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

Weijie Tu¹ Weijian Deng¹ Tom Gedeon² Liang Zheng¹
¹Australian National University ²Curtin 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-  and CIFAR-10 setups.

A. Experimental Setup

A.1. Datasets


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 $1 \times 10^{-2}$ and weight decay $5 \times 10^{-4}$.

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 $(|\rho| = 0.898, 0.823, 0.864)$. 
C. Results of Datasets Testing Difficulty

Full results of correlation study on CIFAR-10.1-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 $|\tau| > 0.81$.

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.
Correlation strength

-0.60
-0.70
-0.80
-0.90
-1.00
-0.68
-0.78
-0.88
-0.98

accuracy (%)

i.e.

shifts (for Inception-V4. In addition to datasets with synthetic
under CIFAR-10 setup and ImageNet setup in Fig. A-4.

of BoP + JS, ATC-MC, DoC and average confidence (AC)
with other methods. We also present the correlation study
overall more predictive of model testing difficulty compared
100
or

codebook size becomes larger.

100

level of CIFAR-10.1-

under CIFAR-10 setup. We report RMSE (%) on each severity
Table A-1. Method comparison in predicting classifier accuracy
with maximum confidence (ATC-MC) [4], Network regression ($\mu + \sigma + FD$) [3] and BoP with codebook size 80 or 100 on CIFAR-10.1-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 original trend of ImageNet-C (grey 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-

with maximum confidence (ATC-MC) [4], Network regression ($\mu + \sigma + FD$) [3] and BoP with codebook size 80 or 100 on CIFAR-10.1-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 original trend of ImageNet-C (grey datasets). This means that BoP still effectively captures the distributional shift of four real-world datasets.

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

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR-10.1-C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
</tr>
<tr>
<td>$\tau = 0.8$</td>
<td>10.43</td>
</tr>
<tr>
<td>$\tau = 0.9$</td>
<td>4.00</td>
</tr>
<tr>
<td>DoC [5]</td>
<td>2.48</td>
</tr>
<tr>
<td>$\mu + \sigma + FD$ [3]</td>
<td>6.46</td>
</tr>
<tr>
<td>BoP ($K = 80$)</td>
<td>1.89</td>
</tr>
<tr>
<td>BoP ($K = 100$)</td>
<td>2.14</td>
</tr>
</tbody>
</table>

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

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.

References


