Pattern Recognition Letters 131 (2020) 8–14

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Variance-preserving deep metric learning for content-based image retrieval

Nikolaos Passalis[a,c,*], Alexandros Iosifidis[b], Moncef Gabbouj[a], Anastasios Tefas[c]

[a] Faculty of Information Technology and Communication, Tampere University, Finland
[b] Department of Engineering, Aarhus University, Denmark
[c] Department of Informatics, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 8 September 2019
Revised 23 November 2019
Accepted 30 November 2019
Available online 4 December 2019

Keywords:
Metric learning
Content-based information retrieval
Deep learning

ABSTRACT

Supervised deep metric learning led to spectacular results for several Content-based Information Retrieval (CBIR) applications. The success of these approaches slowly led to the belief that image retrieval and classification are just slightly different variations of the same problem. However, recent evidence suggests that learning highly discriminative representation for a (limited) set of training classes removes valuable information from the representation, potentially harming both the in-domain, as well as the out-of-domain retrieval precision. In this paper, we propose a regularized discriminative deep metric learning method that aims to not only learn a representation that allows for discriminating between different classes, but it is also capable of encoding the latent generative factors separately for each class, overcoming this limitation. This allows for modeling the in-class variance and, as a result, maintaining the ability to represent both sub-classes of the in-domain data, as well as objects that belong to classes outside the training domain. The effectiveness of the proposed method, over existing supervised and unsupervised representation/metric learning approaches, is demonstrated under different in-domain and out-of-domain setups and three challenging image datasets.

1. Introduction

Deep metric learning is at the cornerstone of most state-of-the-art Content-based Information Retrieval (CBIR) systems, covering several aspects of their operation, ranging from metric learning for image retrieval [24,38,39,47], and supervised and semi-supervised hashing methods [4,21,40,41,53], to multi-view representation learning [1,14,25,32]. Initially, representation/metric learning for information retrieval tasks was handled in an unsupervised way, e.g., by employing reconstruction-based objectives [12], since it was especially difficult to acquire and effectively use annotations for a wide range of different information needs. However, with the recent success of deep learning in supervised learning tasks, along with the availability of larger annotated datasets [2], the interest has gradually shifted into exploring whether supervised objectives, which directly employ class/label annotations, can be effectively used for metric learning for information retrieval purposes [18,39,52]. Indeed, supervised metric learning greatly outperformed the previously used methods in a wide range of settings and setups, ranging from hashing [40,41] to re-identification [5,23]. In fact, the success of these approaches was so spectacular that slowly led to the belief that image retrieval and classification are probably just slightly different variations of the same problem [49].

The aforementioned *classification-retrieval equivalence* assumption has been recently challenged in [33], suggesting that directly optimizing representations for a limited number of classes is not optimal for retrieval tasks, since it ignores that users are actually interested for both the *in-class* variations, as well as for retrieving objects of *novel* classes that are potentially not seen during the training. Learning a highly discriminative representation can indeed lead to excellent results for in-domain queries, i.e., queries from the same classes as the ones used during the optimization. In fact, just using a strong classifier to extract a binary representation works extremely well for this kind of tasks [33]. However, these highly discriminative representations are usually not suitable for retrieving objects that belong to classes that were not seen during the training process [28,33], since most discriminative objectives rely on reducing the in-class variance, throwing away potentially useful information for other (possibly related) information needs. This behavior was also confirmed in the experiments conducted in this paper. These observations lead us to the main research

* Corresponding author.
E-mail address: nikolaos.passalis@tuni.fi (N. Passalis).

https://doi.org/10.1016/j.patrec.2019.11.041
0167-8655/

question of this study: *Is it possible to learn a discriminative representation that will also maintain the in-class variance instead of minimizing it?*. Learning a representation with these properties will allow for encoding both the in-class variations, as well as for better maintaining the ability to represent objects of classes not seen during the training process.

In this paper, we propose a deep supervised metric learning method for content-based image retrieval that is capable of overcoming the aforementioned issues. Therefore, instead of just learning a representation that is capable of discriminating between the different classes used for training, we propose to encode the *latent generative factors* for each class of the data. This allows for a) encoding the information arising from the in-class variance, as well as b) efficiently representing data samples that do not belong to the classes that were used during the training process (out-of-domain retrieval) by (implicitly) modeling the structure of the input space. Any kind of trainable feature extractor can be directly combined with the proposed method, ranging from traditional Bag-of-Words representations [9,29] to state-of-the-art recurrent and convolution neural networks [3,13]. As a result, the proposed method constitutes a powerful metric learning tool that can be used for a wide variety of information retrieval tasks, as well as combined with various types of feature extractors.

Following the suggestions of Sablayrolles et al. [33], we extensively evaluate and demonstrate the effectiveness of the proposed method under different in-domain and out-of-domain setups using three image datasets. Furthermore, two common evaluation setups, i.e., learning image representations using deep Convolutional Neural Networks (CNNs) from scratch as well as fine-tuning a representation extracted from a pre-trained CNN, are employed to demonstrate the effectiveness of the proposed approach. Note that even though the proposed method *does not* use data from the *target* domain during the optimization, it is capable of performing better than both its supervised counterparts trained with in-domain data, as well as other unsupervised methods trained using *both* in-domain and out-of-domain data.

