Attention for Image Registration (AiR): an unsupervised Transformer approach

Zihao Wang, Hervé Delingette

Attention for Image Registration (AiR): an unsupervised Transformer approach

Zihao Wang[1,2] and Hervé Delingette[1]

[1] Inria Sophia-Antipolis, Epione Team, Valbonne, France
zihao.wang@inria.fr
[2] Université Côte d'Azur, Nice, France

Abstract. Image registration as an important basis in signal processing task often encounter the problem of stability and efficiency. Non-learning registration approaches rely on the optimization of the similarity metrics between the fix- and moving images. Yet, those approaches are usually costly in both time and space complexity. The problem can be worse when the size of the image is large or the deformations between the images are severe. Recently, deep learning, or precisely saying, the convolutional neural network (CNN) based image registration methods have been widely investigated in the research community and show promising effectiveness to overcome the weakness of non-learning based methods. To explore the advanced learning approaches in image registration problem for solving practical issues, we present in this paper a method of introducing attention mechanism in deformable image registration problem. The proposed approach is based on learning the deformation field with a Transformer framework that does not rely on the CNN but can be efficiently trained on GPGPU devices also. Our method learns an artificially generated deformation map and be tested on a MINST dataset.

Keywords: Transformer · Images Registration · Deep Learning · Deep Learning

1 Introduction

In many areas such as remote sensing, radar engineering and medical imaging, the problem of matching two images for further processing (eg: images synthesis, stitching or segmentation etc.) is a basic but popular research topic [1, 2, 3]. This task is named image registration which can be divided into rigid and non-rigid registration. The deformable images registration is one of the difficult branches of image registration as the finding of the deformation field between the images to be matched is a highly nonlinear problem. Traditional approaches rely on the collection of images features and using different similarity metrics to measure the matching quality between images pairs during the optimization [4].

With the rapid promotion of deep learning in the field of medical image analysis, a various variant of convolutional neural networks (CNN) are introduced for image registration and obtained eye-catching results in many research works.

Wu *et al.*[5] introduced the CNN for automatic feature extraction to replace the manual feature extraction during registration. Cao *et al.*[6] proposed to learn the non-linear mapping between the input images and the deformation fields with a meticulously designed CNN architecture where the training dataset is prepared with a diffeomorphic demons based registration dataset. De vos *et al.*[7] used an unsupervised Deep Learning Image Registration (DLIR) that using CNN to regress the mapping between the images pairs and the deformation map. The framework is unsupervised as the output of CNN is used directly for warping the moving images to generate the deformed images with transposed convolutions which allow the backpropagate for the resampling operation. Krebs *et al.*[8] proposed to learn a probabilistic deformation model based on variational inference of a conditional variational auto-encoder (CVAE) which maps the deformations representation in a latent space. The trained CVAE can be used to generate deformations for an unseen image.

Since the Transformers based learning models were proposed, it has rapidly expanded from the field of natural language processing (NLP) to the entire machine learning community [9, 10, 11]. As a very successful application of attention conception, the Transformer already conquered the NLP regime and almost replaced the traditional widely-used RNN/LSTM models. The recent advance of Transformers application in image/videos analysis implies that the Transformer have the potential power to show off in structure data processing. The Transformer is a different learning model which does not need convolutional kernels for features representation but learns the data inherent relationships through the **attention mechanism**. Recent works in the computer vision community report that the Transformer based deep learning methods achieved state-of-the-art performance on many datasets. As a representative work that introduces the Transformer in computer vision, the vision Transformer (ViT) was proposed by [12] for image classification. The ViT takes the divided patches p_i of a given image I as the inputs (see Fig. 1 step (a)) and using a Transformer as the attention features extractor to generate features for classification. Another work that using a Transformer for object detection is the DETR [13]. The DETR accepts the image and the detection objects for the Transformer as the inputs and generates the corresponding bounding box information of the target objects on the images. Recent state-of-the-art performance of Transformer for multi-task are achieved by Swin Transformer [14]. The Swin Transformer uses a similar hierarchical structure for building the different attention perception field size on the target images.

