Understanding is Seeing: Metaphorical and Visual Reasoning in Multimodal Large Language Models

Sofia Lugli (sofialugli.personal@gmail.com)
Univeristy of Trento CIMeC, Corso Bettini 31
38068 Rovereto, Italy

Carlo Strapparava (strappa@fbk.eu)
Fondazione Bruno Kessler, Via Sommarive 18
38123 Trento, Italy

Abstract

Drawing from the Conceptual Metaphor Theory and the Structure-Mapping Theory, this paper introduces two exploratory works in the field of metaphorical and visual reasoning using vision models and multimodal large language models. (i) The Multimodal Chain-of-Thought Prompting for Metaphor Generation task aimed to generate metaphorical linguistic expressions from non-metaphorical images by using the multimodal LLaVA 1.5 model and the two-step approach of multimodal chain-of-thought prompting. The results showed the model's ability to generate metaphorical expressions, as 92% of them were classified as metaphors by human evaluators. Additionally, the evaluation revealed interesting patterns in terms of *metaphoricity*, *familiarity* and *appeal* scores across the generated metaphors. (ii) The Metaphorical Visual Analogy (MeVA) task consisted in solving visual analogies of the kind *source domain : target domain :: source element : ?* by choosing the correct target element among three difficult *distractors*, varying in semantic domains and roles. The results showed that all six models and humans performed higher than chance level, with only GPT-4o and ConvNeXt achieving higher than humans. Moreover, the error analysis showed that, in solving the analogies, the most frequent error was the selection of *distractor 1*. These works showed encouraging results for future research in the field of metaphorical and visual reasoning, contributing to the broader question of whether AI models serve as empirical tests of existing cognitive theories.

Keywords: Artificial Intelligence, Analogy; Creativity; Natural Language Processing

Introduction

If on one hand the recent advances in AI have shown impressive potential in modeling human language and cognition, on the other hand they have also raised critical questions about the role of cognitive theories in shaping future research, such as whether AI-driven models serve as empirical tests of existing theories, or whether they call for new theoretical frameworks. Drawing from the Conceptual Metaphor Theory (CMT) (Lakoff & Johnson, 1980b) and the Structure-Mapping Theory (SMT) (Gentner, 1983), this paper serves as an exploratory work in the field of metaphorical and visual reasoning using vision models and multimodal large language models. It aims to provide a different approach to metaphor processing by proposing two specific tasks: the Multimodal Chain-of-Thought Prompting for Metaphor Generation task and the Metaphorical Visual Analogy (MeVA) task. As metaphors are not only pervasive in language but also in everyday life, influencing our thoughts and actions (Lakoff & Johnson, 1980b), and as human meaning representations rely on multiple modalities (Barsalou, 2008), it became relevant to study metaphors in more than one modality, in particular in the vision domain. Recent research has indeed explored multimodal metaphors generation in a variety of ways (Akula et al., 2023; Hwang & Shwartz, 2023; Xu et al., 2022; Chakrabarty et al., 2022). Nevertheless, the common aspect across these studies is that the metaphorical quality was already present in the input employed. Therefore, the first experiment of this paper proposes an alternative approach that involves generating linguistic metaphors from non-metaphorical images, which lack inherent metaphorical qualities. The results show the model's ability to generate metaphorical expressions, with 92% of the generated expressions being classified as metaphors. Despite theoretical views on the connection between analogy and metaphor (Gentner, 1983; Gentner & Clement, 1988; Lakoff & Johnson, 1980b); and despite analogy making being a fundamental component of human intellect (Gentner et al., 2001) and becoming a widely used task to evaluate word embeddings (Mikolov, Chen, et al., 2013), there is limited research on solving metaphorical analogies (Czinczoll et al., 2022; Pitarch et al., 2023), and to the best of the authors' knowledge, a metaphorical analogy task in the visual domain was lacking. Therefore, the second experiment of this paper aims to propose a metaphorical visual analogy task, drawing on the Structure Mapping Theory (Gentner, 1983), which claims that metaphors consist of analogical mappings between source and target domains and their corresponding elements. The aim of this task is to adapt the existing metaphorical textual analogy task (Czinczoll et al., 2022) to the visual domain, incorporating visual-semantic arithmetic and situation recognition. The task consisted in completing the visual analogy of the kind: *source domain image : target domain image :: source element image : ?* by choosing the correct image from a set of candidates images [*target element image, distractor 1, distractor 2, distractor 3*]. The task was solved by five computer vision models in zero-shot arithmetic setting, by GPT-4o in zero-shot multimodal prompting, and by five workers on Amazon Mechanical Turk. For both experiments, the collected dataset and the results obtained are publicly available.[1] [2] By grounding this work in established cognitive theories, we investigate whether AI developments align with or chal-

[1] https://github.com/SofiaLugli/Multi_COT_meta_gen.git
[2] https://github.com/SofiaLugli/MeVA-.git

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

lenge existing frameworks in cognitive science, in particular whether these models exhibit metaphorical reasoning as theorized by CMT and SMT.

