Appendix for
Are Labels Always Necessary for Classifier Accuracy Evaluation?

Weijian Deng Liang Zheng
Australian National University
{firstname.lastname}@anu.edu.au

A.1. Appendix

In this Appendix, we first provide a brief theoretical analysis of our finding from the perspective of theoretical error of domain adaptation. we then show the estimations of using different thresholds for the intuitive solution. Furthermore, we show the negative linear correlation between distribution shift and accuracy based on the other two classifiers (*i.e.*, DenseNet-121 and VGG-16). We also use the proposed regression methods to predict their accuracy on three unseen test sets. In addition, we detail the transformations used to generate the meta dataset. Along with this, we study the impact of image transformation and background change on the diversity of the meta set. Last, we provide more visual examples of sample sets for natural image classification and digit classification.

A.2. Theoretical analysis

We use the implications in domain adaptation: dataset divergence degrades model accuracy. Here, we give a brief analysis of our findings from the perspective of theoretical error of domain adaptation.

Theorem 1 (Ben-David *et al.*, 2010 [1]) Let $\mathcal{H}$ be the hypothesis class. Given two domains $\mathcal{S}$ and $\mathcal{T}$, we have

$$\forall h \in \mathcal{H}, \varepsilon_{\mathcal{T}}(h) \leq \varepsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + C. \quad (1)$$

$\varepsilon_{\mathcal{S}}(h)$ is the expected error on source images. $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ defines the discrepancy distance between two domains $\mathcal{S}$ and $\mathcal{T}$ w.r.t. a hypothesis set $\mathcal{H}$. C is the shared expected loss defined as $\min_{h\in\mathcal{H}} \varepsilon_{\mathcal{S}}(h, f_{\mathcal{S}}) + \varepsilon_{\mathcal{T}}(h, f_{\mathcal{T}})$ where $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ are labeling functions on source and target respectively. C does not depend on particular h and is expected to be negligible [2]. **In AutoEval**, the training set and learned classifier are *given and fixed*, so the target domain $\mathcal{T}$ is the *only factor* that determines the bound: different test sets induce different discrepancy distances, giving different upper bounds. In our work, we empirically observe that classifier accuracy decreases in proportion to Fréchet Distance between the source training set and target sets (sample sets).

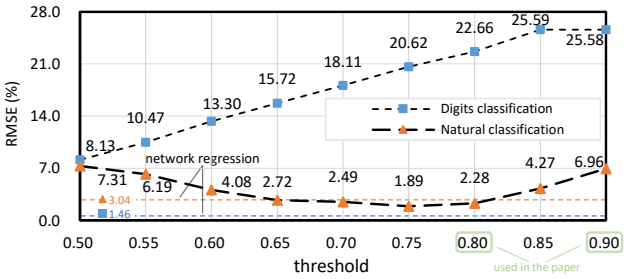

Figure A-1. RMSE of *predicted score* under different thresholds. We observe that it is very threshold-sensitive. Our method does not depend on such a parameter and yields much more stable results.

A.3. Intuitive solution

In Fig. A-1, we show the estimated results of the "predicted score" based intuitive solution with different thresholds. We observe 1) this baseline is sensitive to the threshold. 2) The optimal thresholds differ significantly for the two tasks. The baseline can make a good prediction when an optimal threshold is used. However, it is very threshold-sensitive, and it is infeasible to select the threshold because 1) test labels are unavailable and 2) the test set keeps changing. Our method does not depend on such a parameter and yields much more stable results. That said, it would be interesting to address this drawback in the context of AutoEval.

A.4. Study on different classifiers

A.4.1. Distribution shift vs. Accuracy

In the proof of concept (Section 3.2 of main text), we observe a very strong negative correlation between classifier (ResNet-50) accuracy and distribution shift (Fréchet distance): the Spearman's Rank Correlation (ρ)is about -0.91. It means the strong negative monotonic correlation between accuracy and Fréchet distance.

Here, we show the relationship between distribution and accuracy based on the other classifiers, *i.e.*, Desenet-121, and VGG-16. We show the results on the natural image classification in Fig. A-2. We observe that the range of dis-

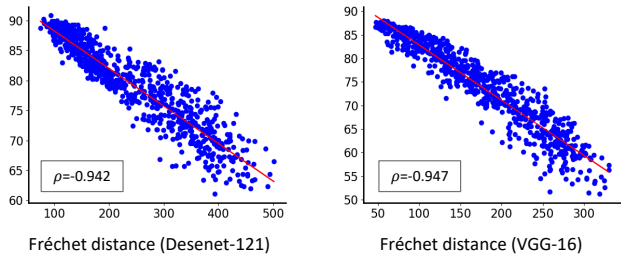

Figure A-2. Relationship between the distribution shift and accuracy based on different classifiers. We show the results on natural image classification. Each point represents a sample set of the meta set. For each classifier, we observe a very strong negative correlation between its accuracy and distribution shift.

