Article in Press IJIMAI journal

Dataset and Baselines for IID and OOD Image Classification Considering Data Quality and Evolving Environments

Zhuo Zhang[1], Yang Li[1,2]*, Yicheng Gong[1], Yue Yang[1], Shukun Ma[1], Xiaolan Guo[1], Sezai Ercisli[3]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin (China)
[2] College of Mechanical and Electrical Engineering, Shihezi University, Shihezi (China)
[3] Department of Horticulture, Faculty of Agriculture, Ataturk University, Erzurum (Turkey)

Received 20 March 2022 | Accepted 3 October 2022 | Early Access 24 January 2023

Abstract

At present, artificial intelligence is in a period of rapid development, and deep learning has begun to be applied in various fields. Data, as a key part of the deep learning, its efficiency and stability, will directly affect the performance of the model, so it is valued by people. In order to make the dataset efficient, many active learning methods have been proposed, the dataset containing independent identically distribution (IID) samples is reduced with excellent performance; in order to make the dataset more stable, it should be solved that the model encounters out-of-distribution (OOD) samples to improve generalization performance. However, the current active learning method design and the method of adding OOD samples lack guidance, and people do not know what samples should be selected and which OOD samples will be added to better improve the generalization performance. In this paper, we propose a dataset containing a variety of elements called a dataset with Complete Sample Elements(CSE), the labels such as rotation angle and distance in addition to the common classification labels. These labels can help people analyze the distribution characteristics of each element of an efficient dataset, thereby inspiring new active learning methods; we also construct a corresponding OOD test set, which can not only detect the generalization performance of the model, but also helps explore metrics between OOD samples and existing dataset to guide the selected method of OOD samples, so that it can improve generalization efficiently. In this paper, we explore the distribution characteristics of efficient datasets in terms of angle element, and confirm that an efficient dataset tends to contain samples with different appearance. At the same time, experiments have proved the positive influence of the addition of OOD samples on the generalization performance of dataset.

Keywords

Active Learning, Data Quality, Efficient Dataset, Evolving Environments, Generalization.

DOI: 10.9781/ijimai.2023.01.007

I. Introduction

In recent years, with the development of artificial intelligence, deep learning is widely used in various fields, such as the detection and prevention of plant diseases and pests, the intelligent recognition of medical images and so on [1]–[6]. However, deep learning needs a large amount of data as the basis, which brings problems such as high data cost, difficult data acquisition and so on. In order to solve the demand problem of a large amount of data, people put forward few-shot learning, which is committed to learning from a small amount of labeled data and obtaining generalization ability [7]–[10]. Methods in the field of few-shot learning have solved the problem of large-scale data dependence to a certain extent. However, although selecting samples with high information quality for training can effectively help neural networks improve performance, little attention has been paid to the quality of sample information. A sample with high information quality not only has less noise, but also has a large difference with the existing samples in the data set. That is, conducting data quality assessments can improve the model performances under the same budget, and reduce the sample collection budget with the same model performances. Therefore, it is of great significance to carry out data quality assessment and establish relevant baselines in typical applications such as identification and classification.

In addition to considering the problem of data quality, the changeable test environment is a practical problem that can not be ignored, which widely exists in industrial processing and manufacturing, automatic driving, field environment and so on [11]. For this problem of poor generalization of the model caused by the change of test data, also known as the difference of out of distribution(OOD), it is necessary to establish a data set that fully considers the change of scene factors to provide a fair comparison platform for relevant research. Although there have been many studies on the generalization of model algorithms, there is still a lack of data set construction. In particular, considering the changes of environmental factors, the construction of high-quality data sets without watermark and error label is of great significance to the promotion of subsequent related research.

* Corresponding author.

E-mail address: liyang328@shzu.edu.cn

Please cite this article in press as:
Z. Zhang, Y. Li, Y. Gong, Y. Yang, S. Ma, X. Guo, S. Ercisli. Dataset and Baselines for IID and OOD Image Classification Considering Data Quality and Evolving Environments, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.01.007

In order to solve the above problems, this work has built an all element image acquisition platform and formed a Complete Sample Elements (CSE) data set, which can support the research and analysis of independent identically distribution (IID) and OOD. Probability entropy and distance entropy are proposed to evaluate the quality of the data set and establish relevant test baselines. In the aspect of OOD test, taking the real shooting data in the actual dynamic environment as the test, the relevant baseline is established, and the impact of distribution differences on the performance of the algorithm is explored.

The structure of this document is as follows: section II introduces the relevant work at home and abroad, section III introduces the CSE data set, section IV is the experimental part of this paper, and section V presents the conclusions and future works.

