SEGA: Semantic Guided Attention on Visual Prototype 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 experiment details for which we mentioned but did not have enough space in the main paper. In the following sections, we show detailed Canonical Correlation Analysis (CCA) results of visual and semantic space in Section 1, we give the datasets summary and training details in Section 2, we provide the results using different word embedding in Section 3, we show more prototypes t-SNE visualization in Section 4, we conduct extra comparison with relevant methods in Section 5, we further explain why using Hadamard product in Section 6, and our advantage on computation complexity in Section 7 followed by an extra discussion in Section 8.

1. CCA Results

As mentioned in §3.1 of the main paper, we apply Canonical Correlation Analysis (CCA), which is trained on base classes, to the visual features and semantic word embeddings of novel classes. More specifically, visual features of base classes are extracted by the "ResNet-12" backbone after being trained in the first training stage on miniImageNet [10], while the corresponding word embedding is derived from the GloVe [6] pre-training model.

The detailed CCA results are shown in Table 1, where the top rows show there is still a relatively high correlation between visual and semantic space on novel classes, and the bottom rows show when using the non-corresponding visual and semantic data to train the CCA model the correlation coefficient will be quite small which means the alignment is not transferable anymore.

Actually, this same phenomenon can be also found in experiments with varying backbones, datasets, and word embedding models, which gives us solid evidence that semantic knowledge can really help FSL tasks.

2. Datasets and More Training Details

In Table 2, we give a summary of datasets we used in the experiments of the main paper (miniImageNet [10], tieredImageNet [7], CIFAR-FS [1], and CUB [11]). They are all common benchmarks for few-shot learning.

Table 1: CCA results on miniImageNet. We train the CCA on 64 base classes which is to find a mapping to maximize the correlation of their visual space and semantic space. This table shows the correlation coefficient calculated by applying this trained CCA model on 16 validation classes and 20 test classes respectively. For comparison, the bottom rows show results when using non-corresponding visual and semantic space to train the CCA mapping. "Corr." denotes the correlation coefficient.

CCA training		CCA test	Corr.
Base classes-visual feature	Corresponding-word embedding	Val cls.	**0.58**
		Test cls.	**0.79**
Base classes-visual feature	**Non**Corresponding-word embedding	Val cls.	0.19
		Test cls.	0.38

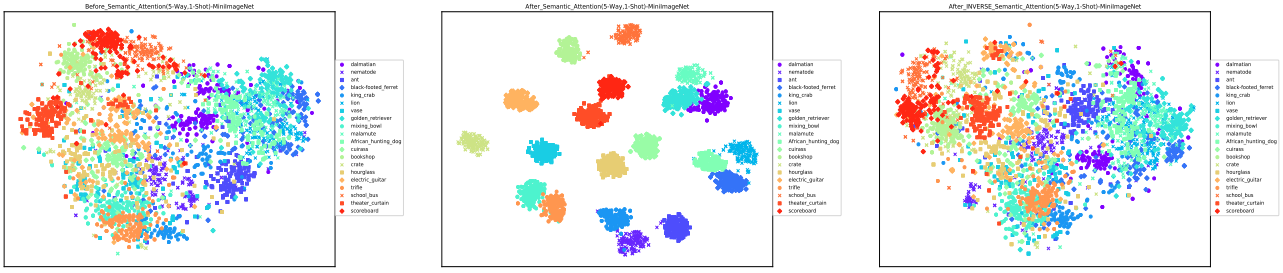
Table 2: Statistics of four few-shot learning datasets.

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. We adopt an empirical learning rate scheduler following the practice of [3, 4, 12]. In the first training stage which is to train *Feature Extractor*, the learning rate is initially set to 0.1 and then changed to 0.006, 0.0012, and 0.00024 at epochs 20, 40, and 50, respectively. Note that the learning rate milestone for tieredImageNet is 30, 60, and 75 since we train 90 epochs on this larger dataset. As for the second training stage, we initially set the learning rate to 1.0, and then change it to 0.5, 0.1, and 0.05 at epochs 5, 10, and 15, respectively. Although we need to train different *Semantic Guided Attention Weight Generator* and *cosine classi-*

Table 3: Results of **different word embedding models** on miniImageNet, tieredImageNet, and CIFAR-FS. We report the average classification accuracies (%) on 5000 test episodes of novel categories (with 95% confidence intervals). "FAKE" means using the non-corresponding label as semantic guidance.

