Deep Metric Learning by Online Soft Mining and Class-Aware Attention

Xinshao Wang,[1,2] Yang Hua,[1] Elyor Kodirov,[2] Guosheng Hu,[1,2] Neil M. Robertson[1,2]
[1]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK
[2]Anyvision Research Team, UK
{xwang39, y.hua, n.robertson}@qub.ac.uk, {elyor, guosheng.hu}@anyvision.co

Abstract

Deep metric learning aims to learn a deep embedding that can capture the semantic similarity of data points. Given the availability of massive training samples, deep metric learning is known to suffer from slow convergence due to a large fraction of trivial samples. Therefore, most existing methods generally resort to sample mining strategies for selecting non-trivial samples to accelerate convergence and improve performance. In this work, we identify two critical limitations of the sample mining methods, and provide solutions for both of them. First, previous mining methods assign one binary score to each sample, i.e., dropping or keeping it, so they only selects a subset of relevant samples in a mini-batch. Therefore, we propose a novel sample mining method, called Online Soft Mining (OSM), which assigns one continuous score to each sample to make use of all samples in the mini-batch. OSM learns extended manifolds that preserve useful intraclass variances by focusing on more similar positives. Second, the existing methods are easily influenced by outliers as they are generally included in the mined subset. To address this, we introduce Class-Aware Attention (CAA) that assigns little attention to abnormal data samples. Furthermore, by combining OSM and CAA, we propose a novel weighted contrastive loss to learn discriminative embeddings. Extensive experiments on two fine-grained visual categorisation datasets and two video-based person re-identification benchmarks show that our method significantly outperforms the state-of-the-art.

Introduction

With success of deep learning, deep metric learning has attracted a great deal of attention and been applied to a wide range of visual tasks such as image retrieval (Wang et al. 2014; Huang, Change Loy, and Tang 2016; Oh Song et al. 2016), face recognition and verification (Schroff, Kalenichenko, and Philbin 2015; Sun et al. 2014; Taigman et al. 2014), and person re-identification (McLaughlin, del Rincon, and Miller 2017; Song et al. 2018). The goal of deep metric learning is to learn a feature embedding/representation that captures the semantic similarity of data points in the embedding space such that samples belonging to the same classes are located closer, while samples belonging to different classes are far away. Various deep metric learning methods have been proposed: doublet-based methods with contrastive loss (Oh Song et al. 2016; Hadsell, Chopra, and LeCun 2006; Shi et al. 2016; Yuan, Yang, and Zhang 2017), triplet-based methods with triple loss (Schroff, Kalenichenko, and Philbin 2015; Wang et al. 2014; Chechik et al. 2010; Hoffer and Ailon 2015; Cui et al. 2016), and quadruplet-based methods with double-header hinge loss (Huang, Loy, and Tang 2016). All of them construct pairs (samples) by using several images and introduce relative constraints among these pairs.

Compared with classification task, deep metric learning methods deal with much more training samples since all possible combinations of images need to be considered, e.g., quadratically larger samples for contrastive loss and cubically larger samples for triplet loss. Consequently, this yields too slow convergence and poor local optima due to a large fraction of trivial samples, e.g., non-informative samples that can be easily verified and contribute zero to the loss function. To alleviate this, sample mining methods are commonly used to select nontrivial images to accelerate convergence and improve performance. Therefore, a variety of sample mining strategies have been studied recently (Schroff, Kalenichenko, and Philbin 2015; Oh Song et al. 2016; Wang and Gupta 2015; Simo-Serra et al. 2015; Huang, Loy, and Tang 2016; Yuan, Yang, and Zhang 2017; Shi et al. 2016; Cui et al. 2016).

Despite the popularity of exploiting sample mining in deep metric learning, there are two problems in the existing methods:

(1) *Not making full use of all samples in the mini-batch* – the first problem concerns with the use of samples in the mini-batch. When most methods perform sample mining, they select samples out of mini-batch, e.g., *hard* [1] difficult negatives mining (Oh Song et al. 2016). Specifically, they assign one binary score (0 or 1) to each sample in order to 'drop it' or 'keep it'. In this way, some samples which may be useful to some extent are removed and thus only a subset of samples are mined. Therefore they cannot make full use of information in each mini-batch.

(2) *Outliers* – the second problem arises from outliers.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]In this paper, 'hard' means 'not soft' as we propose soft mining (continuous mining score) in contrast with 'hard' mining (binary mining score). While in previous work, it means 'difficult'.

