Shaping Visual Representations with Attributes for Few-Shot Recognition

Haoxing Chen, Huaxiong Li, Yaohui Li, Chunlin Chen

Abstract—Few-shot recognition aims to recognize novel categories under low-data regimes. Some recent few-shot recognition methods introduce auxiliary semantic modality, i.e., category attribute information, into representation learning, which enhances the feature discrimination and improves the recognition performance. Most of these existing methods only consider the attribute information of support set while ignoring the query set, resulting in a potential loss of performance. In this letter, we propose a novel attribute-shaped learning (ASL) framework, which can jointly perform query attributes generation and discriminative visual representation learning for few-shot recognition. Specifically, a visual-attribute predictor (VAP) is constructed to predict the attributes of queries. By leveraging the attributes information, an attribute-visual attention module (AVAM) is designed, which can adaptively utilize attributes and visual representations to learn more discriminative features. Under the guidance of attribute modality, our method can learn enhanced semantic-aware representation for classification. Experiments demonstrate that our method can achieve competitive results on CUB and SUN benchmarks. Our source code is available at: https://github.com/chenhaoxing/ASL

Index Terms—Attribute-shaped learning, few-shot learning, attribute-visual attention

I. INTRODUCTION

Deep learning has achieved outstanding performance in many visual tasks [1], [2], [3], [4]. However, training deep models often requires a lot of labeled data, which is not always accessible in real applications [5], [6], [7], [8]. Inspired by the ability of humans, few-shot learning aims to recognize new objects with few labeled training samples.

Recently, various successful few-shot recognition methods have been proposed, which can be roughly divided into three categories: meta-learning methods [9], [10], data augmentation methods [11] and metric-learning based methods [12], [13], [14]. In these methods, metric-learning based methods have attracted extensive attention due to their simplicity and effectiveness. Most of them mainly focus on enhancing the informativeness and discriminability of learned semantic visual representations to improve image recognition accuracy. However, these methods perform image classification in the context of uni-modal visual learning, i.e., only using the images. In the real-world, humans learn new concepts by leveraging multi-modal information rather than a single one [15], [16]. For instance, one can learn about the red-billed blue magpie by some images and attributes, such as red bill and red nape, instead of only seeing many pictures. In other words, attribute information can help humans learn new visual objects [17], [18], [19], [20].

To imitate this ability, some works introduce auxiliary semantic modalities to enhance feature learning in few-shot learning. Pavel et al. [17] took attribute features as supplementary information and enhanced the representation learning ability by adding regularization terms. AM3 [21] utilized convex combination to adaptively mix the semantic structures of visual representations and label semantics. Dual TriNet [22] is an auto-encoder network, which directly synthesizes instance features by leveraging semantics.

Although the above methods [17], [21], [22] have achieved impressive performance by leveraging semantic information, they only learn the feature representations of support set with the help of attribute information and ignore the query set that lacks the attribute information. To the best of our knowledge, there is no specific mechanism to explore the underlying attribute information of query samples, which is hopeful to improve the few-shot recognition performance.

Towards this end, we propose a novel attribute-shaped learning (ASL) model that generates corresponding attributes through image features and learns more discriminative visual features combined with attributes. Specifically, since query images have no auxiliary attribute information, a visual-attribute predictor (VAP) is proposed to generate attribute features. Then, we propose an attribute-visual attention module (AVAM), which can adaptively utilize attributes and visual representations to learn more discriminative features. AVAM contains two sub-modules, i.e., channel attention module (CAM) and pyramid spatial attention module (PSAM). CAM finds important channels, and PSAM finds multi-scale spatial-visual representations. Fig. 1 shows the effectiveness of our ASL method and classical ProtoNets [12]. It can be observed that, by incorporating the generated attribute information, our method can focus on more discriminative local regions for vi-
Pyramid spatial attention aims to generate the prototype for each category and determine the learning based framework. Since ProtoNets \cite{snell2017prototypical} is simple and learning on training set $\mathcal{D}_{\text{train}}$, and then compute prototype $\hat{p}_n$ for support class $n$:

$$p^n = \frac{1}{M} \sum_{i=1}^{M} \mathcal{F}_{\theta}(s_i^n),$$  

where $s_i^n$ is the $i$-th image of support category $n$. For each query image $(q_j, y_j)$, ProtoNets calculates the Euclidean distance between $q$ and each prototype, and use softmax function to obtain the probability distribution:

$$p(\hat{y}_j = n|q_j) = \frac{\exp(-||\mathcal{F}_{\theta}(q_j) - p^n||^2)}{\sum_{k=1}^{N} \exp(-||\mathcal{F}_{\theta}(q_j), p^k||^2)},$$  

