BIDIRECTIONAL GLOBAL TO LOCAL ATTENTION FOR DEEP METRIC LEARNING.

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep metric learning (DML) provides rich measures of content-based visual similarity, which have become an essential component for many downstream tasks in computer vision and beyond. This paper questions a central paradigm of DML, the process of embedding individual images before comparing their embedding vectors. The embedding drastically reduces image information, removing all spatial information and pooling local image characteristics into a holistic representation. But how can we determine for an individual image the characteristics that would render it similar to a particular other image without having seen the other one? Rather than aiming for the least common denominator and requiring a common embedding space for all training images, our approach identifies for each pair of input images the locations and features that should be considered to compare them. We follow a cross-attention approach to determine these meaningful local features in one image by measuring their correspondences to the other image. Overall image similarity is then a non-linear aggregation of these meaningful local comparisons. The experimental evaluation on standard DML benchmarks shows this approach to significantly improve over the state of the art.

1 INTRODUCTION

Similarity learning is important for many different tasks in computer vision: classification, detection, face recognition, zero-shot and few-shot learning. Usually similarity learning is trained on one set of examples of similar and dissimilar pairs and later applied to a different set of pairs. In such a way a certain amount of generalization is required when training a model to find similarities between objects. The main goal of the conventional approach to deep metric learning is to train an encoder function E and an embedding function ϕ such that composition $\phi \circ E$ yields a representation that can fully describe input image. And this representation is later used to measure similarities to other images and to retrieve nearest neighbours, i.e. most similar objects with respect to the notion of similarity.

Moreover, we see that conventional approach focuses a lot on the problem of finding image representation. The comparison to another image is performed via feeding individual image representations to the loss function. What is important here is that the representation of an image is fixed and does not change whatever image it is compared with.

Hence this approach is unnatural to the problem of similarities estimation: given a query image - most decisive parts for similarity estimation may change depending on the image we compare it to.

Let us illustrate this idea with the following example. When we have been working with the SOP dataset we have noticed that images of the same bike vary a lot in viewpoint. One image can focus on a saddle another one on the gears and the wheel, see Fig.1. So it can be hard to determine whether these images are of the same bike if only look on the bike specific details. However, it might be useful to notice a unique joint pattern, for example a green carpet on the floor frame color to amplify those details when perform similarity estimation. unique visual feature that can be amplified and focused on only if we observe two images jointly.

But how do we learn this joint similarity? We need to design a mechanism that will somehow blend two images we want to compare together. Furthermore, we need a mechanism to blend information and we also must decide at which level to fuse images. Taking the input pixel representation can be

too coarse, but if we take the final representation yielded by the $\phi \circ E$, we may already loose too much information at this point. This happens mostly because the output of the encoder E, usually a pretrained on ImageNet(Deng et al., 2009) convolutional part of the Resnet-50(He et al., 2016). For an image of size 224×224px we get a tensor of size $7 \times 7 \times 2048$ as the output of E. The projection ϕ includes a pooling operation of some kind and an embedding projection onto the unit sphere of dimensionality 512. So we have a compression rate of ≥ 200. Moreover, this projection also removes all the spatial information. We see that the image first undergoes a severe compression operation and only afterwards is being compared with another image. This is also bad because it may disregard relations between different image parts. This leads to the following necessity we want to fuse information of a pair of samples as early as possible together and we want our representation to be as rich as possible. First, aggregation methods is on of the crucial thing to redesign. Second, information from both images must be fused at the output of E.

There is also a technical side of the problem of conventional approach: pooling of the features into a single representation is a bottleneck for information flow between the loss and the weights we want to adjust. With the recent advancements in computing hardware the trend for increase of image resolution for deep learning becomes apparent. Regarding our problem, the higher the input resolution is, the more lossy becomes the aggregation method described above. For that reason it becomes necessary to find an alternative to the lossy aggregation operation in particular and to the holistic approach based on finding fixed representation in general. Novel approaches must focus on fusing rich image representation and finding features adjusted to a particular image pair, namely for a particular comparison.

Moreover simple pooling methods(aggregation method) like average pooling or max pooling of features result in information blurring, which becomes a bigger problem when scaling image resolutions which prevents effective training of high resolution input. Talk about other results on this experiment. Mention that we do not need true hi-res, but we use the upsampled version. Also mention that we do not need extra parameters.

We suggest an alternative to the holistic approach. We design a novel bidirectional global to local attention mechanism that facilitates more direct similarity learning between rich image representation and aggregates all individual similarities better better then the conventional approaches. Our attention mechanism can better fuse features together and turn a similarity into a truly pair-based concept.

