arXiv:1901.09054v1 [cs.LG] 25 Jan 2019

Deep Learning on Small Datasets without Pre-Training using Cosine Loss

Björn Barz [1] Joachim Denzler [1]

Abstract

Two things seem to be indisputable in the contemporary deep learning discourse: 1. The categorical cross-entropy loss after softmax activation is the method of choice for classification. 2. Training a CNN classifier from scratch on small datasets does not work well.
In contrast to this, we show that the cosine loss function provides significantly better performance than cross-entropy on datasets with only a handful of samples per class. For example, the accuracy achieved on the CUB-200-2011 dataset without pre-training is by 30% higher than with the cross-entropy loss. Further experiments on four other popular datasets confirm our findings. Moreover, we show that the classification performance can be improved further by integrating prior knowledge in the form of class hierarchies, which is straightforward with the cosine loss.

1. Introduction

Deep learning methods are well-known for their demand after huge amounts of data (Sun et al., 2017). It is even widely acknowledged that the availability of large datasets is one of the main reasons—besides more powerful hardware—for the recent renaissance of deep learning approaches (Krizhevsky et al., 2012; Sun et al., 2017). However, there are plenty of domains and applications where the amount of available training data is limited due to high costs induced by the collection or annotation of suitable data. In such scenarios, pre-training on similar tasks with large amounts of data such as the ImageNet dataset (Deng et al., 2009) has become the de facto standard (e.g., Zeiler & Fergus, 2014; Girshick et al., 2014), for example in the domain of fine-grained recognition (Lin et al., 2015; Zheng et al., 2017; Cui et al., 2018; Simon et al., 2019).

While this so-called *transfer learning* often comes without additional costs for research projects thanks to the availability of pre-trained models, it is rather problematic in at least two important scenarios: On the one hand, the target domain might be highly specialized, e.g., in the field of medical image analysis (cf. Litjens et al., 2017), inducing a large bias between the source and target domain in a transfer learning scenario. On the other hand, most large imagery datasets consist of images collected from the web, whose licenses are either unclear or prohibit commercial use (e.g., Deng et al., 2009; Krizhevsky & Hinton, 2009; Wu et al., 2019). Therefore, copyright regulations imposed by many countries make pre-training on ImageNet illegal for commercial applications. Nevertheless, the majority of research applying deep learning to small datasets focuses on transfer learning. Works aiming at directly learning from small datasets without external data are surprisingly scarce.

Certainly, the notion of a "small dataset" is highly subjective and depends on the task at hand and the diversity of the data, as expressed, for example, in the number of classes. In this work, we will consider datasets with less than 100 training images per class as small, such as the Caltech-UCSD Birds (CUB; Wah et al., 2011) dataset, which comprises at most 30 images per class. The ImageNet LSVRC 2012 dataset (Russakovsky et al., 2015), in contrast, contains between 700 and 1,300 images per class.

Since transfer learning works well in cases where sufficiently large and licensable datasets are available for pre-training, research on new methodologies for learning from small data *without external information* has been very limited. For example, the choice of categorical cross-entropy after a softmax activation as loss function has, to the best of our knowledge, not been questioned. In this work, however, we propose an extremely simple but surprisingly effective loss function for learning from scratch on small datasets: the *cosine loss*, which maximizes the cosine similarity between the output of the neural network and one-hot vectors indicating the true class. Our experiments show that this is superior to cross-entropy by a large margin on small datasets. We attribute this mainly to the L^2 normalization involved in the cosine loss, which is a strong regularizer and apparently compensates the lack of data to a remarkable degree.

In detail, our contributions are the following:

1. We conduct a study on 4 small datasets (CUB, NAB, Stanford Cars, and Oxford Flowers-102) to assess the benefits of the cosine loss for learning from small data.

[1] Computer Vision Group, Friedrich Schiller University Jena, Germany. Correspondence to: Björn Barz <bjoern.barz@uni-jena.de>.

2. We analyze the effect of the dataset size using differently sized subsets of CUB and CIFAR-100.

3. We investigate whether the integration of prior semantic knowledge about the relationships between classes as recently suggested by Barz & Denzler (2019) improves the performance further. To this end, we introduce a novel class taxonomy for the CUB dataset and also evaluate different variants to analyze the effect of the granularity of the hierarchy.

The remainder of this paper is organized as follows: We will first briefly discuss related work in Section 2. In Section 3, we introduce the cosine loss function and briefly review semantic embeddings proposed by Barz & Denzler (2019). Section 4 follows with an introduction of the datasets used for our empirical study in Section 5. A summary of our findings in Section 6 concludes this work.

2. Related Work

Learning from Small Data The problem of learning from limited data has been approached from various directions. First and foremost, there is a huge body of work in the field of *few-shot and one-shot learning*. In this area, it is often assumed to be given a set of classes with sufficient training data that is used to improve the performance on another set of classes with very few labeled examples. *Metric learning* techniques are very common in this scenario. Such methods aim at learning highly discriminative features from a large dataset that generalize well to new classes (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Wang et al., 2018; Wu et al., 2018), so that classification in face of limited training data can be performed with a simple nearest neighbor search. Another approach to few-shot learning is *meta-learning*: training a learner on large datasets to learn from small ones (Li et al., 2017; Qiao et al., 2018).

