On the Adaptability of Unsupervised CNN-Based Deformable Image Registration to Unseen Image Domains

Enzo Ferrante[1], Ozan Oktay[2], Ben Glocker[2], Diego H. Milone[1]

[1] Research institute for signals, systems and computational intelligence, sinc(i), FICH-UNL/CONICET, Argentina
[2]Biomedical Image Analysis Group, Imperial College London, UK

Abstract. Deformable image registration is a fundamental problem in medical image analysis. During the last years, several methods based on deep convolutional neural networks (CNN) proved to be highly accurate to perform this task. These models achieved state-of-the-art accuracy while drastically reducing the required computational time, but mainly focusing on images of specific organs and modalities. To date, no work has reported on how these models adapt across different domains. In this work, we ask the question: can we use CNN-based registration models to spatially align images coming from a domain different than the one/s used at training time? We explore the adaptability of CNN-based image registration to different organs/modalities. We employ a fully convolutional architecture trained following an unsupervised approach. We consider a simple transfer learning strategy to study the generalisation of such model to unseen target domains, and devise a one-shot learning scheme taking advantage of the unsupervised nature of the proposed method. Evaluation on two publicly available datasets of X-Ray lung images and cardiac cine magnetic resonance sequences is provided. Our experiments suggest that models learned in different domains can be transferred at the expense of a decrease in performance, and that one-shot learning in the context of unsupervised CNN-based registration is a valid alternative to achieve consistent registration performance when only a pair of images from the target domain is available.

1 Introduction

Deformable image registration (DIR) is one of the key problems in medical image computing. It is a crucial step in numerous image analysis tasks, ranging from data aggregation for population analysis to atlas based anatomical segmentation. For more than three decades, the research community has made major efforts towards developing more accurate and efficient registration methods. DIR has been modelled through different approaches, ranging from diffusion equations [15] to probabilistic graphical models [8]. During the last years, we have witnessed the birth of new image registration methods learned from data. Since image data became massively available, and computational power grew powerful enough to

process it, learning-based registration algorithms emerged as an alternative to traditional approaches based on iterative optimization.

CNN-Based Deformable Image Registration. Recently, several DIR methods based on deep learning have been proposed [14, 16, 7, 1]. Most of them aim at learning a function (in the form of a CNN) to predict a spatial transformation mapping a *moving* image to a *fixed* image. These approaches can be categorised into supervised [14] and unsupervised [16, 7, 1] techniques based on how they utilise GT deformation fields. Note that, in the context of DIR, the term *unsupervised learning* refers to the case when no ground-truth annotations such as deformation fields are required for training. An alternative term that has been used in [7] to describe this approach is *self-supervised* DIR, given that learning is driven by image similarity metrics computed on the input data. In this work, we will use both terms interchangeably. Regarding supervised methods, since generating manual ground-truth annotations for DIR is an extremely hard and time consuming task, most supervised approaches resorted to using simulated annotations. The main limitation of such approaches is that their capture range is limited by the ground-truth annotations in the training datasets, which may not always be realistic.

On the contrary, unsupervised approaches like [16, 7, 1] do not inherit this limitation. These methods use a differentiable spatial transformer layer [4] to warp the source image during training, performing end-to-end optimization of a similarity metric between the deformed source and the target input images. The resulting CNN learns to predict (in a single forward pass) the transformation that maximizes such similarity. In this work, we follow a similar unsupervised approach and explore how it adapts to unseen scenarios where the images to be registered correspond to a domain different to that used at training time.

Domain Adaptation for CNN-based Deformable Image Registration. Different from CNN-based methods which learn from data, traditional image registration is usually performed through iterative optimization of a (dis)similarity measure. These methods are slower than CNN-based registration, but they are robust, can be used on unseen domains and work independently of the image resolution. Toolboxes like Elastix [6] for example, which use classical iterative image registration, have been widely applied to align different anatomical structures and image modalities[1]. In contrast, most of the aforementioned CNN-based image registration methods were validated only for specific domains such as brain MRI [7, 1] and cardiac cine-MRI [16]. One of the fundamental questions that still needs to be addressed to enable the development of more robust and reusable CNN-based image registration toolboxes is how to adapt such models to new domains. The recent work [1] shows that, when dealing with multiple datasets of the same imaged anatomy (where the only difference among them is the machine used to capture the images or the acquisition parameters), registration models tailored for a specific dataset outperforms more general models trained on all