The rest of this paper is structured as follows. First, the related work is briefly discussed and compared to the proposed method in Section 2, highlighting the shortcoming of existing approaches and providing further motivation for this work. Then, the proposed method is derived in Section 3, while a detailed experimental evaluation is provided in Section 4. Finally, conclusions are drawn in Section 5.

2. Related work

Several supervised metric learning approaches have been proposed, ranging from the seminal contrastive [6] and triplet [44] losses to more advanced and recently proposed methods, such as the N-pair loss [37], lifted structural embedding loss [26], multi-similarity loss [43], and the DDML method [20], that further increase the effectiveness of supervised metric learning. The vast majority of the proposed supervised learning methods works by learning discriminative metric spaces which disentangle the representation of dissimilar objects, while bringing the similar ones as close as possible. However, as we argue through this paper, this approach can negatively affect the precision of an information retrieval system.

In order to better understand the limitations of these discriminative objectives, we can consider a widely used metric learning loss, the contrastive loss [6,35]:

$$\frac{1}{2}\delta_{ij}||\mathbf{x}_i - \mathbf{x}_j||_2^2 + \frac{1}{2}(1-\delta_{ij})\max(0, k - ||\mathbf{x}_i - \mathbf{x}_j||_2^2), \quad (1)$$

where the notation $\mathbf{x}_i$ is used to refer to the vector representation of the ith training sample (as extracted through a deep neural network), k refers to the minimum distance between samples that belong to different classes (or samples marked as dissimilar) and

$$\delta_{ij} = \begin{cases} 1, & \text{if the } i\text{th and } j\text{th samples are similar,} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Therefore, contrastive loss works by forcing all the samples of the same class to be close together (ideally to collapse into one single point), while pushing away points that belong to different classes in order to maintain at least a distance of k between them. The degree up to which the in-class samples collapse depends both on the used regularization methods, as well as on the learning capacity of the employed machine learning model.

Collapsing the in-class samples in this way usually allows for improving the in-domain accuracy, since it leads to reducing the in-class variance and, as a result, learning representations that better discriminate the samples of the training classes between them. However, it is also comes with two important drawbacks. First, reducing the in-class variance leads to stripping away useful information regarding the in-class samples. Therefore, it is usually no longer possible to reliably differentiate between the in-class samples, e.g., retrieve sub-classes of the data. To better understand this consider the following example. Suppose that the representation has been optimized to discriminate between the class “car” and the class “bicycle”. During the optimization process, all the attributes of the data that are irrelevant to these classes are usually discarded, e.g., color, type of vehicle, etc., which is required in order to reliable discriminate between the two classes. Therefore, the learned representation is excellent for classifying cars and bicycles. However, what if a user is interested in retrieving “red cars” or a specific type of cars, e.g., “sedans”? A discriminative representation have probably already discarded this information (recall that in (1) the in-class samples are forced to have as similar representations as possible). As a result, we would probably be unable to reliable retrieve such objects just by relying on this representation.

Second, it can potentially harm the generalization ability of the representation for classes for which it has not been trained (out-of-domain retrieval), since the representation space has been optimized in order to map every input sample into a small set of points (in the perfect case). Indeed, it has been recently demonstrated that these highly discriminative objectives can significant harm the retrieval precision for out-of-domain retrieval tasks [28,33]. It is also quite easy to verify that most of the existing supervised learning objectives, such as triplet-based objectives [27,34], are also prone to this behavior, since they also promote collapsing the representation for similar samples.

There is only a limited number of methods that indeed take into account the aforementioned behavior, despite the fact that this kind of feature space distortion can have a devastating impact on the actual performance of a retrieval system. This problem was studied both in [33], where the precision of supervised hashing methods on retrieving classes that were not seen during the training was examined, as well as in [28,30], where entropy-based objectives were employed to provide better regularized objectives for information retrieval. However, [33] mainly focused on the evaluation metrics that should be employed to evaluate the performance of deep supervised hashing (instead of proposing a method that can overcome the aforementioned limitations), while the continuous approximation proposed in [28] is still vulnerable to class collapse phenomena.

Furthermore, the proposed method is also related to variational learning approaches [11,16,17,19,31]. A variational approach for metric learning, aiming at encoding the intra-class variance, was also proposed in [17]. However, the method proposed in [17] relies on using existing supervised metric learning approaches for disentangling the representation, without providing the ability to

tweak the discrimination-representation trade-off, as the method proposed in this paper, as we also experimentally confirm in Section 4. To the best of our knowledge, this is the first time that the issues arising from the variance-minimization nature of existing supervised metric learning approaches are studied in detail in structured way in the context of information retrieval. Finally, it is worth noting that transfer learning and domain adaptation tasks [27], as well as few-shot learning [36,51], are also related to the aforementioned problem. However, in contrast with most of these approaches, in this paper we revisit the process of supervised deep metric learning, identify critical limitations and propose a novel way to overcome them. The proposed method is, in this sense, orthogonal to existing transfer learning and domain adaptation methods, since it can be combined with most of these methods.

