TEST-TIME AUTOEVAL WITH SUPPORTING SELF-SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Automatic Model Evaluation (AutoEval) framework entertains the possibility of evaluating a trained machine learning model without resorting to a labeled testing set, which commonly isn't accessible nor provided in real-world scenarios. Existing AutoEval methods always rely on computing distribution shift between the unlabelled testing set and the training set. However, this lines of work cannot fit well in some real-world ML applications like edge computing boxes where the original training set is inaccessible. Contrastive Learning (CL) is an efficient self-supervised learning task, which can learn helpful visual representations for down-stream classification tasks. In our work, we surprisingly find that CL accuracy and classification accuracy can build strong linear correlation ($r > 0.88$). This finding motivates us to regress classification accuracy with CL accuracy. In our experiments, we show that without touching training sets, our framework can achieve results comparable to SOTA AutoEval baselines. Besides, our subsequent experiments demonstrate that different CL approaches and model structures can easily fit into our framework.

1 INTRODUCTION

When evaluating classification model performance in real-world scenarios, a common approach is taking its score on labeled test sets as an estimation. However, such ground-truth labels are always expensive to acquire and may lose feasibility in various testing environments. For example, cameras deployed in an autonomous driving car need to recognize objects in the city, countryside or mountains. We can't collect and annotate testing samples from every possible testing environment. But to guarantee the model's performance and reliability, especially for some high-stake applications such as autonomous-driving and medical-diagnosis, we must fully test the model's performance in different environments. To tackle such problem, we need a method to estimate the accuracy of a given trained model under varying testing environments.

This important yet insufficiently researched problem is introduced as **Automatic Model Evaluation (AutoEval)** by Deng & Zheng (2021). Researchers always rely on some widely-recognized benchmarks, which has labeled testing sets, to evaluate and compare the model performance. However, real-world data follow various distributions, radically different from these well-preprocessed benchmark datasets constructed from intentionally selected testing samples. To make the model evaluation more feasible, it's necessary to evaluate on real-world data. However, ground-truth labels of real-world samples are always difficult and expensive to acquire. Overall, the aim of AutoEval can be formulated as: *estimating a given model's accuracy on different unlabeled testing sets*.

Without direct access to ground-truth labels on testing sets, some previous works such as Garg et al. (2022); Guillory et al. (2021); Deng & Zheng (2021) have developed methods based on measuring distribution shifts to predict the model's accuracy or accuracy drop. The core idea of them is choosing an appropriate metric of distribution shift between the training and testing sets (for example, Deng & Zheng (2021) compute the Frechet Distance between the training set and given testing set). Other works such as Jiang et al. (2021a); Corneanu et al. (2020) leverage model's different status in the training stage. Although these methods are well-performed, they may lose feasibility in some real-world scenarios where the training set in inaccessible (for example, on edge computing devices, there is no storage for a large training set). To avoid involving the training set into estimating the

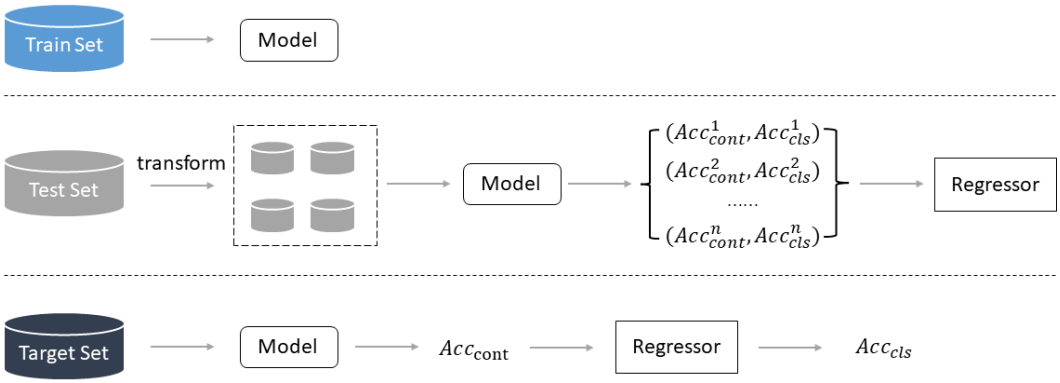


Figure 1: A simple overview of our proposed Test-Time AutoEval framework powered by contrastive learning. The core idea of it is using the contrastive learning accuracy (i.e. how often a CL model is capable of identifying the augmented version of an image early on.) to regress the semantic classification accuracy. The regression model is trained by (Acc_{cont}, Acc_{cls}) pairs on many labeled testing sets (here generated synthetically).

model's performance, in this work, we develop a Test-Time AutoEval framework, which only needs the given trained model and testing sets.