Our work is different from these few-shot learning approaches due to two reasons: First, we aim at learning a deep classifier entirely from scratch on small datasets only, without pre-training on any additional data. Secondly, our approach covers datasets with roughly between 20 and 100 samples per class, which is in the interstice between a typical few-shot scenario with even fewer samples and a classical deep learning setting with much more data.

Other approaches on learning from small datasets employ domain-specific *prior knowledge* to either artificially enlarge the amount of training data or to guide the learning. Regarding the former, Hu et al. (2018), for instance, composite face parts from different images to create new face images. Shrivastava et al. (2017) conduct training on both real images and synthetic images mapped into the domain of realistic images using a GAN for gaze and hand pose estimation. As an example for integrating knowledge into the learning process, Lake et al. (2015) represent classes of handwritten characters as probabilistic programs, which compose characters out of individual strokes and can be learned from a single example. However, this technique cannot be generalized to other types of data straightforwardly.

In contrast to all approaches mentioned above, our work focuses on learning from limited amounts of data without any external data or prior knowledge. This problem has recently also been tackled by incorporating a GAN for data augmentation into the learning process (Zhang et al., 2018). As opposed to this, we approach the problem from the perspective of the loss function, which has not been explored extensively so far for direct fully-supervised classification.

Cosine Loss The cosine loss has already successfully been used for applications other than classification. Qin et al. (2008), for example, use it for a list-wise learning to rank approach, where a vector of predicted ranking scores is compared to a vector of ground-truth scores using the cosine similarity. The cosine loss furthermore enjoys popularity in the area of cross-modal embeddings, where different representations of the same entity, such as images and text, should be close together in a joint embedding space (e.g., Sudholt & Fink, 2017; Salvador et al., 2017).

Various alternatives for the predominant cross-entropy loss have furthermore recently been explored in the field of deep metric learning, mainly in the context of face identification and verification. Liu et al. (2017), for example, extend the cross-entropy loss by enforcing a pre-defined margin between the angle of features predicted for different classes. Ranjan et al. (2017), in contrast, L^2-normalize the predicted features before applying the softmax activation and the cross-entropy loss. However, they found that doing so requires scaling the normalized features by a carefully tuned constant to achieve convergence. Wang et al. (2018) combine both approaches by normalizing both the features and the weights of the classification layer, which realizes a comparison between the predicted features and learned class-prototypes by means of the cosine similarity. They then enforce a margin between classes in angular space. The actual classification, however, is still performed using a softmax activation and supervised by the cross-entropy loss. In contrast to our work, these methods focus more on learning image representations that generalize well to novel classes (such as unseen persons) than on directly improving the classification performance on the set of classes on which the network is trained. Moreover, they introduce new hyper-parameters that must be tuned carefully to obtain satisfactory results.

In the context of content-based image retrieval, Barz & Denzler (2019) recently used the cosine loss to map images onto semantic class embeddings derived from prior knowledge encoded in a hierarchy of classes. While the focus of their work was to improve the semantic consistency of image retrieval results, they also reported classification accuracies and achieved remarkable results on the NAB dataset without pre-training. Since this was the only dataset in their experiments where this approach led to a better classification accuracy than training with cross-entropy loss, they hypothesized that prior semantic knowledge would be particularly useful for fine-grained classification tasks. In this work, we show that the reasons for this phenomenon are completely different ones and that the cosine loss can be applied to any small dataset to achieve significantly better classification accuracy than with the standard softmax and categorical cross-entropy. In contrast to the work of Barz & Denzler (2019), this does not require any prior semantic information. Instead, we will show that simply using one-hot vectors as class embeddings results in similar performance.

3. Cosine Loss

In this section, we introduce the cosine loss and briefly review the idea of hierarchy-based semantic embeddings from Barz & Denzler (2019) for combining this loss function with prior semantic knowledge.

3.1. Cosine Loss

The *cosine similarity* between two d-dimensional vectors $a, b \in \mathbb{R}^d$ is based on the angle between these two vectors and defined as

$$\sigma_{\cos}(a,b) = \cos(a \angle b) = \frac{a^\top b}{\|a\|_2 \cdot \|b\|_2} \;, \tag{1}$$

where $\|\cdot\|_p$ denotes the L^p norm.

Let $x \in \mathfrak{X}$ be a sample from some domain (e.g., images) and $y \in \mathcal{C}$ be the class label of x from the set of classes $\mathcal{C} = \{1, \dots, n\}$. Furthermore, $f_\theta : \mathfrak{X} \to \mathbb{R}^d$ denotes a transformation with learned parameters θ from the *input space* $\mathfrak{X}$ into a d-dimensional Euclidean *feature space* as realized, for instance, by a neural network. The transformations $\psi : \mathbb{R}^d \to \mathcal{P}$ and $\varphi : \mathcal{C} \to \mathcal{P}$ embed features and classes into a common *prediction space* $\mathcal{P}$, respectively. One of the simplest class embeddings, for example, consists in mapping each class to a one-hot vector:

$$\varphi_{\text{onehot}}(y) = \big[\, \underbrace{0 \cdots 0}_{y-1 \text{ times}} \quad 1 \quad \underbrace{0 \cdots 0}_{n-y \text{ times}} \,\big]^\top \;. \tag{2}$$

