Where did you learn that?: Tracing the Impact of Diffusion Model Training Data with Encoded Ensembles

Zheng Dai
CSAIL
MIT
Cambridge, MA 02139
zhengdai@mit.edu

Rui-Jie Yew
Computer Science, CNTR
Brown University
Providence, RI 02912
rui-jie_yew@brown.edu

David Gifford
CSAIL
MIT
Cambridge, MA 02139
gifford@mit.edu

Abstract

The widespread adoption of diffusion models for creative uses such as image, video, and audio synthesis has raised serious legal and ethical concerns surrounding the use of training data and its regulation. Due to the size and complexity of these models, the effect of training data is difficult to characterize with existing methods, confounding regulatory efforts. In this work we propose a novel approach to trace the impact of training data using an encoded ensemble of diffusion models. In our approach, individual models in an ensemble are trained on encoded subsets of the overall training data to permit the identification of important training samples. The resulting ensemble allows us to efficiently remove the impact of any training sample. We demonstrate the viability of these ensembles for assessing influence and consider the regulatory implications of this work.

1 Introduction

Diffusion models have emerged as powerful tools for modeling and sampling from complex natural distributions. These models have garnered significant attention and achieved remarkable results in a wide array of applications. The widespread adoption of diffusion models for creative uses, such as image, video, and audio synthesis, has raised ethical and legal attention surrounding the use of works of authorship as part of training data.

In this work, we propose the use of *encoded ensembles* of diffusion models to trace the contribution of training data. We do this by training the individual members of the ensemble on carefully engineered splits that encode training data in a way that allows us remove the influence of training points by removing ensemble members, which in turn allows us to simulate the effects of leave-one-out retraining.

Contrary to previous approaches, our approach is able to compute ablation based counterfactuals without the need for retraining or approximate methods, both of which limit the applicability of leave-one-out-training approach in generative models, which are too large to efficiently retrain and too complex to approximate [Ho et al., 2020, Rombach et al., 2022]. In addition, we also provide a method for approximate influence calculation, which makes database-scale influence calculations tractable for diffusion models.

1.1 Related Work

Diffusion models were originally introduced to machine learning by Sohl-Dickstein et al. [2015], and were improved by Ho et al. [2020], whose models are the foundation for our models.

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

Leave-one-out retraining is an established paradigm for understanding training data influence, where we analyze a counterfactual scenario where a model is retrained without a given piece of training data. Past approaches have considered influence as either a function of the final resulting model [Koh and Liang, 2017], or the entire training trajectory [Pruthi et al., 2020]. Our method takes the first approach of considering a final model produced in a counterfactual setting.

Fully computing the counterfactual scenario is computationally infeasible, therefore its approximation is an active area of research [Hammoudeh and Lowd, 2022]. However, existing approximation techniques do not scale well to large and complex models [Basu et al., 2020]. Furthermore, many existing approaches are adapted for supervised learning, with relatively fewer works that attempt to analyze influence for unsupervised generative models [Terashita et al., 2021]. Our approach by contrast applies to large and complex generative diffusion models.

2 Contribution

We provide a novel approach to quantify the influence of training data to the output of diffusion models. Furthermore, we propose a methodology that can scale efficiently to training sets with sizes commensurate with those used to train modern generative models [Schuhmann et al., 2022, Ramesh et al., 2022].

2.1 Regulatory Motivation

The framework that we present in this paper traces the causal effect of training images to the resulting model output. This is relevant for several regulatory goals surrounding data provenance and attribution–such as for explainability, privacy, and intellectual property. Our attribution technique could also aid in the discovery of candidate works. Model developers and creative workers alike may be unaware of the contribution of individual copyrighted works to a given output image. Our approach to attribution could aid in the discovery of candidate works that contribute to a given model output. Importantly, this work presents a start at developing technical infrastructure for authors of works used in model training to participate in the market for their work Menell [2012].

3 Results

Rather than tracing influence in a single model, we will instead trace influence through an ensemble of models. One novel finding of our work is that diffusion models can be combined in an ensemble, with each diffusion ensemble member trained on the same or distinct subsets of training data. Each ensemble member is trained as described in Ho et al. [2020]. We sample from the ensemble by averaging the outputs of the models at each denoising step and treating the average as through it were the output of a single model.

To ensure that we can trace influence, the ensemble must be what we refer to as an *encoded ensemble*. We define an encoded ensemble as an ensemble of models with the following property:

Definition 1. *Given a point x in the overall training set, define $S(x)$ as the set of models in the ensemble whose training split included x. Then the ensemble is an encoded ensemble if the set difference $S(x) \setminus S(x')$ is not the empty set of any x, x' in the overall training set.*

We can then remove the influence of any training point by removing all members of the ensemble that was trained on it. If the ensemble is indeed an encoded ensemble, then there exists no other data point whose influence is fully erased. This then allows us to compute counterfactuals by rerunning a diffusion process with the same exogenous noise but on the ablated ensemble. This process is illustrated in Figure 1a. We further demonstrate the validity of this approach by showing that images generated by the counterfactual model behave as expected in the simplified setting where we remove the influence of entire classes, which is provided in Figure 1b and c.

An ensemble that satisfies Definition 1 can be constructed in the following way: assign each datapoint x in the training set a bit vector of length n of some fixed Hamming weight. Include that datapoint in the training split of the ith model if and only if the bit vector is 1 at the ith position. This construction results in an ensemble of models that satisfies Definition 1.