Background

Metaphor Theories

For most people, metaphor is merely a rhetorical device restricted to poetic language; however, according to the Conceptual Metaphor Theory (CMT) (Lakoff & Johnson, 1980b) metaphor is pervasive in everyday language, playing a significant role in communication and cognitive processes (Lakoff & Johnson, 1980b). More precisely, we talk about *conceptual metaphor* and *linguistic metaphor*. Conceptual metaphors consist of systematic sets of mappings across conceptual domains, whereby a target domain, which is usually a more abstract and complex concept, is partly structured in terms of a different source domain, which usually defines a more concrete and common concept (Lakoff & Johnson, 1980b). Conceptual metaphors are then reflected in our everyday language by a wide variety of linguistic metaphors. For instance, ARGUMENT IS WAR is a conceptual metaphor, where ARGUMENT is the target domain and WAR is the source domain; an example of its linguistic metaphor is e.g. *He attacked every weak point in my argument* (Lakoff & Johnson, 1980b). Some of these metaphorical mappings can be defined as *conventional* metaphors, as they are so deep-rooted in our everyday thought and language that they might have become the dominant way of framing a specific concept, and they represent the commonsense (Semino, 2008); while other metaphorical mappings, i.e. *novel* metaphors, are more creative, and they are not (yet) used in everyday discourse, but may become conventionalized if frequently used.

A different but yet complementary view of metaphor is the analogical view. Analogical reasoning is a fundamental mechanism of human intellect, and it consists in the ability to identify and use similarities based on relationships between entities, rather than just on the entities themselves, and it is considered the foundation of metaphor (Gentner, 1983; Lakoff & Johnson, 1980a; Fauconnier & Turner, 2003). According to the *Structure-Mapping Theory* (Gentner, 1983), the analogy maps knowledge from one domain (the base) into another (the target), implying that among both base objects and target objects holds the same system of relations. They propose that interpreting analogies consists in identifying a "common relational structure" independently from the objects in which those relations are contained (Gentner & Clement, 1988). The concept of *systematicity* guides the choice of which relations to match: individuals prefer to match and carry over systems of predicates governed by higher-order constraining relations, as opposed to mapping single predicates (Gentner & Clement, 1988). For instance, the Rutherford analogy between the solar system and the hydrogen atom is based on the common underlying relations between them (i.e., both are central-force systems that attract other entities) and not on their attributes (i.e., color, size). This model assumes that metaphor understanding entails processing of complex representational structures (Gentner & Clement, 1988).

Metaphor Processing

Over the past years, as humans excel at high-level semantic task (E. V. Shutova, 2011), and as statistical corpus analysis (Steen et al., 2010) indicates that in corpora, metaphors occur in approximately one-third of the sentences, metaphor gradually became an important topic also in NLP. Numerous studies have been conducted to investigate metaphors, resulting in three main sub-tasks: metaphor identification (E. V. Shutova, 2011; Tsvetkov et al., 2014; Gao et al., 2018; Mao et al., 2022), metaphor interpretation (E. Shutova, 2010; Su et al., 2017; E. Liu et al., 2022), and metaphor generation (Veale, 2016; Yu & Wan, 2019; Chakrabarty et al., 2021). As human meaning representations rely not only on linguistic exposure, but also on perceptual system and sensorymotor experience, (Barsalou, 2008; Louwerse, 2011); and as metaphors are not merely a matter of language but also of thought and action (Lakoff & Johnson, 1980b), it became relevant to study metaphors through different modalities. In NLP, the shift towards multimodality happened once computational approaches started adding sensory and contextual features which led to a better performance in metaphor processing (P. Turney et al., 2011; E. Shutova et al., 2016). Because of the grounded nature of metaphors, metaphors can occur in different modalities: visual and multimodal metaphors are typically used in mass media communication (Forceville, 2002). In the visual domain, there have been works accounting for metaphor localization, understanding and generation (D. Zhang et al., 2021; Chakrabarty et al., 2023; Xu et al., 2022; Hwang & Shwartz, 2023; Akula et al., 2023). Concerning metaphor generation, Akula et al. (2023) proposed a task that involves generating an image that effectively conveys the metaphorical message provided as the text prompt. Additionally, Chakrabarty et al. (2023) proposed an alternative task for generating visual metaphors from linguistic metaphors using Chain-of-Thought prompting, showing improvements in the quality of generated visual metaphors. Nevertheless, the common aspect across these studies is that the metaphorical quality was already present either in the textual or in the visual input employed. Interestingly, (Özbal et al., 2019) and (Yosef et al., 2023) dealt with literal images and textual metaphors; however their tasks focused on association between the text and images, rather than on metaphor generation.

