Adding Attention to Subspace Metric Learning

Author Name1
ABC@SAMPLE.EDU
Author Name2
XYZ@SAMPLE.EDU
Author Name3
ALPHABETA@EXAMPLE.EDU

Editors: Under Review for MIDL 2020

Abstract

Deep metric learning is a compelling approach to learn an embedding space where the images from the same class are encouraged to be close and images from different classes are pushed away. Current deep metric learning approaches are inadequate to explain visually which regions contribute to the learning embedding space. Visual explanations of images are particularly of interest in medical imaging, since interpretation directly impacts the diagnosis, treatment planning and follow-up of many diseases. In this work, we propose a novel attention-based metric learning approach for medical images and seek to bridge the gap between visual interpretability and deep metric learning. Our method builds upon a divide-and-conquer strategy, where multiple learners refine subspaces of a global embedding. Furthermore, we integrated an attention module that provides visual insights of discriminative regions that contribute to the clustering of image sets and to the visualization of their embedding features. We evaluate the benefits of using an attention-based approach for deep metric learning in the tasks of image clustering and image retrieval using a public benchmark on skin lesion detection. Our attentive deep metric learning improves the performance over recent state-of-the-art, while also providing visual interpretability of image similarities.

Keywords: Metric Learning, Attention Mechanism, Medical Image, Subspace Embedding.

1. Introduction

The perception of similarity and analogy is one basic aspect of human cognition (Vosniadou and Ortony, 1989). Likewise, learning the similarity between arbitrary images is a fundamental problem in many key areas of computer vision such as image retrieval (Sohn, 2016; Movshovitz et al., 2017; He et al., 2018), duplicate detection (Zheng et al., 2016), clustering (Ziko et al., 2018), or zero-shot learning (Zhang and Saligrama, 2016). With the advent of deep neural networks, deep metric learning (DML) has raised as a powerful approach to learn these similarities. Particularly, the goal of DML is to learn an embedding space where images from the same classes are encouraged to be close, whereas images belonging to different classes are pushed away. In recent DML approaches, the loss function can be typically expressed in terms of Euclidean distances or cosine similarities between pairs or tuples of images in the embedding space. Well-known losses employed in DML include: contrastive loss (Hadsell et al., 2006), triplet loss (Wang et al., 2014), lifted structure loss (Oh Song et al., 2016), N-pairs loss (Sohn, 2016), margin loss (Wu et al., 2017) or angular loss (Wang et al., 2017b).
Despite the popularity of DML, there are surprisingly few papers that attempt to explain visually which regions contribute to the similarity between images in embedding networks. These visualizations are of pivotal importance since they provide an efficient mechanism to better understand the learning process underlying the deep models. Recent efforts have been devoted to the interpretability of deep neural networks, resulting in many different approaches (Zeiler and Fergus, 2014; Koh and Liang, 2017; Selvaraju et al., 2017; Chen et al., 2019). Among these methods, Grad-CAM (Selvaraju et al., 2017) has been widely employed for the explainability of deep classification models. This method employs the gradients to highlight the discriminative regions of an image. Nevertheless, since the gradients are not available during testing, directly applying this strategy in embedding networks is not feasible (Chen et al., 2020). On the other hand, few attempts to integrate interpretability in embedding networks have been proposed (Stylianou et al., 2019; Zheng et al., 2019). These attempts also typically require an additional classification branch (Zheng et al., 2019) or a pair or tuple of images during inference (Stylianou et al., 2019). Furthermore, interpretability is of particular interest in medical imaging. Visual explanations of images indeed directly impacts the diagnosis, therapy planning and follow-up of many diseases. In such context, existing DML approaches may, therefore, be inadequate to visually uncover what constitute similarities among a complex set of medical images.

Motivated by this need to bridge a gap between visual interpretability and metric learning, as well as by the scarcity of the DML literature in medical imaging, we propose a novel attention-based deep metric learning approach for medical images. The underlying metric learning method is inspired by the idea of divide and conquer. Particularly, to address the complexity of medical images we propose to follow the approach of (Sanakoyeu et al., 2019) to capture different object attributes, each of them learned in an independent embedding subspace. To specialize the multiple learners we integrate an attention module in feature encoding, which will encourage the learners to focus on the target object. Furthermore, the integrated attention module provides visual insights of discriminative regions that contribute to the clustering of those images and to visualize embedding features.

