

Deep Metric Learning using Similarities from Nonlinear Rank Approximations

Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Klaus Jung

Visual Computing Group, HTW Berlin

Berlin, Germany

{schallk, barthel, hezel, jung}@htw-berlin.de

Abstract—In recent years, deep metric learning has achieved promising results in learning high dimensional semantic feature embeddings where the spatial relationships of the feature vectors match the visual similarities of the images. Similarity search for images is performed by determining the vectors with the smallest distances to a query vector. However, high retrieval quality does not depend on the actual distances of the feature vectors, but rather on the ranking order of the feature vectors from similar images. In this paper, we introduce a metric learning algorithm that focuses on identifying and modifying those feature vectors that most strongly affect the retrieval quality. We compute normalized approximated ranks and convert them to similarities by applying a nonlinear transfer function. These similarities are used in a newly proposed loss function that better contracts similar and disperses dissimilar samples. Experiments demonstrate significant improvement over existing deep feature embedding methods on the CUB-200-2011, Cars196, and Stanford Online Products data sets for all embedding sizes.

Index Terms—Computer vision, Image retrieval, Content-based retrieval, Feature extraction, Machine learning algorithms, Nearest neighbor searches

I. INTRODUCTION

Deep neural networks have led to significant improvements in a wide range of visual computing tasks such as image classification, segmentation and retrieval [3], [7], [20]. In traditional classification networks, the original input pixels are transformed into semantic and robust representations, describing the characteristic of an input image in order to distinguish it from different classes. However, real world visual data is very diverse and it is often not sufficient to describe detailed concepts with high level class labels.

The research field of deep metric learning tries to find a representation space, in which the visual similarity of images is encoded by the distances of respective feature vectors produced by a neural network [9], [23], [25], [28]. Simple nearest neighbor search can then be used for fine grained classification tasks as face recognition [16], [21]. Content based image retrieval is another field that greatly benefits from metric learning. High retrieval quality requires visually similar images (from the same class) to be mapped to similar feature vectors. Deep metric learning based loss functions therefore directly try to optimize the resulting embedding space by enforcing the distances between samples of the same class to be smaller than distances between entities of different classes.

A number of loss functions have been proposed to achieve this goal [6], [9], [23], [25], [28]. The main improvements over

the original ideas of Contrastive loss [6] and Triplet loss [22] have been accomplished by evaluating vector constellations with respect to the entire batch instead of pairs and triplets. This allows a more global view of the actual embedding space and helps to solve difficult local constellations, leading to improved retrieval and clustering results. All these loss functions directly work with the distances in the embedded feature vector space. However, nearest neighbor based problems, as image clustering and retrieval are evaluated by a ranking order. It is therefore not important to have very small distances to all samples of the same class, as long as all these distances are smaller than those of other classes. Instead of calculating the loss from the actual distances, we map all distances to normalized approximated ranks in the interval $[0, 1]$. These rank approximations are then converted to similarities using a nonlinear transfer function, where lower ranks lead to higher similarities. The final loss function is calculated for all batch samples with respect to their two most problematic samples of the batch: the sample from the same class with the lowest similarity and the sample from another class with the highest similarity. These steps allow to optimize each batch from a nearest neighbor perspective, which leads to well distributed embeddings. Our proposed *Nonlinear Rank Approximation (NRA)* loss shows significant improvements compared to other loss functions used for deep metric learning and achieves new state-of-the-art results on the CUB-200-2011 [27] data set. For the Cars196 [14] and Stanford Online Products [25] data sets our approach is only outperformed by an attention based ensemble approach [13] or by more powerful neural networks as used in [17].

II. RELATED WORK

There are a number of other approaches, like Boosting of Independent Embeddings (A-BIER) [19], Hard Aware Deeply Cascaded Embedding (HDC) [29], Attention Based Ensemble (ABE) learning [13] or Query Expansion [2], which can be combined with any loss function to further improve the image retrieval quality. However, in this work we focus on improving deep metric learning based loss and mining functions.

Let \mathcal{X} be the set of input data with samples $\mathbf{x} \in \mathcal{X}$. A neural network with parameters θ takes \mathbf{x} and generates the embedding $f(\cdot; \theta) : \mathcal{X} \rightarrow \mathbb{R}^d$. We omit θ for simplicity and call $f(\mathbf{x})$ the feature vector of \mathbf{x} . The L^2 distance of two feature vectors $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ is denoted by $D_{i,j} = \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2$.