[1] A complete list of configuration parameters for Elastix can be found in http://elastix.bigr.nl/wiki/index.php/Parameter_file_database

of them. This observation calls for a deeper adaptation study, focusing on more diverse datasets consisting of different anatomies and modalities. In this work, we show empirical evidence that such adaptation can be performed

Contributions. We emphasize that the main contributions of this paper are not related to novel CNN-based architectures for image registration, but to addressing a more general question about the adaptability of such models. In that sense, our contributions are two-fold: (i) we present an explorative study of the performance of such models when they are trained, fine-tuned and tested on different organs/modalities and (ii) we show that a simple one-shot learning strategy can be used when the only available data is the pair of images to be registered.

2 Materials and Methods

2.1 Datasets and Clinical Context

For validation, we will focus on two clinically relevant applications of image registration with distinct domains, both in terms of anatomy and modality.

Cardiac Cine-MR Dataset: We employ a simplified version of the Sunnybrook Cardiac Dataset (SCD) [10][2]. It contains 45 cine-MR images (every image composed of 6 to 12 short-axis (SAX) 2D slices) captured at end-systole (ES) and end-diastole (ED) time points, amounting to a total of 416 2D images per cardiac phase. Image registration of 2D slices at different phases is crucial in many cardiac image analysis tasks, e.g. when generating strain fields to study left ventricular (LV) (dys)function [9]. After removing 27 slices because of lack of correspondence, we kept 256 pairs for training and 133 for testing (following the same test/train folds specified in the SegNetCMR site), where both images in every pair correspond to the same spatial location at ED and ES. Image resolution is 256x256, covering a field of view of 320 mm x 320 mm. The dataset includes expert annotations for the LV myocardium, which were used for quantitative evaluation. Image intensities were normalized to range [0,1].

Chest X-Ray Dataset: We used images from the chest X-ray dataset of the Japanese Society of Radiological Technology (JSRT) [13]. It includes 247 chest radiographs: 154 with one lung nodule and 93 healthy cases. We generated 247 pairs of images (with resolution 256x256) for registration (199 for training, 48 for testing, randomly split), by using the original image as fixed target and a left/right reversed version as moving image. In this context, DIR is used to warp the flipped image when applying a contralateral subtraction (C-Sub) technique [5], which consists in enhancing nodules in chest images by subtracting their reversed mirror version from the original. Since here we focus on deformable registration, images were previously aligned using affine registration [6]. At test time, we used expert annotations for left and right lungs (included in the dataset) for quantitative evaluation. Image intensities were normalized to range [0,1].

[2] publicly available at https://github.com/mshunshin/SegNetCMR

2.2 Unsupervised CNN-based Image Registration

Inspired by recent works on unsupervised CNN-based image registration [16, 7, 1], we employ a registration model consisting of two main components. The first one follows the U-Net architecture [11], taking the concatenated moving $\mathcal{M}$ and fixed $\mathcal{F}$ images as input and predicting a deformation field $\mathcal{D}_l = \mathcal{U}_l(\mathcal{M}, \mathcal{F}; \Theta_l)$, where $\mathcal{U}_l$ corresponds to a U-Net like CNN, Θ_l to the CNN parameters that have to be learned and l is the down-sampling factor applied to the input images. We perform down-sampling through an initial average-pooling layer in $\mathcal{D}_l$, where the pooling size is 2^l. Following [3], we reduce the model complexity by implementing skip connections via summations instead of the concatenation originally proposed by [11]. The second component is a differentiable spatial transformer module which warps the input moving image $\mathcal{M}$ using $\mathcal{D}_l$, producing a warped image $\mathcal{M} \circ \mathcal{D}_l$.

