Hybrid-Attention based Decoupled Metric Learning for Zero-Shot Image Retrieval

Binghui Chen\textsuperscript{1, 2}, WeiHong Deng\textsuperscript{1}\textsuperscript{*}
\textsuperscript{1}Beijing University of Posts and Telecommunications
\textsuperscript{2}AI Labs, Didi Chuxing, Beijing 100193, China
chenbinghui@bupt.edu.cn, whdeng@bupt.edu.cn

Abstract

In zero-shot image retrieval (ZSIR) task, embedding learning becomes more attractive, however, many methods follow the traditional metric learning idea and omit the problems behind zero-shot settings. In this paper, we first emphasize the importance of learning visual discriminative metric and preventing the partial/selective learning behavior of learner in ZSIR, and then propose the Decoupled Metric Learning (DeML) framework to achieve these individually. Instead of coarsely optimizing an unified metric, we decouple it into multiple attention-specific parts so as to recurrently induce the discrimination and explicity enhance the generalization. And they are mainly achieved by our object-attention module based on random walk graph propagation and the channel-attention module based on the adversary constraint, respectively. We demonstrate the necessity of addressing the vital problems in ZSIR on the popular benchmarks, outperforming the state-of-the-art methods by a significant margin. Code is available at http://www.bhchen.cn

1. Introduction

In zero-shot image retrieval (ZSIR), the model is required to learn embedding from the seen classes and then to be capable of utilizing the learned knowledge to distinguish the unseen classes without any attributes or semantic information. Most existing works adopt deep embedding (metric) learning and dedicate to exploring powerful loss functions, such as pair-based methods \cite{33, 43, 29, 23, 32, 2, 35}, and hard-example mining strategies \cite{16, 39, 29, 32}, etc. These ideas can be summarised as: learn a good metric on the seen categories so as to boost the performance on the unseen ones. However, in fact, they neglect the existence and significance of some vital problems to ZSIR, thus are easily stuck in the particular data distribution and knowledge zone of the seen training set, some helpful knowledge for unseen classes may have been left out with a high probability, as a result, their performances are almost on par with each other and unsatisfactory.

Specifically, in ZSIR, the ideas above are actually unreasonable as the discrimination and generalization abilities of the learned metric are highly influenced by the following two aspects: (1) the undiscriminating visual input, which contains much noisy information (it is hard to know where should be focused on), resulting in poor descriptors and inducing difficulty in distinguishing the unseen classes. In another word, undiscriminating visual input deteriorates the discrimination of features. (2) partial/selective learning behavior of deep models\cite{3}, specifically, for a functional learner parameterized by CNN, when given an arbitrary input, it will selectively learn partial attributes knowledge that are the easiest ones to decrease the current training empirical risk over the seen categories, instead of learning all-sided details and information, thus yielding over-fitting on seen categories and generalizing worse to the unseen ones. For example on birds, if the head knowledge is enough to distinguish the training seen classes, the deep model will only focus on the head and ignore the other body parts like back, foot and wing, while when testing on the unseen classes (which have similar head features but different wing features), the trained model will easily fail to distinguish them\textsuperscript{1}. This partial learning behavior sometimes

\textsuperscript{*}Corresponding author

\textsuperscript{1}Since in zero-shot settings, the unseen classes have no intersection with the seen classes, this partial learning behavior is common.
is one reason for neural redundancy, since all the neurons only focus on the certain easy knowledge, causing that the performance doesn’t increase much with the feature dimension. Moreover, when optimizing an unified metric as in the aforementioned methods (as in Fig.1.(a)), these two issues are commonly mixed with each other and jointly impede the learning of representation, and without explicit and helpful guidance, the poor discrimination and generalization characteristics of features are greatly exacerbated regardless of the efficacy of objective function and hard-sample mining strategy. And most ZSIR works ignore the importance of learning both discriminative and robust discriminators. To this end, proposing knowledge-preserving and discriminative metric learning framework remains important.

In this paper, inspired by the idea of “divide and rule”, we propose the Decoupled Metric Learning (DeML), a novel and generally applicable framework, to alleviate the above problems by decoupling the embedding representation into multiple attention-specific learners so as to encourage its discrimination and generalization in an explicit manner, as shown in Fig.1.(b). Moreover, equipping each learners with the particular attention tasks will further alleviate the neural redundancy and give chances of performance improvement. Our DeML is a model-agnostic learning framework, and different from the existing metric learning works, it has no need to mine hard samples or construct elaborate sample-pairs. The main contributions are as follows:

- We provide insights of two vital problems to ZSIR that undiscriminating visual input and partial/selective learning behavior of learner will impede the discrimination and generalization of the metric, resp. And propose DeML, a hybrid-attention based decoupling method, to address these challenges individually.
- DeML decouples the commonly used unified metric into object-attention root-learners and channel-attention sub-learners, aiming at recurrently inducing the discrimination and explicitly promoting the diversity together with the generalization, resp.
- The decoupled learners are supported by the corresponding random walk inferred object-attention modules (OAM)\(^2\) and the adversarial channel-attention modules (CAM). They can be easily performed as the drop-in parts and are model-agnostic.
- We demonstrate the necessity to learn discriminative and robust metric via our decoupling idea. Extensive experiments have been performed on several challenging datasets for ZSIR, including CUB [36], CARS [15], Stanford Online Products [23] and InShop [18]. Our DeML achieves state-of-the-art performances and surpasses other methods by a significant margin.