Sem.	miniImageNet		tieredImageNet		CIFAR-FS	
	5Way 1Shot	10Way 1Shot	5Way 1Shot	10Way 1Shot	5Way 1Shot	10Way 1Shot
GloVe **(ours)**	**69.04**±0.26	**52.71**±0.15	**72.18**±0.30	56.82±0.21	**76.24**±0.25	**61.77**±0.17
Word2Vec	68.10±0.26	51.12±0.15	71.85±0.30	**56.96**±0.21	75.25±0.26	60.02±0.18
NO	62.81±0.27	46.73±0.17	68.55±0.31	54.01±0.21	67.78±0.30	53.32±0.21
FAKE	59.04±0.27	43.58±0.16	64.64±0.31	50.07±0.21	63.27±0.29	48.66±0.19



(a) Before semantic attention (b) After semantic attention (c) After **inverse** semantic attention

Figure 1: t-SNE visualization testing on **miniImageNet** under 5-Way 1-Shot scenario. **(a)** and **(b)** are prototypes before ($\mathbf{p}^c$) and after ($\boldsymbol{a}_c \otimes \mathbf{p}^c$) performing the semantic attention. **(c)** shows the result when applying the inverse attention ($(1-\boldsymbol{a}_c) \otimes \mathbf{p}^c$). The setting is in the same manner as Figure 4 of the main paper and the point color represents its category.

fier for different N-Way K-Shot meta-testing scenario, the whole procedure is fast and efficient since *Feature Extractor* is fixed during the second training stage. The best model is chosen based on the accuracy on the validation set.

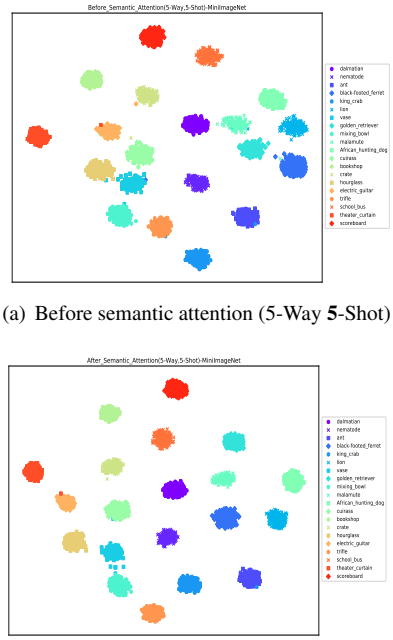
3. Different Word Embedding Models

Note that apart from GloVe [6], there exist some other word embedding models such as Word2Vec [5]. Word2Vec is also trained on large corpus of text and the embedding has 300-dimension which is the same size as GloVe. In this section, we adopt Word2Vec as the semantic knowledge source to verify the generalization ability of our proposed method.

Table 3 shows the results when leveraging different word embedding models. The rows "GloVe **(ours)**", "NO" and "FAKE" exactly are the same results reported in Table 1 of the main paper. As we can see, there is significant performance gain whatever semantic knowledge source is Word2Vec or GloVe (GloVe is slightly better than Word2Vec and it is possibly because that GloVe is a count-based model and captures more global co-occurrence information than Word2Vec). They are both better than "No Semantic" and way better than "Fake Semantic" which uses the GloVe embeddings of fake labels to generate semantic attention. The above phenomenon again verifies the generalization and effectiveness of our proposed method.

4. More t-SNE Visualization Results

In the main paper, we show the t-SNE visualization in Figure 4 which is the results for CIFAR-FS. Here we give the results for miniImageNet in Figure 1 where the same phenomenon can be observed that our SEGA does capture



(a) Before semantic attention (5-Way **5**-Shot)

(b) After semantic attention (5-Way **5**-Shot)

Figure 2: t-SNE visualization of the prototypes in visual space in 5-Way **5**-Shot scenario. **(a)** and **(b)** are prototypes before ($\mathbf{p}^c$) and after ($\boldsymbol{a}_c \otimes \mathbf{p}^c$) performing the semantic attention. The setting is in the same manner as Figure 4 of the main paper and the point color represents its category.

Table 4: Extra comparison results of SEGA and AM3 on miniImageNet, tieredImageNet, and CIFAR-FS. We report the average classification accuracies (%) on 5000 test episodes of novel categories (with 95% confidence intervals).

Sem.	miniImageNet		tieredImageNet		CIFAR-FS	
	5Way 1Shot	10Way 1Shot	5Way 1Shot	10Way 1Shot	5Way 1Shot	10Way 1Shot
SEGA **(ours)**	**69.04**±0.26	**52.71**±0.15	**72.18**±0.30	**56.82**±0.21	**76.24**±0.25	**61.77**±0.17
AM3 in our framework	64.29±0.24	50.85±0.16	70.07±0.30	55.41±0.20	70.34±0.26	58.32±0.17

the class-specific discriminative dimensions.