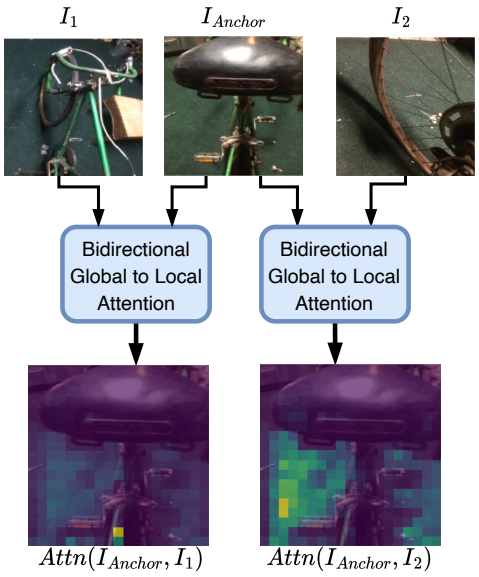


Figure 1: When comparing anchor image I_{anchor} to I_1 and I_2 our bidirectional global to local attention block can spot features present in both images and not just relevant for the class "bicycle". We see that when comparing image I_{anchor} with I_1 model looks at a frame and green floor in I_{anchor} image, since these are only two common details between objects(bottom left). When comparing I_{anchor} with I_w the only detail common is the green floor, so it accummulates all the attention in I_{anchor} image(bottom right).

Through extensive experiments we show that pair-based similarity learning is being superior to the image-based similarity learning in terms of retrieval performance. We study individual elements of the novel bidirectional global to local attention mechanism and provide meaningful insights into the decision making process of our approach. We also show that our method can be combined with classic DML losses and can significantly boosts their performance and make them outperform state-of-the-art approaches which are full of heavy machinery used for training them. We also observe that our method can scales much better with the input image resolution compared to other methods, thus indicating that we have a better training signal.

2 RELATED WORK

2.1 DEEP METRIC LEARNING.

Deep Metric Learning (DML) (Roth et al., 2020b; Musgrave et al., 2020; Milbich et al., 2021) is one of the leading lines of research on similarity learning and related applications, such as image retrieval and search (Sohn, 2016; Wu et al., 2017; Roth et al., 2019; Jacob et al., 2019) or face recognition (Schroff et al., 2015; Hu et al., 2014; Liu et al., 2017; Deng et al., 2019), and even influenced the advance of self-supervised, contrastive representation learning (He et al., 2020; Chen et al., 2020; Misra & Maaten, 2020). With the goal of optimizing individual image projections into an expressive embedding space such that similarity relations between the images are reflected by a given distance metric, a multitude of different approaches for learning have been proposed. The main problem formulation of DML are surrogate ranking tasks over tuples of images, ranging from simple pairs (Hadsell et al., 2006) and triplets (Wu et al., 2017; Schroff et al., 2015) to higher-order quadruplets (Chen et al., 2017) and more generic n-tuples (Sohn, 2016; Oh Song et al., 2016; Hermans et al., 2017; Wang et al., 2019). These ranking tasks sometimes include geometrical constraints (Wang et al., 2017; Deng et al., 2019). To make learning feasible despite the exponential complexity of tuple combinations, such methods are often combined with tuple sampling strategies following either manually defined (Wu et al., 2017; Schroff et al., 2015; Xuan et al., 2020) or learned heuristics (Ge, 2018; Harwood et al., 2017; Roth et al., 2020a). Often, this issue is also successfully alleviated by class proxies representing entire sets of training images such as NCA formulations (Goldberger et al., 2005; Movshovitz-Attias et al., 2017; Kim et al., 2020; Teh et al., 2020; Qian et al., 2019) or classification-based approaches (Deng et al., 2019; Zhai & Wu, 2018). Finally, extensions of these basic formulations further improved the out-of-distribution generalization capabilities of the learned embedding spaces, e.g by leveraging multi-task and ensemble learning (Opitz et al., 2017; 2018; Sanakoyeu et al., 2021; Roth et al., 2019; Milbich et al., 2020; Kim et al., 2018), generating synthetic training samples (Duan et al., 2018; Lin et al., 2018; Zheng et al., 2019; Gu et al., 2021; Ko & Gu, 2020), diverse, complementary feature semantics (Milbich et al., 2020; Milbich et al., 2020), self-distillation (Roth et al., 2021) or sample memory banks (Wang et al., 2020).