We consider the class embeddings φ as fixed and aim at learning the parameters θ of a neural network f_θ by maximizing the cosine similarity between the image features and the embeddings of their classes. To this end, we define the *cosine loss function* to be minimized by the neural network:

$$\mathcal{L}_{\cos}(x,y) = 1 - \sigma_{\cos}\big(f_\theta(x), \varphi(y)\big) \;. \tag{3}$$

In practice, this is implemented as a sequence of two operations. First, the features learned by the network are L^2-normalized: $\psi(x) = \frac{x}{\|x\|_2}$. This restricts the prediction space to the unit hypersphere, where the cosine similarity is equivalent to the dot product:

$$\mathcal{L}_{\cos}(x,y) = 1 - \varphi(y)^\top \psi(f_\theta(x)) \;. \tag{4}$$

3.2. Comparison with Categorical Cross-Entropy and Mean Squared Error

In the following, we discuss the differences between the proposed cosine loss and two other well-known loss functions: the categorical cross-entropy and the mean squared error (MSE). As before, we view each loss function

$$\mathcal{L}(x,y) = \delta\big(\psi(f_\theta(x)), \varphi(y)\big) \tag{5}$$

as a sequence of a transformation $\psi : \mathbb{R}^d \to \mathcal{P}$ and a dissimilarity measure $\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}^+$, which does not need to be a proper metric. The overall loss then equals the mean dissimilarity between features and labels in the prediction space. The main difference between the three loss functions discussed here lies in the type of that space.

MSE is the simplest of these loss functions, since it does not apply any transformation to the feature space and hence considers the prediction space as a Euclidean one. Naturally, the dissimilarity between the representations of samples and those of their classes in this space is measured by the squared Euclidean distance:

$$\psi(x) = x, \quad \delta(x,y) = \|x - y\|_2^2 \;. \tag{6}$$

The cosine loss introduced in the previous section restricts the prediction space to the unit hypersphere through an L^2 normalization applied to the feature space. In the resulting space, the squared Euclidean distance is equivalent to $\mathcal{L}_{\cos}$ as defined in (4), up to multiplication with a constant factor:

$$\psi(x) = \frac{x}{\|x\|_2}, \quad \delta(x,y) = 1 - x^\top y \tag{7}$$

Categorical cross-entropy is the most commonly used loss function for learning a neural classifier. As a proxy for the Kullback-Leibler divergence, it is a dissimilarity measure in the space of probability distributions, though it is not symmetric. The *softmax* activation is applied to transform the network output into this prediction space, interpreting it as the log-odds of the probability distribution over the classes. The cross-entropy between the predicted and the true class distribution is then employed as dissimilarity measure:

$$\psi(x) = \frac{\exp(x)}{\|\exp(x)\|_1}, \quad \delta(x,y) = -y^\top \log(x) \;, \tag{8}$$

where exp and log are applied element-wise.

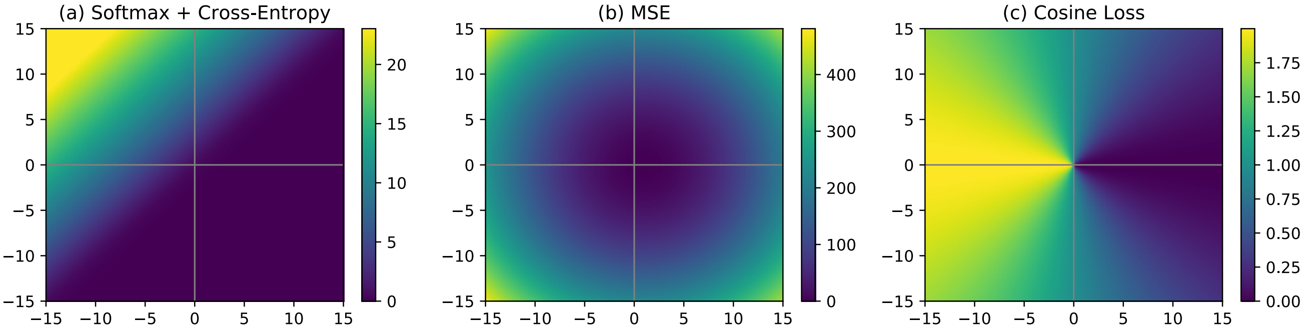

Figure 1. Heatmaps of three loss functions in a 2-dimensional feature space with fixed target $\varphi(y) = \begin{bmatrix}1 & 0\end{bmatrix}$.

A comparison of these loss functions in a 2-dimensional feature space with a fixed target $\varphi(y) = \begin{bmatrix}1 & 0\end{bmatrix}$ is shown in Fig. 1. Compared with cross-entropy and MSE, the cosine loss exhibits some distinctive properties:

1. It is bounded in the interval $[0, 2]$, while cross-entropy and MSE can take arbitrarily high values.

2. It is invariant against scaling of the feature space, since it depends only on the direction of the feature vectors, not on their magnitude.

The cross-entropy loss function, in contrast, exhibits an area of steep descent and two widespread areas of comparatively small variations. Note that the bright and dark regions in Fig. 1a are not constant but the differences are just too small for being visible. This makes the choice of a suitable initialization and learning rate schedule nontrivial. In contrast, we expect the cosine loss to behave more robustly across different datasets with varying numbers of classes thanks to the properties mentioned above.

