Understanding Instance-based Interpretability of Variational Auto-Encoders

Zhifeng Kong
Computer Science and Engineering
University of California San Diego
La Jolla, CA 92093
z4kong@eng.ucsd.edu

Kamalika Chaudhuri
Computer Science and Engineering
University of California San Diego
La Jolla, CA 92093
kamalika@cs.ucsd.edu

Abstract

Instance-based interpretation methods have been widely studied for supervised learning methods as they help explain how black box neural networks predict. However, instance-based interpretations remain ill-understood in the context of unsupervised learning. In this paper, we investigate influence functions [20], a popular instance-based interpretation method, for a class of deep generative models called variational auto-encoders (VAE). We formally frame the counter-factual question answered by influence functions in this setting, and through theoretical analysis, examine what they reveal about the impact of training samples on classical unsupervised learning methods. We then introduce VAE-TracIn, a computationally efficient and theoretically sound solution based on Pruthi et al. [28], for VAEs. Finally, we evaluate VAE-TracIn on several real world datasets with extensive quantitative and qualitative analysis.

1 Introduction

Instance-based interpretation methods have been popular for supervised learning as they help explain why a model makes a certain prediction and hence have many applications [2, 3, 5, 11, 14, 15, 20, 21, 28, 35, 36]. For a classifier and a test sample z, an instance-based interpretation ranks all training samples x according to an interpretability score between x and z. Samples with high (low) scores are considered positively (negatively) important for the prediction of z.

However, in the literature of unsupervised learning especially generative models, instance-based interpretations are much less understood. In this work, we investigate instance-based interpretation methods for unsupervised learning based on influence functions [8, 20]. In particular, we theoretically analyze certain classical non-parametric and parametric methods. Then, we look at a canonical deep generative model, variational auto-encoders (VAE, [16, 19]), and explore some of the applications.

The first challenge is framing the counter-factual question for unsupervised learning. For instance-based interpretability in supervised learning, the counter-factual question is "which training samples are most responsible for the prediction of a test sample?" – which heavily relies on the label information. However, there is no label in unsupervised learning. In this work, we frame the counter-factual question for unsupervised learning as "which training samples are most responsible for increasing the likelihood (or reducing the loss when likelihood is not available) of a test sample?" We show that influence functions can answer this counter-factual question. Then, we examine influence functions for several classical unsupervised learning methods. We present theory and intuitions on how influence functions relate to likelihood and pairwise distances.

The second challenge is how to compute influence functions in VAE. The first difficulty here is that the VAE loss of a test sample involves an expectation over the encoder, so the actual influence function cannot be precisely computed. To deal with this problem, we use the empirical average of

Preprint. Under review.

influence functions, and prove a concentration bound of the empirical average under mild conditions. Another difficulty is computation. The influence function involves inverting the Hessian of the loss with respect to all parameters, which involves massive computation for big neural networks with millions of parameters. To deal with this problem, we adapt a first-order estimate of the influence function called TracIn [28] to VAE. We call our method VAE-TracIn. It is fast because (i) it does not involve the Hessian, and (ii) it can accelerate computation with only a few checkpoints.

We begin with a sanity check that examines whether training samples have the highest influences over themselves, and show VAE-TracIn passes it. We then evaluate VAE-TracIn on several real world datasets. We find high (low) self influence training samples have large (small) losses. Intuitively, high self influence samples are hard to recognize or visually high-contrast, while low self influence ones share similar shapes or background. These findings lead to an application on unsupervised data cleaning, as high self influence samples are likely to be outside the data manifold. We then provide quantitative and visual analysis on influences over test data. We call high and low influence samples proponents and opponents, respectively. [1] We find in certain cases both proponents and opponents are similar samples from the same class, while in other cases proponents have large norms.

We make the following contributions in this paper.

- We formally frame instance-based interpretations for unsupervised learning.
- We examine influence functions for several classical unsupervised learning methods.
- We present VAE-TracIn, an instance-based interpretation method for VAE. We provide both theoretical and empirical justification to VAE-TracIn.
- We evaluate VAE-TracIn on several real world datasets. We provide extensive quantitative analysis and visualization, as well as an application on unsupervised data cleaning.

1.1 Related Work

There are two lines of research on instance-based interpretation methods for supervised learning.

The first line of research frames the following counter-factual question: which training samples are most responsible for the prediction of a particular test sample z? This is answered by designing an interpretability score that measures the importance of training samples over z and selecting those with the highest scores. Many scores and their approximations have been proposed [2, 3, 5, 14, 20, 21, 28, 35]. Specifically, Koh and Liang [20] introduce the influence function (IF) based on the terminology in robust statistics [8]. The intuition is removing an important training sample of z should result in a huge increase of its test loss. Because the IF is hard to compute, Pruthi et al. [28] propose TracIn, a fast first-order approximation to IF.

Our paper extends the counter-factual question to unsupervised learning where there is no label. We ask: which training samples are most responsible for increasing the likelihood (or reducing the loss) of a test sample? In this paper, we propose VAE-TracIn, an instance-based interpretation method for VAE [16, 19] based on TracIn and IF.

The second line of research considers a different counter-factual question: which training samples are most responsible for the overall performance of the model (e.g. accuracy)? This is answered by designing an interpretability score for each training sample. Again many scores have been proposed [11, 15, 36]. Terashita et al. [32] extend this framework to a specific unsupervised model called generative adversarial networks [12]. They measure influences of samples on several evaluation metrics, and discard samples that harm these metrics. Our paper is orthogonal to these works.

The instance-based interpretation methods lead to many applications in various areas including adversarial learning, data cleaning, prototype selection, data summarization, and outlier detection [2, 3, 5, 10, 11, 14, 15, 17, 20, 28, 31, 33–36]. In this paper, we apply the proposed VAE-TracIn to an unsupervised data cleaning task.

Prior works on interpreting generative models analyze their latent space via measuring disentanglement, explaining and visualizing representations, or analysis in an interactive interface

[1] There are different names in the literature, such as helpful/harmful samples [20], excitatory/inhibitory points [35], and proponents/opponents [28].

[1, 4, 6, 9, 18, 26, 27, 30]. These latent space analysis are complimentary to the instance-based interpretation methods in this paper.

2 Instance-based Interpretations

Let $X = \{x_i\}_{i=1}^N \in \mathbb{R}^d$ be the training set. Let $\mathcal{A}$ be an algorithm that takes X as input and outputs a model that describes the distribution of X. $\mathcal{A}$ can be a density estimator or a generative model. Let $L(X; \mathcal{A}) = \frac{1}{N}\sum_{i=1}^N \ell(x_i; \mathcal{A}(X))$ be a loss function. Then, the influence function of a training data x_i over a test data $z \in \mathbb{R}^d$ is the loss of z computed from the model trained without x_i minus that computed from the model trained with x_i. If the difference is large, then x_i should be very influential for z. Formally, the influence function is defined below.

Definition 1 (Influence functions [20])**.** *Let $X_{-i} = X \setminus \{x_i\}$. Then, the influence of x_i over z is defined as $\mathrm{IF}_{X,\mathcal{A}}(x_i, z) = \ell(z; \mathcal{A}(X_{-i})) - \ell(z; \mathcal{A}(X))$. If $\mathrm{IF}_{X,\mathcal{A}}(x_i, z) > 0$, we say x_i is a proponent of z; otherwise, we say x_i is an opponent of z.*

For big models $\mathcal{A}$ such as deep neural networks, doing retraining and obtaining $\mathcal{A}(X_{-i})$ can be expensive. The following TracIn score is a fast approximation to IF.

Definition 2 (TracIn scores [28])**.** *Suppose $\mathcal{A}(X)$ is obtained by minimizing $L(X; \mathcal{A})$ via stochastic gradient descent. Let $\{\theta_{[c]}\}_{c=1}^C$ be C checkpoints during the training procedure. Then, the estimated influence of x_i over z is defined as $\mathrm{TracIn}_{X,\mathcal{A}}(x_i, z) = \sum_{c=1}^C \nabla\ell(x_i; \theta_{[c]})^\top \nabla\ell(z; \theta_{[c]})$.*

We are also interested in the influence of a training sample over itself. Formally, we define this quantity as the self influence of x, or $\text{Self-IF}_{X,\mathcal{A}}(x) = \mathrm{IF}_{X,\mathcal{A}}(x, x)$. In supervised learning, self influences provide rich information about memorization properties of training samples. Intuitively, high self influence samples are atypical, ambiguous or mislabeled, while low self influence samples are typical [10].

3 Influence Functions for Classical Unsupervised Learning

In this section, we analyze influence functions for unsupervised learning. The goal is to provide intuition on what influence functions should tell us in the unsupervised setting. Specifically, we look at three classical methods: the non-parametric k-nearest-neighbor (k-NN) density estimator, the non-parametric kernel density estimator (KDE), and the parametric Gaussian mixture models (GMM). We let the loss function ℓ to be the negative log-likelihood: $\ell(z) = -\log p(z)$.

The k-Nearest-Neighbor (k-NN) density estimator. The k-NN density estimator is defined as $p_{k\mathrm{NN}}(x; X) = k/(NV_d R_k(x; X)^d)$, where $R_k(x; X)$ is the distance between x and its k-th nearest neighbor in X and V_d is the volume of the unit ball in $\mathbb{R}^d$. Then, we have the following influence function for the k-NN density estimator:

$$\mathrm{IF}_{X,k\mathrm{NN}}(x_i, z) = \log\frac{N-1}{N} + \begin{cases} d\log\frac{R_{k+1}(z;X)}{R_k(z;X)} & \|x_i - z\| \le R_k(z; X) \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

See Appendix A.1.1 for proof. Note, when z is fixed, there are only two possible values for training data influences: $\log\frac{N-1}{N}$ and $\log\frac{N-1}{N} + d\log\frac{R_{k+1}(z;X)}{R_k(z;X)}$. As for $\text{Self-IF}_{X,k\mathrm{NN}}(x_i)$, samples with the largest self influences are those with the largest $\frac{R_{k+1}(x_i;X)}{R_k(x_i;X)}$. Intuitively, these samples belong to a cluster of size exactly k, and the cluster is far away from other samples.

Kernel Density Estimators (KDE). The KDE is defined as $p_{\mathrm{KDE}}(x; X) = \frac{1}{N}\sum_{i=1}^N K_\sigma(x - x_i)$, where K_σ is the Gaussian $\mathcal{N}(0, \sigma^2 I)$. Then, we have the following influence function for KDE:

$$\mathrm{IF}_{X,\mathrm{KDE}}(x_i, z) = \log\frac{N-1}{N} + \log\left(1 + \frac{\frac{1}{N}K_\sigma(z - x_i)}{p_{\mathrm{KDE}}(z; X) - \frac{1}{N}K_\sigma(z - x_i)}\right). \tag{2}$$

See Appendix A.1.2 for proof. For a fixed z, an x_i with larger $\|z - x_i\|$ has a higher influence over z. Therefore, the strongest proponents of z are those closest to z in the ℓ_2 distance, and the strongest

Table 1: High level summary of influence functions in k-NN, KDE and GMM.

Method	high self influence samples	low self influence samples
k-NN	in a cluster of size exactly k	–
KDE	in low density (sparse) region	in high density region
GMM	far away to cluster centers	near cluster centers
Method	**strong proponents**	**strong opponents**
k-NN	k nearest neighbours	other than k nearest neighbours
KDE	nearest neighbours	farthest samples
GMM	nearest neighbours	possibly far away samples in the same class

opponents are the farthest. As for $\text{Self-IF}_{X,\text{KDE}}(x_i)$, samples with the largest self influences are those with the least likelihood $p_{\text{KDE}}(x_i; X)$. Intuitively, these samples locate at very sparse regions and have few nearby samples. On the other hand, samples with the largest likelihood $p_{\text{KDE}}(x_i; X)$, or those in the high density area, have the least self influences.

Gaussian Mixture Models (GMM). As there is no closed-form expression for general GMM, we make the following well-separation assumption to simplify the problem.

Assumption 1. *$X = \bigcup_{k=0}^{K-1} X_k$ where each X_k is a cluster. We assume these clusters are well-separated:* $\min\{\|x - x'\| : x \in X_k, x' \in X_{k'}\} \gg \max\{\|x - x'\| : x, x' \in X_k\}$.

Let $|X_k| = N_k$ and $N = \sum_{k=0}^{K-1} N_k$. For $x \in \mathbb{R}^d$, let $k = \arg\min_i d(x, X_i)$. Then, we define the well-separated spherical GMM (WS-GMM) of K mixtures as $p_{\text{WS-GMM}}(x) = \frac{N_k}{N}\mathcal{N}(x; \mu_k, \sigma_k^2 I)$, where the parameters are given by the maximum likelihood estimates

$$\mu_k = \frac{1}{N_k}\sum_{x \in X_k} x,\ \sigma_k^2 = \frac{1}{N_k d}\sum_{x \in X_k}\|x - \mu_k\|^2 = \frac{1}{N_k d}\sum_{x \in X_k} x^\top x - \frac{1}{d}\mu_k^\top \mu_k. \tag{3}$$

For conciseness, we let test sample z from cluster zero: $z \in \text{conv}(X_0)$. Then, we have the following influence function for WS-GMM. If $x_i \notin X_0$, $\text{IF}_{X,\text{WS-GMM}}(x_i, z) = -\frac{1}{N} + \mathcal{O}\left(N^{-2}\right)$. Otherwise,

$$\text{IF}_{X,\text{WS-GMM}}(x_i, z) = \frac{d+2}{2N_0} + \frac{1}{2N_0\sigma_0^2}\left(\frac{\|z - \mu_0\|^2}{\sigma_0^2} - \|z - x_i\|^2\right) - \frac{1}{N} + \mathcal{O}\left(N_0^{-2}\right). \tag{4}$$

See Appendix A.1.3 for proof. A surprising finding is that some $x_i \in X_0$ may have very negative influences over z (i.e. strong opponents of z are from the same class). This happens with high probability if $\|z - x_i\|^2 \gtrapprox (1 + \sigma_0^2)d + 2\sigma_0^2$ for large dimension d. Next, we compute the self influence of an $x_i \in X_k$. According to (4),

$$\text{Self-IF}_{X,\text{WS-GMM}}(x_i) = \frac{d+2}{2N_k} + \frac{\|x_i - \mu_k\|^2}{2N_k\sigma_k^4} - \frac{1}{N} + \mathcal{O}\left(N_k^{-2}\right). \tag{5}$$

Within each cluster X_k, samples far away to the cluster center μ_k have large self influences and vice versa. Across the entire dataset, samples in cluster X_k whose N_k or σ_k is small tend to have large self influences, which is very different from k-NN or KDE.