3. Proposed method

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ denote a collection of N images. A feature extraction model $f_{\mathbf{W}}(\mathbf{x}_i) \in \mathbb{R}^m$, where $\mathbf{W}$ denotes the trainable parameters of the model, is then used in order to extract a compact feature representation $\mathbf{z}_i = f_{\mathbf{W}}(\mathbf{x}_i)$ from each image. Note that $\mathbf{x}_i$ can be any kind of data, e.g., vector, tensor, sequential data, etc., given that the feature extraction model is capable of handling the corresponding data. Also, in this work it is assumed that a typical image retrieval setup is employed: a query image $\mathbf{q}$ is provided and the users are interested in ranking the images in $\mathcal{X}$ according to their relevance to $\mathbf{q}$. Note that images are ranked according to the metric learned by $f_{\mathbf{W}}(\cdot)$ instead of the original space.

It is worth noting that learning feature extractors that are capable of a) keeping as much information as possible for each object, while b) ensuring that objects that fulfill the same information need (i.e., are similar according to a higher level semantic attribute) will be similar according to the learned representation is not an easy task [31]. This happens mainly due to the fact that maintaining the intra-class variance tends to be usually conflicting with the discriminative objective of most related metric methods that aim at minimizing the intra-class variance, especially when an appropriate regularization method is not employed. In this work, we propose learning both a discriminative metric space, along with the latent generative factors of each class, allowing for overcoming the aforementioned limitations. To this end, we assume that a mixture of Gaussians exists in the representation space, where each of these Gaussians is mapped to a different class of the data, as shown in Fig. 1. Pushing the Gaussians apart (instead of the samples) ensures that the representation will be discriminative (repulsive loss in Fig. 1), while encoding the latent generating factors for each class ensures that we will maintain the in-class variance for each class in a meaningful way (reconstruction + KL loss in Fig. 1).

Before delving into the details of handling multiple information needs, let's first consider the case of having just one class. Given a set of images that belong to the same class we are interested in modeling the intra-class differences between them. Perhaps the easiest way to achieve this is to learn the latent generative factors for them, allowing for effective reconstructing each image [11,16,19]. Therefore, each image is reconstructed by sampling a vector $\mathbf{z}$ from the latent space and then using a separate image decoder $f_{dec}(\mathbf{z})$ to reconstruct it. As a result, the proposed method aims at maximizing the probability of correctly reconstructing each training data $\mathbf{x}$ when sampling the latent space:

$$P(\mathbf{x}) = \int_{\mathbf{z}} P(\mathbf{x}|\mathbf{z})P(\mathbf{z})d\mathbf{z}, \tag{3}$$

where $P(\mathbf{x}|\mathbf{z})$ corresponds to a Gaussian distribution centered around the reconstructed sample, after employing an image decoder $f_{dec}(\cdot)$, with a diagonal covariance matrix scaled by σ, i.e.,

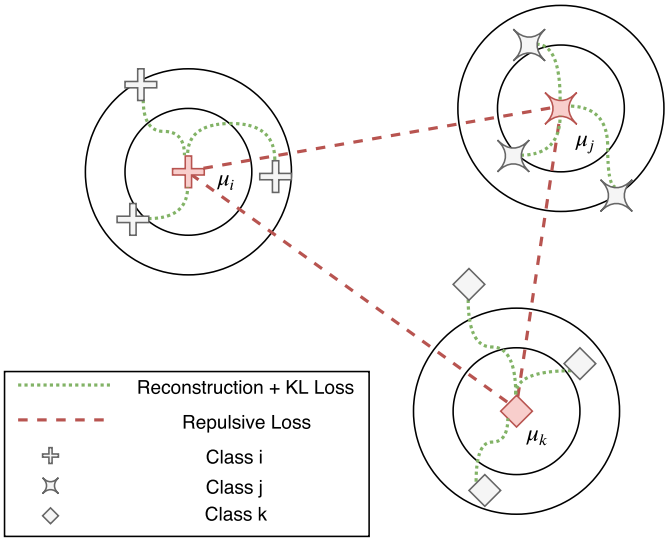

Fig. 1. Proposed method: Two different *forces* act on each training sample (shown in green line). A reconstruction objective ensures that the latent space will model the generative factors of the data, while the KL divergence ensures that each sample will be near the distribution that was used for generating it. Note the difference with existing discriminative objectives (e.g., contrastive loss) that only attract the in-class samples to each other. At the same time, the Gaussians used to model the data should be far enough (repulsive loss, shown in red line), ensuring that the learned metric space will be discriminative. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$P(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; f_{dec}(\mathbf{z}), \sigma\mathbf{I}). \tag{4}$$

Also, $P(\mathbf{z})$ denotes the density function of the latent space, which is again usually set to be a Gaussian distribution with zero-mean and unit variance [11]:

$$P(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}), \tag{5}$$

where

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{k}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{6}$$

and k denotes the dimensionality of $\mathbf{x}$.