II. Related Works

Image quality evaluation is a basic and challenging problem in the field of image processing. Traditional image quality evaluation is realized by human visual system (HVS) or objective image quality assessment (IQA) [12]–[14]. It can evaluate the distorted images such as blur, JPEG compression and noise, and realize the discrimination of distortion types. However, these quality evaluation criteria serve human subjective visual perception and have nothing to do with the improvement of machine vision task performance. The development of artificial intelligence has promoted people to pay attention to data quality from the perspective of improving task performance. Recently, some work has also paid attention to the data quality of classification task guidance, such as the active cleaning of data labels proposed by Bernhardt in 2021 [15]. In addition, some works have paid attention to the influence of sample information quality on model performances, and they have proposed some sample information quality assessment methods on this basis [16], [17]. However, current methods lack validation on large-scale datasets containing constituent elements.

In the real scene, there are differences between train data and test data. How to effectively improve the test effect is a very valuable research direction. Under the guidance of this research direction, a variety of theoretical research methods on task generalization represented by transfer learning have been formed, so as to reduce the dependence on a large number of target domain data [18]. From the perspective of research subjects, transfer learning can be divided into data-based transfer learning, feature-based transfer learning, model parameter based transfer learning and so on. The data-based transfer learning method focuses on the transfer of knowledge through the adjustment and transformation of data; the feature-based method transforms each original feature into a new feature representation; model based transfer learning uses sub modules such as classifier, extractor or encoder to make accurate prediction results for the target domain, such as classification or clustering results [19]–[23]. However, these works focus more on theoretical research, and the data used are still quite different from the real scene.

III. CSE - A Dataset With Complete Sample Elements

In this section, we propose a dataset with complete sample elements, abbreviated as CSE. This dataset will facilitate the following two research topics:

- In addition to the common classification labels, the dataset also labels the remaining elements. When the train and test sets exhibit IID distributions, the element distribution characteristics of efficient datasets can be analyzed.
- When the train and test sets exhibit OOD distributions, the influence of OOD samples on generalization performance can be analyzed.

The dataset is divided into 11 categories, each class can be subdivided into 5 subclasses, so there is a total of 55 subclasses, each subclass is sampled from the same object. In Fig. 1, we show a representation of the images in it. Each subclass has 216 images, with 72 degrees and 3 distances. The background of the dataset is unified as a large checkerboard, and the size of the collected data is unified as 640×480. Since there are objects with small size in this dataset, in order to ensure that the collected sample subject is located in the center, we use a cylindrical heightening pad to support tiny objects. The dataset is publicly available for researchers to download and study[1].

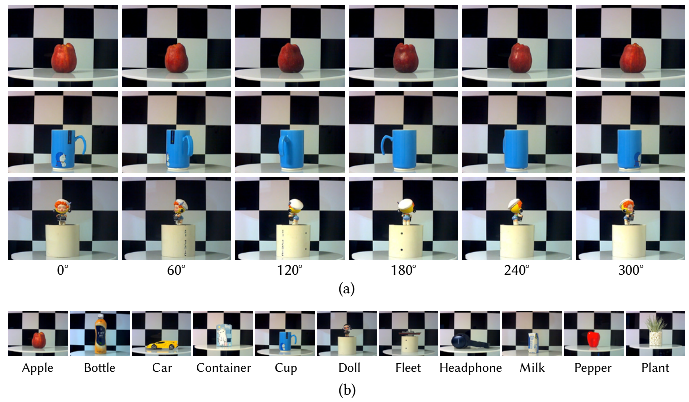


Fig. 1. Some samples of the CSE train set. (a) The examples of rotation; (b) All classes of the CSE dataset.

A. The Need to Structure Element-complete Datasets

1. For IID

For the train set and test set of IID distribution, there is redundancy within the train set. According to our test situation on the CIFAR-10 dataset, the sub-train set selected by the active learning method can obtain performance close to the entire dataset on fewer datasets, which will greatly improve the training efficiency. However, the current active learning methods has shortcomings such as relying on the selection of base classes, and the results are not robust. Therefore, if we can provide a dataset with complete elements, and understand why these samples will improve the performance of the dataset from the distribution level of elements, this will help us in the evaluation of sample information selecting. It should be noticed that in the class of doll and fleet, the objects are too small that it should be supported by a cylinder. To minimize the impact caused by the cylinder, we crop these samples, so it would be like Fig. 2.

Fig. 2. Some samples of the "doll" class in the CSE dataset. Objects in the "doll" class are tiny, so all samples of this class are cropped. The front, side, and back of this class of samples look very different, so if the network has only seen some of them (such as the front and the side), the rest of the samples (the back) are high-informative.

