ContRE: A Complementary Measure for Robustness Evaluation of Deep Networks via Contrastive Examples

Xuhong Li[1], Xuanyu Wu[1], Linghe Kong[2], Xiao Zhang[3], Siyu Huang[4], Dejing Dou[1], and Haoyi Xiong[1,✉]

1. Baidu Research, 2. Shanghai Jiaotong University, 3. Tsinghua University, 4. Clemson University

{lixuhong, wuxuanyu, doudejing, xionghaoyi}@baidu.com; linghe.kong@sjtu.edu.cn;
xzhang19@mails.tsinghua.edu.cn; siyuh@clemson.edu

Abstract—Training images with data transformations, e.g., crops, shifts, rotations and color distortions, have been suggested as *contrastive examples* to evaluate the robustness of deep neural networks against data noises [1]. In this work, we propose a practical framework ContRE (which is the meaning of "against" in French) that uses Contrastive examples for DNN Robustness Estimation. Specifically, ContRE follows the assumption in [2], [3] that robust DNN models with good generalization performance are capable of extracting a consistent set of features and making consistent predictions from the same image under varying data transformations. Incorporating with a set of randomized strategies for well-designed data transformations over the training set, ContRE adopts classification errors and Fisher ratios on the generated contrastive examples to assess and analyze the robustness of DNN models, which correlates to the models' generalization performance. To show the effectiveness and efficiency of ContRE, extensive experiments have been done using various DNN models, e.g., ResNet, VGGNet, DenseNet, EfficientNet, etc., on three open source benchmark datasets, i.e., CIFAR-10, CIFAR-100, and ImageNet, with thorough ablation studies and applicability analyses. Our experiment results confirm that (1) behaviors of deep models on contrastive examples are strongly correlated to what on the testing set, and (2) the robustness that ContRE calculates is a robust measure of generalization performance complementing to the testing set in various settings. Codes is to be publicly available.

Index Terms—Contrastive Examples, Robustness, Generalization Performance

Fig. 1: ContRE compares features and prediction results of original images from the training set with the results of contrastive examples.

I. INTRODUCTION

Deep Neural Networks (DNNs) significantly perform across various domains, including computer vision, natural language processing, and acoustics, but estimating their generalization performance robustly remains a challenge [4]–[7]. The robustness, denoting the capacity to evaluate model performance on unseen data, can be gauged by using samples from a disjoint testing set mirroring the same data distribution as the training set. Thereby, the testing error can approximate the generalization error, bypassing the intractable expectation of classification errors across data distributions. Recent efforts explore intrinsic properties of deep models, such as weight norms, gradient norms, and local minimum sharpness correlated with generalization performance. Nonetheless, employing these properties as metrics facilitates evaluation while preventing overfitting to the testing set in performance tuning [8], although their infirmity persists in contemporary deep learning tasks, concluding in poor generalization performance predictability [7].
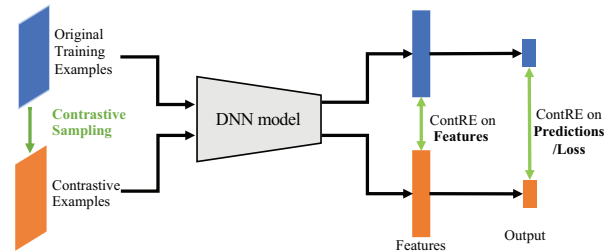
A recent method for evaluating generalization performance in deep learning models utilizes image transformations, termed *contrastive examples*, in training samples [1]. While theoretically supported [9], [10], this approach's effectiveness may wane if similar transformations are employed in the training data augmentation, due to potential overfitting by Deep Neural Networks (DNNs) [11]. Hence, there is a need for adaptable data transformations suited for various generalization performance measures and different training data augmentations.

In this work, we propose ContRE that uses Contrastive examples for the Robustness Evaluation of DNN models. Specifically, ContRE records features and prediction results of the original training set and what of the contrastive examples (see Figure 1 for an illustration). For this evaluation, we follow the assumptions in [2], [12] that a generalizable DNN model should make consistent and robust prediction for the same image of different views transformed by varying strategies, as the model would extract an invariant set of features from the image. In this way, a model with higher/lower consistency in prediction results and extracted features under varying data transformations would be with better/worse robustness. This robustness measurement is shown to be correlated with generalization ability of DNN models. Hereby, ContRE proposes to evaluate the generalization performance of DNN models using the classification accuracy (and Fisher ratio of feature vectors) on a set of contrastive examples. To generate effective contrastive examples, ContRE adopts RandAugment [13], to ensemble tens of commonly-used image transformations in a stochastic manner, where every original training sample would be transformed by a sequence of randomly picked-up operations.

In summary, this work makes contributions as follows:

- We propose a simple yet effective framework `ContRE` that generates contrastive examples using randomized data transformations on the training data and measures classification errors (and Fisher ratios) of DNN models on contrastive samples for the purpose of robustness and generalization performance evaluations.
- We conduct extensive experiments using various DNN models on three benchmark datasets, i.e., CIFAR-10, CIFAR-100 and ImageNet, empirically demonstrating the strong correlations between the accuracy on contrastive examples and the one on the testing set.
- Thorough analyses are provided: (1) By breaking down the data transformation strategies used in `ContRE`, we systematically analyze the effect of every single transformation technique and combination effects of all possible pairs; (2) We provide applicability analyses to confirm the versatility of `ContRE` working with arbitrary models trained with varying data augmentations; (3) We further investigate when and why the `ContRE` an effective way to estimate the generalization performance of DNN models.

To the best of our knowledge, this work is the first to practically assess the relevance of using contrastive examples for generalization performance of DNN models.

II. RELATED WORK

In this section, we introduce the related work from the methodology and application perspectives.

a) Contrastive Learning and Contrastive Examples: The idea of contrastive learning is to pull closer the representations of different views of the same image (positive pair) and yet repel representations of different images (negative pair). Networks are then updated iteratively with positive and negative pairs through various contrastive learning objectives [2], [3], [12], [14], [15]. In a contrastive learning framework, a stochastic data transformation module of generating different correlated views of the same image is proved to be effective to representation learning. Besides usages in contrastive representation learning, data augmentation has also been shown essential to improve the generalization of DNN models. The trivial ones, such as random crop, random flip etc, are widely used for training large DNN models. More recently, automated data augmentation policies including AutoAugment [16] have been proposed to select data augmentation methods for boosting the model performance. Rather than selecting data augmentation methods and compositions from a search space, RandAugment [13] proposes to randomly choose a number of augmentation distortions from a transformation candidate pool with a preset distortion magnitude.