During training, the parameters Θ are learned using stochastic gradient descent (SGD) so that the warped moving image $\mathcal{M} \circ \mathcal{D}_l$ minimizes a particular dissimilarity measure with respect to $\mathcal{F}$. Since we are dealing with monomodal registration, the negative of the global normalized cross correlation $NCC(\mathcal{M} \circ \mathcal{D}_l, \mathcal{F})$ is adopted. NCC is known to perform well for monomodal cases and has been used in the context of CNN-based registration [7, 1]. A regularization term imposing smoothness constraints is adopted to produce more anatomically plausible deformation fields. Following [7], we consider the total variation of the deformation field $TV(\mathcal{D}_l)$. Finally, an extra regularization term taking the L2 norm of $\mathcal{D}_l$ is included, resulting in the following loss function to be minimized during training:

$$\mathcal{L}(\mathcal{M}, \mathcal{T}, \mathcal{D}_l) = -NCC(\mathcal{M} \circ \mathcal{D}_l, \mathcal{F}) + \lambda_1 TV(\mathcal{D}_l) + \lambda_2 \frac{||\mathcal{D}_l||}{n}, \tag{1}$$

where λ_1, λ_2 are weighting factors for the regularization terms and n is the number of pixels in the image.

2.3 Fine-tuning and One-shot Learning in the Context of Unsupervised CNN-based Image Registration

Learning a discriminative classifier or other predictor in the presence of a shift between training and test distributions is known as domain adaptation [2]. In the context of DIR, such shift may be due to a change in the image modality, acquisition parameters or the organs being imaged. This is a rather common scenario for an image registration toolbox: users may download the software and use it to register a single pair of arbitrary images. In the case of iterative image registration algorithms like those implemented in Elastix, this is not a problem since the method will infer the transformation parameters by iteratively minimizing a similarity measure for the pair of images at hand. However, CNN-based DIR methods need to be trained before they can predict a deformation field. This is one of the main drawbacks of learning based image registration: models are

trained on a *source* domain, and their performance decreases when being applied to images from a different *target* domain (see Section 3 for empirical evidence).

In this work, we focus on the case where the only sample available from the target domain is the actual pair of images to be registered. As stated before, this is a rather common scenario for people working in medical image computing, specially those developing toolboxes which could be used to register arbitrary images. In this scenario, we explore two alternatives:

1. **One-shot domain adaptation**: Here the pair of images to be registered is used to update a model pre-trained using images from a source domain (different to the target domain). Since we are in an unsupervised setting (in the sense that our method does not require image annotations at training time), nothing stops us from using the pair of test images for this update of the model parameters before registering them. We call this strategy *one-shot domain adaptation* by analogy with the concept of one-shot learning where the aim is to recognize categories based on very few training examples [12]. Such adaptation is performed by simply fine-tuning the pre-trained model.
2. **One-shot learning from scratch**: In this case, the CNN registration model is trained from scratch using only the pair of images to be registered and no pre-trained model. This is possible given the unsupervised nature of the approach. This scenario resembles the classic non-learning based iterative image registration algorithms, where the transformation parameters are learned from scratch by minimizing a dissimilarity measure on the pair of images to be registered. In our case, such parameters will be the CNN itself.

We also include results when fine-tuning a model pre-trained with images sampled from the target domain, using the pair of images to be registered. Clearly, this is not a case of domain adaptation, since the model has already been trained with images from the target domain. However, we include it to evaluate if fine-tuning an already good model following a one-shot strategy leads to even more accurate results. Finally, in terms of time restrictions, we consider two scenarios: (i) having real time constraints: the models are trained/fine-tuned for only 50 iterations (0.5s in GPU) and (ii) no time constraints: the models are trained/fine-tuned until convergence using the pair of images to be registered. [3]

3 Results and discussion

In this section we present the results for the alternative studies discussed in the previous section. We use the mean of absolute differences (MAD) between warped moving image and fixed target as an indicator of the quality of the registration since we are dealing with monomodal cases. Moreover, additional indicators reflecting complementary information (namely Dice coefficient (DSC)

[3] Fine-tuning for 50 iterations takes only 0.5s on GPU. When training from scratch/fine-tuning until convergence, we update the model for 3000 iterations leading to about 30s per registration case.