Analogical Reasoning

Despite the facts that, within NLP the word analogy task (Mikolov, Yih, & Zweig, 2013) became widely used to evaluate word embedding, and despite the relevance of the analogical view on metaphors (Gentner, 1983), to the best of these authors' knowledge there is a very limited research on metaphorical analogy tasks (Czinczoll et al., 2022; Pitarch et al., 2023). In particular, the SCAN dataset (Czinczoll et

al., 2022) includes metaphorical and scientific analogies, retrieved from literature (Lakoff et al., 1991; Musolff, 2000; Lakoff & Johnson, 1980b; P. D. Turney, 2008; Lakoff & Wehling, 2012). Following this work, Pitarch et al. (2023) created the MEAN dataset of metaphorical textual analogies, each with three distractors, evaluating models for multiple choice classification task. Although less frequent than textual analogy, there have been relevant works on visual analogy. Previous works have focused on representing transformations between pairs of images, (Jacobs et al., 2001; Memisevic & Hinton, 2010; Lovett & Forbus, 2017). Relevant, Radford et al. (2015) developed the new class of DCGANs able to conduct vector arithmetic allowing for manipulation of images. Furthermore, Tewel et al. (2022) introduced the model ZeroCap which was also evaluated on successfully solving multimodal analogies problems based on visual-semantic arithmetic. Moreover, Sadeghi et al. (2015) proposed VISALOGY consisting in solving visual analogies by discovering the mapping between A and B and then extend it to C and D, and choosing the correct image from a set of candidates. Different from previous works, which mostly focused on image transformations, relevant to this paper is the work of Bitton et al. (2023) who introduced the large Visual Analogies of Situation Recognition (VASR) dataset.

Experiment 1: Multimodal Chain-of-Thought Prompting for Metaphor Generation

Dataset and Experimental Setup

We retrieved 300 conceptual metaphors from the MetaNet Metaphor Wiki [3] (Dodge et al., 2015), a comprehensive repository of conceptual metaphors based on the Conceptual Metaphor Theory (Lakoff & Johnson, 1980b). These metaphors follow the standard format, where a target domain is compared to a source domain, e.g., ARGUMENT IS WAR. We collected two images for each metaphor from Google Image Search: one for the target domain and the other for the source domain. For the purpose of this experiment, we employed the multimodal model LLaVA 1.5 (Large Language and Vision Assistant) [4] (H. Liu et al., 2023), considered as the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. We propose an alternative approach for multimodal metaphor generation from non-metaphorical textual and visual inputs. Our approach is based on the multimodal chain-of-thought prompting technique (Z. Zhang et al., 2023; Huang et al., 2023), which follows a two-step process as described in Fig. 1, which provides a visual representation of the task in the case of the conceptual metaphor ENVIRONMENTAL HARM IS PHYSICAL INJURY. Firstly, after being fed with the non-metaphorical images of the target and source domains, the model generates two distinct captions, which are

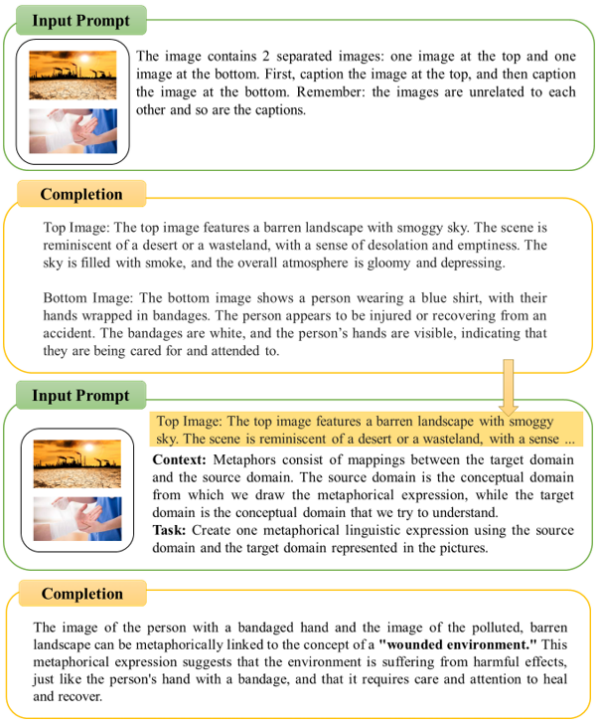

Figure 1: Visual representation of the task for the metaphor ENVIRONMENTAL HARM IS PHYSICAL INJURY.