$\mathcal{F}_{\theta}$ is trained by minimizing the classification loss:

$$\mathcal{L}_{\text{cls}} = -\sum_{j=1}^{Q} \log p(\hat{y}_j = y_j|q_j).$$

\section{II. METHODOLOGY}

In few-shot learning, the model is trained by a series of $N$-way $M$-shot episodes and each episode can be seen as an independent task. Each task $\tau$ is formed by randomly selecting $N$ categories from training set $\mathcal{D}_{\text{train}}$, and then sampling support set $\mathcal{S} = \{(s_j, a_j, y_j)\}_{j=1}^{N\times M}$ and query set $\mathcal{Q} = \{(q_j, a_j, y_j)\}_{j=1}^{Q}$ from these $N$ categories. Here, $x$, $a$, and $y$ represent the image, attribute vector and label respectively. Note that $\mathcal{Q}$ contains different examples from the same $N$ categories and $A$ is determined by the dataset. After learning on training set $\mathcal{D}_{\text{train}}$, a model is evaluated on a test set $\mathcal{D}_{\text{test}}$.

The proposed method is applicable to nearly any metric-learning based framework. Since ProtoNets \cite{snell2017prototypical} is simple and effective, we choose it as our baseline framework. ProtoNets aim to generate the prototype for each category and determine the category of query samples by calculating the distances to each prototype. Specifically, ProtoNets firstly obtain the feature representations of each image through a feature extractor. Then, we evaluate the proposed framework and compare it with existing methods.

\subsection{A. Visual-Attribute Predictor}

To enable both the query images and support categories to have additional attribute descriptions, we generate attribute descriptions for each sample. Specifically, given an image $x$, through $\mathcal{F}_{\theta}$, we can get visual representation $\mathcal{F}_{\theta}(x) \in \mathbb{R}^{H \times W \times C}$, where $W$, $H$ and $C$ are width, height and channel dimension. We first use global-average pooling (GAP) to get pooled fetures for attribute prediction. Then define a visual-attribute predictor $g_{\omega}: \mathbb{R}^{C} \mapsto \mathbb{R}^{A}$, which can predict $A$ number of attributes. To validate the quality of the generated attribute vector, we define the loss function:

$$\mathcal{L}_{\text{attr}} = -\frac{1}{A} \sum_{i=1}^{A} \sum_{k=1}^{N} (\hat{a}_{ik} - a_{ik}),$$

where $a_{ik}$ is the $i$-th observed attribute of the $k$-th sample and $\hat{a}_{ik}$ is the predicted ones. The whole optimization objectives of the whole model are as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{attr}},$$

where $\alpha$ is the weighting factor of $\mathcal{L}_{\text{attr}}$. The training strategy of ASL is illustrated in Algorithm 1.

\subsection{B. Attribute-Visual Attention Module}

To use attribute vectors to generate more discriminating features, we propose an attribute-visual attention module (AVAM), which can simultaneously leverage attribute and visual information to adaptively learn the discriminative features in images. 3) We show that our method achieves competitive results compared to other state-of-the-art methods.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Illustration of attribute-visual attention module}
\end{figure}
Algorithm 1 Training strategy of ASL

Input: training set $D_{train}$;
1: for all iteration=1, ..., MaxIteration do
2: sample N-way M-shot task $(Q,S)$ from $D_{train}$;
3: compute support visual features $\{F_0(s_i)\}_{i=1}^{N \times M}$;
4: compute refined support features Eq. (6)-(9);
5: compute prototypes by Eq. (1);
6: for $j$ in $\{1, ..., Q\}$ do
7: compute query visual feature $F_0(q_j)$;
8: compute refined query features Eq. (6)-(9);
9: obtain the probability distribution by Eq. (2);
10: end for
11: obtain training loss by Eq. (3)-(5);
12: update parameters in ASL by Adam;
13: end for

Output: trained ASL.