Contrastive Loss [6] takes two samples $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ and tries to minimize their L^2 distance if they have the same class label and to maximize otherwise [6]. The loss function is formulated in (1), where $\tilde{y}_{i,j}$ is 1 if the images are from the same class and 0 otherwise.

$$J_{i,j} = \tilde{y}_{i,j} D_{i,j}^2 + (1 - \tilde{y}_{i,j}) \max(0, \alpha - D_{i,j})^2 \quad (1)$$

The optimal solution of such embeddings is reached, if all distances to positive images are zero and the negative distances are greater than a given margin parameter α . This is very hard to achieve and a loss formulation based on actual distances leads to bad convergence properties.

Triplet Loss [22] builds a set of three samples $\mathcal{T} = \{\mathbf{x}^a, \mathbf{x}^+, \mathbf{x}^-\}$ which consists of an anchor \mathbf{x}^a , one positive example \mathbf{x}^+ of the same class as the anchor and one negative example \mathbf{x}^- with a different class. The loss value for a single triplet is calculated by

$$J_{a,+,-} = \max(0, D_{a,+}^2 - D_{a,-}^2 + \alpha) \quad (2)$$

This method generally performs better than Contrastive Loss, since distances are not directly minimized but rather the ratio of the negative and the positive distance w.r.t the anchor feature vector $f(\mathbf{x}^a)$ is maximized here.

Lifted Structured Loss [25] was the first method to introduce *importance sampling* for building batches. However, each batch with m samples focuses on one single class only. k positive samples are chosen and the remaining $m - k$ elements are hard (difficult) negative examples. All negative batch elements \mathcal{N} are weighted by their distances to each positive entity \mathcal{P} with the exponential function and the final loss value per batch is calculated as

$$\begin{aligned} \tilde{J}_{i,j} &= \log \left(\sum_{(i,k) \in \mathcal{N}} \exp(\alpha - D_{i,k}) + \sum_{(j,l) \in \mathcal{N}} \exp(\alpha - D_{j,l}) \right) + D_{i,j} \\ J &= \frac{1}{2|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max(0, \tilde{J}_{i,j})^2 \end{aligned} \quad (3)$$

Focusing on a single or a small number of classes and their neighbors optimizes only a local region of the entire embedding space and can lead to suboptimal constellations.

N-pair Loss: To overcome this problem, [23] introduced a multi-class N-pair loss function that uses N pairs of N different classes, where N is desired to be as close as possible to the total number of classes. This sampling ensures each data point to have exactly one positive example and $N - 2$ negative ones. Similar to the Lifted Structured approach, the entire batch can then be used to calculate a loss value, but instead of the L^2 distance the dot product is used as similarity measure:

$$J = \frac{1}{N} \sum_{i=1}^N \log \left(1 + \sum_{j \neq i} \exp(f(\mathbf{x}_i)^T f(\mathbf{x}_j) - f(\mathbf{x}_i)^T f(\mathbf{x}_i^+)) \right) \quad (4)$$

Ignoring other positive data points and calculating the loss with respect to one single positive distance will always result in a low loss value, if the positive distance is already small in

the N-pair and Triplet approaches. For N-pair one batch with a large positive distance results in a much higher loss value and leads to a more informative gradient, yet with the sampling of pairs, such constellations are only found occasionally.

III. PROPOSED METHOD

Opposed to the described approaches we build batches that contain a larger number of positive examples. This allows finding samples of the same class that are poorly embedded more easily. We use $m = k \cdot n$ samples $\mathbf{x}_i \in \mathcal{X}$, $i \in \{1, 2, \dots, m\}$ of the input data per batch, where k is the number of classes and n the number of samples per class. Subsequently, we refer to this technique as *uniform n-group sampling*. In rare cases where a class does not have enough representative images, more classes are added until the desired batch size is reached. The set of positive samples \mathcal{P}_i and the set of negative samples \mathcal{N}_i w.r.t. an anchor point \mathbf{x}_i with class y_i are given by

$$\mathcal{P}_i = \{j \mid j \neq i \wedge y_j = y_i\} \quad \mathcal{N}_i = \{j \mid y_j \neq y_i\} \quad (5)$$