Furthermore, we also visualize the 5-Way 5-Shot scenario to show why we can only get marginal improvement when the number of shots becomes larger as noted in §4.4 in the main paper. As we can see, the prototypes are already quite stable when given 5 labeled samples per novel class in Figure 2(a) compared to the 1-Shot scenario shown in Figure 1(a). Although the prototypes become more stable and discriminative after performing the semantic attention as shown in Figure 2(b), the improvement of classification accuracy is not as significant as in the 1-Shot scenario. The reason is that when given more samples, the key feature dimensions can be learned and concluded much better which means the prototypes are getting more precise. Therefore, the key feature attention generated by semantic knowledge is not as vital as in the 1-Shot scenario anymore.

5. Extra Comparison with AM3

In §4.4 of the main paper, we already show our advantage over other semantic using methods including AM3 [13], TriNet [2], and MultiSem [8]. Here we even adapt AM3 to our framework for further fair comparison. The origin AM3 is not compatible with our framework since the performance drops dramatically (even no better than baseline) after directly replacing weight in Eq4 with the semantic prototype of AM3. So we adapt our classifier to Euclidean space since the original AM3 uses Euclidean distance. As shown in Table 4, its results are still inferior to SEGA, which further proves our advantage over AM3.

6. Why Hadamard Product for Attention

Hadamard product is chosen by our motivation which is using semantic to guide visual perception about which key features should be focused on. It is the most direct and suitable way to do feature selection and reflects method name SEGA. But we still try element-wise addition and concatenation on miniImageNet to prove they are not better than Hadamard Product:

Different semantic usages	5Way-1Shot	10Way-1Shot
Element-wise addition	67.18±0.24	49.58±0.15
Concatenation attention	67.92±0.25	49.77±0.14
Hadamard (**ours results in** Table1)	**69.04**±0.26	**52.71**±0.15

7. Computation Complexity

SEGA comes with almost no more computation. The extra time cost comes from the attention generation model which is a simple MLP that can be ignored compared with the embedding model.

The economical time cost is actually our advantage over other SOTAs. DeepEMD [14], RFS [9], and SEGA all have 2 training stages. The first stage is common standard embedding model training, so they differ mainly in the second stage where SEGA only costs ∼1hour [0.2second/training-episode] while DeepEMD (solving costly Quadratic Programs) costs >9hours [30second/training-episode] and RFS (self-distillation) costs ∼2hours [0.2second/training-episode] for 5Way-1Shot miniImageNet under same server and GPU.

8. Extra Discussion

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

1) What is the insight of this work?

Semantic guided attention is actually a kind of feature selection and can highlight the key class-specific features and minimize the impact of background noise and large intra-class variation, which is rather important in the few-shot scenario since the key challenge for FSL is incomprehensive cognition of novel category.

2) Is it necessary to use semantic knowledge?

As we have shown in Figure 2 of the main paper, real-world images often contain multiple objects of interest. If not take the semantic label into consideration, even we human beings may get confused about what this novel category exactly is especially in the circumstances where the labeled samples are in dire poverty. However, compared to the large scale of image annotation, semantic information is always easier to be obtained. By leveraging semantic knowledge, the meaning of the novel category can be more clear. Besides, since semantic knowledge is indispensable in the most relevant domain zero-shot learning, we argue that when shot numbers are changed from zero to few, semantic knowledge should also be very helpful.

3) Why does SEGA work? What does it suggest?

SEGA learns the mapping from semantic space to visual space to generate attention over feature dimensions. Each feature dimension can be regarded as an attribute thus the

attention is the class-specific importance of each attribute when distinguishing this category from others. This mapping is transferable since similar kinds of objects always share similar importance of attributes. In addition, the mapping is learnable because this mapping is much easier to be learned than directly reconstructing the prototype from only the word embeddings of class label. Furthermore, SEGA suggests that finding out the key features accurately and ignoring the misleading noise is rather important in the few-shot scenario.

4) Do we use transductive setting or train-val setting?

No. All of our experiments follow the commonly used inductive setting, and we just use validation datasets for model selection.

References

[1] L Bertinetto, J Henriques, PHS Torr, and A Vedaldi. Meta-learning with differentiable closed-form solvers. 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] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10657–10665, 2019.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[7] 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.

[8] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Baby steps towards few-shot learning with multiple semantics. *arXiv preprint arXiv:1906.01905*, 2019.

[9] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, pages 266–282. Springer, 2020.

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

[11] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[12] Zeyuan Wang, Yifan Zhao, Jia Li, and Yonghong Tian. Cooperative bi-path metric for few-shot learning. In *ACM International Conference on Multimedia (ACMMM)*, pages 1524–1532, 2020.

[13] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:4847–4857, 2019.

[14] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12203–12213, 2020.