PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

Towards Bridging Sample Complexity and Model Capacity

Shibin Mei, Chenglong Zhao, Bingbing Ni*, Shengchao Yuan

Shanghai Jiao Tong University, Shanghai 200240, China
{adair327, cl-zhao, nibingbing, sc_yuan}@sjtu.edu.cn

Abstract

In this paper, we give a new definition for sample complexity, and further develop a theoretical analysis to bridge the gap between sample complexity and model capacity. In contrast to previous works which study on some toy samples, we conduct our analysis on more general data space, and build a qualitative relationship from sample complexity to model capacity required to achieve comparable performance. Besides, we introduce a simple indicator to evaluate the sample complexity based on continuous mapping. Moreover, we further analysis the relationship between sample complexity and data distribution, which paves the way to understand the present representation learning. Extensive experiments on several datasets well demonstrate the effectiveness of our evaluation method.

1 Introduction

Over the past decades, the boom in deep neural networks (DNNs) has promoted the rapid development of artificial intelligence communities (Goodfellow et al. 2014; Girshick 2015; Long, Shelhamer, and Darrell 2015). Along with these advances, a large amount of newly proposed datasets (Krizhevsky, Hinton et al. 2009; Deng et al. 2009) and deep neural network structures (Simonyan and Zisserman 2014; He et al. 2016) further promote the breakthrough of deep learning. However, there exist two common concerns in the most basic classification tasks, that is, how to evaluate the sample complexity of a given dataset and how to build the relationship between datasets sample complexity and model capacity.

To tackle these problems, various methods have been proposed to discuss the connection between dataset and model. Some researchers (Branchaud-Charron, Achkar, and Jodoin 2019; Ho and Basu 2002) start from assessing different datasets by a series of complexity measures, which are then used to evaluate the complexity of classification problem. The class entanglement is evaluated based on the assumption that discriminative and separated classes are conducive to classification tasks. Ho et al. (Ho and Basu 2002) propose to evaluate the sample complexity by a combination of 12 descriptors, which is effective for small non-image two-class datasets. These metrics are then generalized by (Orriols-Puig, Macia, and Ho 2010; Sotoca, Mollineda, and Sánchez 2006), which build on strong assumptions for data distribution and complex calculations. Branchaud et al. (Branchaud-Charron, Achkar, and Jodoin 2019) develop CSG metric for dataset complexity assessment based on class overlap and spectral clustering of image datasets. Schmidt et al (Schmidt et al. 2018) utilize a simple setting of mixture of two spherical Gaussians with one component per class. They theoretically analyze the model capacity required for standard training and adversarial training, and further extend their conclusion to complex datasets.

Instead of discussing sample complexity from the whole dataset, we focus on the most representative points in the dataset. These key-points, defined as dominating data, can serve as data anchors, which play an essential role in model fitting. Concretely, the whole dataset can be divided as several clusters based on point distance, and the dominating data can be extracted from these clusters. We assume that if the model fits on these dominating data, the learned patterns can easily generalize to the whole dataset. Similar viewpoints can also be embodied in support vector machine (Noble 2006) and prototype learning (Yang et al. 2018; Zhang et al. 2018). In this paper, we use the concept of dominating data as a medium to explore the relationship between sample complexity and model capacity.

We argue that the amount of dominating data is a good metric for the sample complexity. Furthermore, we give the definition of model capacity, and theoretically give the upper and lower bounds of the model capacity required for training under a certain amount of dominant data. Inspired by continuous mapping, we also provide a qualitative evaluation for sample complexity of a given dataset based on clusters. Besides, considering the dominating data wrapped by clusters can be regarded as unit distribution, we can further analysis the relationship between sample complexity and data distribution, and give a new method to compute the sample complexity by inter-class and intra-class distribution. We think this investigation has broad implications for present representation learning (Bengio, Courville, and Vincent 2013).

Our contributions can be summarized as follows:

- We give a new definition of sample complexity and theoretically build the relationship between sample complexity and model capacity.

*Corresponding author.
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

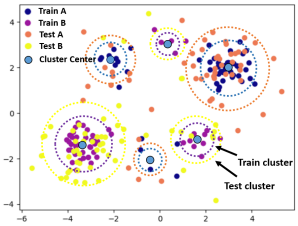
- We further discuss the relationship between sample complexity and data distribution, and measure the sample complexity based on representation learning.
- Extensive experiments well demonstrate that the proposed method can effectively characterize the relationship between sample complexity and model capacity.