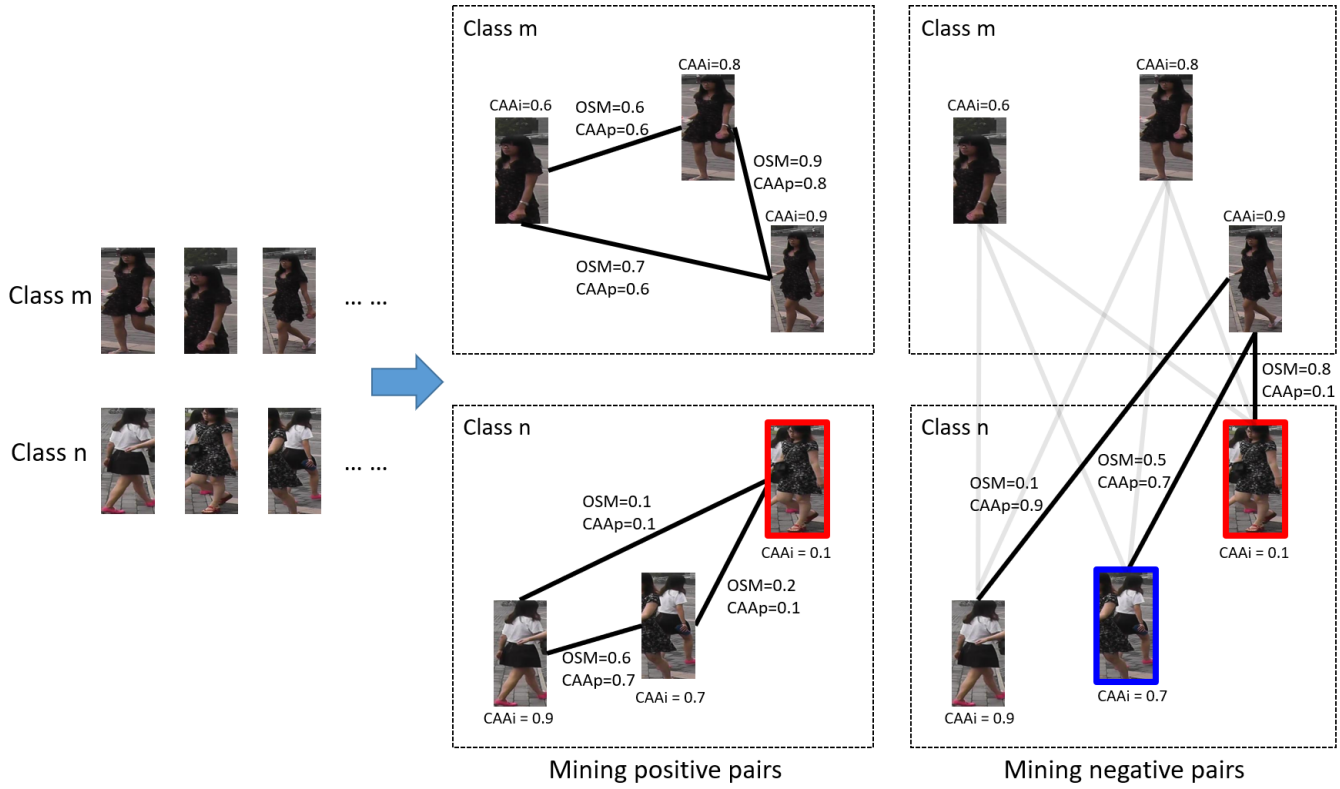


Figure 1: The illustration of Online Soft Mining (OSM) and Class-Aware Attention (CAA) for pair mining. For simplicity, we only consider the scenario with class m and class n. For mining positive pairs, OSM focuses on local positives, i.e., assigning higher scores to the pairs with smaller distance. Meanwhile, CAA helps identify outlier (with red boundary) by CAA score of images (CAA_i) and pairs (CAA_p). For mining negative pairs (only highlighting pairs of one image from class m), OSM generates higher mining scores for more difficult negatives, e.g., outlier (with red boundary) and true difficult negative (with blue boundary), then CAA helps ignore outliers in the difficult negatives mined by OSM. *Best viewed in colour.*

The existing mining methods do not consider outliers. However, the presence of noise in the training set hurts the model's performance significantly as shown in (Zhang et al. 2017). In addition, difficult negatives mining prioritizes samples with larger training loss. As a result, outliers are generally mined in the subset for training as their training losses are more likely to be large.

In this work, we aim to address both problems. Specifically, for the first problem, we propose a novel sample mining method, called Online Soft Mining (OSM), which assigns *one continuous score* to each sample to make use of all samples in the mini-batch. OSM is based on Euclidean distance in the embedding space. OSM incorporates online soft positives mining and online soft negatives mining. For the positives mining, inspired by (Cui et al. 2016), we *assign higher mining scores to more similar pairs* in terms of distance. This mining has a strong connection to manifold learning (Cui et al. 2016), since it preserves helpful intraclass variances like pose and colour in the deep embedding space. For the negatives mining, *more difficult negatives get higher mining scores* by soft negatives mining, which can be considered as *a generalisation of hard difficult negatives mining* (Oh Song et al. 2016; Wang and Gupta 2015; Simo-Serra et al. 2015; Huang, Loy, and Tang 2016; Yuan, Yang, and Zhang 2017; Cui et al. 2016). For the second problem, we propose Class-Aware Attention (CAA) that *pays less attention to outlying data samples*. CAA is simply based on class compatibility score. The illustration of OSM and CAA is given in Figure 1. One attractive property of OSM and CAA is that they are generic, by which we mean they can be integrated to most existing deep metric learning methods (Hadsell, Chopra, and LeCun 2006; Chechik et al. 2010) for mining samples and addressing outliers simultaneously. To demonstrate the effectiveness of OSM and CAA, we propose a novel loss, Weighted Contrastive Loss (WCL), by integrating them into traditional contrastive loss (Hadsell, Chopra, and LeCun 2006). The weight is assigned to each pair of images defined by OSM and CAA.

