GEOMETRIC MEAN IMPROVES LOSS FOR FEW-SHOT LEARNING

Tong Wu[†‡] *and Takumi Kobayashi*[†‡]

† University of Tsukuba, Japan
‡ National Institute of Advanced Industrial Science and Technology, Japan

ABSTRACT

Few-shot learning (FSL) is a challenging task in machine learning, demanding a model to render discriminative classification by using only a few labeled samples. In the literature of FSL, deep models are trained in a manner of metric learning to provide metric in a feature space which is well generalizable to classify samples of novel classes; in the space, even a few amount of labeled training examples can construct a decent classifier. In this paper, we propose a novel FSL loss based on *geometric mean* to embed effective metric into deep features. In contrast to the other losses such as utilizing arithmetic mean in softmax-based formulation, the proposed method leverages geometric mean to aggregate pair-wise relationships among samples for enhancing discriminative metric across class categories. The proposed loss is not only formulated in a simple form but also is thoroughly analyzed in theoretical ways to reveal its favorable characteristics which are favorable for learning feature metric in FSL. In the experiments on few-shot image classification tasks, the method produces competitive performance in comparison to the other losses.

Index Terms— few-shot learning, loss, geometric mean

1. INTRODUCTION

Few-shot learning (FSL) draws inspiration from the remarkable human ability of robust reasoning and analysis, particularly in scenarios where limited information is available. This paradigm has gained significant traction in various applications, e.g., autonomous vehicles and medical analysis, where resource constraints necessitate efficient learning from scarce data. FSL [1] is formulated as a type of machine learning problem where only a limited number of examples with supervised information are available for the target task. Thus, a key challenge in conquering FSL lies in how to efficiently utilize limited data, which has driven research into various approaches to tackle this challenging problem, mainly categorized into *meta learning* and *metric learning*.

For rapid adaptation to new tasks with limited data, meta-learning approaches, such as MAML [2] and Reptile [3], aim to *learn-to-learn* through a complex two-stage process; it is composed of a meta-training phase for learning the model to adapt quickly toward various tasks and a meta-testing phase to deploy the adaptation ability to new tasks. While being potentially flexible, these approaches often suffer from the computation issues that the complex training processes require significant computational resources, involving careful hyperparameter tuning.

On the other hand, metric learning is applied to construct a feature space where semantically similar samples are closely embedded; in the space equipped with such a favorable metric, even novel samples could be discriminated on the basis of a few number of labeled samples. The metric-learning approaches [4, 5, 6] often produce robust performance across various domains without requiring extensive fine-tuning. It is also computationally efficient as the deep models are trained in a rather straightforward way based on a *loss* function that induces effective metric in deep feature representation. Therefore, a loss plays a key role in the metric-based FSL. For embedding better feature metric in deep models, the loss is required to take into account whole training samples, though the FSL losses [5, 6] have difficulty in fully paying attention to whole sample distribution.

We propose a novel loss to learn effective feature metric via a deep model for FSL. The proposed method is built upon softmax-based attention weight [7, 6] to encode pair-wise relationships among samples. We then leverage *geometric mean* to efficiently aggregating those pair-wise weights for taking into account broader structure of sample distribution in a deep feature space. It should be noted that the other FSL losses pay much attention to rather limited structure/amount of samples, impeding metric learning over whole samples. While the proposed method is formulated in a simple loss form, our thorough analysis clarifies various characteristics of the loss in theoretical ways which exhibit superior suitability for FSL metric learning in comparison to the other losses.

Our main contributions are summarized as follows.

- We propose a novel loss for learning effective metric in FSL by means of geometric mean over softmax-based attention weights which efficiently encodes structure of whole samples to facilitate metric learning for FSL.
- We theoretically analyze the proposed loss from various perspectives to reveal favorable characteristics for learning metric of deep features in FSL, while clarifying interesting connections to the other losses.

- The experimental results on few-shot image classification tasks demonstrate the efficacy of the proposed loss in comparison to the other FSL losses.

2. METHOD

We start with briefly reviewing two representative loss functions for FSL, PN Loss [5] and NCA loss [6], and then formulate our proposed loss.

Notations. Suppose we have a *support* set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a target *query* sample $(\boldsymbol{x}_q, y_q)$ for constructing a loss, as shown in Fig. 1; an input image $\mathcal{I}$ is encoded into a D-dimensional feature vector $\boldsymbol{x} = f_\theta(\mathcal{I}) \in \mathbb{R}^D$ via a deep model f_θ equipped with trainable parameters θ, and it is also annotated by a class label $y \in \{1, \cdots, C\}$. Distance metric denoted by $\mathtt{d}(\boldsymbol{x}, \boldsymbol{z})$ measures a discrepancy between two vectors of $\boldsymbol{x}$ and $\boldsymbol{z}$; it can be specified such as by Euclidean distance $\mathtt{d}(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|_2^2$, as will be discussed in Sec. 2.4.

2.1. Metric-based losses for FSL

To embed effective metric into the feature representation $\boldsymbol{x}$ for FSL, it is a key process to train the deep model f_θ on the basis of a loss which works on the distance metric in the feature space.

2.1.1. PN loss

PN loss used in a prototypical network [5] is designed so as to push a query $\boldsymbol{x}_q$ toward the cluster center of class y_q (Fig. 1a), which naturally leads to the loss formulation of

$$\ell_{PN}(\boldsymbol{x}_q, y_q|\mathcal{S}) = -\log \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{\mu}_{y_q}))}{\sum_{c=1}^C \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{\mu}_c))}, \quad (1)$$

where a cluster center of class c is given by $\mu_c = \frac{1}{n_c}\sum_{i|y_i=c} \boldsymbol{x}_i$ and the number of samples in the c-th class is denoted by n_c; $n = \sum_{c=1}^C n_c$.

In the PN loss (1), each class is represented by the center vector and then a query sample $\boldsymbol{x}_q$ is required to reduce the distance against the class center μ_{y_q} that it belongs to, in a manner similar to Center loss [8]. The loss is formulated by means of softmax comparing the distance to those of the other classes $c \neq y$, so that it induces compact feature representation within a class. However, as each class c is described by only a single center μ_c, the PN loss assumes a uni-modal distribution in a class, imposing rather *hard* constraint on feature representation; it is difficult to cope with complicated in-class distribution of multiple modes.

2.1.2. NCA loss

In contrast to the PN loss, the NCA loss [6] is built upon sample-wise relationships more directly. It is derived from neighborhood component analysis (NCA) [9] as

$$\ell_{NCA}(\boldsymbol{x}_q, y_q|\mathcal{S}) = -\log \frac{\sum_{i|y_i=y_q} \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))}{\sum_{j=1}^n \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))}. \quad (2)$$

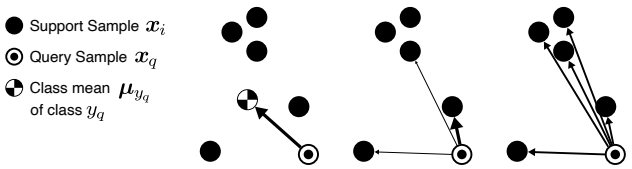
It aggregates pair-wise relationships (distances) $\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i)$ in a direct way without resorting to class centers. In addition, the NCA loss effectively pay much attention to *neighbor* samples due to `log-sum-exp` of

$$\log \sum_j \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j)) =$$

$$-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_{j^*}) + \log\left[1 + \sum_{j \neq j^*} \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))}{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_{j^*}))}\right], \quad (3)$$

where $j^* = \arg\min_j \mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j)$ indicates a nearest-neighbor sample in a support set. Thereby, the (nearest) neighbor samples could dominate the denominator and numerator in (2) since the second term in (3) is significantly decayed for far-away samples $\{k|\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_k) \gg \mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_{j^*})\}$. That is, in (2), the numerator pays attention to *in-class* neighbors while the denominator focuses on *global* neighbors on $\mathcal{S}$. Thereby, this sample-wise approach can effectively deal with a complicated distribution of even multiple modes in contrast to the PN loss.

On the other way, the emphasis on neighbor samples would make the far-away samples less contributive to the loss, hindering the metric learning over *whole* samples; specifically, as shown in Fig. 1b, the metric against those far-away samples is hardly improved in the loss (2).



(a) PN loss (b) NCA loss (c) Ours

Fig. 1. Mechanisms of three types of losses over samples.

2.2. Proposed loss

Toward further effective feature metric, we formulate a metric-based loss by means of *geometric mean*.

We first rewrite the NCA loss (2) into

$$\ell_{NCA} = -\log \frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))}{\sum_{j=1}^n \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))} - \log n_{y_q}, \quad (4)$$

where the second term is just a constant and thus the first term is an intrinsic form of the NCA loss, based on *arithmetic mean* over within-class softmax-based attention weights of

$$\left\{\mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i) \triangleq \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))}{\sum_{j=1}^n \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))}\right\}_{i|y_i=y_q}. \quad (5)$$

As discussed above, far-away samples $\{k|\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_k) \gg \mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_{j^*})\}$ gain less attention weights, thereby hardly contributing to the arithmetic mean.

To remedy it, we leverage *geometric mean* to aggregating the softmax-based attention weights by

$$\begin{aligned}\ell_{ours}(\boldsymbol{x}_q, y_q|\mathcal{S}) &= -\log\big[\textstyle\prod_{i|y_i=y_q} \mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i)\big]^{\frac{1}{n_{y_q}}} \\ &= -\log\left[\prod_{i|y_i=y_q} \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))}{\sum_{j=1}^{n}\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))}\right]^{\frac{1}{n_{y_q}}}. \quad (6)\end{aligned}$$

While it is a simple modification from (4), the proposed loss (6) endows an important characteristic with metric learning that prohibits *any* attention weight $\mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i) = \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))}{\sum_{j=1}^{n}\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))}$ from being close to 0; it is highly contrastive to NCA loss (Sec. 2.1.2) which could lead to sparse attention weights dominated by neighbor samples. Thus, the proposed loss is effective for learning favorable metric to take into account *whole* in-class samples including far-away ones (Fig. 1c). It is also noteworthy that the formulation (6) is further rewritten into a simpler form of

$$\ell_{ours} = \frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i) + \log \sum_{j=1}^{n} \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j)), \quad (7)$$

which is computed by using simple functions of `sum` and `log-sum-exp`.

2.3. Discussion

As described in Sec. 2.2, our method can effectively cope with complicated sample distributions by means of sample-wise softmax attention weights while rendering metric learning to *whole* samples; these two points highlight the contrasts to PN loss (Sec. 2.1.1) and NCA loss (Sec. 2.1.2), respectively.

In addition, we analyze the proposed loss from the following three perspectives, which further clarifies connection not only to PN and NCA losses but also to classification losses.

2.3.1. Relationship to NCA loss

The proposed loss (6) works as an upper bound of the NCA loss (2) as

$$\begin{aligned}\ell_{ours} &= -\log\Big[\prod_{i|y_i=y_q} \mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i)\Big]^{\frac{1}{n_{y_q}}} \\ &\geq -\log \frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i) = \ell_{NCA}, \quad (8)\end{aligned}$$

which is easily proven by using Jensen's inequality. Thus, even in case that the NCA loss is saturated, our loss would be still valid for further learning metrics in a similar way to [10].

To further clarify the mechanism of the proposed loss in comparison to the NCA loss, we analyze them through the lens of loss gradients. The gradients of those two losses with respect to $\boldsymbol{x}_q$ are given by

$$\frac{\partial \ell_{ours}}{\partial \boldsymbol{x}_q} = -\frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \frac{1}{\mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i)} \frac{\partial}{\partial \boldsymbol{x}_q} \mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i), \quad (9)$$

$$\frac{\partial \ell_{NCA}}{\partial \boldsymbol{x}_q} = -\frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \frac{1}{\bar{\mathtt{a}}_{\mathcal{S}}(\boldsymbol{x}_q, y_q)} \frac{\partial}{\partial \boldsymbol{x}_q} \mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i), \quad (10)$$

where $\bar{\mathtt{a}}_{\mathcal{S}}(\boldsymbol{x}_q, y_q) = \frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i)$ is an averaged attention weight in the target class y_q. The key difference between the loss gradients (9, 10) is in the weights for sample-wise gradients $\frac{\partial}{\partial \boldsymbol{x}_q} \mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i)$. Our loss gradient (9) employs adaptive weights based on the attention $\mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i)$; for far-away samples exhibiting less attention, the gradient at the sample is provided with the higher weight $\frac{1}{\mathtt{a}_{\mathcal{S}}(\boldsymbol{x}_q, \boldsymbol{x}_i)}$, which effectively promotes metric learning in a similar way to mean shift [11]. On the other hand, in the NCA loss, the sample-wise gradients (10) are equipped with a *uniform* weight $\frac{1}{\bar{\mathtt{a}}_{\mathcal{S}}(\boldsymbol{x}_q, y)}$, being less efficient especially at the far-away samples which produce gradients of small magnitude. This analysis regarding loss gradients clarifies the efficacy of the proposed loss for learning.

2.3.2. Relationship to PN loss

As shown in (7), our method reduces sum of sample-wise distances $\frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i)$ which, by using L_2 distance $\mathtt{d} = \|\cdot\|_2^2$, is decomposed as

$$\frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \|\boldsymbol{x}_q - \boldsymbol{x}_i\|_2^2 = \|\boldsymbol{x}_q - \boldsymbol{\mu}_{y_q}\|_2^2 + \frac{1}{n_{y_q}} \sum_{i|y_i=y_q} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{y_q}\|_2^2. \quad (11)$$

While the first term indicates the distance to the class center $\boldsymbol{\mu}_{y_q}$, which is the primary target to be minimized in the PN Loss (1), the second term is a within-class variance at the class y_q. Therefore, our method minimizes not only the distance to a class center as in the PN loss but also the within-class variance, in order to further enhance feature discriminativity; the within-class variance that PN loss lacks is also helpful for learning feature metric from discriminative perspective as shown in discriminant analysis [12].

2.3.3. Relationship to classification loss

The proposed loss (6) is also described by

$$\ell_{ours} = -\sum_{i|y_i=y_q} \frac{1}{n_{y_q}} \log \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))}{\sum_{j=1}^{n}\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))}, \quad (12)$$

which can be viewed as a *multi-label* softmax loss in a classification framework where a query $\boldsymbol{x}_q$ is categorized into

Loss	*mini*ImageNet		CIFAR-FS	
	1-shot	5-shot	1-shot	5-shot
BCE [13]	61.61±0.20	76.35±0.16	69.26±0.22	83.82±0.17
ASL [14]	58.01±0.20	70.49±0.17	66.61±0.23	78.82±0.17
Ours	**64.04±0.20**	**79.12±0.15**	**71.19±0.22**	**84.15±0.16**

Table 1. Performance comparison (accuracy %) from the viewpoint of multi-label classification losses (Sec. 2.3.3).

n *samples* which are respectively represented by support samples $\{\boldsymbol{x}_i\}_{i=1}^n$. The classification is based on n logits of $\{-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i)\}_{i=1}^n$ and *multi-hot* labels over n samples, denoted by

$$\hat{\mathtt{p}}_i = \begin{cases} \frac{1}{n_{y_q}} & \text{if } y_i = y_q \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in \{1, \cdots, n\}, \sum_{i=1}^{n} \hat{\mathtt{p}}_i = 1. \tag{13}$$

Therefore, the loss (12) is equivalent to cross-entropy between the multi-hot label (13) and the softmax posterior probabilities over n samples, i.e., the attention weights (5). From this perspective, our loss enforces the softmax probabilities (5) to be close to the multi-hot ones (13), leading to

$$\frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))}{\sum_{j=1}^n \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))} = \frac{\exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_{i'}))}{\sum_{j=1}^n \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))} = \frac{1}{n_{y_q}}$$
$$\Rightarrow \mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i) = \mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_{i'}), \forall (i, i') | y_i = y_{i'} = y_q. \tag{14}$$

This analysis also reveals that our loss pushes $\boldsymbol{x}_q$ toward *medoid* of class y_q on the distance metric $\mathtt{d}$.

2.4. Distance metric

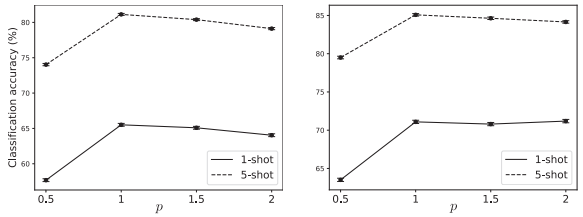
We can arbitrarily design basic distance metric $\mathtt{d}$ used in the loss; in this literature, Euclidean distance $\mathtt{d} = \|\cdot\|_2^2$ is commonly utilized to produce favorable performance [6, 5]. In this work, it is formulated based on L_p norm as

$$\mathtt{d}_p(\boldsymbol{x}, \boldsymbol{z}) = \sum_{d=1}^{D} |x_d - z_d|^p. \tag{15}$$

As discussed in Sec. 2.3.3, different types of distance metric, i.e., p in (15), pull a query $\boldsymbol{x}_q$ to different medoids; for $p = 2$ (Euclidean distance), $\boldsymbol{x}_q$ is moved toward a simple class mean vector while $p = 1$ provides a medoid robust to outliers [12]. We empirically analyze the distance metric in Sec. 3.2.

3. RESULT

We empirically evaluate and analyze the proposed loss (6) on few-shot image classification tasks; in the experiemnts, we primarily focus on performance in terms of loss functions in the FSL framework.



*mini*ImageNet CIFAR-FS

Fig. 2. Performance analysis of various distance metric $\mathtt{d}_p$.

3.1. Experimental settings

Datasets. We employ three FSL benchmark datasets. The *mini*ImageNet [7], derived from the ImageNet, consists of 100 classes with 600 images per class. The CIFAR-FS [19], a variant of CIFAR-100 [20], comprises 100 classes with 600 images per class. They are split in a way of [21]; 64 classes for training, 16 classes for validation, and 20 classes for test sets, which are disjoint sets in terms of class categories. The *tiered*ImageNet [22], based on ImageNet, contains 608 classes which are split into 351 training, 97 validation and 160 test classes. Input images are resized into 84×84 pixels.

FSL framework. Following [6], we train a deep model $\boldsymbol{x} = f_\theta(\mathcal{I})$ by the loss (Sec. 2) computed on mini-batch samples $\mathcal{B} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^B$ randomly drawn from a training set. Then, the mini-batch set $\mathcal{B}$ is divided into one query sample $(\boldsymbol{x}_q, y_q) \in \mathcal{B}$ and a support subset $\mathcal{S} = \mathcal{B} \setminus (\boldsymbol{x}_q, y_q)$, and then we compute the loss repeatedly in a leave-one-out manner as

$$\ell(\mathcal{B}) = \mathop{\mathrm{E}}_{(\boldsymbol{x}_q, y_q) \in \mathcal{B}} \ell(\boldsymbol{x}_q, y_q | \, \mathcal{B} \setminus (\boldsymbol{x}_q, y)). \tag{16}$$

Evaluation protocol. Then, we evaluate the FSL performance of the trained model f_θ by following [23, 24]. For simulating few-shot scenarios, we draw N-way K-shot samples, i.e., K labeled samples of N novel classes, from a test set which contains non-overlapping classes with those of the training set; for each few-shot scenario, we also draw 15 *unlabeled* query samples per class. By embedding those NK labeled samples into a feature space via f_θ, a simple classifier is constructed by means of a nearest-mean classification to categorize the unlabeled query samples into one of N classes. We evaluate performance on $N = 5$-way $K \in \{1, 5\}$-shot scenarios by reporting the averaged classification accuracy with 95% confidence interval over 10,000 trials.

Model architecture. We apply a ResNet12 [25] as a deep model f_θ. We train the model from scratch on the training set which is additionally equipped with a linear projection head [6] to produce $D = 192$-dimensional feature vector $\boldsymbol{x}$ for facilitating metric learning only in the training phase; the projection head is detached at the test phase to produce 640-dimensional features used for novel classification (evaluation). For training the model, we apply to a mini-batch of $B = 512$ samples an SGD optimizer with Nesterov momentum of 0.9, weight decay of 5e-4, and initial learning rate of 0.1 which is decayed by a factor of 10 at 84-th epoch over 120 training epochs including 10 warm-up epochs.

Loss	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	CIFAR-FS 1-shot	CIFAR-FS 5-shot	*tiered*ImageNet 1-shot	*tiered*ImageNet 5-shot
PN [5] (1)	62.42±0.20	79.13±0.15	67.14±0.22	82.36±0.16	66.74±0.23	82.14±0.17
NCA [6] (2)	62.68±0.20	78.93±0.54	69.20±0.21	84.24±0.16	67.21±0.22	83.77±0.16
Ours (6)	**65.51±0.20**	**81.13±0.14**	**71.09±0.22**	**85.08±0.16**	**69.61±0.23**	**84.04±0.16**

Table 2. Performance results of FSL losses. The best results are highlighted in **bold**.

Loss	*mini*ImageNet 5-shot
GM (17)	79.06±0.15
Ours (6)	**81.13±0.14**

Table 3. Performance comparison of geometric-mean based losses.

Method	*mini*ImageNet 1-shot	*mini*ImageNet 5-shot	CIFAR-FS 1-shot	CIFAR-FS 5-shot	tieredImageNet 1-shot	tieredImageNet 5-shot
DN4 [15]	61.23±0.36	75.66±0.29	-	-	-	-
CAN [16]	63.85±0.48	79.44±0.34	-	-	**69.89±0.51**	84.23±0.37
Meta-Baseline [17]	63.17±0.23	72.96±0.17	**72.00±0.70**	84.20±0.50	68.62±0.27	83.74±0.18
RFSIC-simple [18]	62.02±0.63	79.64±0.44	71.50±0.80	**86.00±0.50**	69.74±0.72	**84.41±0.55**
Ours (6)	**65.51±0.20**	**81.13±0.14**	71.09±0.22	85.08±0.16	69.61±0.23	84.04±0.16

Table 4. Comparison to various FSL methods. For the comparison methods [15, 16, 17, 18], the reported scores in respective papers are shown. Scores falling within the confidence interval of the **best** are indicated by underline.

3.2. Performance analysis

We analyze the method from various aspects.

Multi-label classification loss. As discussed in Sec. 2.3.3, our loss is also viewed from the perspective of a multi-label classification loss. Thus, we compare our loss (6) with the other types of multi-label losses which are widely applied in the classification literature, binary cross-entropy loss (BCE) [13] and asymmetric loss (ASL) [14]; for fair comparison, all the losses are computed on the logits of $\{\mathtt{d}_{p=2}(\boldsymbol{x}_q, \boldsymbol{x}_i)\}_{i=1}^n$ with the multi-hot labels (13). Table 1 reports performance comparison on *mini*ImageNet and CIFAR-FS datasets, demonstrating that the proposed loss outperforms those multi-label classification losses. The multi-label losses of BCE and ASL mainly focus on sample-wise relationship $\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i)$ in an individual manner while rather paying less attention to the relationships among whole samples. In contraast, as discussed in Sec. 2.3.3, our loss effectively pushes a query sample toward class medoid, which is favorable for learning effective metric in the FSL framework.

Distance metric $\mathtt{d}_p$. We formulate the distance metric $\mathtt{d}_p$ in (15) based on L_p norms and here empirically evaluate performance of various p in Fig. 2. The results show that $p = 1$ produces favorable performance. The distance metric $\mathtt{d}_{p=1}$ provides an effective medoid representation in (14) robust against some outlier samples which may be included in the training set. The robustness contributes to enhancing metric learning. We apply $\mathtt{d}_{p=1}$ to the losses in the following experiments.

Comparison to FSL losses. In Sec. 2.3, we have analyzed superiority to PN loss (1) and NCA loss (2) from theoretical viewpoints. We also show quantitative performance comparison to those FSL losses in Table 2; based on the above analysis, all the methods are equipped with the distance metric $\mathtt{d}_{p=1}$ for fair comparison. In accordance with the discussion in Sec. 2.3, our loss consistently outperforms PN and NCA losses on various datasets and FSL scenarios. Specifically, comparison to NCA loss clarifies the efficacy of our *geometric mean* in comparison to simple arithmetic mean for aggregating softmax attention weights. It should be noted that the proposed loss is as simple as those two methods without increasing computation cost.

The effectiveness of the geometric mean over the softmax attention is further highlighted through comparison to

$$\ell_{GM} = -\log \frac{\left[\prod_{i|y_i=y_q} \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i))\right]^{\frac{1}{n_{y_q}}}}{\sum_{c=1}^{C}\left[\prod_{j|y_j=c} \exp(-\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_j))\right]^{\frac{1}{n_c}}}, \quad (17)$$

which applies geometric mean to aggregate sample-wise relationships $\mathtt{d}(\boldsymbol{x}_q, \boldsymbol{x}_i)$ within a class. It fails to exhibit the characteristics discussed in Sec. 2.3, thus being inferior to ours as shown in Table 3.

Comparison to others. We also show comparison to the other FSL approaches in Table 4, though our main focus is the loss function as shown in Tables 1-3. While our method just works on a loss, it provides competitive performance even in comparison to the other FSL approaches.

4. CONCLUSION

We have proposed a FSL loss based on geometric mean of softmax-based sample-wise attention weights. While the loss is formulated in a simple form, our theoretical analysis reveals that it renders various favorable characteristics to metric learning for FSL in comparison to the other FSL losses. The experimental results on few-shot image classification tasks empirically demonstrate the efficacy of the proposed loss.

5. REFERENCES

[1] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys*, vol. 53, no. 3, pp. 1–34, 2020.

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

[3] Alex Nichol and John Schulman, "On first-order meta-learning algorithms," *arXiv:1803.02999*, 2018.

[4] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al., "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015.

[5] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017, pp. 4080–4090.

[6] Steinar Laenen and Luca Bertinetto, "On episodes, prototypical networks, and few-shot learning," in *NeurIPS*, 2021, pp. 24581–24592.

[7] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016.

[8] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[9] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov, "Neighbourhood components analysis," in *NeurIPS*, 2005.

[10] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[11] Dorin Comaniciu and Peter Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE TPAMI*, vol. 24, no. 5, pp. 603–619, 2002.

[12] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley, New York, 2001.

[13] Grigorios Tsoumakas and Ioannis Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[14] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor, "Asymmetric loss for multi-label classification," in *ICCV*, 2021, pp. 82–91.

[15] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *CVPR*, 2019, pp. 7260–7268.

[16] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen, "Cross attention network for few-shot classification," in *NeurIPS*, 2019.

[17] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *ICCV*, 2021.

[18] 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 *ECCV*, 2020, pp. 266–282.

[19] Luca Bertinetto, Jo ao Henriques, Philip H.S. Torr, and Andrea Vedaldi, "Meta-learning with differentiable closed-form solvers," in *ICLR*, 2019.

[20] Alex Krizhevsky and Geoffrey E. Hinton, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, 2009.

[21] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2016.

[22] 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 *ICLR*, 2018.

[23] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens Van Der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," *arXiv:1911.04623*, 2019.

[24] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell, "Meta-learning with latent embedding optimization," *arXiv:1807.05960*, 2018.

[25] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*, 2019.