Training-free Design of Augmentations with Data-centric Principles

Jieke Wu [1] Wei Huang [2] Mingyuan Bai [2] Xiaoling Hu [3] Yi Duan [1] Wuyang Chen [4]

Abstract

The remarkable advancements in Artificial Intelligence (AI) and Deep Learning owe significantly to the evolution of informative datasets. There has been a shift in focus from developing deep neural networks (DNNs) to crafting high-quality training datasets. However, current data-centric approaches predominantly rely on empirics or heavy DNN training costs, lacking established design principles. Our work concentrates on data augmentation, a key technique for enhancing data quality. Grounded by the recent development of deep learning theory, we discover principled metrics that effectively gauge both data quality and its interaction with DNNs. Crucially, these principles can be calculated without the need for extensive DNN training, enabling training-free augmentation design with minimal computation costs. Comprehensive experiments validate that our principles are strongly aligned with optimal choices of augmentations used in practice. Our method is particularly beneficial in **domain-specific** fields like **medical** image analysis, where the optimal augmentation strategy and the data's inductive bias are often unclear. Our results demonstrate consistent improvements over existing state-of-the-art segmentation methods across various medical imaging datasets.

1. Introduction

Over the last decade, there has been substantial advancement in Artificial Intelligence (AI) and Deep Learning. The availability of extensive and high-quality data for developing machine learning models has been a key factor in this progress. Lately, the importance of data in AI has increasingly been recognized, leading to the development of the new idea of "Data-centric AI." (Ng, 2021; Zha et al., 2023; Oala et al., 2023). The attention of researchers and practitioners has gradually shifted from advancing model design to enhancing the quality and quantity of the data. There have been extensive works focusing on designing data, such as data distillation (Wang et al., 2018; Nguyen et al., 2021; Zhou et al., 2022b; Cui et al., 2023), data deduplication (Lee et al., 2021; Kaddour, 2023; Biderman et al., 2023), and prompt engineering (Zhou et al., 2022a; Wei et al., 2022; Liu et al., 2023). Among various techniques, data augmentation is perhaps the most widely adopted method for both computer vision and natural languages, as it can improve the data quality and diversity, and the augmentation processing itself is ***almost free***. These augmentations mainly leverage the nature of specific invariance of data and tasks.

Despite the extremely low cost of the data augmentation, designing an appropriate augmentation strategy is ***never free***! In fact, blindly picking an off-the-shelf augmentation scheme or following popular heuristics will very likely fail (Dvornik et al., 2019; Cubuk et al., 2019). In the realms where deep learning has achieved its most notable successes, the prevalent strategy iteratively involves: selecting some types of augmentations (based on the invariance properties of the DNN and the task), picking a magnitude for each augmentation, training the DNN, and adjusting the augmentations based on the training results. However, there are **two bottlenecks** in this standard routine: 1) The heavy computation costs due to the DNN training, which is essentially attributed to the dependence on training performance for choosing augmentations. For example, AutoAugment (Cubuk et al., 2019) costs thousands of GPU hours to find optimal augmentation policies. 2) Possible unknown inductive bias of special data (e.g. biomedical images) will further aggravate the time and computation costs during the design. Therefore, our core question is:

Can we design optimal augmentation strategies without any DNN training cost?

In this work, we provide affirmative answers. We focus on designing augmentations for images. At the core of our method are metrics that can characterize the quality of data. More importantly, these metrics can be calculated at the initialization stage of DNNs, without dependence on any model training or pretrained checkpoints. These metrics

[1] University of Science and Technology of China [2] RIKEN Center for Advanced Intelligence Project [3] MGH/Harvard Medical School [4] Simon Fraser University. Correspondence to: Wuyang Chen <wuyang@sfu.ca>.

AI for Science Workshop, *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024.

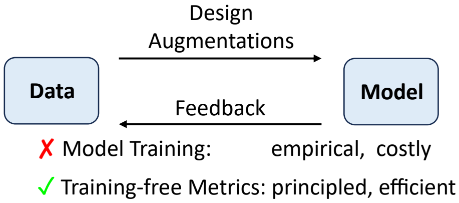

Design
Augmentations
Data
Model
Feedback
✗ Model Training: empirical, costly
✓ Training-free Metrics: principled, efficient

Figure 1. How to design data augmentations? Instead of leveraging model training as feedback, we propose training-free metrics, that can efficiently estimate optimal augmentation choices in principle.

are grounded by recent deep learning theory, thus providing both principled understanding and efficient calculations. We first characterize the impact of data augmentations with our metrics from two perspectives: 1) data quality, being agnostic to any specific model; 2) interactions between data and models. Next, we validate our metrics on standard vision benchmarks and models, showing that augmentations estimated by our metrics are highly aligned with optimal ones. Finally, we demonstrate that our metrics are extremely beneficial in domains where the data's inductive bias is unclear, such as **biomedical image analysis**. We for the first time develop data-centric principles to design data augmentations that are both accurate and easy to use in practice. Our contributions are summarized below:

1. We propose to characterize data quality with training-free metrics, which are principled and inspired by recent theories of deep learning.
2. Our metrics are shown to be strongly aligned with optimal augmentations in practice, and can be calculated at networks' initialization without any training cost.
3. Our metrics show strong benefits when designing augmentations for **scientific domains**, such as **biomedical image analysis**.

2. Related works

2.1. Data Augmentation

Data augmentation is an effective data-centric technique for improving the accuracy of machine learning problems. In computer vision, people develop basic augmentations (random rotations, cropping, color jittering, etc.) and advanced ones (mix-up (Zhang et al., 2017), cut-mix (Yun et al., 2019)). In NLP, widely adopted methods include back translation (Edunov et al., 2018) and random replacement/insertion/swap/deletion (Wei & Zou, 2019).

Efforts to delineate data augmentation are widespread. (Rajput et al., 2019) measure the extent of augmented data required for learning positive margin classifiers. (Dao et al., 2019) demonstrate how augmented k-NN classification converges to a kernel method when augmentations transform each data point into a finite set of possibilities. (Hanin & Sun, 2021) explore the intricate interplay between learning rates and augmentations in the context of stochastic gradient descent (SGD). (Lin et al., 2022) delve into the implicit regularization effect of data augmentations on the spectrum of data covariance. Finally, (Zou et al., 2023) analyze the mixup augmentation mechanism from a feature learning viewpoint. To automate the design of data augmentations and avoid manual efforts, people also design algorithms to optimize augmentation policies (Cubuk et al., 2019; Lim et al., 2019; Ho et al., 2019; Hataya et al., 2020; Cubuk et al., 2020; Hataya et al., 2022; Zoph et al., 2020; Li et al., 2021). However, current understanding and algorithms can hardly contribute to both principled and efficient design of augmentations in practice.

