AN ADVERSARIAL COLLABORATIVE FRAMEWORK FOR COMPREHENSIVE IMAGE CAPTIONING

Dinesh Chowdary Attota, Ying Xie & Linh Le*
College of Computing and Software Engineering
Kennesaw State University
Marietta, GA 30060, USA
dattota@students.kennesaw.edu, {yxie2,lle13}@kennesaw.edu

ABSTRACT

Comprehensive image captioning is a critical task with applications spanning a multitude of domains such as assistive technologies, automated content development, e-commerce, surveillance and security, etc. Research for image captioning has had a long history with numerous successes, however, a challenge remains in obtaining high quality labeled images for model training. While recent large visual language models such as GPT-4 are very capable of both generating detailed captions for images and generating labeled images for smaller models, they have certain issues. First, such models are expensive, either computationally or financially. Second, they require extensive prompt engineering to achieve the desirable outputs. Third, it is difficult to quantitatively evaluate the quality of captions that they generate without a ground truth. Accordingly, we present an automated framework that allows multiple small models to collaborate on the task of comprehensive image captioning without the needs of labeled images. In brief, the system operates by having a captioner generate and continuously improve descriptions of input images so that a generator can synthesize images that are more and more similar to the original ones. The similarity among images is calculated by an evaluator. Through experiment, we show that our framework provides considerable improvements in the comprehensiveness of captions over a standalone visual language model, bridging the gap between small models and larger ones such as GPT-4o.

1 INTRODUCTION

Image captioning occurs at the convergence of computer vision and natural language processing (NLP). This task is applicable across numerous fields, including assistive technology for the visually impaired (Rane et al., 2021; Ahsan et al., 2021; Makav & Kılıç, 2019), automated content development for social media (Wibowo et al., 2024; Kruk et al., 2019), video captioning (Chen et al., 2023; Sarto et al., 2022). In the past a decade, major improvements in image captioning have occurred, motivated by developments in deep learning and the availability of extensive datasets like COCO (Lin et al., 2015), Flickr30k (Plummer et al., 2016), and Visual Genome. Initial methodologies for image captioning utilized template-based techniques (Reale-Nosei et al., 2024) that tend to yield inflexible and generic captions deficient in fluency and detail. The emergence of deep learning has transformed the field of image captioning by introducing encoder-decoder architectures (O'Shea & Nash, 2015; Vinyals et al., 2015) that significantly improved the fluency and contextual relevance of generated captions. While these methods enhanced the fluency and relevancy of captions, they were constrained by their dependence on fixed-length feature representations and their failure to capture intricate visual features. Recent attention-based methods (Monteiro et al., 2017; Zhao et al., 2020; Xu et al., 2015) have enhanced caption quality by facilitating the dynamic alignment of visual and textual information. However, they continue to struggle in creating captions that are both comprehensive and contextually nuanced (Zhao et al., 2024; Tyagi et al., 2024), often yielding descriptions that are either overly general or inadequate in conveying subtle aspects of the context.

*Corresponding author

Another challenge of developing captioner models is the need of quality paired image-text data. However, image labeling tasks that involve human resources are expensive. On the other hand, while pretrained large models such as GPT4 (Achiam et al., 2023) are capable of generate such data, they are not without issues. First, good models tend to be large in sizes which demand expensive computational resources to host or financial resources to service them externally. Second, it could become complicated to design prompts for them to yield the desired outputs. Specifically, they tend to yield generic descriptions, omit details, and introduce irrelevant information, unless getting more specific instructions from users (Zhu et al., 2023; Betker et al., 2023; Ruiz et al., 2023; Bai et al., 2023). Lastly, labels' quality from pretrained models is difficult to quantify and evaluate.

With such motivations, in this paper, we presents an automated collaborative framework based on adversarial exchanges among multiple pretrained models for the task of comprehensive image captioning. In brief, the framework includes a **captioner** component which attempts to describe an input image, a **generator** component which synthesizes images from a given description, and an **evaluator** that computes image similarities. In operation, ***The captioner iteratively updates its captions so that the synthesized images from the generator become more and more similar to the input***. Intuitively, if a description helps the generator create images more similar to the original one, it entails more correct information. This captioning mechanism allows our collaborative to improve itself without the needs of finetuning labeled data while still having a concrete evaluation metric. In terms of architectures, the captioner contains a visual comprehension model that can provide initial descriptions and answer questions for an input image, and a language model to probe extending questions based on the current descriptions. The generator is an image generative model (IGM) that is conditioned by prompt texts or by both texts and images. A metric model is utilized as the evaluator to compute similarities for pair of images. All models in the framework are pretrained. Through experiment study, we show that our framework provides improvement over a standalone visual language model, bridging the gap between small models and larger one such as `GPT-4o`.