**Our Contribution.** We extend a state-of-the-art deep metric learning approach and adapt it for medical imaging. Particularly, we include attention modules to (i) focus the attention of each independent learner, (ii) improve the global performance and (iii) visualize the relevant embedded features. In contrast to previous interpretability methods, our model does not require additional classification or pair of images during testing, which reduces the complexity of the model. To evaluate the proposed method we conduct extensive experiments on a public benchmark of a skin lesion dataset, ISIC (Combalia et al., 2019), on the task of clustering and image retrieval, showing that the proposed method achieves better results than other deep metric learning approaches employed up to date on medical imaging.

1.1. Related Work

**Deep Metric Learning.** Metric learning is a popular research topic that has been widely explored in the learning community (Bromley et al., 1994; Weinberger et al., 2006). The seminal work of Siamese Networks (Bromley et al., 1994) represents the first attempt to use neural networks for feature embedding. The idea behind this work is to employ two identical
neural networks that learn a contrastive embedding from a pair of images. With the advent of deep learning, deep metric learning (DML) has gained popularity, becoming a mainstay in many modern computer vision problems, such as image retrieval (Opitz et al., 2017), person re-identification (Liao et al., 2015) or few-shot learning (Snell et al., 2017). In DML, the images are mapped into a manifold space via deep neural networks. Euclidean or cosine distances can then be directly used as a metric distance between two images in this mapped space. Typical losses employed in MDL include contrastive (Hadsell et al., 2006) or triplet loss (Schroff et al., 2015). The contrastive loss (Hadsell et al., 2006) encourages images from the same class to stay closer—in the learned manifold—while pushing away samples from different classes, which should be separated by a given fixed distance. Nevertheless, forcing the same distance for all pairs of images can discourage any potential distortion in the embedded space. In contrast, this assumption is relaxed in triplet loss (Hadsell et al., 2006), which only imposes that negative pairs of images should be further away than positive pairs. Recent works on DML explore generalizations of these losses, which include: lifted structure loss (Oh Song et al., 2016), multi-class N-pair Loss (Sohn, 2016), margin loss (Wu et al., 2017) or angular loss (Wang et al., 2017b), among many others.

**Metric Learning in Medical Image Analysis.** Despite the interest in other domains, metric learning, and more particularly DML, remains almost unexplored in medical imaging. In (Yan et al., 2018), authors investigate the use of DML to model the similarity relationship between lesions in the context of radiology images. To this end, the triplet loss is employed to learn the lesion embeddings. More recently, Gupta et al. (Gupta et al., 2019) also use the triplet loss to learn the underlying manifold space for the task of Mitotic classification. The embedded features are subsequently used as an input of a Support Vector Machine (SVM) for binary class prediction. In (Teh and Taylor, 2019), authors combine the cross-entropy with a contrastive loss to classify whole slide images in digital pathology. Nevertheless, most of these methods are developed for binary classification tasks and do not exploit the embedding space effectively.

![Figure 1: Schematic of the proposed attention-based pipeline. The data are first grouped into K groups in full embedding space and assigned to subspace learners. In the learning phase, each learner only attends the data from its subgroups, while inference time full embedding space is used. The network in the dotted box shares the weights. Best viewed in color.](image)

**Adding Attention to Subspace Metric Learning**
2. Method

Our proposed approach is depicted in Fig. 1. The main idea is to integrate an attention model in deep metric learning such that it guides the learning of subspace metrics with focused attention maps. To achieve this, an attention model is introduced in a feature encoding module. The underlying metric learning splits the original embedding space into $K$ subspaces so that the global embedding is learned by refining its subspaces. The learning of the embedding space is divided into two steps. First, input images are mapped into the full embedding space, where they are clustered into different groups. Second, the learned embedding space is split into subspaces with an individual learner assigned to each one of them. The key idea is that each subspace learner learns from subgroups of images instead of from the whole embedded representation vector. Finally, all learners are combined to generate a full embedding. In the following sections, we first describe the deep metric learning formulation and then introduce the proposed attentive subspace metric learning.