2.2. Benign Overfitting of Overparameterized Models

Understanding the generalization of overparameterized models has been a long-lasting topic in deep learning (Arora et al., 2018; Zhang et al., 2021). One renowned recent work by (Bartlett et al., 2020) characterizes the conditions for linear overparameterized models to generalize with perfect fitting of data. Importantly, these conditions are rooted in the requirements of the spectrum of the data covariance matrix: 1) the input feature must be of high dimensions; 2) the data signals must reside in a low-dimensional subspace of the input space. More assumptions include orthogonal and variance-bounded noises. Further works extend this research direction to more settings of models (Tsigler & Bartlett, 2023; Frei et al., 2022; Cao et al., 2022; Kou et al., 2023; Xu et al., 2023; Frei et al., 2023; Kornowski et al., 2023) and data (Xu et al., 2023; Meng et al., 2023). In our work, we will focus on how to design these data-related conditions to help models generalize better.

2.3. Persistent Homology and Topological Data Analysis

Persistent homology (PH) is a tool in topological data analysis (TDA) that tracks topological features and identifies geometric patterns in metric spaces. TDA is used to analyze the topological patterns of data (Wasserman, 2018; Bernstein et al., 2020). Recent works in machine learning try to characterize the complexity of training data or DNNs with TDA. (Bianchini & Scarselli, 2014) for the first time gives the lower and upper bound of DNN's complexity, with the dependence of network's depth, using betti number. More works (Guss & Salakhutdinov, 2018; Melodia & Lenz, 2021) empirically characterize the complexity of advanced DNNs with metrics related to persistence homology. People also propose to improve deep learning practices with TDA, such as early stopping (Rieck et al., 2018), architecture selection (Yang et al., 2021), and measuring the similarity between DNNs (Pérez-Fernández et al., 2021). However, previous works separately analyze TDA of only the model or the data, but never jointly compare the topological complexity of both.

3. Methods

In this section, we first motivate our method by explaining common understandings of data augmentations (Section 3.1). We then decompose our characterizations of data augmentations into two perspectives (Section 3.2 and 3.3). Finally, we illustrate simple steps that make our method easy to use in practice (Section 3.4).

3.1. How to Characterize Data Augmentations without DNN Training?

To answer our core question in Section 1, we first need a principled understanding of how augmentations affect training deep networks. We summarize benefits of augmentations from two perspectives.

Data Quality. Data augmentation is commonly believed to improve sample-wise diversity and mutual information (Bachman et al., 2019; Tian et al., 2020; Sordoni et al., 2021; Geiping et al., 2022; Lin et al., 2022; Mohamadi et al., 2023). By explicitly exploring invariance assumptions on data and tasks, augmentations try to cancel noises and task-irrelevant features with randomness, while retraining task-relevant signals. This effect is agnostic to any specific DNN model that will train on the data. It naturally improves the quality of training data.

Model-dependent Regularizations. Besides improving data quality, augmentations are also used as regularizations to reduce DNN's overfitting and improve its generalization (Hernández-García & König, 2018; Balestriero et al., 2022; Yang et al., 2023; Brigato & Mougiakakou, 2023; Dablain & Chawla, 2023). Augmentations will influence the complexity of trained DNNs and encourage DNNs to learn simplified functions. Correspondingly, models with larger sizes typically require stronger augmentations (regularizations) (Tan & Le, 2021; Liu et al., 2021). This effect, and thus the choice of augmentations, highly depends on the specific DNN to be trained on the data.

The above two aspects are usually coupled together and introduce a mixed effect on model training. To characterize the effect of data augmentations, we propose to separately design two training-free metrics that are: 1) model-agnostic (Section 3.2), and 2) model-aware (Section 3.3).

3.2. Data Quality for Augmentation Design

Quality of data is commonly believed to be vital to training DNNs. Explicit input perturbations and label noises can be significantly detrimental to DNN's generalization (Zhang et al., 2021). More commonly, implicit data qualities, such as the strength of task-relevant signals (Cao et al., 2022) and mutual information between different views (Tian et al., 2020), strongly connect to the convergence and generalization of DNNs (Du et al., 2017; Bartlett et al., 2020), while being non-trivial to quantify. People propose different methods to quantify the quality of data, such as intrinsic dimensions (Pope et al., 2021), data valuations (Nohyun et al., 2022), topological complexity (Hu et al., 2019).

Spectral Bias of Data Quality. In our work, we choose to quantify the data quality by analyzing the behavior of its spectrum. Our analysis is inspired by recent works on benign overfitting, which characterizes the bound of excess risk R of minimum norm estimators θ of a linear regression problem $\min_\theta \mathbb{E}\left(y - x^\top\theta\right)^2$. Suppose $x \in \mathbb{R}^{n\times m}, y \in \mathbb{R}^n$ (i.e. we have n data samples), the minimum-norm and optimal solutions are $\hat{\theta}, \theta^* \in \mathbb{R}^m$.

Definition 3.1 (Effective Ranks (Bartlett et al., 2020))**.** Define $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ as eigenvalues of data covariance matrix $\Sigma = \mathbb{E}[xx^\top] \in \mathbb{R}^{n\times n}$ in descending order. If $\lambda_{k+1} > 0$ for $k \geq 0$, define

$$r_k(\Sigma) = \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k}\lambda_i\right)^2}{\sum_{i>k}\lambda_i^2}.$$

Theorem 3.2 (Excess Risk (Bartlett et al., 2020))**.** *Suppose the conditional noise variance is bounded below* $\mathbb{E}\left[\left(y - x^\top\theta^*\right)^2 \mid x\right] \geq \sigma^2$*;* $y - x^\top\theta^*$ *is* σ_y^2*-subgaussian;* $b, c, c_1 > 1$*;* $\delta < 1$ *with* $\log(1/\delta) < n/c$*. Define the excess risk* $R(\theta) := \mathbb{E}_{x,y}\left[\left(y - x^\top\theta\right)^2 - \left(y - x^\top\theta^*\right)^2\right]$, $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$*. If* $k^* < n/c_1$*, then with probability at least* $1-\delta$*:*

$$R(\hat{\theta}) \leq c\left(\underbrace{\|\theta^*\|^2\|\Sigma\|\max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}}_{I_1}\right) + c\log(1/\delta)\sigma_y^2\underbrace{\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)}_{I_2}$$

$$\textit{and} \qquad \mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

We can see that I_1 and I_2 control the upper bound of $R(\hat{\theta})$, and I_2 controls the lower bound of $R(\hat{\theta})$. Since both I_1 and I_2 are properties of data, this implies that it is possible to improve model generalization by improving the data quality via the spectrum of the data.