In this paper, we propose a novel unsupervised **attention mechanism** based image registration (AiR) framework. This paper serves as an attempt to introduce the Transformer model into the field of image registration. Unlike the previous CNN based method, our method framework using a Transformer for learning the deformation information between fix and moving images without convolution operation. In addition, unlike many prior works of using a Transformer in the computer vision community, the proposed framework does not rely on any neural network backbone as a prior feature extractor. To the best of our knowl-

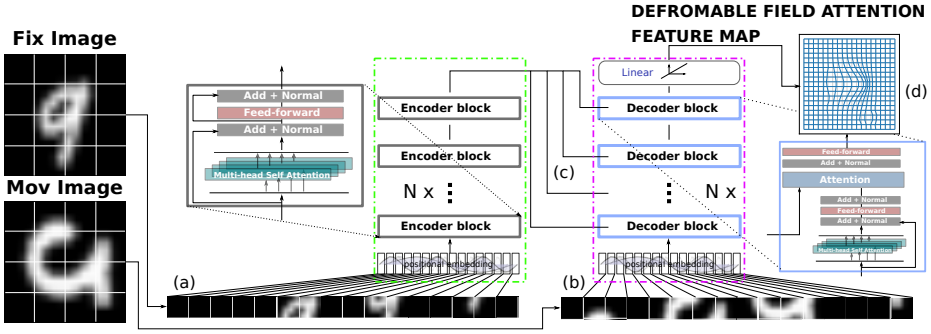

Fig. 1. The transformer framework used for deformation attention feature map prediction. The proposed transformer consists of encoder and decoder modules. The encoder module (shown in the green dotted line box) takes the fixed images patches as input and learns the representation of the memory attention features with self-attention mechanism. The decoder module (shown in the purple dotted line box) takes the attention features of the fixed image from the encoder (memory) and the self-attentions features of the moving image as input for predicting the deformable features that can transform the moving image into a fixed image.

edge, the proposed framework is the first Transformer based image registration method. The proposed Transformer based image registration framework is an unsupervised registration framework that the training deformation fields are not necessary. In addition, we proposed a multi-scale attention parallel Transformer framework for better solving incompetent cross-level feature representation of conventional Transformer.

2 Method

2.1 Attention for Image Registration (AiR)

The proposed attention-based image registration framework is based on the Transformer model [15]. The transformer model can be treated as an dimensional isotropic projection:

$$T_\theta : T(x) \rightarrow z; x, z \in \mathbb{R}^{n \times d} \tag{1}$$

where T is the projection which parameterized by a set of parameters $\theta : \{W_q, W_k, W_h, W_c, W_r\}$. The computational pipeline start from the input tensor x to output tensor z is given through:

$$\begin{aligned} Query = W_q \cdot x; Key = W_k \cdot x; Value = W_v \cdot x; \\ \alpha = softmax(< Query, Key > /\sqrt{k}); \\ z' = \sum_{h=1}^{H} W_c \sum_{j=1}^{J} \alpha \cdot Value; \\ z = NormRelu(z') \end{aligned} \tag{2}$$

To feed the images to Transformer the images need to be processed as sequence data-points [12]. To this end, the fixed I_f and moving I_m images pairs are firstly divided into $i \cdot i$ (see in Fig .1 step (a) where i = 4) patches and then embedded by linear projection with added position embedding to get the position information.

The fixed image patches are fed to transformer encoder network which encoding the attention features of the fixed images. The transformer encoder network contains N recursive blocks as shown in the green box of Fig. 1. Simultaneously, the moving images patches are fed together with the output attention features to the transformer decoder network (step (c) in Fig. 1). The output of the Transformer decoder network is processed with a linear projection layer to generate the deformation attention feature maps.

Similar to the DLIR [7], the generated deformation features is then processed by a spatial Transformer (STN) layer [16] for generating the displacement field to sample the moving images. As shown in Fig. 2, the output of the STN layer is then used for sampling the moving image to generate the warped image. In the end, the backpropagation is achieved by optimizing the similarity metric that measures the similarity between the deformed images and the fixed images.

