Semantic Guided Latent Parts Embedding for Few-Shot Learning
Supplementary Material

Fengyuan Yang^{1,2}, Ruiping Wang^{1,2,3}, Xilin Chen^{1,2}
^{1}Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
^{2}University of Chinese Academy of Sciences, Beijing, 100049, China
^{3}Beijing Academy of Artificial Intelligence, Beijing, 100084, China
fengyuan.yang@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

In this supplementary material, we provide more details that we mentioned but did not have enough space in the main paper. In the following sections, we give the datasets summary and more training details in Section 1, we show more visualization results in Section 2, and provide an additional discussion in Section 3.

1. Datasets and More Training Details

Table 1 shows a summary of datasets we used in the experiments of the main paper. They are miniImageNet [5], tieredImageNet [4], CIFAR-FS [1], and CUB [6]), all commonly used benchmarks for few-shot learning.

Table 1: Statistics of four few-shot learning datasets.

DataSet	Train/Val/Test	Instances	Resolution
miniImageNet	64 / 16 / 20	60,000	84 × 84
tieredImageNet	351 / 97 / 160	779,165	84 × 84
CUB	100 / 50 / 50	11,788	84 × 84
CIFAR-FS	64 / 16 / 20	60,000	32 × 32

In addition to implementation details noted in §4.1 of the main paper, here we provide more details. As mentioned in the main paper, we train our module end-to-end using the stochastic gradient descent optimizer. The initial learning rate is set to 0.1 and will be decayed by 0.05 at epochs 60 and 70. And we set $\lambda = 2$ for miniImageNet and CIFAR-FS, $\lambda = 1.5$ for CUB, and $\lambda = 1$ for tieredImageNet. During the test, we obtain the FSL accuracy by 2000 testing episodes and report the average classification accuracies with 95% confidence intervals. The best model is chosen based on the performance on the validation set.

2. More Visualization Results

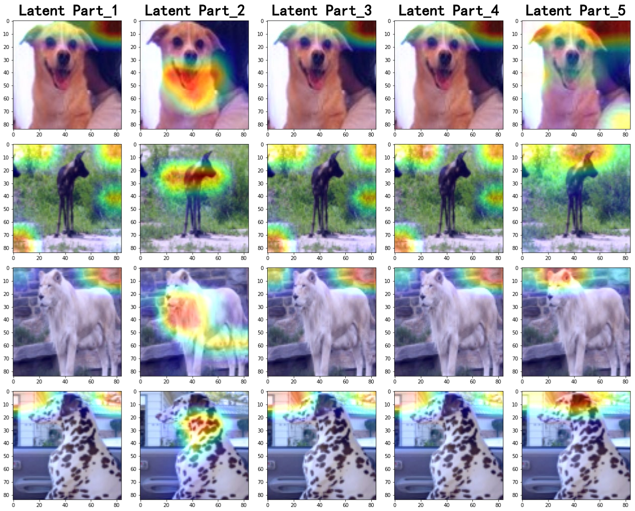

Figure 1: Visualization results of the activation regions of P (=5) latent parts on novel classes of miniImageNet. The redder region means a higher activation value.

In the main paper, we show the visualization results of the latent parts activation on CUB dataset in Fig.6 and it demonstrates that our model tends to discover the same semantic part for different birds. Here we show the results of more coarse-domain classes like some different animal classes from miniImageNet. As shown in Fig.1, the same phenomenon as the main paper can be observed that for each column the similar part of different animals tends to be activated (e.g., neck part in the 2nd column and head part in the 5th column). This again shows that our model indeed learns some shared latent parts across different categories so as to have the potential to go beyond class recognition to fulfill part-based understanding of the novel class.

Fig.2 shows two correlation matrices calculated based on latent parts embeddings of 20 novel classes of miniImageNet. Each element $m_{i,j}$ in these correlation matrices is the cosine similarity between the p-th latent parts embedding of class i (i.e., LPE_p^i) and the p-th latent parts embedding of class j (i.e., LPE_p^j). These two correlation matrices

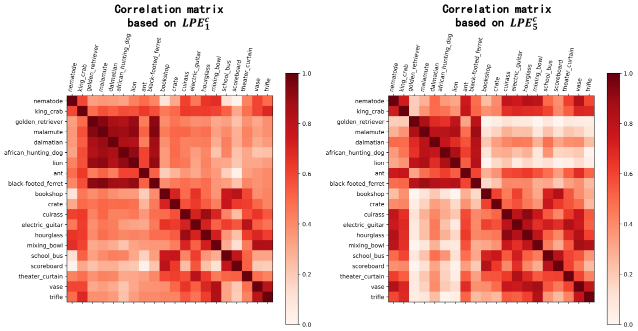

Figure 2: Correlation matrices based on the 1st and 5th latent part embeddings of 20 novel classes of miniImageNet. The redder block means a higher similarity value.

in Fig.2 correspond to $p = 1$ and $p = 5$ respectively. As we can see, the two correlation matrices look different which means under our framework the similarity between novel classes is different with respect to different latent parts. This is reasonable since the similarity between classes is at part level and different classes share different similar parts.

Fig.3 shows that part-level features of the same classes are close to each other, while those of different classes tend to be far away. It again shows our method's effectiveness.

