Title: Text-to-image personalization

URL: https://en.wikipedia.org/wiki/Text-to-image_personalization

PageID: 74705461

Categories: Category:Computer graphics, Category:Deep learning, Category:Text-to-image generation

---

Text-to-Image personalization is a task in deep learning for computer graphics that augments pre-trained text-to-image generative models . In this task, a generative model that was trained on large-scale data (usually a foundation model), is adapted such that it can generate images of novel, user-provided concepts. [1] [2] These concepts are typically unseen during training, and may represent specific objects (such as the user's pet) or more abstract categories (new artistic style [3] or object relations [4]).

Text-to-Image personalization methods typically bind the novel (personal) concept to new words in the vocabulary of the model. These words can then be used in future prompts to invoke the concept for subject-driven generation, [5] inpainting , style transfer [6] and even to correct biases in the model. To do so, models either optimize word-embeddings , fine-tune the generative model itself, or employ a mixture of both approaches.

Technology

Text-to-Image personalization was first proposed during August 2022 by two concurrent works, Textual Inversion [7] and DreamBooth . [8]

In both cases, a user provides a few images (typically 3–5) of a concept, like their own dog, together with a coarse descriptor of the concept class (like the word "dog"). The model then learns to represent the subject through a reconstruction based objective, where prompts referring to the subject are expected to reconstruct images from the training set.

In Textual Inversion, the personalized concepts are introduced into the text-to-image model by adding new words to the vocabulary of the model. Typical text-to-image models represent words (and sometimes parts-of-words) as tokens, or indices in a predefined dictionary. During generation, an input prompt is converted into such tokens, each of which is converted into a ‘word-embedding’: a continuous vector representation which is learned for each token as part of the model's training. Textual Inversion proposes to optimize a new word-embedding vector for representing the novel concept. This new embedding vector can then be assigned to a user-chosen string, and invoked whenever the user's prompt contains this string. [7]

In DreamBooth , rather than optimizing a new word vector, the full generative model itself is fine-tuned. The user first selects an existing token, typically one which rarely appears in prompts. The subject itself is then represented by a string containing this token, followed by a coarse descriptor of the subject's class. A prompt describing the subject will then take the form: "A photo of " (e.g. "a photo of sks cat" when learning to represent a specific cat). The text-to-image model is then tuned so that prompts of this form will generate images of the subject. [8]

Textual Inversion

The key idea in Textual Inversion is to add a new term to the vocabulary of the diffusion model that corresponds to the new (personalized) concept.

Textual Inversion operates by inverting the concepts into new pseudo-words within the textual embedding space of a pre-trained text-to-image model. These pseudo-words can be injected into new scenes using simple natural language descriptions, allowing for simple and intuitive modifications. The method allows a user to leverage multi-modal information — using a text-driven interface for ease of editing, but providing visual cues when approaching the limits of natural language.

The resulting model is extremely light-weight per concept: only 1K long, but succeeds to encode detailed visual properties of the concept. [9]

Extensions

Several approaches were proposed to refine and improve over the original methods. These include the following.

Low-rank Adaptation (LoRA) - an adapter-based technique for efficient finetuning of models. [10] In the case of text-to-image models, LoRA is typically used to modify the cross-attention layers of a diffusion model. [11]

Perfusion - a low rank update method that also locks the activations of the key matrix in the diffusion model's cross attention layers to the concept's coarse class. [12]

Extended Textual Inversion - a technique that learns an individual word embedding for each layer in the diffusion model's denoising network. [13]

Encoder-based methods that use another neural network to quickly personalize a model [14] [15]

Challenges and limitations

Text-to-image personalization methods must contend with several challenges. At their core is the goal of achieving high-fidelity to the personal concept while maintaining high alignment between novel prompts containing the subject, and the generated images (typically referred to as ‘editability’).

Another challenge that personalization methods must contend with is memory requirements. Initial implementations of personalization methods required more than 20 Gigabytes of GPU memory, and more recent approaches have reported requirements of more than 40 Gigabytes. [14] However, optimizations such as Flash Attention [16] have since reduced this requirement considerably.

Approaches that tune the entire generative model may also create checkpoints that are several gigabytes in size, making it difficult to share or store many models. Embedding based approaches require only a few kilobytes, but typically struggle to preserve identity while maintaining editability. More recent approaches have proposed hybrid tuning goals which optimize both an embedding and a subset of network weights. These can reduce storage requirements to as little as 100 Kilobytes while achieving quality comparable to full tuning methods. [12]

Finally, optimization processes can be lengthy, requiring several minutes of tuning for each novel concept. Encoder and quick-tuning methods aim to reduce this to seconds or less. [17]

References