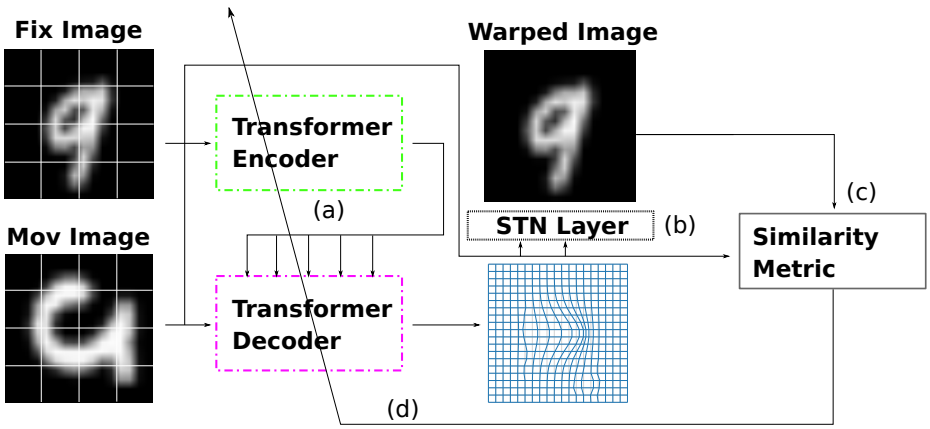

Fig. 2. The Transformer based image registration framework. In step (a): The input images are converted to deformation feature maps by the Transformer. In step (b): Those feature maps are converted to deformation field for a sampler to warp the moving image. Steps (c) -(d): The warped moving image is quantified by the similarity metric for computing the gradients flow for back-propagation.

3 Multi-scale Parallel Attention Transformer

CNN based method can use feature layer hierarchical superposition and different size of convolutional kernels for features expression in different levels. Yet, as the

transformer is not a convolutional based method, the ability of cross feature level feature extraction is relative weaker. This problem can be observed also in our experimental section for a single level Transformer (patch size equal to 2) based AiR framework.

To get the multi-level feature learning for Transformer, we propose a Multi-scale Attention Parallel Transformer (MAPT) that can learn the features from different perception scales. The MAPT consists N Transformers (N decoders and N encoders). For each transformer, they adapt different size of patches as inputs and generates N different attention feature maps F_N. The N feature maps are then sampled into a uniform size for adding together with normalized weighting ratio to get the final deformable feature map F.

4 Experiments and Evaluation

4.1 Dataset and Experiment Details

We evaluate the AiR framework on the MNIST dataset. During the training period, the randomly picked pairs are used to feed the AiR transformer for gradients decent. We use 20% of data for testing the performance of the AiR Transformer and 80% of images for training. The Transformer in our experiment contains 1 block ($N = 1$) for both the Transformer encoder and decoder networks. The dropout ratio of the encoder and decoder layers are set as 0.5. The attention layers of the Transformer consist of 4 heads for both the multi-head self-attention and attention layers. The projection dimension was set as 16 for both the encoder and decoder networks. We train the framework with an Adam optimization with learning rate: $lr = 0.5e-3$. The experiment was run on a Dell Precision 7520 Laptop with an Intel© Core™ i7-7820HQ CPU @ 2.90GHz × 4 and NVIDIA Quadro M2200 Mobile GPU. The program was implemented with Pytorch framework.

4.2 Results and Evaluation

Fig. 4 shows the registration result of the proposed AiR. We can see from the generated results that the proposed framework learns the mapping of the deformation for warping the moving image to the form of fixed images. However, limited by the hardware resources (4 GB GPU memory), the experiment results are constrained by the number of the decoder and encoder blocks [17].

We can see from the Fig. 4 that although the overall shape of the sampled images is similar to the fixed images, there are any defects and noises in geometric details for the warped images. This defect may be related to the quality of the generated deformable fields. Many recent works have pointed out that the transformer is more inferior to CNN in terms of pixel-level details processing ability [17] [18]. This problem is also observed in our experiment section.