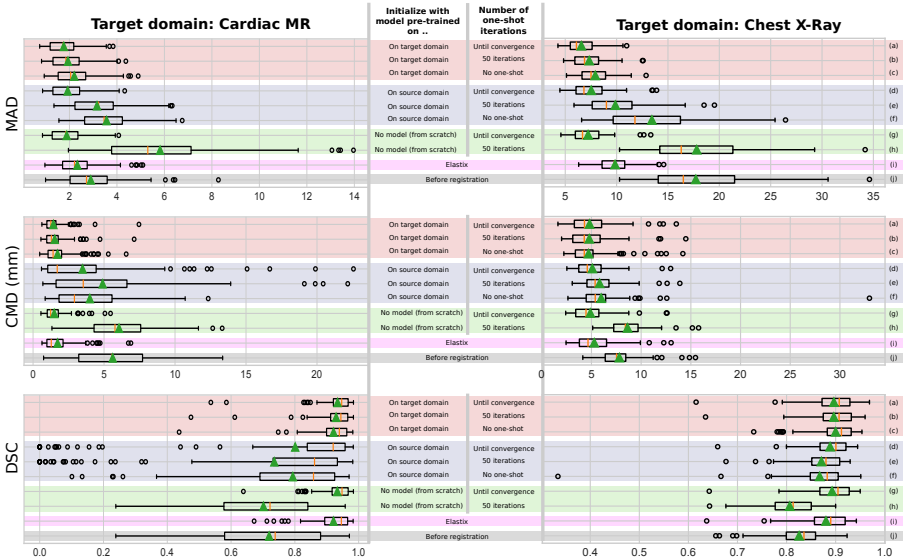

Fig. 1: Experimental validation when fine-tuning/training from scratch using a one-shot strategy. We report results for models pre-trained with images from the target domain (rows a, b, c), from the source domain (rows d, e, f) and without pre-training (rows g, h). See Section 3 for a complete analysis of these results.

and contour mean distance (CMD)) based on warped moving and fixed segmentations are also reported.[4] Results are presented in Fig. 1, considering the cardiac MR images as target domain and chest X-ray as source, and vice versa. We use down-sampling factor $l = 2$ in both cases, since it resulted to be the best performing level in our initial experiments.[5] Let us analyse Fig. 1 in detail.

- Baselines: For comparison, we include results before (row (j)) and after registration using the state-of-the-art Elastix toolbox [6] (row (i)) [6]. At first sight, we can observe that models trained using images from the target domain (row (c)) achieve performance equivalent (and even better) to that of Elastix, significantly outperforming models trained on a different source domain (f). Note that in these experiments, training is performed from scratch as in previous works [1] (i.e., no one-shot strategy is applied).

[4] We used Pyhton and Tensorflow for implementation. Experiments were run in a machine with CPU Intel Core i7-7700, 64GB of RAM and NVidia Titan XP GPU. In order to encourage reproducible research, the project source code and Elastix parameter files can be downloaded from: https://gitlab.com/eferrante/.

[5] The CNN-based models take 0.06s on GPU and 0.08 s on CPU to register a pair of images, while Elastix 2.47s. In all the experiments we used Adam optimization, with LR = 1e-4 and $\lambda_1 = \lambda_2$=1e-6

[6] Elastix parameters were chosen by grid search using the training data and are available online in our project website.

- One-shot domain adaptation: In this case, we fine-tune a model pre-trained on the source domain using just the pair of images to be registered (from a different target domain)[7]. Note that, in order to simulate the one-shot scenario (i.e. just a single pair of images from the target domain is available), we restore the original pre-trained model before performing one-shot domain adaptation for every pair of images. We show results for fine-tuning until convergence (row (d)) and just for 50 iterations (row (e)). On the one side, fine-tuning a model pre-trained on a different source domain does not seem to have a systematic positive impact across all measures in both datasets. On the other side, fine-tuning a model originally trained with other images from the target domain seems to lead to a consistent improvement (row (a)), even when performing just 50 iterations of one-shot fine tuning (row (b)). We hypothesize that initializing the model with weights learned when training in a very different source domain leads the optimization process towards local minima which are not favourable for the target domain registration. However, when test images are closer to the ones used for training, doing one-shot fine-tuning results in systematic improvements. This suggests that one-shot fine-tuning may be a good solution to the multi-site domain adaptation problem reported by [1], when dealing with datasets of the same imaged anatomy captured at different sites or using different machines.
- One-shot learning from scratch: We train the model from scratch using only the pair of images to be registered (i.e. the model is initialized with random weights before registering every pair of images). We show results when training just for 50 iterations (row (h)) and until convergence (row (g)). Interestingly, one-shot learning from scratch (training until convergence) achieves results comparable to Elastix, and even at the level of those obtained when training with other images from the target domain. One could argue that doing one-shot learning from scratch is actually overfitting the model. However, this is not the case since we are following an unsupervised strategy in a one-shot scenario, obtaining a model that is performing well for the data of interest (i.e. the single pair of images from the target domain to be registered). Moreover, this model could be used as initialization to register similar images in the future.