3. Extra Discussion

In this section, we provide more discussion about our method as follows.

1) Why use semantic knowledge in FSL? Is it fair?

As we have discussed in the introduction section of our main paper, images from the real world often contain multiple objects of interest. Without taking the semantic label into consideration, even we humans cannot exactly tell what this novel category exactly is especially in the few-shot scenario. Even if there is only one object of interest, we can never know the exact granularity of this class (e.g., 'Fennec fox' or 'Fox' or 'Canine' or 'Animal') from very few samples either. However, by leveraging its semantic knowledge, the meaning of the novel class can be more clear.

Using semantic knowledge as the additional information is of course unfair compared with those unimodal FSL works. But as we have claimed above, semantic knowledge is indispensable in FSL otherwise the definition of the novel class will be ambiguous. In addition, we also compare with other semantic using FSL works and show our advancement in the main paper.

2) Why use CLIP and is it fair?

As we claimed in the main paper, we only use the semantic encoder of CLIP to help train our visual space from scratch. We never use the visual encoder of CLIP. In another word, here CLIP semantic embedding is just another semantic embedding like the widely used GloVe word embedding in many semantic using FSL works [2, 7–9]. The reason we choose to use it is that CLIP was trained to align the visual space and the semantic space, so it is a more visualized semantic source compared with GloVe. And in experiments, we indeed verify that it is better than GloVe.

It is also worth noting that CLIP semantic embedding is not invincible, as shown in Tab.2 in the main paper the performance of using attributes annotation is way higher than using CLIP semantic on CUB dataset. It shows that in the more fine-grained scenario, CLIP semantic is not good enough compared with more customized semantic sources (e.g., attributes annotation).

3) Why latent parts, not explicit parts?

Because we want to design a more general FSL model instead of just a customized model that can only work on a specific dataset. Thus, a large variety of categories will inevitably cause large diversity of parts, and most importantly we can never pre-define the parts for all novel classes. But still, we can do better with semantics and parts which we will discuss in the next point.

4) What is the motivation and insight of Part-level prior transfer?

We detailedly described our part-level prior transfer mechanism in Sec.3.3 in the main paper, here we give more interpretation. Firstly, the motivation of this module is clear which is to explicitly transfer visual prior from base classes to novel classes. Most importantly, we transfer the prior at the part level which makes more sense because the similarity between two categories is mostly at the part level. This more human-like transfer mechanism is the reason why there is a big drop in FSL performance if not performing the transfer.

In addition, the prior transfer mechanism is also commonly-used in FSL methods like Dynamic-FSL [3]. But we make the first attempt to perform it at the part-level which makes more sense and indeed shows its effectiveness.

5) What can we do better with semantics and parts?

The parts that our model discovers for each class in this work are still latent parts and they are not yet well-aligned with real semantic parts. For future work, although we cannot pre-define parts for every single novel class, we can predefine some general parts for different groups of classes. And by using the semantic knowledge of the novel class, we can predict what group this novel class belongs to and followed by the discovery of the semantic parts. This is a more explainable pipeline for part-based few-shot class understanding.

6) Our contributions.

(a) To the best of our knowledge, we are the first to leverage class semantic knowledge to represent a class as the combination of its several parts. (b) We propose a novel and effective method to extract part-level embeddings and further propose a novel mechanism for part-level prior transfer. (c) We verified that semantic knowledge is effective and the more visualized and customized semantic source is

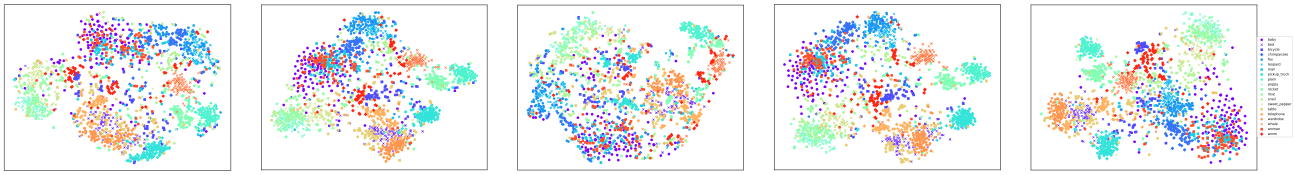

Figure 3: Five t-SNE visualizations on the CIFAR-FS test set, each corresponding to one of the $P = 5$ latent part embeddings.

more useful in FSL. (d) Our method has the potential for real semantic parts discovery in FSL which is a vital step from class-level object recognition to part-level object understanding.

References

[1] L Bertinetto, J Henriques, PHS Torr, and A Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR)*, 2019.

[2] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing (TIP)*, 28(9):4594–4605, 2019.

[3] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4367–4375, 2018.

[4] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2018.

[5] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3630–3638, 2016.

[6] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011.

[7] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. TAFE-Net: Task-aware feature embeddings for low shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1831–1840, 2019.

[8] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4847–4857, 2019.

[9] Kun Yan, Zied Bouraoui, Ping Wang, Shoaib Jameel, and Steven Schockaert. Aligning visual prototypes with bert embeddings for few-shot learning. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR)*, pages 367–375, 2021.