2.1. Preliminaries

Let the training dataset be defined as $X = \{(x_i, y_i)\}_{i=1}^{N}$, where the $i$-th image, $x_i \in \mathbb{R}^m$, and its corresponding class label, $y_i \in \{1, 2, ..., C\}$, are defined by $(x_i, y_i)$. The total number of classes is defined by $C$. The goal of deep metric learning is to learn an embedding function $f_\theta(x_i): \mathbb{R}^m \rightarrow \mathbb{R}^d$ which discriminatively learns embeddings to map semantically similar images in the input space $\mathbb{R}^m$ (i.e., same class) onto metrically close points in the learned manifold $\mathbb{R}^d$. Similarly, semantically dissimilar images in $\mathbb{R}^m$ should be mapped metrically far in $\mathbb{R}^d$. The parameters $\theta$ of the mapping function are typically learned by a convolutional neural network. Formally, the distance metric $d(x_i, x_j): \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ between two images in the embedding space $\mathbb{R}^d$ can be defined as:

$$d(x_i, x_j) = ||f_\theta(x_i) - f_\theta(x_j)||,$$

(1)

where $||\cdot||$ denotes the Euclidean norm. This distance can be minimized in different ways, depending on the loss function employed. In the current work we resort to the Margin loss (Wu et al., 2017):

$$l_{margin}(x_i, x_j) = [\alpha + \mu_{ij}(d(x_i, x_j) - \beta)]_+,$$

(2)

where $\beta$ is the boundary between the similar and dissimilar pairs, $\alpha$ is a separation margin and $\mu_{ij} \in \{-1, 1\}$ indicates whether the images in the pair are similar ($\mu_{ij} = 1$) or different ($\mu_{ij} = -1$).

2.2. Attentive Subspace Metric Learning

The complexity of the original problem can be solved by dividing the problem into smaller sub-problems, which are easier to solve. Specifically, we follow the approach in (Sanakoyeu et al., 2019) and split the learned embedding space $\mathbb{R}^d$ and the data into multiple groups. By doing this, a set of independent metric distances will be learned over the different subspaces and a fraction of the input data, reducing the complexity of the original problem. First,
Adding Attention to Subspace Metric Learning

Let us assume that we have defined $K$ sub-problems. All the original images in $X$ are therein grouped into $K$ groups based on their pairwise distance in the embedding space $\mathbb{R}^d$, for instance, by using K-means. Next, the learned embedded space $\mathbb{R}^d$ is split into $K$ subspaces by slicing the last layer of the CNN into $K$ sub-vectors of the same size, $d/K$. This results in $K$ mapping functions, $f = [f^1, f^2, ..., f^K, ..., f^K]$, where each mapping function $f^k$ will project the images from the original manifold $\mathbb{R}^d$ into the corresponding $d/K$ subspace. Thus, each data sub-group is associated with a particular subspace, where a separate distance metric is learned by an independent learner, $f^K$.

Deep attention has raised as an efficient mechanism to focus the learning on the objects of interest in a wide range of applications, such as person re-identification (Li et al., 2018), object classification (Wang et al., 2017a) or medical image segmentation (Sinha and Dolz, 2019). Inspired by these advances, we introduce an attention module (Fig. 1) to learn attentive features, such that it assists in learning the embedding space. Furthermore, the attentive features also assist in mapping the images into the embedding space $\mathbb{R}^d$, which facilitates the grouping of input images according to the target object. Let the feature encoding be $S$ and the attention model be $A$, then the output of the attention model is given as $A(S(x_i)) \circ S(x_i)$. Note that the attention model adds a few extra parameters, but does not change the mapping dimensionality.