then used as input for the second prompt, which involves generating a corresponding metaphorical expression based on the input, such as *wounded environment* for the example above. Additionally, the model provides an explanation of the new generated metaphor. To prove the utility of the method, the task was performed on a subset of the dataset by using only the second prompt, without Chain-of-Thought, and the results were less satisfactory, as the generation often did not involve a metaphor and failed to consider both images.

Evaluation

The evaluation of the generated metaphors was conducted in two ways: through BERTscore and by five human workers through Amazon Mechanical Turk. Concerning the automatic metaphor evaluation through BERTscore (T. Zhang et al., 2019), each generated metaphorical expression (*candidate*) was paired with a corresponding human metaphorical expression retrieved from MetaNet (*reference*), which provides real world examples of linguistic metaphors.[5] Concerning human evaluation, each generated expression was evaluated by five MTurk workers from English speaking countries.[6] To ensure the quality of the evaluation, the workers were given background knowledge regarding the Conceptual

[3] https://metaphor.icsi.berkeley.edu

[4] More precisely, we used the llava-v1.5 13B-4bit and the parameters were set as follows: temperature=0.2, max_new_tokens=1024. https://github.com/haotian-liu/LLaVA

[5] 75 metaphors were excluded from this evaluation, as they lacked examples.

[6] Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States. The workers were required to had an approval rate greater than 95% on 1000 prior approved HITs; their reward was $0.12 per task.

Agreement	Generated Metaphor	Conceptual Metaphor
5	*Wounded environment*	ENVIRONMENTAL HARM IS PHYSICAL INJURY
	House of thoughts	MIND IS A BUILDING
4	*Climbing the stairs of success*	ACHIEVING POWER IS MOVING UPWARDS
	Fighting the battle against cancer	CANCER PATIENT IS PHYSICAL COMBATANT
3	*Battle of words*	ARGUMENT IS WAR
	Walking down a road to recovery	CANCER IS A JOURNEY
2	*Embracing the warmth of friendship*	AFFECTION IS WARMTH
	Their love was as hot as the sun	PASSION IS HEAT
1	*Shaking hands over a book of contracts is like a marriage of business and legal agreements*	AGREEMENT IS PHYSICAL PROXIMITY
	A political body is like a human body	GOVERNMENT IS A PERSON

Table 1: Examples of the generated metaphors and their corresponding conceptual metaphors. The first column shows the workers' agreement on metaphoricity (with 5 being the highest and 1 the lowest).

Metaphor Theory, as well as positive and negative examples for the task. The workers had to choose whether the generated linguistic expression (e.g., *Wounded environment*) could be accepted as a linguistic metaphor for its corresponding conceptual metaphor (e.g., ENVIRONMENTAL HARM IS PHYSICAL INJURY) with the following Yes or No question: *Can the linguistic expression be considered as a linguistic metaphor for the provided conceptual metaphor?*. Additionally, they were asked other two yes/no questions regarding the familiarity and appeal of the expressions: *Have you encountered this linguistic expression before?* and *Is this linguistic expression appealing to you?*. To consider an expression as metaphorical, it had to be evaluated as such by at least three of the five workers. It is worth noting that it was not mentioned that the metaphors were not human-generated in order to prevent any potential bias.