2 Related Work

Image captioning has been extensively studied in computer vision and natural language processing, with significant progress spurred by deep learning and enormous data sets. Initial methodologies depended on template-based techniques, wherein established sentence frameworks were filled with identified objects and characteristics. Although these systems were comprehensible and efficient in certain contexts, they generated inflexible and generic descriptions that were low in fluency and detail. For instance, Farhadi et al. (2010) (Farhadi et al., 2010) aligned identified objects with predetermined phrase structures, producing captions had difficulties in representing intricate relationships or contextual details.

Over the past decade, image captioning has evolved from simple template-based methods to advanced deep learning methodologies. Early template-based techniques, produced descriptions that were frequently inflexible and generic. The transition to encoder-decoder architectures marked a significant milestone for the field. In these systems, a Convolutional Neural Network (LeCun et al., 1995) encodes the visual content of an image into a fixed-length feature vector, and sequential model, such as Long Short-Term Memory (Aneja et al., 2018) or Gated Recurrent Unit (Chung et al., 2014), generates a descriptive sentence based on that representation. These models, detailed in studies like (O'Shea & Nash, 2015) and (Vinyals et al., 2015), enhanced the state-of-the-art by producing more fluent and contextually aware captions. However, the inherent limitation of fixed-length representations means that many models struggle to capture the fine-grained details of complex images. To mitigate this, some research works has introduced attention mechanisms that allow the model to dynamically focus on relevant image regions during the caption generation process (Geetha et al., 2020; Alzubi et al., 2021; Sairam et al., 2021; Subramanian et al., 2023; Ahmad et al., 2022). Despite these improvements, balancing computational efficiency with the need for detailed visual understanding remains an open challenge.

Recent advancements in extensive pre-training has resulted in the creation of vision-language models, including CLIP (Radford et al., 2021) (Agarwal et al., 2021) and BLIP (Li et al., 2022), which utilize vast datasets to acquire joint representations of visual and textual information, attaining superior performance by correlating visual attributes with natural language descriptions. However, these models lack iterative improvement processes, leading to captions that may overlook complex

aspects or inadequately represent the scene's complexity. To address this, Multiple approaches have employed iterative refining to address this. Cornia et al. (Cornia et al., 2020) suggested a reinforcement learning-based framework that optimizes a reward function to enhance captions, while Xian et al. (Xian et al., 2022) proposed an adaptive captioning model that generates intermediate descriptions and refines them over numerous iterations. Building on these ideas, recent work has explored the use of prompt-based methods to guide the refinement process (Hu et al., 2022). Some studies have employed prompts to generate targeted questions or instructions (Özdemir & Akagündüz, 2024) (Luu et al., 2024) that guide the model to focus on specific aspects of the image, such as object attributes, spatial relationships, or contextual details. An issue with prompt-based methods is that they are limited by the capability of the incorporated language models. Furthermore, evaluations of the captions remain questionable.

With such motivation, in this paper, we present a unique framework that synergistically blends vision and language models to accelerate these advances. Our iterative refinement framework uses a language model to create specific image questions and a vision comprehension model to answer them to improve the caption. Furthermore, the new details are only augmented to a current caption if they help a generator create images more similar to the original one. This iterative feedback loop makes sure the final caption has both high-level and fine-grained features, overcoming the limits of prior approaches and producing high-quality, contextually appropriate captions. Our framework also utilize a metric that directly measures the association between an image and a description to guide the captioning process.

3 THEORETICAL FRAMEWORK

In short, our adversarial collaborative framework comprises three components, 1) a **captioner** which includes a Language Model (LM) and a Visual Comprehension Model (VCM), 2) a **generator** which is an Image Generative Model (IGM), and 3) an **evaluator** which is an Image Metric Model (IMM). For a given image, the operation starts with an initial caption, which can come from the VCM. Next, the LM is provided with the initial caption and prompted to ask the VCM questions to extend it. The VCM answers the questions which form a set of candidate captions, and send those to the IGM. The IGM then generates images for each candidate captions as well as the initial caption. Finally, the IMM compares the synthesized images to the original image. Candidate captions of which reconstructed images are more similar to the original image than that from the initial caption are selected and aggregated to form the new caption. Intuitively, our framework aims to generate a caption such that a generator can reconstruct an image as similar to the original one as possible. Formally, one iteration of operation of the framework is as follows.