2. Related Work

Zero-shot learning: ZSL has been widely studied in many tasks, such as image classification [17, 1, 8, 44], hashing[31], video recognition [6] etc. However, different from these ZSL tasks that are capable of exploiting the extra auxiliary supervision signals of unseen classes (e.g. semantic word embedding of class name and explicit attribute information), our DeML concentrates on a more general version of zero-shot image retrieval (ZSIR), where there are only similarity labels \(s_{pq} \in \{0, 1\}\) available. Therefore, how to capture the discriminative and all-sided information only from the input image is the mainstay of this task.

Deep embedding(metric learning: With only \(s_{pq}\) available, many works seek to exploring deep metric learning for ZSIR task. For example, sampling-Matters [39] proposes distance weighted sampling strategy. Proxy-NCA [21] explains why popular classification loss works from a proxy-agent view, and its implementation is very similar with Softmax. ALMN [2] proposes to optimize an adaptive large margin objective via the generated virtual points instead of mining hard-samples. However, all the above methods are to tackle with the unified metric by designing losses and exploring sample-mining strategies, thus suffer from the aforementioned issues easily. Additionally, HDC [43] employs the cascaded models and selects hard-samples from different levels and models. BIER loss [24, 25] adopts the online gradients boosting methods. Although these two methods try to improve the performances by resorting to the ensemble idea, they can only learn from the undiscriminating input and also suffer from the partial learning behavior, as a result, the discrimination and generalization of holistic features are still limited.

Attention mechanism: Attention serves as a tool to bias the allocation of available resources towards the most informative parts of an input. Many are implemented in combination with a gating mechanism (e.g. softmax or sigmoid) and are widely applied in many tasks, e.g. image captioning [4, 40], lip reading [5], image classification [12, 37, 7]. Schwartz et al. [30] adopt high order attention modules for VQA. Fu et al. [7] propose to learn an attention network so as to produce the attention proposals, while, it is optimized in a two-stage manner, i.e. iteratively train the embedding model and attention model. We emphasize that our DeML is a hybrid-attention system. First, it adopts a parameter-free object-attention module, which is performed by random walk graph propagation and can be directly plugged into any bedrock CNNs without training. Second, different from [12, 4] where the channel-attention operations are cascaded, i.e. only the most informative signal will be captured, our DeML is equipped with a series of channel-attention modules in parallelism, which are optimized by the adversary module and insist to learn various combinations of attributes simultaneously and to capture knowledge as much as possible.

3. Decoupled Metric Learning

As discussed in Sec.1, in zero-shot setting, simply optimizing an unified representation like in most existing works
Figure 2: The framework of our DeML. ** indicates the joint operation of cropping and zooming. The FC layer is first decoupled into two object-attention root-learners (dashed rectangle and ellipse) for coarse and finer scale, resp. Then, each root-learner is further decoupled into three channel-attention sub-learners (best viewed in colors). Each (root or sub) learner is supported by the corresponding attention module (OAM or CAM). The parameters to be shared across streams have been annotated by ‘shared’. \( F_{Net}^{+}GNet \) is the whole GooglenetV1, and we use \( F_{Net}, GNet \) to indicate the nets at scale \( i \). Details of the adversary net is in Fig.3.(c)

Naturally motivated by “divide and rule”, the intuitive solution is to decouple the single representation into a series of artificially defined learners so as to explicitly encourage different learners to attend to different latent features and improve the holistic feature’s discrimination and diversity. Moreover, many research works point out that attention mechanism is a crucial component connecting data embeddings with decision making modules, and diverse attentional cues allows the learning of various knowledge. We follow this inspiring idea and recast the above solution as hybrid-attention learning by decoupling the original unified metric into several attention-specific parts, each part is followed by the corresponding fc learner, i.e. object-attention (OA) root-learner and channel-attention (CA) sub-learner. The overall framework of DeML is illustrated in Fig.2 (for clarity, we depict the network with only two OA root-learners and three CA sub-learners under each root-learner as an example).

Specifically, for the \( i \)-th scale, the embedding part consists of three different components: (1) the bedrock CNN, i.e. Googlenet-V1 [34] (we decompose it into \( F_{Net}^{+}GNet \) (Conv1–Pool3, Incep5–Pool5)), (2) the object-attention module (OAM) is to incrementally pinpoint the informative object-based region proposals for discriminative feature learning, and (3) the channel-attention module (CAM) is to concurrently produce diverse attribute-based channel-attention proposal \( cap(i,j) \). When given an image \( x_n \), we will directly use it for the first scale input and employ its cropped&zoomed version for the next scale input. At the \( i \)-th scale, the input will first be encoded by \( F_{Net} \) and then fed into multiple CAMs. The produced channel-attention proposals \( cap(i,j) \) will then be encoded by the corresponding subsequent \( GNet \) and CA sub-learners. All the CA sub-learners at the same scale constitute one OA root-learner.

