GPT-4 Vision: A Comprehensive Guide for Beginners

With the release of ChatGPT, amidst the AI hype, OpenAI launched GPT-4, its flagship product with extraordinary generative AI capabilities. During its launch in March 2023, it spoke about the possibilities of multi-modal generative AI through GPT-4. Multi-modality in generative AI refers to the ability of a system to process, understand, and/or generate output for more than one type of data.

Learn more about how LLMs work today by starting our Master Large Language Models (LLMs) Concepts Course.

After months of silence, in September 2023, OpenAI announced this multi-modal capability to ChatGPT, which can now see, hear, and speak, indicating new image and voice capabilities rolled out to ChatGPT. The multi-modal capability of AI systems, when perfected, is expected to unlock new possibilities and innovations for multiple industries.

In this tutorial, we will introduce the image capabilities and understand the GPT-4 Vision model, which enables the ChatGPT to "see." We would finally understand the current limitations of the model and leave you with further resources.

What is GPT-4 Vision?

GPT-4 Vision (GPT-4V) is a multimodal model that allows a user to upload an image as input and engage in a conversation with the model. The conversation could comprise questions or instructions in the form of a prompt, directing the model to perform tasks based on the input provided in the form of an image.

The GPT-4V model is built upon the existing capabilities of GPT-4, offering visual analysis in addition to the existing text interaction features. We have covered the beginner-friendly introduction to the OpenAI API that'll help you get up to speed on the developments prior to the release of the GPT-4V model.

Key Capabilities of GPT-4 Vision

- **Visual inputs:** The key feature of the newly released GPT-4 Vision is that it can now accept visual content such as photographs, screenshots, and documents and perform a variety of tasks.
- **Object detection and analysis:** The model can identify and provide information about

objects within images.
- **Data analysis:** GPT-4 Vision is proficient in interpreting and analyzing data presented in visual formats such as graphs, charts, and other data visualizations.
- **Text deciphering:** The model is able to read and interpret handwritten notes and text within images.

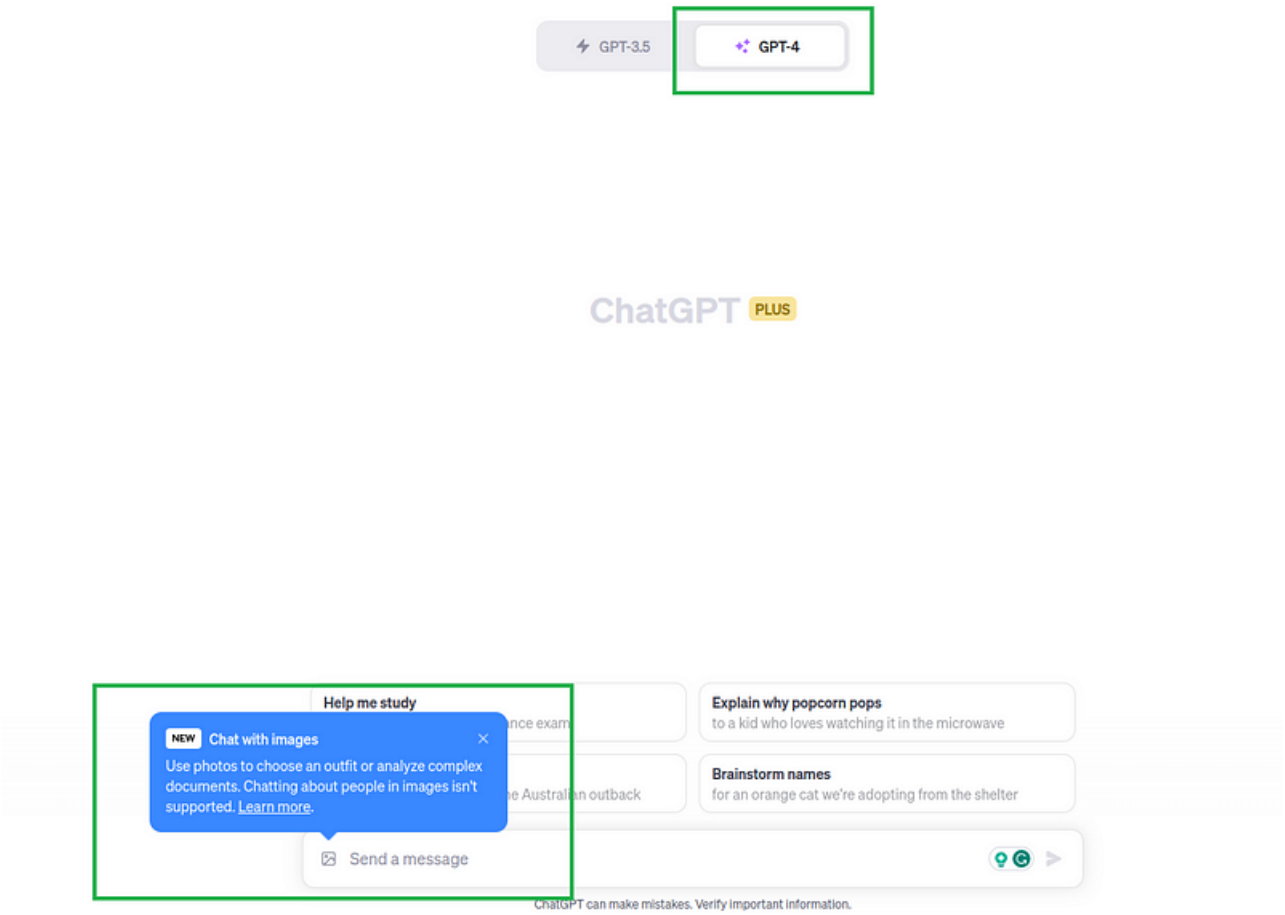
Now that we've understood some of the many capabilities of GPT-4 Vision, let's go hands-on to experience it better ourselves.