All learners are trained jointly in an alternating manner from the corresponding data group. Images are drawn uniformly from each group with a random mini-batch $B$. The learners are updated by their margin loss function defined as follows:

$$L_{\theta_k}^f(x_i, x_j) = \sum_{(x_i,x_j) \sim B} \left[ \alpha + y_{ij}(d_{\theta_k}^f(x_i, x_j) - \beta) \right]_+,$$

where $(x_i, x_j) \sim B$ are sampled from the current mini-batch having both positive and negative classes, and $d_{\theta_k}^f$ is the distance metric (similar to Eq.1) for the $k$-th learner. Assuming the learned embedding space is improving over time by each learner, we re-group the images at every $T$ epochs by mapping the images using the entire embedding space such that images can be better sampled. The full embedding space is then composed by merging all learners. The entire image set is subsequently used to stabilize the embedding space.

3. Experiments

The performance of the proposed attention-based approach is compared to other deep metric learning methods applied in medical imaging (Gupta et al., 2019; Teh and Taylor, 2019), which use Triplet loss and Contrastive loss, respectively. To assess the effectiveness of the attention model in our approach, we compare with the original divide and conquer approach (DivConq) (Sanakoyeu et al., 2019). Furthermore, since we employ the margin loss as our optimization objective (Wu et al., 2017), we also include the results from this work. For a meaningful evaluation, the model architecture and other parameters are fixed across the different methods as described in Sec. 3.2. Note that the baselines based on margin, triplet and contrastive losses rely on single learners to learn the metric distance for the whole embedding space. This contrasts with models based on the divide and conquer strategy which...
employ multiple learners, each of them associated with an embedding subspace. In the next sections, we present the dataset used for evaluating our approach, as well as the implementation details of the framework. Then, we show the quantitative and qualitative results of the proposed approach compared to the different baselines, in the tasks of clustering and image retrieval.

3.1. Dataset and Evaluation Metrics

The proposed method is evaluated on the skin lesion dataset taken from ISIC 2019 Challenges \(^2\) (Combalia et al., 2019). This dataset consists of 25,331 images across 8 different categories. In our experiments, following the standard procedure in deep metric learning, we split our dataset into independent training and testing sets. Specifically, 20,000 images were used for training and the rest (5,331) for testing. For evaluation purposes, we follow the same evaluation protocol typically employed in deep metric learning (Sanakoyeu et al., 2019; Oh Song et al., 2016). We employ a normalized mutual information (NMI) to assess the clustering performance using K-means and Recall score (with k = 1) to evaluate the image retrieval quality.

3.2. Implementation details

Following the work in (Sanakoyeu et al., 2019) we use ResNet-50 (He et al., 2016) as the backbone architecture for our approach. The feature encoding layer uses three residual blocks of ResNet-50 to input sufficient feature maps size to the attention model. The attention model consists of two 3 × 3 convolution layers with filters size of 128, 32 having a ReLU activation, followed by a 3 × 3 convolution layer with a Sigmoid activation to produce an attentive map. It is then multiplied to the feature encoding layer to get the attentive features and finally mapped to the embedding layer through global average pooling. The embedding size is fixed to \(d = 128\) and input image size of 224 × 224 used for all our experiments. All models are trained using the Adam optimizer (Kingma and Ba, 2015) with batch sizes of \(B = 32\). The number of learners is empirically set to \(K = 8\) and the re-clustering parameter is set to \(T = 2\) as in (Sanakoyeu et al., 2019). The margin loss parameters are set to \(\alpha = 0.2, \beta = 1.2\), as in (Wu et al., 2017). In each mini-batch, 8 images per class are sampled to ensure a class-balanced scenario and experiments are trained for 300 epochs. The last 50 epochs are fine-tuned in the full embedding. The code is publicly available at https://github.com/anonymous3578/paper270.

3.3. Results

**Quantitative Evaluation.** We compare our method with our baselines by evaluating NMI and Recall scores. All our experiments are run 5 times and their average performances are reported in Table 1.

The results indicate that our attentive metric learning performs better than the baseline methods, both in terms of NMI and Recall scores. Compared to all single-learner approaches, our method and DivConq performs better due to multiple-subspace learning.