1. A given image X is input into the VCM to obtain an initial description C
2. The LM takes the current description C to form candidate questions $Q_1 \ldots \ Q_k$ asking about more details on the contents of the image X
3. The questions $Q_1 \ldots \ Q_k$ are returned to the VCM to obtain answers $A_1 \ldots \ A_k$, respectively. Each answer is then concatenated to C separately to form a set of *candidate descriptions*: $\hat{C}_1 = C + A_1, \hat{C}_2 = C + A_2, \ldots \ \hat{C}_k = C + A_k$
4. Provide C and all candidate descriptions $\hat{C}_1$ ... $\hat{C}_k$ to the IGM to synthesize images Z and $Z_1 \ldots \ Z_k$.
5. The IMM computes the similarities $S(\cdot)$ for the pairs: $S(X, Z)$, $S(X, Z_1) \ldots \ S(X, Z_k)$. Candidates i that yield improvements in similarity over the original description, i.e., $S(X, Z_i) > S(X, Z)$, are selected - their answers are aggregated with the current description to form the new one $C_{new} = C + \sum\{A_i \mid \forall S(X, Z_i) > S(X, Z)\}$.
6. Repeat steps 2-5 until no significant improvements in similarities can be obtained.

An illustration of the framework's operation is in Figure 1. Overall, this mechanism allows the framework to generate a very detailed description for any input images. Furthermore, the fine details in the caption are added sequentially and only if they help an image generator understand more about the original scene. While, at the moment, we are only focusing on the captioning task, the framework can be used to generate comprehensive captions for images for training or finetuning of other models. Different from arbitrarily using a large model such as `GPT-4o` to generate captions, in this framework, we have a concrete metric – the reconstructed image similarity – to evaluate the quality of the resulted caption.

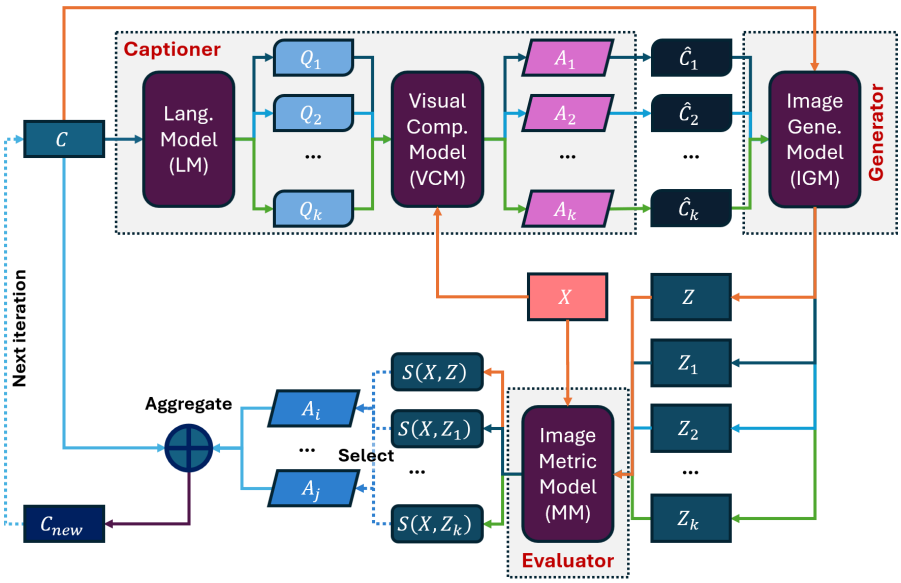


Figure 1: Adversarial framework for visual comprehension

4 Implementation

There are multiple approaches in both model selection and execution to realize the proposed adversarial collaborative framework. In this section, we describe our current implementation that is used for experiments in this paper in terms of model selection for components, and framework operations.

4.1 Components

Our selection for each component as illustrated in Figure 1 is as follows.

