Deep Similarity Learning for Multimodal Medical Images

Xi Cheng, Li Zhang, and Yefeng Zheng

Siemens Corporation, Corporate Technology, Princeton, NJ, USA

Abstract. An effective similarity measure for multi-modal images is crucial for medical image fusion in many clinical applications. The underlining correlation across modalities is usually too complex to be modelled by intensity-based statistical metrics. Therefore, approaches of learning a similarity metric are proposed in recent years. In this work, we propose a novel deep similarity learning method that trains a binary classifier to learn the correspondence of two image patches. The classification output is transformed to a continuous probability value, then used as the similarity score. Moreover, we propose to utilize multi-modal stacked denoising autoencoder to effectively pre-train the deep neural network. We train and test the proposed metric using sampled corresponding/non-corresponding computed tomography (CT) and magnetic resonance (MR) head image patches from a same subject. Comparison is made with two commonly used metrics: normalized mutual information (NMI) and local cross correlation (LCC). The contributions of the multi-modal stacked denoising autoencoder and the deep structure of the neural network are also evaluated. Both the quantitative and qualitative results from the similarity ranking experiments show the advantage of the proposed metric for a highly accurate and robust similarity measure.

1 Introduction

An effective similarity measure for multi-modal medical images is important in many clinical applications such as multi-modal image registration. Statistics-based metrics, such as mutual information [1] and Kullback-Leibler divergence [2], have been proved successful for uni-modal cases where images are similar on intensity and texture. However, they suffer from limitations for the multi-modal case, especially multi-modal deformable registration, mainly because the statistics on the local intensity distribution are insufficient to describe the complex relationship between modalities with different underlying imaging physics.

To tackle this challenge, supervised metric learning is proposed [3, 4]. In contrast to statistics-based similarity metrics, a learning-based method optimizes a metric from training data. This implies that the training phase leads the metric towards specific applications, therefore the metric best separates the observed data. An important factor in machine learning algorithms is data representation. While hand-engineered image features are not guaranteed to work well for all image data, learning-based methods have been developed to learn (shared)

feature representation for uni-modal data [5, 6], or data from different imaging modalities [7] or even different data sources (image and audio)[8, 9].

Unlike traditional image classification or detection problems, where an image/patch can be sent into a standard deep neural network (DNN) classifier directly, learning similarity between images is more difficult. A traditional DNN is not applicable here, as it can be problematic if a concatenation of two images is used as the model input. In addition, the learning objective is not clear. To solve this problem, we propose a supervised deep learning framework to construct a similarity metric across imaging modalities. We also propose to use a multi-modal stacked denoising autoencoder (SDAE) for effectively pre-training the DNN. We use corresponding and non-corresponding CT/MR patches for model learning. We show the advantage of the proposed similarity metric over NMI and LCC by ranking the similarity scores of searched patches within a neighbourhood for a query patch. This strategy is similar to image retrieval. We further evaluate the benefit of using the multi-modal SDAE by replacing it with the uni-modal SDAE, as well as the deep structure of the neutral network by removing one of its hidden layers.

2 Methods

2.1 Similarity Metric Learning

We want to learn a function $f(x_1, x_2)$ to compute a similarity score between a pair of patches x_1 and x_2 from different image modalities. From a supervised learning perspective, we need some kind of ground-truth information for training. In our case, the only information we have between x_1 and x_2 is their state of correspondence. Thus, the goal of our similarity learning is to construct binary similarity function, $g(\cdot)$, that can classify each of the image pairs into its labelled state. To model $g(\cdot)$, which may be very complex, we propose to use a fully connected DNN, whose structure is shown in Fig. 1. The output layer, which has two units, represents the classification result, i.e., "10" for correspondence and "01" for noncorrespondence. It is compared to the label of the training image pairs to drive the optimization of the model parameters.

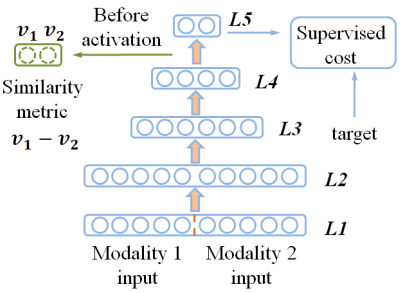

Fig. 1. The structure of a 5-layer DNN. While the 2-unit output is used for supervised training, their values before the activation, i.e., v_1, v_2 are used to form the proposed similarity metric.