Metric 1: Effective Ranks. We will study three quantities of Σ that control I_1 and I_2: r_0, k^*, and R_{k^*}. A Σ of a small r_0, a small k^*, and a large R_{k^*} will lead to a lower excess risk. This states that, from the data perspective, to improve the model's generalization, we need to allocate the spectrum to limited top eigenvalues. As visualized in Figure 3, when we perturb images with inappropriately larger angles, more noises (void background) will be introduced, leading to a less concentrated covariance spectrum and thus a larger r_0.

The effective rank only focuses on data. However, models of different sizes or structures are observed to require different augmentations. Therefore, the effective rank only provides a model-agnostic "prior" of optimal augmentations.

3.3. Model-dependent Complexity for Augmentations

Regularization is commonly believed to be vital to training generalizable DNNs. Data augmentation, which introduces implicit constraints on data invariance via stochasticity, serves as an effective way to regularize the model complexity during training.

Topological Complexity of Data and Models. To quantify model-dependent effects of augmentations, we leverage topological data analysis (TDA), which directly summarizes topologies from a set of points. We introduce core concepts in TDA, and refer readers to (Zomorodian, 2012; Wasserman, 2018) for more details.

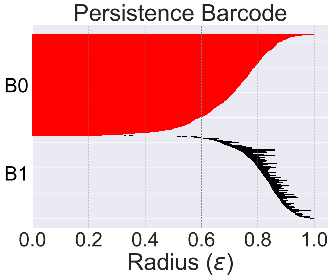


Figure 2. Persistence diagram represents high-dimensional holes as a collection of horizontal line segments. The x-axis corresponds to the radius (ϵ) of spheres around each data point (normalized by the largest death radius in our work), and the y-axis is an ordering of high-dimensional holes. Data: CIFAR-10 images. "B0" and "B1" are 0-th and 1-st betti number, respectively.

Algebraic Topology. Algebraic topology uses algebraic concepts like groups and rings to study the intrinsic qualitative properties of spatial points, focusing on aspects preserved through deformation, twisting, rotation, and extension. Algebraic topology analyzes and computes the "hole" structures of the point cloud, and offers a significant advantage in characterizing data or network structures due to its invariance to features like scaling, rotation, and translation, with the complexity of the topological geometry defined by the count of "holes" in various dimensions.

Betti Number and persistence diagram. Geometrically, the i-th Betti number refers to the number of i-th dimensional holes in the data points. The 0-dimensional hole is the connected component of the space, and 1-dimensional holes are circles. To determine Betti numbers, we first start by envisioning spheres expanding around data points and monitoring the intersection of these spheres as the radius increases. This builds a distance matrix of pair-wise data points (ℓ_2 distance in our work), and constructs a complex where topological features like loops and voids emerge and vanish. These features are then represented in *barcodes* or *persistence diagram* (Figure 2), where each line corresponds to the lifespan of a topological structure across different radii (ϵ). The Betti number at any given radius is calculated by counting these topological features, with the persistence of these features (reflected by the length of bars in the barcode or their position in the persistence diagram) indicating their significance. This approach allows for the analysis of the underlying topology of the data set as the radius changes.

Metric 2: Complexity of Persistence Diagram. The Betti number quantifies the number of "holes" under a specific sphere radius. It is also vital to measure the complexity by considering topological changes over a wide range of radii. We define a diagram of n barcodes $B = (\epsilon_{0,i}, \epsilon_{1,i}), i = 1, \cdots, n$. Each barcode i has a birth radius $\epsilon_{0,i}$ and a death radius $\epsilon_{1,i}$. In our work, we will consider the total survival time $S = \sum_{i=1}^{n} \epsilon_{1,i} - \epsilon_{0,i}$ to summarize the complexity of a persistence diagram. This considers the persistence of all barcodes in the diagram. The intuition is that a diagram of predominant features and robust underlying structures characterized by extended persistence intervals will indicate complicated data points.

Note that the difference in dimensionality of data or the model's hidden features will affect these measurements, as in the Euclidean space the pair-wise distance naturally increases with the dimension of space. This will make measurements of different objects not comparable. Therefore, we adopt normalization during our TDA analysis: 1) Norms of data vectors will be normalized to 1 before calculating pair-wise distances; 2) In the persistence diagram, $(\epsilon_{0,i}, \epsilon_{1,i})$ of all bars will be normalized by the largest death radius (or the birth radius of the last component for 0-dimentional diagram). We also provide a visualization of S of input images in Figure 3. Again, images with noisy and void backgrounds (larger rotation angles) suffer from reduced topological complexity (smaller S).

This metric can be calculated on both *data* ($S_{\mathcal{X}}$) and *model's output* ($S_{\mathcal{M}}$). We calculate $S_{\mathcal{M}}$ based on features output from the model's backbone. Unlike metrics for data quality in Sec. 3.2, TDA analysis will consider interactions between data and different DNNs, and it will contribute to a model-aware design of augmentations (Sec. 3.4). The persistence barcode depends on both model weights and architectures. However, since our work focuses on designing data instead of designing DNN architectures, we do not study the relationship between TDA and architectures in detail. In fact, we find larger models show larger $S_{\mathcal{M}}$, indicating that $S_{\mathcal{M}}$ aligns model complexity. For example, LeNet/ResNet-18/ResNet-34 has $S_{\mathcal{M}}$ as 764/852/856, respectively. We also calculate $S_{\mathcal{M}}$ for DeiT series (Touvron et al., 2021), see Appendix A.