2. For OOD

For the train set and test set of OOD distribution, the information about the test set provided by the train set is insufficient, which will lead to a plummeting performance of the network model. If OOD

[1] Here: http://aimip.tju.edu.cn/rgzn.htm

samples are added to the train set, the network performance of the model will be improved; in future research, we will also try to calculate the distance between the train set and the test set and measure the similarity, and use the parts with high similarity to train the neural network, which will make the network more generalizable to the OOD test set.

B. How to Obtain A Dataset With Complete Elements

In order to make the sampling process automatic and controllable, we built a multi-DOF sampling platform, as shown in Fig. 3. The device is composed of three servo motors, which can realize front and rear, left and right, and up and down transforms, thereby changing the angle of the view captured by the camera.

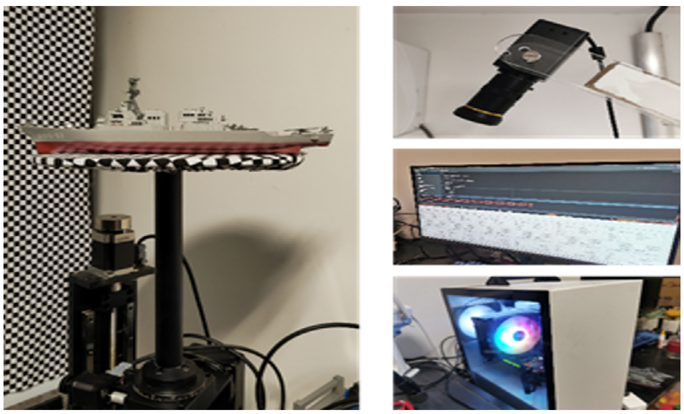


Fig. 3. The platform to sample the CSE dataset.

C. Variable Settings Supported by IID Train Set

After selection, we decided to design the following three variables to control:

1. Rotation Angle

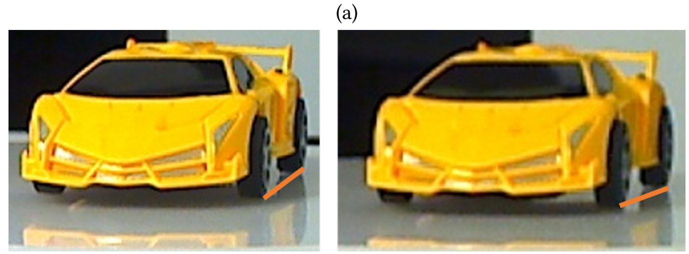
It is well known that for most asymmetric objects, the difference in viewing angle affects the observation results, and the same is true for neural networks. For example, as shown in Fig. 2, this figure shows the different angles of the front, side and back of the doll. For neural networks, especially those without any pre-training, they tend to think that these three angles are of different samples. The process of training is also the process of grouping samples of the same object from different angles into one group. The original intention of our design of the angle variable is how to choose an appropriate angle to reduce the number of samples in the train set as much as possible and improve the test accuracy as much as possible. We believe that this topic will be very helpful for future dataset simplification and dataset information Quantitative work.

When constructing the CSE dataset, we collect a sample every 5˝, and each sample can collect 72 images at the same camera distance. As for why 5˝ was chosen instead of 10˝ or 1˝, we have the following considerations. First, we believe that a full-featured dataset should be linear and smooth, so the angle of acquisition should be as small as possible, or as imperceptible as possible. However, collecting samples from an angle that is too small will greatly increase the number of samples, and many problems will follow: first, an excessively large number of samples will cause a serious burden on storage; second, an excessively large number of samples will prolong the training time; third, and most important point, repeatedly feeding a large number of samples with the same background and similar appearance to the model can easily make the model overfit, and even learn the background, object tray, and other elements incorrectly, which will cause a serious problem with the network model on which the object selecting method(distance entropy, probability entropy, etc.) relies. Therefore, considering the reasons above, we choose to collect a sample every 5˝.

2. Distance Between Object and Camera

The distance of the camera determines the proportion of the object in the sample. The farther the distance between the object and the camera is, the smaller the area of the object in the sample and the larger the area of the background. At the same time, the distance of the camera will also affect the viewing angle, as shown in Fig. 4. The farther the object is, the narrower the range that can be seen, which may bring an extra amount of information to the learning of the network model.

(a)



(b)

Fig. 4. The distance between object and camera also affects the sample. (a) Three distances when sampling a car; (b) Details of the farthest sample and the nearest sample under the same placement angle. As the text says, the connection line (orange) on the underside of the wheel varies with distance.

D. Design of OOD Test Set

Out-of-distribution test samples are encountered in some specific tasks. There are many reasons for the existence of OOD samples. For example, the initial design of the data set is not well thought out, or the data itself is difficult to collect in large quantities, and the train set can only be generated by simulation. How does the model make use of the OOD sample information it encounters? This is also the original intention of our design of the OOD test set.

