Journal Pre-proof

Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study

Suorong Yang, Suhan Guo, Jian Zhao, Furao Shen

PII: S0031-3203(23)00901-9
DOI: https://doi.org/10.1016/j.patcog.2023.110204
Reference: PR 110204

To appear in: *Pattern Recognition*

Received date : 7 June 2023
Revised date : 13 November 2023
Accepted date : 11 December 2023

Please cite this article as: S. Yang, S. Guo, J. Zhao et al., Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study, *Pattern Recognition* (2023), doi: https://doi.org/10.1016/j.patcog.2023.110204.

Highlights

Investigating the Effectiveness of Data Augmentation from Similarity and Diversity: An Empirical Study

Suorong Yang, Suhan Guo, Jian Zhao, Furao Shen

- We propose novel quantitative measures agnostic to model training to investigate the effectiveness of DA methods based on similarity and diversity. Through experiments, we demonstrate that these measures provide a framework for assessing the effectiveness of DA methods based on the similarity and diversity of the augmented data.
- Our quantitative measures formulate the similarity and diversity measures for DA techniques. Through comparisons of our quantifying results with the practical effectiveness of DA methods, we find that the importance of similarity and diversity varies across different datasets.
- The proposed measures are conducted in feature space, rather than raw pixel space, which helps explain why some visually meaningless data augmentation methods are still effective.
- While the top-performing DA methods differ across datasets, our similarity-diversity plane reveals that the majority of these methods are concentrated within a particular region, namely the "candidate interval". The interval encompasses DA methods with the highest potential for achieving optimal performance.
- Our study has the potential to provide a more comprehensive understanding of the mechanisms behind DA methods, as well as guide the design and parameter tuning of DA methods. Additionally, our study can offer an efficient preliminary validation of augmentation method efficacy, saving computational resources and time costs in large-scale model training.

Investigating the Effectiveness of Data Augmentation from Similarity and Diversity: An Empirical Study

Suorong Yang[a,b], Suhan Guo[a,c], Jian Zhao[d], Furao Shen [a,c,*]

[a]*State Key Laboratory for Novel Software Technology, Nanjing University, China*
[b]*Department of Computer Science and Technology, Nanjing University, China*
[c]*School of Artificial Intelligence, Nanjing University, China*
[d]*School of Electronic Science and Engineering, Nanjing University, China*

Abstract

Data augmentation has emerged as a widely adopted technique for improving the generalization capabilities of deep neural networks. However, evaluating the effectiveness of data augmentation methods solely based on model training is computationally demanding and lacks interpretability. Moreover, the absence of quantitative standards hinders our understanding of the underlying mechanisms of data augmentation approaches and the development of novel techniques. To this end, we propose interpretable quantitative measures that decompose the effectiveness of data augmentation methods into two key dimensions: similarity and diversity. The proposed similarity measure describes the overall similarity between the original and augmented datasets, while the diversity measure quantifies the divergence in inherent complexity between the original and augmented datasets in terms of categories. Importantly, our proposed measures are model training-agnostic, ensuring efficiency in their calculation. Through experiments on several benchmark datasets, including MNIST, CIFAR-10, CIFAR-100, and ImageNet, we demonstrate the efficacy of our measures in evaluating the effectiveness of various data augmentation methods. Furthermore, although the proposed measures are straightforward, they have the potential to guide the design and parameter tuning of data augmentation techniques and enable the validation of data augmentation methods' efficacy before embarking on

*Corresponding author. E-mail address: frshen@nju.edu.cn (F. Shen).
Email addresses: sryang@smail.nju.edu.cn (S. Yang), shguo@smail.nju.edu.cn (S. Guo), jianzhao@nju.edu.cn (J. Zhao), frshen@nju.edu.cn (F. Shen)

large-scale model training.

Keywords: Data augmentation, interpretability, generalization, deep learning, image classification.

1. Introduction

Data augmentation has been widely adopted in deep neural networks (DNNs) to alleviate the overfitting risks and enhance models' performance [1, 2]. In the context of deep learning, data augmentation (DA) refers to the technique of artificially expanding a dataset by applying various transformations or modifications to the existing data samples. Notably, DNNs that achieve state-of-the-art (SOTA) performance typically utilize various DA methods, such as AutoAugment [3] and RandAugment [4].

Despite the effectiveness of DA methods, existing DA methods evaluate their efficacy typically based on their performance on specific tasks, such as the classification accuracy of models in image classification tasks [5, 6]. Unfortunately, assessing DA methods based on model training poses limitations. Firstly, it has been observed that different image datasets exhibit varying preferences for augmentation methods [7, 6]. For instance, Cutout [8] has proven effective on CIFAR-10 [9], but not on ImageNet [10]. Similarly, AutoAugment has also demonstrated that optimal combinations of augmentation strategies differ across datasets [3]. Consequently, evaluating a DA method necessitates training numerous deep models on various datasets, resulting in significant computational overhead. Secondly, the lack of model training-agnostic quantitative measures hampers the interpretability of DA methods [11]. The underlying mechanisms of DA methods remain opaque, and there are no established standards to guide the design and parameter tuning of DA methods. For example, traditional DA methods aim to generate images that resemble natural scenes (or scenes that may be seen in the test set but not in the training set), such as rotating images to simulate geometric deformation. However, determining optimal values for parameters of DA methods often requires training many deep models, posing practical challenges for researchers.

Although evaluation metrics, such as Inception Score (IS) [15] and Frechet Inception Distance (FID) [16], have been adopted to assess the quality of synthetic images generated by deep generative models [17], they are not applicable to explain the efficacy of DA methods, especially those that are considered

(a) CutMix [12] (b) Mixup [13] (c) GridMask [6] (d) RandomErasing [14]

Figure 1: Some examples of generated data by data augmentation methods.

unrealistic. These metrics (e.g., IS and FID) primarily focus on the quality and diversity of generated images, assuming a close resemblance to real data. However, in the context of DA research, the focus shifts from high-quality data to the generation of varied training data, aiming to mitigate overfitting risks. As exemplified in Figure 1, GridMask [6] generates images by deleting uniformly distributed square regions, and Mixup [13] generates virtual training samples by creating convex combinations of pairs of inputs and labels. Despite the effectiveness in enhancing model performance, augmented images generated by these methods are not simulations of natural scenes and are visually meaningless; therefore, evaluating the quality of the augmented images using these evaluation metrics is not reasonable. Consequently, the disparity between evaluating deep generative models and DA methods necessitates the development of a quantitative standard to assess the effectiveness of DA techniques. Recently, [18] first proposed an empirical study on affinity and diversity to evaluate DA methods while the calculation of both affinity and diversity involves model training. Based on the results of model training, these two metrics lack interpretability and introduce high computational costs, as determining affinity or diversity requires efforts equivalent to determining the final test accuracy. In light of these challenges, there is an urgent need to establish a quantitative standard for investigating the effectiveness of DA methods.

In this paper, we propose quantitative measures for evaluating the effectiveness of data augmentation (DA) methods from similarity and diversity. Although being straightforward, the proposed measures are highly efficient as they are independent of model training. Due to the inconsistency in the effectiveness of DA methods across different datasets [7, 6, 8], instead of using a single measure, we evaluate DA methods on our proposed similarity-diversity plane. The similarity measure captures the overall resemblance between the original and augmented data, while the diversity measure quantifies the varia-

tion in inherent complexity between the original and augmented data regarding categories. Therefore, we decompose the efficacy of data augmentation into similarity and diversity. To address the challenge of quantifying the distance between image datasets [18, 19], particularly in supervised learning with both image data and corresponding labels, we compare the data-label distributions of the original and the augmented training dataset using the optimal transport (OT) distances in a geometrically meaningful way [19, 20]. Additionally, we introduce a novel diversity measure inspired by principal component analysis (PCA) [21] and biological diversity [22], which quantifies the irrelevance between the original and augmented data in terms of classes. Empirical evidence has shown that visual similarity and diversity have little impact on the effectiveness of DA methods [13, 6], leading us to explore feature space embeddings for measurement purposes. By formally defining similarity and diversity measures, we focus on image-level DA methods and evaluate 220 different augmentations on CIFAR-10 and CIFAR-100 and 143 on ImageNet datasets. The investigated DA methods include basic image transformations that vary both broad transform families and finer transform parameters, as well as some SOTA DA methods, such as Mixup [13], CutMix [12], Cutout [8], Randomerasing [14], GridMask [6], AutoAugment [3], RandAugment [4], and KeepAugment [23], etc. Our experimental results demonstrate a strong correlation between the proposed measures and the performance of DA approaches, highlighting the insufficiency of a single measure, whether it focuses on similarity or diversity, in comprehensively quantifying the effectiveness of a data augmentation method.

In summary, the contributions of this work are as follows:

1. We propose novel quantitative measures agnostic to model training to investigate the effectiveness of DA methods based on similarity and diversity. Through experiments, we demonstrate that these measures provide a framework for assessing the effectiveness of DA methods based on the similarity and diversity of the augmented data.
2. Our quantitative measures formulate the similarity and diversity measures for DA techniques. Through comparisons of our quantifying results with the practical effectiveness of DA methods, we find that the importance of similarity and diversity varies across different datasets.
3. The proposed measures are conducted in feature space, rather than raw pixel space, which helps explain why some visually meaningless data augmentation methods are still effective.

4. While the top-performing DA methods differ across datasets, our similarity-diversity plane reveals that the majority of these methods are concentrated within a particular region, namely the "candidate interval". The interval encompasses DA methods with the highest potential for achieving optimal performance.
5. Our study has the potential to provide a more comprehensive understanding of the mechanisms behind DA methods, as well as guide the design and parameter tuning of DA methods. Additionally, our study can offer an efficient preliminary validation of augmentation method efficacy, saving computational resources and time costs in large-scale model training.