$\mathcal{S}_i = \mathcal{P}_i \cup \mathcal{N}_i$ is the set of all batch samples excluding the anchor. Instead of using a limited set of anchors, all samples of the batch serve as anchors. For each anchor's feature vector $f(\mathbf{x}_i)$ the distances $D_{i,j}$ to all other feature vectors of the batch elements are determined. $D_{i,\max}$ is the maximum of these distances, whereas $D_{i,\min}$ is the minimum distance to all other feature vectors of the batch w.r.t. this anchor:

$$D_{i,\max} = \max_{j \in \mathcal{S}_i} D_{i,j} \quad D_{i,\min} = \min_{j \in \mathcal{S}_i} D_{i,j} \quad (6)$$

As high retrieval quality does not depend on the actual distances and positions of the feature vectors in the embedded space, but rather on the preserved ranking order, we focus on identifying those feature vectors that most strongly affect the retrieval quality. Following the assumption that a perfect clustering is achieved if and only if all distances to negative examples are larger than the maximum distance to positive examples, we search for the two most important distances w.r.t. an anchor: The maximum distance to the embeddings of its positive examples and the minimum distance to its negative examples.

$$D_{i,\max}^+ = \max_{j \in \mathcal{P}_i} D_{i,j} \quad D_{i,\min}^- = \min_{j \in \mathcal{N}_i} D_{i,j} \quad (7)$$

Instead of using these distances directly, we use approximated normalized ranks. For each anchor $f(\mathbf{x}_i)$ the normalized ranks $r_{i,j}$ for the remaining batch elements are approximated by

$$r_{i,j} = \frac{D_{i,j} - D_{i,\min}^-}{D_{i,\max}^+ - D_{i,\min}^-} \in [0, 1] \quad (8)$$

Accordingly we determine the approximated normalized ranks for the most distant embedding of the same class $r_{i,\max}^+$ and the closest embedding of a sample of a different class $r_{i,\min}^-$

$$r_{i,\max}^+ = \frac{D_{i,\max}^+ - D_{i,\min}^-}{D_{i,\max}^+ - D_{i,\min}^-}, \quad r_{i,\min}^- = \frac{D_{i,\min}^- - D_{i,\min}^-}{D_{i,\max}^+ - D_{i,\min}^-} \quad (9)$$

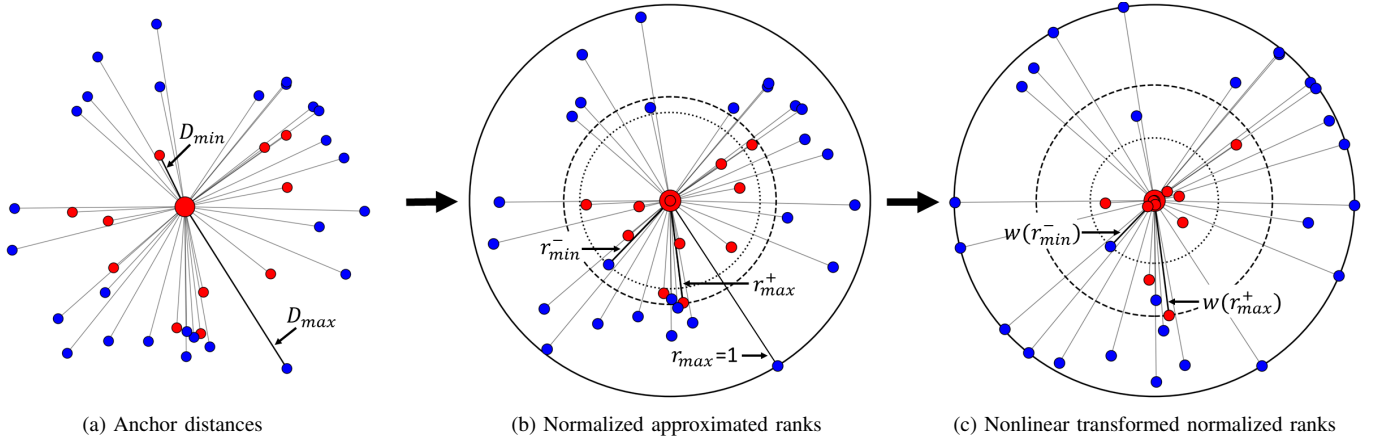


Fig. 1. (a) Example embedding for an anchor point (shown as larger dot) from one class (red) and points from other classes (blue). For the sake of clarity, the indices i, j are omitted. (b) All distances D are converted to normalized approximated ranks $r \in [0, 1]$. (c) These ranks are nonlinearly transformed $w(r)$ to determine the similarities $s = 1 - w(r)$. Perfect clustering for an anchor point is only achieved if all distances of points from other classes are greater than the maximum distance of all points of the same class. Consequently our loss function uses the two most important similarities s_{max}^+ and s_{min}^- .