Furthermore, the optimum value of the cross-entropy loss function is obtained when the feature value for the dimension corresponding to the true class is much larger than that for any other class and approaches infinity (Szegedy et al., 2016; He et al., 2018). This is suspected to result in overfitting (Szegedy et al., 2016), which is a particularly important problem when learning from small datasets. To mitigate this issue, Szegedy et al. (2016) proposed *label smoothing* as a regularization by adding noise to the ground-truth distribution: Instead of projecting the class labels onto one-hot vectors, the true class receives a probability of $1 - \varepsilon$ and all remaining classes are assigned $\frac{\varepsilon}{n-1}$, where ε is a small constant (e.g., 0.1). This makes the optimal network outputs finite and has been found to improve generalization slightly.

With respect to the cosine loss, on the other hand, the L^2 normalization serves as a regularizer, without the need for an additional hyper-parameter that would need to be tuned for each dataset. Furthermore, there is not only one finite optimal network output, but an entire sub-space of optimal values. This allows the training procedure to focus solely on the direction of feature vectors without being confused by Euclidean distance measures, which are problematic in high-dimensional spaces (Beyer et al., 1999). Especially in face of small datasets, we assume that this invariance against scaling of the network output is a useful regularizer.

3.3. Semantic Class Embeddings

So far, we have only considered one-hot vectors as class embeddings, distributing the classes evenly across the feature space. However, this ignores semantic relationships among classes, since some classes are more similar to each other than to other classes. To overcome this, Barz & Denzler (2019) proposed to derive class embeddings φ_{sem} on the unit hypersphere whose dot product equals the semantic similarity of the classes. The measure for this similarity is derived from an ontology such as WordNet (Fellbaum, 1998), encoding prior knowledge about the relationships among classes. They then train a CNN to map images into this semantic feature space using the cosine loss.

They have shown that the integration of this semantic information improves the semantic consistency of content-based image retrieval results significantly. With regard to classification accuracy, however, this method was only competitive with categorical cross-entropy when this was added as an additional loss function:

$$\begin{aligned}\mathcal{L}_{\text{cos+xent}}(x, y) = 1 - \psi(f_\theta(x))^\top \varphi_{\text{sem}}(y) \\ - \lambda \cdot \varphi_{\text{onehot}}(y)^\top \log(g_\theta(\psi(f_\theta(x)))) \;, \quad (9)\end{aligned}$$

where $\lambda \in \mathbb{R}^+$ is a hyper-parameter and the transformation $g_\theta : \mathbb{R}^d \to \mathbb{R}^n$ is realized by an additional fully-connected layer with softmax activation.

Besides two larger ones, Barz & Denzler (2019) also analyzed one small dataset (NAB). This was the only case where their method also provided superior classification accuracy than standard categorical cross-entropy. In the following, we apply the cosine loss with and without semantic embeddings to several small datasets to show that this effect is actually due to the loss function and not due to the prior knowledge. On the contrary, adding the cross-entropy loss to the cosine loss as in (9) is not even necessary if one-hot vectors are used instead of semantic class embeddings.

4. Datasets

We conduct experiments on four small and fine-grained image datasets as well as a larger but not fine-grained one. Statistics for all datasets can be found in Table 1.

4.1. CUB and NAB

The Caltech-UCSD Birds-200-2011 (CUB; Wah et al., 2011) and the North American Birds (NAB; Van Horn et al., 2015) datasets are quite similar. Both are fine-grained datasets of bird species, but NAB comprises four times more images than CUB and almost three times more classes. It is also even more fine-grained than CUB: While CUB distinguishes birds at the species level, NAB also separates male, female, adult, and juvenile birds into individual classes.

In contrast to CUB, the NAB datasets already provides a hierarchy of classes. To enable experiments with semantic class embeddings on CUB as well, we created a hierarchy for this dataset manually. To this end, we used information about the scientific taxonomy of bird species that is publicly available in the Wikispecies project[1]. This resulted in a hierarchy where the 200 bird species of CUB are identified by their scientific names and organized by order, sub-order, super-family, family, sub-family, and genus.

While order, family, and genus exist in all branches of the hierarchy, sub-order, super-family, and sub-family are only available for some of them. This leads to an unbalanced taxonomy tree where not all species are at the same depth (cf. Fig. 2b). To overcome this issue and analyze the effect of the depth of the hierarchy on classification accuracy, we derived two balanced variants of this original hierarchy: a flat one consisting only of the order, family, genus, and species level, (cf. Fig. 2a) and a deeper one comprising 7 levels (cf. Fig. 2c). For the latter one, we manually searched for additional information about missing super-orders, sub-orders, super-families, sub-families, and tribes in the English Wikipedia[2] and The Open Tree of Life[3].

Since CUB is a very popular dataset in the fine-grained classification community and other research projects could benefit from this class hierarchy as well, we make it publicly available at `https://github.com/cvjena/semantic-embeddings/tree/v1.1.0/CUB-Hierarchy`.

Table 1. Dataset statistics. The number of samples per class refers to training samples and numbers in parentheses specify the median.

