), and Midjourney (iii) datasets are shown in Figure 7-Figure 9 for car, calm, and deep, respectively. These are provided to show the different interpretations of the concepts by each generator. They do not form an analogy with one another. Despite some variation in the style, almost everyone could identify each picture in Figure 7 as a car. The generators vary slightly more in Figure 8; however, there are. In which the term calm was used, an abstract concept that cannot be tangibly represented, all three drew on underlying themes of nature and blue hues. The generators divert more in Figure 9 for another abstract concept, "deep," however, there is still an underlying water theme to all the images, perhaps due to the common phrase, "deep blue sea."

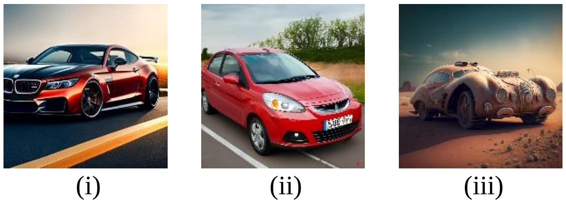
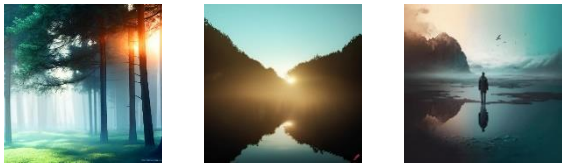
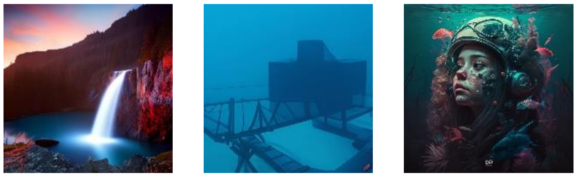


Figure 7. Car Images (Analogy 22)



(i) (ii) (iii)

Figure 8. Calm Images (Analogy 17)



(i) (ii) (iii)

Figure 9. Deep Images (Analogy 18)

4. Methods

The methods to create and evaluate the visual analogy datasets are shown in Figure 1. To process images, they must first be converted into feature vectors. Several methods to extract embeddings from images exist (that is a vector representation); however, img2vec was selected due to its implementation simplicity and variety of compatible models (Safka, 2017). Img2vec is a deterministic model that extracts embeddings from several popular convolutional neural networks (CNNs): ResNet-18, AlexNet, VGG-11, DenseNet, and EfficientNet (B0-B7). Img2vec uses a specific layer of the selected CNN to create the embedding shown in Table 3. The embedding is simply a vector representation of the image with the corresponding dimension shown in the output size. Since each CNN creates different embedding for each image, the analysis for each dataset is repeated for each of the CNNs.

Table 3. CNN Embedding Information

CNN	Layer	Output Size
ResNet-18	Avgpool	512
AlexNet	2	4096
VGG-11	2	4096
DenseNet	1 from features	1024
Efficient Net	1	1024

Once feature vectors are extracted from the images, they are ready for the analogy comparison analysis. Since there is one image for word for Text2Img and Midjourney, that image's feature embedding is what's used to represent the given word. In the case of Craiyon which includes nine images (N=9) for a given word, the average across all nine feature vectors (I_n) is used to represent the image as shown:

$$I_{avg} = \sum_{n=1}^{N} I_n \big/ N. \tag{1}$$

For the comparison between the correct analogy, *A*:*B*::*C*:*D*, and the incorrect one, *A*:*B*::*C*:*D'*, the cosine similarity was used to compare two vector embeddings as previously described in (MASKED). The cosine similarity, denoted *sim*, compares two vectors, v_1, and v_2 by taking their dot product divided by the vector's magnitude as shown in:

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|}. \tag{2}$$

Cosine similarity determines how alike two images are, on a scale of 0 (very dissimilar) to 1 (very similar). To determine whether the correct answer or distractor would be selected, the cosine similarity comparing *A*:*B*, *C*:*D*, and *C*:*D'* are calculated with the feature embeddings for *A*, *B*, *C*, *D*, and *D'*. Next, to determine if *D* or *D'* is selected, the absolute value of the difference between the cosine similarity (called "*SimDiff*") of *A*:*B* and *C*:*D* is calculated along with the absolute value of the difference between the cosine similarity of *A*:*B* and *C*:*D'* shown in

$$SimDiff_{ABCx} = |sim(A, B) - sim(C, x)| \; for \; x \in \{D, D'\}\,. \tag{3}$$