2. Related Work

2.1. Basic Data Augmentation

Data augmentation is a widely used technique to improve the generalization of DNNs. Traditional DA methods generate augmented data by simulating real-scene data through image manipulation such as rotation, flipping, translation, random cropping, etc. For instance, random cropping and horizontal flipping are the most commonly used data augmentation for training deep models. These classic methods are fundamental to obtaining highly generalized deep models.

More advanced DA methods, such as Mixup [13] and CutMix [12], combine two or more images or sub-regions of images into one. These methods modify the input images and labels to fuse information from multiple images and can improve the generalization of models by providing diverse training samples. Recently, researchers have highlighted the importance of occlusion in model generalization and proposed some image-erasing-based methods, including RandomErasing [14], Cutout [8], Hide-and-Seek (HaS) [24], and GridMask [6]. These methods replace random patches in the training samples with some specific values, which can reduce the sensitivity of models, increase the perception field, and enhance the generalization performance. Because occlusion may introduce distribution shift and remove some critical regions in the images, KeepAugment [23] uses the saliency map to detect and preserve the essential regions during augmentation.

Another line of data augmentation research is to leverage the power of ensemble learning in conjunction with data augmentation [25, 26, 27]. Ensemble-based DA typically uses multiple models to guide or optimize the

generation of augmented data. Our work can be applied within this approach to further optimize and evaluate the selection and combination of DA techniques.

2.2. Automated Data Augmentation

Furthermore, automated data augmentation, including AutoAugment [3], Fast-AutoAugment [28], and AWS [29] that automatically searches augmentation policies based on some metrics (e.g., test accuracy of a trained model) has been proposed. Specifically, AutoAugment [3] leverages reinforcement learning to find existing policies for the optimal combination of DA operations on various image datasets. RandAugment [4] utilizes grid search to select and apply a combination of augmentation transformations to training images to improve model robustness. Fast-AutoAugment [28] is motivated by density matching between training and test datasets and proposes an inference-only metric to evaluate the data augmentation. AWS [29] designs an augmentation-wise weight-sharing strategy to search augmentation methods. These automated data augmentation methods introduce the concept of evaluating data augmentation techniques, but this evaluation is based on model training or through the use of the test set, hence lacking interpretability and practicality (as the test set is unseen in practice). Instead, our work aims to decompose the effectiveness of data augmentation methods into interpretable similarity and diversity metrics, thus comprehensively quantifying the effectiveness of DA methods.

2.3. Evaluation of Data Augmentation

Among the various DA methods, the mechanism underlying the effectiveness of these methods is not yet fully understood. To better understand the impact of DA on the performance of DNNs, a recent study [18] first proposes an empirical study on affinity and diversity. Affinity is the difference between the accuracy of a model tested on clean data and an augmented validation set. It can be seen as a measure of distribution shift caused by augmentations. Specifically, affinity is defined as follows:

$$A[a, m, \mathcal{D}_{val}] = Accuracy\left(m, \mathcal{D}'_{val}\right) / Accuracy\left(m, \mathcal{D}_{val}\right), \tag{1}$$

where a is an augmentation method, m is the model trained on the original training set, $\mathcal{D}_{val}$ and $\mathcal{D}'_{val}$ are validation datasets and augmented validation datasets, respectively. At the same time, they propose a model training-based measure for diversity, that is, the final training loss of a model trained

with a given DA method, relative to the final training loss of the model trained on clean data:

$$D[a, m, \mathcal{D}_{train}] = \mathbb{E}_{\mathcal{D}'_{\text{train}}}\left[L_{\text{train}}\right] / \mathbb{E}_{\mathcal{D}_{\text{train}}}\left[L_{\text{train}}\right], \tag{2}$$

where $\mathbb{E}_{\mathcal{D}'_{\text{train}}}\left[L_{\text{train}}\right]$ and $\mathbb{E}_{\mathcal{D}_{\text{train}}}\left[L_{\text{train}}\right]$ are the final training loss of a model, m, trained with the augmented training dataset $\mathcal{D}'_{\text{train}}$ and clean training dataset $\mathcal{D}_{train}$, respectively. a is the applied augmentation method. However, determining the affinity and diversity requires the same amount of work as determining the final test accuracy, which is computationally expensive. Moreover, evaluating DA methods' effectiveness is still indirectly done by assessing the actual performance after the entire training of deep models, making it difficult to use when designing augmentation strategies and tuning parameters in practice. In contrast to previous studies, our approach involves training an embedding model for each dataset only once to embed data into feature space. The proposed measures are based on distance measures and are agnostic to model training, drastically reducing the GPU requirements compared to the aforementioned study. Therefore, our method can be effortlessly employed in practice. To ensure the validity of our proposed metrics, we will perform a comparison with the quantitative measures presented in the previous study in Section 4.6.

2.4. Optimal Transport between Datasets

Optimal transport (OT) is a powerful and principled approach to compare probability distributions with strong theoretical foundations and desirable computational properties [19]. OT considers two probability measures, denoted as α and β in spaces $\mathcal{P}(X)$. The Kantorovich's OT problem [30] is defined as:

$$\mathcal{L}_c(\alpha, \beta) \triangleq \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) \mathrm{d}\pi(x, y), \tag{3}$$

where $c(x, y)$ is a cost function indicating the distance between x and y, and $\mathcal{U}(\alpha, \beta)$ consists of joint distributions $\mathcal{M}_+^1$ over the product space $\mathcal{X} \times \mathcal{Y}$ with marginals α and β:

$$\mathcal{U}(\alpha, \beta) \triangleq \left\{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : P_{\mathcal{X}\sharp}\pi = \alpha, P_{\mathcal{Y}\sharp}\pi = \beta\right\}, \tag{4}$$

where $P_{\mathcal{X}\sharp}\pi = \alpha$ and $P_{\mathcal{Y}\sharp}\pi = \beta$ are the push-forward measures [30] of the projections $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$, repectively. Minimizers of this problem are called optimal transport plans. Since the measures are rarely

known in practice and image datasets contain finite discrete samples, α and β can be defined as discrete measures $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}^{(i)}}$ and $\beta = \sum_{i=1}^{m} \mathbf{b}_i \delta_{\mathbf{y}^{(j)}}$, where $\mathbf{a}$ and $\mathbf{b}$ are vectors in the probability simplex, $\left\{\mathbf{x}^{(i)}\right\} \in \mathcal{X}$ and $\left\{\mathbf{y}^{(j)}\right\} \in \mathcal{X}$, $\delta_{\mathbf{x}^{(i)}}$ and $\delta_{\mathbf{y}^{(j)}}$ are Dirac measures of mass 1 located at the point $\mathbf{x}$ and $\mathbf{y}$, respectively [31]. When smoothing the classic optimal transport problem with an entropic regularization term, the entropy-regularized problem is as follows:

$$\mathrm{OT}_{\epsilon}(\alpha, \beta) \triangleq \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y)\mathrm{d}\pi(x, y) + \varepsilon \mathrm{H}(\pi \mid \alpha \otimes \beta), \tag{5}$$

where $c(x, y)$ is the pairwise costs between x and y, ϵ is the regularization coefficient, and $\mathrm{H}(\pi \mid \alpha \otimes \beta) = \int \log(\mathrm{d}\pi/\mathrm{d}\alpha\mathrm{d}\beta)\mathrm{d}\pi$ is the relative entropy. The entropy-regularized problem can be solved faster with the Sinkhorn algorithm [32].

3. Method

In this section, we provide a comprehensive introduction to investigating the effectiveness of data augmentation in terms of similarity and diversity measures. Codes for reproducing our method are available at [1].

3.1. Similarity

Notably, a fundamental assumption in machine learning algorithms is that the training and test data share the same distribution and feature space. Thus, a model trained on the training set can generalize well on the test set. Ideally, the best data augmentation method should generate an augmented set to approximate the test set as close as possible, which can achieve higher test accuracy.

Although the test set is unseen during the training process, the augmented data can offer insights into the test set's distribution as they are independent and identically distributed. Motivated by this assumption, in our work, we propose the similarity measure to determine the distance $d(\mathcal{D}_{aug}, \mathcal{D}_{train})$, where $\mathcal{D}_{aug}$ and $\mathcal{D}_{train}$ represent the augmented and original training dataset, respectively. However, it has been observed that the visual effect and the actual performance of DA techniques are inconsistent. Visually meaningless

[1] https://github.com/Jackbrocp/Investigating_the_Effectiveness_of_Data_Augmentation

data augmentation methods, such as GridMask, can significantly improve the performance of models. Therefore, we measure similarity and diversity in the feature space rather than in the original image space. Feature maps of image data are obtained by training an embedding model with clean data, and the output of the last fully connected layer is used as the feature map. We will show in the ablation study that the choice of embedding models does not affect our main conclusions.

