Restoring Task-Relevant Information in Synthetic Data: A Small-Scale V-Information View

Sid Bharthulwar [1]

Abstract

This paper investigates synthetic data generation as a mechanism for restoring or reformatting task-relevant information that is obscured or unusable for a specific, computationally bounded learner. We conduct a small-scale, controlled experiment on CIFAR-10, involving pixel permutation to corrupt data, a Convolutional Autoencoder (Conv-AE) synthesizer for information restoration, and a downstream CNN learner. Framed through V-Information, which quantifies information accessible to such a learner, empirical results demonstrate that while permutation drastically reduces usable V-Information, the synthesizer partially restores it, leading to significant performance recovery. We further explore how model capacities interact with this process, finding learner capacity beneficial only when usable information is present. This highlights computation's role in making latent information accessible, a principle highly relevant to current synthetic data practices in capabilities and alignment of foundation models.

1. Introduction

The proliferation of synthetic data in machine learning, pivotal for tasks ranging from instruction-tuning and aligning Large Language Models (LLMs) (Wang et al., 2023; Taori et al., 2023; Bai et al., 2022; **?**) to augmenting computer vision datasets (Geng et al., 2023), presents an intriguing paradox. Classical information theory, through the Data Processing Inequality (DPI), dictates that no deterministic data processing can increase Shannon mutual information with respect to a target variable (Cover & Thomas, 2006). Yet, models trained with synthetic data often exhibit enhanced performance, suggesting data becomes ”more informative.” Our central research question, aligning with the MOSS workshop's focus on scientific understanding from small-scale experiments, is: how does synthetic data improve learning if it cannot create new information? We hypothesize that its primary role, especially in complex systems like LLMs, is often not to introduce fundamentally new information, but rather to restore, reformat, or filter existing, latent task-relevant information into a structure that is more usable by a specific, computationally bounded learning algorithm. This reformatting can involve making implicit knowledge explicit (e.g., generating chain-of-thought reasoning (Mukherjee et al., 2023)) or aligning data with desired behaviors (e.g., through AI feedback (Bai et al., 2022)).

To formalize and quantify this concept of ”usable information,” we leverage the V-Information framework introduced by Xu et al. (2020). V-Information, denoted $I_{\mathcal{V}}(X;Y)$, measures the mutual information between an input X and a target Y that is accessible to a learner constrained to a specific hypothesis class $\mathcal{V}$, as defined by $I_{\mathcal{V}}(X;Y) = \max_{f\in\mathcal{V}} I(f(X);Y)$. Crucially, unlike Shannon information, $I_{\mathcal{V}}$ can be increased by data transformations g if $Z = g(X)$ aligns the data better with the learner's capabilities and inductive biases. For instance, permuting image pixels (Shukla, 2019) preserves $I(X;Y)$ but can catastrophically reduce $I_{\mathcal{V}_{\text{CNN}}}(X;Y)$ for a Convolutional Neural Network (CNN) by destroying the spatial structure vital to its inductive biases. Similarly, raw web text may contain information for a task, but an LLM might require it to be formatted as explicit instruction-response pairs or demonstrations of step-by-step reasoning to effectively learn desired behaviors, thus increasing $I_{\mathcal{V}_{\text{LLM}}}$.

This paper aims to empirically demonstrate and quantify this information restoration/reformatting phenomenon through a minimalistic, reproducible setup. We model a scenario where original ”clean” data X^* undergoes a corruption process C

[1] Department of Computer Science, Harvard University, Cambridge, MA, United States. Correspondence to: Sid Bharthulwar <sbharthulwar@college.harvard.edu>.

Accepted at Methods and Opportunities at Small Scale (MOSS), ICML 2025, Vancouver, Canada.

(pixel permutation) to produce observed data X_{obs}, which has low $I_{\mathcal{V}_{\text{CNN}}}$. A synthesizer model G (a Conv-AE) is then trained to process X_{obs} and produce synthetic data X_{synth}. Our primary goal is to show that $I_{\mathcal{V}_{\text{CNN}}}(G(X_{\text{obs}}); Y) > I_{\mathcal{V}_{\text{CNN}}}(X_{\text{obs}}; Y)$, and to investigate how the capacities of both G and the downstream learner $L \in \mathcal{V}_{\text{CNN}}$ influence this restoration and final task performance. This elementary synthetic task, while in computer vision, serves as an analogy for how SOTA LLM training pipelines use synthetic data: a powerful ”synthesizer” (e.g., GPT-4) refines or generates data that a ”learner” (e.g., a smaller open-source LLM) can more effectively utilize.