The sigmoid output of the learned binary neural network classifier, $g(\cdot)$, indicating the probability of being classified into a particular class, changes too fast between 0 and 1 for a similarity metric. This makes the similarity values almost discrete, which is not desired. Thus, we directly use the values before the

final sigmoid activation, i.e., v_1 and v_2 in Fig. 1. Since the sigmoid function is monotone, v_1 and v_2 still encode the probability of x_1 and x_2 corresponding to each other. After the training stage, $f = v_1 - v_2$ is used as the similarity score.

2.2 DNN Pre-training

Stacked Denoising Autoencoder Autoencoder (AE) is trained to encode an input, x, to a high level representation, y, such that x can be decoded from y. Considering a 1-layer neural network, we can encode the input x into y by $y = \varphi(x) = s(Wx + b)$ and then decode from y to z by $z = \psi(y) = s(W'y + c)$, where $s(\cdot)$ is a non-linear function, e.g., the sigmoid, and z is the reconstruction of x. To minimize the reconstruction error, $L(x, z)$, AE seeks to learn the parameters W, b, and c on the training dataset. Denoising autoencoder (DAE) is an extension of AE. It is trained to reconstruct a clean version of the noisy input. Formally, we first construct x's noisy version, $\tilde{x}$, through a stochastic mapping $\tilde{x} \sim q_D(\tilde{x}|x)$, where q_D can be any function to add noise to the original x. In this work, we use the masking noise, where a fraction of the elements of x are randomly set to 0. The noisy version $\tilde{x}$ is then mapped through AE to a hidden representation $y = \varphi(\tilde{x})$, which is used to reconstruct a clean version of $\tilde{x}$ by $z = \psi(y)$. Stacked denoising autoencoder (SDAE) stacks several AEs. Specifically, all the AEs are trained separately in a layer-wise manner, i.e., the input of a high-level AE is the output from the low-level AE in the previous layer. Once the mapping, φ, has been learned, it is used on the uncorrupted input to produce the representation that will serve as the input for training the next AE. After the SDAE has been built, the output of its highest level can be used as the input to a standalone classifier like Support Vector Machine. Alternatively, a logistic regression or classification layer can be added on top, yielding a DNN amenable to supervised learning.

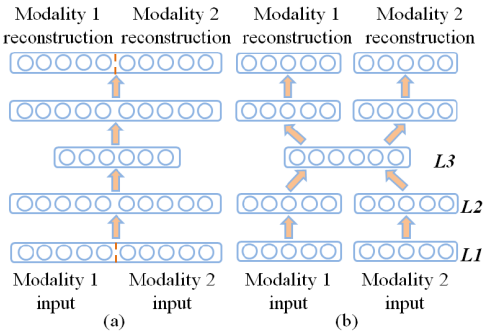

Fig. 2. (a) Uni-modal SDAE with two concatenated images as input; (b) Multi-modal SDAE. Both of them can be used for pre-training layers $L_1 \sim L_3$ of the DNN.

Multi-modal Stacked Denoising Autoencoder In cases where the input contains images from different modalities, a strategy to efficiently model the correlation across modalities is crucial. A direct approach is to train an uni-modal SDAE over the concatenated image data as in Fig. 2(a), which leads to a full connection between the input layer L_1 and the hidden layer L_2. However, due

to the large difference across imaging modalities, it is difficult for the hidden units to model their correlation from the image data directly, which may undermine its merit in initializing the DNN to avoid local optima.

To overcome this limitation, we propose to use a multi-modal SDAE strategy. This model, as shown in Fig. 2(b), applies independent DAEs to each image modality before exploring their correlation. That is, rather than the image data, their higher level feature representations are learned for joint modelling. Since the feature representations are much more similar across modalities than the original image data, the correlation modeling becomes more efficient [8].

The training results from multi-modal SDAE can be used to initialize the bottom three layers ($L_1 \sim L_3$) of DNN. Note that the DNN is fully connected, even between L_1 and L_2. Since the multi-modal SDAE is not fully connected, we initially set those missing connections between L_1 and L_2 to zeros for DNN initialization. The supervised training may further turn some of these connections to be non-zeros, which implies these connections are the correlation that truly exist. For this potential benefit, we propagate the gradients through initially non-existing connections. The higher layer, i.e., L_4, in DNN is still pre-trained with DAE. In a word, the model we propose is a fully connected DNN, whose parameters (weights) connecting its first three layers are initialized using multi-modal SDAE.

