Visual - Semantic Contrastive Alignment for Few-Shot Image Classification

Anonymous ECCV Workshop submission

Paper ID 4

Abstract. Few-Shot learning aims to train and optimize a model that can adapt to unseen visual classes with only a few labeled examples. The existing few-shot learning (FSL) methods, heavily rely only on visual data, thus fail to capture the semantic attributes to learn a more generalized version of the visual concept from very few examples. However, it is a known fact that human visual learning benefits immensely from inputs from multiple modalities such as vision, language, and audio. Inspired by the human learning nature of encapsulating the existing knowledge of a visual category which is in the form of language, we introduce a contrastive alignment mechanism for visual and semantic feature vectors to learn much more generalized visual concepts for few-shot learning. Our method simply adds an auxiliary contrastive learning objective which captures the contextual knowledge of a visual category from a strong textual encoder in addition to the existing training mechanism. Hence, the approach is more generalized and can be plugged into any existing FSL method. The pre-trained semantic feature extractor (learned from a large-scale text corpora) we use in our approach provides a strong contextual prior knowledge to assist FSL. The experimental results done in popular FSL datasets show that our approach is generic in nature and provides a strong boost to the existing FSL baselines.

Keywords: Few-Shot Image Classification, Vision-Language Learning, Contrastive Learning

1 Introduction

In recent years, deep neural networks have already outperformed humans on image classification with enormous labeled samples supported, which may be against human learning behavior. Humans, however, possess a fast adaptive capacity of recognizing novel classes with a handful of annotated samples. For example, a child can easily generalize the concept of cats and quickly recognize them in reality with only one picture from a book or the Internet. In contrast, existing data-driven deep learning algorithms lag far behind humans in versatility and generalization ability. Therefore, how to construct human-like algorithms and perform visual recognition tasks under data scarcity has important practical value, which also has attracted extensive research interest. To overcome this challenge, few-shot learning (FSL) is introduced for image classification which can learn and generalize from limited data.

The main paradigm of FSL is training a model on the base classes and requiring it to accurately classify the novel classes with a limited number of examples, which is still threatened by data scarcity. There are various initial line of works study the problem of few-shot learning for image classification [22, 18, 9, 6] and establish strong baselines to improve on top. Meta-learning used to be predominant approach to solve FSL then. However, some recent works adopted standard supervision setting [20] along with various self-supervised approaches [15, 10, 19] to enhance the quality of the results. However, it is to be noted that visual categories being identified only using class labels (numerical IDs) will seriously limit the contextual features of the category since only a limited number of examples are provided. Identifying this gap, recent line of works [24, 17, 12, 2] adapted using semantic features as a prior knowledge or an auxiliary training mechanism to enhance the FSL performance. RS-FSL [2] is the recent among all to leverage categorical descriptions to perform few-shot image classifcation. However, it is to be noted that our method utilizes contrastive multimodal alignment for FSL which has never been used in the literature to the best of our knowledge. Further, our approach investigates both visual and semantic attributes in the feature level while RS-FSL predicts the descriptions using the hybrid prototype. The goal of our work is to capture the detailed semantic features and feed it to the visual feature extractor which can then be easily adopted novel categories with very few examples.

In this work we study the effectiveness of contrastive learning which has been proved to perform well [5, 4] in standard self-supervised learning. It has also been adapted to multimodal setting as well [14, 13, 1]. We utilize the simple contrastive learning objective as an auxilliary training mechanism in addition to the standard FSL baseline to provide the contextual knowledge to the model via the semantic prototype generated using a designated semantic feature extractor. We align both the semantic and visual prototypes of each class during an episode of training and employ the contrastive learning learning objective such that the corresponding prototypes regardless of the modalities to be embedded close to each other in the multimodal embedding space. This facilitates a prior knowledge to the visual feature extractor on the semantic attributes of the visual category which is crucial in few-shot image classification.

