Joint Representation Learning and Novel Category Discovery on Single- and Multi-modal Data –Supplementary Material–

Xuhui Jia[1] Kai Han[1,2,3*] Yukun Zhu[1] Bradley Green[1]
[1]Google [2]University of Bristol [3]The University of Hong Kong
{xhjia, kaihanx, yukun, brg}@google.com

1. Effects of different WTA hyperparameters on test set

To further validate the effectiveness of the WTA hyperparameters tuning method (described in section 4.4 of the main paper) under our setting, we conduct experiments on real test set. More specifically, we first take the empirical WTA window size $k = 4$ in [3] to run experiment with different threshold μ and report the results in table 1. We find that conclusion is generally consistent with what we observe in "validation set" (i.e., the subset in the labelled data that is pretended to be unlabelled), e.g., the performance is generally stable for μ greater than 200. For kinetics-400, the best performance (55.2) is achieved when $\mu = 200$ in "validation set", whereas the best number (56.5) is observed when $\mu = 240$ in test set. For VGG-Sound, the best performance (51.3) is obtained when $\mu = 240$ in "validation set", whereas the best number (50.2) is observed when $\mu = 200$ in test set. Given that the performance difference between $\mu = 240$ and $\mu = 200$ is small, and they are neighbouring values in the sweeping set, the hyperparemeter tuning method described in section 4.4 of the main paper appears to be an effective method. We choose $\mu = 240$ for both datasets according to the "validation set" to slightly favor the performance on VGG-Sound, as the performance on VGG-Sound is generally worse than that on kinetics-400.

Table 1: **Performance of different WTA threshold.**

Dataset	130	180	200	240	260	300
Kinetcis-400	22.7	40.8	55.3	56.5	56.2	55.8
VGG-Sound	21.2	44.2	50.2	50.0	49.4	49.3

Similarly, given $\mu = 240$, we sweep different k and report the results in table 2. We also find that the $k = 4$ and $k = 8$ perform comparably well, and they are both better than $k = 2$ and $k = 16$. The conclusion remains the same as what we find on "validation set".

Table 2: **Performance of different WTA window size.**

Dataset	$k = 2$	$k = 4$	$k = 8$	$k = 16$
Kinetcis-400	53.4	56.5	56.1	51.2
VGG-Sound	49.2	50.0	51.1	49.7

2. Unknown class number in unlabelled data

Following Han et al. [1], we assume the number of the classes, C^u , in the unlabelled data is known a-priori. When C^u is not known, we can use the method introduced in DTC Han et al. [2] to estimate C^u first, and then substitute the estimated number into our framework. We evaluate the performance of our approach on ImageNet using the unknown category numbers estimated by DTC. The estimates are 34/32/31 and the ground-truth numbers are 30/30/30 on the three unlabelled subsets. The average accuracy over three subsets is 84.1% which outperforms Han et al. [1] by 3.6%.

3. Unsupervised clustering

We further experiment with our approach for pure unsupervised clustering on the unlabelled subset of CIFAR10 and CIFAR100, which contains 5 and 20 classes respectively, by simply dropping the labelled data. Our method achieves 84.6% and 61.5% on the two datasets respectively, while the results by k-means baseline (using features extracted by the model trained on the labelled subset) are 65.5% and 56.6% respectively (see table 1 in the main paper), showing the superiority of our approach. This reveals that our method is also an effective clustering method.

* Corresponding author.

References

[1] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020. 1

[2] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019. 1

[3] Jay Yagnik, Dennis Strelow, David A. Ross, and Ruei sung Lin. The power of comparative reasoning. In *ICCV*, 2011. 1