December 2022

Iterative Image Generation via Voice Interaction with an Image Generation Model

Sarvjeet Singh

Bertrand Damiba

Iterative Image Generation via Voice Interaction with an Image Generation Model

ABSTRACT

Image generation models enable users to generate images by providing instructions. However, such models cannot be invoked with voice commands and are also unable to update a prior image based on the user instruction. This disclosure describes techniques that enable users to obtain and refine images by iteratively interacting with an image generation model in real time, e.g., via voice commands to a virtual assistant. Implementation of the techniques can enable users to use their voice and imagination for artistic visual expression. The techniques can be provided via a virtual assistant available via a smart speaker, smartphone, or other device. The techniques incorporate appropriateness checks for the input query and/or the output image, thus ensuring that the interactive experience is safe and trustworthy.

KEYWORDS

- Generative machine learning
- Generative AI
- Image generation
- Image manipulation
- Artistic expression
- Virtual assistant
- Voice command
- Query safety
- Smart speaker
- Smart display

BACKGROUND

Generative machine learning technology has been applied to create machine learning models that users can use to generate images by providing text instructions. For instance, a user can ask for the image of “a person with books in the background” and receive an image depicting a scene that matches the text specification. Imagen [1] is an example of a generative machine learning model that can be used to generate creative, high-fidelity illustrations with such text input. Current implementations of such models take only text input; users cannot provide an existing image along with the text input to seek adjustments to the input image, rather than obtaining a completely new image.

Users increasingly interact with various devices and applications via voice-based virtual assistants. Apart from being embedded within general-purpose devices such as smartphones, such virtual assistants can also be provided via standalone devices such as smart speakers, smart displays, etc. These standalone devices are typically placed in communal areas of a home and can be used by everyone within the household. The composition of many households includes users of diverse abilities and technical skills, such as children and older adults. Users with lower technical knowledge often find it easier and more convenient to interact with applications via voice-based interfaces. In fact, such interfaces are typically the first interactive experiences for children as they can be used even by children who have not yet learned to read and write.

DESCRIPTION

This disclosure describes techniques that enable users to generate images by providing commands to a trained image generation model. Users can interact with the functionality in real time via voice commands, e.g., provided to a virtual assistant on a smart speaker, smartphone, or other device. In addition to generating an image from scratch based on the user command, the model can also function such that users can specify commands that ask to generate modifications to an existing image provided as input along with the command or previously generated by the model. Such a model enables the users to ask for adjustments to the image generated as output of a previous command by including the previously generated image in the input. For instance, after viewing the image generated for the command "a person with books in the background," a user can issue a command to "increase the number of books."

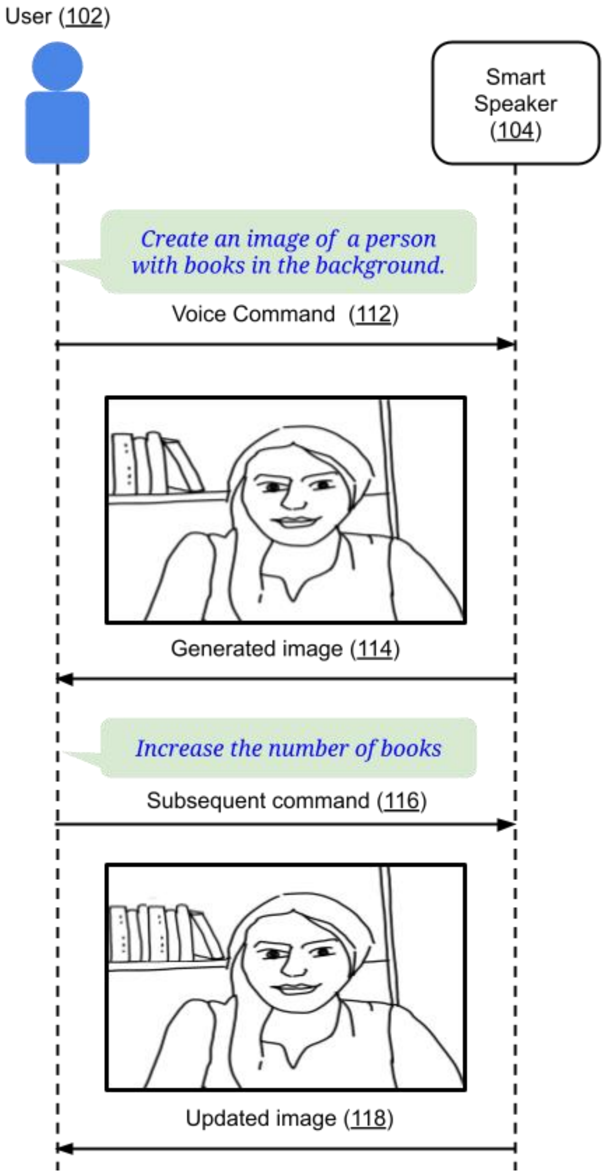

Fig. 1: Obtaining images from an image generation model via voice based virtual assistant