All the above works follow the predominating paradigm of determining image similarity by comparing mutually independent, holistic image projections in the embedding space. Thereby, they rely on the rationale that features shared by similar images are implicitly similarly encoded in the latent encoding. In our work, we break this paradigm and design a bidirectional global to local attention module that explicitly identifies and links local, shared image features for estimating similarity. Most similar to our work is the work of Seidenschwarz et al. (Seidenschwarz et al., 2021) and Elezi et al. (Elezi et al., 2020), which use self-attention, respectively label-propagation to exchange messages between standard, holistic image embeddings to incorporate global structure into the embedding space. Moreover, DIML (Zhao et al., 2021) similarly to our work proposed an interpretable DML framework operating on local features. However, correspondences are established by solving an expensive optimal transport problem. In contrast, our approach is based on an efficient cross-images attention mechanism, thus allowing us to greatly scale the spatial maps of local features.

2.2 ATTENTION MECHANISMS.

The attention mechanism allows neural networks to explicitly focus on dedicated parts of the model input (Jaderberg et al., 2015), feature representations (Vaswani et al., 2017) and even output (Jaegle et al., 2021a). Introduced as hard attention, Spatial Transformers (Jaderberg et al., 2015) proposed a differentiable input sampler. The powerful formulation of soft (self-)attention was pioneered by transformers (Vaswani et al., 2017) which revolutionized the field of natural language processing and recently also gain influence in the vision domain (Dosovitskiy et al., 2021). Finally, cross attention has been shown to be a flexible concept for relating two arbitrary data representations (Jaegle et al., 2021b;a), e.g. for effectively scaling Vision Transformers (Dosovitskiy et al., 2021) to large input images. In our work, we formulate a bidirectional global to local attention mechanism to find correspondences between images.

2.3 EXPLAINABILITY IN DEEP LEARNING.

Deep Metric Learning methods typically are difficult to interpret due to the holistic nature of the optimized latent embedding spaces. ABE (Kim et al., 2018) uses an self-attention mechanism for learning an ensemble of global learners to implicitly focus on different parts of the input image. However, *(i)* attention is not performed between images, thus only masked image regions that are captured by a particular learner can be visualized and *(ii)* those image regions are only consistent for very attention channels. In contrast, our approach explicitly establishes local correspondences between images, which are used to determine individual similarities between object parts. These correspondences naturally allow to visualize fine-grained relations between objects that the model considers crucial for similarity assessment. Similarly, DIML (Zhao et al., 2021) aims at finding local object correspondences, which, however, are limited to coarse object parts only, due to computational restrictions limiting the number of independent image regions to be represented. A widely used visualization in DML are UMAP (McInnes et al., 2018) or tSNE (Maaten & Hinton, 2008) projections of the holistic image embeddings. While such visualizations help to show which images are overall similar and dissimilar, they only implicitly provide insights into why a model puts two images next to each other on the embedding manifold.

3 APPROACH

Lets first recap the conventional approach to Deep Metric Learning. The task is given an input image I find such an embedding e such that it satisfies label relations to the other samples in the dataset. Usually, the image I is fed first into the encoder network E and then mapped onto the manifold using embedding function ϕ. This gives us a representation $e = \phi(E(I))$ in a d dimensional space on a $d-1$ dimensional unit sphere $\mathbf{S}^{d-1} := \{x \in \mathbb{R}^d \mid \|x\| = 1\}$.

To satisfy relationships between dataset labels networks measure similarity between images $I_1,\ I_2$ by computing a distance between embeddings $\phi(E(I_1))$ and $\phi(E(I_2))$.

Thus, it is assumed that image is fully represented using its embedding $\phi(E(I))$. The training signal is computed only after plugging distances between embeddings $d(\phi(E(I_1)), \phi(E(I_2)))$ into the loss function used for optimization. As the reader can notice, the images do not interact until the distance between the points is computed, hence all the computations are performed on the per image basis.

Moreover, training signal passes though the lossy process of compression inside of an embedding function ϕ. However, images contain plenty of information and compressing this information by means of some simple pooling method in the function ϕ can be detrimental to the performance.

To give you exact numbers: the most widely used encoder network E is the convolutional part of the Resnet-50 network. For an input image I_1 of size 224×224 pixels we obtain a spatial tensor $F_1 := E(I_1) \in \mathbb{R}^{h \times w \times d}$, where $h = 7,\ w = 7,\ d = 2048$. This representation has much more space to store useful information compared to the final embedding $e_1 := \phi(E(I_1)) \in \mathbb{R}^d$, where d is usually 128 or 512. This results in a compression rate of ≈ 200 between F_1 and e_1.

These are two flaws of the representation seeking approach when applied to the problem of similarity learning - no interaction between images when computing their embeddings and lossy aggregation procedure. Additionally, a holistic approach can not explain which parts of an image are important for similarity and which are not.