In this work, we propose to measure the robustness of DNN models based on the (dis)similarity on features and final predictions between original images and contrastive examples, where we follow the RandAugment module to generate contrastive examples.

b) Generalization Performance and Robustness: Generalization performance measures the behaviors of models on unseen data. With the hold-out testing set as a proxy for measuring the generalization performance, the generalization gap is defined as the difference between the accuracies on the training set and the testing set. While previous works developed bounds for DNN models' generalization performance [4], [5], some recent works predict generalization gap from the statistics of training data or trained model weights. Jiang et al. [6] approximate the margin in neural networks and train a linear model on margin statistics to predict the generalization gap of deep models. Yak et al. [17] build upon the work by learning DNNs and RNNs in place of linear models. Instead of exploiting the margins, Unterthiner et al. [18] build predictors from model weights. Corneanu et al. [19] define DNNs on the topological space and estimate the gaps from the topology summaries. Note that `ContRE` could be loosely guaranteed through data-dependent theoretical bounds [9] and potentially related to the local elasticity [10], [20] while the theoretical adaptation to DNN models should be more cautiously considered. We leave this as future work.

Another way of analyzing generalization performance is through complexity measures since a lower complexity often implies a smaller generalization gap [7], [21]–[26]. The complexity is roughly defined as the measurement based on the DNN model's properties such as the norm of parameters and "sharpness" of the local minima. We refer interested readers to Jiang et al. [7], where authors did extensive experiments to test the correlation between 40 complexity measures and deep learning models' generalization gaps, suggesting that PAC-Beyasian bounds [21] and sharpness measure [22] perform the best overall.

Different from theoretical investigations and complexity measure of DNN models, `ContRE` measures the consistency of features and prediction results between contrastive examples and original images as generalization performance, which is data-dependent and practically works well.

III. CONTRASTIVE SAMPLING FRAMEWORK FOR DNN ROBUSTNESS EVALUATION

For leveraging large-scale unlabeled images, visual contrastive learning approaches [2], [3], [27]–[29] propose to train DNN models for solving the self-supervised contrastive prediction task that classifies whether each pair of images is similar or dissimilar [30]. Similar pairs are designed by a simple yet effective paradigm SimCLR [2], that transforms original images through various random data augmentation approaches [13], and then considers the pairs of augmented samples derived from the same images as similar pairs and others as dissimilar ones. We call these generated samples as *contrastive examples* and the process of generating contrastive examples as *contrastive sampling*. Contrastive sampling is crucial in the self-supervised training of a DNN model that yields general and discriminative representations.

In this paper, instead of aiming at representation learning, we propose to investigate the robustness and consistency

behaviors of (supervised) trained DNN models through our proposed approach ContRE, the Contrastive sampling framework for DNN Robustness and generalizability Evaluation. Figure 1 shows the process of ContRE. Specifically, ContRE first generates contrastive examples using RandAugment [13] and forwards generated samples into DNN models. Given the intermediate features and predictions of DNN models with contrastive examples as input, ContRE compares the results with original samples as input.

In the rest of this section, we formally introduce the proposed framework ContRE and the evaluation metrics for the investigations and analyses, as well as some discussions about ContRE.

A. Notations

For convenience, we consider a DNN model as composition of the classifier g and the feature extractor f^l, where l indicates the layer index and is omitted for the last layer before the classifier. From an underlying data-generating distribution $\mathcal{D}$, an empirical dataset for training $\boldsymbol{Z}$ and a disjoint testing set $\boldsymbol{Z}_{test}$ are drawn. Given a DNN model (with parameters $\boldsymbol{\theta}$) trained on $\boldsymbol{Z}$, we can simply denote all the forwarding results of this DNN model, including intermediate features $f^l(\boldsymbol{Z}; \boldsymbol{\theta})$, the estimates of the DNN model $g \circ f(\boldsymbol{Z}; \boldsymbol{\theta})$, and the loss/accuracy on the training set with an objective function $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z})$, as well as the generalization performance $\mathbb{E}_{z \sim \mathcal{D}} \mathcal{L}_{\boldsymbol{\theta}}(z)$. In practice, the generalization performance is measured on the disjoint set $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}_{test})$. Moreover, a set of DNN models with different parameters are considered: $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ...\}$.

B. Contrastive Examples

Given the notations, we introduce the proposed framework ContRE, which considers a number of widely-used data transformation approaches to generate a different view of the original dataset as the contrastive sampling process. We note the contrastive sampling strategy as R and the obtained contrastive examples as $\boldsymbol{Z}^s = R(\boldsymbol{Z}) \sim R(\mathcal{D})$. R can be any data transformation method. Besides the trivial ones, ContRE follows the strategy of RandAugment [13] for generating contrastive examples.

According to our preliminary experiments, any relevant image transformation method is effective to the generalization estimation if the transformed images have not been remembered by the model during training. In reverse, the unseen transformed images can be always added as data augmentation techniques for improving the performance. We could not easily identify the used transformation methods nor regulate the data augmentation if the access to the training process is not available. To this end, ContRE adopts the strategy of RandAugment to increase the difficulty of fitting all the possible transformed images by the DNN models.

Given the contrastive examples, the primary objective of ContRE is to exploit the contrastive sampling strategy to analyze and evaluate the robustness of DNN models. To achieve the goal, with contrastive examples or original images as input, ContRE thoroughly compares the behaviors of models on the prediction accuracy and intermediate features. Formally, given $\boldsymbol{Z}^s$, ContRE computes the intermediate features $f^l(\boldsymbol{Z}; \boldsymbol{\theta})$ and $f^l(\boldsymbol{Z}^s; \boldsymbol{\theta})$, prediction or loss $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z})$ and $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}^s)$, across different architectures and parameters $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ...\}$. Given these features and performances of various DNN models, ContRE measures their robustness and analyzes the consistency, with the metrics described below.

