Supplementary Material for
A Bag-of-Prototypes Representation for Dataset-Level Applications

Weijie Tu[1] Weijian Deng[1] Tom Gedeon[2] Liang Zheng[1]
[1] Australian National University [2] Curtin University

In this supplementary material, we first introduce the experimental details including the training schemes, datasets and computation resources. Then, we show the full results of the correlation study on the task: training set suitability in Fig. A-1 and impact of codebook size in Fig. A-2. Last, we report the accuracy estimation results on each severity level of CIFAR-10.1-$\bar{C}$ in Table A-1 and results on datasets with natural distribution shifts in Fig. A-3. After that, we present a correlation study of average thresholded confidence, average confidence, and difference of confidence on ImageNet and CIFAR-10 setups.

A. Experimental Setup

A.1. Datasets

We carefully check the licenses of all datasets used in the experiment and list the open sources to them.
DomainNet [12] (http://ai.bu.edu/M3SDA/);
ImageNet [2] (https://www.image-net.org);
ImageNet-C [7] (https://github.com/hendrycks/robustness);
CIFAR-10 [10] (https://www.cs.toronto.edu/ kriz/cifar.html);
CIFAR-10-C [7] (https://github.com/hendrycks/robustness);
CIFAR-10.1 [13] (https://github.com/modestyachts/CIFAR-10.1);
CIFAR-10.2 [13] (https://github.com/modestyachts/CIFAR-10.1);
CIFAR-10-$\bar{C}$ [11] (https://github.com/facebookresearch/augmentation-corruption) We use the corruption method provided in this link to create CIFAR-10.1-$\bar{C}$ and CIFAR-10.2-$\bar{C}$;

A.2. Experiment: Datasets Training Suitability

Follow the same training scheme as [9], we use ResNet-101 [6] architecture pre-trained on ImageNet [2]. The training epoch is 20, the batch size is 32 and the number of iterations per epoch is 2500. The optimizer is SGD with learning rate 1×10^{-2} and weight decay 5×10^{-4}.

A.3. Experiment: Datasets Testing Difficulty

CIFAR-10 setup. We use ResNet-44, RepVGG-A1, VGG-16-BN and MobileNet-V2 classifiers and their trained weights are publicly released by https://github.com/chenyaofo/pytorch-cifar-models.

ImageNet setup. We use EfficientNet-B1, DenseNet-121, Inception-V4 and ViT-Base-16 classifiers for correlation study. The pretrained models are provided by PyTorch Image Models (timm) [14].

A.4. Computation Resources

PyTorch version is 1.10.2+cu102 and timm version is 1.5. All experiments run on one 2080Ti and the CPU AMD Ryzen Threadripper 2950X 16-Core Processor.

B. Results of Datasets Training Suitability

Full results of correlation study on six test domains. We present a correlation study of BoP + JS divergence, Fréchet distance, maximum mean discrepancy and kernel inception distance. We use ResNet-101 as the feature extractor. The codebook size for BoP is 1000. All methods use the same feature. Based on their formulae, we use mean and covariance feature to compute them. On all domains, we see that BoP has consistent and superior performance.

Impact of codebook size on correlation strength for ResNet-101. In Fig. A-2, we find that BoP + JS divergence gives a relatively low correlation and converges to a high correlation when codebook size becomes larger.

Use Hellinger and Chi-squared to measure the distance between BoP representations. We test other distances than JS, such as Hellinger and Chi-squared under DomainNet setup with ResNet-101 and codebook size 1000. They yield similar results ($|\rho| = 0.928, 0.927$) as JS ($|\rho| = 0.929$). These results further validate the usefulness of BoP.

Use ResNet-34 for model accuracy and ResNet-101 to extract features. We use the extracted features to construct a codebook size 1000 and JS divergence to measure distances. BoP + JS divergence gives $|\rho| = 0.927$, and FD, MMD and KID gives $|\rho| = 0.909, 0.825, 0.899$, respectively. For Pearson's correlation, BoP + JS is still the highest ($|\rho| = 0.959$) compared to FD, MMD and KID ($|\rho| = 0.898, 0.823, 0.864$).