In recent years, self-supervised learning has been deeply explored to deal with unlabeled datasets. The fundamental idea is generating label information in a bootstrap manner by data augmentation or other ways. Naturally, we begin to consider if AutoEval can be powered by self-supervised learning tasks, and which kind of learning task should be chosen. Among these researches, Contrastive Learning (CL) has achieved remarkable results Chen et al. (2020a); He et al. (2019); Chen et al. (2020b); Caron et al. (2020); Grill et al. (2020); Chen & He (2020). Using CL as the pre-training task, the downstream semantic classification task can be improved significantly, which means that CL learns helpful positive semantic features for classification. In other words, a network achieving good performance on contrastive learning is also likely perform well on semantic classification.

Based on this positive correlation mentioned above, we want to know if we can build quantitative relationship between CL accuracy and classification accuracy. Surprisingly, in our empirical experiments, we show that training in a multi-task way, we can build strong linear correlation ($r > 0.88$) between them. This finding encourages us to predict semantic classification accuracy according to contrastive learning accuracy using a linear regression model. Further experimental results show that our method can precisely estimate a given classifier's accuracy on unseen unlabeled testing sets. We also discuss how contrasive learning takes effect in the framework and influences the prediction result. In summary, our contributions are as follows:

- We propose a test-time AutoEval framework (see figure 1), in which we train SimCLR and classification in a multi-task way, and experimental results demonstrate the strong linear correlation between their accuracies.
- Based on the findings above, we use linear regression to predict image classifacation accuracy with contrastive learning accuracy, which reduces the error to an acceptable range ($< 5.3\%$), making a crucial step forward in practical application of AutoEval.
- We explore the influence of some essential factors of contrastive learning in our framework. Experimental results demonstrate that common contrastive learning methods can perfectly fit into our framework.

2 Related Works

Data-centric AutoEval. Data-centric AutoEval Deng & Zheng (2021); Deng et al. (2021); Yu et al. (2022); Chuang et al. (2020a); Chen et al. (2021a); Guillory et al. (2021); Hendrycks & Gimpel (2016); Garg et al. (2022); Chen et al. (2021c); Jiang et al. (2021b) aims to predict a given model's accuracy on a series of invisible testing sets. In this branch of works, a common approach is finding

a metric for the distribution shift between in-domain and out-ouf-domain data, which is then used to regress the model's testing accuracy. For example, Guillory et al. (2021) uses Difference of Confidences (DoC) and Difference of Average Entropy (DoE) to measure the distribution shift. Based on this, Garg et al. (2022) proposes Average Thresholded Confidence (ATC). Deng & Zheng (2021) uses Frechet Distance (FD) to regress OOD accuracies. To compute the distribution shift metric, these methods need to take the training set as an anchor. In contrast, our proposed framework doesn't need access to the original training set, which benefits application cases where the training and testing processes need to be strictly separated.

Model-centric AutoEval. Model-centric AutoEval Corneanu et al. (2020); Jiang et al. (2018); Zhang et al. (2021); Gain & Siegelmann (2020); Unterthiner et al. (2020) aims to evaluate a population or an ensemble of models on an testing set. Compared to Data-centric AutoEval which focuses on measuring the distribution shift between datasets, this branch emphasizes the model's inherent properties. For example, Corneanu et al. (2020) defines DNNs on a topological space, and then calculate a set of compact descriptors of this space to measure important properties of the network's behavior. Jiang et al. (2018) extracts the margin distribution of each layer of the DNN, and compute statistics to regress the testing accuracy.

Contrastive Learning Contrastive Learning (CL) Caron et al. (2020); Chen et al. (2020a); He et al. (2019); Grill et al. (2020); Chen & He (2020) is a kind of self-supervised learning task to learn efficient representation of input samples. Its basic idea is to generate different views of input samples by a series of data augmentation, of which homologous ones are considered as positive samples and the others are negative samples. By aligning positive samples and discriminating negative samples, an embedding for each input sample can be learned. Used as a pre-training task, it can significantly enhance the downstream semantic classification task, which means CL learns helpful features for classification. In this work, we choose Contrastive Learning as the supporting self-supervised learning task, and take SimCLR as an instance to validate its feasibility.