For simplicity, the original data is of the form: $(\boldsymbol{x}, y) \in \mathcal{D}$, where $\boldsymbol{x}$ denotes the feature map of the image data and y corresponds to the label. The dataset $\mathcal{D}$ is sampled from a joint distribution $P(\mathcal{X}, \mathcal{Y})$, noted as $\mathcal{D} = \{(\boldsymbol{x}, y)\} \sim P(\mathcal{X}, \mathcal{Y})$. The dimensions of the label space and feature space of both augmented and original training datasets, denoted by $\mathcal{D}_{aug}$ and $\mathcal{D}_{train}$ respectively, are identical because the former is generated based on the latter. Considering the problem in Equation (3), we define the optimal transport dataset distance as a metric between the feature-label pairs$(\boldsymbol{x}, y)$ and $(\boldsymbol{x}', y')$ as:

$$d((\boldsymbol{x}, y), (\boldsymbol{x}', y')) = \left(d_{\mathcal{X}}\left(\boldsymbol{x}, \boldsymbol{x}'\right)^p + d_{\mathcal{Y}}\left(y, y'\right)^p\right)^{1/p}, \tag{6}$$

where $p \in \mathbb{R}$, $d_{\mathcal{X}}\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ can be defined as the Euclidean distance in the feature space and $d_{\mathcal{Y}}\left(y, y'\right)$ is the distance of labels. The distance of labels $d_{\mathcal{Y}}\left(y, y'\right)^p$ is critical in supervised datasets but difficult to quantify due to the considerable variability among data belonging to different classes. The inter-class variation is not of a similar order of magnitude as the intra-class variation. For instance, the distribution between husky and golden retrievers can be considered an intra-class variation, while the categorical difference between a hot dog and a dog is an inter-class difference. The only information available regarding labels is the corresponding feature maps. Hence, we define an empirical conditional distribution on the feature space: $\mathcal{C}_y(X) = P(X|Y = y)$. Let $X_y = \{x \mid (a, b) \in (\mathcal{X} \times \mathcal{Y}), b = y\}$ denote the set of feature vectors with label y, then X_y becomes a finite sample set of $\mathcal{C}_y(X)$. In this way, the distance between labels becomes the distance between distributions $\mathcal{C}_y(X)$, which can be calculated by a p-Wasserstein distance: $\mathrm{W}_p^p\left(\mathcal{C}_y, \mathcal{C'}_{y'}\right)$. Equation (6) can be calculated as follows:

$$d((\boldsymbol{x}, y), (\boldsymbol{x}', y')) = \left(d_{\mathcal{X}}\left(\boldsymbol{x}, \boldsymbol{x}'\right)^p + \mathrm{W}_p^p\left(\mathcal{C}_y, \mathcal{C'}_{y'}\right)\right)^{1/p}, \tag{7}$$

where $\mathrm{W}_p^p\left(\mathcal{C}_y, \mathcal{C'}_{y'}\right)$ is a p-Wasserstein distance between labels, and $\mathcal{C'}_{y'}$ is the conditional distribution on the augmented set with label y'. The calculation

of the distance between datasets can be achieved through the use of OT as follows:

$$d_{\mathrm{OT}}\left(\mathcal{D}_{aug}, \mathcal{D}_{train}\right) = \min_{\pi \in \mathcal{U}(\mathcal{C}, \mathcal{C}')} \int_{\mathcal{Z} \times \mathcal{Z}} d_{\mathcal{Z}}\left(\boldsymbol{z}, \boldsymbol{z}'\right)^p \pi\left(\boldsymbol{z}, \boldsymbol{z}'\right), \tag{8}$$

where $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ is the joint distribution of data and corresponding labels, $\mathcal{U}$ is defined in Equation (4), $\boldsymbol{z} = (\boldsymbol{x}, y)$ is from $\mathcal{D}_{train}$, and $\boldsymbol{z}' = (\boldsymbol{x}', y')$ is from $\mathcal{D}_{aug}$. Equation (8) defines a proved metric and can be solved with the Sinkhorn algorithm [19, 33]. Finally, the similarity measure can be calculated as follows:

$$similarity(\mathcal{D}_{aug}, \mathcal{D}_{train}) = -d_{\mathrm{OT}}\left(\mathcal{D}_{aug}, \mathcal{D}_{train}\right). \tag{9}$$

The degree of similarity between the training dataset $\mathcal{D}_{train}$ and the augmented datasets $\mathcal{D}_{aug}$ is reflected by the proposed OT distance between them. A smaller OT distance indicates more remarkable similarity between them, resulting in a similarity measure closer to 0 for the DA method employed. In this case, the distribution of the augmented data is similar to that of the training set. Based on the assumption that the training and test datasets share the same distribution, an augmented dataset with a higher similarity measure is more likely to approximate the test dataset, thus enabling the improved performance of deep models. Hence, the proposed similarity measure for DA methods is theoretically linked to the ultimate generalization performance of deep models.

When $\mathcal{D}_{aug}$ and $\mathcal{D}_{train}$ are identical, the proposed similarity measure reaches the maximum value of zero. However, in the context of data augmentation, a higher similarity value is not always desirable because the diversity of the augmented data is limited. As a result, some DA methods with high similarity values severely suffer from overfitting. For instance, if the DA method only duplicates the training set, the maximum similarity value is attained, while such training sets can lead to overfitting easily and hinder the generalization performance of the model. To address this limitation, we propose another measure, diversity, to enhance the credibility of our investigation.

3.2. Diversity

Inspired by the perspective of biodiversity and statistics, we propose a diversity measure that considers the following aspects. Firstly, diversity is a

measure of variation at the species level, as established in the field of biodiversity [22]. For image datasets, we consider the information in the images from different classes unique and characterize the diversity as uncorrelation between classes. Secondly, the diversity is calculated only using the effective components in images [34]. From a statistical perspective, the covariance of random variables is directly proportional to the correlation [35]. In PCA, eigenvectors $\mathbf{u}$ of the covariance matrix represent the principal directions of the data, while the associated eigenvalues λ indicate the variation of the points along the direction. The eigenvalues and eigenvectors of the covariance matrix can characterize the uncorrelation of effective components [35]. Lastly, since visually diverse augmented data may have little correlation with the effectiveness of DA methods, we calculate diversity in the feature space using the same embedding model as the one used for similarity measure.

Let $\mathbf{A}_k \in \mathbb{R}^{m\times n}$ denote the feature matrix of class k, where the j-th column $\mathbf{a}_j$ is the corresponding feature map of j-th input image data. Here, m and n represent the dimension of the feature map and the number of samples from class k, respectively. In this way, we formulate the feature maps of the same class as random variables and use the eigenvalues and eigenvectors of the corresponding covariance matrix to measure the intrinsic diversity. Specifically, let $\boldsymbol{\mu} = \frac{1}{n}\sum_j^n \mathbf{a}_j$ be the sample mean vector. We formulate the empirical covariance matrix $\mathbf{S}$ of the normalized feature matrix as follows:

$$a_{ij}^* = \frac{a_{ij} - \overline{a_i}}{\sqrt{s_{ii}}}, i = 1, 2, ..., m; j = 1, 2, ..., n \tag{10}$$

$$\mathbf{A}^* = [a_{ij}^*]_{m\times n} \tag{11}$$

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A}^*\mathbf{A}^{*T}, \tag{12}$$

where

$$\overline{a_i} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, i = 1, 2, ..., m \tag{13}$$

$$s_{ii} = \frac{1}{n-1}\sum_{j=1}^{n}(a_{ij} - \overline{a_i})^2, i = 1, 2, ..., m. \tag{14}$$

$\mathbf{S} = \sum_{i=1}^{m} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ denotes the eigendecomposition of matrix $\mathbf{S}$, where λ_i is an eigenvalue, and $\boldsymbol{u}_i$ is the associated eigenvector. The eigenvector $\boldsymbol{u}_i$

represents a direction of the effective samples, with the associated eigenvalue λ_i denoting the significance of the samples in that direction. In essence, a larger λ_i implies a greater spread of data along the direction $\boldsymbol{u}_i$. Since the diversity of the augmented data is contingent on that of the original training data, we combine the training and augmented data from a given class k to create $\mathbf{A'}_k$, while $\mathbf{A}_k$ signifies the original training data from class k. In measuring diversity, we focus on the uncorrelation of effective samples between the original training set and the augmented data with the original training set, which can be understood as $d(\mathcal{D}_{train}, \mathcal{D}_{train+aug})$. We perform eigendecomposition on $\mathbf{A}_k$, and $\mathbf{A'}_k$ to obtain $\mathbf{u}_k$, $\boldsymbol{\lambda}_k$ and $\mathbf{u'}_k$, $\boldsymbol{\lambda}'_k$, respectively. We sort the eigenvalues and the corresponding eigenvectors in descending order and select the top t largest eigenvalues and the related eigenvectors for diversity analysis. The value of t is determined as follows:

$$t = \arg\min_{m'} \frac{\sum_{i=1}^{m'} \boldsymbol{\lambda}_{ki}}{\sum_{i=1}^{m} \boldsymbol{\lambda}_{ki}} \geq \theta, \tag{15}$$

where θ is a hyper-parameter and m denotes the total number of eigenvalues of $\mathbf{A}_k$. The main information of the dataset is contained in the largest t eigenvalues and corresponding eigenvectors, while the eigenvectors corresponding to smaller eigenvalues are related to the noise [21, 36]. The value of θ is dataset-dependent, and we will evaluate the effect of different values of θ and discuss the choice of θ in the ablation study.

For simplicity, we use $\boldsymbol{\lambda}^*_k$ and $\boldsymbol{u}^*_k$ to denote the largest t eigenvalues and eigenvectors in class k, respectively. Meanwhile, $\boldsymbol{\lambda'}^*_k$ and $\boldsymbol{u'}^*_k$ denote eigendecomposition results of $\mathbf{A'}_k$. We then obtain the diversity measure of the augmented data as follows:

$$diversity = \frac{1}{K} \sum_{k=1}^{K} ||\boldsymbol{\lambda}^*_k \cdot \boldsymbol{u}^*_k - \boldsymbol{\lambda'}^*_k \cdot \boldsymbol{u'}^*_k||_2^2, \tag{16}$$

where K is the total number of classes, and $|\cdot|$ denotes the element-wise multiplication.

