Can Transformers capture long-range displacements better than CNNs?

Paraskevas Pegios PPEGIOSK@GMAIL.COM
DTU Compute, Technical University of Denmark, Denmark
Steffen Czolbe PER.SC@DI.KU.DK.COM
Department of Computer Science, University of Copenhagen, Denmark

Abstract

Convolutional Neural Networks (CNNs) are well-established in medical imaging tackling various tasks. However, their performance is limited due to their incapacity to capture long spatial correspondences within images. Recently proposed deep-learning-based registration methods try to overcome this limitation by assuming that transformers are better at modeling long-range displacements thanks to the nature of the self-attention mechanism. Even though existing transformers are already considered state-of-the-art in image registration, there is no extensive validation of the key premise. In this work, we test this hypothesis by evaluating the target registration error as a function of the displacement. Our findings show that transformers outperform CNNs on a public dataset of lung 3D CT images with large displacements. Yet, the performance difference stems from transformers registering small displacements with higher accuracy. Contrary to previous beliefs, we find no evidence to support the hypothesis that transformers register long displacements better than CNNs. Additionally, our experiments provide insights on how to train vision transformers effectively for image registration on small datasets with less than 50 image pairs.

Keywords: Image Registration, Vision Transformers, Convolutional Neural Networks

1. Introduction

Image registration aims to find geometric transformations that align images. During the last years, CNN-based methods, such as VoxelMorph (Balakrishnan et al., 2019), have attracted wide attention in the field of deformable registration. After training, these methods can significantly speed up medical image processing pipelines while achieving comparable registration accuracy with traditional optimization approaches. The main limitation of CNNs is that they tend to focus on local aspects of images, which is problematic especially when the displacements between the moving and the fixed images become larger than the effective receptive field. Vision transformers lack the inductive biases of CNNs, such as translation invariance and locally restricted receptive fields and their success is usually ascribed to their ability to capture long-range dependencies within an image, even from the shallowest layers. Very recently, (Park and Kim, 2022) questioned this explanation by revealing new intuitions on how vision transformers work. Following the current trend in computer vision and medical imaging, transformer-based models such as ViT-V-Net (Chen et al., 2021b) and TransMorph (Chen et al., 2021a) have been proposed as strong candidates for better modeling of long-range displacements. Even though these models can have a global view of the entire image (Chen et al., 2021a) achieving state-of-the-art results in image registration, there is no extensive validation of the main hypothesis that transformers can capture long-range displacements better than CNNs.

2. Experimental Setup

Given a fixed volume $\mathbf{F}$, and a moving volume $\mathbf{M}$, we seek to predict a transformation $\Phi = Id + \mathbf{u}$, where $\mathbf{u}$ is the displacement field and Id is the identity transformation. The warping is applied using a spatial transformation function, i.e, $\mathbf{M} \circ \Phi$. We model $\mathbf{u}$ using a deep network which can be either a convolutional (VoxelMorph) or a transformer-based (Vit-V-Net, TransMorph) network. In that sense, a network is used to generate the transformation between the images, i.e, $g_{\boldsymbol{\theta}}(\mathbf{F}, \mathbf{M}) = \mathbf{u}$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{dec}\}$ and subscripts *enc* and *dec* denote the parameters of the encoder and decoder part of network respectively. Instead of naive random initialization, we leverage IXI[1] *pre-trained* weights to initialize $\boldsymbol{\theta}_{enc}$, while $\boldsymbol{\theta}_{dec}$ (task specific) are initialized randomly. To the best of our knowledge, *transfer learning* has not been used for image registration because established CNN-based methods can achieve good performance even for small datasets (Balakrishnan et al., 2019). During training, normalized cross correlation is used as distance metric between $\mathbf{M} \circ \Phi$ and $\mathbf{F}$, together with diffusion regularization, weighted by a hyper-parameter λ. A warm-up phase is evaluated by gradually increasing the learning rate up to a specific point and then using standard schedulers. This is a common technique for fine-tuning transformers because layer normalization in multi-head self-attention (MSA) layers can lead to high gradients at early iterations. Intuitively, by taking small steps we prevent adaptive optimizers from going towards wrong directions. Previous studies focused on evaluating transformers mainly in terms of DICE using large datasets. We conduct the evaluation in terms of Target Registration Error (TRE) since our aim is test to the ability of the models to capture long-range displacements. For evaluation, we use the "Learn2Reg: CT Lung Registration" dataset which contains 30 cases of inhaling and exhaling image pairs. Since there are no available landmarks for the original test pairs, we reorganize the dataset and split it (20/5/5), in order to use available keypoints for our validation (6-10) and test cases (1-5).