4 Conclusions and future works

We present the first study on domain adaptation across different organs/modalities for unsupervised CNN-based DIR, focusing on the extreme case when a single pair of images from the target domain is available at test time. In this context, we evaluate the performance of a model pre-trained with data from a different source domain, observing a clear decrease in performance when used to register images from the target domain. As a potential solution, we propose one-shot domain adaptation, by fine-tuning the original model using the target pair of images, taking advantage of the unsupervised nature of proposed approach. Pre-training does not seem to help when dealing with images from extremely different

[7] We experimented with fine-tuning the model in whole or in part, but we found that fine-tuning the complete model achieved better results in general.

domains, but it achieves systematic improvement when fine-tuning a model pre-trained on similar images. This opens the door to future research where one-shot domain adaptation could alleviate problems when dealing with multi-site data. Last but not least, we show that one-shot learning for CNN-based DIR (trained from scratch using just the pair of images to be registered) achieves very good results, comparable to those produced by state of the art algorithms.

This work constitutes another step towards constructing more robust deep learning models for image registration. In the future, we plan to extend the validation to volumetric image registration and explore one-shot domain adaptation for multi-site datasets. Moreover, more sophisticated strategies focusing on learning features invariant to image domain could also be adopted.

Acknowledgements: EF is beneficiary of an AXA Research Grant. We thank NVIDIA Corporation for the donation of the Titan X GPU used for this project.

References

1. Balakrishnan, G., et al.: An unsupervised learning model for deformable medical image registration. Accepted at CVPR 2018 (2018)
2. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. ICML (2015)
3. Guerrero, R., et al.: White matter hyperintensity and stroke lesion segmentation and differentiation using cnns. NeuroImage: Clinical (2018)
4. Jaderberg, M., et al.: Spatial transformer networks. In: NIPS. pp. 2017–2025 (2015)
5. Kawaguchi, T., Harada, Y., Nagata, R., Miyake, H.: Image registration methods for contralateral subtraction of chest radiographs. In: IEE BMEI (2010)
6. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE TMI 29(1), 196–205 (2010)
7. Li, H., Fan, Y.: Non-rigid image registration using self-supervised fully convolutional networks without training data. Accepted at ISBI 2018 (2018)
8. Paragios, N., et al.: (hyper)-graphical models in biomedical image analysis. Medical Image Analysis 33, 102 – 106 (2016)
9. Phatak, N.S., et al.: Strain measurement in the left ventricle during systole with deformable image registration. Medical image analysis 13(2), 354–361 (2009)
10. Radau, P., Lu, Y., Connelly, K., Paul, G., et al.: Evaluation framework for algorithms segmenting short axis cardiac mri. The MIDAS Journal 49 (2009)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
12. Salakhutdinov, R., Tenenbaum, J., Torralba, A.: One-shot learning with a hierarchical nonparametric bayesian model. In: ICML Workshop Proceedings (2012)
13. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am Jour of Roent 174(1), 71–74 (2000)
14. Sokooti, H., et al.: Nonrigid image registration using multi-scale 3d convolutional neural networks. In: MICCAI 2017. pp. 232–239. Springer (2017)
15. Thirion, J.P.: Image matching as a diffusion process: an analogy with maxwell's demons. Medical Image Analysis 2(3), 243 – 260 (1998)
16. de Vos, B.D., et al.: End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. DLMIA Workshop, MICCAI 2017. LNCS (2017)