However, directly optimizing (3) is not tractable. Therefore, the evidence lower bound (ELBO) will be maximized instead:

$$E_{z\sim Q(\mathbf{z}|\mathbf{x})}[\log P(\mathbf{x}|\mathbf{z}) - \mathcal{D}_{KL}(Q(\mathbf{z}|\mathbf{x})||P(\mathbf{z}))], \tag{7}$$

where $Q(\mathbf{z}|\mathbf{x})$ is a tractable model that we learn to appropriately approximate $P(\mathbf{z}|\mathbf{x})$, and $\mathcal{D}_{KL}$ is the Kullback-Leibler (KL) divergence between two probability distributions. Then, the reparametrization trick, along with Monte-Carlo sampling, are employed to optimize this bound [11]. A neural network is used in this work to model $Q(\mathbf{z}|\mathbf{x})$. This network is used to predict the mean $\mu_Q(\mathbf{x})$ and the covariance matrix $\Sigma_Q(\mathbf{x})$ (constrained to be diagonal) for any given a sample $\mathbf{x}$. Then, a Gaussian is used to model the distribution Q:

$$Q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_Q(\mathbf{x}), \Sigma_Q(\mathbf{x})). \tag{8}$$

Finally, note that the representation of each training sample can be obtained as $f_{\mathbf{W}}(\mathbf{x}) = \mu_Q(\mathbf{x})$. Thus, the feature extractor $f_{\mathbf{W}}(\cdot)$ is learned implicitly through this process.

The aforementioned variational approach allows for learning the latent generative factor for each class. However, at the same time, the learned representation should be able to discriminate between the different classes used during the training process. Therefore, to handle multiple classes/information needs, we propose employing separate Gaussians, each with separate mean $\boldsymbol{\mu}_i$ and unit variance, for generating images of different classes. To this end, we

define $P(\mathbf{z}|c=i)$ to be the probability of reconstructing an image that belongs to the ith class by a sample drawn from the ith Gaussian. This probability, which should be maximized in order to ensure that the latent space will encode the generative factors of the images, is calculated as:

$$P(\mathbf{x}|c=i) = \int_{\mathbf{z}} P(\mathbf{x}|\mathbf{z})P(\mathbf{z}|c=i)d\mathbf{z}, \tag{9}$$

where $P(\mathbf{z}|c=i) = \mathcal{N}(\boldsymbol{\mu}_i, 1)$. Therefore, a different Gaussian is used for each class, allowing for separately modeling the in-class variance of images that belong to different classes. Note that extra Gaussians can be also used to model images that are not annotated, allowing for performing semi-supervised learning.

To optimize (9) the ELBO is again employed leading to the following loss:

$$\mathcal{L}_{var}(\mathbf{x}) = [\log P(\mathbf{x}|\mathbf{z}) - \alpha_{KL}\mathcal{D}_{KL}(Q(\mathbf{z}|\mathbf{x})||P(\mathbf{z}|c=i))], \tag{10}$$

where α_{KL} controls the importance of minimizing the corresponding KL-divergence, which is analytically computed as:

$$\mathcal{D}_{KL}(Q(\mathbf{z}|\mathbf{x})||P(\mathbf{z}|c=i))] = \tag{11}$$

$$\frac{1}{2}\Big(\text{tr}(\Sigma_Q(\mathbf{x})) + (\mu_Q(\mathbf{x}) - \boldsymbol{\mu}_i)^T(\mu_Q(\mathbf{x}) - \boldsymbol{\mu}_i) \tag{12}$$

$$-m - \log\det(\Sigma_Q(\mathbf{x}))\Big), \tag{13}$$

where m is the dimensionality of the latent space. Also, note that $\log P(\mathbf{x}|\mathbf{z})$ is proportional to $-||f_{dec}(\mathbf{z}) - \mathbf{x}||_2^2$ (appropriately scaled by σ).

The loss given in (10) ensures that the latent space will not collapse into a single point for each class, since it models the generative factors for each class (collapsing would not allow for capturing the in-class variability and, as a result, will not allow reconstructing the samples). However, it does not ensure that the learned representation will be discriminative, i.e., that the data in the latent space will be indeed separable. To this end, the Gaussians should be placed at a distance of at least ρ of each other. This allows for separating the different classes, while, at the same time, encoding the semantic similarity between different classes. Therefore, the centers $\boldsymbol{\mu}_i$ are constrained to be at a distance of at least ρ during the optimization, allowing for also encoding possible semantic relationships between different classes. In this work, we propose employing a simple margin-based loss on the representation of the Gaussian means (instead on that of individual samples) to achieve this:

$$\mathcal{L}_{sup} = \frac{1}{\rho}\sum_{i=1}^{N_C}\sum_{j=1, j\neq i}^{N_C}\max(0, \rho - ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||_2^2)^2 \tag{14}$$

where ρ denotes that minimum distance between the used Gaussians and N_C is the number of them.

The final DL model is trained according to the following combined loss:

$$\mathcal{L} = \mathcal{L}_{var} + \mathcal{L}_{sup}, \tag{15}$$

which combines both the reconstruction-based objective (allowing for capturing the in-class variation), as well as the discriminative objective (allowing for discriminating between different classes), as also summarized in Fig. 1. Minimizing the KL divergence between the current and target distributions, as described in (10), ensures that the representation of each sample will be near to the correct Gaussian distribution. As a result, the hyper-parameter α_{KL} (which is typically set to 1, unless otherwise stated) allows for effectively controlling the trade-off between maintaining the intra-class variance and maximizing the inter-class distances. The Gaussian centers are initialized using orthogonal initialization scaled by the desired distance ρ, providing a good starting point for the optimization. Note that, as demonstrated in Section 4, the hyper-parameter ρ also allows for controlling the discriminative power of the learned representations.