Compared to previous hard mining (0/1) in metric learning, soft mining is proposed novelly in this work. Besides, for the first time CAA is proposed to address the influence of outliers, which is an inherent problem of mining methods. Extensive experiments are conducted on two tasks: fine-grained visual categorisation and video-based person re-identification. For the first one, we demonstrate our pro-

posed method's performance by comparing with other metric learning methods on CUB-200-2011 (Wah et al. 2011) and CARS196 (Krause et al. 2013) datasets. For the second one, we use MARS (Zheng et al. 2016) and LPW (Song et al. 2018) datasets, and show that our method surpasses state-of-the-art methods by a large margin.

Related Work

Sample mining in deep metric learning. With the popularity of deep learning (Krizhevsky, Sutskever, and Hinton 2012; Szegedy et al. 2015; Girshick et al. 2016), deep metric learning has been widely studied in many visual tasks and achieved promising results (Schroff, Kalenichenko, and Philbin 2015; Oh Song et al. 2016). Compared with traditional metric learning in which hard-crafted features are widely used (Chechik et al. 2010; Weinberger and Saul 2009), deep metric learning learns feature embeddings directly from data points using deep convolutional neural networks (CNN). Although we can construct a large number of image pairs for deep metric learning, a large fraction of trivial pairs will contribute zero to the loss and gradient once the model reaches a reasonable performance. Thus it is intuitive to mine nontrivial pairs during training to achieve faster convergence and better performance. As a result, sample mining has been widely explored and a range of mining methods have been proposed (Schroff, Kalenichenko, and Philbin 2015; Wang and Gupta 2015; Simo-Serra et al. 2015; Huang, Loy, and Tang 2016; Yuan, Yang, and Zhang 2017; Shi et al. 2016; Cui et al. 2016; Oh Song et al. 2016). For mining negatives, mining difficult negatives is applied in (Wang and Gupta 2015; Simo-Serra et al. 2015; Huang, Loy, and Tang 2016; Oh Song et al. 2016). Mining semi-difficult negatives is studied in (Schroff, Kalenichenko, and Philbin 2015; Cui et al. 2016). For mining positives, mining difficult positives is explored in (Simo-Serra et al. 2015; Huang, Loy, and Tang 2016; Yuan, Yang, and Zhang 2017). Mining semi-difficult positives is explored in (Shi et al. 2016). Local positives mining (i.e., closer positives) is proposed for learning an extended manifold rather than a contracted hypersphere in (Cui et al. 2016). All of them *mine a subset of pairs hardly* by using binary score for each pair, i.e., dropping it or keeping it. Instead, our proposed OSM conducts *soft mining using continuous score*, i.e., assigning different weights for different pairs. In addition, samples with larger training loss are prioritized when difficult pair mining is applied. Therefore, *outliers are generally mined and perturb the model training* due to their large losses. To address this, semi-difficult negatives mining (Schroff, Kalenichenko, and Philbin 2015; Cui et al. 2016) and multilevel mining (Yuan, Yang, and Zhang 2017) are applied to remove outliers. Instead, we propose CAA for assigning smaller weights to outliers in the mini-batch.

In this paper, following (Sohn 2016; Oh Song et al. 2016; Tadmor et al. 2016), we construct one image pair between every two images in the mini-batch. N-pair-mc loss (Sohn 2016) and Multibatch (Tadmor et al. 2016) treat all the constructed pairs equally, i.e., without any pair mining. Lifted-Struct (Oh Song et al. 2016) mines most difficult negatives and treat all positives equally. In contrast, we propose online soft mining for both positives and negatives.

Intraclass variance. In fine-grained recognition (Oh Song et al. 2016; Wah et al. 2011; Krause et al. 2013; Cui et al. 2016) and person re-identification (Zheng et al. 2016; Zhong et al. 2017; McLaughlin, del Rincon, and Miller 2017; Song et al. 2018), intraclass distance could be larger than interclass distance, e.g., images from different categories could have similar colour and shape while the images in the same category can have large variances such as colour, pose and lighting. Several approaches benefit from utilising the inherent structure of data. In particular, local similarity-aware embedding was proposed in (Huang, Loy, and Tang 2016) to preserve local feature structures. (Cui et al. 2016) proposed to learn an extended manifold by using only local positives rather than treating all positives equally. In a similar spirit with (Cui et al. 2016), our OSM assigns larger weights for more similar positives, intending to learn continuous manifolds. The key difference, however, is that a subset composed of local positives are mined in (Cui et al. 2016), whereas our OSM makes use of all positives and gives local positives more attention.

Methodology

Given an image dataset $X = \{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i$ and y_i are i-th image and the corresponding label respectively, our aim is to learn an embedding function (metric) f that takes $\mathbf{x}_i$ as input and outputs its embedding $\mathbf{f}_i \in \mathbb{R}^D$. Ideally, the learned features of similar pairs are pulled together and features of dissimilar pairs are pushed away.