3.1 Summary

We summarize the intuitions of influence functions in classical unsupervised learning in Table 1. Among these methods, the strong proponents are all nearest samples, but self influences and strong opponents are quite different. We then visualize an example of six clusters of 2D points in Figure 5 in Appendix B.1. In Figure 6, We plot the self influences of these data points under different density estimators. For a test data point z, we plot influences of all data points over z in Figure 7.

4 Instance-based Interpretations for Variational Auto-encoders

In this section, we show how to compute influence functions for a class of deep generative models called variational auto-encoders (VAE). Specifically, we look at β-VAE [16] defined below, which generalizes the original VAE by Kingma and Welling [19].

Definition 3 (β-VAE [16])**.** *Let d_{latent} be the latent dimension. Let $P_\phi : \mathbb{R}^{d_{\text{latent}}} \to \mathbb{R}^+$ be the decoder and $Q_\psi : \mathbb{R}^d \to \mathbb{R}^+$ be the encoder, where ϕ and ψ are the parameters of the networks. Let $\theta = [\phi, \psi]$. Let the latent distribution P_{latent} be $\mathcal{N}(0, I)$. For $\beta > 0$, the β-VAE model minimizes the following loss:*

$$L_\beta(X;\theta) = \mathbb{E}_{x\sim X}\ell_\beta(x;\theta) = \beta\cdot\mathbb{E}_{x\sim X}\mathrm{KL}\left(Q_\psi(\cdot|x)\|P_{\text{latent}}\right) - \mathbb{E}_{x\sim X}\mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}\log P_\phi(x|\xi). \quad (6)$$

In practice, the encoder $Q = Q_\psi$ outputs two vectors, μ_Q and σ_Q, so that $Q(\cdot|x) = \mathcal{N}(\mu_Q(x), \mathrm{diag}(\sigma_Q(x))^2 I)$. The decoder $P = P_\phi$ outputs a vector μ_P so that $\log P(x|\xi)$ is a constant times $\|\mu_P(\xi) - x\|^2$ plus a constant.

Let $\mathcal{A}$ be the β-VAE that returns $\mathcal{A}(X) = \arg\min_\theta L_\beta(X;\theta)$. Let $\theta^* = \mathcal{A}(X)$ and $\theta^*_{-i} = \mathcal{A}(X_{-i})$. Then, the influence function of x_i over a test point z is $\ell_\beta(z;\theta^*_{-i}) - \ell_\beta(z;\theta^*)$, which equals to

$$\begin{aligned}\mathrm{IF}_{X,\mathrm{VAE}}(x_i, z) \quad &= \beta\left(\mathrm{KL}\left(Q_{\psi^*_{-i}}(\cdot|z)\|P_{\text{latent}}\right) - \mathrm{KL}\left(Q_{\psi^*}(\cdot|z)\|P_{\text{latent}}\right)\right) \\ &\quad - \left(\mathbb{E}_{\xi\sim Q_{\psi^*_{-i}}(\cdot|z)}\log P_{\phi^*_{-i}}(z|\xi) - \mathbb{E}_{\xi\sim Q_{\psi^*}(\cdot|z)}\log P_{\phi^*}(z|\xi)\right).\end{aligned} \quad (7)$$

Challenge. The first challenge is that IF in (7) involves an expectation over the encoder, so it cannot be precisely computed. To solve the problem, we compute the empirical average of the influence function over m samples. In **Theorem** 1, we theoretically prove that the empirical influence function is close to the actual influence function with high probability when m is properly selected. The second challenge is that IF is hard to compute. To solve this problem, in Section 4.1, we propose VAE-TracIn, a computationally efficient solution to VAE.

A probabilistic bound on influence estimates. Let $\hat{\mathrm{IF}}^{(m)}_{X,\mathrm{VAE}}$ be the empirical average of the influence function over m i.i.d. samples. We have the following result.

Theorem 1 (Error bounds on influence estimates (informal, see formal statement in **Theorem** 4))**.** *Under mild conditions, for any small $\epsilon > 0$ and $\delta > 0$, there exists an $m = \Theta\left(\frac{1}{\epsilon^2\delta}\right)$ such that*

$$\mathrm{Prob}\left(\left|\mathrm{IF}_{X,\mathrm{VAE}}(x_i, z) - \hat{\mathrm{IF}}^{(m)}_{X,\mathrm{VAE}}(x_i, z)\right| \geq \epsilon\right) \leq \delta. \quad (8)$$

Formal statements and proofs are in Appendix A.2.

4.1 VAE-TracIn

In this section, we introduce VAE-TracIn, a computationally efficient interpretation method for VAE. VAE-TracIn is built based on TracIn (**Definition** 2). According to (6), the gradient of the loss $\ell_\beta(x;\theta)$ can be written as $\nabla_\theta\ell_\beta(x;\theta) = [\nabla_\phi\ell_\beta(x;\theta)^\top, \nabla_\psi\ell_\beta(x;\theta)^\top]^\top$, where

$$\begin{aligned}\nabla_\phi\ell_\beta(x;\theta) \quad &= \mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}\left(-\nabla_\phi\log P_\phi(x|\xi)\right) =: \mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}U(x,\xi,\phi,\psi), \text{ and} \\ \nabla_\psi\ell_\beta(x;\theta) \quad &= \mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}\nabla_\psi\log Q_\psi(\xi|x)\left(\beta\log\frac{Q_\psi(\xi|x)}{P_{\text{latent}}(\xi)} - \log P_\phi(x|\xi)\right) \\ &=: \mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}V(x,\xi,\phi,\psi).\end{aligned} \quad (9)$$

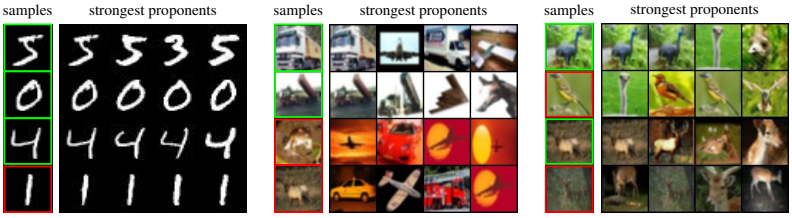
The derivations are based on the Stochastic Gradient Variational Bayes estimator [19], which offers low variance [29]. See Appendix A.3 for full details of the derivation. We estimate the expectation $\mathbb{E}_\xi$ by averaging over m i.i.d. samples. Then, for a training data x and test data z, the VAE-TracIn score of x over z is computed as

$$\begin{aligned}\text{VAE-TracIn}(x, z) \quad &= \sum_{c=1}^{C}\left(\frac{1}{m}\sum_{j=1}^{m}U(x,\xi_{j,[c]},\phi_{[c]},\psi_{[c]})\right)^\top\left(\frac{1}{m}\sum_{j=1}^{m}U(z,\xi'_{j,[c]},\phi_{[c]},\psi_{[c]})\right) \\ &\quad + \sum_{c=1}^{C}\left(\frac{1}{m}\sum_{j=1}^{m}V(x,\xi_{j,[c]},\phi_{[c]},\psi_{[c]})\right)^\top\left(\frac{1}{m}\sum_{j=1}^{m}V(z,\xi'_{j,[c]},\phi_{[c]},\psi_{[c]})\right),\end{aligned} \quad (10)$$

where the notations U, V are from (9), $\theta_{[c]} = [\phi_{[c]}, \psi_{[c]}]$ is the c-th checkpoint, $\{\xi_{j,[c]}\}_{j=1}^m$ are i.i.d. samples from $Q_{\psi_{[c]}}(\cdot|x)$, and $\{\xi'_{j,[c]}\}_{j=1}^m$ are i.i.d. samples from $Q_{\psi_{[c]}}(\cdot|z)$.

Table 2: Sanity check on the frequency that a training sample is the most influential one over itself. Results on MNIST, CIFAR, and the averaged result on CIFAR subclasses are reported.

MNIST		CIFAR		Averaged CIFAR subclass
$d_{\text{latent}} = 64$	$d_{\text{latent}} = 128$	$d_{\text{latent}} = 64$	$d_{\text{latent}} = 128$	$d_{\text{latent}} = 64$
0.992	1.000	0.609	0.602	0.998



(a) MNIST (b) CIFAR (c) CIFAR subclass

Figure 1: Certain training samples and their strongest proponents in the training set (sorted from left to right). A sample x_i is marked in green box if it is more influential than other samples over itself (i.e. it is the strongest proponent of itself) and otherwise in red box.

5 Experiments

In this section, we aim to answer the following questions.

- Does VAE-TracIn pass a sanity check for instance-based interpretations?
- Which training samples have the highest and lowest self influences, respectively?
- Which training samples have the highest influences over (i.e. are strong proponents of) a test sample? Which have the lowest influences over it (i.e. are its strong opponents)?

These questions are examined in experiments on the MNIST [23] and CIFAR-10 [22] datasets.

5.1 Sanity Check

Question. Does VAE-TracIn find the most influential training samples? In a perfect instance-based interpretation for a good model, training samples should have large influences over themselves. As a sanity check, we examine if training samples are the strongest proponents over themselves. This is analogous to the identical subclass test by Hanawa et al. [13].

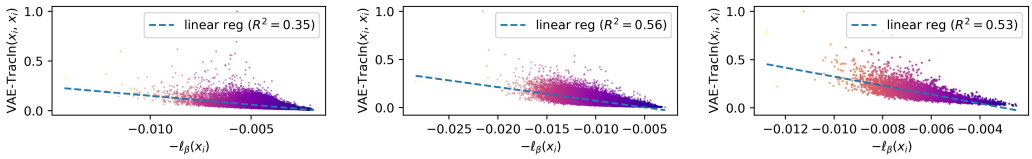
Methodology. We train separate VAE models on MNIST, CIFAR, and each CIFAR subclass (the set of five thousand CIFAR samples in each class). For each model, we examine the frequency that a training sample is the most influential one among all samples over itself. Due to computational limits we examine the first 128 samples. The results for MNIST, CIFAR, and the averaged result for CIFAR subclasses are reported in Table 2. Detailed results for CIFAR subclasses are in Appendix B.3.

Results. The results indicate that VAE-TracIn can find the most influential training samples in MNIST and CIFAR subclasses. This is achieved even under the challenge that many training samples are very similar to each other. The results for CIFAR is possibly due to underfitting as it is challenging to train a good VAE on this dataset. Note, the same VAE architecture is trained on CIFAR subclasses.

Visualization. We visualize some correctly and incorrectly identified strongest proponents in Figure 1. On MNIST or CIFAR subclasses, even if a training sample is not exactly the strongest proponent of itself, it still ranks very high in the order of influences.

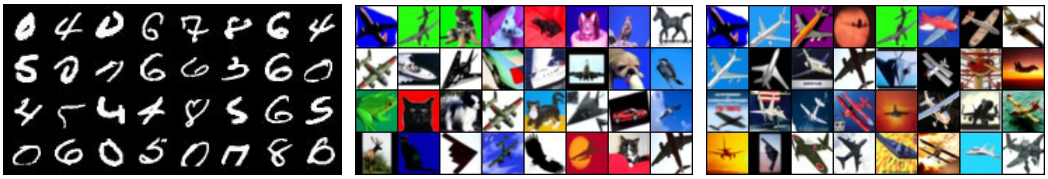
5.2 Self Influences

Question. Which training samples have the highest and lowest self influences, respectively? Self influences provide rich information about properties of training samples such as memorization. In supervised learning, high self influence samples can be atypical, ambiguous or mislabeled, while low self influence samples are typical [10]. We examine what self influences reveal in VAE.



(a) MNIST (b) CIFAR (c) CIFAR-Airplane

Figure 2: Scatter plots of self influences versus negative losses of all training samples in several datasets. The linear regressors show that high self influence samples have large losses.



(a) MNIST (highest self-inf) (b) CIFAR (highest self-inf) (c) CIFAR-Airplane (highest self-inf)

(d) MNIST (lowest self-inf) (e) CIFAR (lowest self-inf) (f) CIFAR-Airplane (lowest self-inf)

Figure 3: High and low self influence samples from several datasets. High self influence samples are hard to recognize or high-contrast. Low self influence samples share similar shapes or background.

Methodology. We train separate VAE models on MNIST, CIFAR, and each CIFAR subclass. We then compute the self influences and losses of each training sample. We show the scatter plots of self influences versus negative losses in Figure 2. [2] We fit linear regression models to these points and report the R^2 scores of the regressors. More comparisons including the marginal distributions and the joint distributions can be found in Appendix B.4 and Appendix B.5.

Results. We find the self influence of a training sample x_i tends to be large when its loss $\ell_\beta(x_i)$ is large. This finding in VAE is consistent with KDE and GMM (see Figure 6). In supervised learning, Pruthi et al. [28] find high self influence samples come from densely populated areas while low self influence samples come from sparsely populated areas. Our findings indicate significant difference between supervised and unsupervised learning in terms of self influences under certain scenarios.