$F^a$ along height and width dimension of $F_c$, and concatenate $F_c$ and $F^a$ to get hybrid feature $F' \in \mathbb{R}^{H \times W \times (C+A)}$. Finally, we feed $F'$ into PSAM to get the final-refined feature $F_f \in \mathbb{R}^{H \times W \times C}$. The overall attribute-visual attention process can be summarized as:

\[
F_c = \mathcal{M}_c(F) \otimes F, \quad (6)
\]

\[
F_f = \mathcal{M}_s(F') \otimes F_c, \quad (7)
\]

where $\otimes$ denotes element-wise multiplication, $\mathcal{M}_c(\mathcal{X}) \in \mathbb{R}^{1 \times 1 \times C}$ denotes channel attention map and $\mathcal{M}_s(\mathcal{X}) \in \mathbb{R}^{H \times W \times 1}$ denotes spatial attention map.

**Channel attention.** To compute the channel-wise attention efficiently, we first use global average pooling and global max pooling to aggregate channel information and get $F_{avg} \in \mathbb{R}^{1 \times 1 \times (C+A)}$ and $F_{max} \in \mathbb{R}^{1 \times 1 \times (C+A)}$. Then, channel attention generating network (i.e., one layer MLP) is adopted to generate a channel-wise attention map $\mathcal{M}_c$. To summarize, the channel attention is computed as:

\[
\mathcal{M}_c(F) = \sigma(\text{MLP}(F_{avg}) + \text{MLP}(F_{max})). \quad (8)
\]

**Pyramid spatial attention.** Since different visual representation of attributes correspond to different sizes in images [23], it is necessary to design a pyramid spatial attention module. Similar to the channel attention module, we first apply average-pooling and max-pooling operations along the channel dimension and concatenate the pooled features. Then we aggregate these two features via $K$ different 2D convolution kernels. Finally, we sum all spatial-wise attention maps to generate the final spatial attention map. The pyramid spatial attention can be described as:

\[
\mathcal{M}_s(F') = \sigma(\sum_{i=1}^{K} \text{Conv}_{K_i}([\text{AvgPool}(F'); \text{MaxPool}(F')])), \quad (9)
\]

where $\text{Conv}_{K_i}$ represents a 2D convolution operation with the filter size of $K_i \times K_i$.

III. EXPERIMENTS

In this section, the effectiveness of ASL is verified by various experiments. Two standard few-shot classification datasets CUB [33] and SUN [34] are selected to compare the performance of our approach with previous few-shot recognition methods.

### A. Experimental Settings

Following [12, 25], we use both shallow four layer convolutional Conv-64F [24] and ResNet-12 [35] as our backbone network. We train our model from scratch by Adam optimizer [36] with an initial learning rate $1 \times 10^{-3}$. For ASL, we set weighting factor $\alpha = 1.0$ in all experiments. For VAP, we only use the categories of training set for training. We train our model for 60,000 iterations. For both CUB and SUN, the images are resized to $84 \times 84$, and no data augmentations are adopted. During the test stage, we report the top-1 mean accuracy over 10,000 tasks. Note that all experiments in Sec. Ex
TABLE III
Ablation study on our model. Backbone: ResNet-12.

| Method                  | 5-way 1-shot | 5-way 5-shot |
|-------------------------|--------------|--------------|
| w/o V AP                | 73.23±0.17   | 82.59±0.13   |
| w/o CAM                 | 69.86±0.19   | 86.64±0.13   |
| w/o PSAM                | 79.21±0.17   | 87.87±0.13   |
| w/o V AP & AVAM         | 69.60±0.19   | 79.83±0.23   |
| Not using attributes    | 69.98±0.19   | 86.48±0.11   |
| Using all-0 attributes  | 69.76±0.17   | 84.54±0.11   |
| ASL                     | **82.12±0.14** | **89.65±0.11** |

Unseen images

Without attributes

ASL

Fig. 3. Grad-CAM visualization of four unseen images.