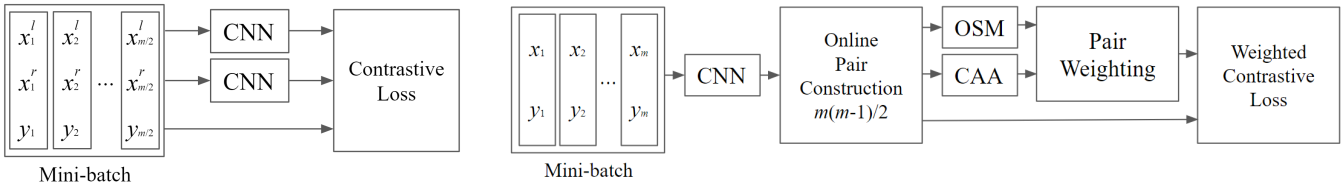
Our proposed weighted contrastive loss for learning the embedding function f is illustrated in Figure 2b. Following the common practice, each mini-batch contains $m = c \times k$ images, where c is the number of classes and k is the number of images per class. CNN is used for extracting the feature of each image in the mini-batch. In online pair construction, we use all images from the same class to construct positive pairs (i.e., similar pairs) and all images from different classes to construct negative pairs (i.e., dissimilar pairs). As a result, we get $m(m-1)/2$ image pairs in total. There are $ck(k-1)/2$ positive pairs included in the positive set, i.e., $P = \{(\mathbf{x}_i, \mathbf{x}_j) | y_i = y_j\}$. Similarly, the negative set, $N = \{(\mathbf{x}_i, \mathbf{x}_j) | y_i \neq y_j\}$, consists of $ck(ck-k)/2$ negative pairs. Then, the weight of each pair is generated by OSM and CAA jointly. Our proposed weighted contrastive loss is based on these pairs and their corresponding weights. For reference, the traditional contrastive loss (Hadsell, Chopra, and LeCun 2006) is:

$$L_{cont}^{\alpha}(\mathbf{x}^i, \mathbf{x}^j; f) = y_{ij} d_{ij}^2 + (1 - y_{ij}) max(0, \alpha - d_{ij})^2, \quad (1)$$

where $y_{ij} = 1$ if $y_i = y_j$, $y_{ij} = 0$ if $y_i \neq y_j$, $d_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|_2$ is the euclidean distance between the pair. It pulls positive pairs as close as possible and pushes negative pairs farther than a pre-defined margin α.

Online Soft Mining

Online Soft Mining contains Online Soft Positives Mining (OSPM) for the positive set and Online Soft Negatives



(a) Traditional contrastive loss

(b) Weighted contrastive loss with OSM and CAA

Figure 2: The comparison of our proposed weighted contrastive loss with traditional contrastive loss. m is the number of images in a mini-batch. Traditional contrastive loss (a) takes as input $m/2$ image pairs and $m/2$ binary labels which indicate the corresponding image pair is from the same class or not. Our proposed weighted contrastive loss (b) applies online pair construction. The weighted contrastive loss takes as input m images and m multi-class labels. We construct non-repeated $m(m-1)/2$ image pairs online rather than m^2 pairs in (Oh Song et al. 2016) or $m(m-1)$ pairs in (Tadmor et al. 2016). Weighted contrastive loss combines Online Soft Mining (OSM) and Class-Aware Attention (CAA) to assign a proper weight for each image pair.

Mining (OSNM) for the negative set.

Online Soft Positives Mining. The goal of OSPM is to generate OSM scores for pairs in the positive set. In many visual tasks, e.g., fine-grained visual categorisation, the interclass distance could be small compared to large intraclass distance. Motivated by manifold learning in (Cui et al. 2016) where *local positives are selected out for learning extended manifolds*, OSPM assigns higher OSM scores to local positives. This is because contracted hyperspheres are learned and large intraclass distance cannot be captured if positives with large distance are treated equally.

Specifically, for each similar pair in the positive set, i.e., $(\mathbf{x}_i, \mathbf{x}_j) \in P$, we compute Euclidean distance d_{ij} between their features after L_2 normalisation. As we want to assign higher mining scores to more similar pairs, we simply transfer the distance d_{ij} to OSM score using a Gaussian function with mean=0. In summary, the OSM score s_{ij}^+ for each positive pair $(\mathbf{x}_i, \mathbf{x}_j)$ is obtained as follows:

$$s_{ij}^+ = \exp(-d_{ij}^2 \big/ {\sigma_{\text{OSM}}}^2) \tag{2}$$

where $d_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|_2$, σ_{OSM} is a hyperparameter for controlling the distribution of OSM scores.