C. Metrics to Quantifying Robustness

Despite the strong non-linearity, DNN models exhibit relatively similar/consistent behaviors with similar samples as input (except the devised adversarial attacks [31]–[34]). *For example, a large crop or a vertical flip of an image rarely changes the DNN estimates; while a rotation or a severe color distortion slightly disturbs its predictions. Nevertheless, the level of robustness differs across models and this difference may be exploited to investigate the relation with the robustness and generalization performance.* Thus, to evaluate the robustness of DNN models given contrastive examples, we benefit from two correlation metrics and one feature clustering quality metric defined as follows

1) Spearman's Rank Correlation Coefficient: The proposed framework ContRE provides a good practical estimator of DNN models' generalization performance by measuring the loss/accuracy with contrastive examples as input. Specifically, given m DNN models, we measure the Spearman's rank correlation coefficient, noted as r_s between two rank variables:

$$r_s = \frac{\text{cov}(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}} \sigma_{\mathbf{Y}}}, \tag{1}$$

where $\text{cov}(\mathbf{X}, \mathbf{Y})$ is the covariance, $\sigma_{\mathbf{X}}$ and $\sigma_{\mathbf{Y}}$ are the standard deviations of the rank variables. Specifically, the rank variables are converted from $\{\mathcal{L}_{\boldsymbol{\theta}_i}(\boldsymbol{Z}_{test})\}_{i=1,2,...,m}$ and $\{\mathcal{L}_{\boldsymbol{\theta}_i}(\boldsymbol{Z}^s)\}_{i=1,2,...,m}$.

We show in the experimental section that this strong correlation exists in various scenarios, across datasets, hyperparameter settings, image transformation techniques, etc.

2) Partial Rank Correlation: Only using correlation as the measure for association between two variables can be misleading because their dependence may come from the associations of each with a third confounding variable. The problem can be solved by controlling the confounding variable via partial correlation. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be random variables taking real values, $r_{\mathbf{XY}}$ be Spearman correlation between $\mathbf{X}$ and $\mathbf{Y}$, then the partial rank correlation given control variable $\mathbf{Z}$ is:

$$r_{\mathbf{XY},\mathbf{Z}} = \frac{r_{\mathbf{XY}} - r_{\mathbf{XZ}} r_{\mathbf{YZ}}}{\sqrt{1 - r_{\mathbf{XZ}}^2} \sqrt{1 - r_{\mathbf{YZ}}^2}}. \tag{2}$$

We use the partial correlation to eliminate the effect of training accuracy $\{\mathcal{L}_{\boldsymbol{\theta}_i}(\boldsymbol{Z})\}_{i=1,2,...,m}$ while investigating ContRE's results. The details can be found in the experiment section.

3) Fisher Ratio: To verify the existence of significant correlations on feature level, ContRE also proposes to investigate the clustering quality of intermediate features. The degree of separation of features from other classes can be measured elegantly by the Fisher ratio. Given g different classes, there are

N_i data points in class π_i, with class mean $\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{Ni} x_{i,j}$. The between class scatter matrix measures mean separation and the within class matrix measures class concentration by pooling the estimates of the covariance matrices of each class, which are defined as:

$$\begin{cases} S_b = \sum_{i=1}^{g} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T, \\ S_w = \sum_{i=1}^{g} \sum_{j=1}^{N_i} N_i(x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T. \end{cases} \quad (3)$$

Then the Fisher ratio is:

$$Fisher Ratio = Trace(S_w^{-1} S_b). \quad (4)$$

Fisher ratio is a powerful tool for analyzing models' generalization capability right before the classifier. We compare features extracted from original samples and contrastive examples using Fisher ratio, i.e., $\{f^l(\boldsymbol{Z}; \boldsymbol{\theta}_i)\}_{i=1,2,\ldots,m}$ and $\{f^l(\boldsymbol{Z}^s; \boldsymbol{\theta}_i)\}_{i=1,2,\ldots,m}$, as part of explanations of why ContRE works well. The detailed findings are located in the experiment section.

IV. EXPERIMENTS AND ANALYSES

In this section, we validate, through extensive experiments, the main claim that $\mathcal{L}_{\boldsymbol{\theta}}(R(\boldsymbol{Z}))$ is strongly correlated to $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}_{test})$. We conduct experiments with various networks on three open source benchmark datasets, and all results with Spearman correlations and partial correlations support our findings. Furthermore, we demonstrate the applicability of ContRE across several practical settings, e.g., model selection across hyper-parameter choices, across RandAugment-Trained models, with small-size datasets or within a practical competition setting. Finally, we provide analyses for investigating the reasons that ContRE works well as an estimator of DNN models' robustness and generalization performance.

A. Experiment Setup

We describe the experiment setups of datasets, networks and the methods for generating contrastive examples.

1) Datasets: Three of the most popular benchmark datasets are considered: CIFAR-10, CIFAR-100 and ImageNet. CIFAR datasets contain 60,000 tiny images of resolution 32×32 while ImageNet includes over 1.2M natural images. Conventional split of training and test sets are followed for CIFAR datasets, while the validation set of ImageNet is used as testing set here.

2) Networks: For evaluating our proposed framework ContRE, the most popular or the most powerful DNN architectures are considered: VGGNet [35], ResNet [36], Wide ResNet [37], DenseNet [38], ResNeXt [39], ShuffleNet [40], [41], MNasNet [42], MobileNet [43], [44] and their variants. We train these DNN models with standard data augmentation methods, i.e., random crop and random flip, and standard training hyper-parameters. We also investigate the different data augmentation methods and different hyper-parameter settings for training DNN models, and the strong correlation between $\mathcal{L}_{\boldsymbol{\theta}}(R(\boldsymbol{Z}))$ and $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}_{test})$ still holds. Generally, models for CIFAR datasets are smaller than those for ImageNet but the architectures are shared among models for these datasets.

TABLE I: Correlations between the test accuracy and contrastive examples based accuracy on the training set, evaluated on three datasets CIFAR-10, CIFAR-100 and ImageNet with 14, 14 and 27 DNN models respectively.

	CIFAR-10	CIFAR-100	ImageNet
Center Crop	0.702	0.677	0.953
Random Crop	0.722	0.789	**0.979**
Random Flip	0.702	0.790	0.955
Random Rotation	0.938	0.653	0.910
ContRE (ours)	**0.965**	**0.974**	0.970

Moreover, many pretrained models on ImageNet are publicly available[1], which greatly accelerate our setup process.