Next for each anchor we transform these ranks to similarities. Instead of calculating the similarity as $s = 1 - r$ we introduce a parametric nonlinear version using a transfer function

$$w(r; \alpha) = \begin{cases} \frac{1}{2}(2r)^\alpha & r \in [0, \frac{1}{2}) \\ 1 - \frac{1}{2}(2(1-r))^\alpha & r \in [\frac{1}{2}, 1] \end{cases} \quad (10)$$

with $\alpha \in \mathbb{R}^+$. The parameter α controls the slope of $w(\cdot; \alpha)$ at $r = 0.5$. A value of $\alpha > 1$ allows to increase the influence of rank errors for the loss function which will be defined next. Figure 3 (left) shows $w(\cdot; \alpha)$ for different values of α .

Using $w(\cdot; \alpha)$ and omitting α in the following notation we define the similarity of a sample x_j to the anchor x_i by

$$s_{i,j} = 1 - w(r_{i,j}) \quad (11)$$

$$s_{i,max}^+ = 1 - w(r_{i,max}^+) \quad s_{i,min}^- = 1 - w(r_{i,min}^-) \quad (12)$$

Our proposed loss function focuses on increasing $s_{i,max}^+$ and reducing $s_{i,min}^-$ for all anchors. The loss function is given by

$$J = -\frac{1}{m} \sum_{i=1}^m \left(\log(s_{i,max}^+ + \varepsilon) + \log(1 - s_{i,min}^- + \varepsilon) \right) \quad (13)$$

Here $\varepsilon > 0$ is a small number that controls the maximum loss contribution if $s_{i,max}^+ = 0$ or $s_{i,min}^- = 1$. We use $\varepsilon = 10^{-4}$ throughout the following experiments. Figure 3 (center) shows the influence of α on the loss component $-\log(s_{i,max}^+ + \varepsilon)$.

Figure 1 schematically demonstrates the transformation of the original feature vector distances (a) for the proposed loss function. The anchor point is displayed as a larger red dot. The distances to the anchor are converted to approximated normalized ranks (b) using Eq. (8). Note that the nearest point to the anchor in (a) has a rank value of 0, plotted on top of the anchor point in (b). Finally the non-linear transformation (10) is applied to obtain (c). Samples with higher ranks are pushed further away whereas samples with lower ranks move towards the anchor. Figure 2 shows two of the 12 configurations for a toy example with three classes and four samples per class.

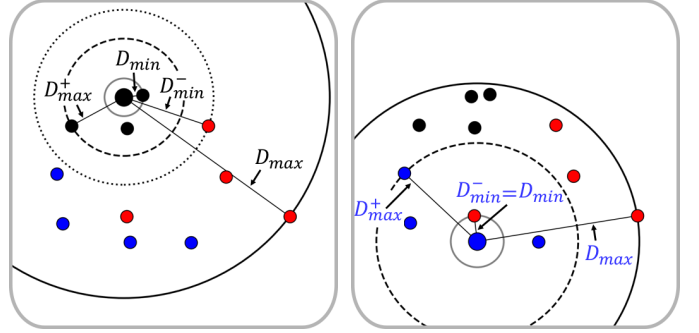


Fig. 2. A toy example of a batch size $m=12$ (3 classes with 4 samples). Again the indices i, j are omitted. As all points from the batch serve as anchors, there are 12 specific configurations, two of which are illustrated above. For each anchor point the four corresponding distances D_{min}^- , D_{min}^+ , D_{max}^- and D_{max}^+ are determined to calculate the similarities s_{max}^+ and s_{min}^- . For the left black anchor point all points of the same class lie within the radius of D_{max}^+ which represents a good embedding. For the right blue anchor point the closest point is from another class leading to a high loss. The proposed loss function focuses on reducing D_{max}^+ and increasing D_{min}^- for all anchors.

IV. EXPERIMENTS

In this section, we first determine the optimal parameter α for our transfer function (10) which will consequently be used throughout the other experiments. Next we compare the embeddings of the MNIST data set [15], generated by using different loss functions. Lastly, we strictly follow the evaluation protocol of [25] and evaluate our method on three fine grained data sets, namely CUB-200-2011 [27], Cars196 [14] and Stanford Online Products [25]. All experiments have been conducted with *TensorFlow* [1], which contains implementations of the Triplet, Lifted Structured and N-pair loss functions in the contrib-package.