2 Theoretical Analysis

2.1 Preliminary

We view the data set as some data clusters according to the distance between data samples. Based on this, firstly, we argue that some small data clusters or outliers (i.e., hard examples (Smirnov et al. 2018)) hinder the generalization of trained models. Namely, there exist few samples surrounding these small clusters or outliers, which resulting in the limited marginal benefit of model generalization during the fitting process. However, for some big clusters, the central of clusters are surrounded by a large amounts of neighboring points. Secondly, some representative key-points can be extracted from clusters of original dataset to enable a favorable performance. As observed in Fig.1, only using central points of clusters can achieve comparable accuracy compared with using the whole train set. Therefore, fitting these key-points can be deemed as an important aspect of model learning. Recent researches on prototype learning (Yang et al. 2018; Zhang et al. 2018; Liu, Song, and Qin 2020) also show that the most representative exemplars can serve as anchors to facilitate efficient learning. Similar standpoint is also revealed in dataset distillation (Hinton, Vinyals, and Dean 2015; Wang et al. 2018).

As discussed above, we want to know how many samples can form a skeleton of the data set which can guarantee the satisfactory performance of models training. To solve this problem, we build a new definition of sample complexity, and further to discuss the relationship between the sample complexity and model capacity theoretically, which paves the movement to understand the current popular deep learning.



Fitting Set	Test Acc
Train Set	91.5
Center Point	94.7

Figure 1: We conduct a simple validation to fit the points of two classes (A and B) in the left graph using SVM classifier with default settings. We respectively fit the central point of each cluster(blue point) and the complete training set (blue and purple points). The variance of clusters in test set is set as two times of that in train set. The size of dotted circle represents the size of clusters.

2.2 Sample Complexity

Many previous methods have been proposed to explore the relationship between sample complexity and model capacity from the perspective of model fitting and generalization (Schmidt et al. 2018; Carmon et al. 2019). Branchaud et al. (Branchaud-Charron, Achkar, and Jodoin 2019) propose spectral metric for dataset complexity assessment based on class overlap for image datasets compared with Ho et al. (Ho and Basu 2002) for non-image datasets. Schmidt et al. (Schmidt et al. 2018) theoretically analysis the generalization of the a linear classifier trained on a toy data set, i.e, a mixture of two spherical Gaussians. However, there exist some limitations on this works: **1).** It is uncertain whether the conclusion established in the linear model on the toy data (low-dimensional gaussian distribution) can be directly extended to high-dimensional data and nonlinear neural network. **2).** Samples drawn from Gaussian distribution tend to have strong distinguish ability, which can be easily classified by a naive linear model. However, the natural data (e.g. images) or training data sets obtained in the wild are really complex, which means that these samples cannot obey to a simple distribution, and even have a complicated manifold structure. Thus, there exits a big gap between the toy data set and real data. As shown in Tab.1, samples from real datasets share less similarity compared with low-dimension Gaussian dataset, even for CNN feature embedding.

To solve these problems, we propose to directly fit representative data points in high-dimensional space, and give a novel definition of sample complexity. Thus, our work is more general compared with prior arts.

CosSim	Gaussian	MNIST	CIFAR10
Original	0.308	0.536	0.825
Embedding	-	0.641	0.889

Table 1: We compute the average cosine similarity on original samples and embedded CNN features in MNIST (LeCun et al. 1998), CIFAR10 (Krizhevsky, Hinton et al. 2009) and constructed Gaussian dataset. We construct Gaussian dataset following Schidmt et al.($\sigma = 1.0$) (Schmidt et al. 2018) and we use ResNet34 (He et al. 2016) and LeNet5 for CIFAR10 and MNIST feature embedding respectively.

For the sake of discussion, we first give some basic definitions as follows,

Definition 1 *(Neighboring points) Let x_1, x_2 be two data points drawn from whole dataset and $\Phi(x_1), \Phi(x_2)$ be their corresponding embedding, such as identity mapping, CNN. Given a distance threshold ϵ, we call x_1, x_2 are neighboring points or x_1 has a neighbor x_2 if,*

$$d(\Phi(x_1), \Phi(x_2)) < \epsilon \tag{1}$$

where d is a distance metric.

Definition 2 *(Cluster) For a group of data points $\mathcal{G} = \{x_1, x_2, ...\}$ as the subset of the whole dataset, we define $\mathcal{G}$ as a cluster if and only if any two points in $\mathcal{G}$ are neighboring points.*

The dataset can hence be divided to several clusters and each data point belongs to a single cluster. Inspired by the concept of dominating set in graph theory (for graph $G = (V, E)$, the dominating set is defined as a subset U of vertex set V that each vertex in $V - U$ has at least one neighbor in U), we obtain the definitions of dominating data and sample complexity of a dataset, as follows,

Definition 3 *(Dominating data) For a given dataset D and corresponding clusters, we extract one point from each cluster thus forming a subset D_d, we call the subset D_d as the dominating data of this dataset.*