Visualization. We visualize high and low self influence samples in Figure 3 (more visualizations in Appendix B.5). High self influence samples are either hard to recognize or visually high-contrast, while low self influence samples share similar shapes or background. These visualizations are consistent with the memorization analysis by Feldman and Zhang [10] in the supervised setting.

Application on unsupervised data cleaning. A potential application on unsupervised data cleaning is to use self influences to detect unlikely samples and let a human expert decide whether to discard them before training. The unlikely samples may include noisy samples, contaminated samples, or incorrectly collected samples due to bugs in the data collection process. For example, they could be unrecognizable handwritten digits in MNIST or objects in CIFAR. Similar approaches in supervised learning use self influences to detect mislabeled data [20, 28, 35] or memorized samples [10]. We extend the application of self influences to scenarios where there are no labels.

To test this application, we design an experiment to see if self influences can find a small amount of extra samples added to the original dataset. The extra samples are from other datasets: 1000 EMNIST [7] samples for MNIST, and 1000 CelebA [24] samples for CIFAR, respectively. In Figure

[2] We use the negative loss because it relates to the log-likelihood of x_i: when $\beta = 1$, $-\ell_\beta(x) \leq \log p(x)$.

15, we plot the detection curves to show fraction of extra samples found when all samples are sorted in the self influence order. The area under these detection curves (AUC) are 0.887 in the MNIST experiment and 0.760 in the CIFAR experiment. [3] Full results and more comparisons can be found in Appendix B.6. The results indicate that extra samples generally have higher self influences than original samples, so it has much potential to apply VAE-TracIn to unsupervised data cleaning.

5.3 Influences over Test Data

Question. Which training samples are strong proponents or opponents of a test sample, respectively? Influences over a test sample z provide rich information about the relationship between training data and z. In supervised learning, strong proponents help the model correctly predict the label of z while strong opponents harm it. Empirically, strong proponents are visually similar samples from the same class, while strong opponents tend to confuse the model [28]. In unsupervised learning, we expect that strong proponents increase the likelihood of z and strong opponents reduce it. We examine which samples are strong proponents or opponents in VAE.

Methodology. We train separate VAE models on MNIST, CIFAR, and each CIFAR subclass. We then compute VAE-TracIn scores of all training samples over 128 test samples.

In MNIST experiments, we plot the distributions of influences according to whether training and test samples belong to the same class (See results on label zero in Figure 18 and full results in Figure 19). We then compare the influences of training over test samples to their distances in the latent space in Figure 20f. Quantitatively, we define samples that have the 0.1% highest/lowest influences as the strongest proponents/opponents. Then, we report the fraction of the strongest proponents/opponents that belong to the same class as the test sample and the statistics of pairwise distances in Table 3. Additional comparisons can be found in Appendix B.7,

In CIFAR and CIFAR subclass experiments, we compare influences of training over test samples to the norms of training samples in the latent space in Figure 22 and Figure 23. Quantitatively, we report the statistics of the norms in Table 4. Additional comparisons can be found in Appendix B.8.

Results. In MNIST experiments, we find many strong proponents and opponents of a test sample are its similar samples from the same class. In terms of class information, many ($\sim 80\%$) strongest proponents and many ($\sim 40\%$) strongest opponents have the same label as test samples. In terms of distances in the latent space, it is shown that the strongest proponents and opponents are close (thus similar) samples, while far away samples have small absolute influences. These findings are similar to GMM discussed in Section 3, where the strongest opponents may come from the same class (see Figure 7). The findings are also related to the supervised setting in the sense that dissimilar samples from a different class have small influences.

Results in CIFAR and CIFAR subclass experiments indicate strong proponents have large norms in the latent space. [4] This observation also happens to many instance-based interpretations in the supervised setting including classification methods [13] and logistic regression [2], where large norm samples can impact a large region in the data space, so they are influential to many test samples.

Visualization. We visualize the strongest proponents and opponents in Figure 4. More visualizations can be found in Appendix B.7 and Appendix B.8. In the MNIST experiment, the strongest proponents look very similar to test samples. The strongest opponents are often the same but visually different digits. For example, the opponents of the test "two" have very different thickness and styles. In CIFAR and CIFAR subclass experiments, we find strong proponents seem to match the color of the images – including the background and the object – and they tend to have the same but brighter colors. Nevertheless, many proponents are from the same class. Strong opponents, on the other hand, tend to have very different colors as the test samples.

5.4 Discussion

VAE-TracIn provides rich information about instance-level interpretability in VAE. In terms of self influences, there is correlation between self influences and VAE losses. Visually, high self influence samples are ambiguous or high-contrast while low self influence samples are similar in shape or

[3] AUC ≈ 1 means the detection is near perfect, and AUC ≈ 0.5 means the detection is near random.

[4] Large norm samples can be outliers, high-contrast samples, or very bright samples.

Table 3: Statistics of influences, class information, and distances of train-test sample pairs in MNIST. "+" means top-0.1% strong proponents, "−" means top-0.1% strong opponents, and "all" means the train set. The first two rows are fractions of samples that belong to the same class as the test sample. The bottom three rows are means ± standard errors of latent space distances between train-test sample pairs.

d_{latent}	64	96	128
same class rate (+)	81.9%	84.0%	82.1%
same class rate (−)	37.3%	43.3%	40.3%
distances (+)	0.94 ± 0.53	0.94 ± 0.55	0.76 ± 0.51
distances (−)	1.78 ± 0.75	1.84 ± 0.78	1.29 ± 0.67
distances (all)	2.54 ± 0.90	2.57 ± 0.91	2.23 ± 0.92

Table 4: The means ± standard errors of latent space norms of training samples in CIFAR and CIFAR-Airplane. Notations follow Table 3. It is shown that strong proponents tend to have very large norms.

CIFAR	(+)	7.42 ± 1.10
	(−)	3.89 ± 1.26
	(all)	5.06 ± 1.18
CIFAR-Airplane	(+)	4.73 ± 0.78
	(−)	4.26 ± 0.91
	(all)	4.07 ± 0.83

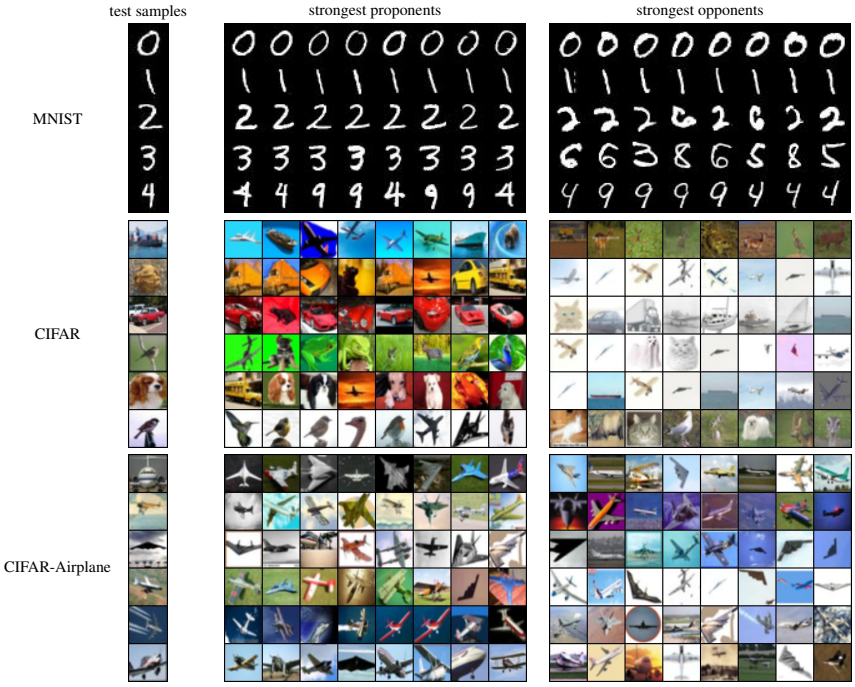


Figure 4: Test samples from several datasets, their strongest proponents, and strongest opponents. In MNIST the strongest proponents are visually similar while the strongest opponents are often the same digit but are visually different. In CIFAR and CIFAR-Airplane the strongest proponents seem to match the colors and are often very bright or high-contrast.

background. In terms of influences over test samples, for VAE trained on MNIST, many proponents and opponents are similar samples in the same class, and for VAE trained on CIFAR, proponents have large norms in the latent space. There are strong connections between these findings and influence functions in KDE, GMM, classification and simple regression models.

6 Conclusion

Influence functions in unsupervised learning can reveal the most responsible training samples that increase the likelihood (or reduce the loss) of a particular test sample. In this paper, we investigate influence functions for several classical unsupervised learning methods and one deep generative model with extensive theoretical and empirical analysis. We present VAE-TracIn, a theoretical sound and computationally efficient algorithm that estimates influence functions for VAE, and evaluate it on real world datasets. One limitation of our work is that it is still challenging to apply VAE-TracIn to modern, huge models trained on millions of samples, which is an important future direction.

Acknowledgements

We thank NSF under IIS 1719133 and CNS 1804829 for research support. We thank Casey Meehan, Yao-yuan Yang, and Mary Anne Smart for helpful feedback.

References

[1] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*, 2018.

[2] Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. Relatif: Identifying explanatory training samples via relative influence. In *International Conference on Artificial Intelligence and Statistics*, pages 1899–1909. PMLR, 2020.

[3] Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pages 715–724. PMLR, 2020.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

[5] Hongge Chen, Si Si, Yang Li, Ciprian Chelba, Sanjiv Kumar, Duane Boning, and Cho-Jui Hsieh. Multi-stage influence function. *arXiv preprint arXiv:2007.09081*, 2020.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

[7] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[8] R Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980.

[9] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012.

[10] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *arXiv preprint arXiv:2008.03703*, 2020.

[11] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR, 2019.

[12] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[13] Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. Evaluation of similarity-based explanations. *arXiv preprint arXiv:2006.04528*, 2020.

[14] Satoshi Hara, Atsushi Nitanda, and Takanori Maehara. Data cleansing for models trained with sgd. *arXiv preprint arXiv:1906.08473*, 2019.

[15] Hrayr Harutyunyan, Alessandro Achille, Giovanni Paolini, Orchid Majumder, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Estimating informativeness of samples with smooth unique information. *arXiv preprint arXiv:2101.06640*, 2021.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[17] Rajiv Khanna, Been Kim, Joydeep Ghosh, and Sanmi Koyejo. Interpreting black box predictions using fisher kernels. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3382–3390. PMLR, 2019.

[18] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[21] Pang Wei Koh, Kai-Siang Ang, Hubert HK Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects. *arXiv preprint arXiv:1905.13289*, 2019.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[25] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. *arXiv preprint arXiv:2004.05675*, 2020.

[26] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

[27] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

[28] Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracking gradient descent. *arXiv preprint arXiv:2002.08484*, 2020.

[29] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[30] Andrew Slavin Ross, Nina Chen, Elisa Zhao Hang, Elena L Glassman, and Finale Doshi-Velez. Evaluating the interpretability of generative models by interactive reconstruction. *arXiv preprint arXiv:2102.01264*, 2021.

[31] Kenji Suzuki, Yoshiyuki Kobayashi, and Takuya Narihira. Data cleansing for deep neural networks with storage-efficient approximation of influence functions. *arXiv preprint arXiv:2103.11807*, 2021.

[32] Naoyuki Terashita, Hiroki Ohashi, Yuichi Nonaka, and Takashi Kanemaru. Influence estimation for generative adversarial networks. *arXiv preprint arXiv:2101.08367*, 2021.

[33] Daniel Ting and Eric Brochu. Optimal subsampling with influence functions. In *Advances in neural information processing systems*, pages 3650–3659, 2018.

[34] Haotian Ye, Chuanlong Xie, Yue Liu, and Zhenguo Li. Out-of-distribution generalization analysis via influence function. *arXiv preprint arXiv:2101.08521*, 2021.

[35] Chih-Kuan Yeh, Joon Sik Kim, Ian EH Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks. *arXiv preprint arXiv:1811.09720*, 2018.

[36] Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pages 10842–10851. PMLR, 2020.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver 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.

Contents

A Omitted Proofs

A.1 Omitted Proofs in Section 3

A.1.1 Proof of (1)

Proof. By definition, we have

$$p_{k\mathrm{NN}}(z;X)=\frac{k}{NV_dR_k(z;X)^d}$$

and

$$p_{k\mathrm{NN}}(z;X_{-i})=\frac{k}{(N-1)V_dR_k(z;X_{-i})^d}.$$

If x_i belongs to k-NN of z, then $R_k(z;X_{-i})$ is $R_{k+1}(z;X)$; otherwise, $R_k(z;X_{-i})$ is $R_k(z;X)$. The result follows by subtracting the logarithm of these two densities. □

A.1.2 Proof of (2)

Proof. By definition, we have

$$p_{\mathrm{KDE}}(z;X)=\frac{1}{N}\sum_{j=1}^{N}K_\sigma(z-x_j)$$

and

$$p_{\mathrm{KDE}}(z;X_{-i})=\frac{1}{N-1}\sum_{j\neq i}^{N}K_\sigma(z-x_j).$$

The result follows by subtracting the logarithm of these two densities. It is interesting to notice when $\max_i K_\sigma(z-x_i)\ll N\cdot p_{\mathrm{KDE}}(z;X)$, we have

$$\sum_{i=1}^{N}\mathrm{IF}_{X,\mathrm{KDE}}(x_i,z)\approx\frac{1}{N}\sum_{i=1}^{N}\left(\frac{K_\sigma(z-x_i)}{p_{\mathrm{KDE}}(z;X)}-1\right)=0.$$

□

A.1.3 Proof of (4)

Proof. By definition, we have

$$p_{\text{WS-GMM}}(z;X)=\frac{N_0}{N}\mathcal{N}(z;\mu_0,\sigma_0^2I).$$

If $x_i\notin X_0$, then

$$p_{\text{WS-GMM}}(z;X_{-i})=\frac{N_0}{N-1}\mathcal{N}(z;\mu_0,\sigma_0^2I).$$

In this case, we have $\mathrm{IF}_{X,\text{WS-GMM}}(x_i,z)=-\log(1+1/N)$.