We show part of the learned weight matrix W connecting L_1 and L_2 from multi-modal SDAE in Fig. 3, which can be interpreted as the filters for feature detection. For CT patches, most of the learned filters (Fig. 3(a)) look like edge detectors. Moreover, they can detect both straight and curved edges with different orientations, which is consistent with the skull structure in the CT images. The filters for MR patches (Fig. 3(b)), are much more diverse, consistent with the complex texture of and around the skull in MR images. These filters show the advantage of the multi-modal SDAE in learning best features for the data, comparing against to engineered features.

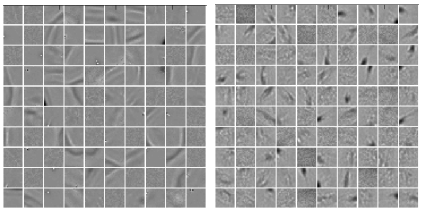

Fig. 3. Filter visualization. First 100 (out of 289) learned filters (17×17 in size) from multi-modal SDAE pre-training for CT image patches (left) and MR image patches (right) using the positive training data.

3 Experimental Results

3.1 Experimental Data and Parameter Selection

We validate the proposed similarity metric with CT and MR head images. To learn the binary DNN classifier, we manually align (rigid transformation) the CT and MR images from a same subject to get positive (matched) and negative (unmatched) training dataset. No apparent deformation is found between

CT/MR scans due to surgery or severe pathological progress. The correctness of the correspondences after rigid alignment was verified manually by the authors. A similar strategy to obtain training data is used in [4].

The CT/MR images are firstly normalized to [0,1]. For positive samples, although we can simply extract patches from all positions of the registered pairs of images, this leads to a too large training dataset to work with. Besides, the similarity measure is only informative in regions with texture and edges rather than homogeneous regions. Therefore, the training dataset is sampled from patches centered in or around skulls. The construction of the negative training samples needs more consideration. There are much more kinds of negative cases than the positive ones, which makes the computational efforts too expensive if using them all. In this regard, we only randomly sample negative MR patches once for each CT skull patch in the positive set, which makes the negative training set the same size of the positive one. On the other hand, as we sample patches centering at all skull voxels, the CT patches present a lot of similarities. That is, some patches are simply translated or rotated versions of others. Therefore, the negative training dataset presents more diversity than the positive set, which increases the discriminating power of the learned classifier. Finally, we use 2000 matched pairs as positive samples and 2000 unmatched pairs as negative samples, resulting in a balanced training dataset. We also did experiment to train a classifier with imbalanced training samples (5 times more negative samples), but didn't achieve better results in the experiments below. With the imbalanced samples, the classifier tends to focus more on the negative samples, therefore becoming inaccurate when classifying a matched pair.

For demonstrative purpose, we perform experiments on 2-D rather than 3-D patches, and experimentally selected a patch size of 17×17. We then construct a DNN with 5 layers (Fig. 1), where the sizes are 578 (i.e., $17 \times 17 \times 2$)$-578-300-100-2$ for $L_1 \sim L_5$. All DAEs for pre-training are applied with a masking noise of 0.3, a learning rate of 0.5, a batch size of 5 for stochastic learning, and 10 loops over the training data. The final parameter tuning by the supervised training is performed with a learning rate of 0.5, a batch size of 5 and in total 5 loops over the training data. This framework is developed using the Deep Learning Toolbox [10], which is implemented in Matlab. The training stage takes about 30 minutes on a Quad-Core processor machine with 2.8 GHz.

3.2 Similarity Metric Evaluation

As we have not incorporated the proposed metric into non-rigid registration, classical target registration error (TRE) cannot be performed to evaluate the proposed similarity metric. Instead, since our training stage is similar to image retrieval, we employ the rank of similarity values as the prediction error to evaluate the proposed metric. The strategy is detailed below. We first randomly select $N = 300$ CT skull patches, and compute the similarity scores for all MR patches centered in a neighborhood of the correct match. The size of the neighborhood is selected to be 81×81, which is larger than the typical search space. As we only show two examples here, a comparison over a larger space can help with a better