3 Identifying the Correlation Between Accuracies

3.1 Adopting SimCLR in Multi-task Training

The key idea of our method is using contrastive learning accuracy to regress semantic classification accuracy. So we need to identify the underlying correlation between them first, and explore the possibility that the correlation can be modeled to train a regressor. Here, we adopt SimCLR Chen et al. (2020a) as the contrastive learning model. Following this, its top-1 accuracy represents how often a CL model is capable of identifying the augmented version of an image early on. The contrastive learning accuracy and classification accuracy can be written as:

$$Acc_{cont} = \frac{\sum_{i=0}^{B-1} \mathbb{I}(\hat{i} = i)}{B}, Acc_{cls} = \frac{\sum_{i=0}^{B-1} \mathbb{I}(\hat{y} = y)}{B} \tag{1}$$

where B is the input batch size, i is the index of an image in original batch and $\hat{i}$ is the index of its most probable augmented view predicted by model. y and $\hat{y}$ are ground-truth labels and predicted labels respectively.

In common approaches, contrastive learning is often used as a pre-training task for image semantic classification. However, we notice some previous works like Wang et al. (2022); Arora et al. (2019); Lee et al. (2021) have theoretically clarified that: contrastive learning cannot guarantee the learning of class-discriminative features, because class-uniform features also minimizes the InfoNCE Loss. When the InfoNCE Loss for contrastive learning achieves its minimum, there is an upper bound for the downstream classification accuracy. In other words, there will be no obvious correlation between them in a pre-training manner. So in this work, we attempt to adopt SimCLR in a multi-task manner (see figure 2), which minimizes:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{InfoNCE} \tag{2}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss for classification and $\mathcal{L}_{InfoNCE}$ represents the InfoNCE loss for contrastive learning.

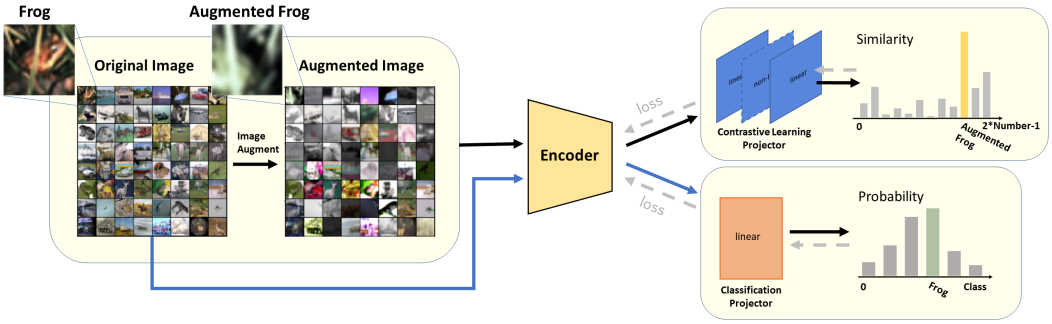


Figure 2: The model structure of training SimCLR and image classification in a multi-task way. A CNN is used as a sharing backbone to encode image features, followed by two task heads for contrastive learning (upper) and semantic classification (lower). Note that here we show the original batch and corresponding augmented batch, but when calculating the classification accuracy, we only consider the original batch.

3.2 Dataset Synthesis

To identify the correlation between Acc_{cont} and Acc_{cls}, we need to apply the framework on many labeled testing sets. According to Deng et al. (2021), these testing sets should have: 1) various distributions. 2) the same classes with the training set. 3) sufficient samples. However, it's quite difficult to collect such testing sets from natural distributions. So following Deng & Zheng (2021); Deng et al. (2021), we synthesize these testing sets by applying transformations on hand-written digit images and natural object images (specifically, MNIST, CIFAR-10 and CIFAR-100) for empirical analysis. Notably, the transformations have no influence for semantic information, so the labels of original images are inherited. Some examples of the synthetic sets are shown in figure 3. For each setup, we generate 200 synthetic testing sets to empirically study the correlation between Acc_{cont} and Acc_{cls}.

(a) original (b) synthetic 1 (c) synthetic 2 (d) synthetic 3

Figure 3: Examples of the synthetic testing sets. Here we use sharpness and color changing. By applying different transformations on the original dataset, we can obtain many testing sets with various distributions.