Definition 4 *(Sample complexity) The sample complexity S is interpreted as the scale of dominating data, i.e.*

$$S = |D_d| \tag{2}$$

We claim that once the model fits the dominating data, it can easily generalize the to the whole data set. So we can use dominating data as a metric to describe the sample (i.e., dataset) complexity. Our assumption is also reflected in other machine learning field. Introducing data augmentation in model training can significantly improve model generalization. Previous data augmentation methods, such as randomly flipping, randomly cropping, adding noise, cutout, can be seen as creating new data patterns compared with the original single style data sets. The newly crafted patterns diversify the space of data expansion, which increase the sample complexity, thus causing much model capacity. In adversarial training, there exists a trade-off between the generalization ability on natural samples and adversarial examples (Madry et al. 2017; Zhang et al. 2019). One rational explanation is that the adversarial examples far deviate from the original data distribution which is easy to generalize, forming more outlier data points as well as dominating data. It is laborious to obtain the generalization of these samples by learning smooth model as before. Moreover, the hard examples (Smirnov et al. 2018), which can be interpreted as samples with small clusters, are not conducive to the model learning. Out of distribution samples (OOD) (Krueger et al. 2021) can be regarded as no corresponding data cluster in the training set and domain adaption (Long et al. 2015) can be viewed as dominating data alignment.

2.3 Model Capacity for Dominating Data

In this section, we discuss the model capacity which need to fit all the dominating data from the perspective of combinatorics, to bridging the relationship between sample complexity and model capacity.

Let n be the amount of dominating data as well as sample complexity $\mathcal{D} : \mathbb{R}^d$. As illustrated above, these samples are relatively independent of each other, and the model needs to consume the fitting ability to memorize these points. For the convenience of the following derivation, we assume that the dominating data are composed of many sub datasets $\mathcal{S}$ with a capacity of w, where $w < n$, $\mathcal{S} \subset \mathcal{D}$. In addition, this assumption stems from the training set and test set in reality are commonly sampled from total data space, and the model is expected to maintain performance under different data combinations.

At the same time, let m be the number of distinct categories $\mathcal{Y}$. Then the number of all the possible mappings $\mathbb{R}^d \to \mathcal{Y}$ in a random data set is $T = m^n$. Note that for a determined data set in reality, the number of mappings corresponds to the amount of dominant data, that is, $T \approx n$. However, considering the preciseness of deduction, we still use $T = m^n$. Correspondingly, suppose that the classifier can be consisted of a series of atomic functions, and each atomic functions can classify or cover t mappings ($t > w$) in the domination data. Thus, we can obtain the definition of model capacity,

Definition 5 *(Model capacity) We assume the classifier can be divided into some atomic function, which can only cover a fixed number of mappings. We call the number of atomic functions $|F|$ required to classify all sub datasets as model capacity M, i.e.*

$$M = |F| \tag{3}$$

Since the sub datasets are sampled from the dominating data, we can find $N = C_n^w$ possible sub datasets. Accordingly, the atomic function corresponds to the subset of mappings, and therefore there are $M = C_T^t$ possible atomic functions. This problem can be transformed into a perfect hash mapping problem, that is, given N sub datasets, how many atomic functions of total M we need to select to achieve complete coverage of all sub datasets.

Referring to the proof of Stein-Lovasz theorem (Deng et al. 2011) in coding theory, we provide the upper bound of model capacity in the above problem.

Theorem 1 *(Upper bound) Let w be the sub dataset capacity sampled from dominating data and let T be the number of all possible mappings. The classifier is composed of atomic functions and each atomic function can cover t mappings with $t > w$. Then the model capacity $|F|$ required to cover all the sub datasets satisfies,*

$$|F| \leq \frac{C_T^t}{C_{T-w}^{t-w}}(1 + \ln C_t^w) \tag{4}$$

At this point, if the fitting ability of a atomic function is close to the size of sub dataset, i.e. $t > w$ and $t \approx w$, the above conclusion can degenerate into $|F| \leq C_T^w$.

All the proofs can be found in the in the supplementary materials.

Note that $\frac{C_T^t}{C_{T-w}^{t-w}} = \frac{T(T-1)...(T-w+1)}{t(t-1)...(t-w+1)}$, as shown clearly in Eqn. 4, the model capacity is related to the total number of mappings in dataset, which is then determined by the amount of dominating data. That is, the more sample complexity, the more atomic functions should be required, the more model capacity is required.

Besides, we use probabilistic approach to obtain a lower bound of model capacity required to coverage of all sub datasets. Due to the limited pages of main body, we put the lower bound and corresponding proof in the supplementary material. Similar views can be drawn from the lower bound of model capacity.

2.4 Sample Complexity and Dataset

Inspired by the concept of continuous mapping (Scarf 1967), we attempt to bridge the sample complexity and dataset. Moreover, we provide the lower bound of the number of small clusters in the dataset measured by data points distance. It can be seen from Fig. 1 that the points in a cluster are easy to learn through model generalization, and consequently the cluster can be served as the basic unit of sample complexity.