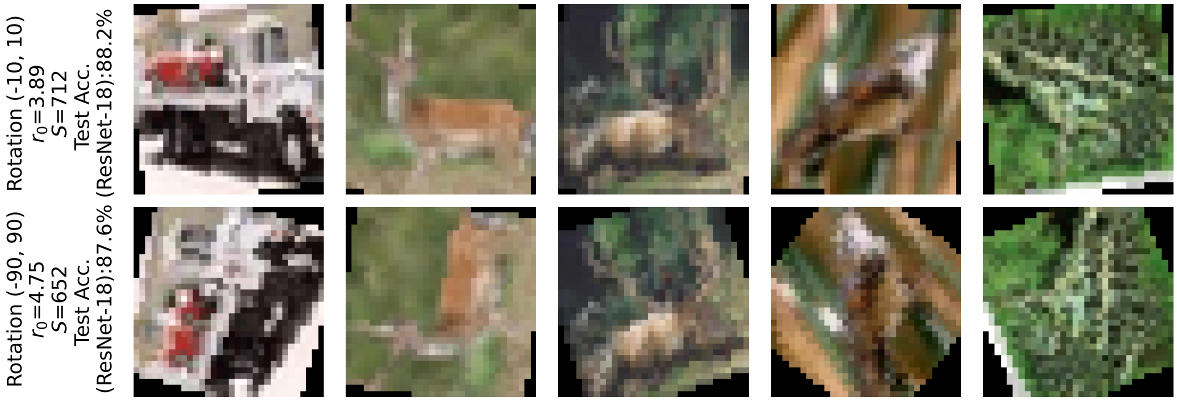


Figure 3. Visualization of augmented images (via random rotations of different angles) show different effective rank (r_0) of data covariance (Section 3.2) and total survival time (S) of persistence diagram (Section 3.3). We can use r_0 (smaller the better) and S (larger the better) to estimate the optimal augmentation magnitude, which is well-aligned with practice (images on the top row lead to better model performance in experiments).

3.4. Training-free Design of Augmentations with Data-centric Principles

The motivations for our choices on effective ranks and topological complexity are deeply rooted in explaining the model's generalization and complexity, particularly the insights into the spectral bias of data and the topological complexity of models. Although the dependency between dataset and model performance is highly complex, we can still find metrics and algorithms that can characterize the data quality and data-model interactions.

For a given type of augmentation, we aim to predict the optimal augmentation magnitude as A^* with our estimation $\hat{A}^*$. $A^*, \hat{A}^* \in [A_{\min}, A_{\max}]$, where $A_{\min}, A_{\max}$ are least and largest magnitudes (e.g. $0 \sim 180$ for random rotations). Estimating optimal augmentation magnitudes is sufficient for choosing optimal types of augmentations, i.e., if the estimated magnitude is 0, then we should not use augmentation. To develop a cost-effective principle while acknowledging the model-agnostic characteristics (Section 3.2) and the model-aware aspects (Section 3.3), we propose the integration of two aforementioned types of training-free metrics, which eliminates the need for any DNN training.

- Step 1: We calculate the spectrum of the covariance matrix of the input data to estimate the optimal augmentation magnitude. For example, for r_0, we estimate $\hat{A}^*_{r_0} = \text{argmin}_A\, r_0$.
- Step 2: Given a target DNN model, we perform model-dependent TDA analysis. For example, for the total survival time, we calculate $S_{\mathcal{X}}$ and $S_{\mathcal{M}}$. We then estimate optimal augmentation magnitudes: $\hat{A}^*_{S_{\mathcal{X}}} = \text{argmax}_A\, S_{\mathcal{X}}$ and model-dependent $\hat{A}^*_{S_{\mathcal{M}}} = \text{argmax}_A\, S_{\mathcal{M}}$.
- Step 3: We suggest accurate augmentation magnitudes by averaging aforementioned augmentation estimations (e.g., $\hat{A}^*_{r_0}$, $\hat{A}^*_{S_{\mathcal{X}}}$, $\hat{A}^*_{S_{\mathcal{M}}}$).

Note that there could be better ways to combine different metrics (like reinforcement learning, evolution, and differentiable search). However, we intentionally make our combination strategy simple, since our focus is to explore and design data-centric training-free metrics, instead of advanced ways of combining them. This follows similar motivations of previous training-free neural architecture search works (Abdelfattah et al., 2021; Chen et al., 2021).

Table 1. Computation costs of r_0 and S. Input: 1000 random matrices of the shape $3 \times 32 \times 32$. CPU: Intel Xeon Platinum 8352V.

	Time (s)	CPU Memory (M)
r_0	1.44	0.11
S	46.09	31.08

We will compare how accurately different metrics can estimate the optimal augmentation magnitude in Section 4. We also benchmark the time and memory cost of our training-free metrics, where we calculate r_0 and S for 1000 random matrices of the shape $3 \times 32 \times 32$ (same as CIFAR-10 images). As shown in Table 1, both r_0 and S are cost-efficient.

4. Experiments

In our experiments, we target answering two questions. *First*, can our training-free metrics accurately estimate optimal augmentation magnitudes? We will decouple the analysis into data quality (Section 4.1) and model-dependent measurements (Section 4.2). These benchmarking experiments are comprehensive and heavy, as we need to exhaustively