MNIST setup. We synthesize test sets by applying various transformations on MNIST dataset. As MNIST contains simple gray-scale images, we first consider change their black background to color ones. Specifically, for each sample of MNIST, we randomly select an image from COCO dataset and randomly crop a patch, which is then used as the background of the hand-written digit. After that, we apply six image transformations on the background-replaced images: *autoContrast, rotation, color, brightness, sharpness, translation*.

CIFAR-10 setup. We synthesize test sets based on CIFAR-10 by applying image transformations on it. The transformations are randomly selected from {*Sharpness, Equalize, ColorTemperature, Solarize, Autocontrast, Brightness, Rotate*}. Then other another random transformations are applied: *RandomCrop, RandomHorizontalFlip*. Note that all possibilities of the operation sequences are not calculated by permutation and combination, because there are many random states when applying these transformations, thus we can generate image sets of various distributions.

CIFAR-100 setup. In order to check if the correlation still exists when there are .Similar to CIFAR-10 above, we use the training split of CIFAR-100 dataset to train our model and the testing split to synthesize many test sets. But for lack of appropriate real-world unseen test sets containing both various features and same classes (as AutoEval is a newly proposed and under-explored problem), we further split the CIFAR-100 testing split into two parts for synthesizing different test sets and the final unseen test.

3.3 Correlation Analysis

For each setup, we train CoLA on the original training set, then test it on each of the synthetic testing sets introduced above and plot the results. Following Deng et al. (2021), the backbone for MNIST setup is LeNet-5, and for CIFAR setup is DenseNet-40-12 (40 layers with growth rate 12). From figure 4, we can see that on various testing sets (or under various testing environments), the contrastive learning accuracy and classification accuracy exhibit strong linear correlation (Pearson's correlation score $r > 0.88$). This finding encourages us to train a linear regressor to predict classification accuracy on unlabelled testing sets.

3.4 Problem Definition

Consider a image classification task. Given a classifier trained on the training set $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i denotes the i-th training image, $y_i \in \{1, 2, \cdots, C\}$ denotes the class label of it, and n denotes the training set size. Our aim is to estimate its classification accuracy on different test sets. Each test set is denoted as $D_{test}^{t} = \{x_j^t\}_{j=1}^{n_t}$, where x_j^t denotes the j-th test image of the t-th test set, $t = 1, 2, \cdots, M$. Notably, our test set holds no supervision as we aim to evaluate model performance automatically. Based on the findings that the two tasks show a strong linear correlation, we are motivated to predict a classifier's accuracy with its contrastive accuracy using a linear regressor:

$$Acc_{cls} = W \cdot Acc_{cont} + b \tag{3}$$

where W, b are parameters learned by robust linear regression Huber (2004).

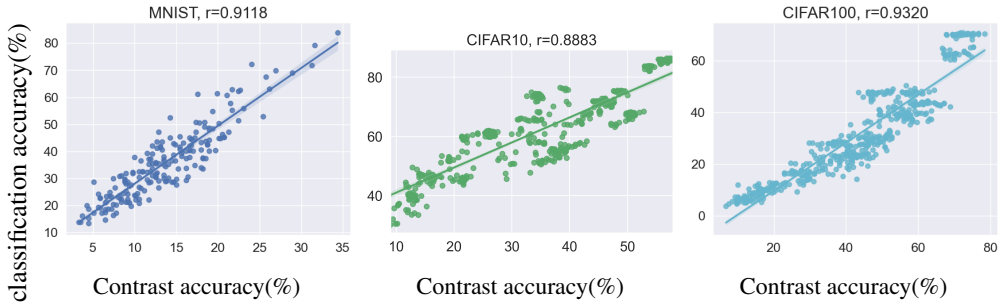


Figure 4: The correlations between contrastive learning accuracy and image classification accuracy. The x-axis represents the contrastive learning accuracy, while the y-axis represents the classification accuracy, respectively. Each point represents a synthetic test set. **Left:** the results of the MNIST setup with LeNet-5, **Mid:** the results of the CIFAR-10 setup with DenseNet-40-12, **Right:** the results of the CIFAR-100 setup with DenseNet-40-12. The trend line is obtained using robust linear regression Huber (2004).