Suppose a CNN model, there exists a continuous mapping from the input space $X : x \in \mathbb{R}^{hwc}$ to the output space $Y : \Phi(x) \in \mathbb{R}^d$ with respect to this model. Data observed in reality are limited and can not fill the whole data space X, however, We assume that the learned CNN model can satisfy local continuous mapping in dataset, i.e., we hope a single point is surrounded by as many neighbor points as possible. Concretely, if our dataset possesses many large clusters, the classifier only need to memory some representative dominating points in these clusters, to obtain generalization for the other points.

Suppose that the number of data points in a dataset is N. We can apply a center threshold stated in Definition. 1 to obtain the neighboring points of x_i, we denote the number of neighboring points of x_i as $d(x_i)$. Then we can get the following theorem,

Theorem 2 *(Lower bound of sample complexity) Let $x_1, ..., x_n$ be the data points of dataset D and $d(x_i)$ be the number of neighboring points of x_i, then the number of clusters $\alpha(D)$ satisfies,*

$$\alpha(D) \geq \sum_{x_i} \frac{1}{1 + d(x_i)} \tag{5}$$

where the size of the maximal cluster satisfies,

$$\omega(D) \geq \sum_{x_i} \frac{1}{n - d(x_i)} \tag{6}$$

From Theorem. 2, it is obvious that the number of neighboring points of each point have negative correlation with the number of clusters, and the cluster number, corresponding to the dominating data, can be regarded as a measurement for sample complexity.

3 Towards Evaluating Data Sets Complexity

As mentioned above, we theoretically analysis the relationship between sample complexity and model capacity. In this section, we explore the impact of the specific data distribution on sample complexity. As discussed above, dominating data can be drawn from clusters, so we can define the unit distribution as a dominating point combined with its corresponding cluster. In order to achieve better generalization, we expect the unit distributions belonging to different classes have less overlap, and the unit distributions of the same class are more collective. From the perspective of representation learning (Bengio, Courville, and Vincent 2013), the inter-class discrimination and intra-class compactness is beneficial for model fitting and generalization.

Then, we will discuss how to evaluate dataset complexity from the perspective of inter-class and intra-class unit distributions.

3.1 Inter-class Distribution Overlap

Given a dataset on domain $\mathcal{X}$, it is infeasible to directly compute the distance of two unit distributions of different classes. Here, JS(Jensen–Shannon) divergence (Menéndez et al. 1997) is applied to depict the overlap of two distribution. Supposing P and Q are two unit distributions,

$$JS(P||Q) = \frac{1}{2}KL(P||\frac{P+Q}{2}) + \frac{1}{2}KL(Q||\frac{P+Q}{2}) \tag{7}$$

Following f-GAN (Nowozin, Cseke, and Tomioka 2016), we define f as the generator function $f : \mathbb{R}_+ \to \mathbb{R}$ of JS divergence, and f^* as its conjugate function, which is defined as $f^*(t) = \sup_{u \in dom\, f} \{ut - f(u)\}$. Then we get the following formulation of JS divergence,

$$JS(P||Q) \geq \sup_{T \in \mathcal{T}} \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))] \tag{8}$$

where T represents a arbitrary mapping from data point x to conjugate variable t and $\mathcal{T}$ represents qualified mapping groups. Proof details can be found in the in the supplementary materials.

We assume that T is an optimized model with learnable parameters θ following f-GAN (Nowozin, Cseke, and Tomioka 2016), and its optimal solution determines a classification for distribution P and Q. We rewrite T as a combination of feature extraction function V_w and activation function g_f, i.e., $T_\theta(x) = g_f(V_w(x))$. Let the activation function be,

$$g_f(v) = \log 2 - \log(1 + e^{-v}) \tag{9}$$

and $D(v) = 1/(1+e^{-v})$, we can obtain the following proposition,

Proposition 6 *Let $V_w(x)$ be a specific feature extraction function, the overlap between unit distributions P and Q measured by JS divergence can be represented as,*

$$D_{PQ} = -\mathbb{E}_{x\sim P}[\log(D(V_w(x)))] - \mathbb{E}_{x\sim Q}[\log(1 - D(V_w(x)))] \tag{10}$$

Note that we omit the constant term $\log 4$. If P and Q are balanced, the overlapping metric is equivalent to cross-entropy loss,

$$\begin{aligned} D_{PQ} &= -\mathbb{E}_x[y_x \log(D(V_w(x))) \\ &\qquad + (1 - y_x)\log(1 - D(V_w(x)))] \\ &= \mathbb{E}_x[\mathbb{CE}_{V_w(x)}(x)] \end{aligned} \tag{11}$$

where $y_x = \mathbb{1}(x \in P)$.