The similarity difference captures how closely the similarity between two vectors (like the relationship between the two) matches the similarity of two other vectors. The similarity difference ranges from 0 (the two sets of vectors have the same relationship) to 1 (the two sets of vectors have the opposite relationship). An ideal analogy has a similarity difference of 0 meaning that the relationship between *A* and *B* is reflected exactly in the relationship between the other set of vectors. Hence, the algorithm will select the option that yields the value closest to 0 when their similarity differences $SimDiff_{ABCD}$ (considers *D*) and $SimDiff_{ABCD'}$ (considers *D'*), respectively, are calculated.

4.1. Performance Metrics

The metric of interest is accuracy which measures how often the correct answer, *D*, was selected compared to the distractor, *D*', divided by the total number of analogies, which is 40 in this case. This allows for the direct comparison to the verbal-only analysis completed in (MASKED) utilizing Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014) as the selection algorithm. Word2Vec and GloVe are standard baselines for the evaluation of natural language processing applications, which is why they were selected for

Table 4. Generator-CNN Performance Broken Down by Antonyms, Synonyms, and Overall Accuracy (%)

Generator	Relation	RN	AN	VGG	DN	EfficientNet								Avg
						B0	B1	B2	B3	B4	B5	B6	B7	
Text2Img	Antonym	45	50	60	40	50	30	45	30	40	40	25	45	42
	Synonym	35	45	60	55	55	45	30	35	50	50	50	40	46
	Overall	40	47.5	60	47.5	52.5	37.5	37.5	32.5	45	45	37.5	42.5	43.8
Craiyon	Antonym	30	45	50	50	40	60	45	60	30	45	40	60	46
	Synonym	20	25	45	30	30	30	30	40	45	40	25	30	33
	Overall	25	35	47.5	40	35	45	37.5	50	37.5	42.5	32.5	45	39.4
Midjourney	Antonym	50	70	70	60	40	40	55	55	55	55	45	60	55
	Synonym	45	45	55	45	50	50	45	70	35	55	60	55	51
	Overall	47.5	57.5	62.5	52.5	45	45	50	62.5	45	55	52.5	57.5	52.7

comparison here. The same cosine similarity analysis was conducted in (MASKED) on the verbal analogies for two popular vector space models, Word2Vec and GloVe (presented in Table 5). Across the board, GloVe performed better than Word2Vec, but both struggled with synonyms more than antonyms.

Table 5. Algorithm Performance on Verbal Analogies

NLP Algorithm	Correctness Percentage Results		
	Overall	Antonyms	Synonyms
Word2Vec	36.8%	45%	27.8%
GloVe	65%	70%	60%

5. Results & Discussion

The accuracy results are shown in Table 4 for each (AI-)generator-CNN combination with the average across all CNNs for a given dataset and the relation displayed in the last column. Key to RQ3 was how these visual analogies compared to when the same analysis was done on their verbal-only counterparts using natural language processing. This analysis was previously done by (MASKED) with key results presented in Table 5. For the same forty analogies, all but three generator-CNN combinations scored above word2vec's overall accuracy score (denoted in Table 4 by light gray); however, none of them beat GloVe's score, with the closest still being 2.5% short at 62.5%. An overwhelming majority of the generator-CNN combinations yielded an accuracy above Word2Vec for antonyms (17/36) (light gray in Table 4) and synonyms (31/36) (dark gray in Table 4). There were also three instances where GloVe's performance was surpassed, which exclusively occurred with the Midjourney dataset. First, a discussion on the performance metrics of the visual analogies compared to the analogies completed on their verbal-only counterparts is presented in Section 5.1. This is followed by a discussion of future work in Section 5.2.

