January 2024

Recreating Photos from Text Descriptions Using Generative Artificial Intelligence

n/a

Recreating Photos from Text Descriptions Using Generative Artificial Intelligence

ABSTRACT

When on trips, users often take near duplicate photos at the same place. Multiple photos with little variation among them can consume a large amount of the limited storage space on the user device and/or cloud storage services where users may choose to store their photos. Retaining only a few of the photos can reduce storage use but makes the photos permanently unavailable. This disclosure describes techniques that enable users to recreate close approximations of discarded photos from corresponding text descriptions. A suitable image-to-text conversion model is applied to photos prior to discarding the photos and the generated text description is stored. The text description, context, and any related photos are used to prompt a suitable image generation Artificial Intelligence (AI) model to recreate the photo. The techniques can additionally support generating variations of photos that did not originally exist. The accuracy, faithfulness, and fidelity of the generated photo can be improved with prompts that contain more detail as well as context from other related photos.

KEYWORDS

- Image generation
- Image synthesis
- Image compression
- Image-to-text
- Image embedding
- Generative artificial intelligence (AI)
- Photo context
- Photo variation

BACKGROUND

When on trips, users often take near duplicate photos at the same place. For instance, users may take multiple successive shots of a scene to ensure that at least one of these meets their preferences. Multiple photos with little variation among them can consume a large amount of the limited storage space on the user device and/or cloud storage services where users may choose to store their photos.

Automated techniques are available that can compare and cluster similar photos, and/or to identify the likely best shot among the cluster that can be retained while deleting the others to free up space. However, such a destructive approach that results in permanent deletion of most of the photos in the cluster means that these photos are unavailable should the user need them in the future.

DESCRIPTION

This disclosure describes techniques that obtain text descriptions from photos to be discarded by employing a suitable image-to-text conversion model. Subsequently, a discarded photo can be recreated based on the text description using an image generation artificial intelligence (AI) model, guided by a prompt. The models can continually improve over time to derive more appropriate and detailed text descriptions, and to generate photos that recreate the discarded photos with greater fidelity to the original photo.

Users can improve the accuracy, faithfulness, and fidelity of the generated photo with prompts that include more detail as well as context from other photos in the user's photo library that are close in time and/or location to the discarded photo that is to be recreated. Such prompts can provide relevant contextual information, such as appearance, clothing, weather, lighting, etc. For instance, a user can add to the prompt information such as "In the photo to be generated, I

was wearing the same outfit as in the first photo and wore my hair in the same way as in the second photo," "I looked the same in the photo to be generated as in the included photo that was taken a few minutes before," etc. With user permission, the process can be further improved by including other photos, e.g., photos shared by the user's friends and family, that are close in time and/or location to the photo to be generated.

In addition, users can expand the prompt by including relevant spatiotemporal, photo-related, and camera-related metadata, such as "The photo was taken on July 31, 2023, at 8:40pm just after sunset with the camera held 5 feet above ground by a person who is 6 feet tall. The camera was titled 34 degrees toward the horizontal with zero yaw and roll. The focal length of the lens was 18mm with a f-stop of 1/16 and shutter speed of 1/30th second." Such information can automatically be added to the prompt based on a photo taken close in time to the photo to be recreated. For example, the text "July 31, 2023, at 8:40 pm" provided by the user can be used to identify a photo that includes metadata that specifies (or that can be used to derive) the rest of the content of the prompt.