The major contribution of this approach can be summarized as follows:

- We show that a simple contrastive alignment of visual and semantic feature vectors in the embedding space formulates a generalizable visual understanding to perform few-shot image classification.
- We introduce an auxiliary contrastive learning objective on top of the existing FSL approach, hence our method is a more generic approach and can be plugged into any of the FSL baselines.
- Our experimental results on two standard FSL benchmarks show that multimodal contrastive alignment improves the performance of the standard baselines in FSL problem.

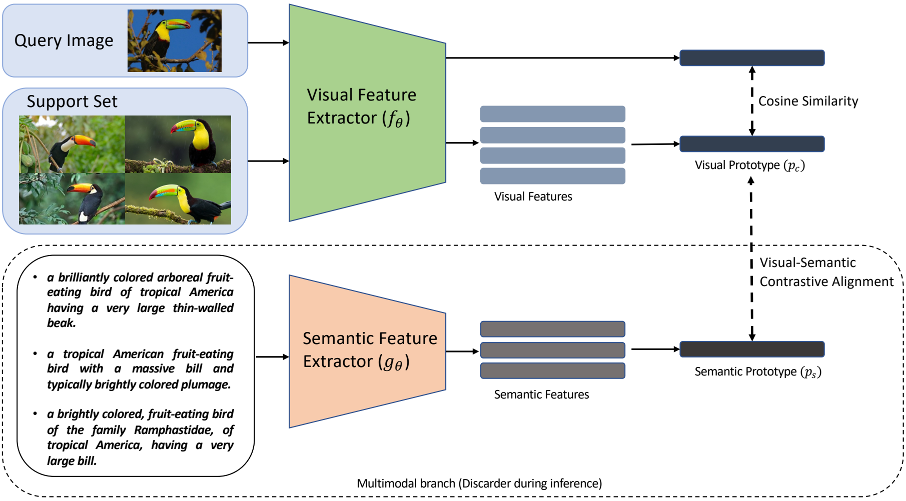
Query Image
Support Set
Visual Feature Extractor (f_θ)
Visual Features
Cosine Similarity
Visual Prototype (p_c)
Visual-Semantic Contrastive Alignment
• a brilliantly colored arboreal fruit-eating bird of tropical America having a very large thin-walled beak.
• a tropical American fruit-eating bird with a massive bill and typically brightly colored plumage.
• a brightly colored, fruit-eating bird of the family Ramphastidae, of tropical America, having a very large bill.
Semantic Feature Extractor (g_θ)
Semantic Features
Semantic Prototype (p_s)
Multimodal branch (Discarder during inference)

Fig. 1. Overall architecture of VS-Alignment for few-shot image classification. Given a support set of images and corresponding class-level descriptions, f_θ will obtain the visual prototype buy averaging the visual vectors and g_θ will create the semantic prototype by averaging the semantic feature vectors. During training we employ both the standard cross-entropy (between visual prototype and ground-truth label of query image) and the contrastive alignment (visual and semantic prototypes) as the learning objectives.

2 Proposed Method

In this work, we revisit the contrastive learning objective and leverage it as an auxiliary learning objective in the baseline few-shot learning approach. Given a support set and query images, we introduce an auxiliary contrastive alignment between visual and semantic prototypes to enhance the contextual visual knowledge of the visual prototypes. We utilize Meta-baseline [7] as our baseline approach and in Sec. 3 we study the generalizability of our approach with multiple standard FSL baselines.

This section begins with defining the few-shot learning problem for image classification (Sec. 2.1), followed by explaining about Meta-baseline [7] for FSL (Sec. 2.2) and finally the descriptions of the proposed add-on architecture to establish the visual-semantic contrastive alignment (Sec. 2.3). We name our approach as VS-Alignment.

2.1 Problem Definition

The standard few-shot image classification paradigm comprises base classes $\mathcal{C}_{base}$, in which there are enough image samples per class and novel classes $\mathcal{C}_{novel}$, where only a limited number of samples are present in each class. The class set between base and novel classes are disjoint i.e., $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \phi$. In general, FSL models are trained on K-shot, and N-way episodes. Each episode is created by first sampling N categories from the training set and then sampling two sets of images from these categories: (i) the support set $\mathcal{S}_e = (s_i, y_i)_{i=1}^{N \times K}$ containing K examples for each of the N categories and (ii) the query set $\mathcal{Q}_e = (q_j, y_j)_{j=1}^{Q}$ containing Q different examples from the same N categories. After training in this episodic training paradigm, the model is then evaluated in the novel classes ($\mathcal{C}_{novel}$) in the same N-way K-shot setting.