A. Parameterizing the Transfer Function

The introduced family of transfer functions (10) is parameterized by α as seen in Figure 3 (left). In the simplest case

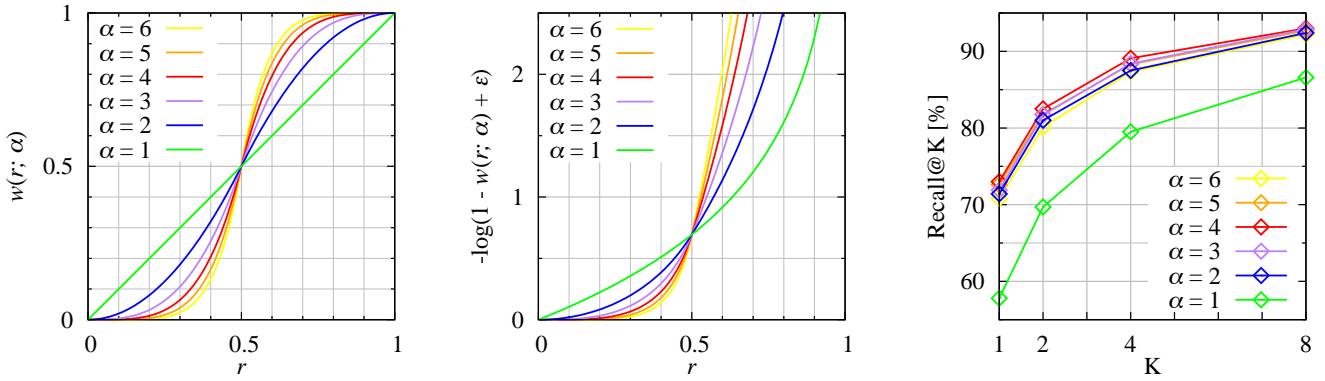


Fig. 3. Transfer function (left), the corresponding loss component (center), and Recall@K results on the Cars196 data set (right)

of $\alpha = 1$ the resulting function is a straight line $w(r; 1) = r$, therefore the use of $\alpha = 1$ is equivalent to optimizing our rank approximations directly. A higher α value leads to a lower slope in the top and bottom ranks and a steeper slope at the center. The corresponding negative log value, as used in the final loss function (8) has a lower slope for $r \in [0, \frac{1}{2}]$ and higher values for $r \in [\frac{1}{2}, 1]$. The optimal value for α is determined empirically. For this purpose we perform a series of experiments with the Cars196 data set [14] and $\alpha \in \{1, 2, \dots, 6\}$. The exact training details are identical with the ones described in the Chapter IV-C. As seen in Figure 3 (right) best results for all Recall@K values are obtained with $\alpha = 4$. We use this α value for all our following experiments.

B. Embedding of MNIST Data in 2D

To demonstrate some of the appealing properties of our approach, we utilize a simple neural network with two convolutional and one fully connected layers to embed the MNIST data set [15] in two dimensions. The network is first trained with the MNIST training set for exactly 20 000 iterations. 5 000 randomly chosen feature vectors of the test set were used for the 2D embeddings. Figure 4 shows the resulting embeddings achieved with the Softmax Cross Entropy, the Triplet with semi hard mining [21], the Lifted Structured and our proposed Nonlinear Rank Approximation loss functions. The Softmax Cross Entropy tries to divide the data into ten separable sections, while ignoring the actual distances. Hence

it achieves a relatively low mean average precision score (mAP). The margin parameter was set to 1 for the Triplet and Lifted Structured loss functions and since both perform distance based optimization, both methods produce clusters of approximately the same size and similar intra-class distances. Even though we do not directly enforce the cluster to be dense, the representations produced by our method for every class are densely packed in the representation space. In addition, the different classes were assigned to regions of varying size, making efficient use of the available space. It appears that our method takes the original variance of each class into account and aligns the embeddings accordingly while also achieving the highest mean average precision score.

C. Fine Tuning for Unseen Object Retrieval

In [25] an evaluation protocol for metric learning was introduced using three fine grained data sets: CUB-200-2011 [27], Cars196 [14] and Stanford Online Products [25]. The goal is to train with the first half of the classes and their corresponding images and to evaluate the retrieval performance with the second half.

CUB-200-2011 contains 11 788 images of 200 different bird species. The first 100 classes are chosen for training, which leads to a number of 5 864 training images. The remaining 5 924 samples from the second half of the classes are used for evaluation.

