

Abstract

Method and System for Identifying Influential Training Images in Diffusion Models Using Gradient-Based Attribution

The present invention relates to a method and system for identifying influential training images in diffusion models by calculating importance scores through gradient-based attribution. The system uses a novel Diffusion-Tracing with the Randomly projected After Kernel (D-TRAK) method for identifying and scoring the influence of individual training data points on the outputs of diffusion models, thereby enabling the accurate and interpretable attribution of data in generative models. This approach allows for the identification of training images that have a significant positive or negative influence on a specific generated output image. By focusing on the final checkpoint data, the system reduces computational costs while providing accurate attribution of image generation results. This invention has applications in copyright protection, and model transparency, particularly in identifying the contribution of specific training data to generated outputs.

Fig. 3