5.1. Accuracy Performance

Looking at Table 4, Midjourney is the best-performing dataset (with an overall average of 52.7%) followed by Text2Img (43.8%) and Craiyon (39.4%). This suggests that paid models may have a slight advantage over their free-tier counterparts concerning RQ2. Since the results in Table 4 are overwhelmingly (84.6%) shaded gray, this suggests that visual analogies can perform better than their verbal-only counterparts. Importantly, this is consistent across Generative AI methods and image classifiers (the CNNs), making this text-to-image analysis agnostic to both the image generator and CNN This result further shows that our image generation method (described in Section 3.2 as the proposed solution to RQ1) results in higher accuracy in evaluating analogies compared to a text-only analysis with NLP algorithms.

Importantly, there are various reasons why the created visual analogies perform similarly to the word-only pairs. Case 1, the ideal reason for this performance, is because the AI-generated images are truly representative of the analogy words and capture the semantic relationships between those words. However, this is hard to quantify in the current analysis without conducting a quality evaluation by a human. Another potential Case 2, that is less ideal, is that the generated images do not represent the analogy words well; however, the images are similar enough to produce the same results when compared. Another question is whether the cost of the AI generator impacts the performance (see Table 2) due to creating more representative images (Case 1) or whether the images are simply more similar in structure/style/etc. (Case 2). Finally, a third question is whether dataset performance represents the quality of the dataset as a whole, which is also dependent on the two cases presented earlier. Each of the generators has its frequently recurring unrelated themes such as cars and robots for Text2Img, portraits for Midjourney, and vehicles (airplanes, trains, etc.) for Craiyon.

5.2. Future Work

There are several potential directions for future work regarding different aspects of this project. One obvious direction is to continue this work through visual expansion of the full Sternberg & Nigro dataset with the generators utilized herein or additional ones. Another avenue to explore is the human factors quality evaluation and selection of the best representative AI-produced image(s), which is currently being conducted by the authors. Once a human baseline is established, this could be automated to guide an end-to-end pipeline for prompt engineering (see (Branwen, 2020; Liu & Chilton, 2022; Oppenlaender, 2022)) for curating future abstract analogy datasets. Prompt engineering would assist in the driving of a particular definition or provide more clarity for a given concept. This would prove the quality of visual analogy datasets by ensuring the concepts are represented accordingly and speeding up the process by reducing the need for human approval of image quality.

6. Conclusions

The rise of generative artificial intelligence (AI) has allowed for its widespread use and applications in various fields such as natural language processing, computer vision, and many more (Cao, et al., 2023; Zhang, Zhang, Zhang, & Kweon, 2023). Specifically, its ability to generate an image from a textual prompt allows for the expansion of small visual datasets and/or the creation of new visual datasets. One domain that suffers from limited datasets and a limited availability of datasets is abstract visual analogical reasoning wherein words are represented visually in an *A:B::C:D*-like format. However, a key problem is that analogies are often abstract, and ambiguous in how to visually represent abstract concepts so that it is universally comprehendible (e.g., “yes,” “deep,” etc.).

This paper utilized generative AI to create three abstract visual *A:B::C:{D,D’}* datasets based on the verbal analogy dataset created by (Sternberg & Nigro, 1980) and later, modified and used by (Morrison, et al., 2004). The datasets we contributed represent 40 analogies (20 antonyms and 20 synonyms) visually based on results from Text2Img, Craiyon, and Midjourney. These datasets had their embeddings extracted by twelve convolutional neural networks (CNNs) such that the cosine similarity could be computed to determine if the algorithms selected the correct answer, *D*, over the incorrect distractor, *D’*. This analysis was repeated for each generator-CNN combination for a total of 36 evaluations. Upon a similar comparison of the analogies' verbal-only representation via natural language processing algorithms, Word2Vec and GloVe (MASKED), a significant majority of the image datasets (84.6%) scored higher than Word2Vec’s performance on antonyms, synonym, and when both antonyms and synonyms are considered (called “overall”). Of this, three evaluations had 70% accuracy which was above the performance of the higher-scoring GloVe algorithm for both antonyms and synonyms.