2. Methodology: Data Corruption, Synthesis, and Learning

Our experimental protocol is designed for clarity and reproducibility with limited computational resources, involving three distinct stages: data corruption, synthesis-based information restoration, and downstream task learning and evaluation on the CIFAR-10 dataset (Krizhevsky, 2009).

First, in the **Data Corruption (**C**)** stage, the original, ”clean” CIFAR-10 images, denoted X^*, serve as our ground truth for pristine information. We apply a fixed, deterministic pixel permutation $P(\cdot)$ to each image $X^{*(i)}$ in the dataset to generate corrupted data $X_{\text{obs}}^{(i)} = P(X^{*(i)})$. The permutation P is generated once using a fixed seed and shuffles the flattened pixel vector (e.g., $3 \times 32 \times 32 = 3072$ elements for CIFAR-10), with the same permutation applied to all images. This specific corruption drastically reduces the $I_{\mathcal{V}}(X_{\text{obs}}; Y)$ for standard CNN learners $L \in \mathcal{V}_{\text{CNN}}$ because it breaks the spatial locality and local correlations that CNNs’ inductive biases rely upon (Shukla, 2019). While the Shannon mutual information $I(X_{\text{obs}}; Y)$ remains theoretically unchanged (as the permutation is invertible), the information becomes ”unusable” for the CNN.

Second, the **Synthesizer (**G**) - Information Restoration** stage employs a Convolutional Autoencoder (Conv-AE) as the synthesizer G. The Conv-AE architecture consists of a convolutional encoder mapping the input image to a lower-dimensional latent representation, and a deconvolutional (or upsampling) decoder that reconstructs the image from this latent code. The Conv-AE is tasked with reconstructing the original, uncorrupted image X^* given the permuted image X_{obs} as input. Its training objective is to minimize the Mean Squared Error (MSE) between the reconstructed image $G_{\theta_G}(X_{\text{obs}})$ and the original image X^*: $\min_{\theta_G} \mathbb{E}_{(X^*, X_{\text{obs}})} \|X^* - G_{\theta_G}(X_{\text{obs}})\|_2^2$. Crucially, following a few-shot or limited supervision paradigm, the Conv-AE is trained on only a small subset of paired data $(X_{\text{obs}}^{(i)}, X^{*(i)})$ from the training set, mimicking scenarios where perfect restoration knowledge or extensive paired data is scarce. After training, the synthesizer G is used to process the entire (or a relevant subset of) permuted training dataset X_{obs} to produce a synthetically restored dataset $X_{\text{synth}} = \{G(X_{\text{obs}}^{(i)})\}$. The success of this stage is measured by whether $I_{\mathcal{V}_{\text{CNN}}}(X_{\text{synth}}; Y) > I_{\mathcal{V}_{\text{CNN}}}(X_{\text{obs}}; Y)$.

Third, in the **Learner (**L**) - Downstream Task Evaluation** stage, a standard CNN architecture (adapted for CIFAR-10 with 3 input channels) is used as the downstream learner L, belonging to the model family $\mathcal{V}_{\text{CNN}}$. The learner L is trained via Empirical Risk Minimization (ERM) with a Cross-Entropy loss function to perform 10-class image classification. This training is conducted separately on three different datasets: (1) the original, clean CIFAR-10 training set (X^*); (2) the permuted CIFAR-10 training set (X_{obs}); and (3) the synthetically restored training set generated by G (X_{synth}). To explore the interplay of model capacity and information usability, the capacity of the learner CNN (C_L) is varied using a multiplier $\alpha_L \in [0.25, 2.0]$, which might adjust the number of filters, layer widths, or depth of the CNN architecture. The primary evaluation metric is the final test accuracy achieved by the learner L on the original, clean CIFAR-10 test set, measuring how well the information available in each training dataset translates to generalization performance.