It is worth noting that the smallest diversity value corresponds to the highest degree of similarity between the augmented and training sets when the augmented set is identical to the training set. However, a higher diversity does not necessarily guarantee better performance. For instance, generating images of cats for the dog category produces a high diversity score, but

such data will impair performance. Therefore, we assess the efficacy of DA methods not solely on the diversity measure but in terms of both similarity and diversity measures.

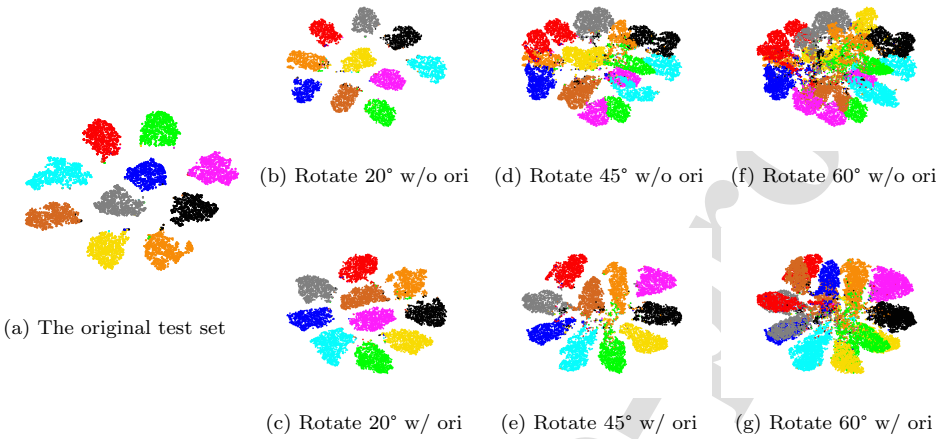
Notably, despite an overall negative correlation between these two measures, either measure is insufficient to measure method efficacy. These two measures evaluate different aspects of DA methods: the similarity measure promotes DA methods to generate augmented data that closely resembles the original training set, which aids in reducing the underfitting risk; the diversity measure encourages DA methods to produce diverse augmented samples, which helps in mitigating the overfitting risk. These two parts serve as complementary tools in investigating the effectiveness of DA methods.

4. Experiment

In this section, we conduct experiments to validate the proposed measures on several widely used benchmark datasets, including MNIST, CIFAR-10, CIFAR-100, and ImageNet.

4.1. Implementation Details and Experiment Settings

To obtain the performance of various augmentations, we closely follow the training settings suggested in [8] and [4] to train various classification models by employing a large number of DA techniques. Specifically, we train ResNet-50 and Wide-ResNet-28-10 models on CIFAR-10 and CIFAR-100 with a batch size of 128. All models are trained for 200 epochs with stochastic gradient descent and momentum. The initial learning rate is set to 0.1. The learning rate starts with a value of 0.1 and is decayed by 20% at epochs 60, 120, and 160. The optimizer uses cross-entropy loss with $l2$ weight decay of 0.0005. To obtain the classification accuracy of ImageNet using different data augmentation methods, the ResNet-50 models are trained for 112.6k steps, with a weight decay of 1e-4, a batch size of 1024, and a learning rate of 0.2, which is decayed by 10% at epochs 30, 60, and 80. For ImageNet, since images have different sizes, all images are also resized to $(224, 224)$. Some classification accuracy results on ImageNet are from [18]. In addition, we use a reduced subset of the ImageNet training set to calculate the similarity and diversity. All images are pre-processed by dividing each pixel value by 255 and normalizing by the dataset statics. To eliminate the influence of commonly used data augmentation techniques such as random cropping and horizontal flipping, we have excluded them from our



(a) The original test set (b) Rotate 20° w/o ori (c) Rotate 20° w/ ori (d) Rotate 45° w/o ori (e) Rotate 45° w/ ori (f) Rotate 60° w/o ori (g) Rotate 60° w/ ori

Figure 2: Visualization of the image embedding space using the MNIST [38] test set. The embedding uses a pre-trained ResNet-18 [37] model on the MNIST training dataset. The t-SNE method is utilized to visualize the image embedding space at various levels of data augmentation. The abbreviation "w/o" denotes the augmented test set without the original test set, while "w/" denotes the combination of both test sets.

implementation unless explicitly stated. For MNIST, we utilize ResNet-18 as the embedding model. For CIFAR-10, CIFAR-100, and ImageNet, we use ResNet-50 [37] as the embedding model. For more detailed implementation and experimental settings, please refer to Appendix B and Appendix C.

4.2. Visualization Results on MNIST Dataset

To gain deep insights into the relationship between our proposed similarity and diversity measures and the distribution of the dataset, as shown in Figure 2, we use the t-distributed stochastic neighbor embedding (t-SNE) algorithm [39] to visualize the embedding results of the MNIST [38] test set rotated by different degrees. Through this experiment, we can see that the similarity decreases and the diversity increases as the rotation angle increases, as also corroborated by the statistical results in Table 1, indicating the proposed measures are consistent with the similarity and diversity of the dataset distribution.

Specifically, Figure 2(a) illustrates the clustering results of the original MNIST test set. Figure 2(b), Figure 2(d), and Figure 2(f) present the rotation augmentation with angles of 20°, 45°, and 60°, respectively. Meanwhile, Figure 2(c), Figure 2(e), and Figure 2(g) show the embedding results of the

Table 1: Similarity and diversity measures on MNIST-test set with different degrees of rotation augmentations.

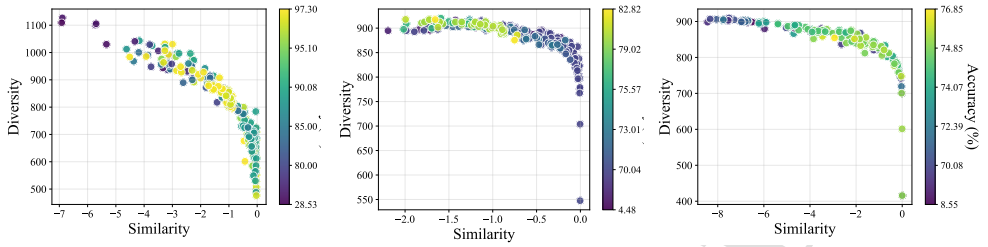
Rotation Angle	Similarity	Diversity
20°	-1.49	719.02
45°	-6.77	861.85
60°	-12.46	1269.45

original and augmented datasets with rotation angles of 20°, 45°, and 60°, respectively. The corresponding statistical results of similarity and diversity measures using different degrees of rotation augmentations are given in Table 1.

When we perform rotation augmentation with angles of 20°, 45°, and 60° in the first row of Figure 2, we can observe that the distribution of the augmented datasets becomes progressively dissimilar to that of the original dataset as the rotation angle increases. Notably, at a rotation angle of 20°, the augmented dataset remains largely invariant, indicating high similarity between the original and augmented datasets. However, as the rotation angle increases to 45° and 60°, the overall distribution of the datasets in the embedding space changes significantly, leading to a decrease in the similarity measure. Especially when the rotation angle reaches 60°, the distribution of data changes significantly, and the classification boundaries of different categories become indistinct.

To scrutinize the connection between the diversity measure and dataset distribution, we present the embedding results of the original and augmented datasets with rotation angles of 20°, 45°, and 60° in the second row of Figure 2. The embedding results reveal that the augmented datasets become increasingly complex with the increase in the rotation angle. Specifically, for rotation angles of 20° and 45°, the distribution of each cluster is akin to that of the original dataset. However, with a 60° rotation, the distribution undergoes significant transformation, and some clusters even become overlapped. Consequently, the augmented dataset's distribution substantially deviates from the training data, which imparts deleterious effects on model training.

Therefore, the proposed measures can be utilized as quantitative indicators to tune the distribution of training data in terms of similarity and diversity. In the following sections, through experiments, we will show how similarity and diversity measures can be used to characterize and understand



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet

Figure 3: Investigation of the efficacy of DA methods in the similarity-diversity plane on CIFAR-10 [9], CIFAR-100 [9], and ImageNet [10], respectively.

the effectiveness of DA methods in detail.

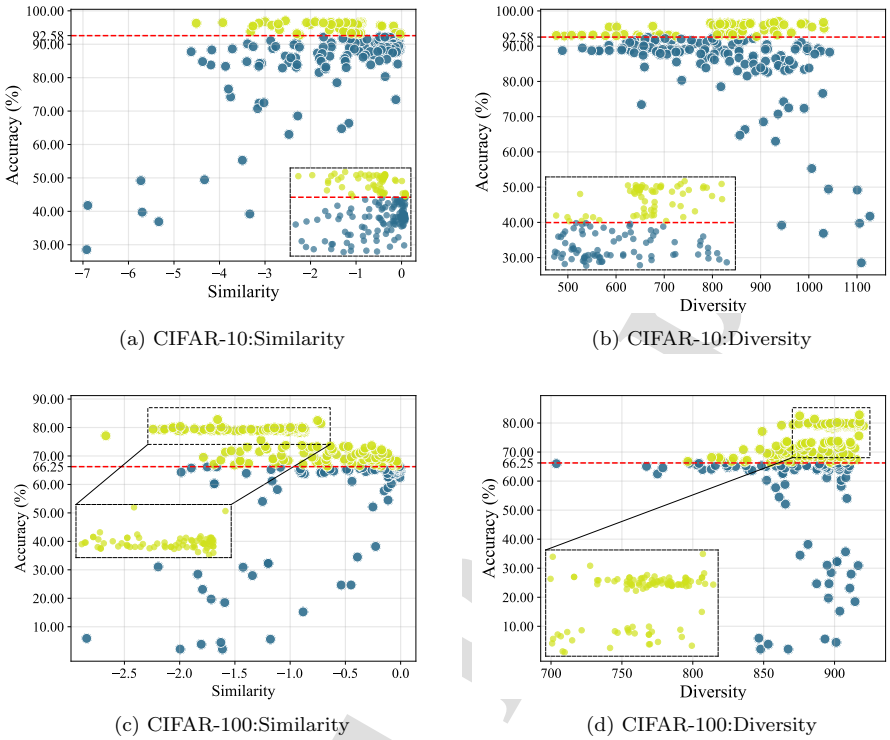
4.3. Investigate the Effectiveness of Augmentations

To investigate the efficacy of the proposed similarity and diversity measures, we conduct experiments using a large number of DA methods on CIFAR-10, CIFAR-100, and ImageNet datasets, respectively. Specifically, Figure 3 illustrates the quantifying results across 220 different augmentation methods for CIFAR-10 and CIFAR-100, and 143 methods for ImageNet, respectively. A complete list of the tested augmentation methods is presented in Appendix H. We also provide the full quantitative results on CIFAR-10, CIFAR-100, and ImageNet given in .csv files, which are uploaded at `https://drive.google.com/file/d/1B9Pms1V9H8fkLxQK9LFvW6fLW3Z3ns0y`.