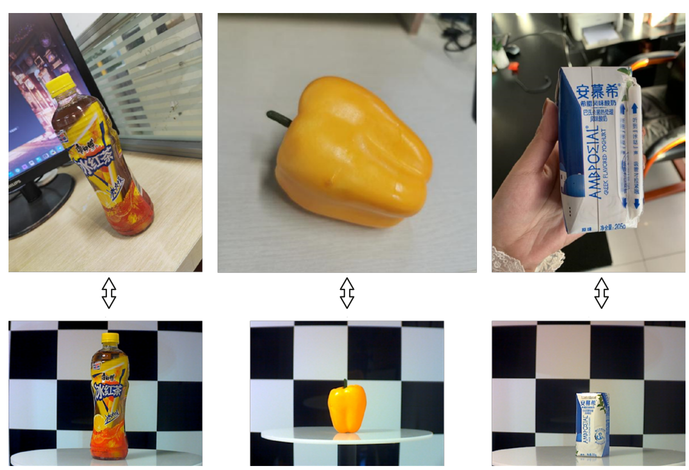


Fig. 5. The comparison between train samples and the corresponding OOD sample.

As shown in Fig. 5, when shooting the OOD test set, we randomly changed the background, randomly rotated the shooting angle, and randomly raised the shooting angle. These operations are very similar to the situation of OOD samples encountered in industry, since there are no such samples in the dataset. At the same time, the OOD samples are taken by mobile phones. Compared with the camera used to collect the train set, the default focal length of the mobile phone is shorter, and the captured samples will have some deformation compared with the train set, which is also a feature of the OOD samples.

IV. Results and Analysis

A. List of Backbones and Training Configuration

In order to verify whether the dataset we construct can effectively reflect the IID distribution and OOD distribution, we conducted several experiments with multiple backbone networks: ResNet [24], VGG [25] and WRN [26]. When we verify IID distribution, we use a uniform sampling of 25% of the train set as the IID test set. The purpose of this is to make the IID test set show a uniform distribution, and because the original train set is uniformly sampled, this sampling method will make this new IID test set present a similar distribution to the original train set, so the new IID train and test set can be approximated. When we verify OOD distribution, we directly use the test set and train set for OOD training and testing. All subsequent experiments are conducted under a single server with an Intel Core i7-12700KF CPU, dual nVidia GeForce RTX 3080Ti GPU and 128 GB memory with PyTorch.

B. Backbone Network Performances on IID and OOD Test Sets

The results obtained by training and testing on the backbone network are shown in Table 1. Note that the three backbone networks we listed all get 100% test accuracy on the IID combination, which proves that the IID train and test sets are IID distributions of each other, which is consistent with our assumption in IV.A. However, the three backbone networks perform poorly in the OOD combination. Given the undisputed high performance of the three backbone networks, it can also be shown that the OOD train and test sets are distributed in OOD.

In particular, we add a set of pre-trained comparison experiments in Table I. The control group has essentially the same parameters as the experiments using the three backbones, but with the ImageNet pre-trained model. It can be seen that a group of experiments using the ImageNet pre-training model is significantly higher in testing accuracy than the group that does not use ImageNet, which confirms the prediction in III.C.1. Using a train set with an interval of 5° has caused the model to overfit, as explained below. The result of VGG is a little bit lower, it is because the performance is weaker than ResNet and WRN.

TABLE I. The Test Accuracy of Several Backbones

settings	ResNet	VGG	WRN
IID, non-pre-training	100%	100%	100%
OOD, non-pre-training	18.808%	17.636%	20.268%
OOD, pre-training	29.256%	26.343%	36.696%

Experiment parameters: batch size: 32; Epoch: 50; initial learning rate: ResNet & WRN: 0.01; VGG:0.1; step learning rate: decay epoch: [20, 30, 40], gamma: 0.1

First, using ImageNet and reducing the learning rate is to make the model "remember" the ImageNet distribution as much as possible. Second, ImageNet is similar to our OOD test set collection method, and the background environment is more variable, which it also does in OOD test set. It can be considered that ImageNet has a similar distribution to the OOD test set. Third, since the pre-trained model achieves convergence in the later stage, it means that the model has also learned the distribution of the OOD train set, and so it does in non-pre-training group. In the comprehensive comparison experiment, the test performance of the ImageNet group is higher than that of the non-pre-training group. Therefore, we have reason to believe that the use of ImageNet pre-trained network can effectively suppress the overfitting phenomenon. The information of ImageNet makes the model not affected by factors such as background and thus overfit. While the non-pre-training group appears some overfitting phenomenon, which further deteriorates the performance on OOD test sets. When we use the Grad-CAM [27] method to visualize the attention of the network, we can see that the non-pre-training group totally cannot pay attention to the objects, but the pre-training group can accurately recognize them, as it can be seen in Fig. 6.