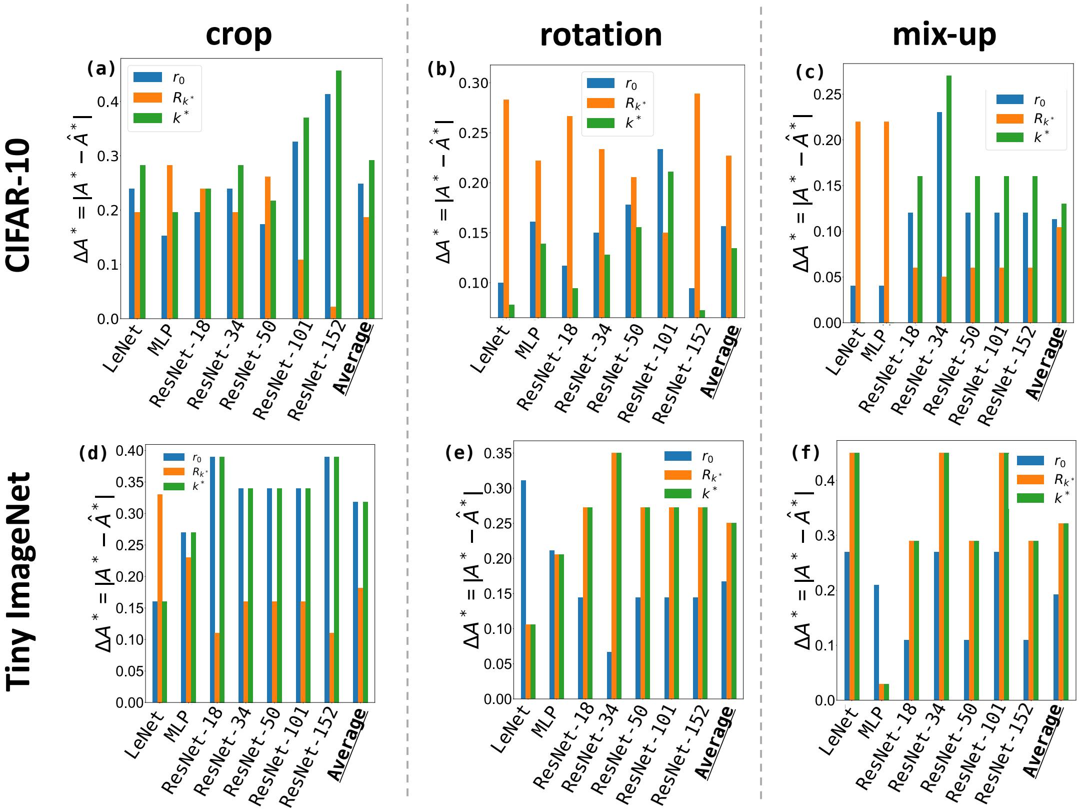


Figure 4. Data-centric training-free metrics based on data quality (effective rank 3.2) can accurately predict the optimal strategies for different augmentations: **(a, d)** random crop; **(b, e)** random rotation; **(c, f)** mix-up. **Top row**: CIFAR-10. **Bottom row**: Tiny ImageNet. We can see r_0 (blue bars) achieves stable and accurate predictions. We compare the estimation error of optimal augmentation magnitudes $\Delta A^* = |A^* - \hat{A}^*|$.

train deep networks with different augmentation magnitudes. *Second*, can our principled design of augmentations improve real-world problems? We will adopt our principle to a medical image analysis with much more challenging data inductive bias (Section 4.3).

We consider widely used augmentations for images, including random crop (followed by re-scale), rotation, and mix-up (Zhang et al., 2017). We study them on both CIFAR-10 and Tiny ImageNet (Le & Yang, 2015). All our metrics are calculated at network's initialization without any training cost. Although they may change during training, we intentionally target avoiding any training costs. Please refer to Appendix B for details about training and implementations.

4.1. Data Quality for Optimal Augmentations

We consider diverse DNN architectures: a three-layer ReLU-Linear network of two hidden widths as 512 and 256 (dubbed "MLP"), LeNet (LeCun et al., 1998), ResNet family (ResNet-18, 34, 50, 101, 152) (He et al., 2016). For each model, we exhaustively train it with different augmentation magnitudes.

To quantify how accurate the estimated augmentation magnitude is, we calculate $\Delta A^* = |A^* - \hat{A}^*|$ for r_0, k^*, and R_{k^*}. Specifically, A^* is chosen to maximize the model's test accuracy via exhaustive search, and $\hat{A}^*_{r_0} = \mathrm{argmin}_A\, r_0$, $\hat{A}^*_{k^*} = \mathrm{argmin}_A\, k^*$, $\hat{A}^*_{R_{k^*}} = \mathrm{argmax}_A\, R_{k^*}$. As shown in Figure 4, across different augmentations and models, on both CIFAR-10 and Tiny ImageNet, we find r_0 shows the best alignment with the optimal augmentation magnitudes.

4.2. Model-dependent TDA for Optimal Augmentations

Next, we study the behaviors of topological complexity (larger the better) of both data ($S_{\mathcal{X}}$) and models ($S_{\mathcal{M}}$) in Figure 5. Importantly, $S_{\mathcal{M}}$ and $S_{\mathcal{X}}$ show different behaviors on different augmentations, indicating that they are complementary to each other. This indicates that we should consider both data's quality and data-model interactions to achieve more informative estimations than only focusing on model-agnostic metrics.