If $x_i\in X_0$, then parameters μ_0 and σ_0 are to be modified to maximize likelihood estimates over $X_0\setminus\{x_i\}$. Denote the modified parameters as μ_0' and σ_0'. Then, we have

$$p_{\text{WS-GMM}}(z;X_{-i})=\frac{N_0-1}{N-1}\mathcal{N}(z;\mu_0',(\sigma_0')^2I).$$

Next, we express μ_0' and σ_0' in terms of known variables. For conciseness, we let $v=z-\mu_0$ and $u=x_i-\mu_0$.

$$\begin{aligned}\mu_0' &= \frac{1}{N_0-1}\sum_{x\in X_0\setminus\{x_i\}}x\\ &= \frac{N_0\mu_0-x_i}{N_0-1}\\ &= \mu_0-\frac{u}{N_0-1}.\end{aligned}$$

$$
\begin{aligned}
(\sigma_0')^2 &= \frac{1}{(N_0-1)d}\sum_{x\in X_0\setminus\{x_i\}} x^\top x - \frac{1}{d}(\mu_0')^\top\mu_0' \\
&= \frac{1}{(N_0-1)d}\left(\sum_{x\in X_0} x^\top x - x_i^\top x_i - (N_0-1)\left(\mu_0^\top\mu_0 - \frac{2u^\top\mu_0}{N_0-1} + \frac{u^\top u}{(N_0-1)^2}\right)\right) \\
&= \frac{1}{(N_0-1)d}\left(N_0 d\sigma_0^2 - x_i^\top x_i + \mu_0^\top\mu_0 + 2u^\top\mu_0 - \frac{u^\top u}{N_0-1}\right) \\
&= \frac{1}{(N_0-1)d}\left(N_0 d\sigma_0^2 - x_i^\top x_i - \frac{N_0 u^\top u}{N_0-1}\right) \\
&= \frac{N_0}{N_0-1}\sigma_0^2 - \frac{N_0 u^\top u}{(N_0-1)^2 d}.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\log p_{\text{WS-GMM}}(z;X) &= \log\frac{N_0}{N} - \frac{d}{2}\log 2\pi - \frac{d}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}(z-\mu_0)^\top(z-\mu_0) \\
&= \log\frac{N_0}{N} - \frac{d}{2}\log 2\pi - \frac{d}{2}\log\sigma_0^2 - \frac{v^\top v}{2\sigma_0^2},
\end{aligned}
$$

and

$$
\begin{aligned}
\log p_{\text{WS-GMM}}(z;X_{-i}) &= \log\frac{N_0-1}{N-1} - \frac{d}{2}\log 2\pi - \frac{d}{2}\log(\sigma_0')^2 - \frac{1}{2{\sigma_0'}^2}(z-\mu_0')^\top(z-\mu_0') \\
&= \log\frac{N_0-1}{N-1} - \frac{d}{2}\log 2\pi - \frac{d}{2}\log\frac{N_0}{N_0-1} - \frac{d}{2}\log\sigma_0^2 \\
&\quad - \frac{d}{2}\log\left(1 - \frac{u^\top u}{(N_0-1)d\sigma_0^2}\right) \\
\left(z-\mu_0' = v + \tfrac{u}{N_0-1}\right) &\quad - \frac{(N_0-1)^2 v^\top v + 2(N_0-1)u^\top v + u^\top u}{2N_0\left((N_0-1)\sigma_0^2 - \frac{1}{d}u^\top u\right)} \\
&= \log\frac{N_0-1}{N-1} - \frac{d}{2}\log 2\pi - \frac{d}{2N_0} - \frac{d}{2}\log\sigma_0^2 + \frac{u^\top u}{2N_0\sigma_0^2} \\
&\quad - \frac{v^\top v}{2\sigma_0^2} - \frac{1}{2N_0\sigma_0^2}\left(2u^\top v - v^\top v + \frac{v^\top v}{\sigma_0^2}\right) + \mathcal{O}\left(N_0^{-2}\right).
\end{aligned}
$$

Subtracting the above two equations, we have

$$
\begin{aligned}
\text{IF}_{X,\text{WS-GMM}}(x_i,z) &= \frac{1}{N_0} - \frac{1}{N} + \frac{d}{2N_0} + \frac{1}{2N_0\sigma_0^2}\left(2u^\top v - v^\top v - u^\top u + \frac{v^\top v}{\sigma_0^2}\right) + \mathcal{O}\left(N_0^{-2}\right) \\
&= \frac{d+2}{2N_0} + \frac{1}{2N_0\sigma_0^2}\left(\frac{\|z-\mu_0\|^2}{\sigma_0^2} - \|z-x_i\|^2\right) - \frac{1}{N} + \mathcal{O}\left(N_0^{-2}\right).
\end{aligned}
$$

□

A.2 Probabilistic Bound on Influence Estimates

Let $\{\xi_j\}_{j=1}^m$ be m i.i.d. samples drawn from $Q_{\psi^*}(\cdot|z)$ and $\{\xi_j'\}_{j=1}^m$ be m i.i.d. samples drawn from $Q_{\psi^*_{-i}}(\cdot|z)$. We can use the empirical influence $\hat{\text{IF}}^{(m)}_{X,\text{VAE}}(x_i,z)$ to estimate the true influence in (7), which is defined below:

$$
\begin{aligned}
\hat{\text{IF}}^{(m)}_{X,\text{VAE}}(x_i,z) &= \beta\left(\text{KL}\left(Q_{\psi^*_{-i}}(\cdot|z)\|P_{\text{latent}}\right) - \text{KL}\left(Q_{\psi^*}(\cdot|z)\|P_{\text{latent}}\right)\right) \\
&\quad - \tfrac{1}{m}\textstyle\sum_{j=1}^m\left(\log P_{\phi^*_{-i}}(z|\xi_j') - \log P_{\phi^*}(z|\xi_j)\right).
\end{aligned} \tag{11}
$$

The questions is, when can we guarantee the empirical influence score $\hat{\text{IF}}^{(m)}_{X,\text{VAE}}(x_i,z)$ is close to the true influence score $\text{IF}_{X,\text{VAE}}(x_i,z)$? We answer this question via an (ϵ,δ)-probabilistic bound: as long as m is larger than a function of ϵ and δ, then with probability at least $1-\delta$, the difference between the empirical and true influence scores is no more than ϵ. To introduce the theory, we first provide the following definition.

Definition 4 (Polynomially-bounded functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^{d'}$. We say f is polynomially bounded by $\{a_c\}_{c=1}^C$ if for any $x \in \mathbb{R}^d$, we have*

$$\|f(x)\| \leq \sum_{c=1}^{C} a_c \|x\|^c. \tag{12}$$

We provide a useful lemma on polynomially-bounded functions below.

Lemma 2. *The composition of polynomially bounded functions is polynomially bounded.*

Next, we show common neural networks are polynomially bounded in the following proposition.

Proposition 3. *Let f be a neural network taking the following form:*

$$f(x) = \sigma_l(W_l \sigma_{l-1}(W_{l-1} \cdots \sigma_1(W_1 x) \cdots)). \tag{13}$$

If every activation function σ_j is polynomially bounded, then f is polynomially bounded.

With the above result, we state the (ϵ, δ)-probabilistic bound on influence estimates below.

Theorem 4 (Error bounds on influence estimates)**.** *Let P and Q be two polynomially bounded networks. For any small $\epsilon > 0$ and $\delta > 0$, there exists an $m = \Theta\left(\frac{1}{\epsilon^2 \delta}\right)$ such that*

$$\text{Prob}\left(\left|\text{IF}_{X,\text{VAE}}(x_i, z) - \hat{\text{IF}}^{(m)}_{X,\text{VAE}}(x_i, z)\right| \geq \epsilon\right) \leq \delta, \tag{14}$$

where the randomness is over all ξ_j and ξ'_j.

A.2.1 Proof of Lemma 2

Proof. Let $f : \mathbb{R}^{m_0} \to \mathbb{R}^{m_1}$ be polynomially bounded by $\{a_c\}_{c=1}^{C_f}$ and $g : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$ be polynomially bounded by $\{b_c\}_{c=1}^{C_g}$. Then, for any $x \in \mathbb{R}^{m_0}$,

$$\|f(x)\| \leq \sum_{c=1}^{C_f} a_c \|x\|^c,$$

and

$$\|g \circ f(x)\| \leq \sum_{c=1}^{C_g} b_c \|f(x)\|^c.$$

Therefore, we have

$$\|g \circ f(x)\| \leq \sum_{c=1}^{C_g} b_c \left(\sum_{c'=1}^{C_f} a_{c'} \|x\|^{c'}\right)^c.$$

This indicates that $g \circ f$ is polynomially bounded. □

A.2.2 Proof of Proposition 3

Proof. First, an affine transformation Wx is polynomially bounded because $\|Wx\| \leq \|W\| \cdot \|x\|$. Then, we show an element-wise transformation $\sigma(x)$ is polynomially bounded. Let σ be polynomially bounded by $\{a_c\}_{c=1}^C$. Then,

$$\|\sigma(x)\| \leq \frac{1}{\sqrt{d}} \left(\sum_{i=1}^{d} |\sigma(x_i)|\right)^2 \leq \frac{1}{\sqrt{d}} \left(\sum_{i=1}^{d} \left|\sum_{c=1}^{C} a_c |x_i|^c\right|\right)^2 \leq \frac{1}{\sqrt{d}} \left(\sum_{i=1}^{d} \left|\sum_{c=1}^{C} a_c \|x\|^c\right|\right)^2.$$

By **Lemma** 2, since f is a composition of polynomially bounded functions, we have f is polynomially bounded. □

A.2.3 Proof of Theorem 4

Proof. According to (7) and (11),

$$\begin{aligned}\mathrm{IF}_{X,\mathrm{VAE}}(x_i,z)-\hat{\mathrm{IF}}^{(m)}_{X,\mathrm{VAE}}(x_i,z) &= \frac{1}{m}\sum_{j=1}^{m}\log P_{\phi^*_{-i}}(z|\xi'_j)-\mathbb{E}_{\xi\sim Q_{\psi^*_{-i}}(\cdot|z)}\log P_{\phi^*_{-i}}(z|\xi)\\ &\quad-\frac{1}{m}\sum_{j=1}^{m}\log P_{\phi^*}(z|\xi_j)+\mathbb{E}_{\xi\sim Q_{\psi^*}(\cdot|z)}\log P_{\phi^*}(z|\xi).\end{aligned}$$

If we have

$$\left|\mathbb{E}_{\xi\sim Q_{\psi^*}(\cdot|z)}\log P_{\phi^*}(z|\xi)-\frac{1}{m}\sum_{j=1}^{m}\log P_{\phi^*}(z|\xi_j)\right|\le\frac{\epsilon}{2}$$

and

$$\left|\mathbb{E}_{\xi\sim Q_{\psi^*_{-i}}(\cdot|z)}\log P_{\phi^*_{-i}}(z|\xi)-\frac{1}{m}\sum_{j=1}^{m}\log P_{\phi^*_{-i}}(z|\xi'_j)\right|\le\frac{\epsilon}{2},$$

then $\left|\mathrm{IF}_{X,\mathrm{VAE}}(x_i,z)-\hat{\mathrm{IF}}^{(m)}_{X,\mathrm{VAE}}(x_i,z)\right|\le\epsilon$. Let ζ and ζ_j be i.i.d. standard Gaussian random variables for $j=1,2,\cdots,m$. First, we provide the probabilistic bound for the first inequality. Let $P=P_{\phi^*}$ and $Q=Q_{\psi^*}$. Then, we can reparameterize $\xi=\mu_Q(z)+\sigma_Q(z)\odot\zeta$ and $\xi_j=\mu_Q(z)+\sigma_Q(z)\odot\zeta_j$. Let

$$f(\zeta)=\|z-\mu_P(\mu_Q(z)+\sigma_Q(z)\odot\zeta)\|_2^2.$$

Since $\log P(z|\xi)$ is a constant α times $\|z-\mu_P(\xi)\|^2$ plus another constant, we have

$$\mathbb{E}_{\xi\sim Q(\cdot|z)}\log P(z|\xi)-\frac{1}{m}\sum_{j=1}^{m}\log P(z|\xi_j)=\mathbb{E}_\zeta f(\zeta)-\frac{1}{m}\sum_{j=1}^{m}f(\zeta_j).$$

By Chebyshev's inequality,

$$\mathrm{Prob}\left(\left|\mathbb{E}_\zeta f(\zeta)-\frac{1}{m}\sum_{j=1}^{m}f(\zeta_j)\right|\ge\frac{\epsilon}{2}\right)\le\frac{4\mathrm{Var}_\zeta f(\zeta)}{m\epsilon^2}\le\frac{4\mathbb{E}_\zeta f(\zeta)^2}{m\epsilon^2}.$$