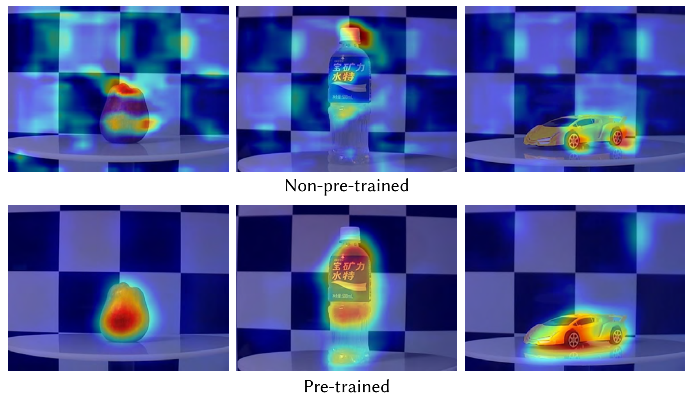


Fig. 6. When visualizing the model by Grad-CAM, we can see clearly that the non-pre-trained group has a stronger overfitting phenomenon than the pre-trained group.

C. Rotation Angle Distribution Features of Efficient Datasets

In order to explore the element distribution characteristics of efficient datasets derived from IID train set, we use two methods in active learning: distance entropy [28] and probability entropy [29]. We design a series of experiments: first, select 88(1%) IID samples as the base, add 88(1%) samples in each round of experiments, a total of 9 rounds. The subsets selected by these methods are re-trained on the ResNet18 network, and the IID test accuracy is shown in Fig. 7. It can be seen that the test accuracy of the subset obtained by selecting 528(6%) samples can already reach 99%. We take the subset with 528(6%) samples selected by distance entropy as an efficient train set for subsequent analysis.

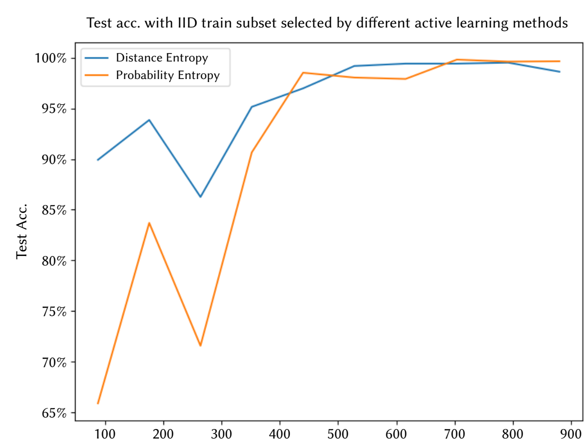


Fig. 7. The IID test accuracy of models trained on subsets selected by distance entropy method and probability entropy method.

1. Irregular Objects

When an irregular object rotates around to collect several samples, each sample has a large change from other samples, such as samples of cars, ships, dolls, etc. We take the sample from the third doll in the nearest position as an example to explore the rotation angle element distribution characteristics of irregular samples in an efficient dataset.

As shown in Fig. 8, the samples with high information content are distributed at 110°-205° and 270°-330°. At these degrees, the object is basically at the front or back angle, and the sample taken after rotation changes greatly, so it brings more information; while the side angle of the object is almost the same as the 205°, 270° samples, so less information. In the repeated experiments, we also did a set of similar experiments with cars, and the results were similar, as shown in Fig. 9, except that the side is high information, and the front and back are low information. This shows that the wider surface with larger rotation variation of irregular samples has high information content.

Fig. 8. The dolls selected by active learning method. In the CSE dataset, the front and back samples of this doll are of much more information, but the side samples are of less information.

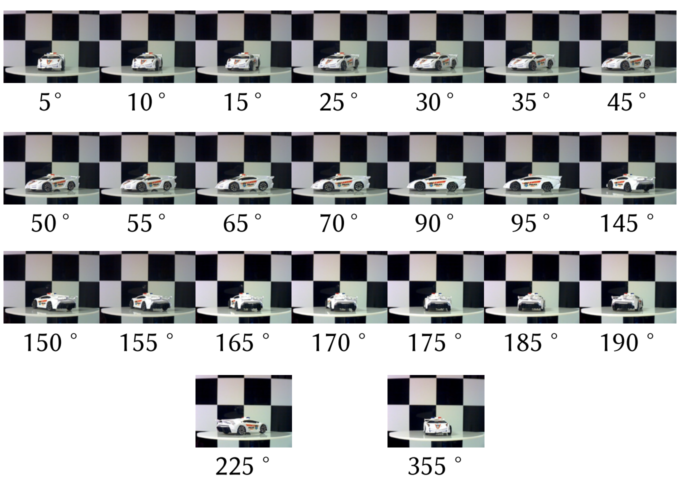


Fig. 9. The cars in the efficent dataset. Different from the dolls in Fig. 8, the side samples are of more information.