As shown in Figure 3, the similarity and diversity measures are inversely proportional, consistent with theoretical expectations where higher similarity tends to obtain lower diversity and vice versa. Despite the negative correlation, neither similarity nor diversity can comprehensively explain the effectiveness of augmentations. As presented in Figure 4, achieving the optimal performance for different datasets does not necessarily require DA methods with the highest or lowest similarity or diversity values. DA methods with either the highest or the lowest similarity or diversity tend to yield inferior performance, which means that any of these two measures can not comprehensively evaluate the efficacy of augmentation methods. For instance, rotating images by 60° and flipping them vertically may produce a high diversity value, but the final accuracies of these two augmentation methods are 36.9% and 39.2% on CIFAR-10, respectively. In contrast, commonly used augmentations such as *Crop* and *HorizaonalFlip* can achieve a much higher

accuracy of 96.2%. This indicates that excessively high similarity between the augmented and training sets may lead to overfitting because the augmented data is very similar to the training set. However, an excessively high diversity can also lead to the distribution of augmented data considerably different from that of the training set, making it challenging for deep models to learn the underlying distribution and consequently leading to suboptimal performance. Therefore, neither similarity nor diversity can comprehensively explain the effectiveness of augmentations. Similarity and diversity are complementary in investigating the effectiveness of DA methods.

While the best-performing methods on these three datasets exhibit different similarity-diversity characteristics, we observe that they are concentrated in a specific region of the similarity-diversity plane, referred to as the "candidate interval." Generally speaking, this region is characterized by a similarity ranging from approximately -2.5 to -1 and a diversity ranging from 800 to 900 using ResNet-50 as the embedding model. To facilitate the selection of DA methods based on the proposed measures, we propose a simple but effective probabilistic indicator function to highlight the area where the best methods are concentrated, which can be found in Appendix F. After further analyzing the similarity-diversity characteristics of the best-performing augmentations on these datasets, we find that the preference of these methods for similarity and diversity measures originates from the varying degrees of ease in fitting the datasets. As shown in Figure 3(a), the best-performing DA methods on CIFAR-10 are mainly concentrated in areas with relatively high similarity and diversity, which can be more clearly observed in Figure 4(a). However, it can be observed in Figure 3(b) that augmentations for CIFAR-100 with high diversity are more beneficial, which can also be verified in Figure 4(c) and Figure 4(d). This is because CIFAR-100 has a more significant number of classes and fewer training samples per class than CIFAR-10. We believe that fitting a model to distinguish 100 categories is more complex than ten categories and therefore requires more diverse samples to reach optimum, making high-diversity DA indispensable for CIFAR-100. As shown in Figure 3(c), the best DA methods for ImageNet achieve relatively high similarity and diversity measures, while higher diversity is more beneficial. This may be because the distribution of the training set and the test set of ImageNet are similar; to avoid overfitting, providing diverse samples under the same distribution is more beneficial. Consequently, a key to devising DA techniques is to balance similarity and diversity concerning the complexity of the dataset, which we will further discuss in Section 5. This is also why some DA



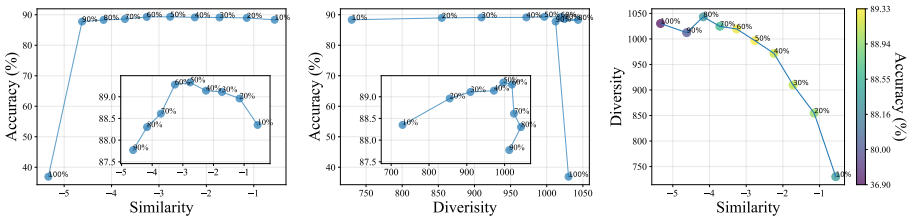
(a) CIFAR-10:Similarity (b) CIFAR-10:Diversity

(c) CIFAR-100:Similarity (d) CIFAR-100:Diversity

Figure 4: The relationship between similarity and accuracy and diversity and accuracy is shown. The Red dashed line indicates the accuracy of the baseline model with no augmentation at all.

techniques (e.g., GridMask and AutoAugment) utilize different parameters and strategies on different datasets.

Finally, we summarize some simple but effective DA methods on these three image datasets. On CIFAR-10, $basic\,()$+$randomerasing\,(80\%)$, $basic()$+$color(0.2, 10\%)$, $basic()$+$patchgaussian(20, 30\%)$, and $basic()$+$posterize(7, 100\%)$ can obtain test accuracy 4% higher than the baseline. On CIFAR-100, $basic\,()$+$contrast\,(0.7, 100\%)$, $basic\,()$+$patchgaussian\,(16, 10\%)$, and $basic\,()$+$color\,(0.2, 10\%)$ can achieve test accuracy 13.61% higher than the baseline, and 1.46% higher than widely used transform $RandomCrop$+$HorizonalFlip$. On ImageNet, $AutoContrast(100\%)$, $Cutout(30, 100\%)$, and $translateX\,(20, 100\%)$ can achieve test accuracy 0.22% higher than the baseline. For the details of the data augmentation methods above, please refer to Appendix C.



(a) Similarity (b) Diversity (c) Similarity and Diversity

Figure 5: By adjusting the ratio of *Rotate*(60°) on CIFAR-10 from 10% to 90%, we show the changing trend of similarity and diversity measures.

4.4. Accuracy with Similarity or Diversity

In this section, we conduct a case study to demonstrate that the proposed measures are consistent with the extent of variation introduced by augmentations. Therefore, the proposed measures can quantitatively indicate the extent of variation introduced by data augmentation approaches, which can then be utilized to guide the parameter tuning of augmentation methods.

As depicted in Figure 5, we evaluate the similarity and diversity measures of the augmentation *rotate* (60°) on CIFAR-10 across different ratios ranging from 10% to 100%. By manipulating the proportion of the augmented data, we can adjust the extent of variation introduced by augmentation techniques. For example, the ratio of 10% means that only 10% of the total data is rotated by 60°, while the remaining 90% is the original data. It can be observed in Figure 5(a) and Figure 5(b) that, as the ratio of the augmented data increases, the similarity measure decreases while the diversity measure increases. Moreover, Figure 5(c) shows that the accuracy first increases and then decreases as the ratio increases. The accuracy is the highest when the ratio is between 50% and 60%. When the similarity and diversity measures are both relatively high, the best generalization performance is obtained, consistent with the results in Figure 3(a).

However, when a small amount of augmented data is used, the augmented dataset is highly similar to the original training set, leading to overfitting and poor generalization performance of deep models. As the ratio of augmented data increases, the augmented dataset becomes more diverse, and the generalization performance of deep models improves. Meanwhile, it is also important to note that if the distribution of the augmented training data deviates too much from the original training and test sets, the model's

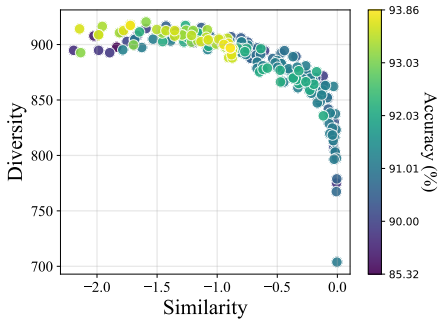


Figure 6: CIFAR-10 transferred test accuracy in the similarity-diversity plane.

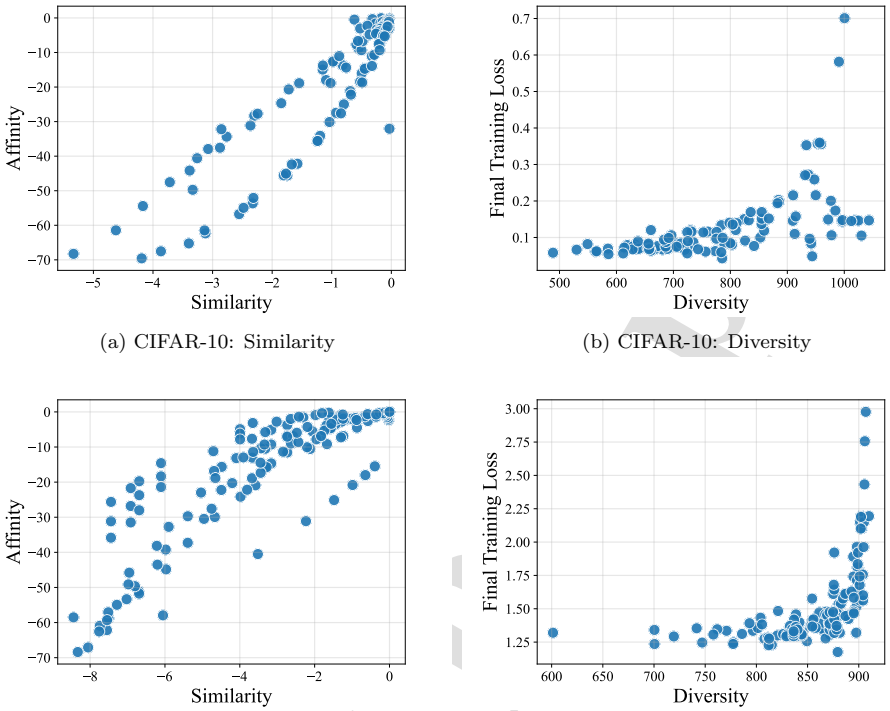
performance may suffer. For instance, when the ratio exceeds 60%, the test accuracy drops sharply. Especially when the ratio reaches 100%, the test accuracy is the lowest.