Dataset	#Classes	#Training	#Test	Samples/Class
CUB	200	5,994	5,794	29 – 30 (30)
NAB	555	23,929	24,633	4 – 60 (44)
Cars	196	8,144	8,041	24 – 68 (42)
Flowers-102	102	2,040	6,149	20
CIFAR-100	100	50,000	10,000	500

4.2. Cars and Flowers-102

The Stanford Cars (Krause et al., 2013) and Oxford Flowers-102 (Nilsback & Zisserman, 2008) datasets are two well-known fine-grained datasets of car models and flowers, respectively. They are not particularly challenging anymore nowadays, but we include them in our experiments to avoid a bias towards bird recognition. Both datasets do not provide a class hierarchy and we will hence only conduct experiments on them in combination with one-hot class embeddings.

4.3. CIFAR-100

With 500 training images per class, the CIFAR-100 (Krizhevsky & Hinton, 2009) dataset does not fit into our definition of a *small dataset* from Section 1. We include this dataset in our experiments nevertheless for two main reasons: First, sub-sampling this dataset allows us to interpolate between small and large datasets for quantifying the effect of the number of samples per class on the gap between the performance of categorical cross-entropy and the cosine loss. Secondly, this dataset is not fine-grained and including it hence eliminates a possible bias of our study towards fine-grained datasets.

A hierarchy for the classes of the CIFAR-100 dataset derived from WordNet (Fellbaum, 1998) has recently been provided by Barz & Denzler (2019). We use this taxonomy in our experiments with semantic class embeddings.

5. Experiments

To demonstrate the performance of the cosine loss on the aforementioned small datasets, we compare it with the categorical cross-entropy loss and analyze the effect of the dataset size and prior semantic knowledge.

Code to reproduce our experiments is available at `https://github.com/cvjena/semantic-embeddings/tree/v1.1.0/CosineLoss.md`.

5.1. Setup

For CIFAR-100, we train a ResNet-110 (He et al., 2016) with an input image size of 32 × 32 and twice the number of channels per layer, as suggested by Barz & Denzler

[1] `https://species.wikimedia.org/`

[2] `https://en.wikipedia.org/`

[3] `https://tree.opentreeoflife.org/`

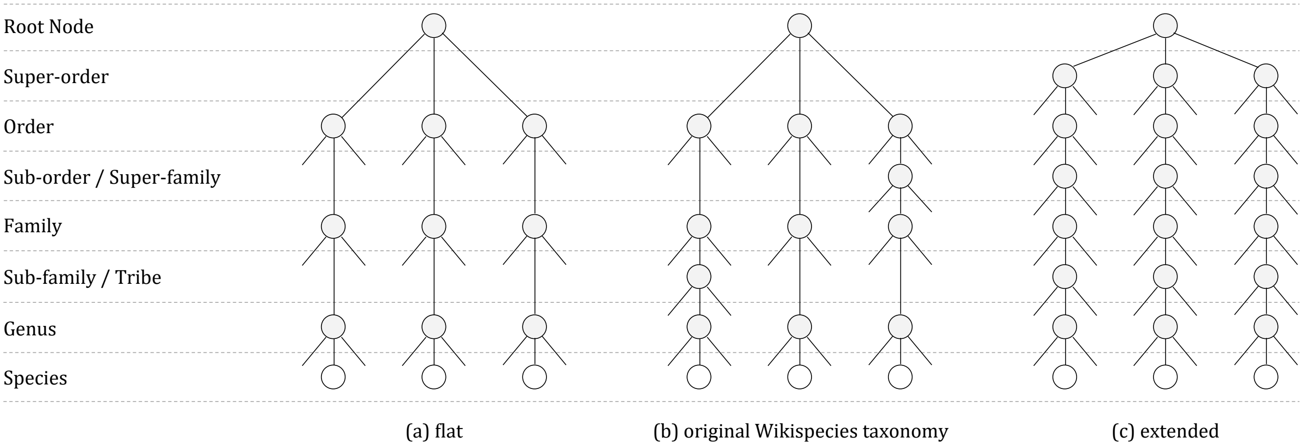

Figure 2. Schematic illustration of the different variants of the CUB hierarchy.

(2019). For all other datasets, we use a standard ResNet-50 architecture (He et al., 2016) with an input image size of 448×448, randomly cropped from images resized so that their smaller side is 512 pixels wide. As pre-processing, the input images are normalized by subtracting the mean channel value computed over the entire training set and dividing by the standard deviation. With regard to data augmentation, we use random horizontal flipping and random shifting/cropping. For all datasets except CIFAR-100, we additionally apply random erasing (Zhong et al., 2017).

For training the network, we follow the learning rate schedule of Barz & Denzler (2018): 5 cycles of Stochastic Gradient Descent with Warm Restarts (SGDR; Loshchilov & Hutter, 2017) with a base cycle length of 12 epochs. The number of epochs is doubled at the end of each cycle, amounting to a total number of 372 epochs. During each epoch, the learning rate is smoothly reduced from a pre-defined maximum learning rate $\mathrm{lr}_{\max}$ down to 10^{-6} using cosine annealing. To prevent divergence caused by initially high learning rates, we employ gradient clipping (Pascanu et al., 2013) to a maximum norm of 10. For CIFAR-100, we perform training on a single GPU with 100 images per batch. For all other datasets, we distribute a batch of 96 samples across 4 GPUs.