2.2 Meta-Baseline for FSL

We adopt the popular and recent baseline named Meta-baseline [7] to validate our argument of incorporating multimodal contrastive alignment for few-shot image classification enhances the contextual knowledge, hence improving the performance. Prior works to meta-baseline investigated the FSL problem using standard supervision setting [20, 15] and episodic learning (meta-learning) setting [22, 18] separately. However, incorporation of both standard supervision and meta-learning arguably produced better results in standard FSL datasets as mentioned by meta-baseline [7].

Classification. During this stage, the model is trained in the base classes $\mathcal{C}_{base}$ in a standard supervision setting. Given a dataset of image (x) and label ($y \in \mathcal{C}_{base}$) pairs: $\mathcal{D}_{base} = x_i, y_i$, the classifier network f maps the input image to a visual feature vector. The visual feature vector is then transformed to the label space to produce the logit p using a linear classifier. This process happens end-to-end and the standard cross-entropy loss is deployed as the learning objective as the following:

$$\mathcal{L}_{class} = -\log \frac{\exp(p_y)}{\sum_j \exp(p_j)} \tag{1}$$

After the classification stage, the last linear classifier layer is removed and the existing embedding module which maps the input image to a visual feature vector is extracted to the meta-learning stage.

Meta-Learning. During this stage, the episodic learning paradigm has been exploited on top of the supervised trained embedding module. Given a few-shot task with support set $\mathcal{S}_e$, a prototype p_c corresponding to class $c \in \mathcal{C}_{base}$ is computed by averaging the embeddings of all support samples belonging to class c:

$$p_c = \frac{1}{|\mathcal{S}_e^c|} \sum_{(s_i, y_i) \in \mathcal{S}_e^c} f_\theta(x) \tag{2}$$

where f_θ is the pre-trained visual embedding module. The evaluation is done on the query set $\mathcal{Q}_e$ with the ability of predicting the probability that sample q_j belongs to class c according to the cosine similarity between the embedding of sample q_j and p_c:

$$p(y = c | q_j, S_e) = \frac{\exp(\tau \cdot \langle f_\theta(q_j), p_c \rangle)}{\sum_k \exp(\tau \cdot \langle f_\theta(q_j), p_k \rangle)} \tag{3}$$

Here, $\langle ., . \rangle$ stands for cosine similarity and k ranges for all the classes in the support set of the episode. The learning objective at this stage is a cross-entropy loss computed from p and the labels of the samples in the query- set. During training, each training batch can contain several tasks and the average loss is computed.

2.3 VS-Alignement for FSL

With the understanding the potential of contrastive learning which has been immensly deployed in multimodal learning [14, 13], we adopted an auxilliary contrastive alignment between the visual and semantic features to enhance the few-shot image classification. To this end, as depicted in Fig. 1 we deploy a semantic feature extractor g_θ which can map the categorical descriptions available to a semantic embedding space as semantic feature vectors. Similar to Eqn. 2, we design a semantic prototype for each class in the support set of the given few-shot episode. We incorporate the proposed multimodal contrastive alignment only on the meta-learning stage of the baseline approach and the classification stage is performed without any modification.

We utilize a transformer model [21] similar to what is in CLIP [14] textual encoder to perform the semantic feature extraction. More details on the implementation will be explained in Sec. 3. For each class $c \in \mathcal{C}_{base}$, it is given that we have access to d_c number of categorical descriptions $(w_1, w_2, ..., w_{d_c})$ which we can use for multimodal contrastive alignment. p_s is formulated by averaging the semantic feature vectors of class c:

$$p_s = \frac{1}{d_c} \sum_{k=1}^{d_c} g_\theta(w_k) \tag{4}$$

We identify that the visual prototype has the knowledge of the understanding of the given visual dataset, while the semantic prototype is able to contextualize the features since the transformer model [21] is able to capture long-range dependencies effectively. Hence, incorporating both the knowledges will yield more generic and fast adoptive understanding of the given visual category. To align both the prototypes, we use simple NT-Xent loss used introduced by Chen *et al.* [5]. The auxiliary loss function at this stage to enhace the few-shot image classification is defined as:

$$\mathcal{L}_{vs}(i, p_c, p_s) = -\log \frac{\exp(\langle p_{c_i}, p_{s_i} \rangle / \tau)}{\sum\limits_{\substack{k=1 \\ k \neq i}}^{N} \exp(\langle p_{c_i}, p_{c_k} \rangle / \tau) + \sum\limits_{k=1}^{N} \exp(\langle p_{c_i}, p_{s_k} \rangle / \tau)} \tag{5}$$

The total learning objective is defined as the weighted combination of both the visual learning objective and the multimodal learning objective:

$$\mathcal{L} = \mathcal{L}_{class} + \lambda \mathcal{L}_{vs} \tag{6}$$

Here, λ is a weighting factor and is a tunable hyperparameter determined using grid-search.

Method	Backbone	Accuracy
MatchingNet [22]	Conv-4	60.52±0.88
MAML [9]	Conv-4	54.73±0.97
ProtoNet [18]	Conv-4	50.46±0.88
RFS [20]	Conv-4	41.47±0.72
L3 [3]	Conv-4	53.96±1.06
LSL [12]	Conv-4	61.24±0.96
Chen *et al.* [6]	Conv-4	60.53±0.83
Meta-Baseline [7]	Conv-4	59.30±0.86
RS-FSL [2]	Conv-4	65.66±0.90
VS-Alignment	Conv-4	**66.73±0.78**

Table 1. Performance comparison on the CUB dataset. We report average 5-way 1-shot accuracy (%) with 95% confidence interval. Table is an extended version adapted from RS-FSL [2].

3 Experiments

Datasets. We conduct experiments on two benchmark datasets for few-shot image classification: mini-ImageNet [22] and CUB [23]. The miniImageNet dataset consists of 100 image classes extracted from the original ImageNet dataset [8]. Each class contains 600 images of size 84 × 84. We follow the splitting protocol proposed by [18], and use 64 classes for training, 16 for validation, and 20 for testing. We obtained the categorical descriptions provided by [2].

The CUB dataset contains 200 classes and 11 788 images in total. We split the dataset into 100 classes for training, 50 for validation, and 50 for testing following the prior standard works [19, 2]. The categorical description for CUB is obtained from [16]. We randomly sample the required number of descriptions.

Implementation Details. To be in fair comparison with the existing works, we deploy the 4-layer convolutional architecture proposed in [18] for CUB and ResNet-12 [11] for miniImageNet. For semantic feature extractor we use pre-trained textual encoder trained using CLIP [14] model in all of our experiments. The model comprises of 12-layer transformer model [21] with 8 attention heads and 512-width. Following [15], during classification stage, we use SGD optimizer with an initial learning rate of 0.05, momentum of 0.9, and weight decay of 0.0005. We train the model for 100 epochs with a batch size of 64 and the learning rate decays twice by a factor of 0.1 at 60 and 80 epochs. During the vs-alignment (meta-learning stage), we use a contant learning rate of 0.001 with Adam optimizer and train the model for 600 epochs. We define $\lambda = 2.5$ based on the grid-search we performed. All the experiments were performed using nvidia Quadro RTX 6000 single-GPU and we report the results form 5-way 1-shot setting.

3.1 Comparison with the baselines

We report the results of our approach along with the comparison of the state-of-the art results in CUB dataset in Tab. 1. It is clear that the proposed visual-semantic alignment method outperforms the baseline approach and some of the existing state of the art approaches. This shows the importance of incorporating the semantic knowledge of the few-shot category in a contrastive style. The results on miniImagenet dataset is reported in Tab. 2. It is to be noted that even though there is an improvement in performance over the baseline [7] in the proposed approach, the gap is not significant. We hypothesize that it could have happened because of the lack of categorical descriptions compared to that of CUB dataset. In both of the experiments, we compare our approach with RS-FSL [2] which is a recent FSL approach that utilizes categorical descriptions as well.