---

Adding Attention to Subspace Metric Learning

<table>
<thead>
<tr>
<th>Method</th>
<th>NMI</th>
<th>R@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrastive loss</td>
<td>47.74</td>
<td>83.92</td>
</tr>
<tr>
<td>Triplet loss</td>
<td>85.90</td>
<td>94.06</td>
</tr>
<tr>
<td>Margin loss</td>
<td>90.09</td>
<td>98.37</td>
</tr>
<tr>
<td>DivConq (Sanakoyeu et al., 2019)</td>
<td>93.08</td>
<td>99.59</td>
</tr>
<tr>
<td>DivConq (Sanakoyeu et al., 2019) + Attention (Ours)</td>
<td><strong>94.13</strong></td>
<td><strong>99.61</strong></td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparison of performance using NMI and Recall@k. Our method yields the best results, in bold, while also providing visual interpretability of image similarities.

However, adding attention to the subspace learners does enhance the performance over of DivConq.

**Qualitative Analysis.** To show the discrimination between the different classes in the embeddings space, a t-SNE mapping (Maaten and Hinton, 2008) is visualized in Fig. 2 for the test set and it is compared with baseline methods. The Contrastive loss (Fig. 2a) fails to cluster the different classes, whereas the Triplet (Fig. 2c) and Margin loss (Fig. 2d) fails to discriminate class 0 (red) and 1 (yellow). The performance is increased when using DivConq (Fig. 2e, NMI of 93.08%) or our method (Fig. 2f, NMI of 94.13%). Our method yields indeed the best results, while also providing attention maps that has learned similarities among such a complex image set of skin lesions.

Figure 2: Visualization of test data in embedding space using t-SNE. Each class is indicated by individual color. The red arrow in (e) DivConq shows the miss classification compared to our method (f). Best seen in color.
The performance of image retrieval quality for the proposed method is shown in Fig. 3 for a query and its 5 nearest neighbors. Retrieved results show that our method retrieves images having similar lesion and skin colors (all rows), comparable lesion shapes (Row 1, 3 and 4), equivalent image scales (Row 2 and 3), and image semantics. For example, if a query image has a lesion with hair (row 1), our method also retrieves images with hair. This illustrates a robustness of our method to variations of similarities between images while grouping them accordingly within classes. The coherence of image retrievals indicates that the intra- and inter-class similarities has been captured by our attention model and thereby demonstrates a robustness of our learned embedding.

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieved nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Query Image 1" /></td>
<td><img src="image2" alt="Nearest Neighbor 1" /></td>
</tr>
<tr>
<td><img src="image3" alt="Query Image 2" /></td>
<td><img src="image4" alt="Nearest Neighbor 2" /></td>
</tr>
<tr>
<td><img src="image5" alt="Query Image 3" /></td>
<td><img src="image6" alt="Nearest Neighbor 3" /></td>
</tr>
<tr>
<td><img src="image7" alt="Query Image 4" /></td>
<td><img src="image8" alt="Nearest Neighbor 4" /></td>
</tr>
<tr>
<td><img src="image9" alt="Query Image 5" /></td>
<td><img src="image10" alt="Nearest Neighbor 5" /></td>
</tr>
</tbody>
</table>

Figure 3: Qualitative image retrieval results on the test set for the query and its 5 nearest neighbors in ascending order of distance (left to right), illustrating robustness to colors, shapes, scales, and semantics of lesions.

**Analysis of Interpretability.** The learned attention maps, enabled by our method, are shown in Fig. 4. These attention maps reveal the shape of the skin lesions. This indicates that each of our subspace learners has indeed captured what constitutes a lesion on an image, despite their variations in sizes, scales, shapes or semantics.
Figure 4: Attention maps learned by our method for the test set of skin lesion images. The embedding space has learned attentions to variations in lesion size (a and e), the scale of the image (d and f), artifacts such as hair (c), background similarity (b and d). Best seen in color.

4. Conclusion

In this paper, we present a novel attention-based subspace metric learning approach with an evaluation on skin lesion images. A key strength of the proposed method is the interpretability of embedding using learned attention maps. Results demonstrated that attention improves the performance of subspace learners in clustering and image retrieval and it is robust to variation in color, size, shape, scale, and semantics of lesions. Our method can be used as a preprocessing step for further downstream tasks such as classification or segmentation of lesions, which has to be explored in future work.

References