It is noteworthy that when we talk about the sample complexity of a dataset, we potentially default to that we are performing a specific task, such as image classification. Therefore, it is rational to discuss the sample complexity relative to a specific model.

Due to the symmetry of JS divergence, it can be verified that the above indicator also holds symmetry property,

$$\begin{aligned} D_{PQ} &= \mathbb{E}_{x\in P}[\log D_p(V_w(x))] + \mathbb{E}_{x\in Q}[\log(1 - D_p(V_w(x)))] \\ &= \mathbb{E}_{x\in P}[\log(1 - D_q(V_w(x)))] + \mathbb{E}_{x\in Q}[\log D_q(V_w(x))] \\ &= D_{QP} \end{aligned} \tag{12}$$

It is shown that the classification loss can be used to measure the overlap between unit distributions. Since we can easily compute this metric on all cluster pairs of different classes, we can thus obtain the index of sample complexity based on inter-class discrimination.

3.2 Intra-class Distribution Dispersity

The dispersity of intra-class distribution can be measured by the concept of information entropy, that is, larger entropy means the chaotic of the distribution, bigger sample capacity and more dominating data. However, due to the high dimension of images, given a specific group of data, it is difficult to directly compute data information entropy. It sounds reasonable to utilize signal-to-noise ratio, but suffers from rough depiction.

In the above section, a metric of overlap has been proposed to measure the distance between different data distributions. Intuitively, considering model generalization, a single peak and more concentrated distribution is superior to a scattered multi peak distribution. Therefore, we can randomly divide data belonging to one class into small groups, and then investigate the distribution overlap among these groups, as follows,

$$D_a = \sup_{K \in \mathcal{K}_a} \frac{2}{a(a-1)} \sum_{k_i, k_j \in K} D(p_{k_i}, p_{k_j}) \qquad (13)$$

where $\mathcal{K}_a$ represents all random divisions of a groups, and k_i, k_j denote two groups in a particular division K. D is a measurement of distribution overlap, as illustrated in Eqn.11.

For a real dataset, given group numbers a, we want to seek out a best division to maximize distribution difference, as an approximation of the above formula. Similar scheme is also applied in DRO (Kuhn et al. 2019; Sagawa et al. 2019), where they group the data set through target and background of images to encourage the model to pay attention to out of distribution images. Data grouping can be applied according to specific tasks. For example, clusters can be used as a division of data. In adversarial training (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017), the adversarial example and clean samples naturally form a division of the data set. More generally, all data can be sorted by model losses and we can make groups according to the order of the loss.

4 Experiments

In this section, we conduct extensive experiments to evaluate the sample complexity and model capacity of classification tasks on various datasets, which demonstrate the validity of the proposed analysis method. Our experiments are based on PyTorch.

4.1 Implementation Details

We conduct our experiments mainly on three datasets, including MNIST (LeCun et al. 1998), CIFAR-10 (Krizhevsky, Hinton et al. 2009) and SVHN (Netzer et al. 2011). Both the original samples and embedded features are utilized to evaluate sample complexity and required model capacity. For CNN embedding, we use ResNet-34 (He et al. 2016) model for CIFAR10 and SVHN datasets, and use LeNet5 (LeCun et al. 1998) model for MNIST dataset. The details of model training can be found in the supplementary materials. In addition, we also apply TSNE (Van der Maaten and Hinton 2008) function to project the original down to 2D space and similarly the CNN_{TSNE} function is applied to the CNN embedding.

Several data augmentation methods are introduced to verify the validity of our method. We compare the original data without applying data augmentation with Random crop, Random flip, Affine transformation, Rotation, Cut-Out, Noise. Due to the limited pages, we put the details and parameters settings about augmentations in the supplementary materials. As a special method of data augmentation, we will separately introduce the generation of adversarial examples in the following subsections. Considering the computational complexity, we will select partial classes in the data set or randomly select partial samples in each category for evaluation without affecting the experimental reliability.

4.2 Sample Complexity and Model Capacity

In this section, we evaluate the sample capacity of different augmentations on CIFAR10 as well as the model capacity to validate the relationship revealed in the theorem above.