Therefore, the proposed measures have the potential to be used as indicators to adjust the extent of variation introduced by augmentations in terms of similarity and diversity.

4.5. Transfer Learning

In the context of data augmentation, the effectiveness of DA methods is closely linked to the model's ability to extract useful features and generalize well on unseen data. In addition to evaluating the performance of models based on test accuracy, the effectiveness of DA methods can also be assessed with transfer learning [40, 41]. To further validate the efficacy of our proposed measures, we associate the transferability of deep models with our proposed quantitative measures. Specifically, we pre-train models on the CIFAR-100 dataset using different augmentations and fine-tune them on the CIFAR-10 dataset. The similarity and diversity are calculated on the CIFAR-100 dataset, and the accuracy is the transferred test accuracy on the CIFAR-10 test set. Theoretically, better DA methods indicate higher transferred test accuracy.

As shown in Figure 6, we illustrate the transferred test accuracy of 220 different DA methods on the similarity-diversity plane. We can see that our proposed measures have demonstrated an excellent ability to investigate the effectiveness of DA methods in terms of model transferability. For instance, compared with the results in Figure 3(b), the test accuracy on CIFAR-100



(a) CIFAR-10: Similarity

(b) CIFAR-10: Diversity

(c) ImageNet: Similarity

(d) ImageNet: Diversity

Figure 7: The relationship between our proposed measures and affinity, FTL on CIFAR-10 and ImageNet datasets.

and the transferred test accuracy on CIFAR-10 share a similar trend. Meanwhile, the DA methods that perform better are mainly concentrated in high-diversity regions, which suggests that models trained with high-diversity DA methods have a stronger feature extraction ability. Therefore, on CIFAR-100, augmentations that achieve relatively high diversity measures are preferred.

Through transfer learning, we have further demonstrated the efficacy of the proposed measures in evaluating the effectiveness of DA methods.

4.6. Comparison with Other Measures

In this section, we compare our proposed measures with the affinity and final training loss (FTL) presented in [18] on both CIFAR-10 and ImageNet datasets. Through experimental results, we show that although similarity

Table 2: The Pearson correlation coefficient and Spearman correlation coefficient between the proposed measures and affinity on CIFAR-10 and ImageNet. **FTL**: final training loss. **PCC**: Pearson correlation coefficient. **SCC**: Spearman correlation coefficient.

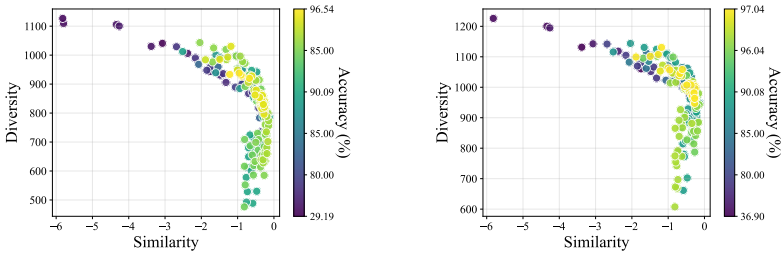
Dataset	Metrics	PCC	SCC
CIFAR-10	Similarity&Affinity	0.91	0.89
	Diversity&FTL	0.75	0.58
ImageNet	Similarity&Affinity	0.86	0.87
	Diversity&FTL	0.70	0.78

and diversity measures are distance-based measures of the datasets, they are closely related to the performance of DA methods on model training and obtain higher practical values than previous works.

Specifically, as presented in Figure 7(a), there is a high positive correlation between the similarity and affinity measure. Meanwhile, we can also observe in Figure 7(b) that DA methods with high diversity values tend to obtain high FTL. To provide a deeper analysis of the correlation, we also calculate the Pearson correlation coefficient [42] and Spearman correlation coefficient [43] between our proposed measures and others in Table 2. As presented in Table 2, there is a strong positive correlation between similarity measure and affinity, as well as diversity and FTL. The p-value of all the correlation coefficients is lower than 1×10^{-9}, indicating high confidence in the strong positive correlation. Therefore, our proposed measures not only investigate the effectiveness of DA methods on the model training but also provide practical values for DA techniques.

However, affinity and FTL require the entire training process, which brings high computation costs and low efficiency, making it difficult to be utilized for the design of DA methods before model training. Meanwhile, the best augmentation policies prefer jointly optimizing affinity and FTL across various datasets [18], regardless of the differences between datasets. Therefore, the nature of the affinity and FTL leads to a lack of transparent methodology of how to utilize these metrics to design augmentation strategies in practice.

In contrast, our proposed distance-based measures formally reveal that the effectiveness of DA techniques originates from generating data with varying degrees of similarity and diversity that are appropriate for specific datasets. Since only one embedding model for each dataset needs to be trained, the pro-



(a) ResNet-50 (b) Wide-ResNet-28-10

Figure 8: Similarity measure is calculated between the augmented set and the test set on CIFAR-10 using both ResNet-50 and Wide-ResNet-28-10. The best-performing DA methods tend to obtain high similarity values.

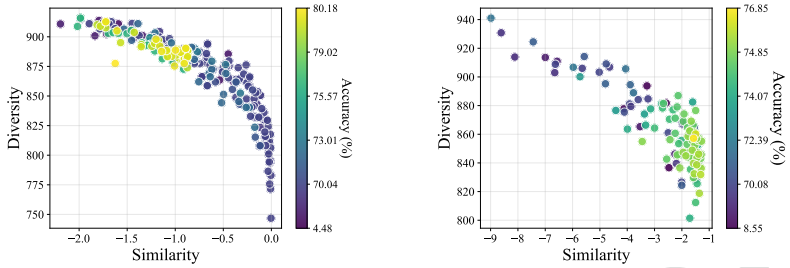
posed measures can be calculated much more efficiently. More importantly, similarity and diversity can serve as a guide for the design of augmentation approaches, parameter tuning, and initial validation of the efficacy of augmentations. For example, by tuning the parameters of a DA method, the similarity or diversity can be targeted to achieve the best performance possible by pushing the method within the "candidate interval".

For more details of the comparison between our proposed measures with other measures commonly used for evaluating synthetic images, including structural similarity index measure (SSIM) [44], inception score (IS) [15], Frechet inception distance (FID) [16], and peak signal-to-noise ratio (PSNR) [45], please refer to Appendix G.

4.7. The Similarity between the Augmented Set and Test Set

In this section, we aim to further validate the effectiveness of our proposed measures by assessing the similarity between the augmented dataset and the test set, despite the test set being entirely unseen during the design of augmentation approaches. The underlying principle suggests that models trained on a training set more similar to the test set tend to exhibit better generalization performance.

As depicted in Figure 8, our proposed measures demonstrate that when the distribution of the augmented dataset and test set is more similar, the model will generalize better on that test set. Contrary to the observations in Figure 3(a), the results indicate that DA methods achieving the highest test accuracy tend to possess both the highest similarity and relatively high



(a) DenseNet121 [46] is utilized as the embedding model on CIFAR-100.

(b) Vision Transformer [47] is utilized as the embedding model on ImageNet.

Figure 9: The effect of embedding models.

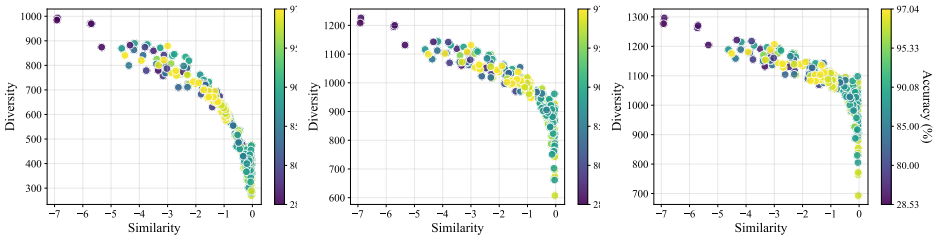
diversity values. In practice, since the test set is unknown during data augmentation design, based on the fundamental assumption in deep learning that the training set and the test set are independent and identically distributed, it is crucial to provide augmented data that resembles the training set. However, to effectively mitigate overfitting, optimal DA methods should also incorporate relatively high diversity values by introducing variations.

Thus, our proposed measures are reasonable and effective in comprehending the effectiveness of DA approaches.