Thus, we need a mechanism to directly compare F_1 with $F_2 := E(I_2)$, not e_1 with $e_2 = \phi(E(I_2))$. Since $F_1, F_2 \in \mathbb{R}^{h \times w \times d}$ are of extremely high dimensionality, we can not just flatten this representation and feed it into the fully connected layer - this would have been computationally ineffective. Instead, we need a mechanism that can effectively estimate which parts across a pair of images to compare and how to weight those similarities.

If we do not know what to compare we may throw information we need before even having a chance to find out this information was useful.

The well established way to estimate which parts of an input must be related and processed jointly is the attention mechanism introduced by (Vaswani et al., 2017).

However, if we compute attention between F_1 and F_2, the result is a matrix of size $hw \times hw$ which indicates correlation between different sites of those images. This set of correlations can be dominated by correlations between irrelevant parts of an image. For example, for birds classification task we can have the highest correlations between blue sky segments in both images, though this information is useless for the task of birds discrimination.

For that reason we must know what to relate - what part is that and how meaningful it is? Additionally, we want to learn how similar two different parts are? For that reason we split the representation $F = E(I)$ into parts embeddings $F^P := \pi^P(F) \in \mathbb{R}^{h \times w \times d}$ and similarities embeddings $F^S := \pi^S(F) \in \mathbb{R}^{h \times w \times d}$. $\pi^P,\ \pi^S$ are defined in the Sec.4.1.

Hence, we need to compare with each other not only on the level of individual parts but with an image as a whole. To have an additional global representation of an image we maxpool the parts representation F^P together across dimension $h \times w$ and obtain $g := \pi^G(F^P)$. Detailed description of π^G is provided in the Sec.4.1.

That means we want to compare g_1 with all parts from F_2^P and g_2 with all parts from F_1^P. This is more efficient then comparing exhaustively individual tokens from F_1^P and F_2^P. For example, sky patches are present in both images and have high correlations but they are not important for discrimination.

For the sake of simplicity from now on we assume that all F^P, F^S are reshaped to the shape $hw \times d$. This should remove ambiguity of the matrix calculus below.

Moreover, we want our method to focus on those details of image I_2 which are important for image I_1. Therefore, we want to relate g_1 with F_2 to enable amplification of tokens of F_2 which are highly correlated with g_1. Even though they might have been unnoticeable in F_2 on its own.

We find the importance of parts of image I_2 to image I_1 as a whole by computing the attention of between local parts of I_2 and global representation g_1 of I_1 $\mathrm{softmax}(\frac{g_1 F_2^p}{\sqrt{d}}) \in \mathbb{R}^{1 \times hw})$. And vice versa we compute attention $\mathrm{softmax}(\frac{g_1 F_2^p}{\sqrt{d}}) \in \mathbb{R}^{1 \times hw})$ for attention between parts of I_2 to image I_1 as a whole. Expressions above tell us which parts must be related. Now we need to estimate similarity between individual local parts. This can be formulated as $S := F_1^S \left(F_2^S\right)^T$.

Now we have similarities between individual parts and importance of inidividual parts. Next we combine those two concepts together:

$$s(F_1, F_2) := softmax\left(\frac{g_1 {F_2^p}^\top}{\sqrt{d}}\right)\left(F_2^s {F_1^s}^\top\right) softmax\left(\frac{F_1^p g_2^\top}{\sqrt{d}}\right). \tag{1}$$

We call this computation block consisting of $\pi^P,\ \pi^S, \pi^G$ a bidirectional global to local attention for similarity estimation.

Reader may note a connection of the equation above to the renown attention mechanism widely used for establishing correlations between objects of different nature. Given queries Q, keys K and values V we estimate first correlation between queries and keys $softmax(QK^\top)$. In our case we attention applied from both sides and values V being individual similarities S between different image parts, while attention weighting matrix is the global to local attention between images.

Given the similarity scores between all pairs of points we plug them into any loss function used as a training objective in DML. We use the multi-similarity loss (Wang et al., 2019) to compute the loss for every batch:

$$\mathcal{L} := \frac{1}{b}\left(\sum_{i=1}^{b} \frac{1}{\alpha} \log\left[\sum_{k \in \mathcal{P}_i} \exp^{-\alpha(s(F_i, F_k) - \lambda)}\right] + \frac{1}{\beta} \log\left[\sum_{k \in \mathcal{N}_i} \exp^{\beta(s(F_i, F_k) - \lambda)}\right]\right). \tag{2}$$

The training algorithm is summarized in Alg.1.