- The Visual Comprehension Model (VCM) in used is `MiniCPM-Llama3-V-2_5` (Yao et al., 2024), a pre-trained vision-language model designed to generate detailed textual descriptions of visual inputs by leveraging a robust multi-modal architecture.
- The Language Model at the moment is the pretrained `Llama-8.1-3-b-Instruct` (AI@Meta, 2024).
- The Image Generative Model (IGM) component is utilizing a pretrained `Stable-Diffusion-3-Medium` model. (Esser et al., 2024)
- The Image Metric Model (IMM) is implemented with an embedding-based metric approach. More specifically, to compute the similarity of two images, first they are transformed into embeddings by a pretrained vision transformer, `vit-base-patch16-224-in21k` (Dosovitskiy et al., 2021). The similarity of the two images are then calculated as the cosine similarity of the two embeddings.

4.2 Operation

The overall operational flow of the implemented framework is mostly similar to that of the theoretical framework that we discuss in Section 3. To start the framework (i.e., step 1 in Section 3), we feed the given image X and the prompt *"What is in the image? Explain in details."* to the VCM to obtain a caption C. While this prompt can be further enhanced for a better starting point, we are not focusing on prompt engineering in this paper. Next, we perform the iterative process of question generation and caption augmentation (i.e., steps 2 to 5 in Section 3). This process includes the steps of 1) question generation and answering, 2) image reconstruction, and 3) answer retrieval and caption augmentation. An illustration of the operational flow of the framework is showed in Figure 2 with detailed discussions in the following subsections.

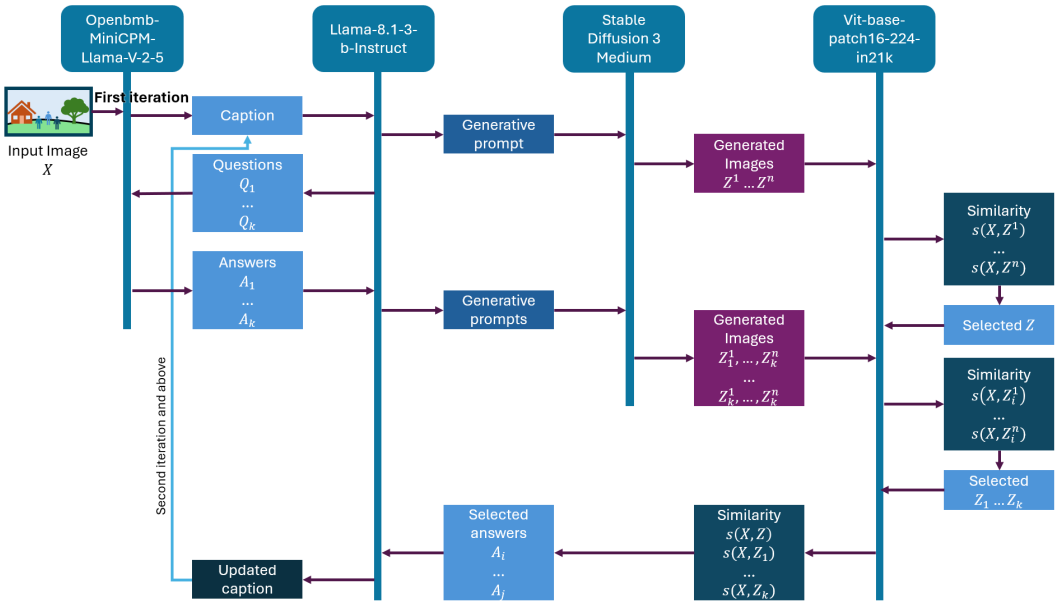


Figure 2: Operational Sequence of the Framework

4.2.1 Question Generation and Answering

The question generation and answering phase is key to our methodology, aimed at improving the first caption by integrating more details via a feedback loop between the language and visual language models. The caption C (that can either be the initial caption or a caption from a previous iteration) is provided to the LM along with a prompt to formulate a series of questions $Q_1, \dots Q_k$. The purpose of these questions is to explore the broadness and variety of the original caption, identifying any absent or incompletely defined elements. To guarantee that the produced questions are thorough and precise, we direct the language model's attention to particular aspects of the image that include:

- **Key Objects and Subjects**: Identify all the primary and secondary objects in the scene.
- **Relationships**: Explore the spatial relationships between different objects (e.g., proximity, above, below, in front).
- **Interactions**: Highlight any visible interactions between objects, animals, or people.
- **Visual Details**: Ask questions about colors, patterns, and textures of objects, including distinctive features.
- **Poses and Actions**: Understand the posture or movement of any animals or people.
- **Background Elements**: Capture information about the setting, such as environmental features (e.g., trees, buildings, sky).
- **Scale and Proportions**: Ask about relative sizes and the spatial alignment of objects.
- **Lighting and Shadows**: Investigate lighting direction, shadows, and any effects that contribute to the realism of the scene.