By **Lemma** 2, if P and Q are polynomially bounded, then f is polynomially bounded and so is f^2. Let

$$f(\zeta)^2\le\sum_{c=1}^{C}a_c\|\zeta\|^c.$$

Then,

$$\begin{aligned}\mathbb{E}_\zeta f(\zeta)^2 &= \int_{\mathbb{R}^d}\mathcal{N}(\zeta;0,I)f(\zeta)^2d\zeta\\ &\le\frac{1}{(2\pi)^{\frac{d}{2}}}\sum_{c=1}^{C}a_c\int_{\mathbb{R}^d}\|\zeta\|^c e^{-\frac{\|\zeta\|^2}{2}}d\zeta\\ \text{(use polar coordinate)}\quad &= \frac{\pi^{\frac{d-1}{2}}}{(2\pi)^{\frac{d}{2}}\Gamma\left(\frac{d+1}{2}\right)}\sum_{c=1}^{C}a_c\int_0^{+\infty}r^{c+d-1}e^{-\frac{r^2}{2}}dr\\ &= \frac{\pi^{\frac{d-1}{2}}}{(2\pi)^{\frac{d}{2}}\Gamma\left(\frac{d+1}{2}\right)}\sum_{c=1}^{C}a_c\int_0^{+\infty}(2r)^{\frac{c+d}{2}-1}e^{-r}dr\\ &= \frac{1}{2\sqrt{\pi}\Gamma\left(\frac{d+1}{2}\right)}\sum_{c=1}^{C}2^{\frac{c}{2}}a_c\Gamma\left(\frac{c+d}{2}\right),\end{aligned}$$

which is a constant. Therefore, there exists an $M_1=\Theta\left(\frac{1}{\epsilon^2\delta}\right)$ such that when $m\ge M_1$,

$$\mathrm{Prob}\left(\left|\mathbb{E}_\zeta f(\zeta)-\frac{1}{m}\sum_{j=1}^{m}f(\zeta_j)\right|\ge\frac{\epsilon}{2}\right)\le\frac{\delta}{2},$$

or

$$\text{Prob}\left(\left|\mathbb{E}_{\xi\sim Q_{\psi^*}(\cdot|z)}\log P_{\phi^*}(z|\xi)-\frac{1}{m}\sum_{j=1}^{m}\log P_{\phi^*}(z|\xi_j)\right|\geq\frac{\epsilon}{2}\right)\leq\frac{\delta}{2}.$$

Similarly, when $P=P_{\phi^*_{-i}}$ and $Q=Q_{\psi^*_{-i}}$, there exists an $M_2=\Theta\left(\frac{1}{\epsilon^2\delta}\right)$ such that when $m\geq M_2$,

$$\text{Prob}\left(\left|\mathbb{E}_{\xi\sim Q_{\psi^*_{-i}}(\cdot|z)}\log P_{\phi^*_{-i}}(z|\xi)-\frac{1}{m}\sum_{j=1}^{m}\log P_{\phi^*_{-i}}(z|\xi'_j)\right|\geq\frac{\epsilon}{2}\right)\leq\frac{\delta}{2}.$$

Taking $m=\max(M_1,M_2)=\Theta\left(\frac{1}{\epsilon^2\delta}\right)$, we have

$$\text{Prob}\left(\left|\text{IF}_{X,\text{VAE}}(x_i,z)-\hat{\text{IF}}^{(m)}_{X,\text{VAE}}(x_i,z)\right|\geq\epsilon\right)\leq\delta.$$

□

A.3 Derivation of (9)

$$\begin{aligned}
\nabla_\phi\ell_\beta(x;\theta) &= -\nabla_\phi\mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}\log P_\phi(x|\xi)\\
&= -\mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}\nabla_\phi\log P_\phi(x|\xi);\\
\nabla_\psi\ell_\beta(x;\theta) &= \nabla_\psi\int_\xi Q_\psi(\xi|x)\left(\beta\log\frac{Q_\psi(\xi|x)}{P_{\text{latent}}(\xi)}-\log P_\phi(x|\xi)\right)d\xi\\
&= \int_\xi\left[\nabla_\psi Q_\psi(\xi|x)\left(\beta\log\frac{Q_\psi(\xi|x)}{P_{\text{latent}}(\xi)}-\log P_\phi(x|\xi)\right)+\beta Q_\psi(\xi|x)\cdot\frac{\nabla_\psi Q_\psi(\xi|x)}{Q_\psi(\xi|x)}\right]d\xi\\
&= \int_\xi\nabla_\psi Q_\psi(\xi|x)\left(\beta+\beta\log\frac{Q_\psi(\xi|x)}{P_{\text{latent}}(\xi)}-\log P_\phi(x|\xi)\right)d\xi\\
&= \int_\xi\nabla_\psi Q_\psi(\xi|x)\left(\beta\log\frac{Q_\psi(\xi|x)}{P_{\text{latent}}(\xi)}-\log P_\phi(x|\xi)\right)d\xi\\
&= \mathbb{E}_{\xi\sim Q_\psi(\cdot|x)}\nabla_\psi\log Q_\psi(\xi|x)\left(\beta\log\frac{Q_\psi(\xi|x)}{P_{\text{latent}}(\xi)}-\log P_\phi(x|\xi)\right).
\end{aligned}$$

B Additional Experiments and Details

B.1 Additional Results on Density Estimators in Section 3

The synthetic data of six clusters are illustrated in Figure 5. The sizes for cluster zero through five are 25, 15, 20, 25, 10, 5, respectively. Each cluster is drawn from a spherical Gaussian distribution. The standard errors are 0.5, 0.5, 0.4, 0.4, 0.5, 1.0, respectively.

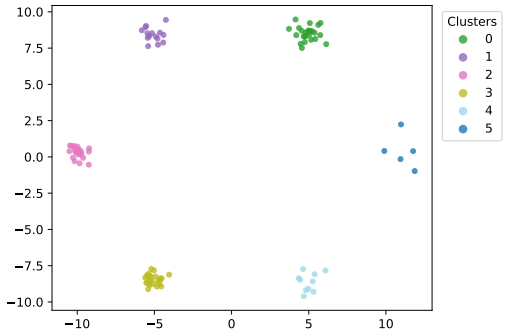
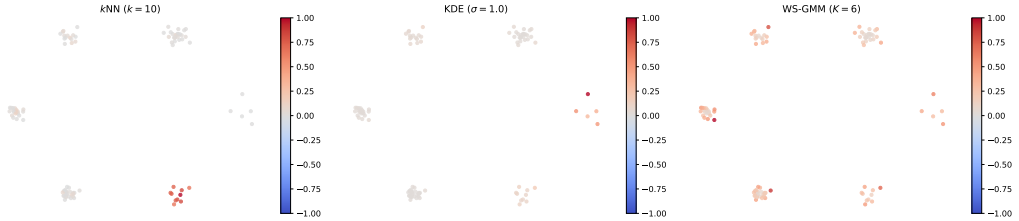
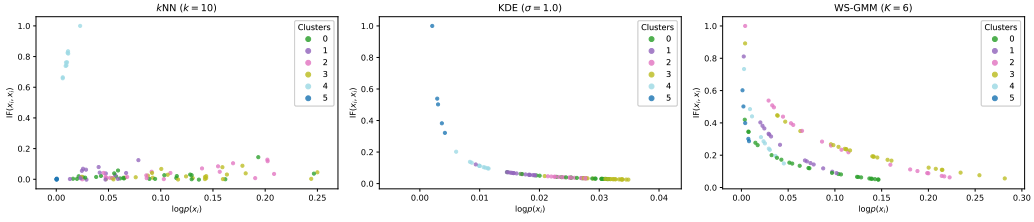


Figure 5: Synthetic data of six clusters in Section 3.1.

We visualize the self influences of all data samples, and compare them to the log likelihood in Figure 6. For k-NN samples from cluster 4 have the highest self influences because the size of this cluster is exactly $k = 10$. For KDE samples in cluster 5 (which is the smallest cluster in terms of number of samples) have the highest self influences. The self influences strictly obey the reverse order of likelihood, which can be derived from (2). For GMM samples far away to cluster centers have high self influences, which can be derived from (5). Samples from clusters 2 and 3 generally have higher self influences because σ_2 and σ_3 are smaller than others.



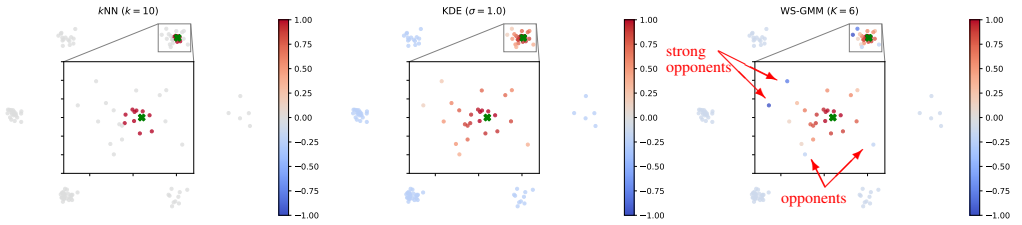
(a) Self influence scores of training samples in different methods. The high self influence samples in k-NN are from a cluster with exactly k samples; those in KDE are from the cluster with the smallest size; and those in GMM are far away to the center of the corresponding cluster.



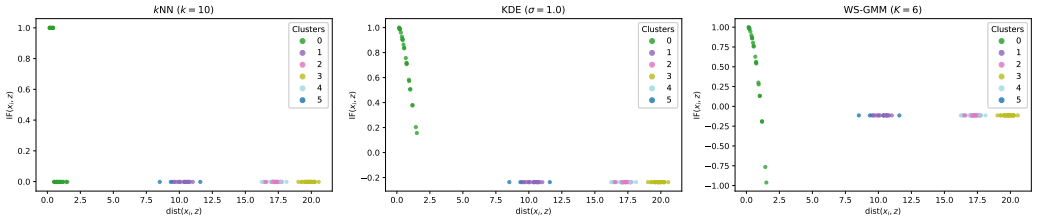
(b) Self influences versus log-likelihood. In k-NN only samples from cluster 4 (which has exatly k points) have large self influences. In KDE and GMM the self influences tend to decrease as the likelihood increases.

Figure 6: Self influences in different density estimators.

We let z be a data point near the center of cluster 0. We then visualize influences of all data samples over z, and compare these influences to the distances between the samples and z in Figure 7. For k-NN the k nearest samples are strong proponents of z, and the rest have little influences over z. For KDE proponents of z are all samples from cluster 0, and the rest have slightly negative influences over z. The influences strictly obey the reverse order of distances to z, which can be derived from (2). For GMM, it is surprising that samples in the same cluster as z can be (even strong) opponents of z. This observation can be mathematically derived from (4). When $\|z - x_i\|^2 > (d+2)\sigma_0^2 + \|z - \mu_0\|^2/\sigma_0^2$, we have $\mathrm{IF}_{X,\text{WS-GMM}}(x_i, z) \lessapprox 0$. When d is large and z is sampled from the mixture $\mathcal{N}(\mu_0, \sigma_0^2 I)$, then with high probability, $\|z - \mu_0\|^2 \approx d\sigma_0^2$. Therefore, the influence of x_i over z is negative with high probability when $\|z - x_i\|^2 \gtrapprox (1 + \sigma_0^2)d + 2\sigma_0^2$.



(a) Influences of training samples over a test sample z (shown as ✖) in different methods. In all cases the strongest proponents are nearest samples. In GMM, surprisingly, samples from the same cluster can be strong opponents.



(b) Influences of training samples over z versus distances to z.

Figure 7: Influences of training samples over a test sample z in different methods. In all methods the strongest proponents are nearest samples. Surprisingly, in GMM strong opponents are also nearby samples.

B.2 Details of Experiments in Section 5

Datasets. We conduct experiments on MNIST and CIFAR-10 (shortened as CIFAR). Because it is challenging to train a successful VAE model on the entire CIFAR dataset, we also train VAE models on each subclass of CIFAR. There are ten subclasses in total, which we name CIFAR_0 though CIFAR_9, and each subclass contains 5k training samples. In the main text, CIFAR-Airplane is CIFAR_0. All CIFAR images are resized to 64×64.

In Section 5.1, we examine influences of all training samples over the first 128 training samples in the trainset. In the unsupervised data cleaning application in Section 5.2, the extra samples are the first 1k samples from EMNIST and CelebA, respectively. In Section 5.3, we randomly select 128 samples from the test set and compute influences of all training samples over these test samples.

Models and hyperparameters. For MNIST, our VAE models are composed of multilayer perceptrons as described by Meehan et al. [25]. In these experiments we let $\beta = 4$ and $d_{\text{latent}} = 128$ unless clearly specified.

For CIFAR and CIFAR subclasses, our VAE models are composed of convolution networks as described by Higgins et al. [16]. We let $\beta = 2, d_{\text{latent}} = 128$ for CIFAR and $\beta = 2, d_{\text{latent}} = 64$ for CIFAR subclass unless clearly specified.

We use stochastic gradient descent to train these VAE models based on a public implementation. [5] In all experiments, we set the batch size to be 64 and train for 1.5M iterations. The learning rates are 1×10^{-4} in MNIST experiments and 3×10^{-4} in CIFAR experiments.

VAE-TracIn settings. In all experiments, we average the loss for $m = 16$ times when computing VAE-TracIn according to (10). We use $C = 30$ (evenly distributed) checkpoints to compute influences in Section 5.1 and Section 5.3. We use the last checkpoint to compute self influences in Section 5.2. For visualization purpose, all self influences are normalized to $[0, 1]$, all influences over test data are normalized to $[-1, 1]$, and all distributions are normalized to densities.

[5] `https://github.com/1Konny/Beta-VAE` (MIT License)

B.3 Sanity Checks for TracIn-VAE

As a sanity check for VAE-TracIn, we examine the frequency that a training sample is the most influential one among all training samples over itself, or formally, the frequency that $i = \arg\max_{1 \le i' \le N} \text{VAE-TracIn}(x_{i'}, x_i)$. Due to computational limits we examine the first 128 training samples. The results for MNIST, CIFAR, and the averaged result for CIFAR subclasses are reported in Table 2. The detailed results for each CIFAR subclass are reported in Table 5. These results indicate that VAE-TracIn can find the most influential training samples in MNIST and CIFAR subclasses.

