CameraForensics

ESGI Proposal - Detecting inappropriate material used to train AI image generation models

Open-source generative AI models, especially diffusion models such as Stable Diffusion and Flux, can be fine-tuned to specialise in generating images aligned with specific visual styles, categories, or subjects. This fine-tuning is often achieved using Low-Rank Adaptation (LoRA), a lightweight method that allows users to efficiently train a small number of additional parameters without altering the full base model.

While this capability offers a wide range of creative and technical applications, it also introduces challenges in model governance and safety. Once a model has been fine-tuned—especially when distributed or shared—it can be difficult to determine what kind of data or concepts it was trained on, particularly when the training data and process are not disclosed.

This problem has broad relevance for model transparency, safety, intellectual property protection, and explainability and raises an important open research question:

Given a fine-tuned generative model, can we develop methods to infer which concepts or kinds of images were likely dominant in its training data?

The core aim of this challenge is to develop **concept attribution** and **auditing techniques** using **image-free approaches**, such as the analysis of latent representations, embedding shifts, or parameter changes.

The following points should be considered:

- All experiments will be conducted using benign, publicly available fine-tuned models and concepts (e.g., food, vehicles, animals, textures, landscapes).
- No models, data, or prompts related to commercially sensitive or inappropriate content will be used or are to be inferred.
- The focus of the research is strictly on technical feasibility and conceptual auditing, not on adversarial or detection applications.
- It is important to understand the distinction between the *capability* of a model and the *intent* of a model, for example, many AI models may be capable of generating inappropriate material if prompted in a certain way, but this does not mean it was necessarily trained on inappropriate material or optimised to create that kind of material in any way.
- The method should not rely on having any information on the training process of the fine-tuned model or how the model is run (e.g. intended generation settings such as samplers/schedulers or intended key words in the prompt).

From a technical perspective, this problem could be approached by adapting or combining techniques such as Concept Attribution in Diffusion models (CAD)[1][2], embedding and latent -

space analysis[3][4], and image-free auditing[5], or any other relevant strategies. Potential goals for the week can include: to understand how a concept that is dominant in training data is represented in a finetuned model's embedding space, to analyse how finetuning a model on a specific concept changes its internal representations, to investigate internal behaviours of the finetuned model to assess the presence of certain concepts. These research directions all contribute towards the overall goal of understanding whether a model was finetuned on a specific concept.

[1] https://arxiv.org/pdf/2412.02542

[2] **CameraForensics Ltd**, 1 Victoria Street, Bristol BS1 6AA, U.K.
Registered in England and Wales 07734891 | VAT Number: 209170036 | DUNS: 217402193

[3] https://arxiv.org/pdf/2501.18877

[4] https://arxiv.org/pdf/2502.02225

[5] https://arxiv.org/pdf/2504.14815