3. Results & Discussion

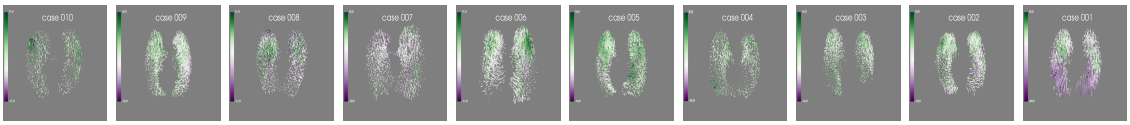

Figure 1: **Comparison of displacement fields between TransMorph^{++} and VoxelMorph-2^{++}.**

For a fair comparison, we tuned λ for baseline (random initialized) and fine-tuned (initialized with encoder pre-trained weights) models using the validation set. Transformers because of the lack of inductive bias required stronger regularization ($\lambda = 1$) than VoxelMorph ($\lambda = 0.5$). We evaluated the irregularity and the smoothness of the transformations and the results are reported in Table 1. Transfer learning proved beneficial not only for transformers but also for VoxelMorph. This can be very useful in practice when working with limited hardware and datasets. As expected, TransMorph benefited the most from transfer learning since its encoder is completely composed of transformer layers. Furthermore, TransMorph outperformed both VoxelMorph and ViT-V-Net, but apart from transfer learning, it required a warm-up phase to improve the smoothness of transformations.

1. https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration/blob/main/IXI/TransMorph_on_IXI.md

A qualitative comparison of the displacements produced by the *fine-tuned* VoxelMorph and TransMorph models is shown in Fig.1. The displacement vectors are colored based on the difference in TRE. The greener the vector the better TransMorph^{++} is while as a vector gets more purple the better VoxelMorph^{++} is. *Diagnostic plots* to measure TRE by the length of the displacement are illustrated in Fig.2. The displacements were binned into approximately evenly-sized bins, in order to determine the mean and a confidence interval for each bin. In this way, diagnostic curves were produced for each *fine-tuned* model aiming to inspect TRE as the displacement length increases. Bin-wise statistical t-tests with Benjamini/Hochberg correction were used to highlight the significant bins (p-value ≤ 0.05) for the pair-wise model comparisons. It is evident that transformers were better at small to medium lengths while for larger displacements there is no such difference.

Warm-up	Model	**TRE** ↓	$\lvert J_\Phi\rvert_{<0}(\%)$ ↓	$\sigma(\lvert J_\Phi\rvert)$
-	Affine	15.34	-	-
-	VoxelMorph-2	11.74	0.61	0.133
-	**VoxelMorph-2^{++}**	**10.58**	**0.45**	**0.120**
-	ViT-V-Net	10.80	0.16	0.090
-	**ViT-V-Net^{++}**	**10.12**	**0.84**	**0.122**
-	TransMorph	11.88	2.29	0.188
✓	**TransMorph^{++}**	**9.91**	**0.34**	**0.105**

Table 1: **Evaluation metrics on our test set (cases 1-5). Transfer learning is denoted with a ++ superscript. TRE is measured in mm.**

Conclusion Overall, transformers outperformed VoxelMorph but the performance gain came from better registering small displacements. To answer the question posed in the title of the paper: *contrary to previous assumptions, we found no evidence to support the claim that transformers register long displacements better than CNNs.* This finding seems to be supported by (Park and Kim, 2022) where it is shown that "the success of MSAs for computer vision is NOT due to their weak inductive bias and capturing long-range dependency".

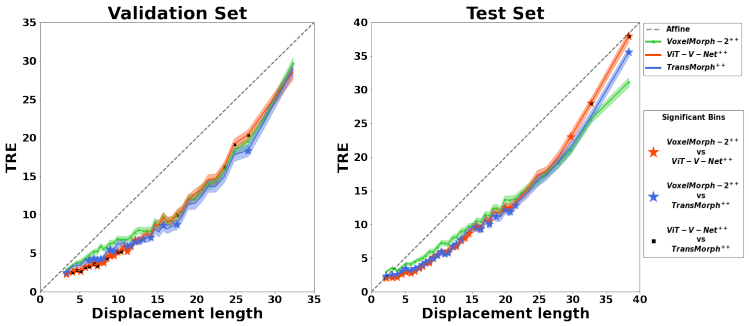

Figure 2: **TRE across the displacement length domain.**

References

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

Junyu Chen, Yong Du, Yufan He, William P Segars, Ye Li, and Eirc C Frey. Transmorph: Transformer for unsupervised medical image registration. *arXiv preprint arXiv:2111.10480*, 2021a.

Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021b.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022.