comparison of these metrics. Ideally, the similarity score between the CT patch and its matched MR patch should be higher than all others. However, the rigid alignment may be imperfect, therefore the corresponding patches in the training dataset could deviate in voxels. Therefore, we allow the highest similarity score to present within a small $s \times s$ neighborhood of the corresponding MR patch. s is set to 3 in this work, which is considered small enough to enforce good alignment. We compute the rank r of the highest similarity score within this $s \times s$ region, denoted as r_s, which is related to the prediction error. If we denote $\psi_s = r_s - 1$ as the prediction error, $r_s = 1$ means the highest similarity score is correctly assigned: therefore there are no prediction errors, i.e., $\psi_s = 0$. We sort ψ_s for the N CT skull patches in ascending order and plot its cumulative summation Ψ_s^n $(n \leq N)$ to compare different similarity metrics. Clearly, the lower the Ψ_s^n, the better the similarity metric. The cumulative ranking is selected over the average ranking as it reveals the distribution of the rankings. The cumulative ranking at $n = 300$ is simply the average ranking if divided by 300.

To show the advantage of the proposed similarity metric, we first compare its Ψ_s^N to two statistics based metrics: NMI and LCC. To show the importance of multi-modal SDAE and the DNN structure, we also compare Ψ_s^N of the following two metrics: 1) DNN initialized by uni-modal SDAE, referred as *unimodal SDAE - deep NN* or *uni-DNN* for short and 2) neural network with layer L_3 removed and initialized with multi-modal SDAE for L_1, L_2 and L_4, referred as *multimodal SDAE - shallow NN* or *multi-shallowNN*. Naturally, the proposed similarity metric is referred as *multimodal SDAE + deep NN* or *multi-DNN*.

We compare Ψ_s^n of the proposed similarity metric to the other four in Fig. 4. We can see that Ψ_3^n of the proposed metric is much smaller than that of NMI and LCC consistently for all $n \leq 300$, showing its advantage in finding most similar patches. The comparison of the black and red lines shows that DNN is better trained if initialized by multi-modal SDAE than uni-modal SDAE. Further comparing the green and the red lines, we see the necessity of using sufficient hidden layers in DNN for modeling the complex correlation between its input and output layers. This also indicates that there is no overfitting in current DNN.

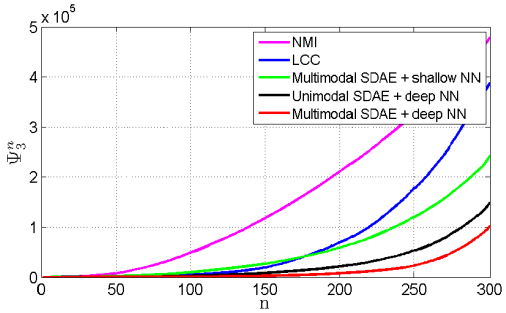

Fig. 4. Comparison of 5 similarity metrics on Ψ_3^n. The cumulative sum of prediction errors is for $n \leq 300$ CT image patches. The worst ψ_3 for a 81×81 neighborhood is 6560, therefore 5×10^5 on the y axis corresponds to a prediction error of about 25% for the 300 patches.

To further examine the prediction error, we choose two representative CT patches with $r_3 = 1$ and $r_3 = 71$, and visualize their 81×81 similarity map

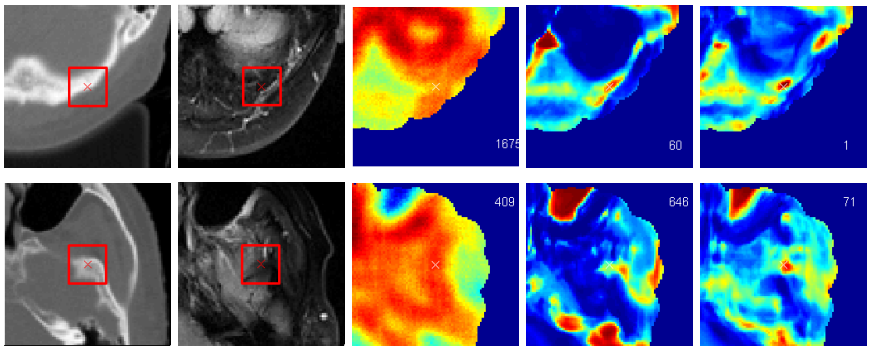

Fig. 5. Two representative examples for comparing the similarity metrics on similarity maps. The 81×81 local similarity values are calculated for the CT patch (the 17×17 red box) and a MR patch within the neighborhood of the corresponding MR match. The CT and MR images are shown in gray scale as in (a)$\sim$(b), while the similarity values are color coded as in (c)$\sim$(e), with red for high values and blue for low values. Within each example, the similarity maps are computed by (c) NMI, (d) unimodal SDAE + deep NN, (e) multimodal SDAE + deep NN (proposed). The similarity ranks of the patch (red box) in (b) given the query patch in (a), computed by different metrics, are listed in (c)$\sim$(e).