Furthermore, questions that are generated in the previous iterations are included in the prompt of the current iteration. This inclusion is for the model to know what has been asked and avoid repeats of poor questions in the set that is being generated. The generated questions $Q_1, \dots Q_k$ are returned to the VCM along with the original image to obtain their answers, i.e., $A_1 \dots A_k$. Finally, each answer A_i is added to a prompt along with the original caption C for the LM to recompose and improve coherency and fluency, yielding a candidate caption $\hat{C}_i$.

4.2.2 IMAGE RECONSTRUCTION

The original caption C and the candidate captions $\hat{C}_1, \dots \hat{C}_k$ are prompted to the IGM to obtain "reconstructred" versions of the image X, Z and $_1 \dots Z_k$, respectively. Furthermore, as the image generation process is highly subjected to randomness, for each caption $C_* \in \{C, \hat{C}_1, \dots \hat{C}_k\}$, we generate n image versions $Z_*^1 \dots Z_*^n$. Then, the version among the n ones that is the most similar to the input image X (the highest embedding similarity to X), Z_*, will be selected for the next step. The output of the image reconstruction phase is a generated image for C, Z, and k generated images $Z_1, \dots Z_k$. In the current implementation, we use $n = 15$.

4.2.3 ANSWER RETRIEVAL AND CAPTION AUGMENTATION

The answer retrieval step aims to select answers A_i that help extending the original caption C. Specifically, if an image Z_i generated by using the candidate caption $\hat{C}_i$ (resulted from A_i) yields a higher similarity value to the original image X than that between Z and X, A_i is selected. Finally, the set of selected answers along with the caption C are input to the LM for recomposing into a coherent caption. The overall architecture and flow of our proposed adversarial collaborative framework is shown in Figure 2 .

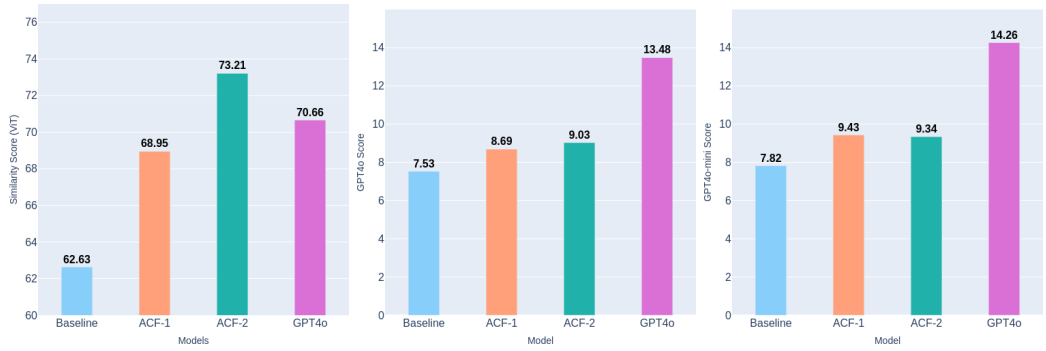
5 EXPERIMENTS

All experiments are conducted in workstations using NVIDIA-A100 Graphical Processing Units. As benchmark data, we randomly sample 500 images from the ImageNet dataset (Deng et al., 2009). A challenge for our method is how to derive evaluation metrics. While datasets with long captions exist, we cannot ensure the comprehensive levels of their captions. Accordingly, a supervised matching metric between our generated captions and the labels in such data is not suitable for our use case. Therefore, we derive two unsupervised evaluation metrics for the comprehensiveness of a caption with respect to an image as follows.

Image Similarity Score. Given an image-caption pair (X, C), the comprehensiveness of the captioned C is evaluated by the similarity between X and an image Z generated using C by an image generation model (IGM). This score follows the intuition of our collaborative framework: If a caption helps a generative model create an image that is more similar to the original one, the caption is more comprehensive. Given a datasets of multiple image-caption pair, we calculate the similarity score for all pairs then average to obtain the overall score for the caption sets. In our experiment, the IGM is `Stable-Diffusion-3-Medium`.