The distribution $Q(\mathbf{z}|\mathbf{x})$, which is optimized through stochastic gradient descent, along with the parameters of the decoder, provides the representation of the samples in the latent space as $f_{\mathbf{W}}(\mathbf{x}) = \mu_Q(\mathbf{x})$. Note that the Adam optimizer is employed in this work for the optimization [10], instead of the plain stochastic gradient descent. The default optimization hyper-parameters are used for all the conducted experiments. Finally, recall that the decoder $f_{dec}(\cdot)$ is not required during the inference process, since it is only employed to provide an auxiliary task during the training process.

4. Experimental evaluation

The proposed method is extensively evaluated and compared to other related methods in this Section. First, the used datasets, network architectures and employed evaluation setups are briefly introduced in the first subsection, while the experimental results are provided in the following subsection.

4.1. Evaluation setup

Three different datasets are used for the evaluation, the Fashion MNIST dataset [48], the Caltech-UCSD Birds 200 (CUB-200) dataset [45], and the Animals with Attributes (AwA2) dataset [46]. The Fashion MNIST dataset contains 70,000 images (60,000 training and 10,000 testing images) that depict fashion-related items. For the in-domain evaluation 5 classes were randomly selected, while the rest of them were used for the out-domain evaluation. As for the Fashion MNIST dataset, for the in-domain setup 5 classes were randomly selected for each experiments, while the rest of them were used for the out-of-domain evaluation. The Animals with Attributes (AwA2) dataset consists of images that belong to 50 different classes. For this dataset, 20 classes were selected for the in-domain evaluation, while for the out-of-domain evaluation the rest of them were employed. The CUB-200 dataset contains 200 classes (100 of them were randomly selected for the in-domain evaluation and the rest of them were used for the out-of-domain evaluation). All the conducted experiments were repeated 5 times. Both the mean and standard deviation of the metrics are reported.

The encoder used for the Fashion MNIST has the following architecture: one 3×3 convolutional layer with 16 filters (stride 2) and one 3×3 convolutional layer with 32 filters, followed by another two convolutional layers with 64 and 128 filters respectively. The output of the convolutional layers was fed to two fully connected layers with 256 neurons and 2 m neurons respectively. The dimensionality of the latent space was set to $m = 30$. For all the layers the ReLU activation function is used, while Batch Normalization [8] is employed for all the convolutional layers. The encoder predicts both the mean vector $\mu_Q(\mathbf{x})$ and the (diagonal) covariance $\Sigma_Q(\mathbf{x})$. For reconstructing each image, a symmetric decoder with transpose convolutional layers is employed (the strides were set appropriately). The sigmoid activation function is used for the last layer of the decoder, while the reconstruction error is measured using the binary cross-entropy loss.

For the AwA2 and CUB-200 datasets a different setup was used. The purpose of the experiments conducted using the AwA2 dataset was to examine the ability of the proposed method to be combined with pre-trained feature extractors. Note that this *transfer learning* setup is commonly used, allowing for rapidly fine-tuning existing deep learning models for different tasks. A ResNet101 model, which was pretrained on the Imagenet dataset [46], was employed to this end. For the CUB-200 dataset, a ResNeXt101 [50] model, again pretrained on the Imagenet dataset, was used. An encoder

Table 1
Evaluation results using the Fashion MNIST dataset (mAP, %).

Method	In-domain	In-domain + Distractors	Out-of-domain
Variational Autoencoder [11]	51.05 ± 4.23	35.41 ± 1.91	49.04 ± 2.12
Contrastive Loss [6]	85.90 ± 5.38	62.47 ± 5.05	58.21 ± 2.43
Triplet [7]	82.04 ± 7.96	57.32 ± 7.87	58.68 ± 3.22
Lifted [26]	88.16 ± 6.20	68.77 ± 3.95	62.10 ± 3.69
N-Pair [37]	88.62 ± 6.33	68.86 ± 4.20	62.02 ± 3.20
Proposed	**90.45 ± 4.19**	**69.42 ± 2.23**	**63.35 ± 2.76**

with three fully connected layers was used: a layer with 512 neurons, followed by a layer with 128 neurons and a final representation layer with 2 m neurons, where $m = 50$. As before, a symmetric decoder was used and the ReLu activation function was employed.

Apart from the proposed method, five other well-known metric learning approaches were used, four supervised ones (contrastive loss [6], triplet loss [7], lifted structural embedding loss [26] (abbreviated as "lifted"), N-pair loss [37]) and an unsupervised one (Variational Autoencoder [11]). The Variational Autoencoder was fitted on the whole Fashion MNIST, while the margin k was set to 10 for the constrastive loss. A online triplet generation method was used for the triplet loss, following the approach described in [7] (the margin was set to $k = 0.5$ without using any hard triplet mining strategy). The implementation provided in [42] was used for the lifted loss (using a margin of 0.5), while the implementation provided in [15] was used for the N-pair loss. Despite their overall better performance compared to other supervised metric learning methods, N-pair and lifted losses were especially unstable despite carefully tuning their hyper-parameters. The best performance was obtained by first pre-training the networks for 5 epochs (30 for the CUB-200 dataset) using the contrastive loss and then resuming training using each of these loss. This strategy lead to significant improved results compared to not using pre-training. The same architecture was used for the encoders for all the evaluated approaches to ensure a fair comparison. The batch size was set to 128 (a reduced batch size of 32 was used for the more resource-intesive triplet-based approaches). For the Fashion MNIST dataset the models were trained for 50 epochs using a learning rate of 0.001. Finally, 10 training epochs were used for the AwA2 dataset ($\eta = 0.001$), while 50 training epochs were used for the CUB-200 dataset.