4 Experiments

Following previous works Deng & Zheng (2021); Deng et al. (2021); Guillory et al. (2021); Garg et al. (2022), in our experiments we evaluate our framework CoLA on both hand-written digits and natural images. Given a trained classifier, we test it on many labelled testing sets with various distribution shifts, and then fit the (Acc_{cont}, Acc_{cls}) to obtain a regressor. Finally, we choose some unseen labelled sets, and evaluate the error between the ground-truth accuracy (calculated by ground-truth labels) and estimated accuracy (predicted by regression) using RMSE.

4.1 EXPERIMENTAL SETUP

Datasets and Models. For hand written digits, we train LeNet-5 on MNIST and test it on SVHN Netzer et al. (2011) and USPS Hull (1994) (both contain digit images with 10 classes). For natural images classification, we build CIFAR-10 setup and COCO setup. Specifically, we train DenseNet-40-12 on CIFAR-10 and test it on CIFAR-10.1 Recht et al. (2018). For the COCO setup, following Deng et al. (2021), we choose 12 object classes (aeroplane, bicycle, bird, boat, bottle, bus, car, dog, horse, tv-monitor, motorcycle, person). To build the training set, object annotation boxes among these 12 classes are cropped from the COCO training images containing them. These images are used to train a ResNet-50 backbone. Similarly, we build unseen testing sets for classification accuracy from Caltech-256 Griffin et al. (2007), PASCAL VOC 2007 Everingham et al. (2010), ImageNet Deng et al. (2009), also select from the 12 classes above.

Baselines. There are few baselines for direct comparison under the AutoEval problem. But in previous works such as Deng & Zheng (2021); Deng et al. (2021); Hendrycks & Gimpel (2016), we notice there are some model-generalization prediction methods based on distribution shifts or softmax probabilities for reference.

CONFIDENCE SCORE. This metric is simple and intuitive. In the output of the last layer of a classifier (i.e. the softmax output), its maximum probability value is refererred to as Confidence Score. If this score is greater than a given threshold value, the corresponding input image is considered correctly classified.

ENTROPY SCORE. Similar to the Confidence Score above, this score is also given by the softmax output. For a K-classes classification task, we compute the entropy of softmax outputs, and then normalize it by $\log K$. If this score is less than a given threshold value (or the negative entropy is greater than this value), then it is considered as a correctly classified sample.

FRECHET DISTANCE. Deng & Zheng (2021) uses Frechet Distance (FD) Dowson & Landau (1982) to measure the domain gap (or distribution shift) between the original training set and testing sets synthesized based on it. They find that there is a strong negative linear correlation between FD and classifier's accuracy on both digit classification and natural image classification scenarios. Motivated by this finding, they estimate the classification accuracy on unlabelled testing sets by a linear regression model trained on many (FD, Acc_{cls}) pairs.

ROTATION PREDICTION. Deng et al. (2021) uses image rotation prediction Gidaris et al. (2018) as a pretext task of semantic classification. Specifically, they learn a multi-task neural network for semantic classification along with an auxiliary rotation prediction task. Their empricial results show that testing a given multi-task model on different synthesized testing sets, the semantic classification accuracy and rotation prediction accuracy have a linear correlationship, which motivates them to regress the classification accuracy with rotation prediction accuracy.

4.2 RESULTS

The performance of our proposed framework is shown in table 1. It is compared with different baselines on RMSE error. Here are some preliminary findings:

Using linear regression to predict classification accuracy is feasible. We can observe that the accuracy on unlabelled testing sets estimated by linear regression is very close to the ground-truth value. For digits classification testing sets (SVHN and USPS), our prediction RMSE is $5.34(\%)$. For natural image classification, our prediction RMSE are $0.48(\%)$ and $2.30(\%)$ under CIFAR-10 setup and COCO setup respectively. Because the previously generated synthetic sets have various distribution shifts, well simulating many kinds of testing environments, we can estimate a classifier's accuracy with its performance on contrastive learning task without much lost of precision.

Contrastive Learning is a more powerful auxiliary task than Rotation Prediction. As shown in the table, our framework outperforms the *Rot+Cls* baselines. Under every experimental setup, the RMSE of ours is less than that of using Rotation Prediction as an auxiliary task. Contrastive Learning

Table 1: Results of the classification accuracy estimation using linear regression with contrastive learning accuracy on several groups of unseen testing sets: 1) SVHN and USPS; 2) CIFAR-10.1; 3) Caltech, Pascal and ImageNet. The training sets of the task settings are MNIST, CIFAR-10 and COCO, respectively. For the baselines and our method, we report the ground-truth accuracy (calculated using ground-truth labels) and linear regression accuracy, with respect to the unseen testing sets. For each group of task setting, we report the RMSE of the linear regression prediction results. Note that different baselines may have different ground-truth accuracies, because multi-task training may influence the classification performance, we just need to keep them as close values.