GPT-Based Score The second metric is based on using a large language model (LLM) such as `GPT-4o` as an evaluator. Specifically, given an image-caption pair (X, C), we perform a segmentation-verification pipeline. First, we segment both caption C into contextually coherent, self-sufficient segments using the LLM. This is done by prompting the LLM with "Split the caption into independent single units of information. Make sure the information is objective.", follows by C. Next, each segment along with the image X is provided to a visual language model to verify whether its content is correctly presented in the corresponding image or not. We also calculate the cosine similarity among the correct segments of to filter out overlapping contents. A segment pair with cosine similarity over 0.8 is considered duplicated and one of the segment is removed. Finally, the count of correct segments remaining is used as the GPT-Based comprehensive score for the caption C. In our experiments, we use two GPT-Based Score, one utilizing `GPT-4o` for both segmentation and verification, and one with `GPT-4o` for segmentation and `GPT-4o-mini` for verification. We want to note that this metric could degrade in quality when using smaller models in either tasks.

To compare our framework against existing method, we utilize two benchmark models, a standalone `MiniCPM-Llama3-V-2_5` (the VCM component of the framework) and `GPT-4o`. Along with the input images, both models are prompted with "What is in the image? Explain in details", which is the same prompt we use to generate the initial captions in our framework. For our method, we report the caption quality after one pass and two passes of augmentation. Figure 3(a) displays the average image similarity scores of the four models, `MiniCPM-Llama3-V-2_5` (Baseline), ours after one pass (ACF-1), and ours after two passes (ACF-2), and `GPT-4o`. Figures 3(b)(c) illustrate the average of the two GPT-Based scores for the captions from the four models. Figure 3(b) shows the score from the `GPT-4o`-`GPT-4o` pipeline, and 3(b), the `GPT-4o`-`GPT-4o-mini` pipeline.



(a) Similarity Score (b) GPT-4o Score (c) GPT-4o-mini Score

Figure 3: Framework Evaluation Results

The baseline similarity score for images produced from the original captions is 62.63. After the initial iteration of our augmentation procedure, the similarity score increases to 68.95, showing that the augmented captions offer a more comprehensive and precise depiction of the images. Following with the second iteration, the similarity score increases to 73.21, so demonstrating the efficacy of our iterative caption enhancement methodology. Interestingly, the captions generated by `GPT-4o` reach an average similarity score of 70.66, which is close to that of our framework after one pass.

The score patterns vary considerably in the two GPT-Based scores. In the first GPT-Based score (Figure 3(b)), the baseline model's captions yield an average of 7.53 correct segments. After one pass of augmentation, our framework reaches an average of 8.69 correct segments, suggesting that the first round of iterative augmentation effectively incorporates additional information that correspond with the image content. After two passes, the count further improves to 9.03, suggesting that subsequent refinement continues to enhance the caption's descriptive accuracy and completeness. Without surprises, captions from the much larger `GPT-4o` reach a score of 13.48. In the second GPT-based score (Figure 3(b)), we observe a fairly similar result. Our framework adds good improvement to the baseline, while all are surpassed by `GPT-4o`. Overall, while our framework certainly do not outperform `GPT-4o`, it does help close the gap of performance between very large models and small ones.

6 Conclusion

Automated image captioning, especially with comprehensive description, is an important task with numerous applications, from assistive technologies, automated content development, e-commerce, to surveillance and security, and so on. Research in image captioning has kept involving with successes and breakthroughs. Regardless, a challenge remains in obtaining high quality labeled images for model training. Labeling is expensive if involve human. Large visual language models, why very capable, also come with issues such as demanding in resources, require extensive prompt engineering to achieve the desirable outputs, and difficulty in evaluation without a ground truth.

In this paper, we proposed an innovative collaborative framework for iterative image captioning, incorporating visual understanding, linguistic modeling, and generative methods. Our methodology enhances captions by iterative question formulation, answer acquisition, and evaluation via reconstructed image similarity, delivering a more precise depiction of visual content. Experiments on 500 ImageNet images has demonstrated considerable improvements in captions' comprehensiveness, which are validated through similarity assessments and GPT-Based verification. Results show the increased similarity scores and more detailed information in captions across iterations, justifies the framework's effectiveness. In contrast to conventional methods, our methodology integrates explicit assessment through image reconstruction similarity, offering a definitive metric for caption quality. In future research, we would investigate on expanding the framework on large scale with fine-tuning approaches and also possible enhancements in image and question-generation methodologies.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

Rana Adnan Ahmad, Muhammad Azhar, and Hina Sattar. An image captioning algorithm based on the hybrid deep learning technique (cnn+ gru). In *2022 International Conference on Frontiers of Information Technology (FIT)*, pp. 124–129. IEEE, 2022.