Online Soft Negatives Mining. For dissimilar pairs in the negative set N, we wish to push each pair away by a margin α inspired by the traditional contrastive loss. Similar to previous hard difficult negatives mining methods using binary mining score (Oh Song et al. 2016; Yuan, Yang, and Zhang 2017; Cui et al. 2016; Huang, Loy, and Tang 2016; Ustinova and Lempitsky 2016; Tadmor et al. 2016), to ignore a large fraction of trivial pairs which do not contribute to learning, we assign higher OSM scores to *negative pairs whose distance is within this margin*. The OSM scores of negative pairs whose distances are larger than the pre-defined margin are set as 0 because *they do not contribute to the loss and gradient*. For simplicity, the OSM score s_{ij}^- of each negative pair $(\mathbf{x}_i, \mathbf{x}_j)$ is computed directly by the margin distance:

$$s_{ij}^- = \max(0, \alpha - d_{ij}) \tag{3}$$

Class-Aware Attention

As indicated in (Schroff, Kalenichenko, and Philbin 2015; Cui et al. 2016), sample mining is easily influenced by outliers, leading the learned model to a bad local minimum. The proposed OSM is also prone to outliers, e.g., we assign more attention to more difficult negatives which may contain outliers.

In general, outliers are generally caused by corrupted labels, which hurt the model's performance significantly (Zhang et al. 2017). In other words, outliers are usually composed of mislabelled images, thus being less semantically related to their labels. Therefore, we propose to identify noisy images by their semantic relation to their labels. As it is guided by the class label, we name it Class-Aware Attention (CAA). The CAA score of image $\mathbf{x}_i$ indicates how much it is semantically related to its label y_i.

To measure image's semantic relation to its label, we compute the compatibility between image's feature vector and its corresponding class context vector. The compatibility between two vectors can be measured by their dot product (Jetley et al. 2018). We apply a classification branch after image embeddings to learn context vectors of all classes. The class context vectors are the trained parameters of the fully connected layer, i.e., $\{\mathbf{c}_k\}_{k=1}^C$, where C is the number of classes in the training set and $\mathbf{c}_k \in \mathbb{R}^D$ is the context vector of class k. Therefore, the CAA score (image's semantic relation to label) of $\mathbf{x}_i$ is computed as:

$$a_i = \frac{\exp(\mathbf{f}_i^\top \mathbf{c}_{y_i})}{\sum_{k=1}^{C} \exp(\mathbf{f}_i^\top \mathbf{c}_k)}. \tag{4}$$

The softmax operation is used for normalising image's semantic relation over all classes.

In (Goldberger and Ben-Reuven 2017), the softmax output (classification confidence/likelihood) is used to predict the true label. In this work, we apply it to estimate image's semantic relation to its label, i.e., the correctness degree of the label.

Weighted Contrastive Loss

Now we have OSM for mining proper pairs and CAA for identifying outliers. To demonstrate the effectiveness of OSM and CAA for deep metric learning, we integrate them into the traditional contrastive loss. Specifically, first we generate OSM scores for both positive and negative pairs $(\mathbf{x}_i, \mathbf{x}_j) \in P \bigcup N$, and to deal with outliers at pair level we generate CAA scores for all pairs . That is,

$$w_{ij}^{+} = s_{ij}^{+} * a_{ij}, \tag{5}$$

$$w_{ij}^{-} = s_{ij}^{-} * a_{ij}, \tag{6}$$

where a_{ij} is the CAA score of pair $(\mathbf{x}_i, \mathbf{x}_j)$ and it is defined by $\min(a_i, a_j)$.

We integrate pair weights computed by Eq. (5) and Eq. (6) into the traditional contrastive loss to formulate our weighted contrastive loss:

$$L_{WCL}(P) = \frac{1}{2} \frac{\sum\limits_{(x_i,x_j)\in P} w_{ij}^{+} {d_{ij}}^2}{\sum\limits_{(x_i,x_j)\in P} w_{ij}^{+}} \tag{7}$$

$$L_{WCL}(N) = \frac{1}{2} \frac{\sum\limits_{(x_i,x_j)\in N} w_{ij}^{-} \max(0, \alpha - d_{ij})^2}{\sum\limits_{(x_i,x_j)\in N} w_{ij}^{-}} \tag{8}$$

$$L_{WCL}(P, N) = (1-\lambda) L_{WCL}(P) + \lambda L_{WCL}(N) \tag{9}$$

where hyperparameter λ controls the contribution of the positive set and the negative set towards final contrastive loss. In our experiments, we fix $\lambda = 0.5$, treating the positive set and negative set equally. The denominators in Eq. (7) and Eq. (8) are used for normalisation. See Figure 2 for the illustration of weigted contrastive loss with comparison to the traditional contrastive loss.

In particular, the number of negative pairs $(ck(ck-k)/2)$ is much larger than the number of positive pairs $(ck(k-1)/2)$. If we normalise all the dissimilar pairs and similar pairs together, the class imbalance problem makes learning unstable. Therefore, we apply independent normalisation for the positive samples P and negative samples N.

Following (Sohn 2016; Tadmor et al. 2016), we construct an unbiased estimate of the full gradient by relying on all possible $m(m-1)/2$ pairs in each mini-batch (batch size $m \ll$ data size). The layers for online pair construction and contrastive loss computation are put behind CNN embeddings, thus each image goes through CNN only once to obtain one embedding. Each embedding vector is used to construct multiple pairs with other embeddings and compute contrastive loss.