2. Rotation-Invariant Objects

Some objects are rotationally invariant, such as apples, containers, etc. Their characteristic is that samples taken from any angle are similar. Our experiments show that the number of rotation-invariant samples in the efficient dataset is much less than the number of irregular samples, such as the first apple corresponding to only seven samples (in contrast, each subclass of dolls generally selects at least 50 samples), and the angle has no regularity. Similar to our previous proof, rotation-invariant samples only need to find a few samples as representatives to obtain most of the information.

3. Approximately Rotation-Invariant Objects

There are also some objects that are approximately rotationally invariant in this dataset. Their main parts are rotationally invariant, but they also have other components that make them rotationally invariant, such as the handle of a cup. The characteristic of this type of object is that when the components that affect its rotation invariance are occluded, the samples have high similarity, as shown in Fig. 10. We take the sample of the second cup at the farthest position as an example to explore the rotation angle element distribution characteristics of approximately rotation-invariant samples in an efficient dataset.

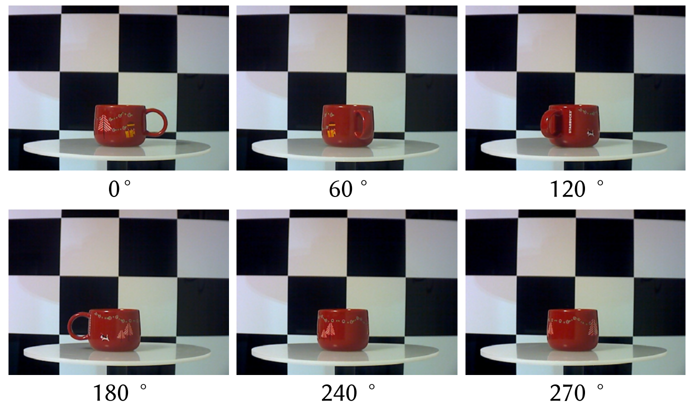


Fig. 10. Examples of approximate rotation-invariant objects. When the rotation angle comes from 240″ to 270″, the cup seems nearly the same, which means the low information in the samples.

As shown in Fig. 11, the samples with high information content are distributed between 45°- 70°, 185°-210°, and 305°-355°. Under these several degrees, the cup handle is on the side or front of the cup body, and the change is more obvious when rotating the object, and the samples of the cup handle behind the cup are relatively similar, so only a few samples can be selected. At the same time, when the handle is in front of the cup, since the color of the handle is closer to the cup, and the difference when it is rotated is smaller than that when the handle is on the side (only the 70° samples are sampled).

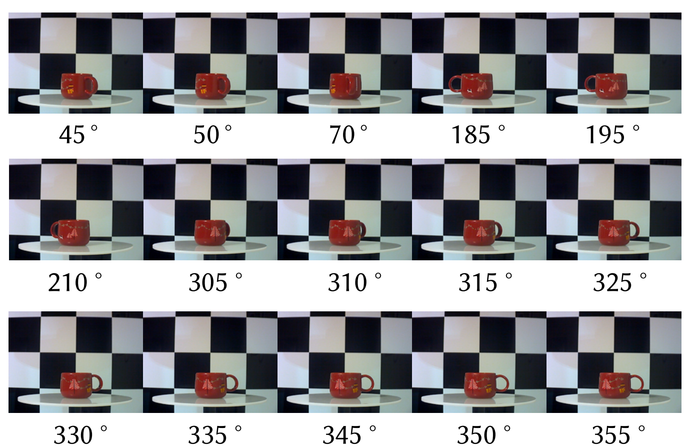


Fig. 11. The cups in the efficent dataset. when the handle is behind the cup, it will be unlikely selected.

At the same time, after our further statistics, we found that the irregular objects samples have the highest ratio in efficient datasets, the approximate rotation-invariant samples are in the middle, and the rotation-invariant samples are the lowest. The above analysis not only verifies the hypothesis that the IID train set has high redundancy, but also classifies the sample features with high information. It is found that the active learning methods will tend to select more differentiated samples to form a dataset, so that the selected samples have high information.