Hands-On: Getting Started with GPT-4 Vision

GPT-4 Vision is currently (as of Oct 2023) available for the ChatGPT Plus and Enterprise users only. ChatGPT Plus costs $20/month, which can be upgraded to from your regular free ChatGPT accounts.

Assuming you're completely new to ChatGPT, here's how to access GPT-4 Vision:

- Visit the OpenAI ChatGPT website and sign up for an account.
- Login to your account and navigate to the "Upgrade to Plus" option.
- Follow through with the upgrade to gain access to ChatGPT Plus (Note: this is a monthly subscription of $20)
- Select the "GPT-4" as your model in the chat window, as shown in the diagram below.



- Click on the image icon to upload the image, and add a prompt instructing the GPT-4 to perform it.

Here, GPT-4 Vision does a pretty good job of understanding that the kids were playing cricket based on the bat in a child's hand. In the world of AI, this task is known as object detection, where the objects identified are children and the bat. Check out our tutorial on YOLO object detection to learn more.

Similarly, you can proceed to perform a different task better depending on your use case.

GPT-4 Vision Real World Use-Cases and Examples

Now that we have understood its capabilities, let us extend them to some practical applications in the industry:

1. Academic research

GPT-4 Vision's integration of advanced language modeling with visual capabilities opens up new

possibilities in academic fields, particularly in deciphering historical manuscripts. This task has traditionally been a meticulous and time-consuming endeavor carried out by skilled paleographers and historians.

We first give an image that appears to be part of an old newspaper article:

GPT-4 Vision does a good job of reading the contents of the image and interpreting it:

The model was able to read, decipher the contents, and provide an analysis of it while providing a realistic answer, that some portions of the image are cut off and obscured.

However, it is also worth noting that the model struggles when given complex manuscripts, especially in other languages (more on this later), as seen below:

2. Web development

The GPT-4 vision can write code for a website when provided with a visual image of the design required. It takes from a visual design to source code for a website. This single ability of the model can drastically reduce the time taken to build websites.

Let us prompt the GPT-4 Vision with a hand-drawn simple design for a blogging website.

Once it provides the source code, we simply copy-paste and create the HTML and CSS files as instructed. Here's how the website looked:

Doesn't it look strikingly similar? Of course, we gave a simple example, but you could take it from here and develop a more complex and tailor-made website in a fraction of the time, thanks to the newly introduced GPT-4 Vision model.

3. Data interpretation

The model is capable of analyzing data visualizations to interpret the underlying data and provide key insights based on the visualizations. To test out this feature, we can simply give a plot and ask for insights.

While it does a good job of understanding the overall context of the plot and the linear trend, it makes errors by mentioning the starting year as 1950, though the data points only start from 1960. The model also derives factors such as population growth and economic development—

while they could be true, those insights cannot be derived from this particular graph alone.

One can ask multiple follow-up questions to refine the initial output from the GPT-4 Vision model. Based on our testing, a human in the loop is still necessary for reviewing the insights, and the model can enhance the productivity of data interpretation use cases.

4. Creative content creation

With the advent of ChatGPT, social media is full of various prompt engineering techniques, and many have found surprising, creative ways to use the generative technology to their advantage.