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. A user (102) issues a voice command (112) to a voice-based virtual assistant application provided via a smart speaker (104) or another device. The command is a request to create an image of a person with books in the background. With user permission, the text of the command (e.g., extracted via standard speech-to-text techniques), is provided as input to a machine learning model trained suitably for generating images matching the given text input. The model can be implemented locally on the smart device or on a remote server. User permission is obtained to provide the user query as input to the server-based model. The user is shown a generated image (114) obtained as the output of the model.

The user can accept the image as is, or can choose to provide additional commands to modify the image. In the example of Fig. 1, the user provides a subsequent command (116) to increase the number of books in the image. The subsequent command as well as the previously generated image (114) that the command refers to are provided as inputs to the model. The model then generates an updated image (118) that fulfills the user request. The interaction can be performed any number of times to update the image per further user requests.

The ability to refine a previous image can provide an improved interactive user experience (UX) for creating images using image generation models. Users can specify any relevant adjustments to the image such as adding, removing, or moving objects (e.g., “add a basketball in the middle”), refining image properties (e.g., “make it less dark”), etc. Such an interactive experience that enables users to make iterative enhancements can be much more engaging and effective than having users compose long and detailed text input at the outset without any visual feedback. Apart from images previously generated by the image generation model, the techniques can additionally incorporate functionality that allows users to include as input images obtained via other mechanisms, such as images created by other models, photos taken with a user device, etc.

With user permission, the described techniques can also take into account pertinent user characteristics, such as vocabulary. For instance, when a child issues a voice command, the speech-to-text conversion may not precisely transcribe each utterance and the output can be adjusted based on the vocabulary for children of similar ages. For instance, a command to generate an image containing a barn can imply that it is more likely to contain words such as “cow,” “pigs,” “chicken,” etc. Further, any inappropriate or prohibited content within the voice command can be detected (e.g., using standard language classification models) and filtered out prior to sending the command to the image generation model.

Similarly, the content of the images generated by the image generation model can be adjusted to output scenes that are fitting for the user context. For instance, if a child user asks for an image containing a drawing, the image generation model can be instructed to generate output images that match the properties of drawings of children of a similar age, rather than providing photorealistic or professional quality images. The generated output images can be analyzed with suitable image classifiers to perform appropriateness checks similar to those for the input commands. For instance, image classifiers can be applied to verify that images generated for work contexts do not include content unsuitable for professional settings, to ensure that images generated by children are age appropriate, etc.

Any suitably trained and distilled image generation model that can scale to provide a satisfactory real time user experience for the volume of queries to a virtual assistant (or other interface) can be used. Further, the generated image may be shown on a different device than the device that received the voice command. With user permissions, voice commands can be processed on-device or in the cloud.

The described techniques add to the functionality provided by voice-based virtual assistants (or query response engines in general), thus enhancing their utility and effectiveness. With user permission, the techniques can be implemented within any device, application, or platform that provides voice-based interaction with a virtual assistant or other forms of interaction with a query response engine. Implementation of the techniques can enable users to use their voice and imagination for artistic visual expression created via image generation models. The experience can additionally serve to facilitate learning and enjoyment, especially for specific user populations such as children. Moreover, implementations can incorporate appropriateness checks for the input and the output as described above, thus ensuring that the interactive experience is safe and trustworthy.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user’s voice commands/queries, a user’s images, profession, a user’s preferences, or a user’s current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user’s identity may be treated so that no personally identifiable information can be determined for the user, or a user’s geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques that enable users to obtain and refine images by iteratively interacting with an image generation model in real time, e.g., via voice commands to a virtual assistant. Implementation of the techniques can enable users to use their voice and imagination for artistic visual expression. The techniques can be provided via a virtual assistant available via a smart speaker, smartphone, or other device. The techniques incorporate appropriateness checks for the input query and/or the output image, thus ensuring that the interactive experience is safe and trustworthy.

REFERENCES

1. “Imagen” available online at https://imagen.research.google/