5.2. Performance Comparison

First, we examine the performance obtained by training with the cosine loss and investigate the additional use of prior semantic knowledge. Therefore, we report the classification accuracy of the cosine loss with semantic embeddings and with one-hot embeddings in Table 2 and compare it with the performance of standard softmax with cross-entropy. For the CUB dataset, we have used the deep variant of the class hierarchy for this experiment (cf. Fig. 2c). Other variants will be analyzed in Section 5.3.

With regard to cross-entropy, we additionally examine the use of label smoothing as described in Section 3.2. We use $\varepsilon = 0.1$ for the hyper-parameter of label smoothing, as suggested by Szegedy et al. (2016).

As an upper bound, we also report the classification accuracy achieved by fine-tuning a network pre-trained on the ILSVRC 2012 dataset (Russakovsky et al., 2015) with any of the two loss functions. We used the pre-trained weights provided by He et al. (2016) for the cross-entropy loss and by Barz & Denzler (2019) for the cosine loss.

Regarding the cosine loss, we report the performance of two variants: the cosine loss alone as in (4) and combined with the cross-entropy loss after an additional fully-connected layer as in (9). In the latter case, we fixed the combination hyper-parameter $\lambda = 0.1$, following Barz & Denzler (2019).

To avoid any bias in favor of a certain method due to the maximum learning rate $\mathrm{lr}_{\max}$, we fine-tuned it for each method individually by reporting the best results from the set $\mathrm{lr}_{\max} \in \{2.5, 1.0, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$. While the optimum $\mathrm{lr}_{\max}$ varied between 0.05 and 1.0 for the cross-entropy loss on the 4 small datasets, the cosine loss exhibited a more stable behavior and always achieved the best performance with $\mathrm{lr}_{\max} = 0.5$. Since overfitting also began at different epochs for different loss functions, we do not report the final performance after all 372 epochs in Table 2, but the best performance achieved after any epoch.

Results It can be seen that the classification accuracy obtained with the cosine loss outperforms cross-entropy after softmax significantly on all the small datasets, with the largest relative improvements being 30% and 21% on the CUB and NAB dataset. On Cars and Flowers-102, the relative improvements are 8% and 6%, but these datasets are easier in general. The label smoothing technique of Szegedy et al. (2016), on the other hand, leads to an improvement on CUB and NAB only and still falls behind the cosine loss

Table 2. Test-set classification accuracy in percent (%) achieved with different loss functions on various datasets. The best value per column that does not use external data or information is set in bold font.

	CUB	NAB	Cars	Flowers-102	CIFAR-100
softmax + cross-entropy	51.9	59.4	78.2	67.3	77.0
softmax + cross-entropy + label smoothing	55.5	68.3	78.1	66.8	**77.5**
cosine loss (one-hot embeddings)	67.6	71.7	84.3	**71.1**	75.3
cosine loss + cross-entropy (one-hot embeddings)	**68.0**	**71.9**	**85.0**	70.6	76.4
cosine loss (semantic embeddings)	59.6	72.1	—	—	74.6
cosine loss + cross-entropy (semantic embeddings)	70.4	73.8	—	—	76.7
fine-tuned softmax + cross-entropy	81.5	80.4	89.4	95.0	—
fine-tuned cosine loss (one-hot embeddings)	78.4	77.1	89.7	94.2	—
fine-tuned cosine loss + cross-entropy (one-hot embeddings)	81.3	79.3	90.3	95.1	—

by a large margin. When a sufficiently large dataset such as CIFAR-100 is used, however, the cross-entropy loss and the cosine loss perform similarly well.

Still, there is a large gap between the performance achieved by training from scratch and fine-tuning a network pre-trained on a million of images from ImageNet. Our results prove, however, that this gap can in fact be reduced.

5.3. Effect of Semantic Information

In their work on hierarchy-based image embeddings for semantic image retrieval, Barz & Denzler (2019) also observed that learning semantic embeddings using the cosine loss resulted in better classification accuracy on the NAB dataset than with cross-entropy. They attributed this phenomenon to the prior knowledge about the relationships between classes incorporated into the learning process, which they assumed to be particularly beneficial for small datasets. However, our results from the previous sub-section shown in Table 2 have falsified this hypothesis: While the classification performance improves indeed with semantic embeddings by one percent point on NAB and 3 percent points on CUB, this difference is rather small compared to the 17 percent points improvement over cross-entropy on CUB achieved by the cosine loss alone. This supports our hypothesis that the main benefit actually lies in the cosine loss itself, while the effect of prior semantic information is only complementary.

To analyze the influence of semantic information derived from class taxonomies further, we have experimented with three hierarchy variants of different depth on the CUB dataset. These have been described in detail in Section 4.1 and are illustrated in Fig. 2. The classification performance obtained with the cosine loss for each of the hierarchies can be found in Table 3.

When using one-hot embeddings, the difference between the cosine loss alone ($\mathcal{L}_{\text{cos}}$) and the cosine loss combined with cross-entropy ($\mathcal{L}_{\text{cos+xent}}$) is smallest. When the class hierarchy grows deeper, however, the performance of $\mathcal{L}_{\text{cos}}$ decreases, while the classification accuracy achieved by $\mathcal{L}_{\text{cos+xent}}$ improves.