We now define \( OA_i, CA_{i,(j)} \) as the corresponding learners and \( OA_i(·), CA_{i,(j)}(·) \) as the corresponding encoding functions, then give the holistic feature representation \( \varphi[x_n] \) as:

\[
\varphi[x_n] = \{ OA_i[x_n^i]; OA_2(x_n^2); \cdots \} \tag{1}
\]

where \( OA_i[·] = \{ \cdots; CA_{i,(j)}[·]; \cdots \} \)

where the input \( x_n \in \mathcal{X}_{seen} \equiv \{ x_n^i \}_{n=1}^N \), \( x_n^i \) is the \( i \)-th scale image which is cropped and zoomed from \( x_n^{-1} \) by \( OA_{i-1} \), and \( x_n^i = x_n \) is the original scale image. Notably, we use square brackets “[·]” to show the conditional relation, e.g. \( CA_{i,(j)}[x_n^i] \) indicates the CA learner \( CA_{i,(j)} \) is conditional on input \( x_n^i \), and this relation can also be formulated as \( CA_{i,(j)}[x_n^i] = CA_{i,(j)}(g^i(CAM_{i,(j)}(f^i(x_n^i)))) \) (shows the encoding process from image to feature space), where \( f^i \) and \( g^i \) indicate the encoding functions of \( F_{Net} \) and \( GNet \), resp. \( CAM_{i,(j)} \) denotes the corresponding channel-attention module (CAM), \( i \in [1 \cdots I], j \in [1 \cdots J] \) (in Fig.2, \( I=2, J=3 \)). The total size of \( \varphi[x_n] \) is \( d \)-dim, thus for \( I \) OA learners and \( I \times J \) CA learners, \( OA_i(·) \in \mathbb{R}^d \) and \( CA_{i,(j)}(·) \in \mathbb{R}^{d_J} \).

Below, we will introduce our OAM and CAM in Sec.3.1 and Sec.3.2 resp, and then show the objective metric loss together with the regularization term in Sec.3.3.

---

3Notably, these shortcomings also reside in conventional full-shot tasks, while they heavily exacerbate the performances in zero-shot settings.
As illustrated in Fig.2, our OAM takes as input many convolutional response maps \( U^{l_k} \in \mathbb{R}^{C_{l_k} \times H_{l_k} \times W_{l_k}}, \ k \in [1, 2, \cdots] \), \( C_{l_k}, H_{l_k} \) and \( W_{l_k} \) denote the number of channels, height and width of the \( l_k \)-th layer, respectively, and then gives the final object-attention proposal. For clarity, we give a toy example and omit the subscript \( k \) for written convenience (if not specified), each spatial location \((h, w)\) on \( U^{l} \) has a feature vector \( u^{l}_{h, w} \in \mathbb{R}^{C_{l}} \). To produce the spatial-object-attention proposal \( M^{l} \in \mathbb{R}^{H_{l} \times W_{l}}, \) a fully connected directed graph \( G^{l} \) is first constructed by connecting every location on \( U^{l} \) with the weight matrix \( D^{l} \in \mathbb{R}^{H_{l} \times W_{l} \times H_{l} \times W_{l}} \), where \( D_{h_{l} + w_{l}, h', w'} \) indicates the weight of the edge from node \((h, w)\) to \((h', w')\) and is defined as:

\[
D_{h_{l} + w_{l}, h', w'} \triangleq \| u^{l}_{h, w} - u^{l}_{h', w'} \|. \tag{2}
\]

Before performing propagation, the weights of the outbound edges of each node are normalized to 1, i.e. \( D_{a, b}^{l} \leftarrow \frac{D_{a, b}^{l}}{\sum_{a', b'} D_{a', b'}^{l}} \). Then, a random walk graph propagation algorithm [20] is employed to produce the object-attention proposal \( M^{l} \). As shown in Fig.3.(a), the node \((h, w)\) receives and diffuses mass via the inbound and outbound directed edges respectively. Then this random walk propagation iteratively collects mass at the nodes that have high dissimilarity with their surroundings (i.e. it will highlight the saliency foreground object region), since the edge weight (i.e. transition probability) is proportional to the dissimilarity between two nodes as defined by Eq.2, transitions into such subgraphs is likely, and unlikely if nodes have similar deep features \( u' \). By the globally objectness flow, it can not only collects object evidence but also depresses noise regions. For the convenience of random walk calculation, the proposal map \( M^{l} \) is first reshaped to a \( H_{l}W_{l}\)-dim vector, and then it is updated by the iterative multiplication with the weight matrix \( D^{l} \), i.e. \( M \leftarrow D \times M \).