For this tutorial, we will use the DALL-E-3 (which is also available in ChatGPT Plus) together with GPT-4 Vision, to creatively create a social media post.

Step 1: Ask GPT-4 to create a prompt to generate an image. Let's say you want to create a post contrasting the differences between a data scientist role in a startup vs a corporate one.

Step 2: Use the prompt and generate an image from DALL-E. You can tweak and refine the prompt till you're happy with the output.

Step 3: Use the image and ask GPT-4 Vision to create a post that goes alongside the image.

By tweaking and providing a more detailed prompt, a better output could be obtained, and creative content generation could be explored further. It's worth noting that spamming the internet or social media with AI-generated content isn't recommended, as this content comes with its own limitations. Instead, fact-check and refine it with your own experiences.

Of course, this isn't an exhaustive list of use cases that's possible—GPT-4 Vision is capable of many more. Instead, treat this as an inspiration and a starting point to exploring your curiosity by applying the technology to a domain of your choice.

Limitations and Mitigating Risks of GPT-4 Vision

There's one last thing you need to be aware of before using GPT-4 Vision in use cases—the limitations and risks associated with it.

This is particularly important because OpenAI themselves have taken a few extra months since the launch of GPT-4 in March 2023 to test it with their internal and external "red-teaming" exercise to determine the shortcomings of this generative technology, which they have outlined in the system card.

1. Accuracy and reliability

While the GPT-4 model represents significant progress toward reliability and accuracy, it's not always the case. According to OpenAI, based on the internal tests, the GPT-4 Vision can still be unreliable and inaccurate at times. The team goes as far as even mentioning that "ChatGPT can make mistakes. Verify important information." beneath the chat bar when we input text and images.

Thus, it is of utmost importance that users critically assess the output of the model and remain vigilant.

2. Privacy and bias concerns

According to OpenAI, similar to its predecessors, GPT-4 Vision continues to reinforce social biases and worldviews, including harmful stereotypical and demeaning associations for certain marginalized groups. Thus, it is important to understand this limitation and take other necessary steps to handle the bias within the use case itself and not rely on the model to solve it.

In addition to the bias concerns, the data shared with ChatGPT may be used to train models unless opted out; hence, it's important to be mindful not to be sharing any sensitive or private information with the model. Users also may choose to opt out of sharing data for improving models by going into "Data Controls" under the "Settings & Beta section."

3. Restricted for risky tasks

GPT-4 Vision is unable to answer questions that ask to identify specific individuals in an image. This is an expected "refusal" behavior by design. Furthermore, OpenAI advises refraining from using GPT-4 Vision on high-risk tasks, which include:

- **Scientific proficiency:** The model could miss text or characters, overlook mathematical symbols from the images provided with scientific information, and be unable to recognize spatial locations and color mappings.
- **Medical advice:** The model at times provides correct responses for questions based on medical imaging, but falters at times for the same question. Provided the inconsistencies in responses, the model's answers or outputs should not be relied upon as a replacement for professional medical advice.
- **Disinformation risks:** People are said to believe statements (irrelevant to whether they're true or not) when accompanied by images. The model can be used to generate plausible, realistic, and targeted text content tailored for an image input and thus possesses the disinformation risk.

- **Hateful content:** The model refuses to respond to questions with hate symbols and extremist content in some instances, but that is not always the case. This remains a challenging problem for OpenAI to solve.

Thus, as users, we need to be vigilant in using GPT-4 Vision responsibly, particularly in above mentioned high-risk tasks and sensitive contexts.

Conclusion

This tutorial provided you with a comprehensive introduction to the newly released GPT-4 Vision model. You also were cautioned on the limitations and risks the model poses, and now understand how and when to use the model.

The most practical way to master the new technology is to get your hands on it and experiment by providing various prompts to evaluate its capabilities, and over time, you'd grow more comfortable with it.

While this is a relatively new and one-month-old tool, it's built upon the principles of Large Language Models and GPT-4. Here are some more related resources if you're looking to dive deep into the underlying foundational and related concepts of the GPT-4 Vision model:

- What is GPT-4 and Why Does it Matter?
- A Beginner's Guide to ChatGPT Prompt Engineering
- A Beginner's Guide to GPT-3
- Master Large Language Models (LLMs) Concepts Course

As a senior data scientist, I design, develop and deploy large-scale machine-learning solutions to help businesses make better data-driven decisions. As a data science writer, I share learnings, career advice, and in-depth, hands-on tutorials.