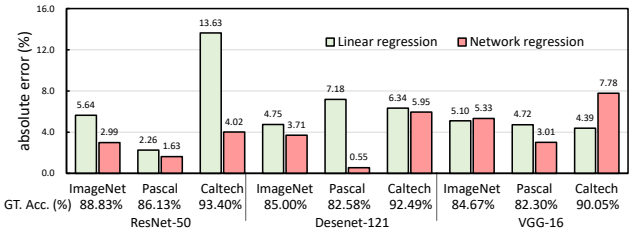

Figure A-3. Absolute errors (%) of two regression methods for three classifiers on three unseen test sets. From left to right is ResNet-50, DenseNet-121, and VGG-16, respectively. We observe the two regression methods are able to make reasonably good predictions for each classifier on three test sets.

tribution shift (horizontal axis) can be varied with different classifiers. However, the overall relationship between the classifier's accuracy and distribution shift is the same. They have a very strong negative correlation: the rank correlation is -0.942 and -0.947 using DenseNet-121 and VGG-16, respectively. That is, the classifier tends to achieve a low accuracy on the sample set which has a high distribution shift with training set $\mathcal{D}_{ori}$. Moreover, the strong negative linear correlation also indicates that it is feasible to predict classifier accuracy based on the distribution difference between training and test set.

A.4.2. Accuracy estimation for different classifiers

We use two regression methods to predict the accuracy of different classifiers on natural image classification. We train DenseNet-121 and VGG-16 on COCO training set, and test them on three unseen test sets. *i.e.*, PASCAL, Caltech, and ImageNet. For each classifier, we use it to extract features needed for two regression models on every sample set. Then, we train two regression models. In Fig. A-3, we report the accuracy estimated results for each classifier. We observe the two methods are able to make accurate predictions for each classifier on three test sets.

Operation	Description	Magnitude
AutoContrast	Maximize the image contrast, by making the darkest pixel black and lightest pixel white.	
Rotate	Rotate the image *magnitude* degrees.	[-30, 30]
Color	Adjust the color balance of the image. A *magnitude*=0 gives a black & white image, whereas *magnitude*=1 gives the original image.	[0.1, 1.9]
Brightness	Adjust the brightness of the image. A *magnitude*=0 gives a black image, whereas *magnitude*=1 gives the original image.	[0.1, 1.9]
Sharpness	Adjust the sharpness of the image. A *magnitude*=0 gives a blurred image, whereas *magnitude*=1 gives the original image.	[0.1, 1.9]
TranslateX/Y	Translate the image in the horizontal (vertical) direction by *magnitude* number of pixels.	[-150, 150]
Cutout	Set a random square patch of side-length *magnitude* pixels to gray.	[0, 60]
ShearX/Y	Shear the image along the horizontal (vertical) axis with rate *magnitude*.	[-0.3, 0.3]
Equalize	Equalize the image histogram.	
ColorTemp	Change the temperature of an image to a given *magnitude* in Kelvin.	[1000, 11000]

Table A-1. List of all image transformations that we choose from during the meta set construction. The magnitude range of each transformation is shown in the third column. Some transformations do not use the magnitude information (*e.g.*, AutoConstrast).

A.5. Meta set study

A.5.1. Image transformation

During meta-set construction, we adopt a two-step procedure: perform background change, and then image transformations. For the transformation, we use six image transformations, including autoContrast, rotation, color, brightness, sharpness, and translation. In practice, we randomly select and combine three out of the six transformations. The image transformations are listed in Table. A-1. We briefly describe each transformation (the second column), and introduce its magnitude information (the third column).

A.5.2. Meta set diversity

In our work, we use a combination of background change and *three* random image transformations for the meta set construction. Both image transformation and background can introduce many visual differences, and thus create diverse sample sets. Here, we study the impact of these two techniques on the diversity of the meta set. Specifically, we construct another three meta sets, 1) Meta set A, construc-