TABLE IV
Influence of kernel combination. Backbone: ResNet-12.

| Kernel Size | 5-way 1-shot | 5-way 5-shot |
|-------------|--------------|--------------|
| (3, 5)      | 80.65±0.16   | 86.61±0.12   |
| (5, 7)      | 80.49±0.15   | 87.05±0.12   |
| (3, 5, 7)   | 78.30±0.17   | 86.69±0.11   |
| (3, 5, 7, 9)| **82.12±0.14** | **89.65±0.11** |

TABLE V
Influence of weighting factor $\alpha$. Backbone: ResNet-12.

| $\alpha$ | 5-way 1-shot | 5-way 5-shot |
|----------|--------------|--------------|
| 0.001    | 76.93±0.15   | 83.60±0.11   |
| 0.01     | 77.85±0.14   | 85.91±0.12   |
| 0.05     | 78.25±0.18   | 86.61±0.13   |
| 0.1      | 80.24±0.18   | 87.69±0.12   |
| 0.5      | 80.45±0.16   | 87.13±0.13   |
| 1.0      | 82.12±0.14   | **89.65±0.11** |
| 2.0      | 79.01±0.17   | 86.62±0.12   |

B. Results and Analysis

TABLE I lists the recognition results of different methods on CUB and SUN datasets. From the experimental results, we have the following observations:

1) It can be observed that ASL achieves the best performance among all approaches. To be more specific, our model is around 19.2%/8.1% better than FEAT [28] on CUB with ResNet-12 under 1-shot and 5-shot settings.

2) Multi-modal based methods (e.g., Mul. ProtoNets, AM3, Dual TriNet and Comp.) generally outperform classic meta-learning based methods that rely on the single visual modality by a large margin, which validates the effectiveness of using auxiliary semantic information.

3) Among these methods, AM3, Dual TriNet, Comp., and AGAM only augment the representations of support classes, while query images have no semantic modalities information to enhance representations. ASL generally achieves better performance than these methods, demonstrating that generating and leveraging the attribute information of query images is beneficial to improve representation learning.

4) The performance improvement of the multi-modal based methods on CUB is greater than that on SUN. This may be because there are more attribute categories in CUB, which can better guide visual attention.

C. Discussion

Pluggability of ASL. To verify the effectiveness of our proposed ASL, we embed it into three classic metric-based approaches. Table II shows the gains obtained by incorporating ASL into each approach on CUB, and for all three approaches, incorporating ASL leads to a significant improvement. These experimental results verify that our model is a flexible plug-and-play method.

Framework Design. As shown in Table III, each module in ASL has a significant contribution to performance improvement. 1) Removing the V AP drops the performance by 10.8%/7.9% on CUB under 1-shot/5-shot settings respectively, which attests that the attribute generation task brings extra useful information and significant improvement. 2) Removing either of the CAM and the PSAM leads to 32.0%/3.5% performance drops on the 1-shot setting. This demonstrates that using both attention modules simultaneously can capture more useful information. 3) To prove the effectiveness of our whole model ASL, we also remove both V AP and AVAM. Removing these two modules drops the performance by 15.2%/11.0% on 1-shot/5-shot settings respectively. 4) If we only use visual features, the improvement brought by pure visual attention mechanism is limited.

Influence of kernel combination. We select various kernels and their combinations to conduct experiments on CUB dataset. As shown in Table IV, it is better to use a combination of different kernels. A possible explanation is that different attribute features of objects correspond to different sizes in images, and using a single size cannot precisely locate attribute features.

Influence of weighting factor $\alpha$. $\alpha$ is a pre-defined parameter that influences the recognition performance. Experimental results on CUB dataset are shown in Table V. It can be seen that when we set $\alpha$ to 1, ASL can achieve the best performance.

More discussions. We further inserted AVAM after each ResBlock and achieve 83.42%/90.32% accuracy on CUB under 1-shot/5-shot settings with ResNet-12 respectively, indicating that using AVAM at different layers can provide more useful auxiliary information.