3. Results and Discussion

Our experiments, conducted on CIFAR-10 using the described small-scale protocol, yield compelling evidence for the role of synthetic data in restoring task-relevant information and highlight the critical impact of model capacity on leveraging this information. The key findings are visualized in Figure 1.

The classification performance of the downstream CNN learner, as depicted in Figure 1 (Right Panel), exhibits a clear hierarchy contingent on the training data’s nature. Training the CNN on the original, clean CIFAR-10 images (X^*) yields the highest test accuracy (blue solid line), establishing our upper-bound benchmark where task-relevant information is fully and directly accessible to a learner with appropriate inductive biases. Conversely, when the CNN is trained on pixel-permuted CIFAR-10 images (X_{obs}), the test accuracy plummets dramatically (orange dashed line). This aligns with our hypothesis: the permutation corruption C renders the spatial information critical for the CNN’s convolutional filters unusable.

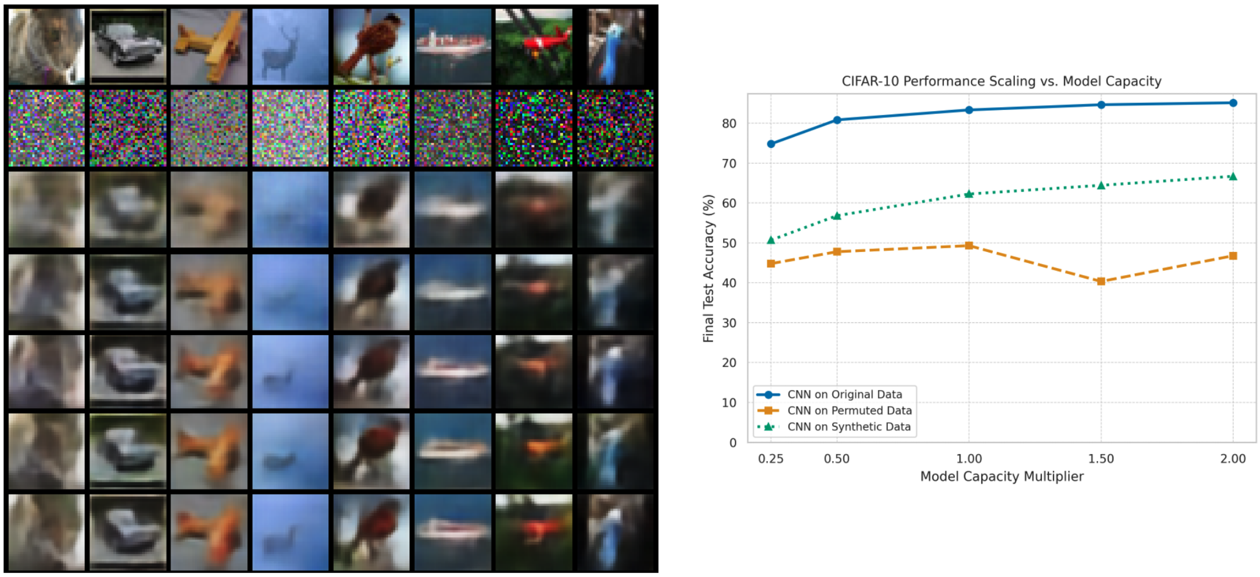

Figure 1. Experimental results on CIFAR-10. **Left Panel:** Example images illustrating Original CIFAR-10 samples (Top Row), Pixel-Permuted versions (Second Row, appearing as random noise), and Synthesized/Reconstructed images by the Conv-AE (Subsequent Rows, showing iterative refinement or different qualities), demonstrating partial restoration of objects and spatial coherence. **Right Panel:** CIFAR-10 performance scaling (Final Test Accuracy %) versus Learner (CNN) model capacity multiplier. It compares models trained on Original data (blue solid line with circles), Permuted data (orange dashed line with squares), and Synthesized data (green dotted line with triangles). The plot demonstrates the performance hierarchy (Original $\gg$ Synthesized $>$ Permuted) and that learner capacity is beneficial primarily when usable information is present, mirroring how powerful ”teacher” models generate more ”usable” data for smaller ”student” LLMs.