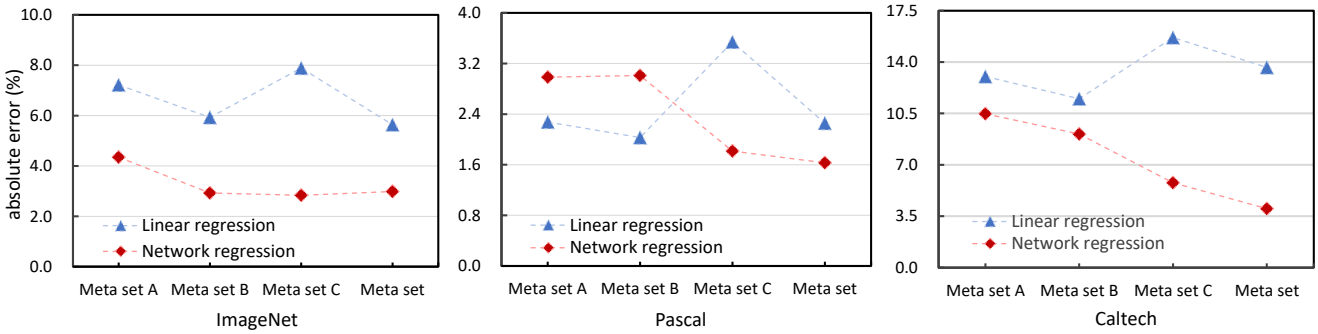

Figure A-4. Absolute errors (%) of regression methods trained on different meta sets. The four meta sets are, 1) **Meta set A**, construction only with background change; 2) **Meta set B**, construction only with three random image transformations; 3) **Meta set C**, construction with background change and only one random image transformation; 4) **meta set**, construction with background change and three random image transformations. The classifier used in this experiment is ResNet-50.

Figure A-5. Seed set and examples of fifteen sample sets for the task of natural image classification. The seed set is sampled from the same distribution as the original training set; they share the same classes but do not have image overlap. The sample sets are generated from the seed by background replacement and image transformations. The sample sets exhibit distinct data distributions, but inherit the foreground objects from the seed, and hence are fully labeled.

tion only with background change; 2) Meta set B, construction only with three random image transformations; 3) Meta set C, construction with background change and only one random image transformation. Using different meta sets, we learn different regression models and then compare their estimation accuracy in Fig. A-4.

We observe both regression methods produce a high absolute error when using meta set A for training. This indicates only changing background cannot introduce sufficient visual changes for constructing a diverse meta set. Moreover, only using image transformations (Meta set B) also insufficient. We note that network regression gains more desirable accuracy when meta set becomes more diverse (using more transformations). *The comparison demonstrates that learning a mapping function from the distribution shift to the classifier's accuracy requires a diverse meta set.*

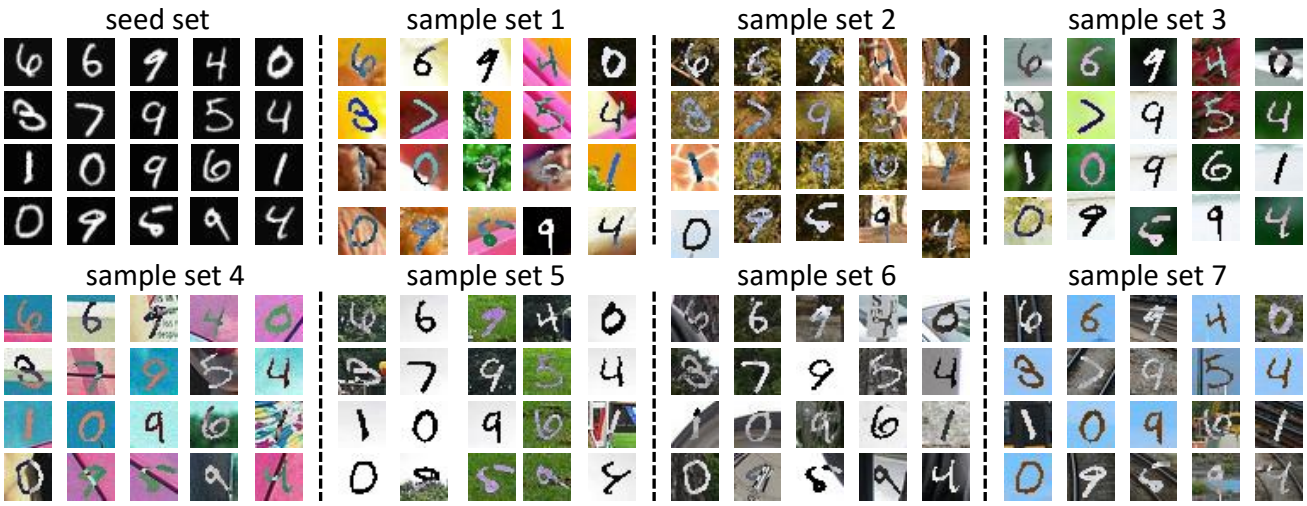

Figure A-6. Seed set and examples of seven sample sets for the task of digit classification.

A.5.3. Sample set example

We show more visual examples of sample sets for natural image classification and digit classification in Fig. A-5 and Fig. A-6, respectively. Each sample set is generated by background change and a combination of three image transformations. Compared with the seed set, each sample set has many visual differences. Thus, each sample set exhibits a distinct data distribution. Moreover, the foreground object is preserved, so the sample set is fully labeled.

A meta set consists of many diverse sample sets. With it, we can learn robust regression models to predict classifier accuracy based on distribution-related statistics (mean and covariance in this work).

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 1

[2] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. ICML*, pages 1180–1189, 2015. 1