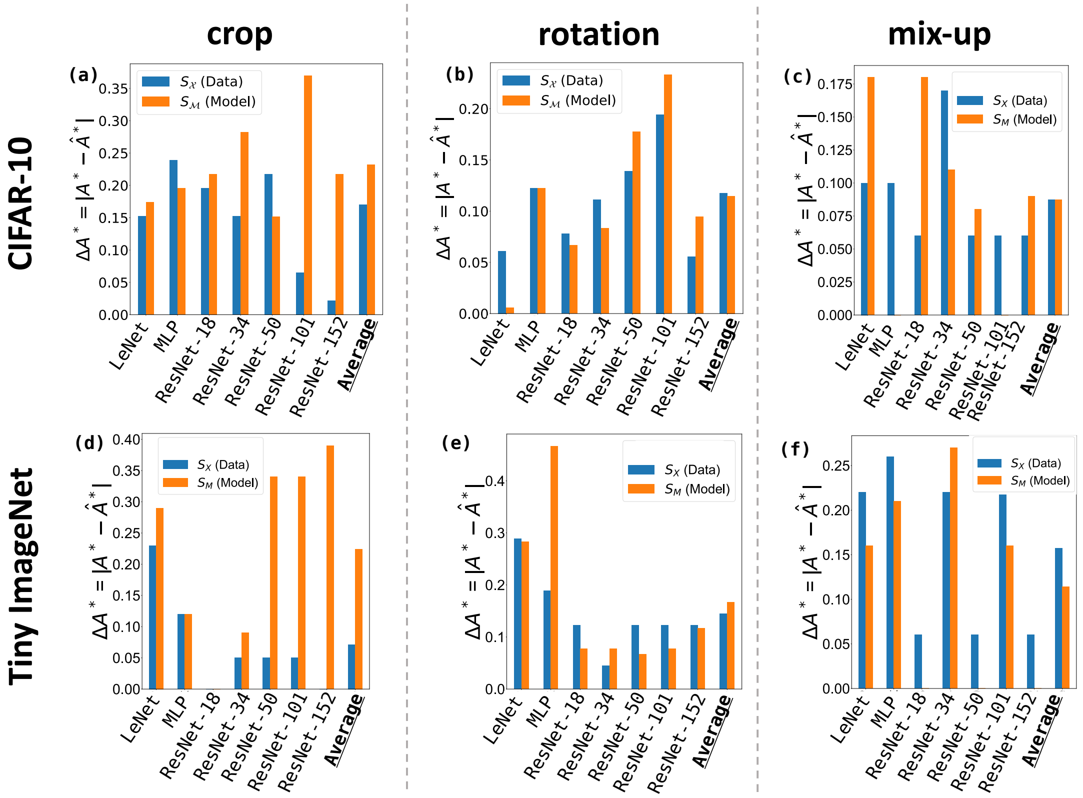


Figure 5. Data-centric training-free metrics based on persistence homology (Sec. 3.3) can accurately predict the optimal strategies for different augmentations: **(a, d)** random crop; **(b, e)** random rotation; **(c, f)** mix-up. **Top row**: CIFAR-10. **Bottom row**: Tiny ImageNet. We can see that $S_{\mathcal{X}}$ and $S_{\mathcal{M}}$ are complementary to each other. We compare the estimation error of optimal augmentation magnitudes $\Delta A^* = |A^* - \hat{A}^*|$.

4.3. From Theory to Practice: Medical Image Analysis

We now move to a more practical case for biomedical image analysis because of two **core motivations**:

1. Augmentations are rarely systematically studied on medical images, and augmentations are designed in different works without principle.
2. Different domains of medical images encompass different inductive biases (object locations, color shifts, illuminations, texture, shape, etc., largely deviating from natural images like CIFAR-10.

Datasets. We choose two medical image segmentation datasets, and visualize images and masks in Fig. 6. We split the train-validation set as 4:1.

ISIC 2018 (International Skin Imaging Collaboration) (Gutman et al., 2016) contains 2594 camera-acquired dermatologic images and corresponding segmentation maps of skin lesion regions. We resize all the images to a resolution of 512 × 512.

BUSI (Breast UltraSound Images) (Al-Dhabyani et al., 2020) consists of ultrasound images of normal, benign and malignant cases of breast cancer along with the corresponding segmentation maps. We use only benign and mailgnant images which results in a total of 647 images resized to a resolution of 256 × 256.

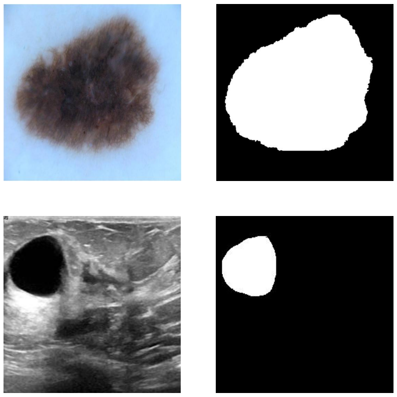


Figure 6. Visualization of medical images and segmentation masks. Top: ISIC 2018 (Gutman et al., 2016) for skin lesion. Bottom: BUSI (Al-Dhabyani et al., 2020) for breast ultrasound.

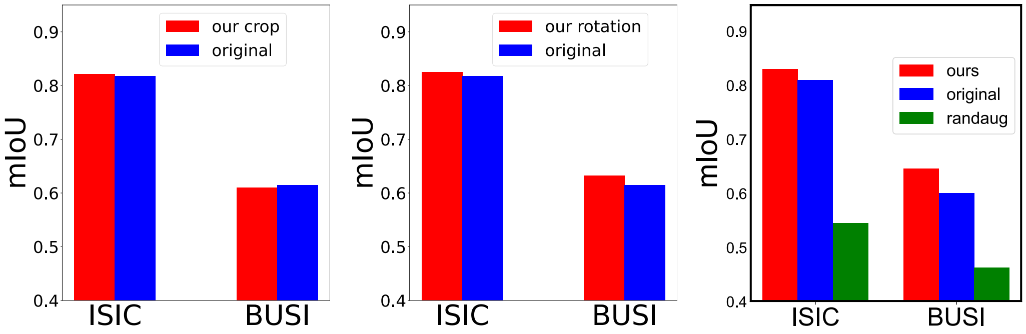


Figure 7. Our training-free data-centric principles can design augmentations for real-world medical image segmentation problems. Comparison of segmentation performance between (Valanarasu & Patel, 2022) and UNext trained with our designed augmentations (from left to right): random crop, random rotation, crop+rotation. Our training-free data-centric principles can also outperform RandAug (Cubuk et al., 2020).

Table 2. Augmentation magnitudes (crop ratio and rotation angle) estimated by r_0, $S_{\mathcal{X}}$, and $S_{\mathcal{M}}$ on medical images.

Random Crop	r_0	$S_{\mathcal{X}}$	$S_{\mathcal{M}}$	Average
ISIC 2018	0.92	0.56	0.5	0.66
BUSI	0.97	0.53	0.53	0.67
Random Rotation	r_0	$S_{\mathcal{X}}$	$S_{\mathcal{M}}$	Average
ISIC 2018	64	2	2	23
BUSI	2	169	144	105

Model. We focus on improving UNext (Valanarasu & Patel, 2022), which is a widely adopted UNet-like convolutional neural network. UNext integrates UNet (Ronneberger et al., 2015) and MLP-Mixer (Tolstikhin et al., 2021) for efficient and accurate medical image analysis.

Augmentations. Originally, (Valanarasu & Patel, 2022) empirically considered two augmentations: random flips and rotations with $90 \times \{0, 1, 2, 3\}$ degrees. We target adopting random crop and rotation augmentations with magnitudes estimated by r_0, $S_{\mathcal{X}}$, and $S_{\mathcal{M}}$.