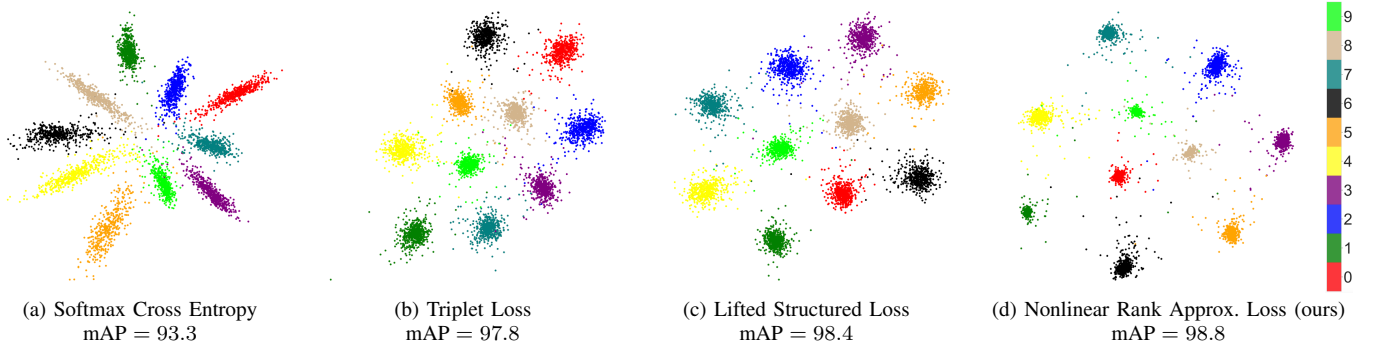


Fig. 4. Visualization of the 2D embedding space for different loss functions using MNIST data

TABLE I
RECALL@K [%] ON CUB-200-2011, CARS196 AND THE STANFORD ONLINE PRODUCTS DATA SET FOR 64 AND 512 DIMENSIONAL EMBEDDINGS.

Method	CUB-200-2011				Stanford Cars196				Stanford Online Products		
	Recall@1	Recall@2	Recall@4	Recall@8	Recall@1	Recall@2	Recall@4	Recall@8	Recall@1	Recall@10	Recall@100
Triplet	46.3 / 51.6	58.6 / 63.5	70.9 / 74.2	80.2 / 83.9	56.5 / 58.4	69.2 / 71.4	78.7 / 80.2	86.8 / 88.0	57.2 / 59.8	76.1 / 78.4	89.4 / 91.0
Lifted	45.7 / 55.7	58.0 / 67.4	70.0 / 77.7	79.4 / 85.5	48.8 / 50.7	62.3 / 63.9	72.8 / 73.7	82.3 / 83.5	61.6 / 63.8	79.9 / 81.4	91.5 / 92.7
N-Pair	51.8 / 56.4	63.8 / 69.1	74.9 / 79.1	83.8 / 86.6	63.3 / 68.3	74.4 / 78.8	82.8 / 86.1	89.4 / 91.1	63.6 / 65.4	81.5 / 83.8	92.6 / 94.0
NRA (ours)	57.6 / 64.3	69.4 / 74.8	78.8 / 83.4	87.3 / 89.9	73.0 / 81.9	82.5 / 88.8	89.1 / 92.8	93.0 / 95.4	71.9 / 75.6	86.3 / 88.8	94.3 / 95.7

Cars196: This collection of car images contains 196 classes and a total number of 16 185 images. Again, the first half of the classes are used in the fine tuning stage and the second half, for evaluation purposes. This leads to 8 054 training and 8 131 test images.

Stanford Online Products (SOP): The last set consists of 120 053 images of 22 634 different products. While the CUB-200-2011 has an average of 58.44 images per class and the Cars196 set 82.58, the Stanford Products data set has only 5.3 images per class. Again, the first half of the images is used for training and the second half for evaluation. This splits the data into 59 551 training images from 11 318 classes and 60 502 test images from 11 316 classes.