Table 5: Sanity check on the frequency of a training sample being more influential than other samples over itself. Results on CIFAR subclasses (CIFAR$_i$, $0 \le i \le 9$) are reported.

i	0	1	2	3	4	5	6	7	8	9
Top-1 scores	1.000	1.000	0.992	1.000	0.984	1.000	1.000	1.000	1.000	1.000

We then conduct an additional sanity check for VAE-TracIn on MNIST. For two training samples $x_{\text{major}} = x_i$ and $x_{\text{minor}} = x_j$, we synthesize a new sample $\hat{x} = \alpha x_{\text{major}} + (1-\alpha)x_{\text{minor}}$, where $\alpha = 0.75$. Then, $\hat{x}$ is very similar to x_{major} but the minor component x_{minor} can also be visually recognized. For each pair of different labels, we obtain x_{major} and x_{minor} by randomly picking one sample within each class. The entire 90 samples are shown in Figure 8. We expect a perfect instance-based interpretation should indicate x_i and x_j have very high influences over $\hat{x}$. We report quantiles of the 90 ranks of x_{major} and x_{minor} sorted by influences over $\hat{x}$ in Table 6. We then compute the frequency that x_{major} is exactly the strongest proponent of $\hat{x}$, namely the top-1 score of the major component. We compare the results to a baseline model that finds nearest neighbours in a perceptual autoencoder latent space (PAE-NN, [25, 37]). Although VAE-TracIn does not detect x_{major} as well as PAE-NN, it still has reasonable results, and performs much better in detecting x_{minor}. The results indicate that VAE-TracIn can capture potentially influential components.

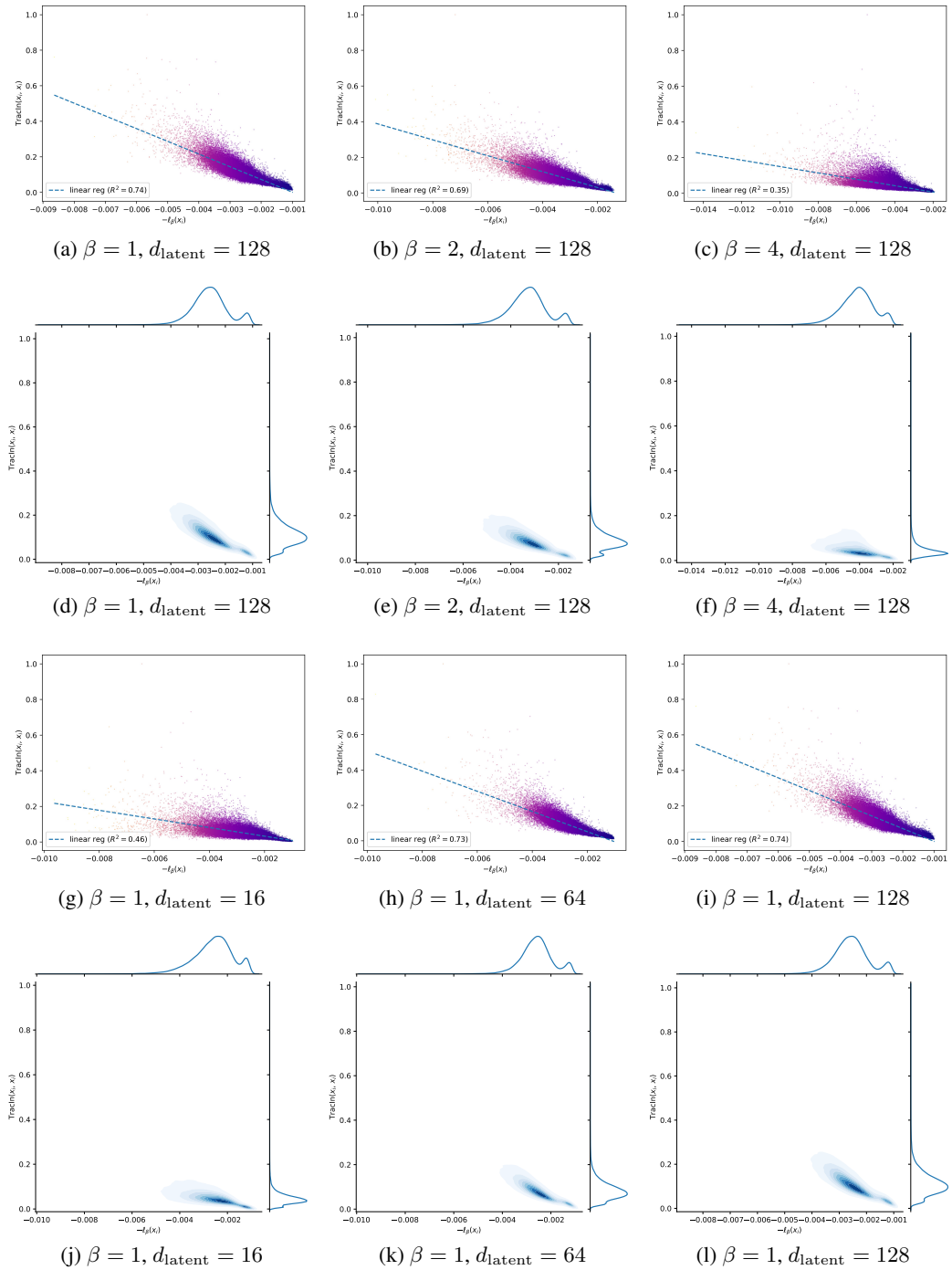
Figure 8: Synthesized samples $\hat{x} = \alpha x_{\text{major}} + (1-\alpha)x_{\text{minor}}$, where $\alpha = 0.75$.

Table 6: Quantiles of ranks of x_{major} and x_{minor} sorted by influences over $\hat{x}$, and top-1 scores of the major components. We let 0 to be the highest rank.

Method	rank(x_{major}) quantiles			Top-1	rank(x_{minor}) quantiles		
	25%	50%	75%	scores	25%	50%	75%
PAE-NN	0	0	1	0.633	6943	13405	29993
VAE-TracIn ($d_{\text{latent}} = 64$)	0	2	146	0.422	1097	5206	10220
VAE-TracIn ($d_{\text{latent}} = 96$)	0	1	44	0.456	1372	4283	15319
VAE-TracIn ($d_{\text{latent}} = 128$)	0	1	18	0.467	1203	6043	13873

B.4 Self Influences (MNIST)

In MNIST experiments, we compare self influences and losses across different hyperparameters. The scatter and density plots are shown in Figure 9. We fit linear regression models to these points and report R^2 scores. In all settings high self influence samples have large losses. We find R^2 is larger under high latent dimensions or smaller β.



(a) $\beta = 1$, $d_{\text{latent}} = 128$ (b) $\beta = 2$, $d_{\text{latent}} = 128$ (c) $\beta = 4$, $d_{\text{latent}} = 128$

(d) $\beta = 1$, $d_{\text{latent}} = 128$ (e) $\beta = 2$, $d_{\text{latent}} = 128$ (f) $\beta = 4$, $d_{\text{latent}} = 128$

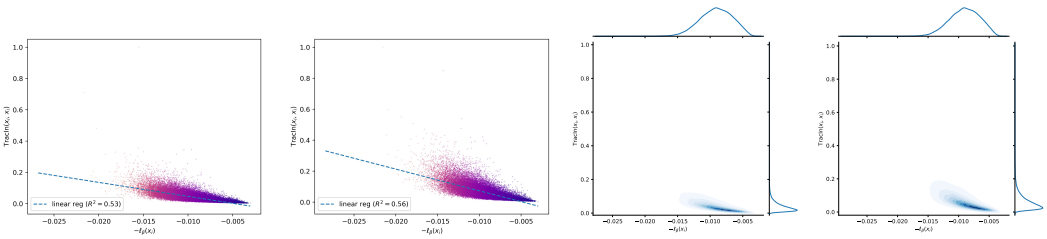
(g) $\beta = 1$, $d_{\text{latent}} = 16$ (h) $\beta = 1$, $d_{\text{latent}} = 64$ (i) $\beta = 1$, $d_{\text{latent}} = 128$

(j) $\beta = 1$, $d_{\text{latent}} = 16$ (k) $\beta = 1$, $d_{\text{latent}} = 64$ (l) $\beta = 1$, $d_{\text{latent}} = 128$

Figure 9: Scatter and density plots of self influences versus negative losses of all training samples in MNIST. The linear regressors show that high self influence samples have large losses.

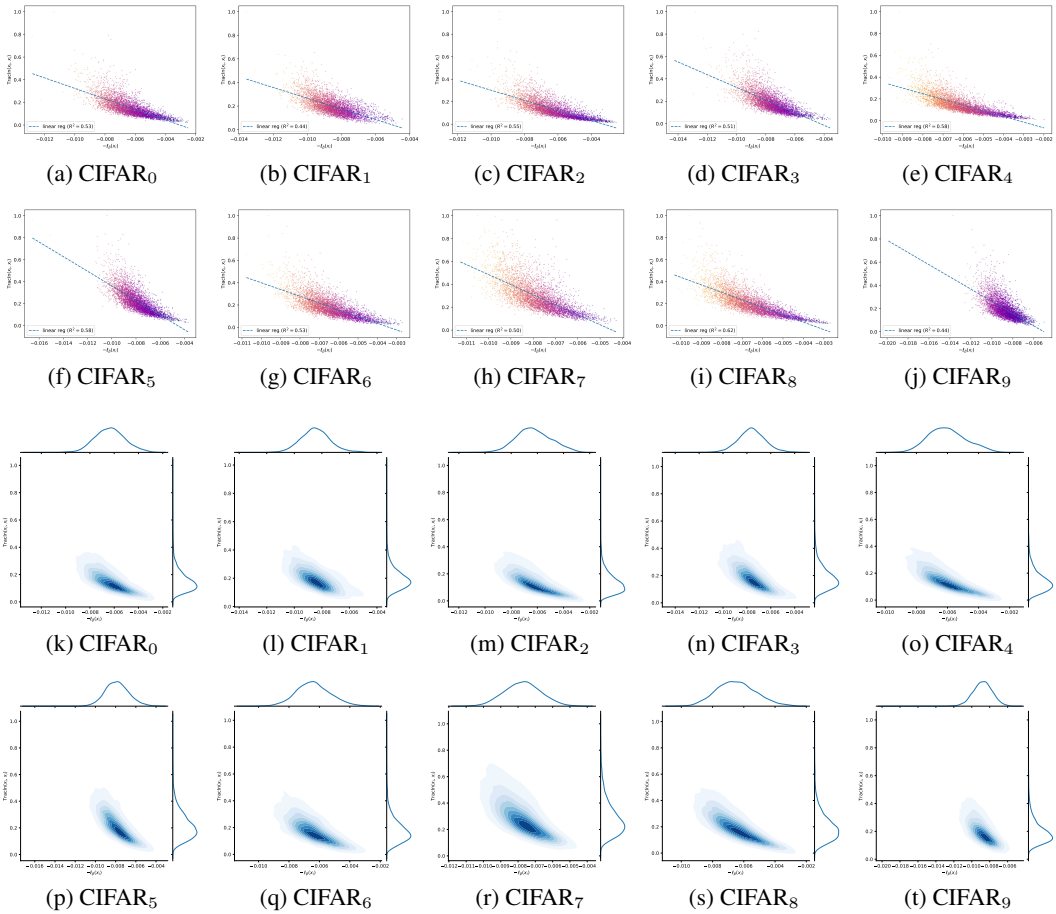
B.5 Self Influences (CIFAR)

In CIFAR and CIFAR subclass experiments, we compare self influences and losses across different hyperparameters. Similar to Appendix B.4, we demonstrate scatter and density plots, and report R^2 scores of linear regression models fit to these data. Comparisons on CIFAR are shown in Figure 10, and CIFAR subclasses in Figure 11. In all settings high self influence samples have large losses. We then visualize high and low self influence samples from each CIFAR subclass in Figure 12 and Figure 13, respectively.



(a) $d_{\text{latent}} = 64$ (b) $d_{\text{latent}} = 128$ (c) $d_{\text{latent}} = 64$ (d) $d_{\text{latent}} = 128$

Figure 10: Scatter and density plots of self influences versus negative losses of all training samples in CIFAR. The linear regressors show that high self influence samples have large losses.



(a) CIFAR_0 (b) CIFAR_1 (c) CIFAR_2 (d) CIFAR_3 (e) CIFAR_4

(f) CIFAR_5 (g) CIFAR_6 (h) CIFAR_7 (i) CIFAR_8 (j) CIFAR_9

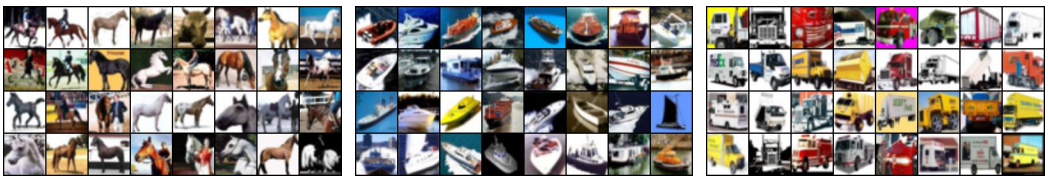
(k) CIFAR_0 (l) CIFAR_1 (m) CIFAR_2 (n) CIFAR_3 (o) CIFAR_4

(p) CIFAR_5 (q) CIFAR_6 (r) CIFAR_7 (s) CIFAR_8 (t) CIFAR_9

Figure 11: Scatter and density plots of self influences versus negative losses of all training samples in each CIFAR subclass. The linear regressors show that high self influence samples have large losses.