Results and Discussion

Regarding the automatic evaluation, it is important to note that, overall the BERTscore between the generated and the human metaphors was Precision= 0.41, Recall= 0.43, and F1= 0.42. This suggests that there is a discrepancy between the model's generations and human examples, which may indicate that the generated metaphors may not be capturing the same semantic meaning as the human-generated ones. Additionally, this might be due to the difference in contexts. Human-generated metaphors often reference real-world examples, including real people and events; whereas the generated metaphors tend to be more generic and less nuanced. Moreover, BERTscore, while it has been shown to be effective for paraphrase detection (Devlin et al., 2019), might still have limitations in capturing the subtle differences and similarities in metaphorical language, which is typically subjective and context-dependent. Concerning the human evaluation, it was conducted across three criteria: *metaphoricity*, *familiarity* and *appeal* of the generated linguistic expressions. First, the generated expressions achieved an average *metaphoricity* score of 3.8, with 92% classified as metaphors by at least three workers. Notably, 92 expressions were unanimously recognized as metaphors, e.g., *Wounded environment*. All expressions were considered metaphors by at least one worker, indicating LLaVA 1.5's successfully generated metaphorical language from non-metaphorical visual inputs (Table 1). For *familiarity*, the average score was 2.95, with 67% deemed familiar by at least three workers. A total of 22 expressions were unanimously considered as familiar, e.g., *A journey through life* for PROGRESSING THROUGH LIFE IS MOVING ALONG A PATH. This indicates that the model generated not only familiar expressions but also novel and uncommon metaphors. Regarding *appeal*, the average score was 3.32, with 78% of expressions liked by at least three workers. A total of 37 metaphors were unanimously appealing, e.g., *Walking down a road to recovery* for CANCER IS A JOURNEY. These results suggest that the generated expressions were well-received overall. It might be meaningful to examine the distribution of the mean familiarity and appeal agreement scores in relation to the metaphoricity one. As illustrated in Fig. 2, the observed pattern seems to suggest that the mean familiarity and appeal scores exhibit contrasting trends across different metaphoricity scores. Interestingly, as the metaphoricity score increases, the familiarity score decreases while the appeal score increases. Metaphoricity scores 5 and 1 represent the extremes, with distinct differences in both familiarity and appeal. For the generated metaphorical expressions evaluated as such by all five workers, the mean score of familiarity is 2.92 and of appeal is 3.6; whereas for the expressions considered metaphorical only by one worker, the mean familiarity score is 3.67 and appeal is 3.0. With the exception of the expressions with metaphoricity score 2, which registered the lowest score (2.71) both for familiarity and appeal, the pattern seems to indicate that metaphoric expressions with higher metaphoricity scores tend to have lower familiarity and higher appeal. This means that the evaluators found the literal generated

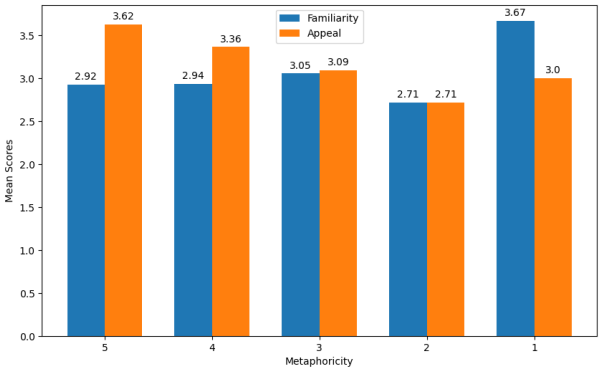

Figure 2: Mean familiarity and appeal scores for each metaphoricity score.

expressions (metaphoricity scores 1 and 2) to be more familiar compared to the metaphorical ones. The results suggest that the model was able to create novel metaphorical expressions which may differ from the more conventional metaphors, which the evaluators might have been more familiar with. Despite being less familiar, the metaphorical expressions were preferred over the non-metaphorical ones. These findings show that the model exhibited a degree of creativity in metaphor generation, as it generated novel or unconventional metaphorical expressions which were appreciated by human evaluators.

Experiment 2: Metaphorical Visual Analogy

Dataset and Experimental Setup

The goal of this second experiment is to complete metaphorical visual analogies of the kind *A : A' :: B : ?* by choosing the image that best fits as the target element *B'* from a set of four candidate images, which includes the correct image for *B'* and three well chosen *distractors*. The metaphorical analogies were selected from the SCAN dataset (Czinczoll et al., 2022), following the format of *battle : debate :: soldier : opponent* where the first word pair is the cross-domain mappings and the second word pair is the corresponding attribute mappings, it can be read as "If A is like A', then B is like B'". To adapt the dataset to the visual domain, images for each analogy component were retrieved from Google Image Search. Following Pitarch et al. (2023), distractors were selected controlling for the *semantic domain*, i.e., the general category shared by the target domain and target element, and *semantic attribute*, i.e., the role an element plays within that domain. A target element fits the analogy when it shares the target domain's semantic domain and the source element's semantic role. For instance, the target element *opponent* properly fits the analogy *battle : debate :: soldier : opponent* as it shares the same semantic domain of *debate* and the same semantic role of *opponent*. There are three distractors possibilities, examples for the analogy *battle : debate :: soldier : opponent* are provided: (i) distractor 1: the target element shares the same semantic domain as the target domain of the metaphor, but has a different attribute than the source element (e.g., *knowledge*); (ii) distractor 2: the target element shares the same attribute as the source element, but does not share the semantic domain with the target domain (e.g., *patient*); (iii) distractor 3: the target element has both different semantic domain and attribute (e.g., *medicine*). Following these criteria, we selected three distractors for each analogy. The MeVA dataset contains analogies that require to understand the full scene, as in the VASR dataset (Bitton et al., 2023), *situation recognition*, is the task of predicting the different *semantic role labels* (SRL) in an image. In addition to the manual verification, to ensure that the selected images had clear reference to specific semantic roles, they were verified through the Grounded Situation Recognition system [7]