3) Contrastive Examples: As introduced previously, RandAugment [13] is used in the ContRE framework to generate contrastive examples, where we take all the available transformation operations from the open-sourced implementation[2]. We note RA_NX_MY, where X and Y are the two hyperparameters to configure in RandAugment: the number of sequential distortion operations and the magnitude of each distortion. For example, RA_N2_M9 means that 2 operations will be chosen randomly from all the available operations and the distortion of magnitude 9 will be applied[3].

B. Experiment Results

Strong correlations between the generalization performance $\{\mathcal{L}_{\boldsymbol{\theta}_i}(\boldsymbol{Z}_{test})\}_{i=1,2,\ldots,m}$ and the accuracy on contrastive examples $\{\mathcal{L}_{\boldsymbol{\theta}_i}(\boldsymbol{Z}^s)\}_{i=1,2,\ldots,m}$, are found across the three evaluated datasets, CIFAR-10, CIFAR-100 and ImageNet. Here we present the detailed experimental results for these three datasets in comparison with trivial image transformations, such as center crop, random crop, random flip and random rotation.

As shown in Figure 2 (top and middle), on the CIFAR datasets, most of the models have achieved close to 100% accuracy on the training set for training samples transformed by crop or flip, which have been used during the training process. For these transformations, the correlations are around 0.7. Random rotation is not used during the training process, while the samples transformed by random rotation do not achieve stable correlations: a high correlation coefficient is obtained for CIFAR-10 while a relatively low value for CIFAR-100. Meanwhile, our proposed approach ContRE consistently outperforms others by reaching the highest correlation coeeficients for both CIFAR datsets.

The results are different for models on ImageNet. Due to the tremendous amount of samples and classes, ImageNet models with the best validation accuracies failed to perfectly fit the training data. As shown in Figure 2 (bottom), training accuracies based on the listed operations all correlate well with test accuracies. Results given by random crop attains the first place with a correlation of 0.979, followed closely by ContRE's 0.970.

[1]https://pytorch.org/vision/stable/models.html
[2]https://github.com/ildoonet/pytorch-randaugment
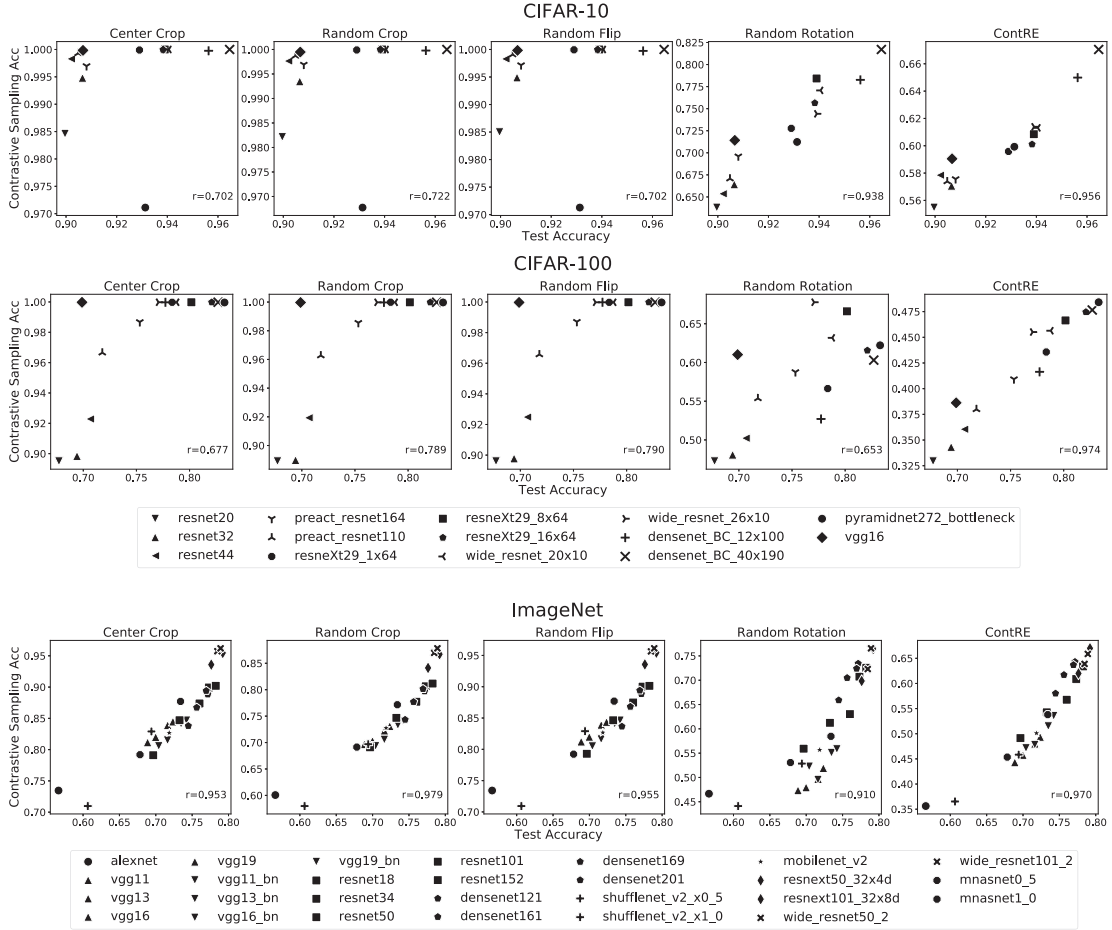[3]The maximum magnitude is 30, even which does not transform the image into a total noise.

Fig. 2: Spearman correlations between test accuracies and contrastive examples based accuracies on training data, using four trivial data augmentation techniques and our proposed approach `ContRE`, evaluated on CIFAR-10, CIFAR-100 and ImageNet. For DNN models on ImageNet, we use a single mark for the same family of networks due to the clarity and notation reasons.

The experimental results show that `ContRE` gets high correlation coefficients $Corr$ (Equation 1) of 0.969 on CIFAR-10 and CIFAR-100 and 0.970 on ImageNet, indicating that the grades of the models under the evaluation of `ContRE` are significantly correlated to the generalization performances of these models. Note that this observation is hard to get from trivial approaches, including random crop, rotation, flip etc., because of randomness and instability. Considering these trivial approaches, "Random Rotation" works well on CIFAR-10 but yields a relatively low correlation coefficient on ImageNet. Similarly, "Random Crop" works well on ImageNet but not on CIFAR-10 or CIFAR-100. However, across three tasks and datasets, our approach `ContRE` perfectly grades the models, probably thanks to the stability from the expectation over the randomness. We thus reasonably conclude that `ContRE` might have the potential to estimate the generalization performance of models with the use of training data and the contrastive techniques.