computed by 3 (out of 5) metrics in Fig. 5. A good similarity measure should have an unique local maxima for the correct match, i.e., the patch centered within the 3×3 neighborhood of the white cross as in Fig. 5. In the 1^{st} example, this goal is achieved by the proposed metric ($r_3 = 1$) as in Fig. 5 (e), but failed by the others. In the 2^{nd} example, the ranking ($r_3 = 71$) of the proposed metric is much smaller than the others. It is worthwhile to note that if we reduce the neighbourhood size to 20×20 which is a reasonable search space, both uni-DNN and multi-DNN can find the correct match for the two examples. NMI is still incapable though. However, given the better performance of multi-DNN in a larger search region, we can predict it works generally better for other patches.

Further investigation on Fig. 5 (e) indicates that the wrong predictions are located in the eye region (top-left) which has low intensity in MR. Due to the very small size of this region compared to the head, the random extraction of negative dataset fails to sample the patches there, resulting in erroneous similarity measure. A way to handle this issue is to train another binary classifier with the same positive dataset but a different negative dataset consisting of pairs of CT skull patches and eye patches, and cascade it to the existing classifier.

4 Conclusion and Future Work

In this work, we have presented a novel similarity metric for multi-modal images. Specifically, we use the corresponding states of a pair of patches to form a

classification setting, and train a DNN via supervised learning. We also propose to utilize a multi-modal SDAE to effectively pre-train the neural network. We finally construct a continuous and smooth similarity metric based on the output of the neural network, but before the activation in the last layer. We evaluate the proposed similarity metric on 2-D CT and MR patches. The investigation of the assigned similarity scores shows the great advance of the new metric over traditional statistics based metrics such as NMI and LCC, in terms of correctly finding correspondences. Finally, we evaluate the contribution of the multi-modal SDAE for pre-training and the deep structure of the neural network, which also validates the novelty of the proposed framework.

In our previous experiments, we found the ineffectiveness of the traditional similarity measures is the root cause of unsatisfied results for the entire multi-modal deformable registration workflow. Therefore we only focus on similarity measure and evaluate it as a separate component from registration using prediction error for similarity ranking. The advantage of the proposed similarity metric in correctly finding corresponding patches indicates it is promising to improve registration. We will extend the proposed similarity metric from 2-D to 3-D, and incorporate it into our nonrigid registration framework for further evaluation.

References

1. Studholme, C., Hill, D., Hawkes, D.: An overlap invariant entropy measure of 3d medical image alignment. Pattern Recognition **32**(1) (1999) 71–86
2. So, R.W.K., Chung, A.C.S.: Multi-modal non-rigid image registration basd on similarity and dissimilarity with the prior joint intensity distributions. In: ISBI. (2010) 368–371
3. Lee, D., Hofmann, M., Steinke, F., Altun, Y.: Learning similarity measure for multi-modal 3d image registration. In: CVPR. (2009) 186–193
4. Bronstein, M.M., Brostein, A.M., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: CVPR. (2010) 3594 – 3601
5. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.: Stacked denoising autoencoders: Learning useful representations in a deep network with local denoising criterion. Journal of Machine Learning Research **11** (2010) 3371–3408
6. Kavukcuoglu, K., Sermanet, P., Boureau, Y., Gregor, K., Mathieu, M., LeCun, Y.: Learning convolutional feature hierarchies for visual recognition. In: NIPS. (2010) 1090–1098
7. Suk, H., Lee, S., Shen, D.: Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. NeuroImage **101** (2014) 569–582
8. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: International Conference on Machine Learning (ICML), Bellevue, USA. (June 2011)
9. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: NIPS. (2012) 2231–2239
10. Palm, R.B.: Prediction as a candidate for learning deep hierarchical models of data. Master's thesis (2012)