	Train set	MNIST			CIFAR-10		COCO			
Manner	Unseen test set	SVHN	USPS	RMSE	CIFAR-10.1	RMSE	Caltech	Pascal	ImageNet	RMSE
Single-task	Ground-truth	25.46	64.08	-	91.24	-	93.40	86.13	88.83	-
	Confidence	22.07	30.39	23.94	86.85	4.39	84.30	78.00	79.83	8.75
	Entropy	26.63	33.23	21.83	89.20	2.04	86.80	80.14	82.50	6.31
	FD	26.28	50.14	9.87	-	-	79.77	83.87	83.19	8.62
Multi-task	Ground-truth (Rot+Cls)	23.06	65.52	-	88.15	-	92.61	86.43	87.83	-
	Linear-reg (Rot+Cls)	24.84	53.10	8.87	91.89	3.74	90.70	89.29	90.98	2.68
	Ground-truth (Ours)	19.69	63.57	-	86.22	-	91.81	90.33	88.28	-
	Linear-reg (Ours)	26.40	67.04	**5.34**	86.70	**0.48**	94.64	88.90	85.87	**2.30**

can learn more representative features for semantic classification, and help obtain an estimation closer to the ground-truth classification accuracy.

5 Discussion and Analysis

Multi-task training will not degrade the performance of classification when trained along with contrastive learning. In Lee et al. (2019), the authors point out that learning through methods such as data augmentation or multi-task learning enforces a certain invariance of features, which will make learning more difficult and may lead to performance degradation. In Deng et al. (2021), they also emphasize that the auxiliary task should satisfy the following requirements: 1) introduces no learning complexity for the main classification. 2) requires minimal network structural change. 3) does not degrade classification accuracy.

In this work, we need to ensure that SimCLR does not influence the model performance on classification task. From Table 1, we can see that there are no significant differences between the classification accuracy of our framework and baseline. Specifically, under MNIST setup, our ground-truth accuracy is 63.57(%) on USPS, almost the same as baseline 65.52(%). Under CIFAR-10 setup and COCO setup, our ground-truth accuracy is very close to that of baseline (with an absolulte error $< 4\%$). That means using contrastive learning as the auxiliary task of classification will not degrade the main task performance.

Optimal settings of Contrastive Learning can be directly inherited in our framework for best performance. The success of Contrastive Learning can be largely attributed to various data augmentations for generating positive and negative examples. As we adopt SimCLR in our framework, we want to know if our overall performance is consistent to its basic settings (i.e. the linear correlation coefficient achieves its maximum under the best training parameters of SimCLR). According to Wang et al. (2022), among the data augmentations adopted in SimCLR, RandomResizedCrop is the most important augmentation, and ColorJitter is the second. So we study the influence of these two kinds of augmentations in our work.

For RandomResizedCrop, to quantify its influence, we use the augmentation strength defined in Wang et al. (2022). For a RandomResizedCrop operator with scale range $[a, b]$, its aug-strength can be defined as $r = (1 - b) + (1 - a)$. In figure 5, we show that on the synthetic testing sets generated from MNIST, KMNIST and FashionMNIST, under different augmentation strengths, the

linear correlation coefficient achieves its maximum at the default strength value 0.92. For ColorJitter, we study its parameters: brightness, contrast, saturation and hue, where the augmentation strength is corresponding to the parameter value. Note that all other augmentations in SimCLR are kept default. In figure 6, for each of these parameters, we plot the changes of the linear correlation coefficient under the CIFAR-10 setup. Here we have similar observation that the default parameter values in SimCLR yield best linear correlation.

