Illuminating the Shadows - Challenges and Risks of Generative AI in Computer Vision for Brands

Alex Kaplunovich[1]

[1] University of Maryland, Baltimore, MD 21250, USA
akaplun1@umbc.edu

Abstract. The rapid advancements in generative AI have significantly transformed computer vision, presenting both opportunities and challenges for brands. This paper digs into the risks associated with the use of generative AI in computer vision applications, particularly focusing on brand integrity, detection and security. One primary concern is the ethical implications, where LLMs can amplify biases, produce fake product images and propagate harmful stereotypes, affecting brand reputation. The rise of deepfakes and AI-generated content poses a substantial risk of disinformation, leading to potential misuse in creating misleading advertisements or damaging a brand's image through falsified media.
Legal challenges are another critical aspect, especially concerning intellectual property rights and copyright issues. The ability of generative AI to produce content indistinguishable from original works raises questions about ownership, detection techniques and the legal frameworks required to protect brands. To address these tasks, the paper explores various generation, detection, and mitigation strategies emphasizing the importance of developing responsible and trustworthy generative AI technologies. By highlighting these issues, the paper aims to foster a balanced discourse on the ethical and practical aspects of generative AI in computer vision for brands, shares detection results, and suggests mitigation strategies.

Keywords: Generative AI, computer vision, LLM, prompt engineering, text-to-image, image-to-text, brand detection, deepfakes.

1 Introduction

Large Language Models are changing our lives helping to solve many complex tasks. Until recently, it was primarily text generation. However, with introduction of several state-of-the art models (DALL-E 3 or MidJourney), image generation has become an easy task, just a matter of providing a prompt. While generated image quality grows drastically, there are no reliable tools to detect whether an image is generated or real. In the light of that, brand generation, detection and altering becomes a huge problem. Nowadays, good actors can use advances of Gen AI for dataset augmentation, brand improvement, virtual try-ons, marketing and advertising. However, bad guys potentially can temper with brand's intellectual property, commit fraud and disinformation, falsify products and damage corporate reputation. We will investigate modern techniques that are available for generation, detection and possible mitigation.

Generative AI models such as GPT-4o and DALL-E, have transformed content creation by leveraging vast datasets to generate text, images, and other media forms. While these technologies offer numerous benefits, they also pose substantial risks to brands and IP. This paper aims to shed light on the darker aspects of generative AI, focusing on issues related to brand misrepresentation, copyright infringement, and the ethical use of AI-generated content.

2 Generative AI Techniques for Images and Brands

2.1 LLM models (text-to-image)

The greatest value of generative AI is its accessibility: anyone can open a web browser, type a prompt, and receive a response in the form of text or an image. This task can also be accomplished by writing a few lines of code to call a specific LLM API. Just two years ago (ChatGPT was released in November 2022), solving a logical puzzle, researching a subject, or drawing a complex image required substantial time. Today, AI can provide a satisfactory answer in mere seconds.

Leading image generative models include DALLE-3 (OpenAI), MidJourney, and Stable Diffusion. The latter is an open-source model available for free. These models offer web interfaces as well as API access, enabling dynamic image generation through code. DALLE-3 pricing is 8 cents per high-resolution 1792×1024 generated image.

While image generation is relatively easy and inexpensive, it still takes time. The generation time depends on the complexity and length of the prompt [6], as well as model server's utilization. On average, it takes between 5 to 20 seconds per image. Our experiments confirm that using DALLE-3, it takes approximately 11 seconds per image generation.

This means creating a new, massive set of images or generating a video remains time-consuming. For example, one minute of video (60 seconds * 24 frames per second = 1440 frames) would take approximately 4.4 hours (given an average of 11 seconds per image) and cost around $115 (at $0.08 per image).