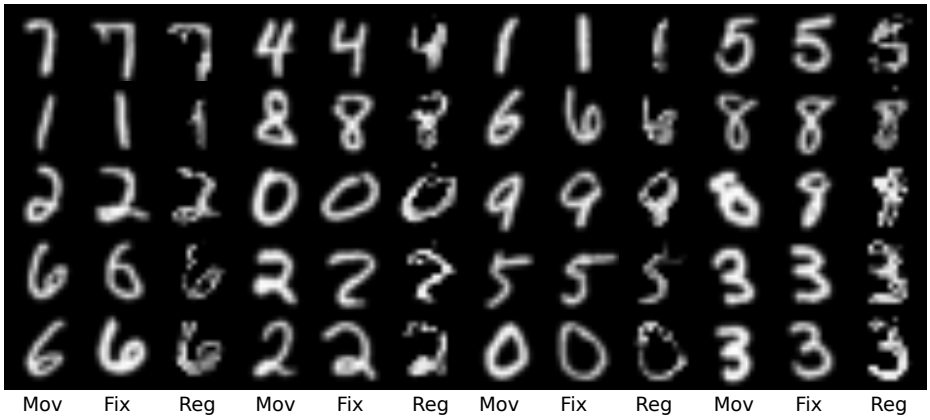

Fig. 3. Registration results of the proposed method of single level Transformer.

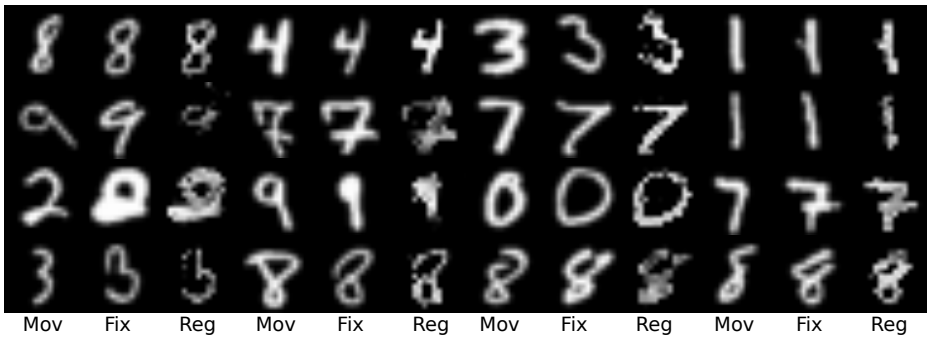

Fig. 4. Registration results of the proposed method of single level Transformer.

5 Conclusion

In this work, we presented an Attention mechanism-based Image Registration framework: AiR. The framework was built based on the Transformer model which is different from current commonly used CNN based approaches. We have shown in the experiment section that the proposed framework can achieve the goal of unsupervised deformable images registration. The results show that the Transformer model can achieve plausible performance for image registration. However, as the Transformer is a young research topic in computer vision, there are still many limitations for wide application compared to current CNN-based methods. A typical problem is the current Transformer is GPU-bound which leads to the prohibition for conventional computing systems. There is still a lot of work to be done to deal with the computational complexity for the current Transformer framework.

Bibliography

[1] Armand Zampieri, Guillaume Charpiat, Nicolas Girard, and Yuliya Tarabalka. Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In *European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018.

[2] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable Medical Image Registration: A Survey. Research Report RR-7919, INRIA, September 2012.

[3] I. Yanovsky, B. Holt, and F. Ayoub. Deriving velocity fields of submesoscale eddies using multi-sensor imagery. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 1921–1924, 2020.

[4] Francisco P.M. Oliveira and João Manuel R.S. Tavares. Medical image registration: a review. *Computer Methods in Biomechanics and Biomedical Engineering*, 17(2):73–93, 2014. PMID: 22435355.

[5] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Transactions on Biomedical Engineering*, 63(7):1505–1516, 2016.

[6] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pages 300–308, Cham, 2017. Springer International Publishing.

[7] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, 2019.

[8] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE Transactions on Medical Imaging*, 38(9):2165–2176, 2019.

[9] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21665–21674. Curran Associates, Inc., 2020.

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[11] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021.
[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.
[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.
[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
[17] Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive Transformer. In *ICLR 2020 - Eighth International Conference on Learning Representations*, pages 1–14, Addis Ababa, Ethiopia, April 2020.
[18] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer, 2021.