Experiments

We compare our method with state-of-the-art embedding methods on two tasks from different domains, i.e., fine-grained visual categorisation and video-based person re-identification. For both tasks, testing and training classes are disjoint. This setting is common as the true deep metric should be class-agnostic.

Implementation details. We use GoogleNet with batch normalisation (Ioffe and Szegedy 2015) as the backbone architecture. For all datasets, the images are resized to 224×224 during training and testing. No data augmentation is applied for training and testing. For fine-grained categorisation, we set $c = 8$ and $k = 7$ for each mini-batch, while $c = 3$ and $k = 18$ for video-based person re-identification. We set empirically $\sigma_{\text{OSM}} = 0.8$ and $\sigma_{\text{CAA}} = 0.18$ for all experiments to avoid adjusting them based on the specific dataset, making it general and fair for comparision. When training a model, the weights are initialised by the pretrained model on ImageNet (Russakovsky et al. 2015). For optimisation, Stochastic Gradient Decent (SGD) is used with a learning rate of 0.001, a momentum of 0.9. The margin of weighted contrastive loss α is set to 1.2 for all experiments. We implement our method in Caffe (Jia et al. 2014) deep learning framework. We will release the source code and trained models for the ease of reproduction.

Fine-Grained Visual Categorisation

Fine-grained visual categorisation task is widely used for evaluating deep metric learning methods.

Datasets and evaluation protocol. We conduct experiments on two benchmarks:

- CUB-200-2011 contains 11,788 images of 200 species of birds. The first 100 classes (5,864 images) are used for training and the remaining classes (5,924 images) for testing.
- CARS196 is composed of 16,185 images of 196 types of cars. The first 98 classes containing 8,054 images are for training and the remaining 98 classes (8,131 images) for testing.

Both datasets come with two versions: raw images (containing large background with respect to a region of interest) and cropped images (containing only regions of interest). For fair comparison, we follow standard evaluation protocol (Oh Song et al. 2016) and report the retrieval results using Recall@K metric. For each query, the Recall@K is 1 if at least one positive in the gallery appears in its top K nearest neighbours. During testing, every image is served as a probe in turn and all the other images compose the gallery set.

Comparative evaluation. We compare our method with state-of-the-art methods in Table 1. All the compared methods use GoogleNet as the base architecture. We have the following observations:

- In both settings of datasets, i.e., raw images and cropped images, our method outperforms all the compared methods. Generally, our method improves state-of-the-art performance by around 1.5%.
- The only method that is competitive to ours is HDC (Yuan, Yang, and Zhang 2017). However, it is important to note that HDC (Yuan, Yang, and Zhang 2017) is based on cascaded embedding requiring several models, while ours is a single model.

Table 1: Comparison with state-of-the-art methods on CARS196 and CUB-200-2011 in terms of Recall@K (%). The raw images are used for training and testing for methods in the first group. The cropped images are used for training and testing for methods in the second group. * indicates cascaded models are applied for sample mining and learning embeddings.

	CARS196						CUB-200-2011					
K	1	2	4	8	16	32	1	2	4	8	16	32
Contrastive (Bell and Bala 2015)	21.7	32.3	46.1	58.9	72.2	83.4	26.4	37.7	49.8	62.3	76.4	85.3
Triplet (Schroff, Kalenichenko, and Philbin 2015)	39.1	50.4	63.3	74.5	84.1	89.8	36.1	48.6	59.3	70.0	80.2	88.4
LiftedStruct (Oh Song et al. 2016)	49.0	60.3	72.1	81.5	89.2	92.8	47.1	58.9	70.2	80.2	89.3	93.2
Binomial Deviance (Ustinova and Lempitsky 2016)	–	–	–	–	–	–	52.8	64.4	74. 7	83.9	90.4	94. 3
Histogram Loss (Ustinova and Lempitsky 2016)	–	–	–	–	–	–	50.3	61.9	72.6	82.4	88.8	93.7
Smart Mining (Harwood et al. 2017)	64.7	76.2	84.2	90.2	–	–	49.8	62.3	74.1	83.3	–	–
HDC* (Yuan, Yang, and Zhang 2017)	73.7	83.2	89.5	93.8	96.7	98.4	53.6	65.7	77.0	85.6	91.5	**95.5**
Ours	**74.0**	**83.8**	**90.2**	**94.8**	**97.3**	**98.6**	**55.3**	**67.3**	**77.5**	**85.8**	**91.8**	95.4
PDDM+Triplet (Huang, Loy, and Tang 2016)	46.4	58.2	70.3	80.1	88.6	92.6	50.9	62.1	73.2	82.5	91.1	94.4
PDDM+Quadruplet (Huang, Loy, and Tang 2016)	57.4	68.6	80.1	89.4	92.3	94.9	58.3	69.2	79.0	88.4	93.1	95.7
HDC* (Yuan, Yang, and Zhang 2017)	83.8	89.8	93.6	96.2	97.8	98.9	60.7	72.4	81.9	89.2	93.7	96.8
Ours	**85.5**	**91.5**	**95.1**	**97.2**	**98.5**	**99.2**	**62.3**	**73.2**	**83.3**	**89.6**	**94.1**	**96.9**

Table 2: Video-based person re-identification datasets: MARS and LPW. # indicates the number of corresponding terms. DT failure denotes whether there exists detection or tracking failure in the tracklets.