Moreover, as a common knowledge that the response maps at different layers are of different information, e.g. the bottom layers have better predictive power for shape while the top layers are robust to noisy regions [19, 27], our OAM takes as input multiple Conv. layers \( U^{l_k}, k \in [1, 2, \cdots] \) so as to avoid the influence of noise surroundings (e.g. trees around birds), and each object-attention proposal \( M^{l_k} \) can be obtained by performing random walk graph propagation over \( U^{l_k} \) as follows:

\[
M^{l_k} \leftarrow (\prod_{t=1}^{T} D^{l_k})M^{l_k} \tag{3}
\]

\( M^{l_k} \) is first initialized with value \( \frac{1}{H_{l_k}W_{l_k}} \), and the weight matrix \( D^{l_k} \) is conditional on the response map \( U^{l_k} \), then the iterative update of \( M^{l_k} \) like \( M \leftarrow D \times M \) can be rewritten as the above equation and each \( M^{l_k} \) can reach its stable state with \( T = 10 \). After Eq.3, all \( M^{l_k} \) are reshaped to 2D maps and resized to the same size as the input image, then they are averaged to a single map, i.e. the final object-attention proposal. In experiments, we find using \{ \( k \in [1, 2], \text{and} \ l_1 = \text{Incep4}, l_2 = \text{Incep5b} \} \) is enough, As in Fig.2, with the help of this proposal, the coarse scale image is processed into the finer scale by operation \( \bigodot \), and then the finer scale input will benefit the discriminative learning of the corresponding OA learner.

Remark: The procedure in Eq.3 can be viewed as a variant of the eigenvector centrality measure [22], which outputs a proposal map to indicate the mass of each location on the deep response maps. It can also be regarded as a Markov chain that can reach unique stable state since the chain is ergodic, a property which emerges from the fact that the graph \( G \) is by construction strongly connected. [11, 45]. Moreover, the random-walk based informative-region search is computationally lightweight and parameter-free, thus it can be easily employed in any bedrock CNNs and has no need to
be additionally optimized, i.e. when given the well trained convolutional maps \( U \), it can directly infer the important object region in an unsupervised manner. Finally, recurrently executing OAMs will incrementally improve the discrimination of the corresponding OA learners.

### 3.2. Adversarial CAM

Although the OA learners can produce discriminative representations, each single OA learner will still be likely to suffer from the partial/selective learning behavior, i.e. only focuses on certain attributes (e.g. heads of birds, bellows of car, etc.) which are the easiest ones to reduce the current training risk. Therefore, in order to improve the diversity of each OA root-learner, we further decouple it into \( J \) CA sub-learners which insist to capture various attributes information coming from diverse channel-attention proposals \( cap_{i,j} \) as in Fig.2, and each \( cap_{i,j} \) is produced by a particular channel-attention module (CAMs\(_{i,j}\)). For clarity, we take one CAM for example, as shown in Fig.3.(b), it takes as input \( U^l \in \mathbb{R}^{C_l \times H_l \times W_l} \) of the \( l \)-th layer (omit subscript \( k \) for convenience) and outputs a channel-weighted counterpart of \( U^l \), i.e. \( cap_{i,j} \). We formulate it as:

\[
\begin{align*}
\text{cap}_{i,j} = \text{CAM}_{i,j}(U^l) = U^l \circ \sigma(W^l_{i,j} \delta(W^l_{i,j} \Psi(U^l)))
\end{align*}
\]

where \( U^l \) is first aggregated to a channel descriptor through spatial average pooling \( \Psi \), then is passed through two fully-connected layers (parameterized by \( W^l_{i,j} \in \mathbb{R}^{64 \times C_l} \), \( W^2_{i,j} \in \mathbb{R}^{C_l \times 64} \)) that intend to capture the interactions among channels and are followed by \( \text{ReLU}(\delta) \) and \( \text{Sigmoid}(\sigma) \) respectively. Finally, \( U^l \) is reweighted by the channel-wise multiplication \( \Box \) between 2D map \( I_{i,j} \) and scalar \( \sigma(W^2_{i,j} \delta(W^1_{i,j} \Psi(U^1)))c \), where \( c \) is the channel index.

Since the filters in certain convolutional layers serve as the local attribute detectors to some extent, the resulting deep response maps will reserve the attribute information in channels. In another word, different attributes reside in different channels. Therefore, the reweighting operation in our CAM can be regarded as an attributes selector(e.g. in Fig.3.(b) it selects ‘foot’ and ‘wing’ yet ignores ‘body’), and the concerned cues(within \( cap_{i,j} \)) will be passed to the corresponding CA sub-learner. If there are diverse selectors, i.e. multiple CAMs, capable of capturing different combinations of attributes information in parallel, the corresponding CA learners will focus on different attributes knowledge, and the richness and diversity of each OA root-learner will be significantly improved.