Hiba Ahsan, Nikita Bhalla, Daivat Bhatt, and Kaivankumar Shah. Multi-modal image captioning for the visually impaired. *arXiv preprint arXiv:2105.08106*, 2021.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Jafar A Alzubi, Rachna Jain, Preeti Nagrath, Suresh Satapathy, Soham Taneja, and Paras Gupta. Deep image captioning using an ensemble of cnn and lstm based deep neural networks. *Journal of Intelligent & Fuzzy Systems*, 40(4):5761–5769, 2021.

Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2:3, 2023.

Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Hongyang Chao, and Tao Mei. Retrieval augmented convolutional encoder-decoder networks for video captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1s):1–24, 2023.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL https://arxiv.org/abs/1412.3555.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10578–10587, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pp. 15–29. Springer, 2010.

G Geetha, T Kirthigadevi, G Godwin Ponsam, T Karthik, and M Safa. Image captioning using deep convolutional neural networks (cnns). In *Journal of Physics: Conference Series*, volume 1712, pp. 012015. IOP Publishing, 2020.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*, 2019.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Duc-Tuan Luu, Viet-Tuan Le, and Duc Minh Vo. Questioning, answering, and captioning for zero-shot detailed image caption. In *Proceedings of the Asian Conference on Computer Vision*, pp. 242–259, 2024.

Burak Makav and Volkan Kılıç. A new image captioning approach for visually impaired people. In *2019 11th international conference on electrical and electronics engineering (ELECO)*, pp. 945–949. IEEE, 2019.

João Monteiro, Asanobu Kitamoto, and Bruno Martins. Situational awareness from social media photographs using automated image captioning. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 203–211. IEEE, 2017.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. URL `https://arxiv.org/abs/1511.08458`.

Övgü Özdemir and Erdem Akagündüz. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1562–1571, 2024.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL `https://arxiv.org/abs/1505.04870`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

Chinmayi Rane, Amol Lashkare, Aarti Karande, and YS Rao. Image captioning based smart navigation system for visually impaired. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pp. 1–5. IEEE, 2021.

Gabriel Reale-Nosei, Elvira Amador-Domínguez, and Emilio Serrano. From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, pp. 103264, 2024.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Gourishetty Sairam, Mounika Mandha, Penjarla Prashanth, and Polisetty Swetha. Image captioning using cnn and lstm. In *4th Smart Cities Symposium (SCS 2021)*, volume 2021, pp. 274–277. IET, 2021.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th international conference on content-based multimedia indexing*, pp. 1–7, 2022.

R Raja Subramanian, Patan Khamrunnisa, B Sai Krishna Vikas Reddy, and Chilakuri Vishnu Chaithanya. Image caption generation using cnn-gru approach. In *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, pp. 1–6. IEEE, 2023.

Shourya Tyagi, Olukayode Ayodele Oki, Vineet Verma, Swati Gupta, Meenu Vijarania, Joseph Bamidele Awotunde, and Abdulrauph Olanrewaju Babatunde. Novel advance image caption generation utilizing vision transformer and generative adversarial networks. *Computers*, 13 (12):305, 2024.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Jessica Lynn Wibowo, Gabriel Seemore Gunawan, and Ivan Sebastian Edbert. Enhancing social media accessibility: Automatic alternative text generation in x by image captioning. In *2024 14th International Conference on System Engineering and Technology (ICSET)*, pp. 123–128. IEEE, 2024.

Tiantao Xian, Zhixin Li, Zhenjun Tang, and Huifang Ma. Adaptive path selection for dynamic image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5762–5775, 2022.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/xuc15.html`.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint 2408.01800*, 2024.

Fengzhi Zhao, Zhezhou Yu, Tao Wang, and Yi Lv. Image captioning based on semantic scenes. *Entropy*, 26(10):876, 2024.

Wentian Zhao, Xinxiao Wu, and Jiebo Luo. Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Transactions on Image Processing*, 30:1180–1192, 2020.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.