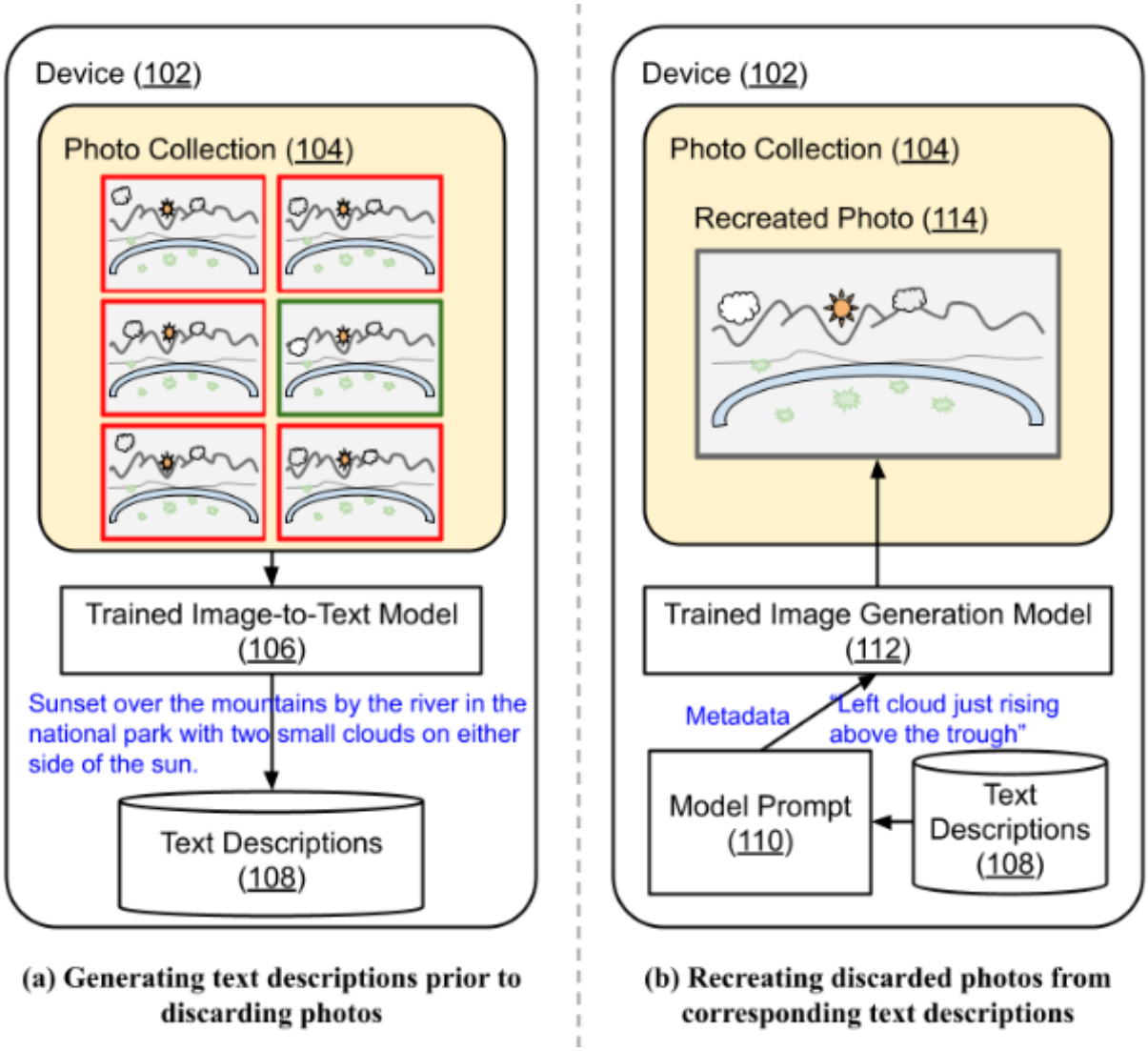

Fig. 1: Recreating discarded photos based on text descriptions saved at an earlier time

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. A user's photo collection (104) on a device (102) contains multiple similar (near duplicate) photographs of a nature scene. Prior to retaining only one of the similar photos by discarding the others, text descriptions (108) for each photo are generated via a trained image-to-text model (106). The generated text descriptions are stored in a local database. As shown in Fig. 1(a), the process results in the scene being described as "Sunset over the mountains by the river in the national park with two small clouds on either side of the sun." Individual descriptions may

specify additional attributes, e.g., position of the clouds with respect to the mountain peaks, with respect to each other, position of the sun, etc. After generating and storing the text descriptions, the user can discard the corresponding photos. In the example of Fig. 1(a), photos with a red border are discarded, while the photo with a green border is retained in storage. As seen in Fig. 1(a), the retained photo has a cloud on the left side that is just beneath the mountain range in the back, while other photos have that cloud above the mountain range.