While, to explicitly encourage the diversity of CAMs, directly adding diversity constraints after CAMs might be detrimental, since in fact it is hard to know which channel contains what attribute signal and which kind of channel combination is beneficial to the embedding learning. To this end, as shown in Fig.2, all \( cap_{i,j} \) are first fed into the corresponding subsequent **shared** embedding layers, and the diversity of CAMs are naturally transferred to that of the CA learners. And in order to explicitly encourage the diversity of the CA learners. Then, an adversarial optimization strategy is introduced as the diversity constraint by playing a \( \max \min \) game, i.e. the adversary network tries to minimize the discrepancies among CA learners while CAMs try to maximize these discrepancies as follows:

\[
\begin{align*}
\max_{f^t, g^t; \text{CAMs}} \min_{F^{t'}} \lambda_0 \sum_{j \neq j'} \left\| f^t(CA_{i,j}[x_{n_i}]) - f^{t'}(CA_{i,j'}[x_{n_i}]) \right\|^2_2
\end{align*}
\]

where \( f^t \), \( g^t \) and \( F^t \) indicate the mapping functions of \( F_{Net}^t \), \( G_{Net}^t \) and the adversary network respectively, \( \lambda_0 \) is the loss weight. The above equation is a discrepancy measurement over all \( J \) CA sub-learners that belong to the \( i \)-th OA root-learner. Moreover, to simplify the two-stage optimization in this \( \max \min \) game, a gradients reverse layer (GRL) \([9]\) is introduced as shown in Fig.3.(c), so as to make Eq.4 be equivalent to \( \min_{G^{adv}} \), \( \max_{f^t, g^t; \text{CAMs}} \) and \( \min_{f^t, g^t; \text{CAMs}} \) (\( -L_{adv} \)).

**Remark:** Since at each scale \( t \) the params of \( G_{Net}^t \) in different channel-attention branches are shared as shown in Fig.2, when optimizing \( \min_{f^t, g^t; \text{CAMs}} \) (\( -L_{adv} \)), the only differentiating parts are CAMs, in other words, only diverse CAMs will make \( -L_{adv} \) small. Moreover, the reason of using adversary objective instead of directly constrain CA learners to be different with each other is because the definition of difference between learners is likely to be hard to be artificially made than the similarity, we thus play the above adversarial game so as to allow the automatic learning of the optimal discrepancies between CAMs.

### 3.3. Objective Loss and Regularization Term

Employing multiple OAMs and CAMs, the holistic feature representation \( \varphi[x_{n_i}] \) can be decoupled into a series of attention-specific parts. Then, the binomial deviance loss \([41]\) is utilized and the empirical metric loss over all CA sub-learners can be formulated as:

\[
\begin{align*}
L_{\text{metric}} = \frac{1}{IJ} \sum_{i=1}^{J} \sum_{j=1}^{I} \sum_{p, q} \frac{1}{w_{pq}} \log (1 + e^{-(2\gamma_{pq} - 1)\alpha(D^{(i,j)}_{pq} - \beta)}}
\end{align*}
\]

where \( \alpha = 2 \), \( \beta = 0.5 \) are the scaling and translation parameters resp. \( \gamma_{pq} \) is the penalty coefficient and is set to 1 if \( s_{pq} = 1 \)(i.e. images \( x_p, x_q \) are from the same class), otherwise \( \gamma_{pq} = 35 \). \( w_{pq} \) is the number of positive pairs(negative pairs) if \( s_{pq} = 1 (s_{pq} = 0) \). \( D^{(i,j)}_{pq} = \frac{<CA_{i,j}[x_p]^T, CA_{i,j}[x_q]^T>}{\|CA_{i,j}[x_p]\| ||CA_{i,j}[x_q]\|} \) measures the distance between inputs \( x_p \) and \( x_q \) at learner \( CA_{i,j} \). Notably, if \( J = 1 \), the OA root-learner has no need to be decoupled, thus \( CA_{i,j} = OA_i \). Additionally, the learners are trained on the relatively small datasets (compared to ImageNet \([28]\) from scratch and thus are easier to be over-fitting, following the commonly used weight decay strategy which regularizes the value of parameters not to be too large so as to reduce
over-fitting, we present a regularization term for optimizing each learner, called activation decay, to constraint the representation value as follows:

\[ L_{\text{act}} = \frac{\lambda_1}{2IJN} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{n=1}^{N} \| [CA(i,j) [x_n]]_2 \|^2 + L_{\text{act}i} \]  

(6)

where \( L_{\text{act}i} = \lambda_2 \sum_{n=1}^{N} \text{tr}[(\omega_{ij}w^T_{ij} - I) \odot (\omega_{ij}w^T_{ij} - I)] \)

where \( L_{\text{act}i} \) is used to avoid the trivial solution (i.e. all parameters are optimized to 0) when minimizing \( L_{\text{act}} \). \( \lambda_1, \lambda_2 \) are trade-off hyper-parameters. \( \omega_{ij} \in \mathbb{R}^{d_i \times d_i} \) are the parameters in learner \( CA(i,j) \); \( d_i \) indicates its input channel number. \( \odot \) refers to the Hadamard-Product.