(a) CIFAR$_1$ (b) CIFAR$_2$ (c) CIFAR$_3$

(d) CIFAR$_4$ (e) CIFAR$_5$ (f) CIFAR$_6$

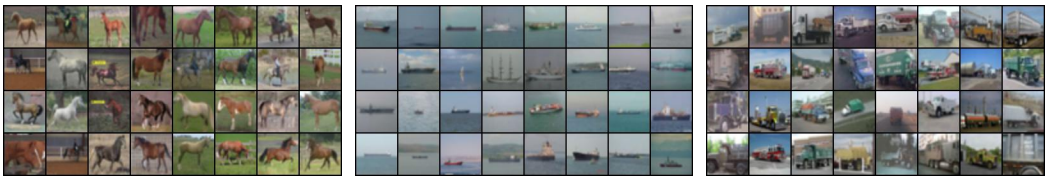


(g) CIFAR$_7$ (h) CIFAR$_8$ (i) CIFAR$_9$

Figure 12: High self influence samples in each CIFAR subclass. These samples are visually high-contrast and bright.

(a) CIFAR$_1$ (b) CIFAR$_2$ (c) CIFAR$_3$

(d) CIFAR$_4$ (e) CIFAR$_5$ (f) CIFAR$_6$

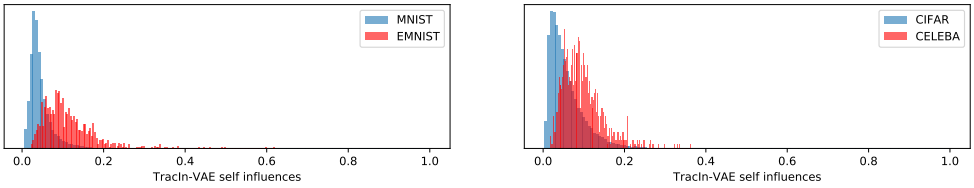


(g) CIFAR$_7$ (h) CIFAR$_8$ (i) CIFAR$_9$

Figure 13: Low self influence samples in each CIFAR subclass. These samples are visually similar in shape or background.

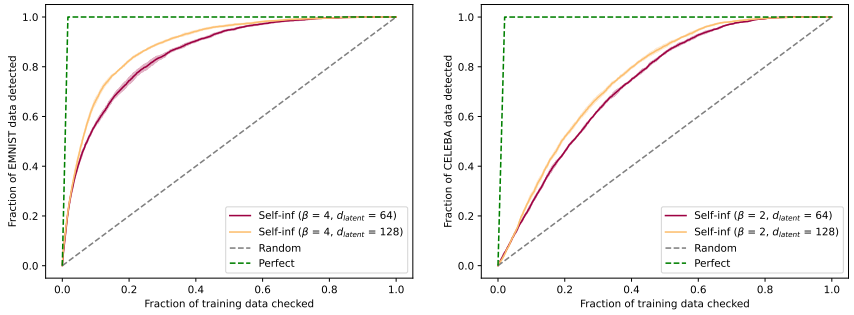
B.6 Application on Unsupervised Data Cleaning

We plot the distribution of self influences of extra samples (EMNIST or CelebA) and original samples (MNIST or CIFAR) in Figure 14. We plot the detection curves in Figure 15, where the horizontal axis is the fraction of all samples checked when they are sorted in the self influence order, and the vertical axis is the fraction of extra samples found. The area under these detection curves (AUC) are reported in Table 7. These experiments are repeated five times to reduce randomness. The results indicate that extra samples have higher self influences than original samples. This justifies the potential to apply VAE-TracIn to unsupervised data cleaning.



(a) EMNIST versus MNIST (b) CelebA versus CIFAR

Figure 14: Distributions of self influences of 1k extra samples versus samples from the original dataset. Distributions are normalized as densities for better visualization. It is shown that extra samples have higher self influences.



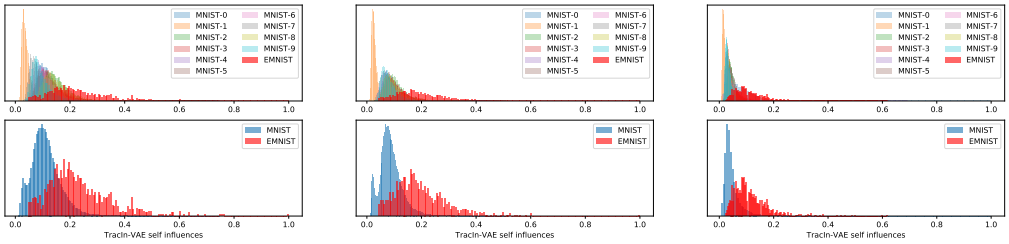
(a) EMNIST versus MNIST (b) CelebA versus CIFAR

Figure 15: Detection curves (fraction of extra samples detected versus fraction of training data checked in the self influence order) with standard errors. It is shown that extra samples can be detected by sorting self influences.

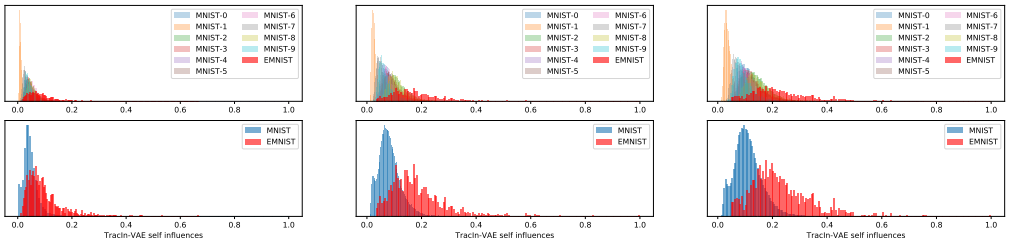
Table 7: Mean area-under-curve (AUC) $\pm$ standard errors of detection curves in Figure 15. A higher AUC indicates extra samples can be better detected by sorting self influences, where AUC≈ 1 implies perfect detection and AUC ≈ 0.5 implies random selection. The results indicate detection on the simple MNIST + EMNIST datasets is better than the more complicated CIFAR + CelebA datasets. In addition, a higher d_{latent} leads to slightly better AUC.

Original dataset	Extra samples	d_{latent}	AUC
MNIST	EMNIST	64	0.858$\pm$0.003
MNIST	EMNIST	128	0.887$\pm$0.002
CIFAR	CelebA	64	0.735$\pm$0.002
CIFAR	CelebA	128	0.760$\pm$0.001

We then conduct extensive comparison among different hyperparameters under the MNIST + EMNIST setting. Distributions of self influences are shown in Figure 16 and detection curves are shown in Figure 17. In all settings, extra samples have higher self influences than original samples. Increasing β or d_{latent} can slightly improve detection.



(a) $\beta = 1$, $d_{\text{latent}} = 128$ (b) $\beta = 2$, $d_{\text{latent}} = 128$ (c) $\beta = 4$, $d_{\text{latent}} = 128$



(d) $\beta = 1$, $d_{\text{latent}} = 16$ (e) $\beta = 1$, $d_{\text{latent}} = 64$ (f) $\beta = 1$, $d_{\text{latent}} = 128$

Figure 16: Distribution of self influences of extra sample and samples from the original MNIST dataset. In all settings, extra samples have higher self influences than original samples.

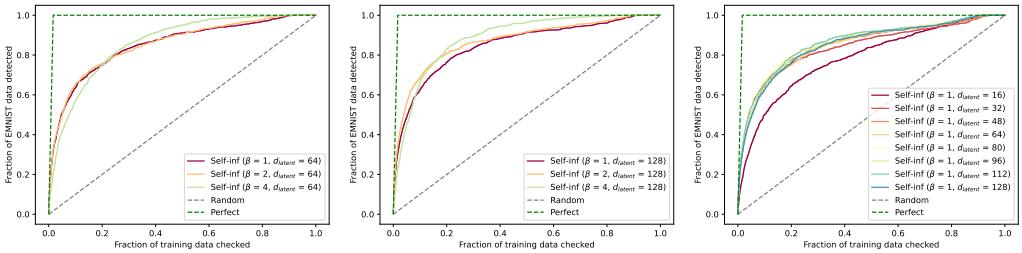


Figure 17: Detection curves (EMNIST versus MNIST). It is shown that extra samples can be detected by sorting self influences. Increasing β or d_{latent} can slightly improve detection.

B.7 Influences over Test Data (MNIST)

In Figure 18, we plot the distributions of influences of training zeroes (red distributions) and non-zeroes (blue distributions) over test zeroes. Full results on all classes are shown in Figure 19. For most labels including 0, 2, 4, 6, 7, and 9, the strongest proponents and opponents are very likely from the same class. For the rest of the labels including 1, 3, 5, and 8, the strongest opponents seem equally likely from the same or a different class.

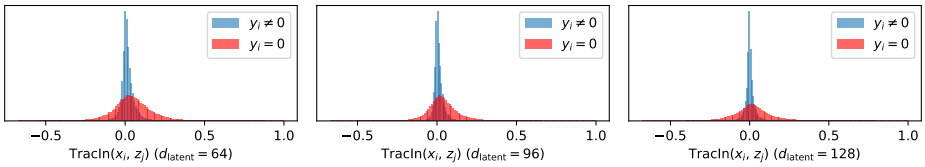
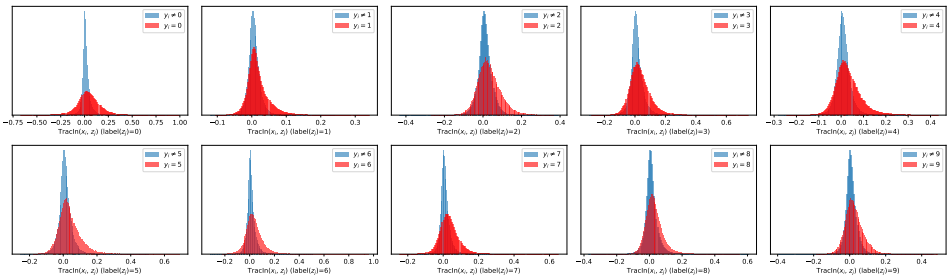
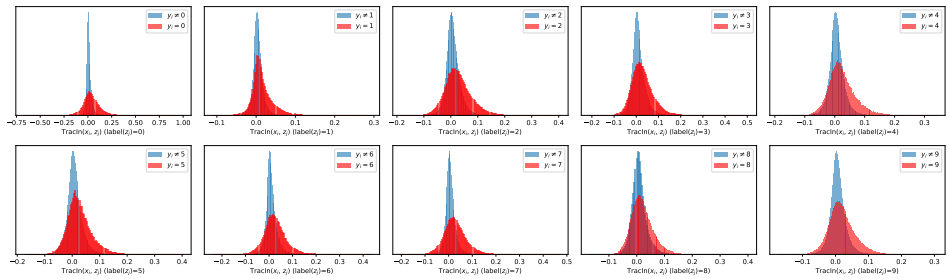


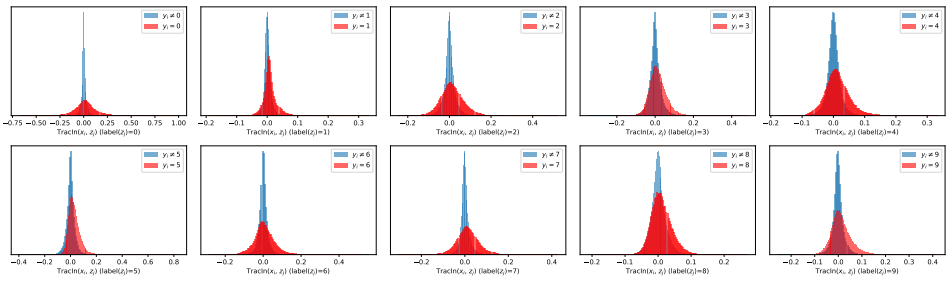
Figure 18: Distributions of influences of training samples (x_i) over test zeroes (z_j). The red distributions are training zeroes and the blue distributions are training non-zeroes. It is shown that a large proportion of strong proponents and opponents of test zeroes are training zeroes.



(a) $\beta = 4$, $d_{\text{latent}} = 64$



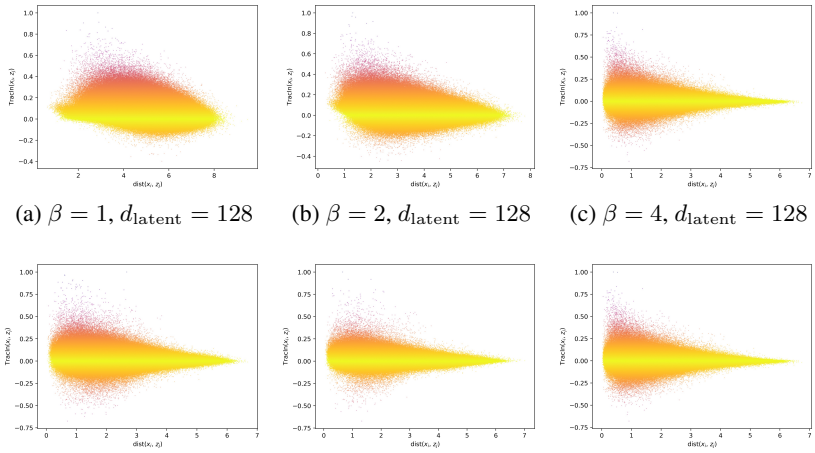
(b) $\beta = 4$, $d_{\text{latent}} = 96$



(c) $\beta = 4$, $d_{\text{latent}} = 128$

Figure 19: Distributions of influences of training samples (x_i) over test samples (z_j). The red distributions are x_i in the same class as z_j, and the blue distributions are x_i in a different class as z_j. For most labels the strongest proponents and opponents are very likely from the same class. For the rest the strongest opponents seem equally likely from the same or a different class.