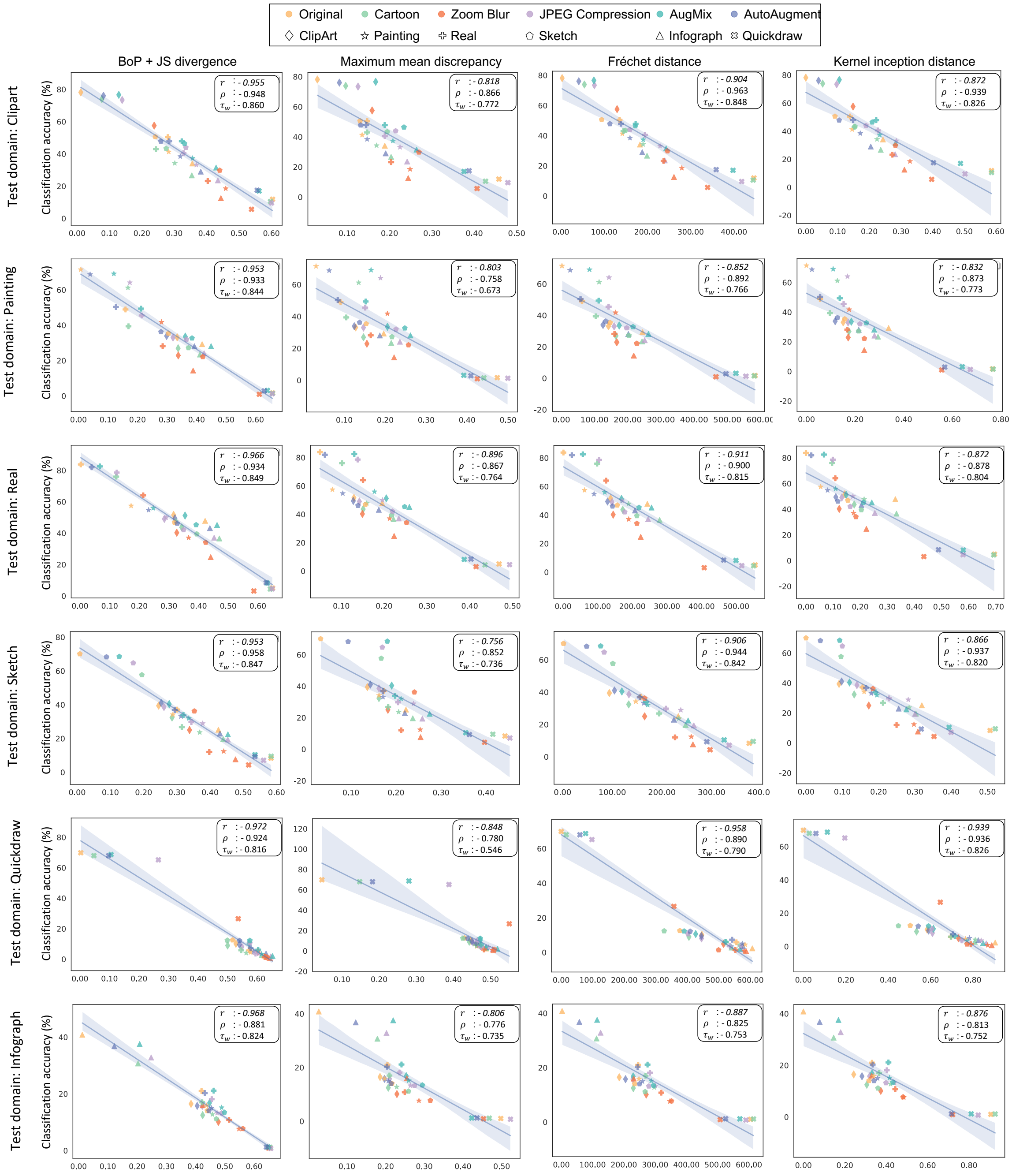

Figure A-1. **The full results of comparing training suitability of datasets.** Each point denotes a model. The models trained on transformed training sets are marked with shapes and each shape denote one specific transformation operation (*e.g.*, AutoAugment [1]). The straight lines are fit with robust linear regression [8].

C. Results of Datasets Testing Difficulty

Full results of correlation study on CIFAR-10.1-$\bar{C}$ is shown Fig. A-1. We see that BoP + JS divergence consistently well correlates with classification accuracy on six test domains: ClipArt, Painting, Real, Sketch, Quickdraw, Infograph with $|r| > 0.95$, $|\rho| > 0.88$ and $|\tau_w| > 0.81$.