4.8. Ablation Study

4.8.1. The Effect of the Embedding Models

To study the impact of different embedding models, in Figure 9(a) and Figure 9(b), we train a DenseNet121 [46] model on CIFAR-100 and a Vision Transformer [47] model on ImageNet to obtain the feature map, respectively. The embedding results on the datasets using ResNet-50 can be seen in Figure 3. It can be observed that the candidate interval exists no matter which embedding model is used. Specifically, similar to the results in Figure 3, there is a generally inverse proportionality between the similarity and diversity measures, and most of the best-performing methods are concentrated in a specific area of the similarity-diversity plane. In Figure 9(a), CIFAR-100 also favors DA methods with high diversity. The average Spearman correlation coefficient between measures in Figure 3(b) and Figure 9(a) is as high as 0.91. Similar to the results shown in Figure 3(c), in Figure 9(b), the best DA methods for ImageNet are also concentrated in the region where both similarity and diversity are relatively high. Despite using different embedding



(a) $\theta = 20\%$ (b) $\theta = 60\%$ (c) $\theta = 80\%$

Figure 10: Quantifying results with different settings of θ.

models, we can still obtain similar results; the best-performing methods are still concentrated in the "candidate interval". Consequently, the similarity-diversity metric can still be used on any model, just the specific range of the "candidate interval" may vary slightly. We will also show in Appendix F a simple unified probabilistic metric to identify a narrow "candidate interval" containing the best-performing DA methods.

4.8.2. The Effect of the Parameter θ

As described in Equation (15) in Section 3.2, parameter θ determines the number of chosen eigenvalues and corresponding eigenvectors. Here, we investigate the effect of the parameter θ on the quantification of the diversity of DA methods for CIFAR-10 with various θ values, including 20%, 40%, 60%, and 80%. The results obtained using $\theta = 40\%$ are shown in Figure 3(a).

As demonstrated in Figure 10, increasing the value of θ leads to a decrease in the variation of diversity measure among different data augmentation methods. When θ is set to a smaller value, such as 20%, the diversity measure does not fully capture the critical information in the dataset as insufficient eigenvectors are retained. Conversely, when θ exceeds 60%, the diversity measure fails to capture the difference in diversity among various data augmentation methods, as it considers the information related to small eigenvalues that do not contribute much to the diversity of the augmented data. Therefore, we select $\theta = 40\%$ as the optimal value for CIFAR-10.

Parameter Setting. The parameter θ is essential for the diversity measure yet easy to set. In the calculation of diversity measure, the feature map A_i is of the size $m \times n$, where m represents the total number of classes and n represents the number of samples in each category. Therefore, the

covariance matrix S in Equation (12) is of the size $m \times m$. The number of eigenvalues and eigenvectors of the covariance matrix S are closely related to m. For larger covariance matrices, more principal components are typically required to explain the variance in the data, resulting in an increased need for principal components. This is because larger covariance matrices typically contain more subtle variations that require additional principal components to capture these details and provide a more accurate data representation. For instance, in the case of the MNIST and CIFAR-10 dataset, which has an m value of 10, we have set θ to 40%. If additional eigenvalues and eigenvectors are selected, there is a higher likelihood of introducing noisy information. In contrast, the CIFAR-100 dataset has an m value of 100, and we have set θ to 80% accordingly, as the primary information of the dataset is contained in more eigenvalues and eigenvectors. Consequently, θ can be easily set by referencing commonly used datasets in our research, including MNIST, CIFAR-10, CIFAR-100, and ImageNet.

5. Discussion and Future Work

In this work, we propose two model training-agnostic quantitative measures that can effectively investigate the performance of DA methods. These measures are designed to uncover the key factors contributing to successful DA methods without large-scale model training. Our investigation reveals that different datasets have different similarity-diversity preferences for DA methods and that the effectiveness of DA methods originates from the similarity and diversity of the augmented data. Prior research on data augmentation has often focused explicitly on one of these measures while implicitly accounting for the other. For instance, Cutout, HaS, and GridMask ensure the similarity of augmented data by masking sub-regions in the original image and increasing diversity through changes in the masking regions. To enhance diversity, automated data augmentation uses combinations of various augmentations that involve minor changes to maintain similarity. Our findings underscore the importance of balancing similarity and diversity to design effective DA methods. The proposed similarity and diversity measures can explicitly capture the adjustments of these properties. Additionally, we observe that high-accuracy augmentations tend to be concentrated in a specific region of the similarity-diversity plane. Augmentations in the "candidate interval" are more likely to achieve high accuracy, which informs the design of future augmentations by adjusting the parameters of a given method to

fall within the "candidate interval". Besides strengths, the limitations of our method should also be mentioned.

5.1. The Impact of Dataset Complexity on Data Augmentation Methods

While the proposed metrics directly reflect the effectiveness of DA methods, we believe that the dataset complexity plays a vital role in determining the preferences for the DA method. However, the dataset complexity is challenging to quantify because it involves multiple factors such as dataset size, number of categories, sample size per category, noises, label quality, etc. Our approach tries to decompose the dataset complexity by similarity and diversity measures. These two measures work in an adversarial manner, and both are indispensable for a complete evaluation. Therefore, future work should further analyze dataset complexity's influence on DA approaches.

5.2. The Range of Similarity and Diversity Values

While our work has demonstrated its efficacy across various benchmark datasets, it's important to underscore that the range of similarity and diversity values may vary for certain specific datasets, such as medical MRI images. This can lead to an inconsistency in the "candidate interval". This potential limitation can be mitigated by referring to our unified metric in Appendix F, which empirically harnesses existing DA methods to adaptively pinpoint an approximate range for the candidate interval. Through this approach, we aim to minimize the impact of the limitation by providing a broader understanding of the expected metric values across diverse dataset scenarios. In the future, based on the similarity and diversity measures, we would explore the development of an even more robust unified metric to evaluate the effectiveness of DA methods.

5.3. Sample Sizes in the Diversity Measure

In the calculation of the diversity measure, the sample size within each category significantly influences the outcomes of PCA [48]. Specifically, let the feature map be denoted as A_i with the size $m \times n$, where m is the total number of classes and n is the total number of samples in each category. We define the sample abundance s_a as $s_a = \frac{n}{m}$, which highlights the abundance of samples in each category. Determining the appropriate value of s_a is challenging but crucial to PCA. There are only a few studies focusing on the sample size of PCA [49, 50]. Furthermore, empirical evidence from studies such as [51, 52] suggests that the minimum sample size should be larger than

five times the number of variables, recommending a lower bound of 5 for s_a. Consequently, the efficacy of the diversity measure is constrained when the sample size per category is exceedingly small. This limitation implies that our method may not be suitable in scenarios characterized by minimal samples per category. Therefore, future work could delve deeper into the evaluation of data augmentation techniques specifically on smaller sample datasets.

6. Conclusion

In this paper, we proposed efficient quantitative measures that enable a comprehensive investigation into the effectiveness of data augmentation methods by considering both similarity and diversity aspects. Departing from the traditional evaluation of DA methods based solely on model performance, our approach offers a model training-agnostic and easily implementable framework. To validate the efficacy of our proposed measures, we conduct extensive experiments on multiple benchmark datasets, including MNIST, CIFAR-10, CIFAR-100, and ImageNet. The experimental results demonstrate that the proposed measures provide a practical framework for investigating the effectiveness of DA methods by decomposing the performance into similarity and diversity. Furthermore, the proposed measures can be employed to understand the effectiveness of DA methods, guide the design and parameter tuning of such methods, and offer an efficient initial validation of DA methods' efficacy. Moving forward, future research directions should explore the implications of our quantitative measures for the design of novel augmentation methods.

Acknowledgments

This work was supported in part by the STI 2030-Major Projects of China under Grant 2021ZD0201300, and by the National Science Foundation of China under Grant 62276127.

References

[1] M. Xu, S. Yoon, A. Fuentes, D. S. Park, A comprehensive survey of image augmentation techniques for deep learning, Pattern Recognition 137 (2023) 109347.

[2] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, F. Shen, Image data augmentation for deep learning: A survey, arXiv preprint arXiv:2204.08610 (2022).

[3] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation strategies from data, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 113–123.

[4] E. D. Cubuk, B. Zoph, J. Shlens, Q. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Proc. Adv. Neural Inf. Process. Syst., Vol. 33, 2020, pp. 18613–18624.

[5] S. Yang, J. Li, T. Zhang, J. Zhao, F. Shen, Advmask: A sparse adversarial attack-based data augmentation method for image classification, Pattern Recognition (2023) 109847.

[6] P. Chen, S. Liu, H. Zhao, J. Jia, Gridmask data augmentation, arXiv preprint arXiv:2001.04086 (2020).

[7] Y. Pang, J. Lin, T. Qin, Z. Chen, Image-to-image translation: Methods and applications, IEEE Transactions on Multimedia 24 (2021) 3859–3881.

[8] T. DeVries, G. W. Taylor, Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552 (2017).

[9] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[11] S. Wickramanayake, W. Hsu, M. L. Lee, Explanation-based data augmentation for image classification, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Proc. Adv. Neural Inf. Process. Syst., Vol. 34, Curran Associates, Inc., 2021, pp. 20929–20940.

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

[13] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.
URL https://openreview.net/forum?id=r1Ddp1-Rb

[14] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proc. AAAI, Vol. 34, 2020, pp. 13001–13008.

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Advances in neural information processing systems 29 (2016).

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[18] R. Gontijo-Lopes, S. Smullin, E. D. Cubuk, E. Dyer, Tradeoffs in data augmentation: An empirical studyl, in: Proc. Int. Conf. on Learning Representations, 2020.

[19] D. Alvarez-Melis, N. Fusi, Geometric dataset distances via optimal transport, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Proc. Adv. Neural Inf. Process. Syst., Vol. 33, Curran Associates, Inc., 2020, pp. 21428–21439.

[20] N. Courty, R. Flamary, A. Habrard, A. Rakotomamonjy, Joint distribution optimal transportation for domain adaptation, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Proc. Adv. Neural Inf. Process. Syst., Vol. 30, Curran Associates, Inc., 2017, p. 3733–3742.
URL https://proceedings.neurips.cc/paper/2017/file/0070d23b06b1486a538c0eaa45dd167a-Paper.pdf

[21] I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. Roy. Soc. A, Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202.