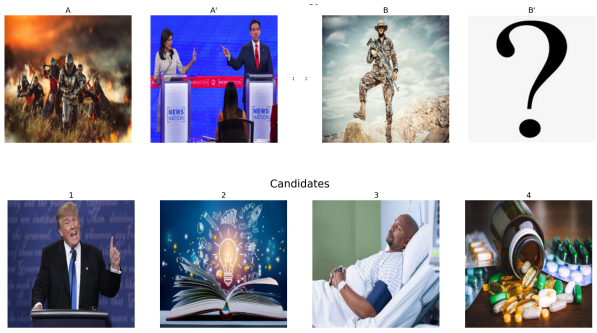

Figure 3: Visual representation of the task. The top row shows the analogy and below it the candidates.

(Pratt et al., 2020), which allows to identify the situation observed in the image and also visually ground the identified roles within the corresponding image with bounding boxes. Finally, the MeVA dataset ended up consisting of 333 analogies, each consisting of 7 images corresponding to a source domain, a target domain, a source element, and four target element candidates within which, one is the correct target element and the other three are the difficult distractors: e.g., battle : debate :: soldier : ? + [opponent, knowledge, patient, medicine]. Fig. 3 shows the visual representation of the MeVA task for the previous example.

Evaluation

Following the work of Bitton et al. (2023), the zero-shot arithmetic baseline consists in extracting visual features from pre-trained model for each image and represent the input in an arithmetic structure by taking the embedding of *B + A' - A*. Then, the cosine similarity is computed to each of the candidates and the most similar one is selected. Fig. 4 shows a visual representation of this operation. The task was conducted with the following models: ViT (Dosovitskiy et al., 2020), Swin Transformer (Z. Liu et al., 2021), DeiT (Touvron et al., 2021), ConvNeXt (Z. Liu et al., 2022) and EVA-02 (Fang et al., 2024).[8]. Additionally, the task was solved by GPT-4o, [9] to which was given the following multimodal zero-shot prompt to solve the task: `The following images form the analogy A : A' :: B : B'. image[A], image[A'], image[B]. Which of the following images best completes the analogy A : A' :: B : B'?. Candidates: image[1], image[2], image[3], image[4].` Moreover, we asked MTurk workers to solve the complete MeVA dataset. Each analogy was solved by 5 workers from English speaking countries, and they were given the same prompt as the one given to GPT-4o.

[7] https://vision-explorer.allenai.org/grounded_situation_recognition

[8] The exact versions are available in timm library: ViT Large patch32-384, Swin Large patch4 window7-224, DeiT Base patch16 384, ConvNeXt Large, EVA-02 large patch14 448

[9] https://openai.com/index/hello-gpt-4o/

Figure 4: Visual representation of the arithmetic structure of *B + A' - A*.

Methods	Models	Accuracy
Zero-Shot Arithmetic	ConvNeXt	**47.7%**
	ViT	45.0%
	EVA-02	42.6%
	Swin	41.1%
	Deit	40.2%
Zero-Shot Prompting	GPT-4o	**52.3%**
Humans		46.5%
Random		25%

Table 2: Humans' and models' performance in solving the MeVA dataset.