At a later time, as shown in Fig. 1(b), the user retrieves the text description and provides it to a trained image generation model (112) along with a prompt (110) that additionally includes relevant metadata and context for the photo. For example, the user can specify a difference between a retained photo and the desired recreation, e.g., “left cloud just rising above the trough.” Additionally, the prompt can also include one or more related saved photos, e.g., the photo with the green border in Fig. 1(a), if such a photo is available.

The detailed text prompt and context results in generating a matching recreation of the original photo (114) being generated by the image generation model. As can be seen, the generated photo is very close to the bottom-right photo in Fig. 1(a) that was discarded. Any suitable contextual information can be included in the prompt. As appropriate, users can choose to generate text descriptions for the photos to be discarded for an entire photo and/or specific regions of a photo (e.g., foreground, background, etc.) and/or a collection of photos and/or for differences between a pair of photos.

The techniques described herein can additionally support generating variations of photos that were not originally captured. For instance, uses can generate photos of a hypothetical visit to a place (e.g., a museum) that the user could not visit during a trip because of a lack of time, add a person who missed the trip to the photo to generate a “you were missed” variation, alter people’s

appearance (e.g., expressions, poses, attire, hair styles, etc.). Similarly, users can generate variations of existing photos by employing suitable models (e.g., [1], [2]) to generate the scene from alternative angles with information about the surroundings retrieved from photos close in time and location. For instance, a user can choose to generate a photo of a scene as seen from behind a landmark instead of the view captured in the original photo while standing at an angle.

Since language and text descriptions of photos are relatively stable over long periods, the resolution, quality, and form of the generated photos can be improved with continued technological improvements in storage capacity, network bandwidth, hardware capabilities and speed, and memory resources. For instance, the text description of a low resolution photo coupled with relevant contextual metadata (e.g., people, time, place, weather, etc.) and example photos close in time and place can be used to generate higher resolution photos of the scene with such capabilities improving over time. Similarly, technological improvements can make it possible to generate extremely high resolution photos with three-dimensional (3D) views. In addition, it may be possible to add interactive capabilities to the generated photos and/or turn them into short video segments.

The recreation/ generation of photos can be performed via any suitable generative AI models, such as variational autoencoder (VAE), generative adversarial network (GAN), diffusion, etc. The generating model architecture can be a deep neural network, such as a visual transformer. The model can be trained using a training dataset (obtained with appropriate permissions) that includes a large number of images with attached text, such as captioned images on webpages, photos on social media with accompanying comments, photos linked with corresponding alt text, public datasets containing photos with respective descriptions, collections of manually labeled images, etc.

During training, photos and corresponding text can be used together in the same latent embedding space such that the text description fed into the photo can create an embedding compatible with an image generation model. Trained with image and text tokens appearing in the same latent embedding or token space can make the neural network invariant to whether the input is an image or text.

The embedded text description can be input to a suitable image generator model for deconvolution. The generator model can be provided with additional information, such as one or more reference images related to the text. The information can be concatenated into a separate single larger embedding used as context. Alternatively, the information can be used as smaller embeddings (e.g., one or more tokens) provided to the network/transformer similar to how text is input to a recurrent model such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) where self-attention is applied to the text as well as context/image tokens referenced by the text tokens. For instance, the text prompt “use the first photo as reference for the background,” references a token connected to a specific photo.

As appropriate and with specific user permission, the techniques can additionally incorporate embeddings from existing models, such as face embeddings from a face recognition model (e.g., [3]), person embeddings from a person comparison/recognition model (e.g., [4]), etc. Voice embeddings, sample waveforms, and action- or video-based embeddings can be used for video. Video-based embeddings can consist of one or more images sampled over time and combined into a single embedding.

The generation process can be further improved by including relevant depth and stereo information, such as: depth inferred from a photo; real stereo 3D depth obtained from multiple cameras; depth obtained from an autofocus or LIght Detection And Ranging (LIDAR) sensors;

lighting information from light sensors; sound levels or audio data; etc. If users permit, such information can be obtained from the corresponding sensors included in mobile devices, such smartphones.