2.2 Visual Prompt Engineering

Prompt engineering - the new uniform programming language. By issuing well-crafted questions to LLMs, it is possible to accomplish a great deal. Numerous research papers ([2] and [3]) discuss prompting techniques. However, most of this research has focused on NLP and textual information until very recently. There are far fewer papers related to prompts for image generation ([4] and [5]). It will take some time to finalize visual prompt engineering techniques, tricks, and best practices, simply because the field is relatively new compared to text generation.

2.3 Image analysis (image-to-text and vision transformer models)

There are numerous models and tools that can analyze images and generate text. These models include OpenAI's ChatGPT-4o, Amazon's Rekognition, and the open-source Salesforce/blip-image-captioning-large. As of July 2024, there are 602 image-to-text

models available in the Hugging Face hub. These models take an image and describe it in human words.

Vision Transformers (ViT) ([7], [8]), such as open-ai/clip-vit-large-patch14, calculate the similarity between an image and a user-provided string, returning a confidence classification score. Huggingface hosts 486 Zero-Shot Image Classification models. These models are widely used by leading AI companies to automatically identify image content or filter out potentially harmful or obscene scenes. The details and quality of these models are impressive. They can not only detect brands but also identify image objects and describe scenes. Figure 1 demonstrates the image-to-text capabilities.

Fig. 1. ChatGPT-4o output for the prompt "describe the image please": "The image shows a kitchen scene with a brown paper bag labeled **"NIKE"** on a countertop. The bag contains various items, including fresh vegetables in plastic bags. In the background, the kitchen counter features several items, such as a teal woven basket with tomatoes, a cutting board, a white plate with remnants of food, and a large knife. The kitchen appliances, including a stove and a microwave, are visible in the background, indicating a typical domestic kitchen setting."

2.4 Object/Brand detection models

There is a class of models that can detect certain objects in images. Usually, these models are trained for a specific set of classes and can be finetuned for user-specified categories. Examples of these models are facebook/detr-resnet-50 [9] and YOLOv8 [10]. The power of these models lies in their ability to be trained via transfer learning to recognize specific brands with relatively few examples (50+). These models not only detect objects but also return bounding boxes and confidence scores.

We asked ChatGPT-4o to generate images with popular brand products. Fig. 2 demonstrates results of Detr50 model finetuned for brand detection. It identified brands with very high confidence scores (up to 99%). As of July 2024, the Huggingface hub contains 2,248 assorted object detection models available for free download.

3 Dark side or Malicious Generative AI for Brands

Given the advanced stage of vision LLMs, it is relatively simple to use state-of-the-art

models for brand misrepresentation and misinformation. As these models become more accessible to the general public, legal entities should be aware of potential fakes and take appropriate measures to protect their intellectual property. While there are several research papers ([11] and [12]), the scientific community is still developing standards and best practices to address this problem, especially considering the rapid development in the field, which is continually improving the quality and speed of image generation.

Brands are built on trust, consistency, and the unique identity they project to consumers. LLMs can undermine these principles by creating counterfeit or misleading content that appears to be endorsed by or associated with reputable brands. Below are several case studies.

Fig. 2. facebook/detr-resnet-50 model finetuned for brand detection identifies popular brands in AI generated images (see red bounding box, and detected brand name on grey background).

3.1 Fake Brand products

By design, text-to-image models could generate anything user asks. Fake products of well-known brands can easily damage reputation and confuse consumers. Fig. 3 demonstrates several “unusual” products from popular brands that were generated by LLMs.

3.2 Brand Tampering and Modification

Similarly, it is possible to modify a brand's logo - its font, color, or shape. While authorized personnel may use this capability for brand improvement, unauthorized modifications constitute copyright infringement. We will discuss possible mitigation strategies later in section 4.

Fig. 3. Fake toilet products from Apple, Ford and Mercedes-Benz produced by ChatGPT-4o

3.3 Unauthorized Brand usage

Visual Large Language Models were trained on immense set of images. That is why they could generate images containing known brand products and logos. However, each AI generated image might contain hallucinations, non-existent products (Fig. 3), and can confuse customers if published somewhere. Such practice is likely illegal, especially if generation was not authorized by the brand rights holders.