Algorithm 1 Training

Require: E - pretrained ResNet-50,
X - dataset with images and class labels,
b - batch size
Initialize E
Initialize layers π^S, π^P, π^G of the similarity cross attention.
while not converged **do**
  Sample b Images with labels $(I_i, l_i) \in X, i \in \{1, .., b\}$
  for $\forall i \in \{1, .., b\}$ **do**
    Compute backbone output $\bar{F}_i$
    Compute similarities $F_i^S = \pi^S(F_i)$, parts $F_i^P = \pi^P(F_i)$
    Compute global representation $g_i = \pi^G(\pi^P(F_i))$
  end for
  for $\forall i, j \in \{1, .., b\} \mid i \neq j$ **do**
    Compute local similarities $S_{ij} = F_j^S {F_i^S}^\top$
    Compute global to local attentions $softmax\left(\frac{g_i {F_j^P}^\top}{\sqrt{d}}\right)$ and $softmax\left(\frac{F_i^P g_j^\top}{\sqrt{d}}\right)$
    Compute final similarity $s(F_i, F_j)$ using Eq.1
  end for
  Compute loss $\mathcal{L}$ specified in Eq.2
  Backpropagate gradients of $\mathcal{L}$ into weights $\theta_{\pi^S}, \theta_{\pi^P}, \theta_{\pi^G}$.
end while

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS.

Implementation details. We follow the common training protocol (Wu et al., 2017; Roth et al., 2019; Sanakoyeu et al., 2021) for DML and utilize an ResNet50 (He et al., 2016) encoder E pretrained on the ImageNet dataset. The model is implemented in the Tensorflow2 framework. All the experiments are conducted on a single RTX 8000 or a single RTX 6000 GPU.
For training, we use the Adam (Kingma & Ba, 2015) optimizer with a fixed learning rate of 10^{-5} and default $\beta_1,\ \beta_2$ parameters with no learning rate scheduling being applied. A default batch size of 32 is used unless stated otherwise. We choose the popular multi-similarity loss (Wang et al., 2019) as our DML objective function using default parameters stated in the original paper. For all the experiments unless stated otherwise we first resize input images to the size 256×256px following standard practice (Musgrave et al., 2020; Roth et al., 2020a) and afterwards artificially upsample them to size 608×608px. At inference time, to further follow standard protocol, we apply center cropping to size 224×224px after the initial resize to 256×256px and then upsamle it back to the our final input size of 608×608px. We discuss the rationale of the upsampling and its benefit for our approach in Sec. 4.3.1.
Datasets. We evaluate the performance on three standard DML benchmark datasets using the default train-test splits:

- *CARS196*(Krause et al., 2013), which contains 16,185 images from 196 car classes. The first 98 classes containing 8054 images are used for training, while the remaining 98 classes with 8131 images are used for testing.
- *CUB200-2011*(Wah et al., 2011) with 11,788 bird images from 200 classes. Training/test sets contain the first/last 100 classes with 5864/5924 images respectively.
- *Stanford Online Products (SOP)*(Oh Song et al., 2016) provides 120,053 images divided in 22,634 product classes. 11318 classes with 59551 images are used for training, while the remaining 11316 classes with 60502 images are used for testing.

Architecture design. The design of the mappings π^P, π^S, π^G is inspired by the design of the transformer encoder of the vision transformers(Dosovitskiy et al., 2021). Both π^P, π^S perform

layer normalization of the input and follow that by a single fully connected layer. π^G performs max pooling across hw channels, followed by another fully connected layer and L_2-normalization.

Evaluation procedure. Our method computes similarity score directly between a pair of images images. In order to compute R@k for every query image we need to compute its similarities to all the other neighbours in the dataset. This results in a quadratic complexity at evaluation step, since we need to porcess all pairs of images. To circumvent this nuisance we compute and store all intermediate embeddings F and the global parts embeddings g. The latter is used to compute nearest 100 neighbours using these global embeddings. And only for those approximate nearest neighbours we compute similarities with our full method. Using these similarities we rerank approximate neighbours accordingly and compute final retrieval scores. This gives a reasonable time overhead, especially when compared to the exhaustive pairwise similarity computation for all pairs in the dataset. In practice it results in 15% increase in evaluation time.

4.2 Comparison to the state of the art methods

First of all we present how our approach stands against other methods. We evaluate performance on three standard datasets i.e. CUB200 (Wah et al., 2011), CARS196 (Krause et al., 2013) and SOP (Oh Song et al., 2016). We measure the retrieval performance using the widely used Recall@k score (Jegou et al., 2011). Results are summarized in Tab.1. They indicate that our approach significantly outperforms other approaches and validates efficiency of our cross-image similarity estimation. Please note that for the sake of fairness all experiments are performed after applying standard DML image preprocessing - image is first scaled to the size of 256×256px , then we take a central crop of size 224×224px and only afterwards image is upsampled to the size 608×608px. Thus our approach can not benefit from minuscule details visible only in high-resolutional input, see Sec.4.3.1 for the detailed study on the importance of the resolution and fine details.