D. Effect of Adding OOD Samples on Generalization Performance of Models

In order to explore whether the addition of OOD samples will affect the generalization performance of the classification model, we designed the following experiments: 2% of the OOD train set samples of each class were selected to join in the OOD train set for training and the rest of the samples were used for testing, adding a total of 10 rounds, up to 20%. The experimental results are shown in Table II. As with our assumption, a small number of OOD samples can greatly improve the generalization performance. It is worth mentioning that after reaching a certain threshold, more OOD samples will not improve the accuracy. In contrast, using the added 20% samples for training and testing the remaining samples, the accuracy even slightly exceeds the results of the OOD train set + 20% OOD samples. This is because IID samples bring little information gain to the OOD test set classification problem, and may even drag back.

TABLE II. The Test Accuracy of Adding OOD Samples, Tested on ResNet18

train IID	Add OOD	Test
100%	0%	18.808%
100%	2%	46.684%
100%	4%	62.132%
100%	6%	61.507%
100%	8%	71.484%
100%	10%	74.069%
100%	12%	83.008%
100%	14%	83.112%
100%	16%	80.446%
100%	18%	82.188%
100%	20%	84.635%
0%	20%	84.659%

V. Conclusion and Future Works

In this paper, we construct a dataset called CSE, which has various element labels such as rotation angle, object category, distance, etc., which can be used to explore the distribution of each element of the high-informative train set samples, and how to add samples to the OOD test set, which can improve the generalization of the model. We believe that the CSE dataset we constructed can promote the development of active learning interpretability and active learning algorithm design. At the same time, we also believe that analyzing and designing active learning algorithms from the perspective of elements will be a direction of active learning development.

We use the backbone network to obtain the performance of the dataset under the IID and OOD test sets, confirming that the dataset we constructed exhibits IID and OOD distributions. We put forward the conclusion that the active learning model tends to select more differentiated samples. In the IID experiment, after using the active learning algorithm on the IID train set to extract the efficient train set, the rotation angle is divided into three basic types for statistical analysis, which confirms this assertion. We put forward the conclusion that adding OOD samples will greatly improve the generalization performance of the model, and verified this thesis by gradually adding OOD samples for training and testing in the OOD experiment. Among them, the experimental results of IV.B and IV.D also prove that OOD will bring low performance, and even if the model complexity increases, it will not bring noticeable performance improvement. So when creating datasets in various fields, researchers should pay strict attention to how well the training data fits the ground truth or testing data distribution.

However, we also noticed that batch addition of active learning algorithms brings some problems, such as the possibility of similarity between samples added in the same batch. In future work, we will expand the element information (such as pitch angle, background, etc.) of the CSE dataset to establish a more complete dataset for subsequent research.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under grant No.32101612 and No.61871283.

References

[1] A. Diaz-Pinto, N. Ravikumar, R. Attar, A. Suinesiaputra, Y. Zhao, E. Levelt, E. Dall'Armellina, M. Lorenzi, Q. Chen, T. D. Keenan, *et al.*, "Predicting myocardial infarction through retinal scans and minimal personal information," *Nature Machine Intelligence*, pp. 1–7, 2022.

[2] Y. Li, J. Yang, Z. Zhang, J. Wen, P. Kumar, "Healthcare data quality assessment for cybersecurity intelligence," *IEEE Transactions on Industrial Informatics*, 2022.

[3] V. Srivastava, S. Gupta, G. Chaudhary, A. Balodi, M. Khari, V. García-Díaz, "An enhanced texture- based feature extraction approach for classification of biomedical images of ct-scan of lungs," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 18-25, 2021.

[4] J. Nie, Y. Wang, Y. Li, X. Chao, "Sustainable computing in smart agriculture: survey and challenges," *Turkish Journal of Agriculture and Forestry*, vol. 46, no. 4, pp. 550– 566, 2022.

[5] S. Qiu, K. Cheng, T. Zhou, R. Tahir, L. Ting, "An EGG signal recognition algorithm during epileptic seizure based on distributed edge computing," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 5, pp. 6-13, 2022.

[6] S.-H. Chen, C.-W. Wang, I. Tai, K.-P. Weng, Y.-H. Chen, K.-S. Hsieh, *et al.*, "Modified YOLOv4-densenet algorithm for detection of ventricular septal defects in ultrasound images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol.6, no.7, pp. 101-108, 2021.

[7] Y. Li, J. Yang, "Few-shot cotton pest recognition and terminal realization," *Computers and Electronics in Agriculture*, vol. 169, p. 105240, 2020.

[8] J. Yang, X. Guo, Y. Li, F. Marinello, S. Ercisli, Z. Zhang, "A survey of few-shot learning in smart agriculture: developments, applications, and challenges," *Plant Methods*, vol. 18, no. 1, pp. 1–12, 2022.

[9] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[10] Y. Li, J. Yang, "Meta-learning baselines and database for few-shot classification in agriculture," *Computers and Electronics in Agriculture*, vol. 182, p. 106055, 2021.