While the Shannon information content regarding the label is preserved, the V-Information for a CNN, $I_{\mathcal{V}_{\text{CNN}}}(X_{\text{obs}}; Y)$, is severely diminished. Most importantly, training the CNN on the data reconstructed by the Convolutional Autoencoder ($X_{\text{synth}} = G(X_{\text{obs}})$) results in a significant recovery of performance compared to the permuted data (green dotted line). The accuracy achieved on synthesized data is substantially higher than on permuted data, though it does not fully reach the level of the original data. This demonstrates that the synthesizer G is indeed capable of restoring a significant portion of the task-relevant information that was obscured by the permutation. The performance gap between the ”Synthesized” curve and the ”Original” curve can be attributed to imperfections in the autoencoder’s reconstruction, likely due to its own limited capacity C_G, the limited (few-shot) training data provided for its training, or the inherent difficulty of achieving perfect reconstruction from severely corrupted inputs. This process is analogous to how SOTA LLMs like GPT-4 are used to generate instruction-following data (e.g., for Alpaca (Taori et al., 2023)) or complex reasoning traces (e.g., for Orca (Mukherjee et al., 2023)), which are then used to train smaller, more specialized models. The powerful ”synthesizer” LLM reformats or elucidates information in a way that increases its V-Information for the ”learner” LLM.

The interaction between learner model capacity (C_L, varied by the ”Model Capacity Multiplier” α_L on the x-axis of Figure 1, Right Panel) and the different data conditions further illuminates these dynamics. For both the original clean data and the synthesized data, the test accuracy of the CNN learner generally improves as its capacity increases. This follows a power-law-like trend, typical of learning curves where more capable models can better fit the data and extract more complex patterns, provided the information is present and usable. This indicates that as C_L increases, the learner can leverage more of the available $I_{\mathcal{V}_{\text{CNN}}}$ from these datasets. The synthesized data curve, while lower in absolute terms, still shows a consistent positive scaling trend, suggesting that the restored information is indeed useful and can be further exploited by larger learners. In stark contrast, the performance curve for the CNN trained on permuted data is significantly lower and shows much less pronounced, and somewhat erratic, scaling with learner capacity (note the dip around a capacity multiplier of 1.50). This observation is critical: it suggests that when the V-Information $I_{\mathcal{V}_{\text{CNN}}}(X_{\text{obs}}; Y)$ is extremely low due to severe corruption, additional model capacity provides limited benefit. The learner cannot learn effectively if the necessary cues are not presented in a usable format compatible with its inductive biases. This highlights that capacity alone is insufficient; the data must contain accessible task-relevant information for the specific learner class. This principle is echoed in LLM training: simply scaling up a model might not be effective if the training data lacks the necessary structure or quality to guide its learning towards desired capabilities; data curation and synthesis are often key to unlocking performance at scale.

These empirical results can be coherently interpreted through the lens of V-Information. The pixel permutation C acts to drastically reduce $I_{\mathcal{V}_{\text{CNN}}}(X^*; Y)$ to $I_{\mathcal{V}_{\text{CNN}}}(X_{\text{obs}}; Y) \approx 0$ (relative to achieving non-trivial learning, as performance is low). The synthesizer G functions as an information restoration mechanism. By learning to (approximately) invert the permutation, it performs computation to transform X_{obs} into X_{synth} such that $I_{\mathcal{V}_{\text{CNN}}}(X_{\text{synth}}; Y) > I_{\mathcal{V}_{\text{CNN}}}(X_{\text{obs}}; Y)$. The quality of this restoration, dependent on the synthesizer's own capacity C_G and its training, determines how close $I_{\mathcal{V}_{\text{CNN}}}(X_{\text{synth}}; Y)$ can approach $I_{\mathcal{V}_{\text{CNN}}}(X^*; Y)$. The scaling with learner capacity C_L then demonstrates that the *realized* V-Information (as proxied by accuracy) depends on whether L is sufficiently powerful to exploit the information made available. If $I_{\mathcal{V}_{\text{CNN}}}(X; Y)$ is high (as in X^* or well-restored X_{synth}), increasing C_L is beneficial. If it's low (as in X_{obs}), increasing C_L offers much less improvement. The qualitative appearance of the images (Figure 1, Left Panel) supports this: the top row shows clear original CIFAR-10 images; the second row displays the pixel-permuted versions, which appear as random noise, visually confirming the destruction of spatial structure; the subsequent rows exhibit the autoencoder's reconstructions (X_{synth}), which, while blurry and imperfect compared to the originals, visibly restore some semblance of the original objects and spatial coherence, especially compared to the permuted images. This visual restoration of structure is precisely what allows the CNN learner to regain significant predictive performance. Our initial findings affirm the core hypothesis: computation, embodied by a trained synthesizer, can restore task-relevant information that was rendered unusable by a corruption process, thereby improving downstream learning performance, with the extent of this restoration and its utility being modulated by the capacities of both the synthesizer and the learner.