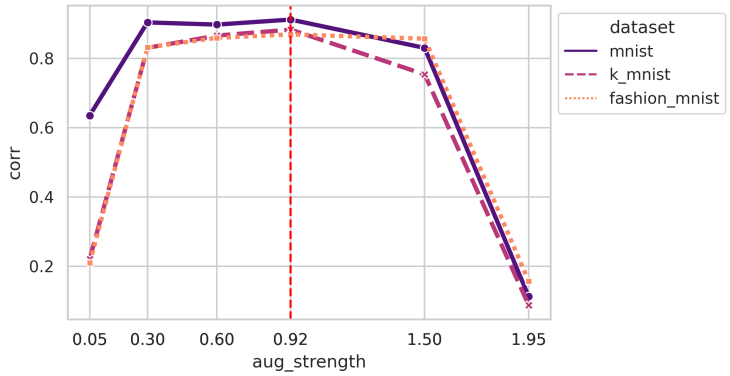
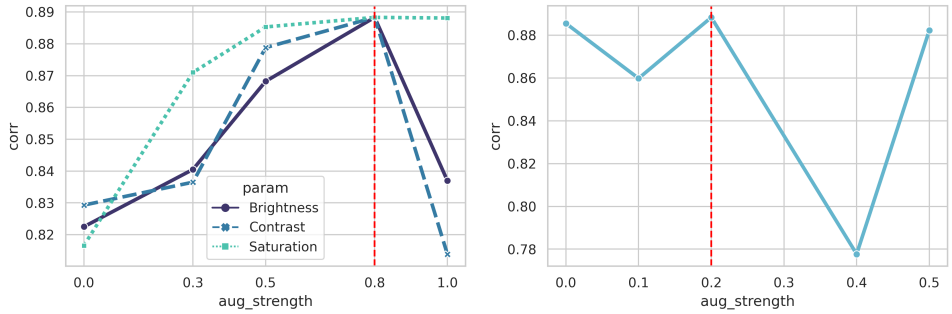


Figure 5: Changes of the linear correlation coefficient under different augmentation strengths. When using the default RandomResizedCrop strength 0.92, the linear correlation coefficient achieves its maximum value.

Besides data augmentations, temperature scaling is also an import factor during the training process of SimCLR. We study the temperature parameter τ on MNIST and CIFAR-10. As figure 7 shows, when using default temperature value $\tau = 0.07$, we can obtain best linear correlation. The empirical results demonstrate that when adopting contrastive learning frameworks, keeping default optimal settings is most likely to build strong linear correlation between the CL accuracy and classification accuracy.

The linear correlation is robust against different training settings. Here we empirically study the robustness of the linear correlation between the contrastive learning accuracy and semantic classification accuracy. Specifically, we consider different model structures, testing set amounts, and different augmentation of contrastive learning frameworks.

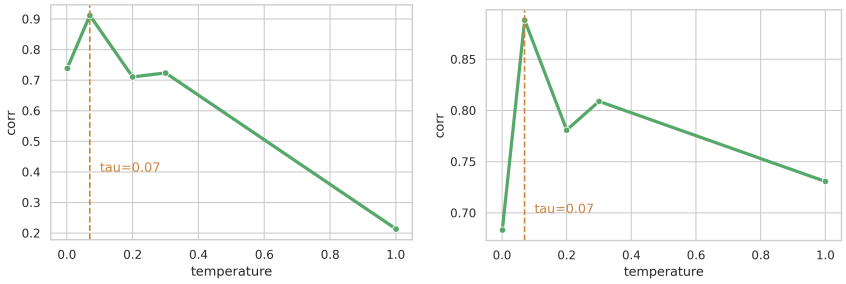
DIFFERENT BACKBONES. To study if the linear correlation between contrastive learning accuracy and classification accuracy relies on specific model structures, we change different CNN backbones as the feature extractor. In figure 8a, we plot the changes of linear correlation coefficient using



(a) Brightness, Contrast, Saturation

(b) Hue

Figure 6: Changes of linear correlation coefficient against different ColorJitter strengths under CIFAR-10 setup. In SimCLR, the default parameter values of brightness, contrast, saturation and hue are 0.8, 0.8, 0.8 and 0.2 respectively, which yield best linear correlation.



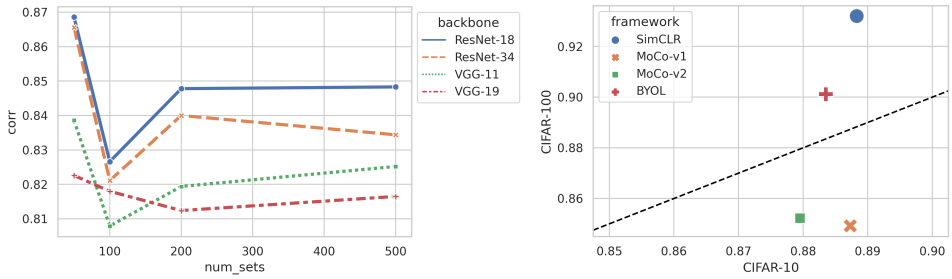
(a) MNIST (b) CIFAR-10

Figure 7: Changes of linear correlation coefficient on MNIST and CIFAR-10 with different temperature values.

different backbones (ResNet-18, ResNet-34, VGG-11 and VGG-19) under CIFAR-10 setup. We can observe that the linear correlation is robust against changing backbones and the amount of testing sets. Meanwhile, the robustness is not influenced by the number of testing sets. This means our framework can be used to evaluate various kinds of models.