Results and Discussion

The results (Table 2) indicate that the best zero-shot arithmetic model in the MeVA task was ConvNeXt, which identified the correct candidate in 47.7% of cases, outperforming the other models ViT (45%), EVA-02 (42.6%), Swin (41.1%), and Deit (40.2%). GPT-4o emerged as the top performer overall, selecting the correct candidate in 52.3% of the dataset. Human participants demonstrated a 46.5% success rate, which is lower than both ConvNeXt and GPT-4o. Overall, the performance of both models and humans was relatively similar, ranging from 40% to 47%, with only GPT-4o reaching 52.3%. These scores suggest that both models and humans encountered challenges when attempting to solve these metaphorical visual analogies in a zero-shot setting with difficult distractors. However, it is important to note that both humans and models scored significantly higher than the random baseline of 25%. The task's difficulty was intentionally increased by requiring to choose the target element from a set of candidate images specifically curated to account for difference in semantic roles and domains. To understand with which semantic feature the models and humans encountered more difficulties (role distinction or domain distinction), an error analysis of the most frequently selected incorrect candidates was conducted. The percentages of the different error types made by the models and humans are shown in Table 3. Both models and humans most frequently selected the correct candidate image corresponding to the correct target element to solve the analogies. Considering the errors, among all models, the most errors were made by predicting *distractor 1* as the target element, which has the same semantic domain but different semantic role than the correct target element. This error ranged from the lowest 33% of VIT to the highest 41.1% of EVA-02. The prevalence of *distractor 1* as the most frequent error could point to a lesser knowledge in semantic roles comprehension. Following, the second most selected candidate was *distractor 2*, ranging from 9.3% of GPT-4o to 17.1% of Deit. Despite sharing the same semantic role as the correct target element, it exhibited lower selection rates compared to *distractor 1*, which shared the correct semantic domain. This suggests that the discrepancy in semantic domains played a greater role than the discrepancy in semantic roles. Finally, the least predicted target element was *distractor 3*, ranging from the lowest 2.4% of GPT-4o to the highest 7.8% of VIT. This was likely due to its differences in both semantic role and domain, making it more visually incongruous among the other candidates of the analogy.

Model	Distractor 1	Distractor 2	Distractor 3
ConvNeXt	35.1%	11.7%	5.4%
VIT	33.0%	14.1%	7.8%
EVA-02	41.1%	10.2%	6.0%
Swin	37.8%	13.8%	7.2%
Deit	35.7%	17.1%	6.9%
GPT-4o	36.0%	9.3%	2.4%
Humans	30.9%	15.0%	7.5%

Table 3: Percentage of errors per error type, calculated for each model.

Conclusion

Drawing from cognitive theories, this paper explored alternative approaches to metaphorical and visual processing by developing two tasks. The first task showed the model's ability to construct metaphorical meaning without direct metaphorical input. The second task showed that both models and humans were able to solve complex metaphorical visual analogies, reinforcing the relation between metaphor and analogy. It is important to state again that this is an exploratory work, and as such it has certain limitations. Firstly, the choice and quality of the images influence the results, and representing nuanced concepts visually introduced challenges, due to their inherent complexity. Additionally, while human evaluators received some background knowledge, they may still have lacked the necessary expertise, including expert reviewers could improve reliability. Finally, as the models occasionally generated hallucinations, future work could explore their performance in few-shot or supervised settings. Despite the limitations, this work showed promising results for future research in the field of metaphorical and visual reasoning, providing insights in AI's role in testing cognitive theories.

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program.

References

Akula, A. R., Driscoll, B., Narayana, P., Changpinyo, S., Jia, Z., Damle, S., ... others (2023). Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 23201–23211).

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Bitton, Y., Yosef, R., Strugo, E., Shahaf, D., Schwartz, R., & Stanovsky, G. (2023). Vasr: Visual analogies of situation recognition. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 241–249).

Chakrabarty, T., Choi, Y., & Shwartz, V. (2022). It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, *10*, 589–606.

Chakrabarty, T., Saakyan, A., Winn, O., Panagopoulou, A., Yang, Y., Apidianaki, M., & Muresan, S. (2023, July). I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 7370–7388). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.465/ doi: 10.18653/v1/2023.findings-acl.465

Chakrabarty, T., Zhang, X., Muresan, S., & Peng, N. (2021). Mermaid: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4250–4261).

Czinczoll, T., Yannakoudakis, H., Mishra, P., & Shutova, E. (2022). Scientific and creative analogies in pretrained language models. In *Findings of the association for computational linguistics: Emnlp 2022* (pp. 2094–2100).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Dodge, E. K., Hong, J., & Stickles, E. (2015). Metanet: Deep semantic automatic metaphor analysis. In *Proceedings of the third workshop on metaphor in nlp* (pp. 40–49).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., & Cao, Y. (2024). Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, *149*, 105171.

Fauconnier, G., & Turner, M. (2003). Conceptual blending, form and meaning. *Recherches en communication*, *19*, 57–86.

Forceville, C. (2002). *Pictorial metaphor in advertising*. Routledge.

Gao, G., Choi, E., Choi, Y., & Zettlemoyer, L. (2018). Neural metaphor detection in context. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 607–613).