Following the evaluation protocol from [25] various metric learning papers have been published with incomparable results. Some utilize more advanced network architectures, while others report results for deviating embedding sizes. For this reason we have reproduced the experiments of important deep metric learning based loss functions using the GoogLeNet [26] network and feature vector sizes of 64 and 512 dimensions. The network was initialized with the parameters from the corresponding TensorFlow Slim Model [5] (top-1 accuracy of 69.8% on the ImageNet validation set [20]) and is fine tuned with the training images of each set. The network generates a set of 1024 7x7 feature maps at the last convolutional layer, which are average pooled to produce a 1024 dimensional embedding. This embedding is then linearly transformed to a lower dimensional space with the help of a fully connected layer without any activation function. The input images are resized to 256 pixels at their larger dimension, before a random crop of size 224 x 224 is chosen. The crop is randomly flipped horizontally for data augmentation purposes, before it is fed into the network. The retrieval quality is measured by the Recall@K metric, as proposed in [12] and is used in the evaluation protocol. This measure computes the proportion of queries in which at least one positive example is within the retrieved K nearest neighbors.

In compliance with the settings described in [25] the number of images processed per batch is 128. For N-pair the batches are built of 64 pairs in all three data sets. The remaining methods are trained with the *uniform n-group sampling* technique. As for the CUB-200-2011 and the Cars196 data sets, 16 groups of 8 samples are chosen. Since the average number of samples per class is much lower in the Stanford Online Products data set, 32 groups with 4 representatives are used in each batch. For a fair comparison, a stochastic gradient descent optimizer with momentum of 0.9 is used for all loss functions. A learning

rate of 0.0001 was chosen for CUB-200-2011, 0.0005 for Cars196 and 0.001 for SOP. The produced feature vectors have not been L^2 -normalized during the training phase, but they are L^2 -normalized for testing.

Sohn [23] has shown that their proposed N-pair batch sampling improves the results over the original triplet approach. The *uniform n-group sampling* further improves the Triplet loss results (see Table I). Since more than one positive sample per class is present, a higher number of informative triplets in the online mining process can be found. Triplet loss even outperforms Lifted Structured loss ("Lifted"), despite both using the same batching method. Our proposed Nonlinear Rank Approximation (NRA) loss function significantly outperforms the other loss functions for both dimensions and all Recall@K results.

D. Comparison to the State-of-the-Art

To put our results in context with other deep metric learning methods Table II presents an overview of previously published Recall@1 values. The middle section of the table lists other not directly comparable approaches. They use either more sophisticated networks, protocol violating test/training techniques or network ensembles. Most of these modifications are applicable with our loss function and would further improve the retrieval quality. The superscript denotes the embedding size. In [24] Song et al. claim the results in the N-pair [23] paper have been achieved by an average of ten extracted embeddings from ten random crops. The usage of such a crop averaging technique is marked with \boxplus . In some cases bounding box cropping (denoted with \square) is applied during training and testing of the CUB-200-2011 and Cars196 data set. Not all listed approaches employ the GoogLeNet architecture [26]. A ResNet50 v2 [8] with a top-1 accuracy of 75.6% on the ImageNet validation set [20] is used by Margin and InceptionBN [11] with 73.9% by Proxy-NCA [18] and Clustering [24]. Compared to the GoogLeNet, the two more advanced architectures might give a better general image retrieval performance. If we take just the loss functions trained with GoogLeNet, using no bounding boxes and no test augmentation we can see that our Nonlinear Rank Approximation (NRA) approach achieves state-of-the-art results. It also outperforms N-pairs [23] testing method and PDDM [10] bounding box training. In summary our method is only outperformed by attention based ensembles like ABE [13]. As mentioned before, HDC, A-BIER [19], ABE-8 and Proxy-NCA [18] use more complex/multiple network architectures. These networks are applicable in combination with our loss function and will lead to improved retrieval results. This is subject to future research.

TABLE II

RECALL@1 [%] RESULTS. THE METHODS IN THE SECOND SECTION APPLY MORE ADVANCED NETWORKS, TESTING OR TRAINING TECHNIQUES.