[22] A. E. Magurran, Measuring biological diversity, Current Biology 31 (19) (2021) R1174–R1177.

[23] C. Gong, D. Wang, M. Li, V. Chandra, Q. Liu, Keepaugment: A simple information-preserving data augmentation approach, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 1055–1064.

[24] K. K. Singh, Y. J. Lee, Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), IEEE, 2017, pp. 3544–3553.

[25] R. Bravin, L. Nanni, A. Loreggia, S. Brahnam, M. Paci, Varied image data augmentation methods for building ensemble, IEEE Access 11 (2023) 8810–8823.

[26] A. Ç. Demir, S. Caton, P. Dondio, Subnetwork ensembling and data augmentation: Effects on calibration, Expert Systems (2023) e13252.

[27] E. J. Snider, S. I. Hernandez-Torres, R. Hennessey, Using ultrasound image augmentation and ensemble predictions to prevent machine-learning model overfitting, Diagnostics 13 (3) (2023) 417.

[28] S. Lim, I. Kim, T. Kim, C. Kim, S. Kim, Fast autoaugment, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Proc. Adv. Neural Inf. Process. Syst., Vol. 32, Curran Associates, Inc., 2019.
URL https://proceedings.neurips.cc/paper/2019/file/6add07cf50424b14fdf649da87843d01-Paper.pdf

[29] K. Tian, C. Lin, M. Sun, L. Zhou, J. Yan, W. Ouyang, Improving auto-augment via augmentation-wise weight sharing, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Proc. Adv. Neural Inf. Process. Syst., Vol. 33, Curran Associates, Inc., 2020, pp. 19088–19098.
URL https://proceedings.neurips.cc/paper/2020/file/dc49dfebb0b00fd44aeff5c60cc1f825-Paper.pdf

[30] L. V. Kantorovich, On the translocation of masses, in: Dokl. Akad. Nauk. USSR (NS), Vol. 37, 1942, pp. 199–201.

[31] L. Chizat, G. Peyré, B. Schmitzer, F.-X. Vialard, Unbalanced optimal transport: Dynamic and kantorovich formulations, Journal of Functional Analysis 274 (11) (2018) 3090–3123.

[32] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Proc. Adv. Neural Inf. Process. Syst., Vol. 26, Curran Associates, Inc., 2013, pp. 2292–2300.
URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf

[33] G. Peyré, M. Cuturi, Computational optimal transport: With applications to data science, Foundations and Trends® in Machine Learning 11 (5-6) (2019) 355–607. doi:10.1561/2200000073.
URL http://dx.doi.org/10.1561/2200000073

[34] A. A. Rahane, A. Subramanian, Measures of complexity for large scale image datasets, in: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 2020, pp. 282–287.

[35] P. Xie, A. Singh, E. P. Xing, Uncorrelation and evenness: a new diversity-promoting regularizer, in: Proc. Int. Conf. Mach. Learn., PMLR, 2017, pp. 3811–3820.

[36] S. Bailey, Principal component analysis with noisy and/or missing data, Publications Astronomical Soc. Pacific 124 (919) (2012) 1015.

[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[38] Y. LeCun, C. Cortes, C. J. Burges, The MNIST database of handwritten digits, 1998, URL http://yann. lecun. com/exdb/mnist 10 (34) (1998) 14.

[39] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Machine Learning Research 9 (11) (2008).

[40] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359. doi:10.1109/TKDE.2009.191.

[41] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (1) (2021) 43–76. doi:10.1109/JPROC.2020.3004555.

[42] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, Noise reduction in speech processing (2009) 1–4.

[43] L. Myers, M. J. Sirois, Spearman correlation coefficients, differences between, Encyclopedia of statistical sciences 12 (2004).

[44] Z. Wang, A. C. Bovik, A universal image quality index, IEEE signal processing letters 9 (3) (2002) 81–84.

[45] Q. Huynh-Thu, M. Ghanbari, Scope of validity of psnr in image/video quality assessment, Electronics letters 44 (13) (2008) 800–801.

[46] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[48] J. W. Osborne, A. B. Costello, Sample size and subject to item ratio in principal components analysis, Practical Assessment, Research, and Evaluation 9 (1) (2004) 11.

[49] E. Saccenti, M. E. Timmerman, Approaches to sample size determination for multivariate data: Applications to pca and pls-da of omics data, Journal of proteome research 15 (8) (2016) 2379–2393.

[50] S. S. Shaukat, T. A. Rao, M. A. Khan, Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure, Ekológia (Bratislava) 35 (2) (2016) 173–190. doi:doi:10.1515/eko-2016-0014.
URL https://doi.org/10.1515/eko-2016-0014

[51] N. O'Rourke, L. Hatcher, A Step-By-Step Approach to Using SAS System for Factor Analysis and Structural Equation Modeling, 2013.

[52] F. B. Bryant, P. R. Yarnold, Principal-components analysis and exploratory and confirmatory factor analysis. (1995).

Investigating the Effectiveness of Data Augmentation from Similarity and Diversity: An Empirical Study

Suorong Yang[a,b], Suhan Guo[a,c], Jian Zhao[d], Furao Shen [a,c,*]

[a]*State Key Laboratory for Novel Software Technology, Nanjing University, China*
[b]*Department of Computer Science and Technology, Nanjing University, China*
[c]*School of Artificial Intelligence, Nanjing University, China*
[d]*School of Electronic Science and Engineering, Nanjing University, China*

Abstract

Data augmentation has emerged as a widely adopted technique for improving the generalization capabilities of deep neural networks. However, evaluating the effectiveness of data augmentation methods solely based on model training is computationally demanding and lacks interpretability. Moreover, the absence of quantitative standards hinders our understanding of the underlying mechanisms of data augmentation approaches and the development of novel techniques. To this end, we propose interpretable quantitative measures that decompose the effectiveness of data augmentation methods into two key dimensions: similarity and diversity. The proposed similarity measure describes the overall similarity between the original and augmented datasets, while the diversity measure quantifies the divergence in inherent complexity between the original and augmented datasets in terms of categories. Importantly, our proposed measures are model training-agnostic, ensuring efficiency in their calculation. Through experiments on several benchmark datasets, including MNIST, CIFAR-10, CIFAR-100, and ImageNet, we demonstrate the efficacy of our measures in evaluating the effectiveness of various data augmentation methods. Furthermore, although the proposed measures are straightforward, they have the potential to guide the design and parameter tuning of data augmentation techniques and enable the validation of data augmentation methods' efficacy before embarking on

*Corresponding author. E-mail address: frshen@nju.edu.cn (F. Shen).
Email addresses: sryang@smail.nju.edu.cn (S. Yang), shguo@smail.nju.edu.cn (S. Guo), jianzhao@nju.edu.cn (J. Zhao), frshen@nju.edu.cn (F. Shen)

large-scale model training.

Keywords: Data augmentation, interpretability, generalization, deep learning, image classification.

Highlights

Investigating the Effectiveness of Data Augmentation from Similarity and Diversity: An Empirical Study

Suorong Yang, Suhan Guo, Jian Zhao, Furao Shen

- We propose novel quantitative measures agnostic to model training to investigate the effectiveness of DA methods based on similarity and diversity. Through experiments, we demonstrate that these measures provide a framework for assessing the effectiveness of DA methods based on the similarity and diversity of the augmented data.
- Our quantitative measures formulate the similarity and diversity measures for DA techniques. Through comparisons of our quantifying results with the practical effectiveness of DA methods, we find that the importance of similarity and diversity varies across different datasets.
- The proposed measures are conducted in feature space, rather than raw pixel space, which helps explain why some visually meaningless data augmentation methods are still effective.
- While the top-performing DA methods differ across datasets, our similarity-diversity plane reveals that the majority of these methods are concentrated within a particular region, namely the "candidate interval". The interval encompasses DA methods with the highest potential for achieving optimal performance.
- Our study has the potential to provide a more comprehensive understanding of the mechanisms behind DA methods, as well as guide the design and parameter tuning of DA methods. Additionally, our study can offer an efficient preliminary validation of augmentation method efficacy, saving computational resources and time costs in large-scale model training.

Suorong Yang received the B.S. degree in the department of Computer Science and Technology from Nanjing University in 2019. He is pursuing the Ph.D. degree in Computer Science and Technology at Nanjing University under the supervision of Prof. Furao Shen. His research interests include computer vision, data augmentation, generative adversarial network, etc.

Suhan Guo received the B.A degree from The Johns Hopkins University, Maryland, U.S., in 2016 and MPH degree from New York University, New York, U.S., in 2019. She is currently pursuing the Ph.D. degree at Nanjing University, Nanjing, China. Her current research interests include medical image analysis.

Jian Zhao (Senior Member, IEEE) received the B.S. degree from Nanjing University, Nanjing, China, the M.Sc. degree from the Hamburg University of Technology, Hamburg, Germany, and the Dr. Sc. degree in electrical engineering from the Swiss Federal Institute of Technology (ETH) Zurich, Switzerland. From 2010 to 2015, he was a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore. Currently, he is an Associate Professor with the School of Electronic Science and Engineering, Nanjing University. His research interests include deep neural networks, mathematical optimization, and wireless communication networks. Dr. Zhao was honored with the Dengfeng Scholars Program of Nanjing University in 2015, IEEE Globecom 2008 Best Paper Award, and the 2009 Chinese Government Award for Outstanding Self-Financed Students Abroad.

Furao Shen (Member, IEEE) received the B.Sc. and M.Sc. degrees in mathematics from Nanjing University, Nanjing, China, in 1995 and 1998, respectively, and the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2006. He is currently a Full Professor of computer science and technology with Nanjing University. His current research interests include neural computing and robotic intelligence.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: