Analyzing the Accuracy, Representations, and Explainability of Various Loss Functions for Deep Learning

Tenshi Ito, Hiroki Adachi, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi
Chubu University
1200 Matsumoto-cho, Kasugai, Aichi, Japan
{tenchi, ha618, hirakawa}@mprg.cs.chubu.ac.jp, {takayoshi, fujiyoshi}@isc.chubu.ac.jp

***Abstract*—Deep learning utilizes a vast amounts of training data and updates weight parameters so as to minimize the loss between a predicted probability and a ground truth label. Generally, we use cross-entropy as the loss function. Although loss functions for image classification other than cross-entropy exist, their efficacy has not been adequately investigated. In this work, we extensively analyze models trained with different loss functions and clarify the properties of each. Specifically, we analyze the feature space and explainability as well as the classification accuracy on various benchmark datasets and network architectures. For feature space and explainability, we investigate the effectiveness of each loss function by quantitative and qualitative evaluations. We then discuss the properties and improvements of each.**

***Index Terms*—loss function, attention, accuracy, representations, explainability**

I. Introduction

In computer vision, convolutional neural networks (CNNs) not only improve image classification but also aim to explain the classification reasons visually. CNNs can achieve performances on par with that of human beings by updating the weight parameters of a network so as to minimize the loss of vast amounts of training data. While the loss during is typically computed with cross-entropy, various other loss functions have been proposed for obtaining excellent feature representations [1]–[8].

Cross-entropy computes with predicted distribution and a ground truth label. It considers only the correct class probability because weight parameters are optimized so as to be close to the one-hot vector of the ground truth. By minimizing cross-entropy loss, the correct probability is high and incorrect probabilities inevitably decrease. Cross-entropy loss plays a key role in separating each feature and is thus an essential loss function for image classification.

Center loss [2] and prototype conformity loss (PC loss) [5] are loss functions related to features utilizing trainable class centroids prepared in equal number of the classes. Specifically, center loss minimizes the Euclidean distance so as to keep each feature close to the correct class centroid in an arbitrary feature space. By introducing center loss, the feature space in CNNs inhibits the intra-class variance by gathering the same class features and improving the classification accuracy. Feature space applied center loss becomes a large inter-class variance indirectly because of gathering same class features. As for PC loss, it can pull apart directly between classes by including the maximization of inter-class distance. PC loss has been proposed for adversarial training [9], [10] and can obtain a robust model by pulling apart a margin between features and the decision boundary. Achieving the best performance with these two loss functions requires them to be combined with cross-entropy loss.

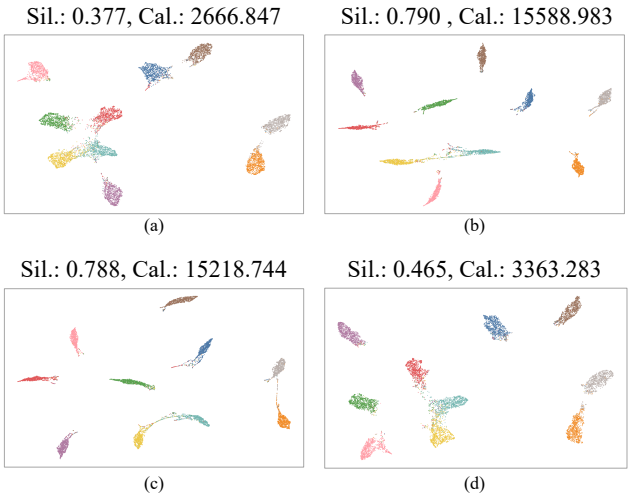

Fig. 1. Feature spaces on CIFAR-10 dataset compressed with UMAP. (a), (b), and (c) are feature spaces with only cross-entropy loss, cross-entropy loss + center loss, and cross-entropy loss + PC loss, respectively. (d) is the feature space with cross-entropy loss + COT.

Chen *et al.* [6] proposed complement objective training (COT) as a training method that maximizes complement entropy. Complement entropy is the sum of the entropy of all classes except the correct class. COT flattens incorrect class probabilities by maximizing complement entropy after minimizing cross-entropy loss. With this training approach, COT can lead to excellent classification by feature space that has a narrower intra-class variance than standard training and improves the separateness of the inter-class.

Although these loss functions have similar effectiveness, little is known about their respective benefits and drawbacks. Moreover, past evaluations have typically focused on clas-

sification accuracy or qualitative demonstrations of feature representations, while explainability and quantitative investigation of a feature space are frequently neglected. Since feature representations compress high-dimensional features to just 2 or 3 dimensions with t-SNE [11] or UMAP [12], it is difficult to compare appropriately when only qualitative evaluation is used because the feature spaces of each model exhibit the same trends (see Fig. 1). In the current work, we fairly train with these loss functions by utilizing various network architectures and datasets and then evaluate the trained models in terms of *accuracy*, *representations*, and *explainability* to determine the effectiveness of each loss function. Classification accuracy is the match rate between the predicted class and the ground truth. For feature representations, we perform not only qualitative evaluation of a feature space compressed with UMAP but also quantitative evaluation for raw features without any dimension reduction methods by using the Calinski & Harabasz index (Cal.) [13] and Silhouette score (Sil.) [14]. By leveraging these metrics, we can interpret models quantitatively. Explainability visualizes attention maps with Grad-CAM [15], which we evaluate quantitatively with an insertion/deletion score [16]. Following these three analyses, we discuss the properties and effectiveness of each loss function.

Contributions. In this work, we comprehensively evaluate loss functions that have been used unconsciously by many researchers thus far. We also investigate not only individual loss functions but also how they behave when combined with COT. Our analyses examine the effectiveness of models trained with each loss function in terms of *accuracy*, *representations*, and *explainability*, and using the results as a basis, we clarify the properties of the loss functions and the effectiveness of various combination that have hitherto been ignored.

II. Preliminaries and Related Works

In this section, we define the variables and empirical risk minimization utilized in this paper. We then describe the loss functions in II-A and explainability in computer vision in II-B.

Notations. We use a training dataset $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^{c\times h\times w}$ is an image and $y_i \in \mathcal{Y} := \{0, 1, \dots, K-1\}$ is a ground truth to $\boldsymbol{x}_i$. We denote a model $f : \mathbb{R}^{c\times h\times w} \to \mathbb{R}^K$ parameterized by $\boldsymbol{\theta}$ that maps $\boldsymbol{x}_i$ to a K-dimension vector. The loss between a predicted distribution and the ground truth can be computed with the following cross-entropy loss:

$$\mathcal{L}(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i) = -\log \sigma_{y_i}(f(\boldsymbol{x}_i;\boldsymbol{\theta})), \tag{1}$$

where $\sigma : \mathbb{R}^K \to [0,1]^K$, $\sum_{k=0}^{K-1} \sigma_k(f(\boldsymbol{x}_i;\boldsymbol{\theta})) = 1$, and σ_{y_i} is the true class probability. To achieve an excellent performance, we update the weight parameters $\boldsymbol{\theta}$ so as to minimize the loss to training data:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}} \left[\mathcal{L}(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)\right]. \tag{2}$$

The model trains various images by applying a geometric transform to the training data $\boldsymbol{x}_i$ rather than directly using it.

A. Loss function for image classification

Image classification in deep learning is typically done by computed the loss between a predicted probability and the ground truth label with cross-entropy Eq. (1). Cross-entropy loss plays a key role in separating features and achieves excellent classification by minimizing the loss. Recent methods utilizing CNNs have attempted to improve the classification performance by training with loss functions into not only likelihood space but also feature space.

Center loss. Center loss [2] is the loss function that inhibits intra-class variance by concentrating the same class features on one point in an arbitrary feature space. Specifically, it prepares trainable centroids equal to the number of classes and computes the ℓ_2 norm between each feature and the correct class centroid by

$$\mathcal{L}_{\text{center}} = \frac{1}{2}\sum_{i}^{n} \|g(\boldsymbol{x}_i) - \boldsymbol{w}_{y_i}^c\|_2^2, \tag{3}$$

where $g : \mathbb{R}^{c\times h\times w} \to \mathbb{R}^d$ is part of f and $\boldsymbol{w}_{y_i}^c \in \mathbb{R}^d$ is the correct centroid to $g(\boldsymbol{x}_i)$. Center loss implicitly introduces the effect of separate inter-class by inhibiting intra-class variance.

PC loss. PC loss [5] also utilizes trainable centroids but differs from center loss in that it ncludes the ℓ_2 norm between each feature and incorrect class centroids and between the centroids from the center loss. PC loss is represented as

$$\mathcal{L}_{\text{pc}} = \sum_{i}^{n}\{\|g(\boldsymbol{x}_i) - \boldsymbol{w}_{y_i}^c\|_2 - \frac{1}{K-1}\sum_{j\neq y_{y_i}} (\|g(\boldsymbol{x}_i) - \boldsymbol{w}_j^c\|_2 + \|\boldsymbol{w}_{y_i}^c - \boldsymbol{w}_j^c\|_2)\}, \tag{4}$$

where $\boldsymbol{w}_j^c \in \mathbb{R}^d$ indicates an incorrect class centroid. As aforementioned, although PC loss computes the ℓ_2 norm between centroids, among the centroids separates inevitably if it can pull apart each feature to an incorrect cluster. Thus, PC loss can obtain the same effect without its computing.

Complement entropy. Complement entropy is a loss function utilized in complement objective training (COT) [6]. It is the sum of entropy to all classes except the correct class and is represented by

$$\mathcal{C} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, j\neq y_i}^{K} \frac{p_j(\boldsymbol{x}_i)}{1 - p_{y_i}(\boldsymbol{x}_i)} \log\left(\frac{p_j(\boldsymbol{x}_i)}{1 - p_{y_i}(\boldsymbol{x}_i)}\right), \tag{5}$$

where $p(\boldsymbol{x}_i) := \sigma(f(\boldsymbol{x}_i;\boldsymbol{\theta}))$. COT maximizes the complement entropy after minimizing the cross-entropy so that a predicted distribution and the ground truth can be matched. By minimizing complement entropy, the predicted distribution becomes sharp because incorrect class probabilities are flattened. Thus, CNNs that apply COT lead to accurate classification due to improvements in the separateness for inter-class. Guided complement entropy (GCE) [7] is a derivative of COT that attempts to enhance adversarial robustness.

While these loss functions show good potential in terms of improving the classification performance, there has been virtually no detailed but analysis or discussion of their effect. In this work, we therefore focus on investigating trends related

TABLE I
NETWORK ARCHITECTURE ON EACH DATASET.

Dataset	Network architecture
SVHN CIFAR-10 CIFAR-100	ResNet-20, ResNet-56, WRN-28-10
Tiny ImageNet	ResNet-18, ResNet-50, WRN-28-10

to the *accuracy*, *representations*, and *explainability* of loss functions, both individually and combined with COT.

B. Explainability

Research over the past several years has explored how to visually explain the decision reasons [15]–[22]. Class activation mapping (CAM) [17] applies the feature maps output from the last convolution to global average pooling (GAP) [23] and classifies images by inputting them to a fully connected layer. Although this approach can effectively visualize attention maps for each class thanks to utilizing the response of a convolution layer and the weight at the last fully connected layer, CAM induces performance degradation. Attention branch networks (ABN) [22] achieve an excellent performance by obtaining attention maps during the training and multiplying them by the features with the attention mechanism. Gradient-weighted CAM (Grad-CAM) [15] obtains attention maps utilizing only positive gradients w.r.t. the specific class. Grad-CAM often leverages the analysis method of CNNs, as this enables attention maps from various trained models to be obtained.

In the current work, we obtain attention maps of each trained model with Grad-CAM so as to analyze the explainability of trained models with various loss functions.

III. INVESTIGATING EFFECT OF VARIOUS LOSS FUNCTIONS

In this section, we compare and evaluate the loss functions discussed in II-A from the following three viewpoints:

- Accuracy: the classification performance of the trained model with an arbitrary loss function.
- Representation: quantitative and qualitative evaluation for arbitrary feature space.
- Explainability: quantitative and qualitative evaluation of attention maps.

A. Experimental details

We use SVHN [24], CIFAR-10, CIFAR-100, and Tiny ImageNet as training datasets. SVHN is a digits classification benchmark dataset with ten classes including 73,257 images for training and 26,032 for inference, with images sized 32×32. CIFAR-10 is a natural image dataset with ten classes including 50,000 images for training and 10,000 for inference, with images sized 32×32. CIFAR-100 is the same as CIFAR-10 except for the number of classes and images assigned to each class. Tiny ImageNet is a general object recognition dataset with 200 classes including 100,000 images for training and 10,000 for inference, with images sized 64×64. We train the models using residual networks (ResNet) [25] and WideResNet [26] on these datasets. Table I lists the network architectures used for each dataset.

For all datasets and network architectures, we train for 300 epochs with a batch size of 128 (100 epochs are delegated for pre-training). The optimizer for updating the whole model utilizes stochastic gradient descent (SGD) with the weight decay of 1×10^{-4}, momentum of 0.9, and learning rate of 0.1. Centroids of center loss and PC loss updated with SGD have the learning rate and weight decay of 0.5 and 1×10^{-4}, respectively. The optimizer for complement entropy uses SGD with the weight decay of 1×10^{-4}, momentum of 0.9, and learning rate of 0.01. The learning rate of all optimizers is multiplied by $1/10$ at $\{100, 150\}$ epochs.

B. Evaluation metrics

Classification accuracy of the trained models is computed by

$$\mathrm{Acc} = \frac{1}{n'} \sum_{i=1}^{n'} \mathbf{1}[\arg\max_j p_j(\boldsymbol{x}_i) = y_i], \tag{6}$$

where n' is the amount of inference data and $\mathbf{1}[\cdot]$ is the indicator function.

We qualitatively evaluate the feature space by compressing high-dimensional features of an arbitrary layer to 2D with UMAP. For the quantitative evaluation of the feature space, we use the Calinski & Harabasz index (Cal.) [13] and Silhouette score (Sil.) [14], which can quantitatively evaluate raw features. Cal. is a metric representing the degree of condensation of the intra-class and the dispersion of the inter-class, and is computed by

$$\mathrm{Cal.} = \frac{\sum_{k=1}^{K} n'_k \times \|c_k - c\|^2}{\sum_{k=1}^{K} \sum_{i=1}^{n'_k} \|g_k(\boldsymbol{x}_i) - c_k\|} \times \frac{n' - K}{K - 1}, \tag{7}$$

where c and c_k are the centroid of the whole dataset and the centroid of class k, respectively. Sil. is a metric that can represent the separation distance of the inter-class, and is computed by

$$\mathrm{Sil.}(\boldsymbol{x}_i) = \frac{b(\boldsymbol{x}_i) - a(\boldsymbol{x}_i)}{\max\left(b(\boldsymbol{x}_i), a(\boldsymbol{x}_i)\right)}, \tag{8}$$

where $a(\boldsymbol{x}_i)$ is the average intra-class distance and $b(\boldsymbol{x}_i)$ is the average nearest-cluster distance for each feature.

Attention maps are evaluated using an insertion/deletion score [16] that gradually injects/eliminates a pixel from high attention. We compute the insertion/deletion score with the area under the curve drawn for the classification accuracy of each ratio. Insertion/deletion score are defined within $[0, 1]$, where a high/low value indicates a good performance.

C. Classification accuracy

Table II lists the classification accuracy on each dataset with various combinations of loss functions. First, focusing on the results of ResNet-20/18, we can see confirmed that combining the center loss or PC loss with cross-entropy loss achieved equal or better results than using only cross-entropy. For

TABLE II

CLASSIFICATION ACCURACY ON EACH DATASET WITH VARIOUS COMBINATIONS OF LOSS FUNCTION [%]. BOLD SYMBOLS INDICATE THE BEST PERFORMANCE. $\mathcal{L}_{xent}$ INDICATES THE MODEL WITH CROSS-ENTROPY LOSS.

	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet
	w/o COT				w/ COT			
ResNet-20/18								
$\mathcal{L}_{xent}$	97.18	95.00	76.58	59.69	97.33	95.19	76.60	59.86
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	97.27	**95.32**	**77.77**	60.07	**97.36**	95.26	**77.93**	60.28
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	**97.35**	95.21	77.50	**60.20**	97.31	**95.36**	77.89	**60.70**
ResNet-56/50								
$\mathcal{L}_{xent}$	**97.56**	**95.50**	77.48	**62.91**	97.47	**95.64**	76.60	**62.50**
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	97.51	95.20	**77.79**	60.91	**97.49**	94.78	77.83	61.03
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	97.52	95.04	77.51	60.98	97.48	94.57	**77.84**	61.11
WRN28-10								
$\mathcal{L}_{xent}$	97.33	**95.64**	78.53	**64.30**	**97.48**	95.74	78.40	**64.45**
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	97.38	95.63	79.84	63.35	97.44	95.58	79.44	63.11
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	**97.47**	95.38	**79.91**	63.51	97.42	**95.80**	**79.55**	63.95

CIFAR-100, there was an accuracy improvement of over the 1 point compared to cross-entropy loss only. We also confirmed that combining COT with each loss function resulted in higher accuracy than the results without COT.

Next, focusing on the results of ResNet-56/50, the models without COT had the highest accuracy compared to when only cross-entropy loss was used, except for CIFAR-100. In particular, using center loss or PC loss on CIFAR-10 and Tiny ImageNet significantly dropped the performance. While center loss and PC loss showed excellent performances on CIFAR-100, the results on CIFAR-10 and Tiny ImageNet showed no improvement in accuracy despite combining with COT.

Finally, for WRN28-10, the model without COT showed a significant drop in accuracy by including center loss and PC loss on Tiny ImageNet, whereas the performance on CIFAR-100 improved. The model with COT also exhibited the same trends as the results with ResNet-56/50.

These results indicate that datasets for easy classification (e.g., SVHN) did not benefit much from the loss function of feature space. In contrast, the datasets with many classes obtained excellent performances by narrowing the intra-class variance and improving the separateness of inter-class with center loss and PC loss. As for Tiny ImageNet, it achieved a sufficient performance with only cross-entropy when the model had enough capacity.

D. Representation in feature space

CIFAR-100 or Tiny ImageNet are not suitable for visualization because of the many classes and few data in each class. Therefore, we visualize the feature space compressed with UMAP on CIFAR-10. The quantitative evaluation compares all datasets because it can compute using raw features without any dimensional reduction.

First, we discuss the compressed feature space with UMAP, as shown in Fig. 2. These visualized results are the compressed features output at the last convolution layer of ResNet-20. In a nutshell, we confirmed visually that the feature space narrowed the intra-class variance by introducing center loss or PC loss. As mentioned above, there were no major differences with and without COT exhibited a slightly larger inter-class variance than the vanilla result. As shown in Fig. 2, it was difficult to qualitatively observe the effect that separates among features incorporating PC loss.

Next, we discuss the quantitative evaluation of the feature space. Focusing on the results of Sil. listed in Table III, we can see that introducing center loss and PC loss improved the score on all datasets. Center loss and PC loss scores are comparable regardless of the dataset. We confirmed that increasing the network capacity results in the same class features being gathered despite only cross-entropy loss. The results of Cal. listed in Table IV showed the same as the trends as Sil.. We presume that Cal. and Sil. are weakly correlated because both had high scores. Moreover, the scores of both Cal. and Sil. decreased with an increase in the number of classes. This phenomenon suggests that it is difficult to obtain excellent feature space on such a highly complex dataset.

Although the results of the qualitative evaluation in Fig. 2 do not indicate any major differences, the quantitative evaluation revealed significant differences. This demonstrates that utilizing both quantitative and qualitative evaluations can provide provide deep insights into the feature space, since only the 2D-plane compressed high-dimension features are difficult to determine excellent representations for.

E. Explainability for attention maps

We visualize the attention maps with Grad-CAM for the models trained on Tiny ImageNet and evaluate them qualitatively. Datasets other than Tiny ImageNet are quantitatively evaluated with only insertion/deletion scores, as the image sizes are small image sizes and the classification target is included in the center of the image.

The attention maps of ResNet-18 obtained with Grad-CAM are shown in the Fig. 3. Attention maps obtained by center loss and PC loss had greater attention to local regions than those with only cross-entropy loss. The attention maps obtained with and without COT are almost identical. Fig. 3(a) shows an image where the classification target occupies almost the entire image, and the model with cross-entropy loss widely attended to the classification target. Although center loss and PC loss could both attend to the local area of the classification target,

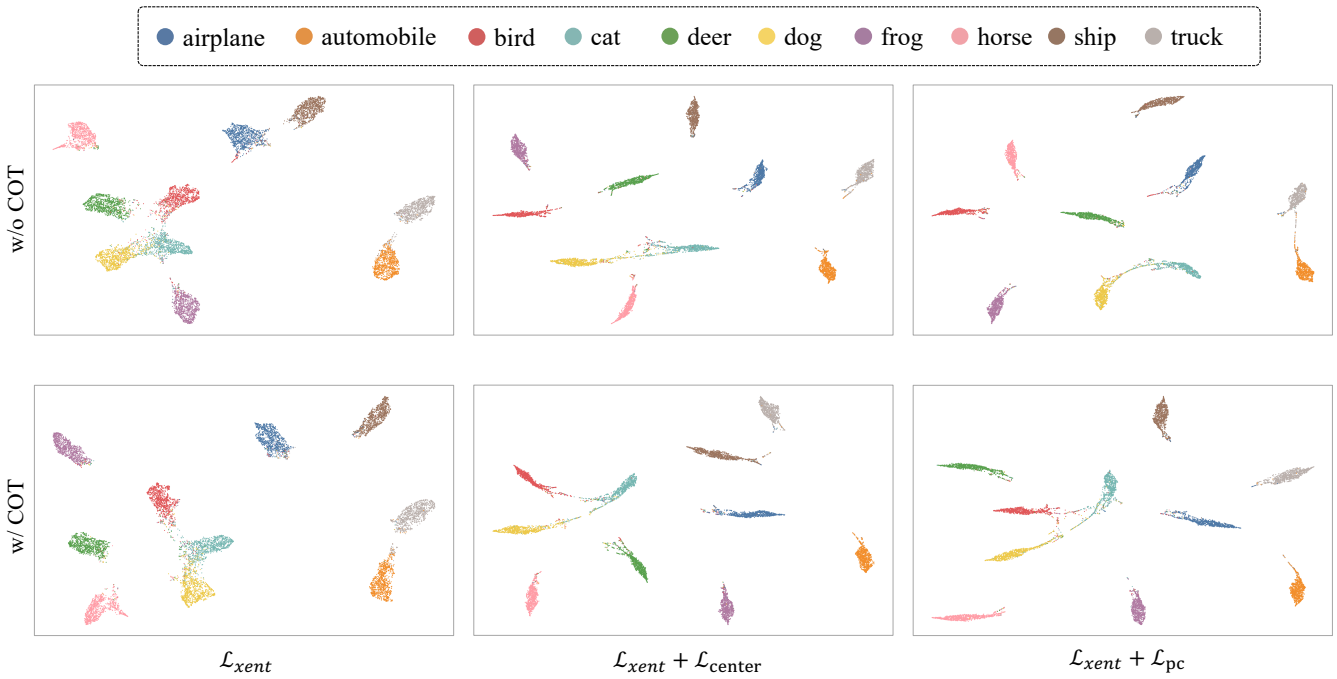

Fig. 2. Visualized feature space of each loss function on inference data of CIFAR-10. These features have been compressed to 2 dimensions with UMAP. Top and bottom examples are for feature space without and with COT, respectively.

TABLE III
QUANTITATIVE EVALUATION RESULTS OF FEATURE SPACE WITH SILHOUETTE SCORE. THIS SCORE DEFINES A RANGE OF $[-1, 1]$, AND A HIGHER VALUE IS BETTER. BOLD SYMBOLS INDICATE THE BEST RESULT.

	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet
	w/o COT				w/ COT			
ResNet-20/18								
$\mathcal{L}_{xent}$	0.544	0.377	0.063	0.003	0.684	0.465	0.065	0.004
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	0.862	**0.790**	**0.255**	**0.032**	**0.877**	**0.793**	0.242	0.032
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	**0.874**	0.788	0.254	0.026	0.868	0.792	**0.256**	**0.037**
ResNet-56/50								
$\mathcal{L}_{xent}$	0.556	0.391	0.085	0.020	0.677	0.484	0.090	0.023
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	**0.889**	0.794	**0.330**	0.036	**0.873**	**0.801**	0.294	**0.047**
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	0.887	**0.801**	0.327	**0.038**	0.858	0.787	**0.334**	0.044
WRN28-10								
$\mathcal{L}_{xent}$	0.616	0.445	0.104	0.018	0.748	0.582	0.108	0.019
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	0.862	**0.812**	0.349	**0.053**	**0.883**	0.818	**0.362**	**0.051**
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	**0.871**	0.791	**0.362**	0.052	0.866	**0.822**	0.361	0.050

they were not able to provide an adequate explanation for their decision-making process. The image in Fig. 3(b) includes an object other than the classification target, and the classification target itself is extremely small. All models trained with each loss function could perform classification correctly, especially since they all focused locally on the appropriate area by center loss or PC loss, and therefore improved explainability. Fig. 3(c)–(e) show more interesting results in that they are misclassifications or unconfident samples with only cross-entropy. For the models without COT, we can see in Fig. 3(c) that they induced misclassification due to focusing on wide regions of the image. In contrast, center loss or PC loss led to accurate classification thanks to focusing only on the characteristic regions of the classification target. The results in (d)(e) show the same trend. In Fig. 3(c), the classification target could be appropriately focused on by combining COT with PC loss.

Fig. 4 shows attention maps of ResNet-18 and WRN28-10. The attention maps with only cross-entropy loss tended to focus on the whole classification target, the same as with ResNet-18. Although center loss, PC loss, and COT all encouraged focusing on the local area, they were unable to capture the characteristics of the target. Interestingly, the classification result or confidence score of each data is adequate thanks to the correct by the benefit of large-scale network architecture, despite different attention areas to the target. Moreover, local attention maps that have inappropriate attention regions to the target induced misclassification.

The quantitative evaluation results of the attention maps are listed in Table V. As we can see, the results with

TABLE IV

QUANTITATIVE EVALUATION RESULTS OF FEATURE SPACE WITH CALINSKI & HARABASZ INDEX. THIS SCORE IS DEFINED WITH A VALUE GREATER THAN 0, AND A HIGHER VALUE IS BETTER. BOLD SYMBOLS INDICATE THE BEST RESULT.

	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet
	w/o COT				w/ COT			
ResNet-20/18								
$\mathcal{L}_{xent}$	12420.282	2666.847	72.687	22.855	22939.190	3363.283	72.644	22.855
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	53209.170	**15588.983**	226.769	**46.636**	**54516.244**	14269.112	220.540	47.130
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	**54939.280**	15218.744	**227.184**	46.568	52848.282	**14280.560**	**227.568**	**47.974**
ResNet-56/50								
$\mathcal{L}_{xent}$	12994.215	2847.820	87.742	30.635	22127.795	3617.251	88.059	31.741
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	**59083.890**	**14513.632**	**295.858**	52.465	55486.307	**12893.707**	271.807	50.010
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	57367.503	14410.582	294.515	**52.711**	**57315.224**	12698.176	**291.533**	**51.320**
WRN28-10								
$\mathcal{L}_{xent}$	17127.821	3549.024	96.0172	30.834	32332.632	5931.204	94.714	31.643
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	53761.992	**16395.141**	**321.329**	**53.119**	**55713.822**	15607.673	317.700	50.218
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	**57868.697**	16096.414	319.900	52.903	54049.729	**16046.791**	**319.667**	**50.626**

TABLE V

INSERTION/DELETION SCORES ON EACH DATASET. BETTER ATTENTION MAPS HAVE HIGHER/LOWER INSERTION/DELETION SCORES. BOLD SYMBOLS INDICATE THE BEST SCORE ON EACH DATASET.

	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet	SVHN	CIFAR-10	CIFAR-100	Tiny ImageNet
	w/o COT				w/ COT			
Insertion score								
$\mathcal{L}_{xent}$	**0.585**	**0.545**	**0.385**	**0.386**	**0.552**	**0.544**	**0.392**	**0.390**
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	0.275	0.234	0.267	0.302	0.259	0.231	0.260	0.308
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	0.339	0.233	0.263	0.298	0.300	0.227	0.272	0.314
Deletion score								
$\mathcal{L}_{xent}$	**0.258**	**0.422**	**0.267**	**0.419**	**0.238**	**0.439**	**0.265**	**0.413**
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	0.514	0.710	0.322	0.495	0.413	0.726	0.319	0.492
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	0.457	0.721	0.328	0.489	0.424	0.696	0.317	0.494
Difference between insertion score and deletion score								
$\mathcal{L}_{xent}$	**0.327**	**0.123**	**0.118**	**-0.033**	**0.314**	**0.105**	**0.127**	**-0.023**
$\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$	-0.239	-0.476	-0.055	-0.193	-0.154	-0.495	-0.059	-0.184
$\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$	-0.118	-0.488	-0.065	-0.191	-0.124	-0.469	-0.045	-0.180

only cross-entropy loss had the best performance across all results. We found that uncomplicated data (e.g., small image sizes) had significantly lowered insertion/deletion scores due to incorporating center loss or PC loss. This trend implies that their attention maps are difficult to classify and have low explainability with only these regions because of too much local attention. We conclude that both center loss and PC loss provided average attention maps of the correct class for easily classifiable data because they narrow the intra-class variance of features.

IV. DISCUSSION

The experiment in Section III clarifies that the loss function directly pulling away inter-class features (e.g., PC loss) does not significantly improve classification accuracy or feature representations. Indeed, PC loss is at risk for training collapse due to divergence of the loss, since the loss is maximized to pull apart the inter-class. We therefore trained PC loss by multiplying it by an extremely small coefficient, but unfortunately this coefficient reduced the positive effect of pulling apart the inter-class, and the PC loss thus performed as poorly as the center loss.

Moreover, we found that directly flattening incorrect class probabilities by maximizing the complement entropy (e.g., COT) resulted in no significant benefit in terms of accuracy, representations, or explainability. COT does not presuppose that incorrect class probabilities decrease, although it does aim to maximize complement entropy. In other words, COT has difficulty pulling apart inter-class variance because it simply flattens the probabilities of all incorrect classes. We therefore conclude that it is important to flatten incorrect class probabilities by considering the correct class probability (e.g., GCE [7]).

In terms of explainability, both center loss and PC loss can improve the classification accuracy or the explainability of its decision-making process for images including multiple objects or small targets. As mentioned earlier, the effects of PC loss and COT for maximizing inter-class distance are insufficient. Therefore, we expect that if it is possible to inject a powerful effect into them, we can obtain excellent attention maps capturing specific features, which is extremely important in fine-grained image classification. We leave this improvement to future work.

V. CONCLUSION

In this paper, we examined the accuracy, representations, and explainability of cross-entropy loss, center loss, PC loss, and COT. Experimental results demonstrated that the feature representations and explainability could be significantly improved by minimizing the intra-class variance. In particular,

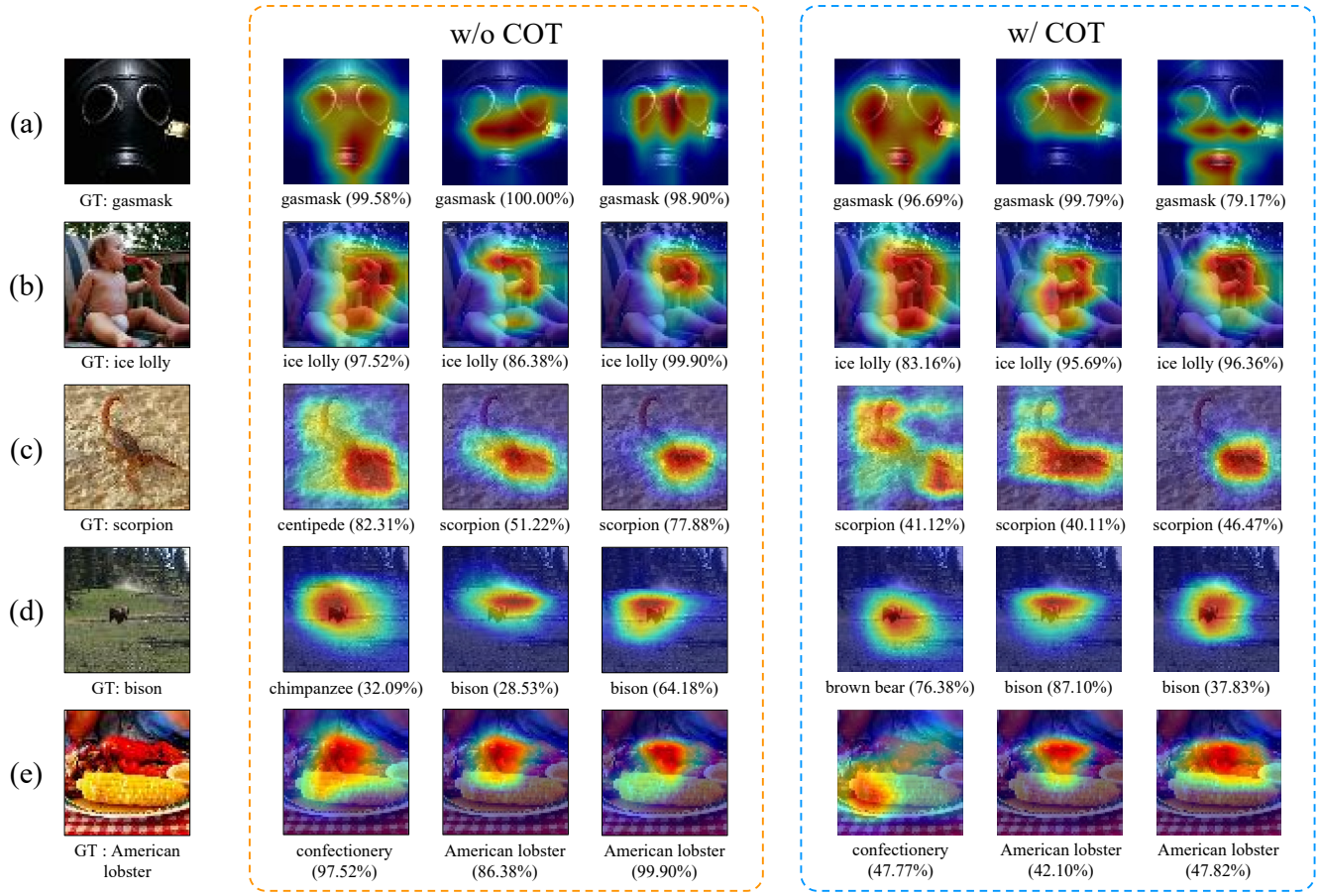

Fig. 3. Attention maps of trained ResNet-18 with each loss function on Tiny ImageNet. Left-most images are input images, and attention maps in sky blue and orange dashed lines indicate the results on models trained with and without COT, respectively. Attention maps in each group are (left to right) $\mathcal{L}_{xent}$, $\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$, and $\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$.

center loss and PC loss both lead to improved accuracy and better attention maps for images including multiple objects or small targets. However, we also observed that they degrade explainability because of too much local attention to images that include the classification target in the center. Loss functions that directly pull apart inter-class features do not perform well in the current design. Overall, our findings suggest that these loss functions are suitable for fine-grained image classification. Furthermore, we should extensively investigate more loss functions (e.g., hinge loss and binary cross-entropy loss) and network architectures (e.g., DenseNet and MobileNet) rather than only the four types of loss functions targeted in this paper. In future works, we plan to not only carry out fine-grained datasets but also reveal the nature of more loss functions and network architectures.

REFERENCES

[1] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015, pp. 84–92.

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, 2016, pp. 499–515.

[3] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 403–412.

[4] B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao, "Correcting the triplet selection bias for triplet loss," in *European Conference on Computer Vision*, 2018, pp. 71–86.

[5] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "Adversarial defense by restricting the hidden space of deep neural networks," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3384–3393.

[6] H.-Y. Chen, P.-H. Wang, C.-H. Liu, S.-C. Chang, J.-Y. Pan, Y.-T. Chen, W. Wei, and D.-C. Juan, "Complement objective training," in *International Conference on Learning Representations*, 2019, pp. 1–11.

[7] H.-Y. Chen, J.-H. Liang, S.-C. Chang, J.-Y. Pan, Y.-T. Chen, W. Wei, and D.-C. Juan, "Improving adversarial robustness via guided complement entropy," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4881–4889.

[8] H. Xuan, A. Stylianou, and R. Pless, "Improved embeddings with easy positive triplet mining," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2474–2482.

[9] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018, pp. 1–23.

[11] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[12] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold

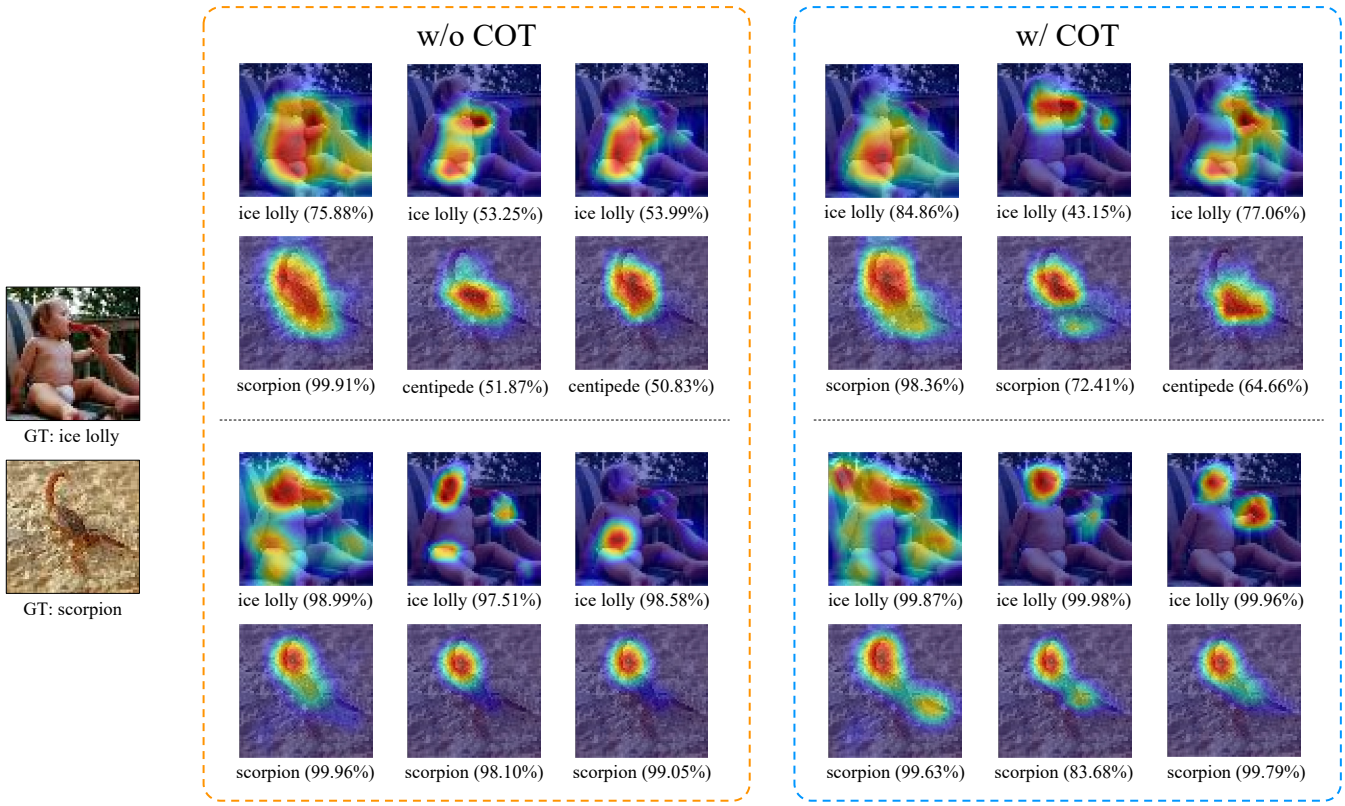

Fig. 4. Attention maps of trained models with each loss function on Tiny ImageNet. Left-most images are input images, and attention maps in sky blue and orange dashed lines indicate the results on models trained with and without COT, respectively. Attention maps in each group are (left to right) $\mathcal{L}_{xent}$, $\mathcal{L}_{xent} + \mathcal{L}_{\text{center}}$, and $\mathcal{L}_{xent} + \mathcal{L}_{\text{pc}}$. Top two examples are ResNet-50 and bottom two examples are WRN28-10.

approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426, 2018.

[13] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[16] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *British Machine Vision Conference*, 2018.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[19] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," arXiv preprint arXiv: 1706.03825, 2017.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4768–4777.

[21] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 839–847.

[22] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 705–10 714.

[23] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations*, 2014, pp. 1–10.

[24] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[26] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference*, 2016, pp. 87.1–87.12.