The following evaluation setups were used for the conducted experiments:

1. "**In-domain**', where in-domain queries were employed to retrieve images from a database that was composed of in-domain data (the in-domain queries belong to the same classes as the training data, but the actual queries were not seen during the training),
2. "**In-domain + Distractors**", where a database that contains both in-domain, as well as out-of-domain data (data from classes not seen during the training), was queried by in-domain queries,
3. "**Out-of-domain**" (out-of-domain), where a database that contains out-of-domain data was queried using out-of-domain queries, and
4. "**Out-of-domain + Distractors**", where a database that contains both in-domain and out-of-domain data was queried using out-of-domain queries.

The actual setup used depends on the available annotations for each dataset. Following the standard retrieval evaluation approach, we report the 11-recall point-based mean Average Precision (mAP) [22]. Note again that only in-domain data were used for training all the supervised methods evaluated in this paper.

4.2. Evaluation results

First, the proposed method is evaluated using the Fashion MNIST dataset. The experimental results are reported in Table 1. Several conclusions can be drawn from the reported results. First, all the supervised learning methods lead to significant improvements over the unsupervised one (variational AE), regardless the employed evaluation setup. Also, the contrastive loss seems to be more efficient than the triplet loss, leading to slightly more discriminative representations, while the N-pair and lifted losses further increase the retrieval precision. On the other hand, the proposed method leads to better generalization, increasing the in-domain retrieval precision. At the same time, it also leads to significant precision improvements for the out-of-domain and distractor-based evaluation setups. These results confirm the hypothesis that employing additional auxiliary tasks, that allow for modeling the in-class variance, allow for improving both the in-class performance, significantly increasing the generalization abilities of the representation, as well as provide the ability to better represent objects that belong to classes not seen during the training.

The aforementioned observations are also confirmed in the plot shown in Fig. 2, which demonstrates the trade-off between the ability of a representation to discriminative between in-class samples and to represent unknown samples. Note that as the minimum distance between the Gaussians increases, the in-domain precision also increases. However, after a certain point this happens at the expense of the out-of-domain performance. The hyper-parameter ρ was set to 2 according to these results for all the conducted experiments (except for the CUB-200, where a larger value of $\rho = 20$ led to consistently better results for all the evaluation setups) Table 2. Similar results were also obtained when varying the

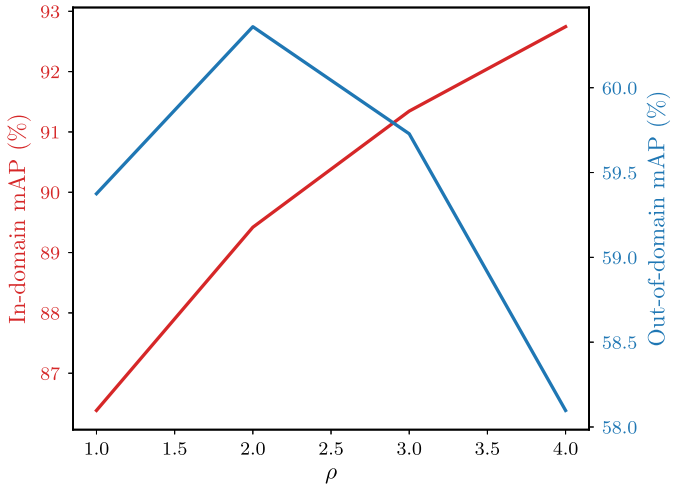

Fig. 2. Effect of varying the hyper-parameter ρ (minimum distance between the Gaussians) on the in-domain and out-of-domain retrieval precision (the optimization ran for 30 training epochs).

Table 2
Evaluation results using the CUB-200 dataset (mAP, %).

Method	In-domain	In-domain + Distractors	Out-of-domain
Variational Autoencoder [11]	18.52 ± 0.74	15.63 ± 0.77	18.69 ± 0.64
Contrastive Loss [6]	32.85 ± 0.09	27.57 ± 0.11	27.98 ± 0.67
Triplet [7]	22.50 ± 0.67	18.66 ± 0.48	19.97 ± 0.46
Lifted [26]	27.36 ± 0.76	22.80 ± 0.52	23.97 ± 0.56
N-pair [37]	33.90 ± 0.69	28.74 ± 0.58	28.89 ± 0.62
Proposed	**40.52 ± 0.28**	**34.90 ± 0.44**	**29.43 ± 0.36**