The correlation study of BoP + JS, ATC, DoC and AC under CIFAR-10 and ImageNet setups for ResNet-44 and ViT-Base-16, respectively. We include results of (1) Prediction score ($\tau = 0.8$, $\tau = 0.9$), (2) Difference of confidence (DoC) [5], (3)Average thresholded confidence

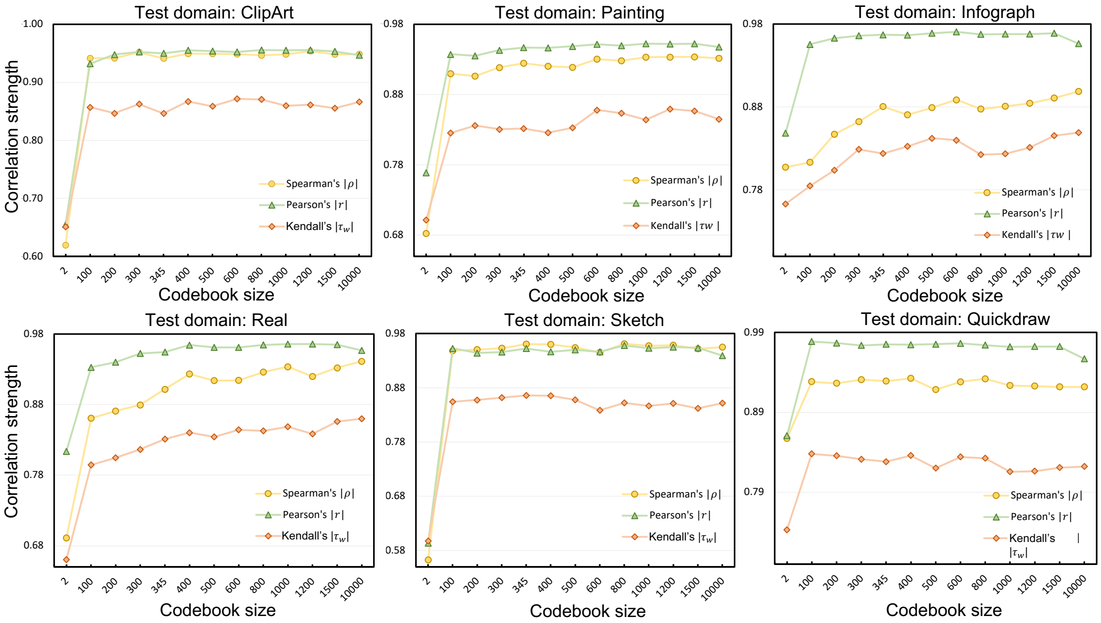

Figure A-2. **The impact of codebook size on correlation strength on six test domains: ClipArt, Painting, Infograph, Real, Sketch and Quickdraw.** We observe on six domains that BoP+JS gives low correlation with a small codebook and maintains stably high when codebook size becomes larger.

Method	CIFAR-10.1-$\bar{C}$					
	L1	L2	L3	L4	L5	Overall
$\tau = 0.8$	10.43	13.02	16.44	21.35	24.44	17.90
$\tau = 0.9$	4.00	5.26	8.36	11.90	13.98	9.49
ATC-MC [4]	3.47	4.63	7.64	11.15	12.85	8.73
DoC [5]	2.48	2.90	6.30	8.80	6.87	5.98
$\mu + \sigma + FD$ [3]	6.46	6.12	5.51	4.58	5.52	5.67
BoP ($K = 80$)	1.89	2.34	3.56	5.92	6.32	4.40
BoP ($K = 100$)	2.14	2.83	3.98	3.68	5.52	3.81

Table A-1. Method comparison in predicting classifier accuracy under CIFAR-10 setup. We report RMSE (%) on each severity level of CIFAR-10.1-$\bar{C}$.

with maximum confidence (ATC-MC) [4], Network regression ($\mu + \sigma + FD$) [3] and BoP with codebook size 80 or 100 on CIFAR-10.1-$\bar{C}$ in Table A-1. We see that BoP is overall more predictive of model testing difficulty compared with other methods. We also present the correlation study of BoP + JS, ATC-MC, DoC and average confidence (AC) under CIFAR-10 setup and ImageNet setup in Fig. A-4.

Results on datasets with natural distribution shifts for Inception-V4. In addition to datasets with synthetic shifts (*i.e.* ImageNet-C), we explore the effectiveness of BoP on datasets with natural shifts. Under ImageNet setup, we include results of ImageNet-V2-A/B/C and ImageNet-Sketch for Inception-V4. From Fig. A-3, we observe that four datasets (red dots) still lie on the trend of ImageNet-C (grey dots). This means that BoP still effectively captures the distributional shift of four real-world datasets.