TABLE II: Partial correlations between test accuracy and contrastive examples based accuracy on the training set while controlling training accuracy.

	CIFAR-10	CIFAR-100	ImageNet
Random Crop	0.241	0.551	<u>0.766</u>
Random Flip	N/A[4]	<u>0.619</u>	0.228
Random Rotation	<u>0.887</u>	0.190	0.566
`ContRE` (ours)	**0.938**	**0.962**	**0.774**

C. Analyses

We provide investigations from the partial correlations controlling the confounding variable and the feature-level correlations, to dissect the reasons that `ContRE` works for estimations of DNN models' robustness and generalization performance.

[4]The correlation between accuracy of training examples after random flip and original training accuracy is 1, leading to division by 0 in Equation (2). Similar for "Center Crop" on the training set that is used as training accuracy.

TABLE III: Correlations between test accuracy and Fisher ratios of contrastive examples on the training set, evaluated on three datasets with 14, 14, 27 DNN models respectively.

	CIFAR-10	CIFAR-100	ImageNet
Center Crop	0.495	0.754	0.662
Random Crop	0.455	0.780	0.686
Random Flip	0.495	0.754	0.662
Random Rotation	0.736	0.640	0.774
ContRE (ours)	**0.767**	**0.890**	**0.816**

1) Partial Correlations: We observe that the training accuracy is well correlated to the testing accuracy on ImageNet, where the best model cannot fit the training set. This shows that the training accuracy might be a good estimator of the generalization performance in some cases. On the other hand, training accuracy is also related to the accuracy on contrastive examples because they are directly generated from the training set. A question naturally follows: *is the training accuracy the key factor for ContRE being a good estimator?*

To answer this question, we first define the notion of *consistency* $C_{\boldsymbol{\theta}}(\boldsymbol{Z}^s)$ as the difference between training accuracy and contrastive examples based accuracy on the training set $C_{\boldsymbol{\theta}}(\boldsymbol{Z}^s) = \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}) - \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}^s)$. The testing accuracy and the accuracy on contrastive examples contain a common variable, i.e., training accuracy $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z})$, shown in the decomposition below:

$$\begin{cases} \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}^s) & = \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}) - C_{\boldsymbol{\theta}}(\boldsymbol{Z}^s), \\ \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}_{test}) & = \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}) - gap. \end{cases} \quad (5)$$

With controlling the effect of this compound variable, training accuracy $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z})$, we compute the partial correlation between the accuracy on contrastive examples $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}^s)$ and the testing accuracy $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{Z}_{test})$. Directly estimating the gap with the consistency fails on ImageNet for the reason that models fit to ImageNet training set in varying degrees. This is not a practical issue for two reasons: (1) it will be less needed to estimate the gap if the generalization performance is strongly correlated to the training accuracy; (2) few datasets are in the scale of ImageNet. Nevertheless, the results in Table II show ContRE still gets relatively strong (partial) correlations on all the three datasets, while a moderate difference is observed from the comparison between CIFAR datasets and ImageNet. We note that in the case of ImageNet, where the best model cannot perfectly fit the training set, the training accuracy is directly correlated to the testing accuracy and plays an important role for generalization performance estimation. However, for CIFAR datasets, where all (or some) models get 100% training accuracy, the training accuracy loses the predictability and the partial correlations controlling the training accuracy have no large difference compared to the standard correlations. In summary, the training accuracy is not the key factor that ContRE works well for the generalization performance estimation in most cases.

2) Feature-Level Correlations: For complex DNN models, we further conduct analysis experiments by looking into the intermediate features of DNN models, computed on the origi-

nal and contrastive examples. For these examples, we compute the Fisher ratios of features $f(\boldsymbol{Z}; \boldsymbol{\theta})$ before the classifier for all models, and get the Spearman correlations with test accuracies. Note that for ImageNet and CIFAR-100 datasets, the number of data instances in each class is less than the number of dimensions of feature vectors when the DNN model, for instance DenseNet, is wide, leading to the non-invertibility of the within class matrix during the computation of Fisher ratio. We therefore perform the dimension reduction techniques on the feature vectors using singular value decomposition (SVD), reducing the number of dimensions to 64 or 512 for CIFAR-100 or ImageNet respectively, with most models containing over 95% of the total variance.

The results in Figure 6 (a) show that Fisher ratios from ContRE are highly correlated with the generalization performance of DNN models on each of the three datasets. We also conduct experiments to compute the results on models with the same architecture but different hyper-parameters, shown in Figure 6 (b). The conclusion also holds for models trained with different hyper-parameters. The overall numerical comparison is shown in Table III. It suggests that, among contrastive examples based on various operations, models that are able to generate better class-separated features from ContRE samples are more likely to have better generalization performance. Intuitively, Fisher ratio, which can be seen as a measure of the clustering quality, is highly correlated with the final classification accuracies. Therefore, the high correlation between ContRE accuracies and generalization performance inherently comes from the association between ContRE accuracies and its features' degree of separation, and that between Fisher ratio of features and model generalization performance.

D. Ablation Studies

For further showing the stability and effectiveness of ContRE, we carry out the experiments for ablation studies on the choices of two hyper-parameters in RandAugment when generating contrastive examples, and on the comparisons between contrastive examples from RandAugment and from single data transformations or the compositions of two transformations.

1) Choices of N and M: In our framework ContRE, two hyper-parameters in RandAugment can be tuned: N denotes the number of transformations that are to be applied on the original examples, and M denotes the magnitude of distortions. Though they both represent the amount of distortions, their effects on ContRE are varied. As shown in IV, we have tested all the compositions of $N = \{2, 3, 4\}$ and $M = \{4, 9, 15, 20\}$. A large ($M = 15$) magnitude of each distortion leads to a higher correlation but an additional increase from 15 to 20 does not change much. The results match our previous observations, implying that a weak distortion on the samples cannot help to tell good classifiers from the bad ones. On the other hand, applying more transformations at the same time seems to hurt the performance, especially when M is small. Therefore, our approach adapts a large value of $M = 20$ and

TABLE IV: Spearman's correlations between test accuracy and contrastive examples based accuracy on the training set, with different choices of N and M.