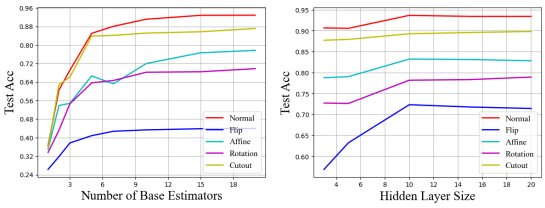
For a certain dataset, we utilized the number of clusters (Clusters), which can be interpreted as dominating data, to indicate the sample capacity as illustrated above. For each data point pair, we compute the L_2 distance and get a distance matrix. We then apply Definition. 1 to identify neighboring points (ϵ is empirically set as 12 for original images and 2 for CNN embedding) thus obtaining a graph, where vertexes represent data points and edges represent points neighboring. The dual relationship between the maximum independent set and the number of cliques in the graph is used to obtain the number of clusters, as defined in Definition.2. For more comprehensive validation, we also compute the size of the maximal cluster (MCS), which is defined as clique number in the graph theory. To evaluate the model capacity, several classifiers are utilized, including decision tree (Safavian and Landgrebe 1991), random forest (Biau and Scornet 2016) and 3 layer multi-layer perceptron (MLP). Hyper-parameters about classifiers can be found in the supplementary materials. For decision tree, the number of nodes and leaves are used to measure model capacity. For random forest and MLP, the number of base estimators and the hidden layer size are used to measure the model capacity.

The experimental results on CIFAR10 dataset in shown in Tab. 2. Average Clusters and MCS are collected from each classes. The decision tree fitting accuracy on train set of different augmentation is 100%. We only pay attention to the accuracy on train set since we are currently discussing the relationship between model fitting and data capacity. We can observed from the results that data augmentations can promote the dominating data, thus increasing the model capacity. For CNN embedding, we apply random forest and MLP classifier and evaluate the test set accuracy related with the number of base estimators and the hidden layer size, as shown in Fig.2. It clearly shows that augmentation datasets that possess more dominating data require more complex

classifier to obtain the comparable performance. Similar results can be observed from Fig.3, where the original images and CNN embedding are projected to 2D space.

Augmentation	Clusters	MCS	Nodes	Leaves
Normal	493.2	60.4	4993	2497
Crop	636.1	18.3	5251	2626
Flip	492.7	60.4	4995	2498
Affine	564.3	28.6	5063	2532
Rotation	579.6	28.2	5135	2575
Cutout	730.0	15.6	5303	2652

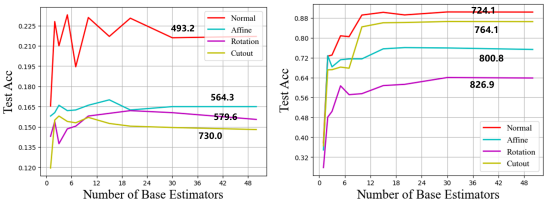
Table 2: Sample complexity measured by average Clusters and MCS and model capacity measured by decision tree nodes and leaves on CIFAR10 original images. The decision tree fitting accuracy of all augmentations is 100.0%. We evaluate totally 10000 samples where 1000 samples from each class.



(a) Random Forest (b) 3 Layer MLP

Aug.	Normal	Flip	Affine	Rotation	Cutout
Clusters	724.1	960.2	800.8	826.9	764.1

Figure 2: We report the test accuracy related with the model (Random Forest, 3 Layer MLP) capacity measured by number of base estimators and hidden layer size respectively for different augmentations on CIFAR10. Corresponding Clusters of different augmentation CNN embedding are shown in table below.



(a) TSNE (b) CNN_{TSNE}

Figure 3: We apply TSNE to the original images and CNN_{TSNE} to the CNN embedding. Corresponding clusters are marked in above figure.

To evaluate the data sample complexity based on representation learning, we conduct experiments of several augmentations on CIFAR10 to compute the metric of distribution overlapping and distribution dispersity. The distribution overlapping is calculated by Eqn.11 and the distribution dispersity is calculated by Eqn.13, where the dataset is grouped by clusters. Since the trained model has applied random flip and random crop augmentations, for fairly comparison, we only use other four augmentations for evaluation. As shown in Fig.4, the indicators used to measure sample complexity by distribution are aligned with the data capacity embodied by clusters. The experimental results shows the rationality of utilizing representation learning to depict the sample complexity.

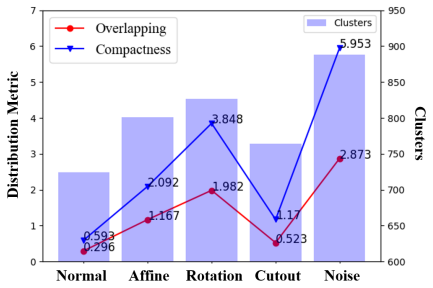


Figure 4: We show the distribution overlapping and distribution dispersity related with clusters of different augmentations.