Finally, DeML can be trained by jointly optimizing Eq.4, 5 & 6.

3.4. Discussion

Why decoupling? In ZSIR, the traditional idea of devising loss functions for optimizing an unified metric omits the issues of (1) indiscriminating visual input regions and (2) the partial/selective learning behavior of learner, and thus exhibits limitations in discriminative and knowledge-preserving feature learning that is crucial for ZSIR. However, in our DeML, decoupling this unified metric into multiple separate learners gives us chances and flexibilities to mitigate the above problems individually, i.e. recurrently improve the discrimination via OA root-learners and explicitly encourage the diversity and richness via CA sub-learners, and as a result, the redundancy in features is reduced and the representation ability is encouraged as well. Additionally, in Sec.4, we show that our decoupling idea outperforms the existing methods in ZSIR by a significant margin even without hard-samples mining.

Relation to other ensemble methods: In deep metric learning community, there are some other ensemble works related to us. Yuan et al [43] employ multiple layers at different depths for hard-ware samples mining and then cascade the learned embeddings together. Optiz et al [24, 25] adopt the online gradients boosting and optimize different learners with the reweighted data. Kim et al [13] try to increase the feature diversity via contrastive loss but ignore the importance of learning discriminative metric in ZSIR task. In summary, although these methods aim at learning ensemble metric, they easily suffer from the indiscriminating inputs, and the partial learning behaviors are not effectively constrained as well. We emphasize that our DeML recasts the ensemble problem as a hybrid-attention model that has a clear object of addressing the aforementioned issues individually, and decouples the unified metric into multiple attention-specific parts, so as to explicitly intensify both the discrimination and diversity within each part of features, thus capturing rich knowledge and being capable of generalizing to unseen categories.

Relation to other attention mechanism: Many other attention works, such as spatial-attention [40, 8], channel-attention [4, 12] and semantic-attention [42, 30], employ the task-specific attention modules to mine the most informative signals. While, we emphasize that our DeML is a hybrid-attention system, which dedicates to capturing not only the informative object-attention regions via OAMs but also the diversity of attributes via CAMs. Moreover, OAM differs from works [4, 40, 7] in that it is performed by the random walk graph propagation, which is in a parameter-free manner and can be directly employed without training. And different from [4, 12] where the channel-attention modules are to extract the most informative channel-signals, our parallel CAMs are constrained by the adversary task such that they are different with each other and capable of extracting rich and complementary attributes knowledge. In summary, CAMs can work in conjunction with OAMs, and the generalization and discrimination of the holistic metric are explicitly and naturally ensured.

4. Experiments

Notation: For clarity, the conventional unified metric learning (i.e. directly train \( d \)-dim features with Eq.5) is denoted by \( Ud \), e.g. U512, and we set it as our baseline. The proposed method is denoted by DeML(I,J), where there are \( I \) OA root-learners and \( J \) CA sub-learners in total, \( d \) is set to 512 by default and each OA(CA) learner is \( \frac{512}{I} (\frac{512}{J}) \)-dim, if not divisible, operation floor is performed.

Implementation details: Following the codes released by [3, 23], we choose the pretrained GooglenetV1 [34] as our bedrock network and adopt the same data preprocess method so as to make fair comparisons with other works. The learners, CAMs and adversary networks are initialized with random weights. CAMs are placed after pool3 layer, OAM takes as inputs \( \text{incep4e}_\text{output} \) and \( \text{incep5b}_\text{output} \).

Training: The optimizer is Adam [14] with learning rate \( 1e-5 \) and weight decay \( 2e-4 \). The training iterations are \( 15k(\text{CUB}), 25k(\text{CARS}), 100k(\text{Stanford Online Products and In-shop}) \) resp. We set \( \lambda_0 = 1, \lambda_1 = 0.014, \lambda_2 = 0.25 \) and 10 times learning rate for the learners.

Evaluation and datasets: For fair comparison, we follow [23] to evaluate the retrieval performance by Recall@K metric. For testing unseen images, the outputs of \( I \times J \) learners will be concatenated into the holistic feature after L2 normalization. Then we use the simple cosine distance rule for similarity measurement. The datasets include \( \text{CUB}[36], \text{CARS}[15], \text{Stanford Online Products}[23] \) and \( \text{In-Shop}[18] \). There is no intersection of classes between training and testing sets.

4.1. Results

Effect of activation decay: From Tab.1, one can observe that with the help of our activation decay \( L_{\text{act}} \), the performances of unified metric learning under different feature sizes are consistently improved over both CUB
Figure 5: Training (seen) and testing (unseen) curves on CUB.

Table 2: Effect of attention modules. H and O indicate the retrieval results of features coming from the holistic learner and OA root-learners, resp.