There is another popular metrics in DML is the NMI (Manning et al., 2010) (Normalized Mutual information) score. We do not report it because our approach yields a single similarity score and essentially eliminates a concept of embedding, thus making NMI score inapplicable to our approach.

4.3 Components of the bidirectional global to local attention module

Let us have a closer look at Eq.1 closely. It consists of two main components: attention between holistic parts embeddings of the first image and parts embedding of the second image $softmax(g_1 F_2^p)\mathbb{R}^{1\times hw}$ and the matrix of local similarities $S = F_2^s F_1^s$.

We can study the effect of each individual component separately. At first we can assume that we do not need any attention between image parts across images. In that case our similarity boils down to the average of the local similarities S, namely final similarity is $\mathbf{1}^T F_2^s F_1^s \mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{d\times 1}$ is a vector of all ones. The R@1 score drops by 8.9pp on CUB dataset and by 6.5pp on the Cars196 dataset for the image resolution 608×608px. We conclude that the parts embeddings F^P are crucial for similarity learning.

We can also ablate effect of individual similarities between global embeddings g_1 and F_2^P and replace it with attention between local parts, namely replace eq.1 with

$$softmax\left(\frac{F_1^p (F_2^p)^\top}{\sqrt{d}}\right) \odot \left(F_2^s (F_1^s)^\top\right) \odot softmax\left(\frac{F_2^p (F_1^p)^\top}{\sqrt{d}}\right). \tag{3}$$

. This has less effect on the final score with 3.5pp and 2.9pp drop in R@1 on CUB200 and Cars196 datasets respectively. This indicates that relation between local and global representation in Eq.1 helps similarity learning.

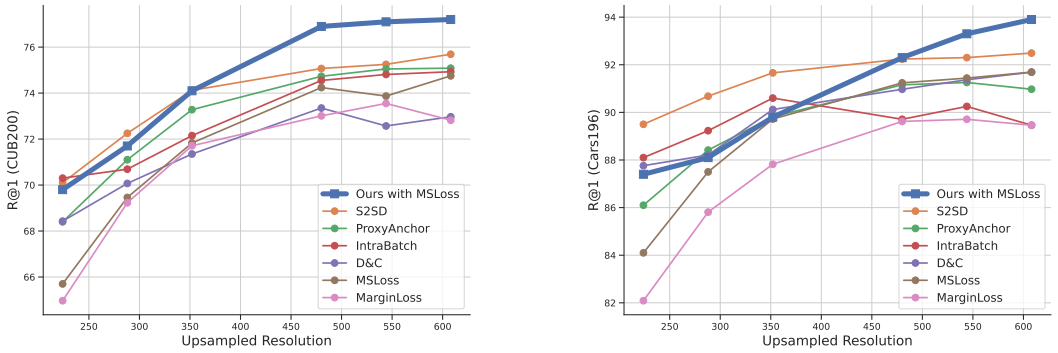
We can completely remove the bidirectional global to local attention mechanism and use baseline projection function ϕ for finding the representation and use cosine similarity for computing the similarity between points. This experiment is provided in Sec.4.3.2. Where we study how does our model performs if coupled with different losses.

		CUB200-2011			CARS196			SOP		
Method	BB	R@1	R@2	NMI	R@1	R@2	NMI	R@1	R@10	NMI
Margin^{128}Wu et al. (2017)	R50	63.6	74.4	69.0	79.6	86.5	69.1	72.7	86.2	90.7
Multi-Sim^{512}Wang et al. (2019)	BNI	65.7	77.0	-	84.1	90.4	-	78.2	90.5	-
MIC^{128}Roth et al. (2019)	R50	66.1	76.8	69.7	82.6	89.1	68.4	77.2	89.4	90.0
HORDE^{512}Jacob et al. (2019)	BNI	66.3	76.7	-	83.9	90.3	-	80.1	91.3	-
Softtriple^{512}Qian et al. (2019)	BNI	65.4	76.4	69.3	84.5	90.7	70.1	78.3	90.3	92.0
ABE^{512}Kim et al. (2018)	G	60.6	71.5	-	85.2	90.5	-	76.3	88.4	-
ProxyNCA++^{512} Teh et al. (2020)	R50	69.0	79.8	71.3	86.5	92.5	71.5	80.7	92	-
XBM^{128} Wang et al. (2020)	BNI	65.8	75.9	-	82.0	88.7	-	80.6	91.6	-
PADS^{128} Roth et al. (2020a)	R50	67.3	78.0	69.9	83.5	89.7	68.8	76.5	89.0	89.9
GroupLoss^{1024} Elezi et al. (2020)	BNI	65.5	77.0	69.0	85.6	91.2	72.7	75.1	87.5	90.8
DIML^{512}Zhao et al. (2021)	R50	67.9	-	-	87.0	-	-	78.5	-	-
D&C^{512}Sanakoyeu et al. (2021)	R50	68.2	-	69.5	87.8	-	70.7	79.8	-	89.7
SynProxy^{512} Gu et al. (2021)	R50	69.2	79.5	-	86.9	92.4	-	79.8	90.9	-
DiVA^{512} Milbich et al. (2020)	R50	69.2	79.3	71.4	87.6	92.9	72.2	79.6	91.2	90.6
ProxyAnchor^{512} Kim et al. (2020)	R50	69.7	80.0	-	87.7	92.9	-	80.0	91.7	-
Intra-Batch^{512} Seidenschwarz et al. (2021)	R50	70.3	80.3	74.0	88.1	93.3	74.8	81.4	91.3	92.6
S2D2^{512} Roth et al. (2021)	R50	70.1	79.7	71.6	89.5	93.9	72.9	80.0	91.4	90.8
Ours ($256 \xrightarrow{\text{upsmpl}} 608$)	R50	77.2	84.7	-	93.9	95.2	-	83.0	92.4	-