Datasets	#identities	#boxes	#tracklets	#cameras	DT failure	resolution	annotation	evaluation
MARS	1261	1,067,516	20,715	6	Yes	256x128	detector	CMC & mAP
LPW	2731	590,547	7694	11	No	256x128	detector	CMC

Table 3: Comparison with state-of-the-art methods on MARS in terms of CMC(%) and mAP(%). The methods using attention mechanism are indicated by 'Attention' column.

Methods	Attention	1	5	20	mAP
IDE (ResNet50) (Zhong et al. 2017)	No	62.7	–	–	44.1
IDE (ResNet50)+XQDA (Zhong et al. 2017)	No	70.5	–	–	55.1
IDE (ResNet50)+XQDA+Re-ranking (Zhong et al. 2017)	No	73.9	–	–	68.5
CNN+RNN (McLaughlin, del Rincon, and Miller 2017)	No	43.0	61.0	73.0	–
CNN+RNN+XQDA (McLaughlin, del Rincon, and Miller 2017)	No	52.0	67.0	77.0	–
AMOC+EpicFlow (Liu et al. 2017)	No	68.3	81.4	90.6	52.9
ASTPN (Xu et al. 2017)	Yes	44.0	70.0	81.0	–
SRM+TAM (Zhou et al. 2017)	Yes	70.6	90.0	**97.6**	50.7
RQEN (Song et al. 2018)	Yes	73.7	84.9	91.6	51.7
RQEN+XQDA+Re-ranking (Song et al. 2018)	Yes	77.8	88.8	94.3	71.1
DRSA (Li et al. 2018)	Yes	82.3	–	–	65.8
CAE (Chen et al. 2018)	Yes	82.4	92.9	–	67.5
Ours	Yes	**84.7**	**94.1**	97.0	**72.4**
Ours + Re-ranking(Zhong et al. 2017)	Yes	**86.0**	**94.4**	97.1	**81.0**

Table 4: Comparison with state-of-the-art methods on LPW in terms of CMC (%). The methods using attention mechanism are indicated by 'Attention' column.

Methods	Attention	1	5	20
GoogleNet(Song et al. 2018)	No	41.5	66.7	86.2
RQEN(Song et al. 2018)	Yes	57.1	81.3	91.5
Ours	Yes	**71.7**	**89.8**	**96.6**

Video-based Person Re-identification

To further evaluate the effectiveness of our method, we conduct experiments on video-based person re-identification task.

Datasets and evaluation protocol. In this task, we evaluate our methods on two large-scale benchmark datasets: MARS (Zheng et al. 2016) consists of images (frames) with huge variations due to camera setup, yielding many noisy images and outliers, and LPW (Song et al. 2018) is a cross-scene video dataset. Note that LPW is more challenging than MARS dataset, since data is collected under different scenes, whereas MARS is collected in only one scene, around university campus. The detailed information about MARS and LPW datasets are summarised in Table 2. We follow exactly the evaluation setting in (Zheng et al. 2016) and (Song et

Table 5: Ablation studies on CUB-200-2011, LPW, MARS in terms of Recall@K (%) and mAP (%).

	CUB-200-2011						LPW			MARS			
K	1	2	4	8	16	32	1	5	20	1	5	20	mAP
GoogleNet	47.8	61.1	73.2	83.2	90.4	95.4	–	–	–	–	–	–	–
Baseline	52.0	65.0	76.2	84.8	90.9	95.0	61.9	84.4	94.3	79.1	91.5	95.9	66.6
OSM	54.2	66.8	76.9	85.9	91.6	95.5	69.3	89.8	95.4	82.3	93.1	96.0	70.5
OSM+CAA	**55.3**	**67.3**	**77.5**	**85.8**	**91.8**	**95.4**	**71.7**	**89.8**	**96.6**	**84.7**	**94.1**	**97.0**	**72.4**

al. 2018), respectively. We report the Cumulated Matching Characteristics (CMC)[2] results for both datasets. We also compute the mean average precision (mAP) for MARS.

Comparative evaluation. The results of MARS and LPW are shown in Table 3 and Table 4 respectively. The results on both datasets demonstrate the effectiveness of our proposed approach. Our observations are as follows:

- On MARS dataset, our method outperforms all the methods by a large margin. Compared with CAE (Chen et al. 2018)[3], ours improves 2.3% and 4.9% in terms of CMC-1 and mAP respectively; When using re-ranking as post-processing in the test phase, the performance of our method further increases.
- On LPW dataset, we can see that our method again outperforms state-of-the-art methods by a large margin.