Method	network	CUB	Cars196	SOP
Contrastive ¹²⁸ [6], [25]	GoogLeNet	26.4	21.7	42.0
Triplet ⁶⁴ [22], [25]	GoogLeNet	36.1	39.1	42.1
Lifted Struct. ¹²⁸ [25]	GoogLeNet	47.2	49.0	60.8
Lifted Struct. ⁵¹² [25]	GoogLeNet	-	-	62.1
Angular Loss ⁵¹² [28]	GoogLeNet	54.7	71.4	70.9
Clustering ⁶⁴ [24]	InceptionBN	48.2	58.1	67.0
N-pair ⁶⁴ [23]	GoogLeNet	51.0	71.1	-
N-pair ⁵¹² [23]	GoogLeNet	-	-	67.7
PDDM+Quad. ¹²⁸ [10]	GoogLeNet	58.3	57.4	-
Proxy-NCA ⁶⁴ [18]	InceptionBN	49.2	73.2	73.7
HDC ³⁸⁴ [29]	GoogLeNet	53.6	73.7	69.5
HTL ⁵¹² [4]	InceptionBN	57.1	81.4	74.8
Margin ¹²⁸ [17]	ResNet50v2	63.6	79.6	72.7
A-BIER ⁵¹² [19]	GoogLeNet	57.5	82.0	74.2
ABE-8 ⁵¹² [13]	GoogLeNet	60.6	85.2	76.3
NRA (ours) ⁶⁴	GoogLeNet	57.6	73.0	71.9
NRA (ours) ⁵¹²	GoogLeNet	64.3	82.1	75.6

V. CONCLUSION

In this paper, we propose Nonlinear Rank Approximation loss (NRA) for deep metric learning, which significantly improves upon existing approaches like Triplet, Lifted Structured, and N-pair loss. Our approach uses all batch elements as anchor points. Instead of having to deal with all other batch elements for each anchor we focus on those two that mostly impact the clustering and retrieval performance: the furthest point of the same class and the closest point of a different class as the anchor. By using nonlinear transformed ranks instead of distances our proposed loss function is better able to optimize the embedding for image retrieval tasks. We demonstrate the effectiveness of the new approach on fine-grained visual recognition, as well as visual object clustering and retrieval. The NRA loss function generates very compact embeddings, therefore this new approach can also be used for dimensionality reduction and feature compression.

REFERENCES

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [2] H. K. Azad and A. Deepak. Query expansion techniques for information retrieval: a survey. *CoRR*, abs/1708.00247, 2017.
- [3] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. *CoRR*, abs/1404.1777, 2014.
- [4] W. Ge, W. Huang, D. Dong, and M. R. Scott. Deep metric learning with hierarchical triplet loss. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *ECCV (6)*, volume 11210 of *Lecture Notes in Computer Science*, pages 272–288. Springer, 2018.
- [5] S. Guadarrama and N. Silberman. Tensorflow-slim: a lightweight library for defining, training and evaluating complex models in tensorflow. Technical report, Google, 2016.
- [6] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR (2)*, pages 1735–1742. IEEE Computer Society, 2006.
- [7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, pages 630–645. Springer, 2016.

- [9] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In Y. Bengio and Y. LeCun, editors, *ICLR (Workshop)*, 2015.
- [10] C. Huang, C. C. Loy, and X. Tang. Local similarity-aware deep feature embedding. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *NIPS*, pages 1262–1270, 2016.
- [11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. R. Bach and D. M. Blei, editors, *ICML*, volume 37, pages 448–456. JMLR, 2015.
- [12] H. Jgou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, 2011.
- [13] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon. Attention-based ensemble for deep metric learning. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *ECCV (1)*, volume 11205 of *Lecture Notes in Computer Science*, pages 760–777. Springer, 2018.
- [14] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.
- [15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.
- [16] J. Lu, J. Hu, and Y.-P. Tan. Discriminative deep metric learning for face and kinship verification. *IEEE Trans. Image Processing*, 26(9):4269–4282, 2017.
- [17] R. Manmatha, C.-Y. Wu, A. J. Smola, and P. Krhenbhl. Sampling matters in deep embedding learning. In *ICCV*, pages 2859–2867. IEEE Computer Society, 2017.
- [18] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh. No fuss distance metric learning using proxies. In *ICCV*, pages 360–368. IEEE Computer Society, 2017.
- [19] M. Opitz, G. Waltner, H. Possegger, and H. Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *CoRR*, abs/1801.04815, 2018.
- [20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.
- [21] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823. IEEE Computer Society, 2015.
- [22] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *NIPS*, pages 41–48. MIT Press, 2003.
- [23] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *NIPS*, pages 1849–1857, 2016.
- [24] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy. Deep metric learning via facility location. In *CVPR*, pages 2206–2214. IEEE Computer Society, 2017.
- [25] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012. IEEE Computer Society, 2016.
- [26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9. IEEE Computer Society, 2015.
- [27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [28] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep metric learning with angular loss. In *ICCV*, pages 2612–2620. IEEE Computer Society, 2017.
- [29] Y. Yuan, K. Yang, and C. Zhang. Hard-aware deeply cascaded embedding. In *ICCV*, pages 814–823. IEEE Computer Society, 2017.