After a sufficiently large generic model is trained using a sufficiently large dataset, it can be improved with transfer learning. Further, the model can be fine-tuned on a smaller or more specific dataset of photos related to a specific task. For instance, fine-tuning can be achieved using any suitable generic techniques such as over-training an existing trained model by unfreezing one or more layers in the network and providing more training data specific to a user and task. Alternatively, or in addition, the fine-tuning process can use distillation by providing the same photos to a smart teacher and a blank student model. The outputs of the two models are compared using cosine similarity or a similar metric. The results of the comparison are used to update the weights for the student model until it converges to the same weights/performance as the teacher model.

Fine-tuning the model can enable users to have a smaller personalized generative AI model fine-tuned for the user's photo collection. The fine-tuning can be performed locally on the user device. Alternative, if users permit, the fine-tuning can take place externally in the cloud, with the use of privacy-preserving techniques such as differential privacy and/or federated learning (where the data never leaves the user device and only ML training gradients and optionally lossy tokens of image and text are sent to the cloud in order to fine-tune or personalize the model performance). Model personalization can make the model better suited for generating photos of specific users and their specific contexts as well as photos of people and places, in general.

The process of generating text descriptions of the photos to be discarded can be the reverse of the image generation process described above. The trained image generation model can be modified to learn to generate text descriptions of photos using the same or similar embeddings. The first half of the model can be retained as a visual transformer for tokenizing the input images and any relevant metadata. The second half of the model can be composed by replacing the VAE, GAN or Stable Diffusion component with a transformer neural network capable of generating text. With such an architecture, the first half of the model can be used for compression (deriving text descriptions of photos) as well as decompression (generating photos from text descriptions). As a result, the techniques can operate with three models: image-to-text generator, text-to-image generator, and a common text and image embedder that can provide information to either generator depending on whether compression or decompression is being performed.

The techniques can be implemented within any application, service, or platform that involves the use of photos, such as photo storage and organization applications, photo editors, image search engines, digital maps that include photographic views, etc. The techniques can be implemented to run locally on the user device which can achieve substantial savings in the computational resources and bandwidth needed on the cloud or server side. In addition, the local operation can be performed when the device is offline. The ability to generate the photos without requiring a network connection can help create an enjoyable user experience (UX) for exploring original photos and their hypothetical and realistic variations before, during, or after a trip.

Implementation of the techniques can help users save device and cloud storage space for photos while retaining a way to recreate (close approximations of) the discarded photos in the future. For further storage savings, the text descriptions of the discarded photos can be

compressed using any commonly used compression techniques, such as Huffman coding. In addition to recreating discarded photos and variations thereof, the techniques can enable users to leverage advances in technology to generate photos at higher resolutions and fidelity than that of the originally captured photos.

Users are provided with controls allowing them to select if and when systems, programs or features described herein may perform collection of user information (e.g., information about a user's photo library, a user's current or prior locations, a user's preferences, text descriptions of a user's photos, user-provided prompts). The user can also control whether and how a server may be used to implement features described herein. 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.

CONCLUSION

When on trips, users often take near duplicate photos at the same place. Multiple photos with little variation among them can consume a large amount of the limited storage space on the user device and/or cloud storage services where users may choose to store their photos. Retaining only a few of the photos can reduce storage use but makes the photos permanently unavailable. This disclosure describes techniques that enable users to recreate close approximations of discarded photos from corresponding text descriptions. A suitable image-to-text conversion model is applied to photos prior to discarding the photos and the generated text description is stored. The text description, context, and any related photos are used to prompt a

suitable image generation Artificial Intelligence (AI) model to recreate the photo. The techniques can additionally support generating variations of photos that did not originally exist. The accuracy, faithfulness, and fidelity of the generated photo can be improved with prompts that contain more detail as well as context from other related photos.

REFERENCES

1. Wang, Zirui, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. "NeRF--: Neural radiance fields without known camera parameters." *arXiv preprint arXiv:2102.07064* (2021).
2. Xu, Hongyi, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. "GHUM & GHUML: Generative 3d human shape and articulated pose models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6184-6193. 2020.
3. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815-823. 2015.
4. Wu, Lin, Chunhua Shen, and Anton van den Hengel. "Personnet: Person re-identification with deep convolutional neural networks." *arXiv preprint arXiv:1601.07255* (2016).