Table 3. Classification accuracy in % on the CUB test set obtained by the cosine loss with class embeddings derived from taxonomies of varying depth. The best value per column is set in bold font.

Embedding	Hierarchy Levels	$\mathcal{L}_{\text{cos}}$	$\mathcal{L}_{\text{cos+xent}}$
one-hot	1	**67.6**	68.0
flat	4	66.6	68.8
Wikispecies	4-6	61.6	69.9
deep	7	59.9	**70.4**

We attribute this to the fact that semantic embeddings do not enforce the separability of all classes to the same extent. With one-hot embeddings, all classes are equally far apart. With semantic embeddings, however, separating similar classes is not as important as distinguishing between dissimilar ones. This is why the additional integration of the cross-entropy loss improves classification accuracy in such a scenario.

5.4. Effect of Dataset Size

To examine the behavior of the cosine loss and the cross-entropy loss depending on the size of the training dataset, we conduct experiments on various sub-sampled versions of CUB and CIFAR-100. We specify the size of a dataset in the number of samples per class and vary this number from 1 to 30 for CUB, while we use 10, 25, 50, 100, 150, 200, and 250 samples per class for CIFAR-100. For each experiment, we choose the respective number of samples from each class at random and increase the number of iterations per training epoch, so that the total number of iterations is approximately constant. The performance is always evaluated on the full test set. For CIFAR-100, we report the average over 3 runs. To facilitate comparability between experiments with different dataset sizes, we fixed the maximum learning rate $\mathrm{lr}_{\max}$

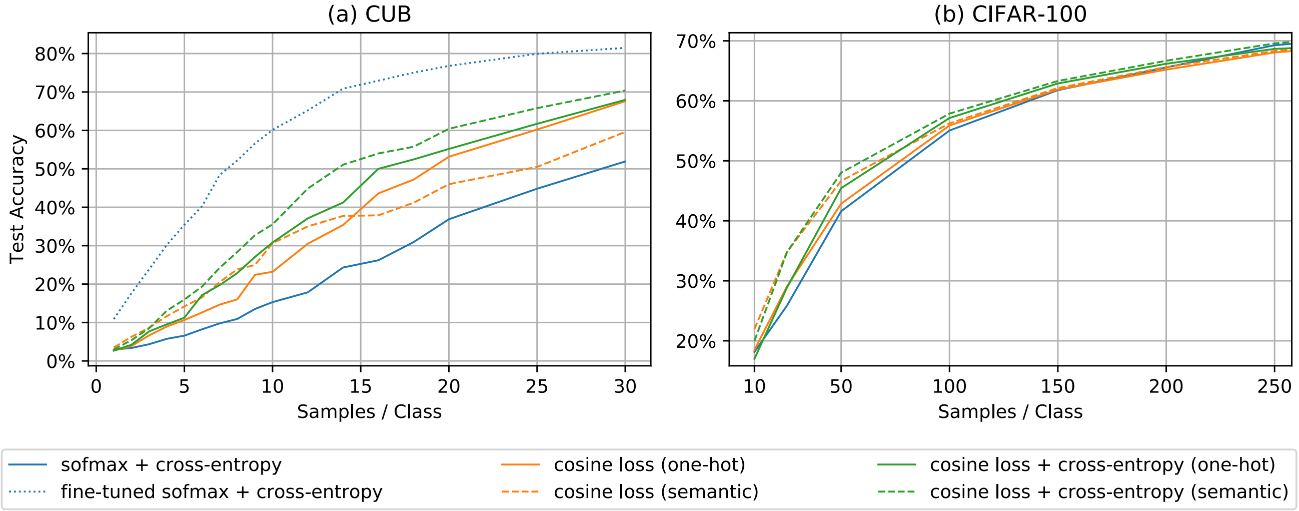

Figure 3. Classification performance depending on the dataset size.

to the best value identified for each method individually in Section 5.2.

The results depicted in Fig. 3 emphasize the benefits of the cosine loss for learning from small datasets. On CUB, the cosine loss results in consistently better classification accuracy than the cross-entropy loss and also improves faster when more samples are added. Additionally including semantic information about the relationships among classes seems to be most helpful in scenarios with very few samples. The same holds true for the combination of the cosine loss and the cross-entropy loss, but since this performed slightly better than the cosine loss alone in all cases, we would recommend this variant for practical use in general.

Nevertheless, all methods are still largely outperformed on CUB by pre-training on ILSVRC 2012. This is barely a surprise, since the network has seen 200 times more images in this case. We have argued in Section 1 why this kind of transfer learning can sometimes be problematic (e.g., domain shift, legal restrictions). In such a scenario, better methods for learning from scratch on small datasets, such as the cosine loss proposed here, are crucial.

The experiments on CIFAR-100 allow us to smoothly transition from small to larger datasets. The gap between the cosine loss and cross-entropy is smaller here, but still noticeable and consistent. It can also be seen that the cross-entropy loss starts to take over from 150–200 samples per class.

6. Conclusions

We have proposed the cosine loss for training deep neural networks for classification from scratch on limited data. Experiments on four widely used small datasets have shown that this loss functions outperforms the traditionally used categorical cross-entropy loss after softmax activation by a large margin. On the other hand, both loss functions perform similar if a sufficient amount of training data is available or the network is initialized with weights pre-trained on a large dataset.