Method	Backbone	Accuracy
ProtoNet [18]	Conv-4	55.50±0.70
Matching Net [22]	Conv-4	43.56±0.78
MAML[9]	Conv-4	48.70±1.84
Chen *et al.* [6]	Conv-4	48.24±0.75
Boosting [10]	WRN-28-10	63.77±0.45
RFS-Simple [20]	ResNet-12	62.02±0.63
RFS-Distill [20]	ResNet-12	64.82±0.60
Meta-Baseline [7]	ResNet-12	63.17±0.23
RS-FSL [2]	ResNet-12	65.33±0.83
VS-Alignment	ResNet-12	**65.89±0.80**

Table 2. Comparison with prior works on the miniImageNet.

Tab. 3 reports the results of FSL across multiple standard baselines. It is descriptive that the addition of visual-semantic alignment boosts the performance in both ProtoNet [18] and Meta-baseline [7] while it depreciates the performance in RFS and SKD [20, 15]. Since RFS and SKD are of non-episodic FSL methods, we come to a conclusion that we can plug in pur method and boost the performance only episodic few-shot learning paradigm.

Baseline	Backbone	Without VS-Alignment	With VS-Alignment
ProtoNet [18]	Conv-4	57.97±0.96	61.43±0.83
RFS [20]	Conv-4	44.93±0.76	42.36±0.64
SKD [15]	Conv-4	58.75±0.96	56.43±0.43
Meta-Baseline [7]	Conv-4	59.30±0.86	**66.73±0.78**

Table 3. Performance of different baselines both with and without the proposed visual-semantic contrastive alignment on the CUB dataset.

4 Conclusion

In this work, we introduce a simple contrastive alignment between visual and semantic prototypes of visual categories which acts as an auxiliary task to faciliate few-shot image classification. Our approach is generic in nature and can be plugged into any meta-learning based few-shot baselines. We also prove that our approach outperforms multiple standard baselines in the 5-way 1-shot few-shot setting hence establishing a new research direction to solve few-shot image classification task.

References

1. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9902–9912 (2022)
2. Afham, M., Khan, S., Khan, M.H., Naseer, M., Khan, F.S.: Rich semantics improve few-shot learning. 32nd British Machine Vision Conference (2021)
3. Andreas, J., Klein, D., Levine, S.: Learning with latent language. arXiv preprint arXiv:1711.00482 (2017)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607 (2020)
6. Chen, W., Liu, Y., Kira, Z., Wang, Y., Huang, J.B.: A closer look at few-shot classification. ArXiv **abs/1904.04232** (2019)
7. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: Exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9062–9071 (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
9. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1126–1135 (2017)
10. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 8058–8067 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Jesse Mu, P.L., Goodman, N.: Shaping visual representations with language for few-shot classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
13. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
15. Rajasegaran, J., Khan, S., Hayat, M., Khan, F.S., Shah, M.: Self-supervised knowledge distillation for few-shot learning. arXiv preprint arXiv:2006.09785 (2020)
16. Reed, S., Akata, Z., Schiele, B., Lee, H.: Learning deep representations of fine-grained visual descriptions. In: IEEE Computer Vision and Pattern Recognition (2016)
17. Schwartz, E., Karlinsky, L., Feris, R., Giryes, R., Bronstein, A.M.: Baby steps towards few-shot learning with multiple semantics. arXiv preprint arXiv:1906.01905 (2019)

18. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems (2017)
19. Su, J.C., Maji, S., Hariharan, B.: When does self-supervision improve few-shot learning? In: ECCV (2020)
20. Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J.B., Isola, P.: Rethinking few-shot image classification: a good embedding is all you need? arXiv preprint arXiv:2003.11539 (2020)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
22. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29** (2016)
23. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Tech. rep. (2010)
24. Xing, C., Rostamzadeh, N., Oreshkin, B.N., Pinheiro, P.O.: Adaptive cross-modal few-shot learning. arXiv preprint arXiv:1902.07104 (2019)