Results. We compare the segmentation performance by (Valanarasu & Patel, 2022) versus UNext trained with our augmentation strategy. For our methods, we only change the data augmentations while keeping other model/training protocols unchanged. From the bar plots in Figure 7, we can see that on ISIC 2018, by adopting our estimated augmentations, we can consistently improve performance. mIoU is also improved on BUSI when we adopt our random rotations. When we adopt both random crop and rotations, our segmentation performance outperforms UNext on both two datasets. Notably, our augmentation strategy is designed without involving any model training on these medical images, just using our training-free metrics.

We also compare with RandAug (Cubuk et al., 2020), which was proposed as a principled strategy for designing image augmentations without introducing any human or DNN training efforts. To fairly compare our method with RandAug, we directly adopt RandAug on medical images, and show the results in Figure 7 right. We can see that, since RandAug is mainly tested on natural images like ImageNet (Krizhevsky et al., 2012), it fails to adapt to images of different domains. In contrast, our training-free data-centric metrics outperform RandAug and can design better augmentations based on the inductive bias of domain-specific data without any training costs.

5. Conclusions

In this work, we propose effective training-free augmentation designs. Our approach, grounded in deep learning theory concerning the spectral bias of data and the topological complexity of models, illustrates the potential for more efficient and principled augmentation strategies. This is particularly valuable in domains where the data's inductive bias is less understood, such as biomedical imaging. By employing metrics that do not require deep neural network (DNN) training, we successfully design augmentations for real-world medical imaging datasets, demonstrating significant improvements. Our results show that these training-free augmentations can enhance model performance without the need for extensive computational resources typically required for traditional training processes. We anticipate that our work will inspire the community to explore more training-free design methods, potentially leading to a new wave of research focused on data-centric machine learning.

References

Abdelfattah, M. S., Mehrotra, A., Dudziak, Ł., and Lane, N. D. Zero-cost proxies for lightweight nas. *arXiv preprint arXiv:2101.08134*, 2021.

Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. Dataset of breast ultrasound images. *Data in brief*, 28: 104863, 2020.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Balestriero, R., Bottou, L., and LeCun, Y. The effects of regularization and data augmentation are class dependent. *Advances in Neural Information Processing Systems*, 35: 37878–37891, 2022.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bernstein, A., Burnaev, E., Sharaev, M., Kondrateva, E., and Kachan, O. Topological data analysis in computer vision. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, volume 11433, pp. 673–679. SPIE, 2020.

Bianchini, M. and Scarselli, F. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Brigato, L. and Mougiakakou, S. No data augmentation? alternative regularizations for effective training on small datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 139–148, 2023.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Chen, W., Gong, X., and Wang, Z. Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. *arXiv preprint arXiv:2102.11535*, 2021.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Cui, J., Wang, R., Si, S., and Hsieh, C.-J. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pp. 6565–6590. PMLR, 2023.

Dablain, D. A. and Chawla, N. V. Towards understanding how data augmentation works with imbalanced data. *arXiv preprint arXiv:2304.05895*, 2023.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. In *International conference on machine learning*, pp. 1528–1537. PMLR, 2019.

Du, S. S., Lee, J. D., and Tian, Y. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

Dvornik, N., Mairal, J., and Schmid, C. On the importance of visual context for data augmentation in scene understanding. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2014–2028, 2019.

Edunov, S., Ott, M., Auli, M., and Grangier, D. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

Frei, S., Chatterji, N. S., and Bartlett, P. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Frei, S., Vardi, G., Bartlett, P., and Srebro, N. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 3173–3228. PMLR, 2023.

Geiping, J., Goldblum, M., Somepalli, G., Shwartz-Ziv, R., Goldstein, T., and Wilson, A. G. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. *arXiv preprint arXiv:2210.06441*, 2022.

Guss, W. H. and Salakhutdinov, R. On characterizing the capacity of neural networks using algebraic topology. *arXiv preprint arXiv:1802.04443*, 2018.

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., and Halpern, A. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

Hanin, B. and Sun, Y. How data augmentation affects optimization for linear regression. *Advances in Neural Information Processing Systems*, 34:8095–8105, 2021.

Hataya, R., Zdenek, J., Yoshizoe, K., and Nakayama, H. Faster autoaugment: Learning augmentation strategies using backpropagation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 1–16. Springer, 2020.

Hataya, R., Zdenek, J., Yoshizoe, K., and Nakayama, H. Meta approach to data augmentation optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2574–2583, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hernández-García, A. and König, P. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018.

Ho, D., Liang, E., Chen, X., Stoica, I., and Abbeel, P. Population based augmentation: Efficient learning of augmentation policy schedules. In *International conference on machine learning*, pp. 2731–2741. PMLR, 2019.

Hu, X., Li, F., Samaras, D., and Chen, C. Topology-preserving deep image segmentation. *Advances in neural information processing systems*, 32, 2019.

Kaddour, J. The minipile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*, 2023.

Kornowski, G., Yehudai, G., and Shamir, O. From tempered to benign overfitting in relu neural networks. *arXiv preprint arXiv:2305.15141*, 2023.

Kou, Y., Chen, Z., Chen, Y., and Gu, Q. Benign overfitting for two-layer relu networks. *arXiv preprint arXiv:2303.04145*, 2023.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Li, P., Liu, X., and Xie, X. Learning sample-specific policies for sequential image augmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4491–4500, 2021.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. Fast autoaugment. *Advances in Neural Information Processing Systems*, 32, 2019.

Lin, C.-H., Kaushik, C., Dyer, E. L., and Muthukumar, V. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *arXiv preprint arXiv:2210.05021*, 2022.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Melodia, L. and Lenz, R. Estimate of the neural network dimension using algebraic topology and lie theory. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V*, pp. 15–29. Springer, 2021.

Meng, X., Zou, D., and Cao, Y. Benign overfitting in two-layer relu convolutional neural networks for xor data. *arXiv preprint arXiv:2310.01975*, 2023.

Mohamadi, S., Doretto, G., and Adjeroh, D. A. More synergy, less redundancy: Exploiting joint mutual information for self-supervised learning. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1390–1394. IEEE, 2023.

Ng, A. From model-centric to data-centric ai. *DeepLearningAI [Online*, 2021.

Nguyen, T., Novak, R., Xiao, L., and Lee, J. Dataset distillation with infinitely wide convolutional networks. *Advances in Neural Information Processing Systems*, 34: 5186–5198, 2021.