Ablation Study

Our key contributions in this paper are OSM and CAA. To evaluate the contribution of each module, we conduct ablation studies on CUB-200-2011, LPW, and MARS datasets. The details are as follows:

- Baseline: Neither OSM or CAA is used. Online pair construction is applied. The weight of each pair is fixed as one in the loss functions Eq. (7) and Eq. (8). Specifically, $w_{ij}^{+} = 1$ and $w_{ij}^{-} = 1$.
- OSM: When mining negatives, OSM generates higher scores for difficult negatives and lower scores for trivial negative samples. When mining positives, OSM generates higher scores for local positives to preserves intraclass variances. CAA is not used. Specifically, $w_{ij}^{+} = s_{ij}^{+}$ and $w_{ij}^{-} = s_{ij}^{-}$.
- OSM+CAA: As mentioned before, OSM is prone to outliers similar to other mining methods, which is an inherent problem of mining approaches. Therefore, we combine OSM with CAA to conduct soft mining as well as address outliers. Specifically, $w_{ij}^{+} = s_{ij}^{+} * a_{ij}$ and $w_{ij}^{-} = s_{ij}^{-} * a_{ij}$.

The results are summarized in Table 5. We can see that OSM alone can achieve very competitive performance compared with the baseline. After integrating CCA to OSM, the accuracy increases on all datasets by about 1-2%. It is easy to see that both OSM and CCA contribute positively toward the final performance.

Furthermore, we notice that test set classes of CUB-200-2011 overlaps with ImageNet (Russakovsky et al. 2015) dataset. To verify how much actually we improve, we show the performance of pre-trained GoogleNet (Ioffe and Szegedy 2015) without any fine-tuning. The result is shown in the first row of Table 5. Just using GoogleNet with pre-trained weights, the performance is 47.8% – already beating most of the methods shown in Table 1. In contrast, our method outperforms pre-trained GoogleNet by a large margin.

Conclusion

In this work, we propose a simple yet effective mining method named OSM. Importantly, OSM not only mines nontrivial sample pairs for accelerating convergence, but also learns extended manifolds for preserving intraclass variances. Besides, we propose CAA to address the impact of outliers. Finally, to evaluate the effectiveness of these two modules in action, we propose weighted contrastive loss by combining OSM and CAA to learn discriminative embeddings. The experiments on fine-grained visual categorisation and video-based person re-identification with four datasets demonstrate the superiority of our method. For future work, it is interesting to see the integration of OSM and CAA for other kinds of losses such as triplet loss.

References

Bell, S., and Bala, K. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics* 98.

Chechik, G.; Sharma, V.; Shalit, U.; and Bengio, S. 2010. Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research* 1109–1135.

Chen, D.; Li, H.; Xiao, T.; Yi, S.; and Wang, X. 2018. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*.

Cui, Y.; Zhou, F.; Lin, Y.; and Belongie, S. 2016. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *CVPR*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2016. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 142–158.

[2]CMC@K is the same as Recall@K. Generally, the term CMC is used in person re-identification.

[3]For CAE, we present the results of complete sequence instead of multiple snippets so that it can be compared with other methods. Multiple snippets can be regarded as data augmentation in the test phase.

Goldberger, J., and Ben-Reuven, E. 2017. Training deep neural-networks using a noise adaptation layer. *ICLR*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

Harwood, B.; Kumar, B.; Carneiro, G.; Reid, I.; Drummond, T.; et al. 2017. Smart mining for deep metric learning. In *ICCV*.

Hoffer, E., and Ailon, N. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*.

Huang, C.; Change Loy, C.; and Tang, X. 2016. Unsupervised learning of discriminative attributes and visual representations. In *CVPR*.

Huang, C.; Loy, C. C.; and Tang, X. 2016. Local similarity-aware deep feature embedding. In *NIPS*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Jetley, S.; Lord, N. A.; Lee, N.; and Torr, P. H. 2018. Learn to pay attention. *ICLR*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACMMM*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV Workshop*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Li, S.; Bak, S.; Carr, P.; and Wang, X. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*.

Liu, H.; Jie, Z.; Jayashree, K.; Qi, M.; Jiang, J.; Yan, S.; and Feng, J. 2017. Video-based person re-identification with accumulative motion context. *IEEE Transactions on Circuits and Systems for Video Technology*.

McLaughlin, N.; del Rincon, J. M.; and Miller, P. 2017. Video person re-identification for wide area tracking based on recurrent neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 211–252.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Shi, H.; Yang, Y.; Zhu, X.; Liao, S.; Lei, Z.; Zheng, W.; and Li, S. Z. 2016. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*.

Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; and Moreno-Noguer, F. 2015. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*.

Song, G.; Leng, B.; Liu, Y.; Hetang, C.; and Cai, S. 2018. Region-based quality estimation network for large-scale person re-identification. In *AAAI*.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *NIPS*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*.

Tadmor, O.; Rosenwein, T.; Shalev-Shwartz, S.; Wexler, Y.; and Shashua, A. 2016. Learning a metric embedding for face recognition using the multibatch method. In *NIPS*.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*.

Ustinova, E., and Lempitsky, V. 2016. Learning deep embeddings with histogram loss. In *NIPS*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, X., and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *ICCV*.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research* 207–244.

Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; and Zhou, P. 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *CVPR*.

Yuan, Y.; Yang, K.; and Zhang, C. 2017. Hard-aware deeply cascaded embedding. In *ICCV*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *ECCV*.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*.

Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; and Tan, T. 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*.