Claims

1. A method of identifying an influential training image in a diffusion model for image generation, comprising the steps of:

training the diffusion model using a set of training images;

generating one or more output images using the trained diffusion model;

computing a gradient matrix for each image in the set of training images;

calculating an importance score for the each image in the set of training images based on the computed gradient matrix, wherein the importance score indicates an importance of the set of training images to the one or more output image; and

identifying one or more of the training images as the influential training image when the importance score exceeds a predetermined threshold,

wherein the importance score is τ_{D-TRAK} and is calculated based on the equation of

$$\tau_{D-TRAK}(x,\mathcal{D}) = \left[\frac{1}{\mathrm{S}}\sum_{S=1}^{S}\phi^{s}(x)^{\top}\left(\Phi^{\mathrm{s}}_{\mathrm{D-TRAK}}{}^{\top}\Phi^{\mathrm{s}}_{\mathrm{D-TRAK}}\right)^{-1}\Phi^{\mathrm{s}}_{\mathrm{D-TRAK}}{}^{\top}\right]$$

wherein the score $\tau_{\mathrm{D-TRAK}}$ represents how a training sample $\mathrm{x}^{\mathrm{n}} \in \mathcal{D}$ affects the diffusion model for the image generation, a total of S subsets are initially sampled from the training dataset $\mathcal{D}$, wherein $\Phi^{\mathrm{s}}_{\mathrm{D-TRAK}} = [\phi^{s}(\mathrm{x}^{1});\cdots;\phi^{s}(\mathrm{x}^{\mathrm{N}})]^{\top}$, which is the gradient matrix.

2. The method of claim 1, wherein $\phi^{s}(x) = \nabla_{\theta}\mathcal{L}(x,\theta_{s}^{*})$, and function of $\mathcal{L}$ comprises any one of $\mathcal{L}_{Square}, \mathcal{L}_{Avg}$ *or* $\mathcal{L}_{p-norm}$, wherein:

$\mathcal{L}_{square}(x,\theta) = E_{t,\epsilon,}[\,||\epsilon_{\theta}(x_t,t)||_2^2]$;

$\mathcal{L}_{Avg}(x,\theta) = E_{t,\epsilon,}[\,Avg\,(\epsilon_{\theta}(x_t,t))]$;

$\mathcal{L}_{p-norm}(x,\theta) = E_{t,\epsilon,}\left[\,||\epsilon_{\theta}(x_t,t)||_p\right]$;

wherein $x_t = \sqrt{\overline{\alpha}_t}x + \sqrt{1-\overline{\alpha}_t},\epsilon$ and Avg is an average pooling operation.

3. The method of claims 1 or 2, wherein the gradient matrix computation is performed at different training data checkpoints.
4. The method of any of claims 1 to 3, further comprising identifying training data checkpoints that contributed most to the generated one or more output images.
5. The method of any of claims 1 to 4, wherein the gradient matrix computation is performed at a final checkpoint.
6. The method of any of claims 1 to 5, wherein the diffusion model is a Denoising Diffusion Probabilistic Model (DDPM) or a Latent Diffusion Model (LDM).
7. The method of any of claims 1 to 6, further comprising ranking the set of training images based on their importance scores to identify most important images.
8. The method of claim 7, further comprising a verification step, which comprises:
   retraining the diffusion model after removing the most important images identified in the ranking process;
   generating images using the retrained model; and
   measuring a pixel wise ℓ_2 distance or Contrastive Language-Image Pre-training (CLIP) cosine similarity between images generated by the original and retrained models to confirm an influence of the removed most important images.
9. The method of any of claims 1 to 8, wherein positive importance scores identify proponent images, and negative scores identify opponent images among the set of training images.
10. The method of any of claims 1-9, further comprising displaying the identified influential training image and/or its corresponding importance score.
11. The method of any of claims 1-10, further comprising
    removing training images with low importance score, where the low importance score is below the predetermined threshold value;
    finetuning or retraining the diffusion model after removing the training images with low importance score.
12. The method of claim 11, further comprising
    regenerating the one or more output image using the retrained diffusion model.
13. A system for identifying an influential training image in a diffusion model for image generation, comprising:

a processor;

a memory in electronic communication with the processor; and

instructions stored in the memory and executable by the processor to cause the system to:

train the diffusion model using a set of training images;

generate one or more output images using the trained diffusion model;

compute a gradient matrix for each image in the set of training images;

calculate an importance score for the each image in the set of training images based on the computed gradient matrix, wherein the importance score indicates an importance of the set of training images to the one or more output image; and

identify one or more of the training images as the influential training image when the importance score exceeds a predetermined threshold,

wherein the importance score is τ_{D-TRAK} and is calculated based on the equation of

$$\tau_{D-TRAK}(x, \mathcal{D}) = \left[\frac{1}{S}\sum_{S=1}^{S} \phi^{s}(x)^{\top}\left(\Phi^{s}_{\mathrm{D-TRAK}}{}^{\top}\Phi^{s}_{\mathrm{D-TRAK}}\right)^{-1}\Phi^{s}_{\mathrm{D-TRAK}}{}^{\top}\right]$$

wherein the score $\tau_{\mathrm{D-TRAK}}$ represents how a training sample $\mathrm{x}^{\mathrm{n}} \in \mathcal{D}$ affects the diffusion model for the image generation, a total of S subsets are initially sampled from the training dataset $\mathcal{D}$, wherein $\Phi^{s}_{\mathrm{D-TRAK}} = [\phi^{s}(\mathrm{x}^{1}); \cdots; \phi^{s}(\mathrm{x}^{\mathrm{N}})]^{\top}$, which is the gradient matrix.

14. The system of claim 13, wherein $\phi^{s}(x) = \nabla_{\theta}\mathcal{L}(x, \theta_{s}^{*})$, and function of $\mathcal{L}$ comprises any of $\mathcal{L}_{Square}, \mathcal{L}_{Avg}$ *or* $\mathcal{L}_{p-norm}$, wherein:

$\mathcal{L}_{square}(x, \theta) = E_{t,\epsilon,}[\,||\epsilon_{\theta}(x_t, t)||_2^2]$;

$\mathcal{L}_{Avg}(x, \theta) = E_{t,\epsilon,}[\,Avg\,(\epsilon_{\theta}(x_t, t))]$;

$\mathcal{L}_{p-norm}(x, \theta) = E_{t,\epsilon,}\big[\,||\epsilon_{\theta}(x_t, t)||_p\big]$;

wherein $x_t = \sqrt{\overline{\alpha}_t}x + \sqrt{1-\overline{\alpha}_t}, \epsilon$ and Avg is an average pooling operation.

15. The system of claims 13 or 14, wherein the gradient matrix computation is performed at a final checkpoint.

16. The system of any of claims 13-15, further comprising displaying the identified influential training image and/or its corresponding importance score.

17. An apparatus for identifying an influential training image in a diffusion model for image generation, comprising:

means for training the diffusion model using a set of training images;

means for generating one or more output images using the trained diffusion model;

means for computing a gradient matrix for each image in the set of training images;

means for calculating an importance score for the each image in the set of training images based on the computed gradient matrix, wherein the importance score indicates an importance of the set of training images to the one or more output image; and

means for identifying one or more of the training images as the influential training image when the importance score exceeds a predetermined threshold,

wherein the importance score is τ_{D-TRAK} and is calculated based on the equation of

$$\tau_{D-TRAK}(x, \mathcal{D}) = \left[\frac{1}{S}\sum_{S=1}^{S} \phi^{s}(x)^{\top}\left(\Phi^{s}_{\mathrm{D-TRAK}}{}^{\top}\Phi^{s}_{\mathrm{D-TRAK}}\right)^{-1}\Phi^{s}_{\mathrm{D-TRAK}}{}^{\top}\right]$$

wherein the score $\tau_{\mathrm{D-TRAK}}$ represents how a training sample $\mathrm{x}^{\mathrm{n}} \in \mathcal{D}$ affects the diffusion model for the image generation, a total of S subsets are initially sampled from the training dataset $\mathcal{D}$, wherein $\Phi^{\mathrm{s}}_{\mathrm{D-TRAK}} = [\phi^{s}(\mathrm{x}^{1}); \cdots; \phi^{s}(\mathrm{x}^{\mathrm{N}})]^{\top}$, which is the gradient matrix.

18. A computer-readable storage medium, on which a computer program is stored, wherein the computer program, when executed in a computer, causes the computer to perform the method of any of claims 1-12.