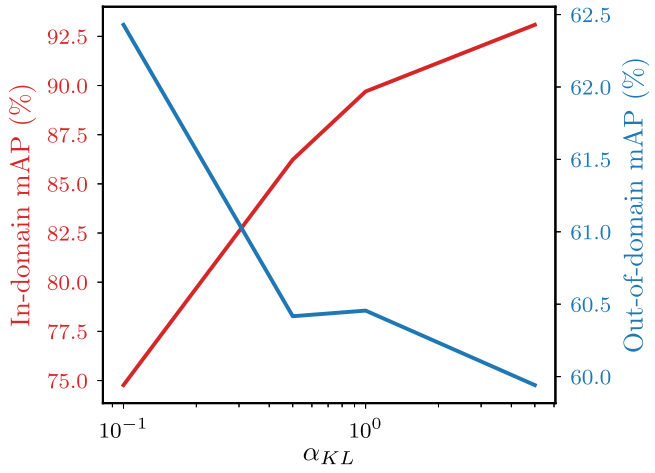

Fig. 3. Effect of varying the hyper-parameter α_{KL} on the in-domain and out-of-domain retrieval precision (the optimization ran for 30 training epochs).

α_{KL} hyper-parameter, i.e., higher values lead to more discriminative representations, while lower values lead to better performance on out-of-domain tasks, as shown in Fig. 3.

Next, the performance of the method was evaluated using the CUB-200 dataset. Again, the supervised metric learning approaches lead to significant performance improvements over the plain unsupervised Variational AE, with the N-pair method leading to the overall highest mAP over the rest of the supervised metric learning approaches. Note that triplet loss, despite improving the in-domain precision, yields only margin improvements for out-of-domain retrieval. On the other hand, the proposed method yields an impressive 20% relative improvement over the next best performing method (N-pair) for in-domain retrieval, while it also improves the retrieval precision in the presence of distractors (over 20% relative improvement). At the same time, it leads to the overall best performance for out-of-domain retrieval tasks, confirming the positive effect of modeling the intra-class variance.

Finally, experiments using the AwA2 dataset were conducted. The results are reported in Table 3. Since no training/test split was employed for this dataset, only the "out-of-domain" (abbreviated as "Out" in Table 3) and "out-of-domain + in-domain distractors" (abbreviated as "Out+" in Table 3) results are provided. Note that for the AwA2 evaluation, three out of the four supervised learning methods actually lead to lower retrieval performance compared to a fully unsupervised Variational Autoencoder (VAE), while the best performing N-pair loss leads only to slight improvements over the VAE. This implies that the more discriminative representations are not necessarily the most suitable for out-of-domain retrieval tasks, confirming again the findings reported in the literature [28,33]. Again, the proposed method leads to better performance compared to all the other evaluated methods, improving the out-of-domain retrieval precision by about 5% for both scenarios (relative improvements).

Table 3
Evaluation results using the AwA2 dataset (mAP, %).

Method	Out	Out+
VAE	45.39 ± 2.49	39.20 ± 1.03
Contrastive	43.42 ± 0.53	37.68 ± 1.23
Triplet	41.02 ± 2.18	35.53 ± 1.67
Lifted	45.50 ± 1.68	39.19 ± 1.53
N-pair	46.32 ± 1.71	40.27 ± 2.10
Proposed	**48.60 ± 1.58**	**42.56 ± 0.65**

5. Conclusions

A regularized discriminative deep metric learning method for content-based image retrieval, which works by modeling the latent generative factors for each class, was proposed in this paper. The proposed method is capable of learning efficient representations that can discriminate between different classes, modeling both the in-class variance and possible sub-classes, as well as representing objects from classes that were not seen during the training process. The effectiveness of the proposed method was demonstrated using several experiments under different different in-domain and out-of-domain setups and three challenging image datasets, outperforming the evaluated supervised and unsupervised representation/metric learning approaches.

Declaration of Competing Interest

The authors declare that have no conflicts of interest.

References

[1] G. Cao, A. Iosifidis, K. Chen, M. Gabbouj, Generalized multi-view embedding for visual recognition and cross-modal retrieval, IEEE Trans. Cybern. 48 (9) (2018) 2542–2555.
[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[3] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.
[4] Y. Duan, J. Lu, Z. Wang, J. Feng, J. Zhou, Learning deep binary descriptor with multi-quantization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1183–1192.
[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition, 2010, pp. 2360–2367.
[6] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, 2006, pp. 1735–1742.
[7] A. Hermans, L. Beyer, B. Leibe, In defense triplet loss for person re-identification (2017). arXiv: 1703.07737.
[8] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 448–456.
[9] A. Iosifidis, A. Tefas, I. Pitas, Multidimensional sequence classification based on fuzzy distances and discriminant analysis, IEEE Trans. Knowl. Data Eng. 25 (11) (2012) 2564–2575.
[10] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Proceedings of the International Conference for Learning Representations, 2015.