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155–170.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. MIT press.

Gentner, D., & Clement, C. (1988). Evidence for relational selectivity in the interpretation of analogy and metaphor. *Psychology of learning and motivation*, *22*, 307–358.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., ... others (2023). Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, *36*, 72096–72109.

Hwang, E., & Shwartz, V. (2023, December). MemeCap: A dataset for captioning and interpreting memes. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 1433–1445). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.89/ doi: 10.18653/v1/2023.emnlp-main.89

Jacobs, C., Salesin, D., Oliver, N., Hertzmann, A., & Curless, A. (2001). Image analogies. In *Proceedings of siggraph* (pp. 327–340).

Lakoff, G., Espenson, J., & Schwartz, A. (1991). Master metaphor list (technical report). *Cognitive Linguistics Group University of California, Berkeley*.

Lakoff, G., & Johnson, M. (1980a). The metaphorical structure of the human conceptual system. *Cognitive science*, *4*(2), 195–208.

Lakoff, G., & Johnson, M. (1980b). *Metaphors we live by*. University of Chicago Press. Retrieved from https://books.google.it/books?id=r6nOYYtxzUoC

Lakoff, G., & Wehling, E. (2012). *The little blue book: The essential guide to thinking and talking democratic*. Simon and Schuster.

Liu, E., Cui, C., Zheng, K., & Neubig, G. (2022). Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies.*

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). *Improved baselines with visual instruction tuning.*

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the ieee/cvf conference on computer vision and pattern*

recognition (pp. 11976–11986).

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302.

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological review*, *124*(1), 60.

Mao, R., Li, X., Ge, M., & Cambria, E. (2022). Metapro: A computational metaphor processing model for text preprocessing. *Information Fusion*, *86*, 30–43.

Memisevic, R., & Hinton, G. E. (2010). Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural computation*, *22*(6), 1473–1492.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Musolff, A. (2000). *Mirror images of europe: Metaphors in the public debate about europe in britain and germany*. Iudicium Munich.

Özbal, G., Pighin, D., Strapparava, C., et al. (2019). A proverb is worth a thousand words: learning to associate images with proverbs. In *Proceedings of the 41st annual conference of the cognitive science society (cogsci'19)* (pp. 2515–2521).

Pitarch, L., Bernad, J., & Gracia, J. (2023). Mean: Metaphoric erroneous analogies dataset for ptlms metaphor knowledge probing. In *Proceedings of the 4th conference on language, data and knowledge* (pp. 147–152).

Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., & Kembhavi, A. (2020). Grounded situation recognition. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part iv 16* (pp. 314–332).

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Sadeghi, F., Zitnick, C. L., & Farhadi, A. (2015). Visalogy: Answering visual analogy questions. *Advances in Neural Information Processing Systems*, *28*.

Semino, E. (2008). *Metaphor in discourse*. Cambridge University Press. Retrieved from `https://books.google.it/books?id=QT1uilVRDTYC`

Shutova, E. (2010). Automatic metaphor interpretation as a paraphrasing task. In *Human language technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 1029–1037).

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 160–170).

Shutova, E. V. (2011). *Computational approaches to figurative language* (Tech. Rep.). Cambridge, UK: University of Cambridge, Computer Laboratory.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., & Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics*.

Su, C., Huang, S., & Chen, Y. (2017). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, *219*, 300–311.

Tewel, Y., Shalev, Y., Schwartz, I., & Wolf, L. (2022). Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 17918–17928).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347–10357).

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 248–258).

Turney, P., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 680–690).

Turney, P. D. (2008). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, *33*, 615–655.

Veale, T. (2016). Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the fourth workshop on metaphor in nlp* (pp. 34–41).

Xu, B., Li, T., Zheng, J., Naseriparsa, M., Zhao, Z., Lin, H., & Xia, F. (2022). Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international acm sigir conference on research and development in information retrieval* (pp. 2887–2899).

Yosef, R., Bitton, Y., & Shahaf, D. (2023). Irfl: Image recognition of figurative language. In *Findings of the association for computational linguistics: Emnlp 2023* (pp. 1044–1058).

Yu, Z., & Wan, X. (2019). How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 861–871).

Zhang, D., Zhang, M., Zhang, H., Yang, L., & Lin, H. (2021). Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th annual meeting of the as-*

sociation for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers) (pp. 3214–3225).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In *International conference on learning representations.*

Zhang, Z., Zhang, A., Li, M., Karypis, G., Smola, A., et al. (2023). Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.