This research provides the third-ever semantically abstract visual analogy dataset, through the creation of a method that transfers textual analogies into visual ones by leveraging generative AI. We believe this could be a source for additional datasets but must be referenced as synthetic or AI-generated. Our next contribution to the literature is the analysis of these visual analogies, which has only been completed in two previous studies on a different dataset. Finally, we compared these results to the analysis of their verbal-only counterpart done previously, which has only occurred in one research paper. This analysis showed that AI-generated images for abstract analogies provide added knowledge compared to the evaluation of their verbal-only counterparts. This shows the promising nature of generative AI in the expansion, creation, and development of visual analogy datasets. This leads to more potential for multi-modal research on analogies, which is a direct application of information fusion.

7. Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily represent any views of the U.S. Government, U.S. Department of Defense, or U.S. Air Force. This work was cleared for unlimited release under AFRL-2023-2746.

8. References

Alston, E. (2023, March 21). Text2Img vs. GPT-3 and GPT-4: What's the difference? *Zapier*.

Barrett, D. G., Hill, F., Santoro, A., Morcos, A. S., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. *International Conference on Machine Learning.* Stockholm.

Bongard, M. M. (1967). *Pattern Recognition.* Moscow: Nauka Press.

Borji, A. (2022). Generated faces in the wild: Quantitative comparison of Stable Diffusion, Midjourney and DALL-E 2. *arXiv preprint arXiv:2210.00586*, 1-9.

Branwen, G. (2020, June 19). *Prompts as Programming*. Retrieved from GPT-3 Creative Fiction: https://gwern.net/gpt-3#prompts-as-programming

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to Text2Img. *arXiv preprint arXiv:2303.04226*, 1-44.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., . . . Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 1-87.

Christie, S., Gao, Y., & Ma, Q. (2020). Development of analogical reasoning: A novel perspective from cross-cultural studies. *Child Development Perspective, 14*(3), 164-170.

Combs, K., Bihl, T. J., Ganapathy, S., & Staples, D. (2022). Analogical reasoning: An algorithm comparison for natural language processing. *Proceedings of the 55th Hawaii International Conference on System Sciences* (pp. 1310-1319). HICSS.

Combs, K., Lu, H., & Bihl, T. J. (2023). Transfer learning and analogical inference: A critical comparison of algorithms, methods, and applications. *Algorithms, 16*(3), 146.

Dayma, B., Patil, S., Cuenca, P., Saifullah, A. T., Le, P., Melas, L., & Ghosh, R. (2022, July 4). DALL-E Mini Explained. *Weights & Biases*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Minneapolis.

Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. *Proceedings of the April 21-23, 1964, spring joint computer conference.* New York City.

Gentner, D., & Maravila, F. (2017). Analogical Reasoning. In L. J. Ball, & V. A. Thompson, *International Handbook of Thinking & Reasoning* (pp. 186-203). London: Routledge.

Gentner, D., & Smith, L. A. (2013). Analogical Learning and Reasoning. In D. Reisberg, *Oxford Handbook of Cognitive Psychology* (pp. 668-681). Oxford: Oxford University Press.

Goranson, T. E. (2001). *On diagnosing Alzheimer's disease: Assessing abstract thinking and reasoning.* Dissertation, University of Victoria.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child development, 62*(1), 1-22.

Griffith, E. (2023, March 16). GPT-4 vs. ChatpGPT3.5: What's the difference? *PC Magazine*.

Guerin, J. M., Wade, S. L., & Mano, Q. R. (2021). Does reasoning training improve fluid reasoning and academic achievement for children and adolescents? A systematic review. *Trends in Neuroscience and Education, 23*, 1-15.

IBM. (2023, April 4). *What is artificial intelligence (AI)?* Retrieved from IBM Cloud: https://www.ibm.com/topics/artificial-intelligence

Ichien, N., Lu, H., & Holyoak, K. J. (2020). Verbal analogy problem sets: An inventory of testing materials. *Behavior research methods, 52*(5), 1803-1816.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics. 2,* pp. 427-431. Valencia: Association for Computational Linguistics.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410). Long Beach: IEEE/CVF.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of Stylegan. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110-8119). Seattle: IEEE/CVF.

Khatena, J. (1972). The use of analogy in the production of original verbal images. *Journal of Creative Behavior,* 209-213.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., . . . Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia, 46,* 2020-2032.