[11] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: Proceedings of the International Conference on Learning Representations, 2014.
[12] A. Krizhevsky, G. Hinton, Using very deep autoencoders for content-based image retrieval., in: Proceedings of the European Symposium on Artificial Neural Networks, 1, 2011, p. 2.
[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[14] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Cross-view and multi-view gait recognitions based on view transformation model using multi-layer perceptron, Pattern Recognit. Lett. 33 (7) (2012) 882–889.
[15] S. Lee, M. Kook, S.-W. Kim, A pytorch framework for image retrieval, 2019. Available at https://github.com/leeesangwon/PyTorch-Image-Retrieval.
[16] J. Lim, Y. Yoo, B. Heo, J.Y. Choi, Pose transforming network: learning to disentangle human posture in variational auto-encoded latent space, Pattern Recognit. Lett. 112 (2018) 91–97.
[17] X. Lin, Y. Duan, Q. Dong, J. Lu, J. Zhou, Deep variational metric learning, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 689–704.
[18] H. Liu, R. Wang, S. Shan, X. Chen, Deep supervised hashing for fast image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2064–2072.
[19] G. Lu, X. Zhao, J. Yin, W. Yang, B. Li, Multi-task learning using variational auto-encoder for sentiment classification, Pattern Recognit. Lett. (2018).
[20] J. Lu, J. Hu, Y.-P. Tan, Discriminative deep metric learning for face and kinship verification, IEEE Trans. Image Process. 26 (9) (2017) 4269–4282.
[21] C. Ma, C. Gong, Y. Gu, J. Yang, D. Feng, Shiss: supervised hashing with informative set selection, Pattern Recognit. Lett. 107 (2018) 105–113.
[22] C. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Nat. Lang. Eng. 16 (1) (2010) 100–103.
[23] R. Mazzon, S.F. Tahir, A. Cavallaro, Person re-identification in crowd, Pattern Recognit. Lett. 33 (14) (2012) 1828–1837.
[24] H. Müller, W. Müller, D.M. Squire, S. Marchand-Maillet, T. Pun, Performance evaluation in content-based image retrieval: overview and proposals, Pattern Recognit. Lett. 22 (5) (2001) 593–601.
[25] X. Nie, W. Jing, C. Cui, J. Zhang, L. Zhu, Y. Yin, Joint multi-view hashing for large-scale near-duplicate video retrieval, IEEE Trans. Knowl. Data Eng. (2019).
[26] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4004–4012.
[27] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 10 (22) (2010) 1345–1359.
[28] N. Passalis, A. Tefas, Entropy optimized feature-based bag-of-words representation for information retrieval, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1664–1677.
[29] N. Passalis, A. Tefas, Learning bag-of-features pooling for deep convolutional neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5755–5763.
[30] N. Passalis, A. Tefas, Learning deep representations with probabilistic knowledge transfer, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 268–284.
[31] N. Passalis, A. Tefas, A. Iosifidis, M. Gabbouj, Class-based variational representation learning for robust image retrieval, in: Proceedings of the IEEE International Conference on Image Processing, 2019, pp. 854–858.
[32] H. Peng, J. He, S. Chen, Y. Wang, Y. Qiao, Dual-supervised attention network for deep cross-modal hashing, Pattern Recognit. Lett. (2019).
[33] A. Sablayrolles, M. Douze, N. Usunier, H. Jégou, How should we evaluate supervised hashing? in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 1732–1736.
[34] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[35] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, F. Moreno-Noguer, Discriminative learning of deep convolutional feature point descriptors, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 118–126.
[36] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Proceedings of the of Advances in Neural Information Processing Systems, 2017, pp. 4077–4087.
[37] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 1857–1865.
[38] M. Tzelepi, A. Tefas, Deep convolutional learning for content based image retrieval, Neurocomputing 275 (2018) 2467–2478.
[39] J. Wan, D. Wang, S.C.H. Hoi, P. Wu, J. Zhu, Y. Zhang, J. Li, Deep learning for content-based image retrieval: acomprehensive study, in: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 157–166.
[40] J. Wang, S. Kumar, S.-F. Chang, Semi-supervised hashing for large-scale search, IEEE Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2393–2406.
[41] J. Wang, T. Zhang, N. Sebe, H.T. Shen, et al., A survey on learning to hash, IEEE Trans. Pattern Anal. Mach.Intell. 40 (4) (2017) 769–790.
[42] X. Wang, Deep metric learning in pytorch, 2019. Available at https://github.com/bnu-wangxun/Deep_Metric.
[43] X. Wang, X. Han, W. Huang, D. Dong, M.R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.
[44] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (Feb) (2009) 207–244.
[45] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-UCSD Birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
[46] Y. Xian, C.H. Lampert, B. Schiele, Z. Akata, Zero-shot learning-a comprehensive evaluation good, the bad and the ugly, IEEE Trans. Pattern Anal. Mach.Intell. (2018).
[47] L. Xiang, G. Zhao, X. Shen, F. Li, Adaptive multi-graph hashing for scalable multimedia retrieval, Pattern Recognit. Lett. (2019).
[48] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.
[49] L. Xie, R. Hong, B. Zhang, Q. Tian, Image classification and retrieval are one, in: Proceedings of the ACM on International Conference on Multimedia Retrieval, 2015, pp. 3–10.
[50] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.
[51] H. Zhang, Y. Long, L. Shao, Zero-shot hashing with orthogonal projection for image retrieval, Pattern Recognit. Lett. 117 (2019) 201–209.
[52] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1239–1248.
[53] X. Zhang, L. Zhou, X. Bai, X. Luan, J. Luo, E.R. Hancock, Deep supervised hashing using symmetric relative entropy, Pattern Recognit. Lett. 125 (2019) 677–683.