4. Conclusion and Future Work

This study, employing a minimalistic and reproducible experimental setup on CIFAR-10, investigated synthetic data generation as an information restoration process, analyzed through V-Information and model capacity scaling. We demonstrated that severe data corruption via pixel permutation drastically reduces usable information ($I_{\mathcal{V}_{\text{CNN}}}$) for a CNN learner, leading to poor performance irrespective of learner capacity. However, a Conv-AE synthesizer, even when trained with limited supervision, can partially restore this unusable information by transforming the corrupted data into a more structured format, resulting in significant improvements in downstream task performance. Crucially, increased learner capacity proves beneficial if and only if the training data contains accessible, task-relevant information; capacity is largely ”wasted” on data with very low $I_{\mathcal{V}}$. These findings support the interpretation that computation performed by a synthesizer makes latent task-relevant information accessible, effectively reducing the learning complexity for a subsequent, constrained learner. This principle is highly relevant to current practices in SOTA LLM development, where powerful models generate or refine data (e.g., instruction tuning, chain-of-thought, AI feedback for alignment (**?**Bai et al., 2022)) to enhance the V-Information for target LLMs, rather than creating fundamentally new information. Our work contributes to understanding that synthetic data's power often lies in this computational reformatting and restoration.

Several exciting directions for future work emerge from this small-scale investigation. First, exploring the dynamics of **model collapse** (Shumailov et al., 2023; Alemohammad et al., 2023) in an iterative synthesizer-learner loop (e.g., CAE → CNN → Pseudo-Labels from CNN → Retrain CAE with Pseudo-Labels + Permuted Data) is a critical next step. Key questions include whether $I_{\mathcal{V}}$ decays over iterations and if there's a critical rate of ”real” data needed to prevent decay. Second, systematically varying the **capacity of the synthesizer** (**C_G**) alongside learner capacity (C_L) would allow for a more complete mapping of the capacity-information landscape. Third, incorporating **verifiers or critics** into the synthetic data pipeline offers another avenue for increasing $I_{\mathcal{V}}$. For instance, a verifier model could perform rejection sampling on the synthesizer's output, filtering for high-quality or high-relevance examples based on learned criteria (e.g., adherence to safety principles, factual correctness, or stylistic consistency). This process, akin to how human feedback or AI critics refine LLM outputs in Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022), doesn't create new raw information from scratch but performs computation to select or weight data points where task-relevant signals are stronger or better aligned, thus increasing the $I_{\mathcal{V}}$ of the selected subset for the learner. Such verifiers can be seen as injecting additional, targeted V-Information by guiding the data distribution towards regions more beneficial for the learner. Finally, extending this V-Information framework to more directly quantify information restoration in LLM-specific tasks (e.g., instruction following, complex reasoning, summarization) where synthetic data from teacher models is paramount, would be a valuable contribution. Understanding how these teacher models effectively ”distill” knowledge by reformatting it into more usable forms for student LLMs is crucial for advancing the field efficiently and robustly. The V-Information framework provides a robust theoretical and empirical tool for these future analyses.

References

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.04018*, 2023.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. Wiley-Interscience, 2nd edition, 2006.

Geng, X., Chen, Y., Liu, Z., Chen, K., Wei, Z., Liu, T., Tao, F., Zhang, W., Chen, C., Liu, S., Liu, D., Yuan, B., Zhao, T., Chandra, S., Athey, S., and Anandkumar, A. The unmet promise of synthetic images. 2023.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Shukla, A. Effect of scrambling images on the performance of convolutional neural network. *Medium blog post*, Aug 2019. URL https://medium.com/@aayushmnit/effect-of-scrambling-images-on-the-performance-of-convolutional-neural-network-5ba6d675f2f5

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2023.

Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=S1xNqjHFDr.