Attention-based Dynamic Subspace Learners

Sukesh Adiga V SUKESH.ADIGA-VASUDEVA.1@ETSMTL.NET
Jose Dolz JOSE.DOLZ@@ETSMTL.CA
Herve Lombaert HERVE.LOMBAERT@ETSMTL.CA
ETS Montreal, Canada

Abstract

Deep metric learning methods are widely used to learn similarities in the data. Most methods use a single metric learner, which is inadequate to handle the variety of object attributes such as color, shape, or artifacts in the images. Multiple metric learners could focus on these object attributes. However, it requires a number of learners to be found empirically for each new dataset. This work presents a Dynamic Subspace Learners to dynamically exploit multiple learners by removing the need of knowing *apriori* the number of learners and aggregating new subspace learners during training. Furthermore, the interpretability of such subspace learning is enforced by integrating an attention module into our method, providing a visual explanation of the embedding features. Our method achieves competitive results with the performances of multiple learners baselines and significantly improves over the classification network in clustering and retrieval tasks.

1. Introduction

Learning similarity is a key aspect in uncovering the interpretation of anatomical data in medical images. Deep metric learning (DML) technique can learn such similarities in an embedding space by encouraging same class images to be close to each other while pushing away the images belonging to different classes. Most DML methods use a single learner, which is inadequate to characterize the complex distributions of images having different object attributes such as color, shape, size, or artifacts. A multiple learners method has been proposed to address this complexity by splitting the manifold into several subspaces (Sanakoyeu et al., 2019). However, it needs to empirically find the optimal number of learners, which requires a new validation for every new setting, including every use of a new dataset. Another main limitation in the existing DML approaches is visually explaining what constitutes similarities among a complex set of medical images. In contrast to classification models, applying GradCAM (Selvaraju et al., 2017) in embedding networks is not feasible (Zhu et al., 2021) since the gradients are not available during inference.

These limitations motivate our work, which offers a novel dynamic learning strategy in DML approaches. More specifically, we contribute to a novel training strategy that (i) explores the dynamic learning of an embedding and (ii) overcomes the empirical search of an optimal number of subspaces in approaches based on multiple learners. Furthermore, the visual interpretation of the embedding is addressed by integrating an attention module, encouraging the learners to focus on the discriminative areas of target objects.

2. Method

An overview of the proposed approach is depicted in Fig. 1. The main idea is to split the embedding space into multiple subspaces (K) such that the original embedding space can

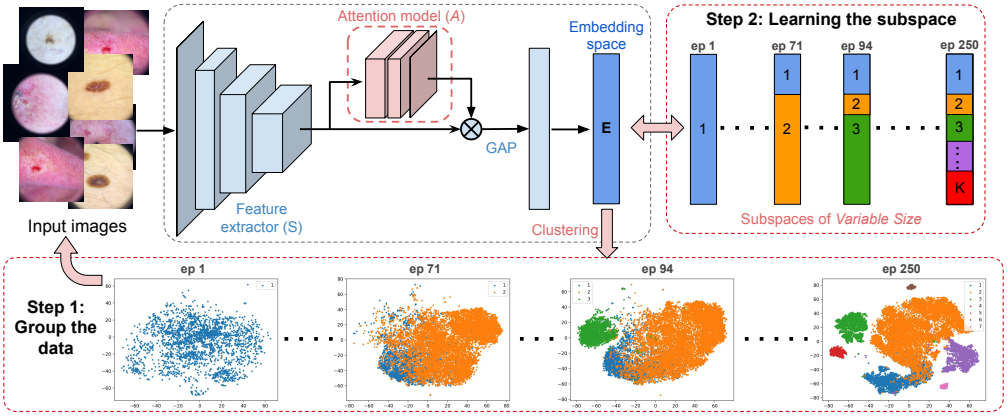

Figure 1: Overview of our proposed method. An embedding space E is learnt by aggregating new subspaces during training. Suppose there are K subspaces, the training data are grouped into K groups in the embedding space (step 1) and assigned to an individual subspace learner. Each learner later only attends the data from its subgroup in the learning stage (step 2). In inference time, the entire embedding space is used to map an image.

be learned by refining its subspaces. Contrary to (Sanakoyeu et al., 2019), the embedding space is split dynamically, which removes the need to search for the optimal number of learners K in each scenario. Training consists of two iterative steps. First, input images are mapped into the lower dimensional space using the entire embedding layer E, where images are clustered into different groups. Second, the clustered data is consequently assigned to an individual subspace learner so that each learner learns a part of the embedding space from a subgroup of images. Each subspace is formed by grouping high-scoring neurons of the embedding layer E, as and when the network accuracy plateaus and assigns a new metric learner. Scoring function of neurons (e_i) is defined as $s(e_i) = \left|\frac{\partial f_\theta}{\partial e_i} e_i\right|$, similar to pruning strategy (Molchanov et al., 2017). The training data is eventually re-clustered with the updated K. The remaining neurons of the embedding layer are reset and used to explore the new subspace. Finally, all subspaces are combined to generate a full embedding space.