Table 1: *Comparison to the state-of-the-art methods on CUB200-2011*Wah et al. (2011), *CARS196*Krause et al. (2013) *and SOP*Oh Song et al. (2016). 'BB' denote the backbone architecture being used ('R50'=ResNet50, 'BNI'=BN-InceptionNet, 'G'=GoogleNet). For our model, '$x \rightarrow y$' indicates the image resizing process (x initial downsampling resolution, y subsequent upsampling resolution.) If y is not specified, the intial downsampling resolution x is used.

4.3.1 RESOLUTION EFFECT

We see an increase in performance with the increase of the image size. In Fig.2 we summarize effect of the increase in image resolution for different methods on different datasets. Majority of the methods benefit to some extend from the increase in image size. However, our attention mechanism that replaces pooling operation helps to unleash the benefits of hi-resolution training.



(a) Our results on the CUB200 dataset.

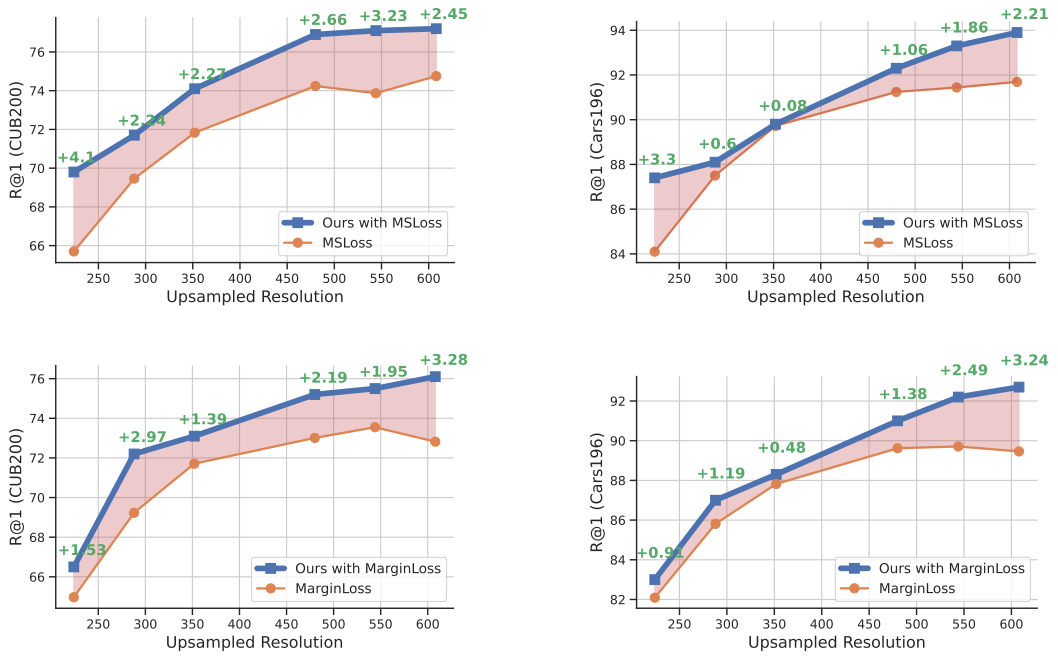
(b) Our results on the Cars196 dataset.

Figure 2: Effect of image resolution of our approach against other methods.

Fine-grained details importance.
As an additional experiment we verify how much performance is lost due to the intermediate downsampling (no downsampling) to the size 256×256px. When no downsampling is performed we can reach 0.7pp higher on R@1 on the CUB200 dataset and only 0.15pp R@1 on the Cars1-196. As we see, our model does not significantly suffer from the missing information of real high-resolution

input. Hence, not additional, fine-grained information is crucial for performance, but the increased number of "tokens" entailed by larger input image resolutions of tensors F^S and F^P.