This leads to the hypothesis, that the L^2 normalization involved in the cosine loss is a strong regularizer. Previous works have found that direction bears substantially more information in high-dimensional feature spaces than magnitude (e.g., Husain & Bober, 2017; Zhe et al., 2018). The magnitude of feature vectors can hence mainly be considered as noise, which is eliminated by L^2 normalization. Moreover, the cosine loss is bounded between 0 and 2, which facilitates a dataset-independent choice of a learning rate schedule and limits the impact of misclassified samples, e.g., difficult examples or label noise.

We have analyzed the effect of the dataset size by performing experiments on sub-sampled variants of two datasets and found the cosine loss to perform better than cross entropy for datasets with less than 200 samples per class.

Furthermore, we investigated the benefit of using semantic class embeddings instead of one-hot vectors as target values. While doing so did result in a higher classification accuracy, the improvement was rather small compared to the large gain caused by the cosine loss itself.

While some problems can in fact be solved satisfactorily by simply collecting more and more data, we hope that applications which have to deal with limited amounts of data and cannot apply pre-training can benefit from our approach. Moreover, we hope to motivate future research on different loss functions for classification, since there obviously are viable alternatives to categorical cross-entropy.

Acknowledgements

This work was supported by the German Research Foundation as part of the priority programme "Volunteered Geographic Information: Interpretation, Visualisation and Social Computing" (SPP 1894, contract number DE 735/11-1).

We also gratefully acknowledge the support of NVIDIA Corporation with the donation of Titan Xp GPUs used for this research.

References

Barz, B. and Denzler, J. Deep learning is not a matter of depth but of good training. In *International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI)*, pp. 683–687. CENPARMI, Concordia University, Montreal, 2018. ISBN 1-895193-06-0.

Barz, B. and Denzler, J. Hierarchy-based image embeddings for semantic image retrieval. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 638–647, 2019. doi: 10.1109/WACV.2019.00073.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. When is "nearest neighbor" meaningful? In *International conference on database theory*, pp. 217–235. Springer, 1999.

Cui, Y., Song, Y., Sun, C., Howard, A., and Belongie, S. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4109–4118, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

Fellbaum, C. *WordNet*. Wiley Online Library, 1998.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. *arXiv preprint arXiv:1812.01187*, 2018.

Hu, G., Peng, X., Yang, Y., Hospedales, T. M., and Verbeek, J. Frankenstein: Learning deep face representations using small data. *IEEE Transactions on Image Processing*, 27 (1):293–303, 2018.

Husain, S. S. and Bober, M. Improving large-scale image retrieval through robust aggregation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(9):1783–1796, 2017.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3D object representations for fine-grained categorization. In *IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-SGD: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Lin, T.-Y., RoyChowdhury, A., and Maji, S. Bilinear CNN models for fine-grained visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1449–1457, 2015.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. SphereFace: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 212–220, 2017.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1310–1318, 2013.

Qiao, S., Liu, C., Shen, W., and Yuille, A. L. Few-shot image recognition by predicting parameters from activations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7229–7238, 2018.

Qin, T., Zhang, X.-D., Tsai, M.-F., Wang, D.-S., Liu, T.-Y., and Li, H. Query-level loss functions for information retrieval. *Information Processing & Management*, 44(2): 838–855, 2008. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2007.07.016.

Ranjan, R., Castillo, C. D., and Chellappa, R. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., and Torralba, A. Learning cross-modal embeddings for cooking recipes and food images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3068–3076. IEEE, 2017.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2242–2251. IEEE, 2017.

Simon, M., Rodner, E., Darell, T., and Denzler, J. The whole is more than its parts? from explicit to implicit pose normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4077–4087, 2017.

Sudholt, S. and Fink, G. A. Evaluating word string embeddings and loss functions for CNN-based word spotting. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pp. 493–498. IEEE, 2017.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852. IEEE, 2017.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1199–1208, 2018.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., and Belongie, S. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 595–604, 2015.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3630–3638, 2016.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. CosFace: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265–5274, 2018.

Wu, B., Chen, W., Fan, Y., Zhang, Y., Hou, J., Huang, J., Liu, W., and Zhang, T. Tencent ML-Images: A large-scale multi-label image database for visual representation learning. *arXiv preprint arXiv:1901.01703*, 2019.

Wu, Z., Efros, A. A., and Stella, X. Y. Improving generalization via scalable neighborhood component analysis. In *European Conference on Computer Vision*, pp. 712–728. Springer, 2018.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833. Springer, 2014.

Zhang, X., Wang, Z., Liu, D., and Ling, Q. DADA: Deep adversarial data augmentation for extremely low data regime classification. *arXiv preprint arXiv:1809.00981*, 2018.

Zhe, X., Chen, S., and Yan, H. Directional statistics-based deep metric learning for image classification and retrieval. *arXiv preprint arXiv:1802.09662*, 2018.

Zheng, H., Fu, J., Mei, T., and Luo, J. Learning multi-attention convolutional neural network for fine-grained image recognition. In *IEEE International Conference on Computer Vision (ICCV)*, volume 6, 2017.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.