	$N=2$	$N=3$	$N=4$
$M=4$	0.837	0.763	0.503
$M=9$	0.855	0.833	0.714
$M=15$	0.974	0.974	0.960
$M=20$	0.965	0.978	0.938
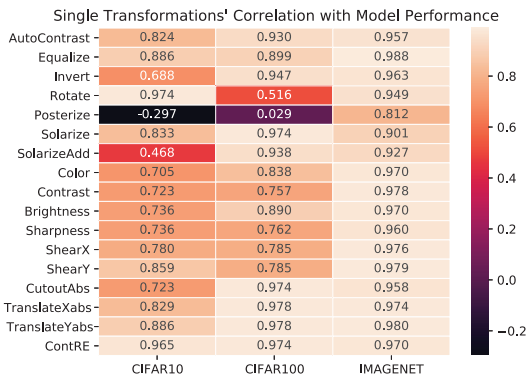
Fig. 3: Spearman correlations between test accuracy and the accuracy on contrastive examples generated by single transformations and `ContRE`.

a smaller value of $N = 2$, which is proved to be quite robust across various tasks and datasets.

2) Single Transformations: `ContRE` generalizes the idea of contrastive learning to the evaluations of DNN models' robustness and generalization performance. However, is it really necessary to use a stochastic module for generating the contrastive examples for `ContRE`? Is there any single transformation that could achieve the desired results? We answer these questions through the comparison of the effects by `ContRE` and each single contrastive technique.

Following the same experimental setup, contrastive examples are generated under each single transformation with the same magnitude as `ContRE` and the experiments are performed using all DNN models considered previously. The obtained Spearman correlation coefficients range from -0.297 to 0.988. While a part of single transformations yield correlations higher than 0.9, especially for ImageNet-trained models, there is no single operation that performs well across the three datasets. Therefore, choosing a best transformation depends on the dataset and it is difficult to find a single transformation that works well across different datasets. In contrast, `ContRE` introduces robustness and consistently achieves high performance through the expectation over the stochastic module and by combining several competent candidates.

3) Compositions of Two Transformations: To further prove the effectiveness and efficiency of `ContRE`, we conduct similar experiments on CIFAR datasets using contrastive examples that are generated by all the possible compositions of two transformations, obtain the accuracy on various contrastive

Fig. 4: Spearman correlations between test accuracy and the accuracy on contrastive examples generated by two transformations on CIFAR-10 (upper) and CIFAR-100 (lower).

examples and show the Spearman correlations with the testing accuracy in Figure 4. Some observations are shared with the results from the comparison using single transformations, that any composition with "Posterize" does not work well for CIFAR datasets and that any composition with "Rotate" works well for CIFAR-10 but not for CIFAR-100. A small part of compositions obtain higher correlations than `ContRE`. We argue that searching the optimal composition demands much computation resource and the optimal choice varies from different datasets, while `ContRE` benefits the expectation over the stochastic module to consistently produce precise estimation results. We will present in the following that `ContRE` is consistently applicable, efficient and effective in many practical complex situations.

E. Performance on Partial Training Data

In some practical settings, training data are very scarce while the experiments we have shown are using a large dataset

with over 100 samples per class. So we here would like to test the robustness of `ContRE` by reducing the training data size to small fraction (20%) and training the DNN models. When provided with 20% training data, models have limited information of the true data distributions and the generalization performances drop by 3-10 percent. Under such setting, `ContRE` outperforms trivial approaches by achieving a correlation coefficient of 0.754. The results indicate the potential of `ContRE` in predicting model performances when models are not trained with abundant data.

F. Comparisons with RandAugment-Trained Models

Given `ContRE`'s high correlations between the generalization performance and the accuracy on contrastive examples of DNN models that are trained with standard data augmentation methods, we have observed that the "Random Crop" and "Random Flip" that are used for the standard training process do not provide desired results in our framework. For over-parametric DNN models, including more data augmentation methods could usually improve the generalization performance. We thus ask if the correlation obtained by `ContRE` mainly benefits from the image transformation methods that are not used during training and whether `ContRE` still performs well when the generated contrastive examples have already been used during training.

To answer the questions, we have trained the same set of DNN models on CIFAR-10 with replacing the standard data augmentation by the RandAugment module while leaving other training settings unchanged. Then, we repeat the processing of `ContRE` and get that `ContRE` still achieves a correlation higher than 0.9, with the results in Figure 5 (right) and comparisons in Table V. We note that "Random Flip" and "Random Rotation" get higher correlations than `ContRE`. We argue that single transformations could get high correlations if they are not used during training, but the correlation coefficients largely decrease if they have been used during training, as shown in the previous results. In contrast, `ContRE` combines a number of single transformations to generate contrastive examples and introduces the robustness from the expectation over the stochastic module. Even with the models that have seen the contrastive examples during the training stage, `ContRE` can also successfully estimate the generalization performances.

V. Evaluations for Applicability Analyses

We have validated our proposed method `ContRE` on three popular datasets with standard training processes. In this subsection, we carry out experiments with practical settings and demonstrate the applicability of `ContRE`.

A. Applications to Hyper-parameter Selection

A direct application of `ContRE` is to find the best model among models with the same architecture but trained with different hyper-parameters. We have trained 18 models, all of them is of the network structure of Wide-ResNet-26-10, on CIFAR-10 with various choices of batch size (32, 64,

TABLE V: Spearman correlations between test accuracy and the accuracy on the contrastive examples, for models trained with different hyperparameter choices, trained with 20% training data, and trained with RandAugment.

	Hyperparameter	Subset	RandAugment
Center Crop	0.268	0.465	0.723
Random Crop	0.221	0.535	0.738
Random Flip	0.268	0.583	**0.978**
Random Rotation	0682	0.055	0.960
`ContRE` (ours)	**0.861**	0.754	0.916

128), learning rate (0.01, 0.05, 0.1) and weight decay (0.0001, 0.001). Based on the resulting 18 models, `ContRE` outperforms other approaches by achieving a correlation coefficient Corr of 0.859, see the results in Figure 5 (left) and the comparison with other approaches in Table V. The results illustrate that, besides helping with selecting the best architecture, our approach can be helpful in the hyper-parameter-tuning stage.