3.4 Fake Brand Endorsement

Deepfake technology, powered by generative AI, has been used to create videos or images that falsely represent brand endorsements or statements. Potentially, by utilizing multimodal LLMs, any person's image and voice can be used to promote any brand.

4 Mitigation

Fortunately, there are techniques addressing brand integrity and deepfake issues. It is possible to integrate guardrails into generative AI pipelines. For generation, we can use models like meta-llama/LlamaGuard-7b or Amazon Bedrock's Guardrails to filter out unwanted generative prompts. More and more ML companies are adopting these filtering technologies to promote safe and ethical AI, thereby avoiding potentially expensive and damaging lawsuits. For example, ChatGPT does not generate images with Disney products, providing the following message*: 'I was unable to generate images of Disney products due to our content policy. If you have any other requests or need images of different themes, please let me know!'* It is likely that more brands will establish similar legal agreements with vendors.

There are also several techniques to identify deepfakes [13]. Modern image-to-text models (like ChatGPT-4o) can answer the question 'Is the image AI-generated?' We generated 100 images with brands (see Fig. 2, Fig. 3), and the AI detected 99% of deepfakes with the following reasoning: *'Yes, the image appears to be AI-generated.'*

Watermarking and tracking can help intellectual property holders by placing unique metadata to track content. We believe that more legal regulations, policies, practices, and AI tools will be implemented shortly to preserve brands' copyright and integrity.

5 Conclusion

Generative AI offers transformative potential for brands, enhancing marketing, product design, and customer engagement with unprecedented creativity and personalization. However, this technology also presents significant risks, including deepfakes, counterfeit products, and unauthorized use of brand logos, which can damage consumer trust and brand reputation.

To mitigate these risks, it is crucial to establish robust legal frameworks and ethical guidelines that define acceptable use and protect intellectual property. Advanced technological solutions, such as real-time detection algorithms and digital watermarking, must be integrated into corporate processes to proactively defend against misuse.

Collaboration among governments, attorneys, industry leaders, and technologists is essential to create a comprehensive approach to the responsible use of generative AI.

By addressing the darker aspects of generative AI, we can harness its benefits while safeguarding brand integrity and protecting intellectual property rights. This balanced approach ensures that LLMs contribute positively to the brand landscape, driving innovation and efficiency, while maintaining trust and authenticity.

References

1. OpenAI pricing web site, https://openai.com/api/pricing/
2. White, Jules, et al. "A prompt pattern catalog to enhance prompt engineering with chatgpt." *arXiv preprint arXiv:2302.11382* (2023).
3. Meskó, B., 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. Journal of medical Internet research, 25, p.e50638.
4. Wang, Jiaqi, et al. "Review of large vision models and visual prompt engineering." *Meta-Radiology* (2023): 100047.
5. Strobelt, Hendrik, et al. "Interactive and visual prompt engineering for ad-hoc task adaptation with large language models." IEEE transactions on visualization and computer graphics 29.1 (2022): 1146-1156.
6. Lin, Kevin, et al. "Designbench: Exploring and benchmarking dall-e 3 for imagining visual design." arXiv preprint arXiv:2310.15144 (2023).
7. Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
8. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
9. Carion, Nicolas, et al. "End-to-end object detection with transformers." *European conference on computer vision*. Cham: Springer International Publishing, 2020.
10. Yolov8 model documentation, https://docs.ultralytics.com/
11. Brundage, Miles, et al. "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation." *arXiv preprint arXiv:1802.07228* (2018).
12. Blauth, et al. "Artificial intelligence crime: An overview of malicious use and abuse of AI." *Ieee Access* 10 (2022): 77110-77122.
13. Rana, Md Shohel, et al. "Deepfake detection: A systematic literature review." *IEEE access* 10 (2022): 25494-25513.