<table>
<thead>
<tr>
<th>Method</th>
<th>CUB (R@1)</th>
<th>CARS (R@1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U512</td>
<td>59.9↑</td>
<td>54.0↑</td>
</tr>
<tr>
<td>U192</td>
<td>52.4</td>
<td>58.3↑</td>
</tr>
<tr>
<td>U384</td>
<td>53.0</td>
<td>69.1↑</td>
</tr>
<tr>
<td>U512+act</td>
<td>55.6↑</td>
<td>77.4↑</td>
</tr>
<tr>
<td>U512</td>
<td>52.9</td>
<td>68.4</td>
</tr>
<tr>
<td>U512+act</td>
<td>56.2↑</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Table 1: Effect of $L_{act}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>CUB (R@1)</th>
<th>CARS (R@1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U64</td>
<td>50.0</td>
<td>65.4</td>
</tr>
<tr>
<td>U64+L_{act}</td>
<td>67.1↑</td>
<td>68.3</td>
</tr>
<tr>
<td>U192</td>
<td>54.9↑</td>
<td>74.6↑</td>
</tr>
<tr>
<td>U384</td>
<td>53.0</td>
<td>69.1</td>
</tr>
<tr>
<td>U512</td>
<td>55.6↑</td>
<td>77.4↑</td>
</tr>
<tr>
<td>U512+act</td>
<td>56.2↑</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Effect of attention modules: We also conduct quantitative comparisons on attention modules in Tab.2. By default, dimension $d$ is set to 512. The model DeML($I=1, J=1$) is very similar to model (U512+$L_{act}$) with only a small difference of an extra single CAM, and from Tab.1 and Tab.2, one can observe that their performances are almost the same (56.2% vs. 56.1% on CUB, 77.6% vs. 77.9% on CARS), implying that capturing the single attentional knowledge via attention module might not be helpful for ZSIR. Moreover, in Tab.2 one can observe that, by fixing $J=1$, the holistic performance (H) will significantly increase with $I$ (56.1% $\rightarrow$ 60.9% $\rightarrow$ 64.9% on CUB, 77.9% $\rightarrow$ 83.4% $\rightarrow$ 85.6% on CARS), and the OA learner at finer scale will be more discriminative than that at coarse scale, e.g. in DeML($I=3, J=1$) on CUB, the third OA learner achieves 60.1%, the second gets 58% and the first gets 54%, indicating that the more discriminable the input is, the more discriminative embedding is. Furthermore, when fixing $I$ to a certain value, decoupling each OA root-learner into multiple CA sub-learners will consistently improve the performances over both CUB and CARS, e.g. when $I=1$, the result (H) increases with $J$ from 56.1% to 59.0% (77.9% to 82.5%) on CUB (CARS), and the similar improvements can be also observed when $I=2$ or 3, revealing that explicitly encouraging the diversity and capturing rich knowledge indeed improve the generalization of metric in ZSIR. This...
Table 3: Retrieval results on CUB [36], CARS [15] and Stanford Online Products [23]. Here, ○ refers to other ensemble methods, the superscript * indicates the results are from model DeML(I=2,J=4) not DeML(I=3,J=3).