4.3 Data Combination

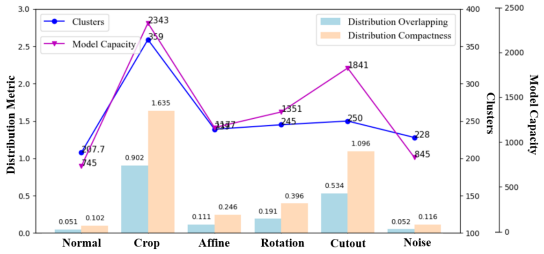
In this section, we conduct experiments on several self-constructed datasets extracted from CIFAR10 dataset to further verify large intra-class distribution dispersity can lead to more sample complexity. We construct four kind of datasets, 1) Merge 5: merge the adjacent categories like (0,1), (2,3), (4,5), (6,7), (8,9) to total 5 classes, 2) Keep 5: drop the class 5,6,7,8,9, and only keep the first 5 classes, 3) Merge 5 (Sim): merge the categories according to class similarity, like (airplane, ship), (automobile, truck), (bird, frog), (cat, dog), (deer, horse), to total 5 classes, 4) Keep 5 (Sim): keep one class from each pair above, like airplane, truck, bird, dog, deer. Details about class pairs in 3) can be found in supplementary materials. We make each processed class contain the same number of samples for fair comparison, and here we set the number of samples each class as 1000. Tab.3 shows our experimental results. Model capacity is evaluated on decision tree(Nodes, Leaves), random forest with 10 base estimators and MLP with hidden neural size 10. It is reported that the merged datasets, which damages the intra-class compactness, increase the model capacity compared with non-merged settings. The results holds even when we merge the similar classes. Note that since similar classes will also reduce sample complexity, the model capacity for Merge 5 (Sim) is a bit smaller than that of Merge 5, which further validate our assumption.

4.4 Results on MNIST and SVHN

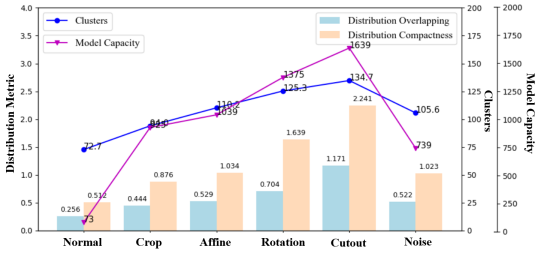
To further validate the proposed evaluation methods, we perform experiments on MNIST and SVHN datasets. CNN embedding is used to compute the point neighboring. We empirically set the distance threshold ϵ as 10 for MNIST and 2 for SVHN. Number of clusters, as an indicator for dominating data, is used to measure sample complexity. The

Combinations	Nodes	Leaves	RF_{10}	MLP_{10}
Merge 5	67	34	90.2	93.8
Keep 5	23	12	95.3	96.6
Merge 5 (Sim)	65	33	92.7	95.3
Keep 5 (Sim)	27	14	95.1	96.1

Table 3: Model capacity of different data combination methods. Nodes and leaves are used to indicate the model capacity of decision tree. RF_{10} represents the test set accuracy of random forest with 10 base estimators and MLP_{10} represents the test set accuracy of MLP with hidden neural size 10.



(a) MNIST



(b) SVHN

Figure 5: Experimental results on MNIST and SVHN datasets. We use the number of clusters, decision tree nodes to measure sample complexity and model capacity.

model capacity is measured by the number of nodes in decision tree. For distribution metric, we compute the distribution overlapping and as Eqn.11 and Eqn.13. We make a simplification here. We divide the samples of each class into two groups according to the order of classification loss, and approximate the intra-class by computing the distribution overlapping between the two groups of samples. All the experiments are conducted on test set. For sample complexity and model capacity evaluation, we randomly select 500 (MNIST) and 1000 (SVHN) samples from each class and compute the average results. For distribution metric, we use the whole test set. As displayed in Fig.5, the sample complexity, model capacity and distribution metric are almost aligned, which well demonstrate the effectiveness of the proposed evaluation methods.

4.5 Evaluation on Adversarial Examples

We conduct experiments on adversarial examples using CNN embedding. Project Gradient Descent(PGD) (Madry et al. 2017) are apply to attack clean images. As shown in Tab.4, adversarial examples can significantly raise the sample complexity, especially for distribution overlapping, thus increasing model capacity.

Data Process	Nodes	RF10	Overlap	Dispersity
Normal	99	92.15	0.296	0.592
PGD	1835	47.05	37.791	6.889

Table 4: Evaluation on adversarial examples. Decision tree and random forest are used to measure model capacity. We use distribution overlapping and distribution dispersity to show sample complexity.