B. Applications to Generalization Gap Prediction [1]

`ContRE` directly estimates the generalization performance instead of the generalization gap (the difference between the training accuracy and the generalization performance). In the previous experiments, we mainly discussed the estimations of generalization performance in this work. Meanwhile, under the general condition that all DNN models equally fit on the training set [7], `ContRE` is also capable of providing a good estimator of the gap.

To further demonstrate the feasibility of `ContRE`, we use the experiment settings of "Predicting Generalization in Deep Learning Competition" at NeurIPS 2020 [1] to evaluate `ContRE` on predicting the generalization gap of DNN models using the contrastive examples of the training set. The competition offers a large number of deep models trained with various hyper-parameters and DNN architectures trained on CIFAR-10 or SVHN, while the evaluator of the competition first estimates the generalization performance of every model using the proposed measure, then computes the mutual information score (the higher the better) between the proposed measures and the (observable) ground truth of generalization gaps.

In the experiments, we propose to use the difference between the accuracy based on the original training samples and the one using *contrastive examples* generated from the original training samples with `ContRE`. For the comparison reason, we include several baseline measures in generalization gap predictors, including VC Dimension [45], Jacobian norm w.r.t intermediate layers [46], distance from the convergence point to initialization [24], and the sharpness of convergence point [47], all of these baselines provide a theoretical bound for the estimated gap. Table VI presents the comparisons between the proposed measures and baselines. It shows that the proposed `ContRE` framework, with estimating the gap between the accuracy on original training samples and the one on contrastive examples, significantly outperforms the baseline methods. Note that our proposed approach `ContRE` is an
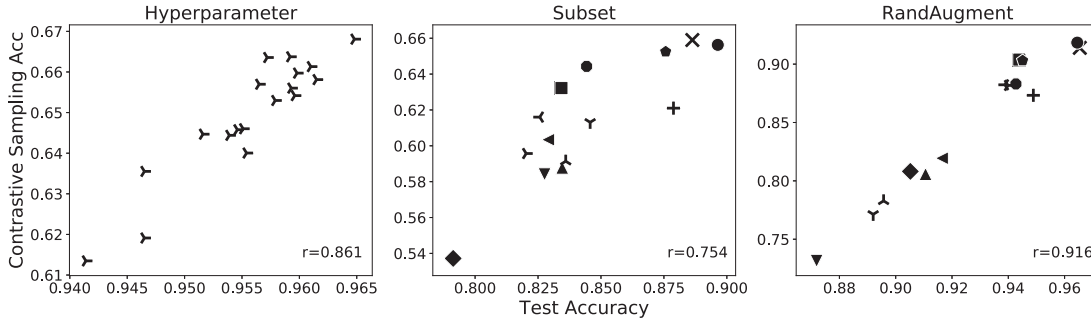
Fig. 5: Spearman correlations between test accuracy and contrastive sampling accuracy on the training set for Wide-ResNet-26-10 with different hyperparameter choices (left); for all the models trained with 20% training data of CIFAR10 (middle); for all the models trained with RandAugment on CIFAR10 (right). Symbols for DNN models are the same as in Figure 2.

TABLE VI: Mutual information scores of different methods to predict the generalization gap using CIFAR-10 and SHVN.

Methods	Prediction Score ↑
VC Dimension [45]	0.020
Jacobian norm w.r.t intermediate layers [46]	2.061
Distance to Initialization [24]	4.921
Sharpness of the convergence point [47]	10.667
ContRE (ours)	**13.531**

efficient and effective estimator for the generalization performance in practice, and that `ContRE` is a good estimator for the generalization gap in condition that the training accuracy be controlled on the same level.

VI. CONCLUSION AND DISCUSSION

In this work, we build a framework `ContRE` to generate contrastive examples and measure the consistency of DNN models' behaviors with contrastive examples or training examples as input. This consistency can be exploited to estimate the generalization performance of DNN models, endowed by the assumption that robust DNN models with good generalization performance tends to giving consistent features and predictions from the same image under varying data transformations. We adopt RandAugment to generate contrastive examples and have practically assessed that the proposed `ContRE` is able to consistently estimate the generalization performance of various DNN models on three benchmark datasets. Systematical ablation studies and thorough analyses have also been provided to demonstrate the versatility of `ContRE` in complex real-world situations and to dissect the reasons that `ContRE` works well for the estimation.

We further discuss the potential limitations in the estimation of generalization performance by our proposed framework: (1) The data transformations may spoil the data and lead to total random guess in the worst case. We show that the transformations used by `ContRE` would not significantly affect the generalization estimation, by additionally reporting the correlations between transformed training and testing data, i.e., $\mathcal{L}_{\boldsymbol{\theta}}(R(\boldsymbol{Z}))$ and $\mathcal{L}_{\boldsymbol{\theta}}(R(\boldsymbol{Z}_{test}))$ (0.983, 0.903 and 0.989 for CIFAR10, CIFAR100 and ImageNet respectively), but for further extensions, transformations should be cautiously chosen. (2) Theoretical analyses of local elasticity [9], [10],

[20] seems related to our approach, but theoretical links among contrastive examples, generalization performance and this notion are not available yet. (3) Lack of appropriate data transformations, it would be difficult to extend to other data formats, such as texts, audios, graphs etc. Future investigation on the existing transformations and exploration on new ones would address this limitation.

REFERENCES

[1] Y. Jiang, P. Foret, S. Yak, D. M. Roy, *et al.*, "Neurips 2020 competition: Predicting generalization in deep learning", *arXiv preprint arXiv:2012.07976*, in press, 2020.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations", in *International Conference on Machine Learning (ICML)*, 2020.

[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.9729–9738, 2020.

[4] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach", in *International Conference on Machine Learning (ICML)*, pp.254–263, 2018.

[5] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning", in *Advances in Neural Information Processing Systems (NeurIPS)*, pp.5949–5958, 2017.

[6] Y. Jiang, D. Krishnan, H. Mobahi, and S. Bengio, "Predicting the generalization gap in deep networks with margin distributions", in *International Conference on Learning Representations (ICML)*, 2019.

[7] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them", in *International Conference on Learning Representations (ICML)*, 2020.