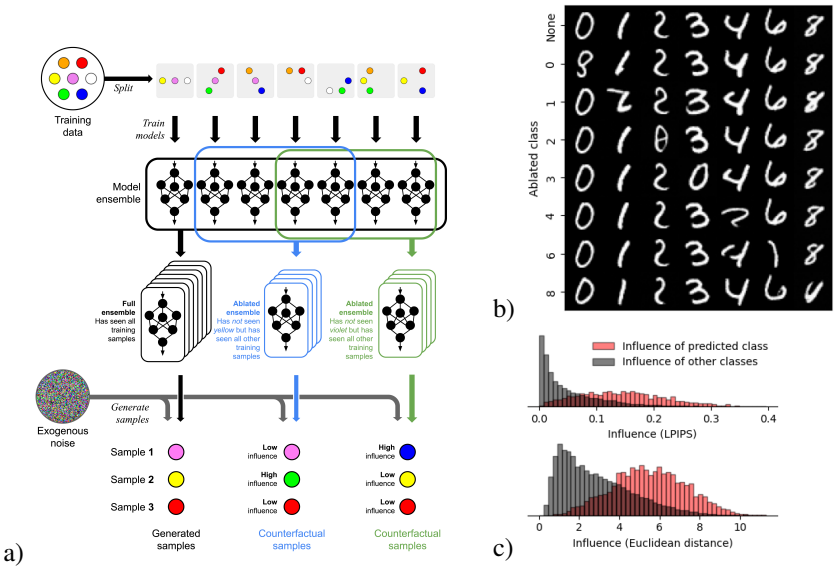

Figure 1: Overview of our method and its viability in producing counterfactuals
a) We split our data such that, for each piece of training data, there exists some subset of the trained ensembles that has, collectively, seen every other piece of training data except the one in question. This allows us to efficiently simulate retraining by ablating the ensemble. These ablated ensembles can then be used to generate counterfactual samples, which can then be used to assess influence.
b) As control, we consider all images of a given MNIST LeCun et al. [2010] class to be a single training sample. Note that we also remove images that belong to classes 5, 7, and 9. We then ablate classes as described in a). The top row corresponds to generated samples, while the lower rows correspond to counterfactuals of the top row with certain classes ablated. Visually inspecting these cherry picked examples shows that ablating a class has more significant consequences on generated images that are of that class, compared to ablating a different class which in most cases leaves the generated sample mostly unchanged.
c) Since examples in b) are cherry picked, we conduct a more quantitative evaluation, where 3000 generated samples are classified. Influence of a class is measured by ablating that class and measuring the distance of the counterfactual to the original, measured using both LPIPS Zhang et al. [2018] (top) and Euclidean distance (bottom). In both cases the predicted class has significantly more influence on generated images than other classes (the Mann-Whitney U test gives $p \leq 10^{-300}$ for both pairs of distributions).

3.1 Database Scale Influence Tracing

While an encoded ensemble allows us to trivially produce a counterfactual model, we still need to run a full diffusion run to produce a counterfactual sample. This can be very expensive. To mitigate the costs involved, we estimate the effects of ensemble ablation by differentially downweighting the removed models and upweighting the unremoved models, and then linearly extrapolate to the point where the weights are zero for the removed models and sum to one for the remaining models. We call the process of removing training data influence by weighting ensemble members *synthetic ablation*. Synthetic ablation can be accomplished via a Jacobian-vector product, and computing the Jacobian requires n diffusion runs with forward mode automatic differentiation, where n is the size of the ensemble. Once computed, the Jacobian can be reused for the entire training set, enabling database-scale evaluation. Results are presented in Figure 2. We can see that when the training set is small, attribution is mostly quite obvious since the model directly copies the training set. For larger training sets, the models become more creative (note how in Figure 2b some of the digits are not quite digits). This can make the attributions far less intuitive. We note that this is consistent with prior findings that show training set copying becomes less prevalent at larger training set sizes [Somepalli et al., 2022].

Figure 2: Methods and experiments
Given a set of generated images, we use our method of differentially downweighting ensemble members to compute leave-one-out counterfactuals over the entire training set. For each generated image, we take the counterfactual with the largest Euclidean distance to the original image to be the most influential. The generated image is shown above the most influential training image in the above figure.
a) We show the attributed image for 16 images generated from an encoded ensemble of 16 models trained on a dataset of 300 randomly selected images from MNIST.
b) We show the attributed image for 16 images generated from an encoded ensemble of 16 models trained on the 10000 images from MNIST's validation split.
c) We show the attributed image for 16 images generated from an encoded ensemble of 20 models trained on the 60000 images from CIFAR-10 [Krizhevsky et al., 2009].

4 Limitations and Ongoing Work

The proposed approach in this work is to construct a model that has attributability build into it as a feature. It is not a black box method that is meant to be applicable to an existing model. The adoption of this method therefore requires the training of an ensemble of new models.

Currently, we demonstrate attribution for unconditional pixel space diffusion models. The same methods can be applied to text conditioned latent diffusion models, which are currently the subject of regulatory concern [Jiang et al., 2023].

References

S. Basu, P. Pope, and S. Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

Z. Hammoudeh and D. Lowd. Training data influence analysis and estimation: A survey. *arXiv preprint arXiv:2212.04612*, 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

H. H. Jiang, L. Brown, J. Cheng, M. Khan, A. Gupta, D. Workman, A. Hanna, J. Flowers, and T. Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference*

on AI, Ethics, and Society, AIES '23, page 363–374, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604681. URL https://doi.org/10.1145/3600211.3604681.

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

P. S. Menell. Design for symbiosis. *Communications of the ACM*, 55(5):30–32, 2012.

G. Pruthi, F. Liu, S. Kale, and M. Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022.

N. Terashita, H. Ohashi, Y. Nonaka, and T. Kanemaru. Influence estimation for generative adversarial networks. *arXiv preprint arXiv:2101.08367*, 2021.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.