Nohyun, K., Choi, H., and Chung, H. W. Data valuation without training of a model. In *The Eleventh International Conference on Learning Representations*, 2022.

Oala, L., Maskey, M., Bat-Leah, L., Parrish, A., Gürel, N. M., Kuo, T.-S., Liu, Y., Dror, R., Brajovic, D., Yao, X., et al. Dmlr: Data-centric machine learning research–past, present and future. *arXiv preprint arXiv:2311.13028*, 2023.

Pérez-Fernández, D., Gutiérrez-Fandiño, A., Armengol-Estapé, J., and Villegas, M. Characterizing and measuring the similarity of neural networks with persistent homology. *arXiv preprint arXiv:2101.07752*, 2021.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Rajput, S., Feng, Z., Charles, Z., Loh, P.-L., and Papailiopoulos, D. Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pp. 5321–5330. PMLR, 2019.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. Neural persistence: A complexity measure for deep neural networks using algebraic topology. *arXiv preprint arXiv:1812.09764*, 2018.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Sordoni, A., Dziri, N., Schulz, H., Gordon, G., Bachman, P., and Des Combes, R. T. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pp. 9859–9869. PMLR, 2021.

Tan, M. and Le, Q. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pp. 10096–10106. PMLR, 2021.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Tsigler, A. and Bartlett, P. L. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24 (123):1–76, 2023.

Valanarasu, J. M. J. and Patel, V. M. Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 23–33. Springer, 2022.

Wang, T., Zhu, J.-Y., Torralba, A., and Efros, A. A. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Wasserman, L. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.

Wei, J. and Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Xu, Z., Wang, Y., Frei, S., Vardi, G., and Hu, W. Benign overfitting and grokking in relu networks for xor cluster data. *arXiv preprint arXiv:2310.02541*, 2023.

Yang, J., Sang, L., and Cremers, D. Dive into layers: Neural network capacity bounding using algebraic geometry. *arXiv preprint arXiv:2109.01461*, 2021.

Yang, S., Dong, Y., Ward, R., Dhillon, I. S., Sanghavi, S., and Lei, Q. Sample efficiency of data augmentation consistency regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3825–3853. PMLR, 2023.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022a.

Zhou, Y., Nezhadarya, E., and Ba, J. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022b.

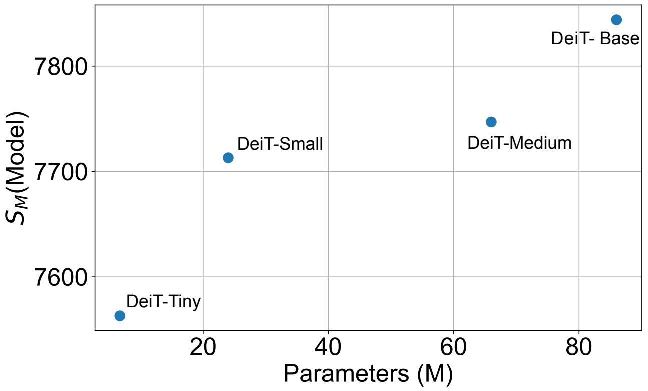
Zomorodian, A. Topological data analysis. *Advances in applied and computational topology*, 70:1–39, 2012.

Zoph, B., Cubuk, E. D., Ghiasi, G., Lin, T.-Y., Shlens, J., and Le, Q. V. Learning data augmentation strategies for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pp. 566–583. Springer, 2020.

Zou, D., Cao, Y., Li, Y., and Gu, Q. The benefits of mixup for feature learning. *arXiv preprint arXiv:2303.08433*, 2023.

A. $S_{\mathcal{M}}$ for Different Models

As the calculation of $S_{\mathcal{M}}$ mainly focuses on model's output and is agnostic model architectures, it can be adopted to different models beyond CNNs. We further test $S_{\mathcal{M}}$ on vision transformers. As shown in Figure 8, $S_{\mathcal{M}}$ can faithfully characterizes the model capacity of DeiT (Touvron et al., 2021).



font=small

Figure 8. $S_{\mathcal{M}}$ for vision transformers (DeiT (Touvron et al., 2021)).

B. Implementation Details

B.1. Training Settings

All our experiments were run on A100 GPUs. The GPU memory cost of our training is low (less than 12GB). We use A100 mainly to accelerate our training, as we need to benchmark seven models on two datasets and three augmentation types, with a dense grid over augmentation magnitudes.

B.1.1. CIFAR-10 and Tiny ImageNet

We train models on CIFAR-10 and Tiny ImageNet for 400 epochs. We use the Adam optimizer with the weight decay as 5×10^{-4}, and `ReduceLROnPlateau` learning rate scheduler (patience as 100, reduction factor as 0.5). One CIFAR-10, we choose the batch size as 32, initial learning rate as 2.5×10^{-4} and decay to 2.5×10^{-9}. On Tiny ImageNet, due to the larger image size, we use the batch size as 16, initial learning rate as 1.25×10^{-4} and decay to 1.25×10^{-9}.

B.1.2. ISIC and BUSI

Following (Valanarasu & Patel, 2022), We train the UNext model on ISIC and BUSI for 500 epochs, with the batch size as 8, initial learning rate as 0.0001 and decay to 0 with the `CosineAnnealingLR` scheduler. We use the SGD optimizer weight decay as 5×10^{-4}.

B.2. Effective Ranks r and R

To calculate effective ranks, we randomly sample 1000 images, and compute their covariance matrix (after images are normalized). We calculate eigenvalues of the covariance matrix, and calculate r and R based on Definition 3.1.

B.3. Topological Complexity $\mathcal{S}$

To calculate $S_{\mathcal{X}}$ and $S_{\mathcal{M}}$, we first need to compute the persistent diagram. We first randomly sample 200 images, and build their point cloud by calculating the pair-wise ℓ_2 distance matrix among images (for $S_{\mathcal{X}}$) and among their features from the model's backbone (for $S_{\mathcal{M}}$). Next, we build the persistent diagram using the `ripser` library. When we focus on the 1st-order betti number, the death radius of the last connected component is infinity. We thus remove the last connected component by default. When we calculate $S_{\mathcal{M}}$, since the distance matrix can be affected by the dimension of model's embedding (large models bias to larger distance), we normalize the persistent barcodes with the largest birth radius, such that the range of the persistent diagram is within $0 \sim 1$.