[8] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?", in *Proceedings of ICML*, pp.5389–5400, 2019.

[9] O. Bousquet and A. Elisseeff, "Stability and generalization", *The Journal of Machine Learning Research*, vol.2, pp.499–526, 2002.

[10] Z. Deng, H. He, and W. J. Su, "Toward better generalization bounds with locally elastic stability", *arXiv preprint arXiv:2010.13988*, in press, 2020.

[11] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization", *arXiv preprint arXiv:1611.03530*, in press, 2016.

[12] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning", *arXiv preprint arXiv:2003.04297*, in press, 2020.
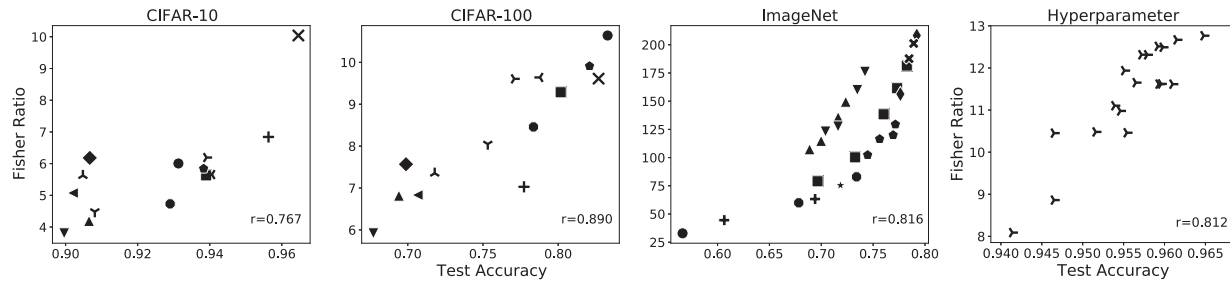
Fig. 6: Spearman correlations between test accuracy and Fisher ratios on contrastive examples. The same DNN models are used as previously for the left three figures, while models across hyper-parameter choices of the wide-ResNet-26-10 on CIFAR-10 is used for the last figure, corresponding to Section V-A. Symbols for DNN models are the same as in Figure 2.

[13] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp.702–703, 2020.

[14] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners", in *Advances in Neural Information Processing Systems (NeurIPS)*, vol.33, pp.22243–22255, 2020.

[15] M. Caron, I. Misra, J. Mairal, P. Goyal, *et al.*, "Unsupervised learning of visual features by contrasting cluster assignments", in *Advances in Neural Information Processing Systems (NeurIPS)*, vol.33, pp.9912–9924, 2020.

[16] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudeva, and Q. V. Le, "Autoaugment: Learning augmentation policies from data", *arXiv preprint arXiv:1805.09501*, in press, 2018.

[17] S. Yak, J. Gonzalvo, and H. Mazzawi, "Towards task and architecture-independent generalization gap predictors", *arXiv preprint arXiv:1906.01550*, in press, 2019.

[18] T. Unterthiner, D. Keysers, S. Gelly, O. Bousquet, and I. Tolstikhin, "Predicting neural network accuracy from weights", *arXiv preprint arXiv:2002.11448*, in press, 2021.

[19] C. A. Corneanu, S. Escalera, and A. M. Martinez, "Computing the testing error without a testing set", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] H. He and W. J. Su, "The local elasticity of neural networks", *arXiv preprint arXiv:1910.06943*, in press, 2019.

[21] D. A. McAllester, "Pac-bayesian model averaging", in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, Association for Computing Machinery, New York, NY, USA, COLT '99, artilce no.164–170, 1999, ISBN 1581131674.

[22] N. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima", *arXiv preprint arXiv:1609.04836*, in press, 2016.

[23] T. Liang, T. A. Poggio, A. Rakhlin, and J. Stokes, "Fisher-rao metric, geometry, and complexity of neural networks", *arXiv preprint arXiv:1711.01530*, in press, 2017.

[24] V. Nagarajan and J. Z. Kolter, "Generalization in deep networks: The role of distance from initialization", *arXiv preprint arXiv:1901.01672*, in press, 2019.

[25] B. Neyshabur, R. Salakhutdinov, and N. Srebro, "Path-sgd: Path-normalized optimization in deep neural networks", in *Advances in Neural Information Processing Systems (NeurIPS)*, pp.2422–2430, 2015.

[26] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks", in *Advances in Neural Information Processing Systems (NeurIPS)*, pp.6241–6250, 2017.

[27] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding", *arXiv preprint arXiv:1807.03748*, in press, 2018.

[28] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding", *arXiv preprint arXiv:1906.05849*, in press, 2019.

[29] O. Henaff, A. Srinivas, J. De Fauw, A. Razavi, *et al.*, "Data-efficient image recognition with contrastive predictive coding", in *International Conference on Machine Learning (ICML)*, pp.4182–4192, 2020.

[30] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[31] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, *et al.*, "Intriguing properties of neural networks", *arXiv preprint arXiv:1312.6199*, in press, 2013.

[32] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples", *arXiv preprint arXiv:1412.6572*, in press, 2014.

[33] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, *et al.*, "Adversarial examples are not bugs, they are features", in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[34] H. Eghbal-zadeh, K. Koutini, P. Primus, V. Haunschmid, *et al.*, "On data augmentation and adversarial risk: An empirical analysis", *arXiv preprint arXiv:2007.02650*, in press, 2020.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in *International Conference on Learning Representations (ICLR)*, 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] S. Zagoruyko and N. Komodakis, "Wide residual networks", in *Proceedings of the British Machine Vision Conference (BMVC)*, pp.87.1–87.12, 2016.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4700–4708, 2017.

[39] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[40] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design", in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.116–131, 2018.

[42] M. Tan, B. Chen, R. Pang, V. Vasudevan, *et al.*, "Mnasnet: Platform-aware neural architecture search for mobile.", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2820–2828, 2019.

[43] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications", *arXiv preprint arXiv:1704.04861*, in press, 2017.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 2013.

[46] Y. Jiang, P. Foret, S. Yak, B. Neyshabur, *et al.*, "Predicting generalization in deep learning, competition at NeurIPS 2020", Available at: https://sites.google.com/view/pgdl2020/home?authuser=0, , 2020.

[47] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them", *arXiv preprint arXiv:1912.02178*, in press, 2019.