DIFFERENT CONTRASTIVE LEARNING AUGMENTATION GROUPS. In this paper, we adopt SimCLR as the contrastive learning framework. To study if other frameworks can fit in well, we change SimCLR to MoCo-v1, MoCo-v2, and BYOL. From figure 8b, we can observe that the linear correlations are all strong across different CL frameworks ($r > 0.87$).

DIFFERENT AMOUNT OF SEMANTIC CLASSES. For the image classification task, more semantic classes usually mean higher difficulty, which may weaken the linear correlation. To study this, we compare the results on CIFAR-10 and CIFAR-100 (with similar input images but different amounts of classes). As figure 8b shows, all the points distribute near the identity line, which means the linear correlation is robust against the amount of semantic classes.



(a) Backbones (b) CL frameworks

Figure 8: Study of linear correlation robustness. 8a shows the changes of linear correlation coefficient computed on different amounts of testing sets using different CNN backbones. 8b shows the linear correlation trained by SimCLR, MoCo-v1, MoCo-v2 and BYOL (the x-axis and y-axis represent the linear correlation coefficient on CIFAR-10 and CIFAR-100, respectively).

6 CONCLUSION

In this paper, we propose a novel framework for estimating classifier accuracy on invisible test sets without ground-truth labels. We find that training in a multi-task manner, there is a strong linear correlation between contrastive learning accuracy and classification accuracy, which indicates that it is feasible to estimate classifier accuracy using linear regression. We train SimCLR and image classification in a multi-task way, and use the contrast accuracy to estimate classification accuracy. Experimental results show that our method outperforms previous works. We hope our work can motivate future research on AutoEval techniques assisted by self-supervised learning.

REFERENCES

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020. URL https://arxiv.org/abs/2006.09882.

Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *Advances in Neural Information Processing Systems*, 34:14980–14992, 2021a.

Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles, 2021b. URL https://arxiv.org/abs/2106.15728.

Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. In *International Conference on Machine Learning*, pp. 1617–1629. PMLR, 2021c.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020. URL https://arxiv.org/abs/2011.10566.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. *arXiv preprint arXiv:2007.03511*, 2020a.

Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1984–1994. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/chuang20a.html.

Ciprian A Corneanu, Sergio Escalera, and Aleix M Martinez. Computing the testing error without a testing set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2677–2685, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proc. CVPR*, 2021.

Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning*, pp. 2579–2589. PMLR, 2021.

DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Alex Gain and Hava Siegelmann. Abstraction mechanisms predict generalization in deep neural networks. In *International Conference on Machine Learning*, pp. 3357–3366. PMLR, 2020.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234*, 2022.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020. URL https://arxiv.org/abs/2006.07733.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1134–1144, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. *arXiv preprint arXiv:1810.00113*, 2018.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter. Assessing generalization of SGD via disagreement. *CoRR*, abs/2106.13799, 2021a. URL https://arxiv.org/abs/2106.13799.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021b.

Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Rethinking data augmentation: Self-supervision and self-distillation. 2019.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2021. URL https://arxiv.org/abs/2110.05208.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training, 2021. URL https://arxiv.org/abs/2112.12750.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Thomas Unterthiner, Daniel Keysers, Sylvain Gelly, Olivier Bousquet, and Ilya Tolstikhin. Predicting neural network accuracy from weights. *arXiv preprint arXiv:2002.11448*, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap, 2022. URL https://arxiv.org/abs/2203.13457.

Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. *arXiv preprint arXiv:2202.05834*, 2022.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021. URL https://arxiv.org/abs/2111.11432.

Yi Zhang, Arushi Gupta, Nikunj Saunshi, and Sanjeev Arora. On predicting generalization using gans. *arXiv preprint arXiv:2111.14212*, 2021.