We then compare the influences of training over test samples to the distances between them in the latent space in Figure 20. We observe that both proponents and opponents are very close to test samples in the latent space, which indicates strong similarity between them. This phenomenon is more obvious when $\beta = 4$.

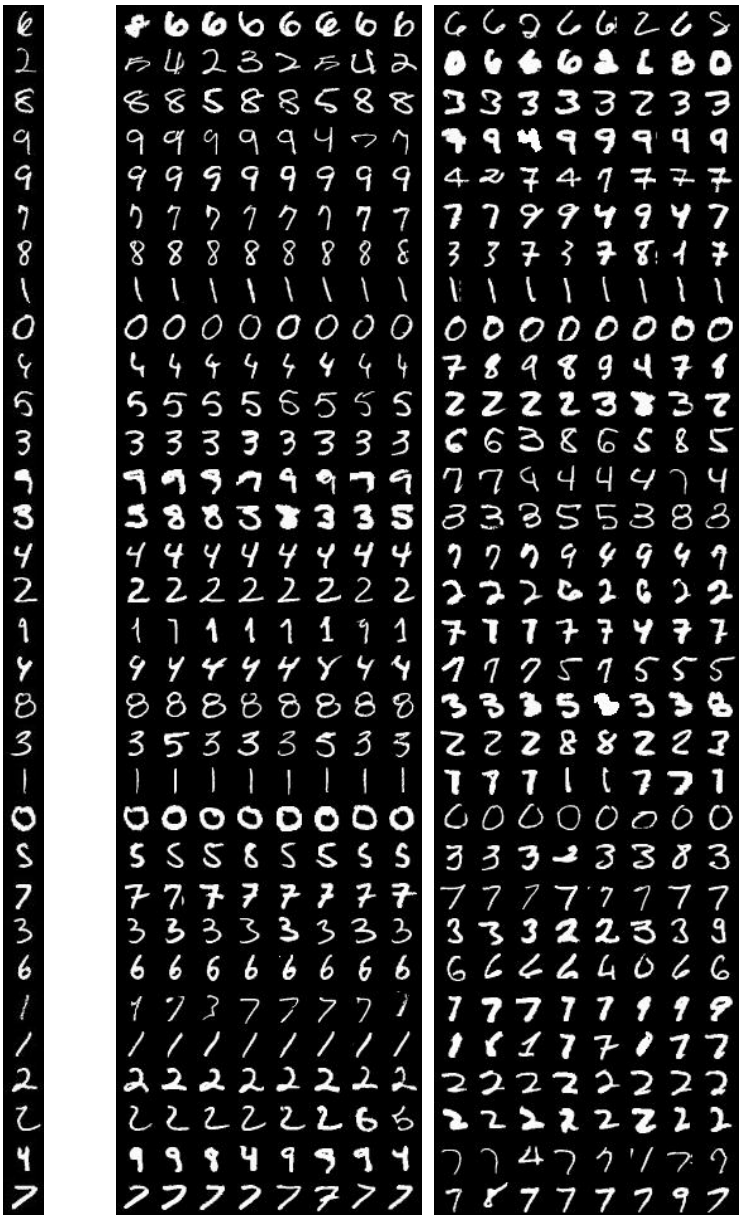


(a) $\beta = 1$, $d_{\text{latent}} = 128$ (b) $\beta = 2$, $d_{\text{latent}} = 128$ (c) $\beta = 4$, $d_{\text{latent}} = 128$

(d) $\beta = 4$, $d_{\text{latent}} = 64$ (e) $\beta = 4$, $d_{\text{latent}} = 96$ (f) $\beta = 4$, $d_{\text{latent}} = 128$

Figure 20: Influences of training over test samples versus pairwise distances between them in the latent space. It is shown that both proponents and opponents are very close to test samples in the latent space especially when $\beta = 4$.

We display the first 32 test samples in MNIST, their strongest proponents, and their strongest opponents in Figure 21. The strongest proponents look very similar to test samples. The strongest opponents are often the same digit but are visually very different. For instance, many strong opponents have very different thickness, shapes, or styles.

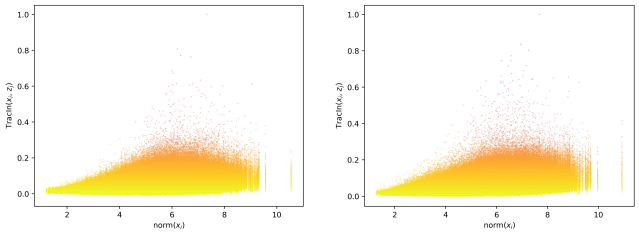


(a) z (b) Strongest proponents (c) Strongest opponents

Figure 21: Test samples z from MNIST, their strongest proponents, and their strongest opponents. The strongest proponents look very similar to test samples, while the strongest opponents are often the same digit but are visually very different.

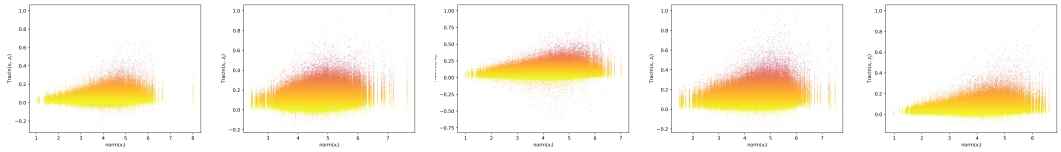
B.8 Influences over Test Data (CIFAR)

We compare the influences of training over test samples to the norms of training samples in the latent space. Results for CIFAR are shown in Figure 22 and results for CIFAR subclasses are shown in Figure 23. We observe that strong proponents tend to have very large norms. This indicates they are high-contrast or very bright samples. This phenomenon occurs to CIFAR and all CIFAR subclasses.

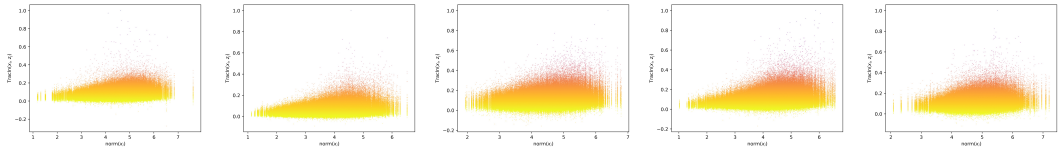


(a) $d_{\text{latent}} = 64$ (b) $d_{\text{latent}} = 128$

Figure 22: Influences of training samples over test samples (CIFAR) versus norms of training samples in the latent space. It is shown that strong proponents have large norms.



(a) CIFAR_0 (b) CIFAR_1 (c) CIFAR_2 (d) CIFAR_3 (e) CIFAR_4



(f) CIFAR_5 (g) CIFAR_6 (h) CIFAR_7 (i) CIFAR_8 (j) CIFAR_9

Figure 23: Influences of training samples over test samples (CIFAR subclass) versus norms of training samples in the latent space. It is shown that strong proponents have large norms.

For 128 test samples in each CIFAR subclass, we report the statistics of the latent space norms of their strongest proponents, strongest opponents, and all training samples in Table 8.

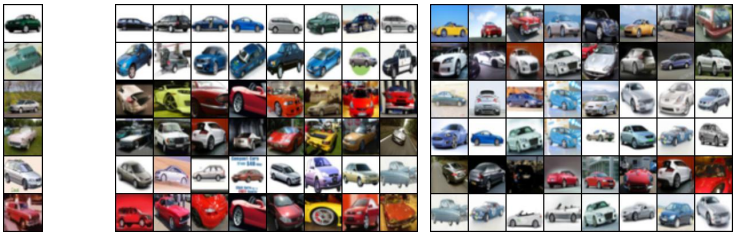
Table 8: The means $\pm$ standard errors of latent space norms of training samples in CIFAR subclasses. Strong proponents tend to have large norms.

Dataset	top-0.1% strong proponents	top-0.1% strong opponents	all training samples
CIFAR_0	4.73 ± 0.78	4.26 ± 0.91	4.07 ± 0.83
CIFAR_1	5.30 ± 0.71	4.54 ± 0.65	4.65 ± 0.64
CIFAR_2	4.89 ± 0.78	4.18 ± 0.88	4.09 ± 0.93
CIFAR_3	5.09 ± 0.75	4.47 ± 0.78	4.42 ± 0.79
CIFAR_4	5.06 ± 0.72	3.96 ± 1.00	4.01 ± 0.89
CIFAR_5	5.25 ± 0.75	4.33 ± 0.94	4.54 ± 0.81
CIFAR_6	4.73 ± 0.66	3.99 ± 0.80	3.95 ± 0.82
CIFAR_7	5.11 ± 0.76	4.42 ± 0.73	4.43 ± 0.74
CIFAR_8	5.10 ± 0.72	4.19 ± 0.88	4.22 ± 0.84
CIFAR_9	5.48 ± 0.62	4.62 ± 0.69	4.79 ± 0.64

For each CIFAR subclass, we display test samples, their strongest proponents, and their strongest opponents in Figures 24 $\sim$ 33. The strongest proponents seem to match the color of test samples in terms of background and the object. In addition, they tend to have the same but brighter colors.

(a) z (b) Strongest proponents (c) Strongest opponents

Figure 24: Test samples z from CIFAR_0, their strongest proponents, and their strongest opponents.

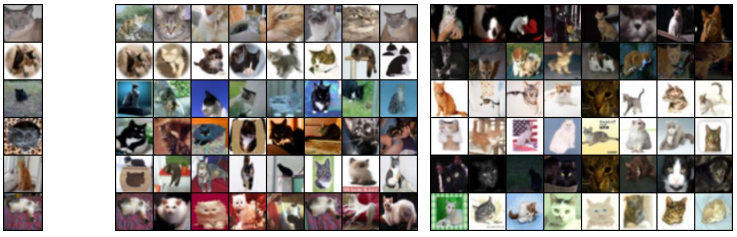


(a) z (b) Strongest proponents (c) Strongest opponents

Figure 25: Test samples z from CIFAR_1, their strongest proponents, and their strongest opponents.

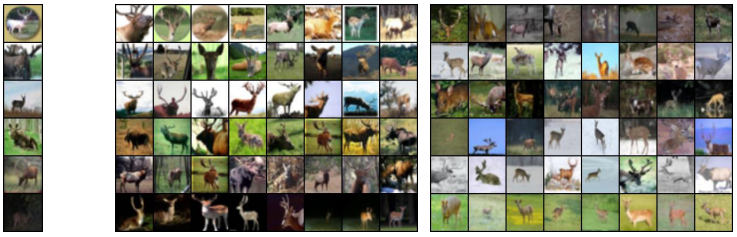
(a) z (b) Strongest proponents (c) Strongest opponents

Figure 26: Test samples z from CIFAR_2, their strongest proponents, and their strongest opponents.



(a) z (b) Strongest proponents (c) Strongest opponents

Figure 27: Test samples z from CIFAR_3, their strongest proponents, and their strongest opponents.

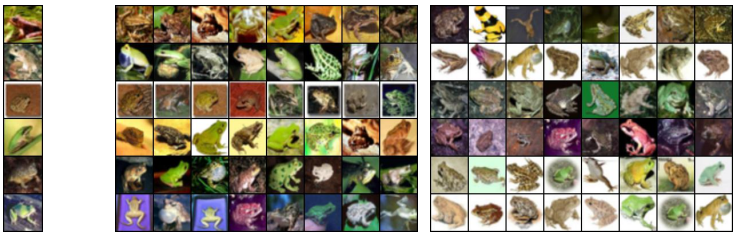


(a) z (b) Strongest proponents (c) Strongest opponents

Figure 28: Test samples z from CIFAR$_4$, their strongest proponents, and their strongest opponents.

(a) z (b) Strongest proponents (c) Strongest opponents

Figure 29: Test samples z from CIFAR$_5$, their strongest proponents, and their strongest opponents.

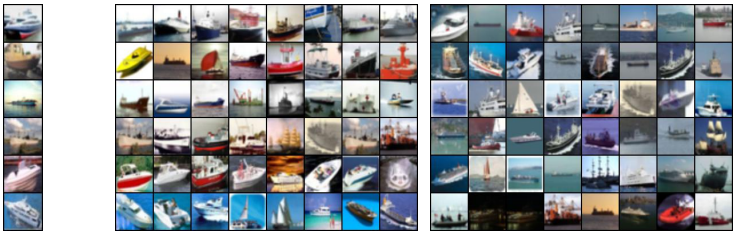


(a) z (b) Strongest proponents (c) Strongest opponents

Figure 30: Test samples z from CIFAR$_6$, their strongest proponents, and their strongest opponents.

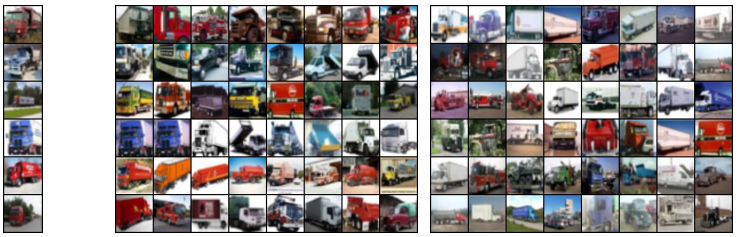
(a) z (b) Strongest proponents (c) Strongest opponents

Figure 31: Test samples z from CIFAR$_7$, their strongest proponents, and their strongest opponents.



(a) z (b) Strongest proponents (c) Strongest opponents

Figure 32: Test samples z from CIFAR_8, their strongest proponents, and their strongest opponents.



(a) z (b) Strongest proponents (c) Strongest opponents

Figure 33: Test samples z from CIFAR_9, their strongest proponents, and their strongest opponents.