Visualization. To make an intuitive understanding of our ASL, Fig. 3 visualizes the Grad-CAM [37]. It can be seen that ASL improves the recognition performance by introducing additional attribute vectors to help the model focus on more representative local features.

IV. CONCLUSION

In this letter, we argue that auxiliary semantic modalities are necessary for query images. We propose an attribute-shaped learning method to generate corresponding attributes through visual features and learn more discriminative visual features. Experimental results illustrate the encouraging performance of our ASL, which has achieved competitive results with other state-of-the-art few-shot learning methods.
REFERENCES

[1] Huaxiong Li, Chao Zhang, Xiuyi Jia, Yang Gao, and Chunlin Chen, “Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval,” *IEEE Trans. Knowl. Data Eng.*, 2021, to be published, doi:10.1109/TKDE.2021.3102119.

[2] Ge Gao, Pei You, Rong Pan, Shunya Han, Yuanyuan Zhang, Yuchao Dai, and Hojae Lee, “Neural image compression via attentional multi-scale back projection and frequency decomposition,” in *ICCV*, 2021, pp. 14677–14686.

[3] Tao Li, Zheng Zhang, Lisen Pei, and Yan Gan, “Hashformer: Vision transformer based deep hashing for image retrieval,” *IEEE Signal Process. Lett.*, to be published, doi:10.1109/LSP.2022.3157517.

[4] Lei Zhang, Feng-Heng Li, Hao-Zhi Huang, Yong Zhang, Shao-Ping Lu, and Jue Wang, “Aesthetic-guided outward image cropping,” *ACM Trans. Graph.*, vol. 40, no. 6, pp. 211:1–211:13, 2021.

[5] Nikita Dvornik, Cordelia Schmid, and Julien Mairal, “Selecting relevant features from a multi-domain representation for few-shot classification,” in *ECCV*, 2020, vol. 12355, pp. 769–786.

[6] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao, “Lofgan: Fusing local representations for few-shot image generation,” in *ICCV*, 2021, pp. 8443–8451.

[7] Chunhao Cai, Minglei Yuan, and Tong Lu, “IFSM: an iterative feature selection mechanism for few-shot image classification,” in *ICPR*, 2020, pp. 9429–9436.

[8] Debasis Mitra, Sungrack Yun, and Fatih Porikli, “ConfeSS: A framework for single source cross-domain few-shot learning,” in *ICCV*, 2021, pp. 8443–8452.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *ICML*, 2017, vol. 70, pp. 1126–1135.

[10] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and Ilya Sutskever, “Learning transferable architectures for few-shot learning,” in *NeurIPS*, 2017, pp. 4077–4087.

[11] Haoxing Chen, Huaxiong Li, Chao Zhang, Xiuyi Jia, Yang Gao, and Chunlin Chen, “Distribution consistency based covariance metric networks for few-shot learning,” in *AAAI*, 2019, pp. 8642–8649.

[12] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal, “Multi-level semantic feature augmentation for one-shot learning,” in *ICCV*, 2019, pp. 8642–8649.

[13] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, “Matching networks for one shot learning,” in *NeurIPS*, 2016, pp. 3630–3638.

[14] Tie-Yan Liu, Guosheng Lin, and Chunhua Shen, “Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification,” *IEEE Trans. Multim.*, vol. 23, pp. 1666–1680, 2021.

[15] Huaxiong Li, Chao Zhang, Xiuyi Jia, Yang Gao, and Jiebo Luo, “Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval,” *IEEE Trans. Knowl. Data Eng.*, 2021, to be published, doi:10.1109/TKDE.2021.3102119.

[16] Ge Gao, Pei You, Rong Pan, Shunya Han, Yuanyuan Zhang, Yuchao Dai, and Hojae Lee, “Neural image compression via attentional multi-scale back projection and frequency decomposition,” in *ICCV*, 2021, pp. 14677–14686.

[17] Tao Li, Zheng Zhang, Lisen Pei, and Yan Gan, “Hashformer: Vision transformer based deep hashing for image retrieval,” *IEEE Signal Process. Lett.*, to be published, doi:10.1109/LSP.2022.3157517.