<table>
<thead>
<tr>
<th>Method</th>
<th>CUB-200</th>
<th>CARS-196</th>
<th>Stanford Online Products</th>
<th>In-Shop</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@K(%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labeled [33]</td>
<td>47.2 58.9 70.2 80.2</td>
<td>49.0 60.3 72.1 81.5</td>
<td>67.7 83.8 93.0 97.8</td>
<td>- - - -</td>
</tr>
<tr>
<td>N-pair [32]</td>
<td>51.0 63.3 74.3 83.2</td>
<td>71.1 79.7 86.5 91.6</td>
<td>- - - -</td>
<td>- - - -</td>
</tr>
<tr>
<td>Angular [38]</td>
<td>53.6 65.0 75.3 83.7</td>
<td>71.3 80.7 87.0 91.8</td>
<td>79.9 85.0 93.5 98.0</td>
<td>- - - -</td>
</tr>
<tr>
<td>Proxy NCA [21]</td>
<td>49.2 61.9 67.9 72.4</td>
<td>73.2 82.4 86.4 88.7</td>
<td>73.7 - - -</td>
<td>- - - -</td>
</tr>
<tr>
<td>ALMN [2]</td>
<td>53.4 64.8 74.5 84.3</td>
<td>71.6 81.3 88.2 93.4</td>
<td>69.9 84.8 92.8 -</td>
<td>- - - -</td>
</tr>
<tr>
<td>ECAML [3]</td>
<td>55.7 66.5 76.7 85.1</td>
<td>84.5 90.4 93.8 96.6</td>
<td>71.3 85.6 93.6 98.0</td>
<td>- - - -</td>
</tr>
<tr>
<td>HDC(43) ○</td>
<td>53.6 65.7 77.0 85.6</td>
<td>73.7 82.2 89.5 93.8</td>
<td>69.8 84.4 92.7 97.7</td>
<td>- - - -</td>
</tr>
<tr>
<td>BIER [24] ○</td>
<td>55.3 67.2 76.9 85.1</td>
<td>78.0 85.8 91.1 95.1</td>
<td>72.7 86.5 94.0 98.0</td>
<td>- - - -</td>
</tr>
<tr>
<td>ABE [13]* ○</td>
<td>58.6 69.9 79.4 87.1</td>
<td>82.7 88.3 93.1 96.0</td>
<td>74.7 87.9 94.2 98.0</td>
<td>- - - -</td>
</tr>
<tr>
<td>A-BIER [25]* ○</td>
<td>57.5 68.7 78.3 86.0</td>
<td>82.0 89.0 93.2 96.1</td>
<td>74.2 86.9 94.0 97.8</td>
<td>- - - -</td>
</tr>
<tr>
<td>baseline(U512)</td>
<td>52.9 65.0 75.4 83.6</td>
<td>68.4 78.7 86.0 91.2</td>
<td>68.7 84.0 92.7 97.6</td>
<td>81.8 94.1 95.7 96.5 97.9 97.3</td>
</tr>
<tr>
<td>DeML(I=3,J=3)</td>
<td>65.4 75.3 83.7 89.5</td>
<td>86.3 91.2 94.3 97.0</td>
<td>76.1* 88.4* 94.9* 98.1*</td>
<td>88.2 97.0 98.0 98.3 98.6 98.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>DeML(I=1,J=8)</th>
<th>DeML(I=1,J=8) without $L_{adv}$</th>
<th>U64+Laet</th>
<th>DeML(I=3,J=3) without $L_{adv}$</th>
<th>U64+Laet</th>
</tr>
</thead>
<tbody>
<tr>
<td>R@K(%)</td>
<td>59.0 70.2 79.7 87.4</td>
<td>-0.08 ±0.01</td>
<td>50.9 63.9 74.0 82.3</td>
<td>0.92 ±0.02</td>
<td>-</td>
</tr>
<tr>
<td>sim</td>
<td>0.92 ±0.02</td>
<td>-</td>
<td>0.95 ±0.01</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: Effect of adversary loss. *sim indicates the mean cosine similarity between CA learners.

Effect of adversary loss: Unlike OAMs that can automatically extract the object-attention regions via random-walk graph propagation, the CAMs should be extra constrained so as to explicitly encourage the learning of complementary attributes knowledge instead of the partial one. From Tab.4, one can observe that without the constraint of $L_{adv}$, the performances of DeML(I=1,J=8) are drastically decreased over both CUB and CARS (i.e. from 59.0% to 52.0% on CUB and 82.5% to 68.2% on CARS), and the similarities between different CA learners are obviously increased from 0 to nearly 1, verifying that our adversary loss $L_{adv}$ indeed encourages the discrepancies between CA learners (and CAMs) and reduces the neural redundancy, as a result, the diverse and rich information are captured such that the generalization ability is enhanced. Moreover, one can observe that even if there are 8 64-dim learners in (DeML(I=1,J=8) without $L_{adv}$), the performances are close to (U64+Laet), since these 8 learners are supervised independently and thus are all likely to focus on the same attributes knowledge that are the easiest one to reduce the training risk, and then have little differences with each other (‘sim’ is close to 1), implying the redundant learners will not contribute much and the partial learning behavior will turn up easily without the diversity constraint.

Comparison with State-of-the-Art: In order to highlight the significance of our decoupling idea for zero-shot retrieval, we compare DeML with some remarkable embedding works⁴, over the popular used datasets CUB, CARS, Stanford Online Products and In-Shop Clothes. The comparison results are listed in Tab.3. From these tables, one can observe that our baseline(U512) can only achieve the general performances, and the performances of traditional ideas of loss-designing and samples-mining are almost on par with each other, while, by explicitly intensifying both the discrimination and diversity within the deep metric, our DeML can significantly improve the performances over baseline model and outperforms other state-of-the-art methods by a noteworthy margin, demonstrating the necessity of our explicit enhancement for discrimination and generalization via the decoupling idea. Moreover, different from the listed ensemble methods [43, 24, 25, 13], DeML has clear objects of jointly mitigating the aforementioned issues that are vital to ZSIR, and thus can easily surpass them. Worthy of mention is that, we find DeML(I=2,J=4) is enough for Stanford Online Products dataset, since as in Fig.4 the second scale has been capable of localizing the discriminative regions.

5. Conclusion
In this paper, we propose an interpretable DeML, a hybrid-attention based decoupling method, for zero-shot image retrieval task by explicitly intensifying the discrimination and generalization within the learned metric with the proposed OAMs and CAMs. Extensive experiments validate the necessity and effectiveness of our decoupling idea.

Acknowledgments: This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61871052, 61573068, 61471048, and 61375031, and by the Beijing Nova Program under Grant Z161100004916088, DiDi GAIA Research Collaboration Initiative and BUPT Excellent Ph.D. Students Foundation CX2019307.

⁴For fair comparison, we report the reproducing results of [13] when using the same training settings as us, e.g. batchsize and loss function.
References