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-23). New Orleans: ACM.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 1-13.

Lu, H., Liu, Q., Ichien, N. Y., & Holyoak, K. J. (2019). Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1-7). Austin: Cognitive Science Society.

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, 116*(10), 4176-4181.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems,* 3111-3119.

Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences, 1505*(1), 79-101.

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience, 16*(2), 260-271.

Morsanyi, K., Stamenkovic, D., & Holyoak, K. J. (2020). Analogical reasoning in autism: A systematic review and meta-analysis. *Thinking, reasoning, and decision making in autism,* 59-87.

OpenAI. (2022). *DALL-E 2*. Retrieved from OpenAI: https://openai.com/product/dall-e-2

OpenAI. (2023). GPT-4 Technical Report. *arXiv:2303.08774,* 1-100.

OpenAI. (2023, February 2023). *Introducing Text2Img Plus*. Retrieved from OpenAI: https://openai.com/blog/Text2Img-plus

Oppenlaender, J. (2022). A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988*, 1-20.

Ota, K., Shirai, K., Miyao, H., & Maruyama, M. (2022). Multimodal analogy-based image retrieval by improving semantic embeddings. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 26*(6), 995-1003.

Patadia, D., Kejriwal, S., Shah, R., & Katre, N. (2021). Review of VQA: Datasets and approaches. *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)* (pp. 1-6). Greater Noida: IEEE.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543). Doha: Association for Computational Linguistics.

Radford, A., Wook, K. J., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 1-28.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1-27.

Raven, J. C., & Court, J. H. (1938). *Raven's progressive matrices.* Los Angeles: Western Psychological Services.

Ray, T. (2019, February 23). *Intel's neuro guru slams deep learning: 'It's not actually learning'*. Retrieved from ZDNet: https://www.zdnet.com/article/intels-neuro-guru-slams-deep-learning-its-not-actually-learning/

Reed, S. E., Zhang, Y., Yuting, Z., & Lee, H. (2015). Deep visual analogy-making. *Advances in neural information processing systems.* New York City.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring modelings and data for image question answering. *Advances in Neural Information Processing Systems 28* (pp. 1-9). Montreal: NeurIPS.

Sadeghi, F., Zitnick, C. L., & Farhadi, A. (2015). Visalogy: Answering visual analogy questions. *Advances in Neural Information Processing Systems.* Montreal.

Safka, C. (2017, November 3). Extract a feature vector for any image with PyTorch. *Becoming Human: Artificial Intelligence Magazine*. Retrieved from Becoming Human: https://becominghuman.ai/extract-a-feature-vector-for-any-image-with-pytorch-9717561d1d4c

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 1-5.

Spearman, C. (1923). *The Nature of Intelligence and the Principles of Cognition.* London: Macmillan.

Sternberg, R., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development, 19,* 27-38.

Vendetti, M. S., Matlen, B. J., Richland, L. E., & Bunge, S. A. (2015). Analogical reasoning in the classroom: Insights from cognitive science. *Mind, Brain, and Education, 9*(2), 100-106.

Wong, E. F., Schauer, G. F., Gordon, P. C., & Holyoak, K. J. (2019). Semantic and visual interference in solving pictorial analogies. *Proceedings of the 41st Annual Conference of the Cognitive Science Society.* Austin: Cognitive Science Society.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding, 163*, 21-40.

Yun, X., Bohn, T., & Ling, C. (2020). A deeper look at Bongard problems. *Canadian conference on artificial intelligence.* Virtual.

Zaremba, W., & Brockman, G. (2021, August 10). *OpenAI Codex*. Retrieved from OpenAI: https://openai.com/blog/openai-codex

Zhang, C., Gao, F., Baoxiong, J., Zhu, Y., & Song-Chun, Z. (2019). RAVEN: A dataset for relational and analogical visual reasoning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5317-5327). Long Beach: IEEE.

Zou, Y., & Xie, Q. (2020). A survey on VQA: Datasets and approaches. *2020 2nd International Conference on Information Technology and Computer Application (ITCA)* (pp. 289-297). Guangzhou: IEEE.