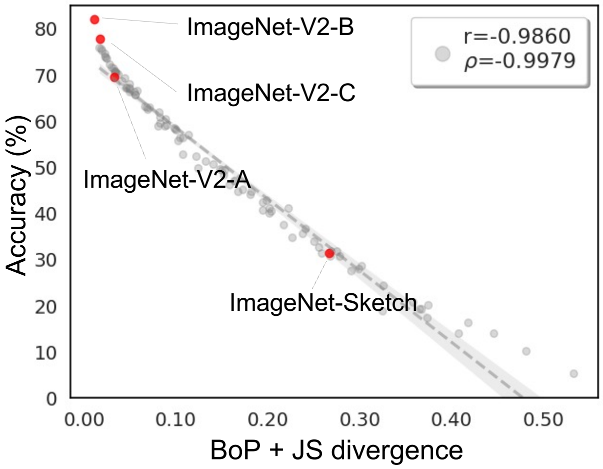

Figure A-3. **Results on four datasets with natural shifts for Inception-V4: ImageNet-V2-A/B/C and ImageNet-Sketch.** We find that four new datasets (red dots) still lie on the original trend of ImageNet-C (grey datasets).

DomainNet setup for test set difficulty. We also use 36 domains from DomainNet setup for datasets testing dif-

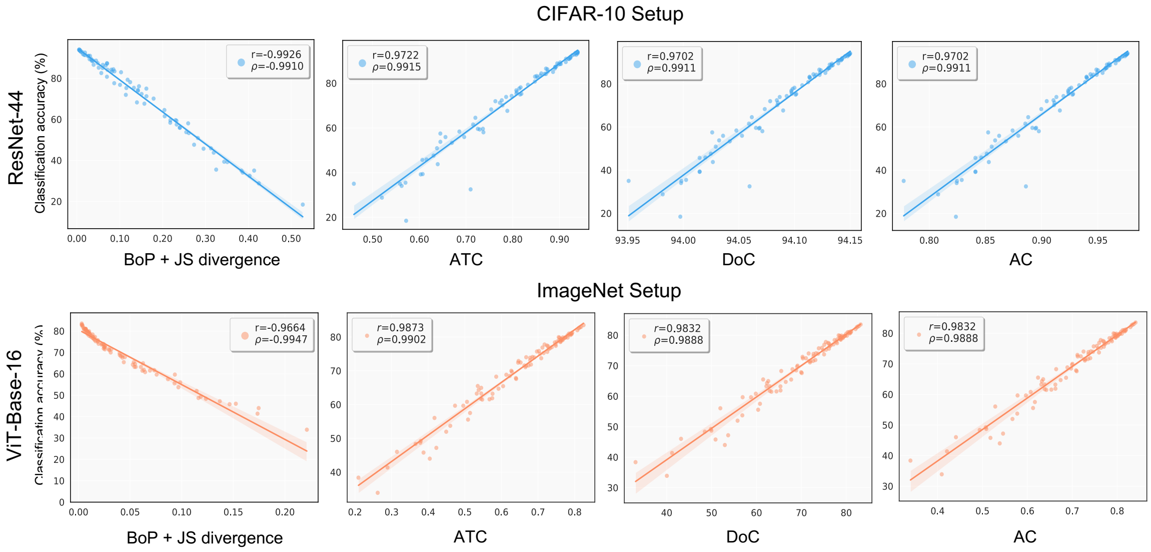

Figure A-4. **The correlation study of BoP + JS, ATC, DoC and AC under CIFAR-10 and ImageNet setups. Top:** Correlation study under CIFAR-10 setup using ResNet-44. In each figure, a point denotes a dataset from CIFAR-10-C benchmark. **Bottom:** Correlation study under ImageNet setup using ViT-Base-16. In each figure, a point denotes a dataset from ImageNet-C benchmark. The straight lines are fit with robust linear regression [8]. We observe that BoP+JS gives a higher Spearman's correlation on both setups.

ficulty. We use ResNet-101 trained on 'Real' domain to extract features and construct a codebook of size 1000. We conduct a correlation study between BoPs of the rest 35 domains and model accuracy. We compare BoP + JS with ATC, DoC and AC. We find that BoP + JS shows the strongest correlation ($|r| = 0.959, |\rho| = 0.972$) while the second best ATC has $|r| = 0.957, |\rho| = 0.924$.

Analysis of the sensitivity of BoP to dataset size. Datasets in DomainNet also vary widely in size (from $48k$ to $173k$) where BoP works well on both dataset-level applications. We further studied the impact of dataset size on DomainNet in dataset testing difficulty by randomly sampling 1% to 10% of data from each test set. Correlation of all methods decreases but BoP remains the best: $|\rho|$ is $0.909, 0.857, 0.849$ (decrease from $0.973, 0.914, 0.925$) for BoP, DoC and ATC, respectively.

References

[1] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[4] Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

[5] 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 and Pattern Recognition*, pages 1134–1144, 2021.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019.

[8] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[9] Junguang Jiang, Baixu Chen, Bo Fu, and Mingsheng Long. Transfer-learning-library. https://github.com/thuml/Transfer-Learning-Library, 2020.

[10] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[11] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems*, 2021.

[12] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

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

[14] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.