Furthermore, an attention module $A(\cdot)$ is integrated into the learning process to enhance the learning of the embedding space. The attention module generates the attention map, which is element-wise multiplied with each feature map, resulting in the set of attentive features. These attentive features are subsequently combined with global average pooling (GAP) and mapped into the embedding space using a dense layer (Fig. 1).

3. Experiments and Results

Our method is evaluated for clustering and image retrieval tasks using NMI and Recall scores on ISIC19 (Combalia et al., 2019) and MURA (Rajpurkar et al., 2018) datasets with a training/testing split of 20,000/5,331 and 36,808/3,197 images, respectively. The performance of our method is compared with single and multiple metric learner (MML) (Sanakoyeu et al., 2019) methods. Note that, MML and our method employ a margin loss, while MML with $K = 1$ is equivalent to the margin loss method. From Table 1, our proposed method consistently achieves the best results in terms of NMI while performing on par with

	ISIC19 dataset		MURA dataset	
Method	**NMI (↑)**	**R@1 (↑)**	**NMI (↑)**	**R@1 (↑)**
Classification network	45.41 ± 1.95	77.85 ± 0.86	71.09 ± 1.25	74.21 ± 0.27
Contrastive loss (single-learner)	31.47 ± 0.39	78.13 ± 0.59	74.28 ± 0.53	71.65 ± 0.53
Triplet loss (single-learner)	50.97 ± 0.61	79.84 ± 0.49	74.41 ± 0.27	74.51 ± 0.78
MML (worst K = 1, 10)	50.53 ± 1.01	**82.84 ± 0.39**	72.88 ± 0.40	73.55 ± 0.16
MML (best K = 6, 1)	55.08 ± 0.83	82.29 ± 0.56	74.67 ± 0.35	75.36 ± 0.79
Ours (free from K)	**55.14 ± 0.87**	82.39 ± 0.11	**74.88 ± 0.09**	**75.52 ± 0.18**

Table 1: Quantitative results on ISIC19 and MURA datasets. The best and second-best results are highlighted in bold and underlined. Performance of MML varies with K while our method yields consistently better results without requiring to find an optimal K value.

the best MML settings on the recall metric. The performance of MML heavily depends on the value of K. For instance, the difference between the worst and best MML model in NMI score can be up to 5% on the ISIC19 dataset. Our method outperforms the classification network and single-learner methods in both scores across datasets, highlighting the potential of exploring embeddings via multiple subspaces. We also evaluated our generated attention maps in a weakly supervised setting, which improves segmentation accuracy up to 15% in Dice score when compared to the recent interpretation method (Wu et al., 2021).

4. Conclusions

This work presents a novel attention-based dynamic subspace metric learning approach for medical image analysis. Our novel training strategy overcomes the empirical search for the optimal number of metric learners while achieving competitive results in clustering and retrieval tasks. Our experiments have shown consistent results with a smaller standard deviation across the datasets, demonstrating the robustness of our method.

References

M. Combalia, S. Puig, et al. BCN20000: Dermoscopic lesions in the wild. arXiv, 2019.

P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning convolutional neural networks for resource efficient inference. ICLR, 2017.

P. Rajpurkar, J. Irvin, A. Bagul, D. Laird, R. L. Ball, A. Y. Ng, et al. MURA: Large dataset for abnormality detection in musculoskeletal radiographs. MIDL, 2018.

A. Sanakoyeu, V. Tschernezki, U. Buchler, and B. Ommer. Divide and conquer the embedding space for metric learning. In CVPR, 2019.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In ICCV, 2017.

T. Wu, J. Huang, G. Gao, X. Wei, X. Wei, X. Luo, and C. H. Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In CVPR, 2021.

S. Zhu, C. Chen, et al. Visual explanation for deep metric learning. IEEE TIP, 2021.