5 Related Work

Classical framework to abstractly characterize the machine learning ability is Probably Approximately Correct (PAC) learning (Haussler 1990), which studies the conditions required to obtain a favorable model and the scale of corresponding training samples under a certain hypothesis space. For the infinite hypothesis space, VC (Vapnik-Chervonenkis) dimension (Blumer et al. 1989) is introduced to measure the complexity of the hypothesis space and the relationship between hypothesis space and data space upper bound is thus established. With the wide application of deep neural networks (He et al. 2016; Goodfellow et al. 2014; Girshick 2015) and the research on the sensitivity of model to adversarial examples (Szegedy et al. 2013; Carlini and Wagner 2017; Madry et al. 2017), a variety of works on sample complexity and model capacity have emerged. Schmidt et al. (Schmidt et al. 2018) theoretically show the data complexity to train a standard classification model and an adversarially robust model. Carmon et al. (Carmon et al. 2019) supplemented that the unlabeled data can be used to fill the gap between standard model and robust model through semisupervised learning. Moreover, some researchers study on assessing different datasets through a series of complexity measures (c-measures) to evaluate the difficulty of classification problems (Branchaud-Charron, Achkar, and Jodoin 2019). They assess class entanglement and assume that overlapping data distributions in dataset are more difficult to classify than well separated discriminative distributions. Related works include (Anwar, Jones, and Ganesh 2014; Baumgartner and Somorjai 2006; Duin and Pekalska 2006; Ho and Basu 2002; Sotoca, Mollineda, and Sánchez 2006).

6 Conclusion and Further Work

In this paper, we theoretically develop a new qualitative analysis between sample complexity and model capacity bridged by dominating data, and then propose how to compute sample complexity based on representation learning. Extensive experiments demonstrated the effectiveness of our evaluation method. The idea of analysing representative supporting data can be viewed as a new approach to explore the properties of datasets. Since the dominating data serve as the information about sample importance, in the further work, we can dynamically evaluate the importance of training data during model training, and correspondingly customize the data batch to improve efficiency.

Acknowledgments

This work was supported by National Science Foundation of China (U20B2072, 61976137).

References

Anwar, N.; Jones, G.; and Ganesh, S. 2014. Measurement of data complexity for classification problems with unbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(3): 194–211.

Baumgartner, R.; and Somorjai, R. L. 2006. Data complexity assessment in undersampled classification of high-dimensional biomedical data. *Pattern Recognition Letters*, 27(12): 1383–1389.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Biau, G.; and Scornet, E. 2016. A random forest guided tour. *Test*, 25(2): 197–227.

Blumer, A.; Ehrenfeucht, A.; Haussler, D.; and Warmuth, M. K. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4): 929–965.

Branchaud-Charron, F.; Achkar, A.; and Jodoin, P.-M. 2019. Spectral metric for dataset complexity assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3215–3224.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Carmon, Y.; Raghunathan, A.; Schmidt, L.; Liang, P.; and Duchi, J. C. 2019. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*.

Deng, D.; Li, P.; van Rees, G.; and Zhang, Y. 2011. The Stein-Lovasz theorem and its applications to some combinatorial arrays. *JCMCC-Journal of Combinatorial Mathematicsand Combinatorial Computing*, 77: 17.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Duin, R. P.; and Pekalska, E. 2006. Object representation, sample size, and data set complexity. In *Data complexity in pattern recognition*, 25–58. Springer.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Haussler, D. 1990. *Probably approximately correct learning*. University of California, Santa Cruz, Computer Research Laboratory.

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

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, T. K.; and Basu, M. 2002. Complexity measures of supervised classification problems. *IEEE transactions on pattern analysis and machine intelligence*, 24(3): 289–300.

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

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 5815–5826. PMLR.

Kuhn, D.; Esfahani, P. M.; Nguyen, V. A.; and Shafieezadeh-Abadeh, S. 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, 130–166. INFORMS.

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

Liu, J.; Song, L.; and Qin, Y. 2020. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 741–756. Springer.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Menéndez, M.; Pardo, J.; Pardo, L.; and Pardo, M. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2): 307–318.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Noble, W. S. 2006. What is a support vector machine? *Nature biotechnology*, 24(12): 1565–1567.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 271–279.

Orriols-Puig, A.; Macia, N.; and Ho, T. K. 2010. Documentation for the data complexity library in C++. *Universitat Ramon Llull, La Salle*, 196(1-40): 12.

Safavian, S. R.; and Landgrebe, D. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3): 660–674.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Scarf, H. 1967. The approximation of fixed points of a continuous mapping. *SIAM Journal on Applied Mathematics*, 15(5): 1328–1343.

Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smirnov, E.; Melnikov, A.; Oleinik, A.; Ivanova, E.; Kalinovskiy, I.; and Luckyanets, E. 2018. Hard example mining with auxiliary embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 37–46.

Sotoca, J. M.; Mollineda, R. A.; and Sánchez, J. S. 2006. A meta-learning framework for pattern classication by means of data complexity measures. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 10(29): 31–38.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

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

Yang, H.-M.; Zhang, X.-Y.; Yin, F.; and Liu, C.-L. 2018. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3474–3482.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 7472–7482. PMLR.

Zhang, X.; Zhu, Z.; Zhao, Y.; and Kong, D. 2018. Self-Supervised Deep Low-Rank Assignment Model for Prototype Selection. In *IJCAI*, 3141–3147.