[11] T. Isomura, T. Toyoizumi, "Dimensionality reduction to maximize prediction generalization capability," *Nature Machine Intelligence*, vol. 3, no. 5, pp. 434–446, 2021.

[12] J. Yang, Y. Zhao, J. Liu, B. Jiang, Q. Meng, W. Lu, X. Gao, "No reference quality assessment for screen content images using stacked autoencoders in pictorial and textual regions," *IEEE transactions on cybernetics*, 2020.

[13] K. Sim, J. Yang, W. Lu, X. Gao, "Mad-dls: mean and deviation of deep and local similarity for image quality assessment," *IEEE Transactions on Multimedia*, vol. 23, pp. 4037–4048, 2020.

[14] J. Yang, A. Li, S. Xiao, W. Lu, X. Gao, "Mtd-net: learning to detect deepfakes images by multi-scale texture difference," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4234–4245, 2021.

[15] M. Bernhardt, D. C. Castro, R. Tanno, A. Schwaighofer, K. C. Tezcan, M. Monteiro, S. Bannur, M. P. Lungren, A. Nori, B. Glocker, *et al.*, "Active label cleaning for improved dataset quality under resource constraints," *Nature Communications*, vol. 13, no. 1, pp. 1–11, 2022.

[16] Y. Li, X. Chao, "Toward sustainability: trade-off between data quality and quantity in crop pest recognition," *Frontiers in plant science*, vol. 12, 2021.

[17] Y. Li, X. Chao, S. Ercisli, "Disturbed-entropy: A simple data quality assessment approach," *ICT Express*, 2022.

[18] S. J. Pan, Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[19] J. Yang, J. Wen, B. Jiang, H. Wang, "Blockchain- based sharing and tamper-proof framework of big data networking," *IEEE Network*, vol. 34, no. 4, pp. 62–67, 2020.

[20] S. Kornblith, J. Shlens, Q. V. Le, "Do better imagenet models transfer better?," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2661–2671.

[21] G. Kang, L. Jiang, Y. Yang, A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902.

[22] Z. Hou, B. Yu, Y. Qiao, X. Peng, D. Tao, "Affordance transfer learning for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 495–504.

[23] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43– 76, 2020.

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

[25] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] S. Zagoruyko, N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[28] Y. Li, X. Chao, "Distance-entropy: An effective indicator for selecting informative data," *Frontiers in Plant Science*, vol. 12, 2022, doi: 10.3389/fpls.2021.818895.

[29] Y. Li, J. Yang, J. Wen, "Entropy-based redundancy analysis and information screening," *Digital Communications and Networks*, 2021, doi: https://doi.org/10.1016/j.dcan.2021.12.001.

Zhuo Zhang

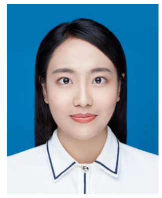
Zhuo Zhang received the M.S. degree in electronic information engineering from Tianjin University, Tianjin, China, in 2021. He is currently pursuing the Ph.D. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. His research interests include Data Information Assessment and Deep Learning.

Yang Li

Yang Li received the M.S. degree in Electrical Engineering from Dalian University of Technology, China, in 2016. He is currently a Lecturer at Shihezi University, China. His research interests include information processing, pattern recognition, few-shot learning, and data quality assessment. He is an Associate Editor of Plant Methods, Precision Agriculture, and Data Technologies and Applications.

Yicheng Gong

Yicheng Gong received B.S. degree in communication and information engineering from Tianjin University, Tianjin, China, in 2021. He is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. His research interests include Pattern Recognition and Deep Learning.



Yue Yang

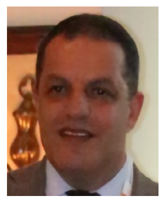
Yue Yang received B.S. degree in electronic information engineering from Tianjin University, Tianjin, China, in 2018. She is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. Her research interests include deep Learning and image information quality evaluation.

Shukun Ma

Shukun Ma received B.S. degree in communication engineering from Tianjin University, Tianjin, China, in 2021. He is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. His research interests include deep Learning and image information quality evaluation.

Xiaolan Guo

Xiaolan Guo received B.S. degree in communication engineering from Tianjin University, Tianjin, China, in 2020. She is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. Her research interests include deep Learning and image information quality evaluation.



Sezai Ercisli

Sezai Ercisli received his PhD degree at Ataturk University, Erzurum, Turkey in 1996. He is currently a Professor at the Department of Horticulture, Agricultural Faculty of Ataturk University in Turkey. His research interests include the modern technology in plant production, such as deep learning and machine learning. He is the Editor-in-Chief of Turkish Journal of Agriculture and Forestry.