4.3.2 OTHER LOSSES



(a) CUB200 dataset.

(b) Cars196 dataset.

Figure 3: Effect of the attention module for different image resolutions. We see that the our mechanism has bigger effect with the increase in image resolution.

We also apply our method using other losses used for similarity learning and observe consistent improvement when scaling to larger image size. Thus, our bidirectional global to local attention mechanism for similarity learning is applicable to other methods as well. Though other methods increase the recall scores with the increase in resolution, our method helps to boost this effect. This becomes especially prominent when we go for higher resolutions rates, reaching image size 608×608. In Fig.3 we visualize results for multi-similarity loss and for margin loss(Wu et al., 2017) on the Cars-196 and CUB200 datasets.

5 CONCLUSIONS

We have presented a novel approach to visual similarity learning by abandoning the common paradigm of holistic image encodings. Rather we have framed a similarity learning task as a pair-based approach and not an image-based approach more suitable for a general representation learning. We have designed a novel way to learn and utilize similarities between local regions of the image without any extra labels. Our novel bidirectional global to local attention module splits the task into two parts: what is related and how similar is that. We have provided a visual evidence that the similarity learning may alter its focus within the same image depending on the image we compare it to. On a technical side, we fight a problem of high compression rate of the embeddings mapping function. We have shown that our bidirectional global to local attention similarity learning scales better with increase in resolution compared to the other state-of-the-art approaches and significantly outperform them in retrieval metrics on all three datasets. Our approach is generic and easy to combine with other losses or even more sophisticated approaches to DML. We have also studied the effect of each individual block of our bidirectional global to local attention block.

REFERENCES

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *International Conference on Machine Learning*, volume 119, 2020.

Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. Deep adversarial metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ismail Elezi, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixe. The group loss for deep metric learning. In *European Conference on Computer Vision (ECCV)*, 2020.

Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–285, 2018.

Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In L. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005. URL https://proceedings.neurips.cc/paper/2004/file/42fe880812925e520249e808937738d2-Paper.pdf.

Geonmo Gu, Byungsoo Ko, , and Han-Gyu Kim. Proxy synthesis: Learning with synthetic classes for deep metric learning. In *AAAI Conference on Artificial Intelligence*, 2021.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

Ben Harwood, BG Kumar, Gustavo Carneiro, Ian Reid, Tom Drummond, et al. Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2821–2829, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv e-prints*, art. arXiv:1703.07737, March 2017.

J. Hu, J. Lu, and Y. Tan. Discriminative deep metric learning for face verification in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Pierre Jacob, David Picard, Aymeric Histace, and Edouard Klein. Metric learning with horde: High-order regularizer for deep embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*. 2015.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. *CoRR*, abs/2107.14795, 2021a.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021b.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2011.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Wonsik Kim, Bhavya Goyal, Kunal Chawla, Jungmin Lee, and Keunjoo Kwon. Attention-based ensemble for deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.

Byungsoo Ko and Geonmo Gu. Embedding expansion: Augmentation in embedding space for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

T. Milbich, K. Roth, S. Sinha, L. Schmidt, M. Ghassemi, and B. Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Timo Milbich, Karsten Roth, Homanga Bharadhwaj, Samarth Sinha, Yoshua Bengio, Björn Ommer, and Joseph Paul Cohen. Diva: Diverse visual feature aggregation for deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Timo Milbich, Karsten Roth, Biagio Brattoli, and Björn Ommer. Sharing matters for generalization in deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 360–368, 2017.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Bier-boosting independent embeddings robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5189–5198, 2017.

Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Karsten Roth, Biagio Brattoli, and Bjorn Ommer. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Karsten Roth, Timo Milbich, and Bjorn Ommer. Pads: Policy-adapted sampling for visual similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, 2020b.

Karsten Roth, Timo Milbich, Björn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. S2sd: Simultaneous similarity-based self-distillation for deep metric learning. 2021.

Artsiom Sanakoyeu, Pingchuan Ma, V. Tschernezki, and Björn Ommer. Improving deep metric learning by divide and conquer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Jenny Seidenschwarz, Ismail Elezi, and Laura Leal-Taixé. Learning intra-batch connections for deep metric learning, 2021.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pp. 1857–1865, 2016.

Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *European Conference on Computer Vision (ECCV)*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2593–2601, 2017.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.

Hong Xuan, Abby Stylianou, and Robert Pless. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning, 2018.

Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with structural matching. In *ICCV*, 2021.

Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.