IMAGE REGISTRATION USING THE MNIST DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, a model is presented for deformable image registration using an approach based on the VoxelMorph algorithm. This model is trained on the MNIST dataset as a lightweight demonstration of the convolutional approach to learnable image registration. The model was successful at rough reconstruction of target images from the fixed images, but showed difficulty reproducing discontinuous image features. Mixed success was achieved at generalizing the results beyond the scope of the original dataset. x

1 INTRODUCTION

Image registration is an important step in many medical imaging applications. Deformable image registration, where the shape of the image is allowed to be warped to achieve congruence between the initial and target images, is particularly useful in applications such as neuroimaging and organ segmentation, where images may represent cross-sections of volumes that shift in a 3-dimensional space Sheikhjafari et al. (2022). By mapping a deformation field from one image to another, annotated boundaries in one image can be mapped to their corresponding boundaries in different layers of a scanned volume, allowing for important clinical and research processes such as the 3-dimensional mapping of different organ systems, or the aggregation of multi-modal data into one aligned space.

Previous deterministic algorithms for deformable image registration have been slow and computationally intensive, limiting their practical clinical application Balakrishnan et al. (2019). By training a convolutional neural network (CNN) to approximate the deformation field between pairs of images, this field can be generated significantly faster by the VoxelMorph algorithm than older methods. Efficiency improvements not only provide for more convenient application for the user, but also allow models to be applied practically on much larger datasets, and in use cases where long runtimes present a barrier to deployment Manolis Kellis & Dalca (2021).

In this paper, this approach is demonstrated using the MNIST dataset. This dataset, which is made up of grayscale images of handwritten integers from 0 to 9, is both freely available and relatively lightweight, making it ideal for training on consumer-grade hardware.

2 METHODS

2.1 DATASET

The MNIST dataset is a freely-available set of 70,000 single-channel, 28 x 28 grayscale images of handwritten integers from 0 to 9, with 60,000 original test images and 10,000 test images. The original training split was further divided into 48,000 training and 12,000 validation images. For model training, the training and validation data were subset to 6,000 and 1,000 images respectively of only the number 7. A batch size of 179 was used in the training dataset, where each image in each batch was randomly matched to another image in the training data. Self-matching was allowed in order to allow the model to learn null mappings.

2.2 NETWORK DESIGN

Given a fixed image F and a target image T in $\mathbb{R}^2$ from the MNIST dataset, we attempt to find a continuous vector field ϕ describing the deformation from F to T. The goal is for this vector field to maintain continuity in the deformation such that, when applied to medical imaging, anatomical structures in the fixed image are preserved in the reconstructed target image. Because this model is trained using the MNIST dataset, interpretability of the mapping is not a focus of this project.

To learn a function $f(F, \phi) \approx T$, we use a CNN approach based on the VoxelMorph algorithm. Three convolutional layers are used to encode the image information and reduce the complexity of the represented dataset using max pooling. Three convolutional layers are then used to decode the encoded information via upsampling and convert it into a two-layer field representing the x- and y-axis deformation for each pixel in F. All layers were batch-normalized with a 20% dropout rate, and the hidden layers used a leaky ReLU activation function over 256 filters. The output layer used a tanh activation function to normalize the final field estimate.

2.3 Model Training

The loss of the final output was calculated using a two-part function:

$$Loss = \sum (f(F, \phi) – T)^2 + \lambda_{smooth} ||\nabla\phi||^2]$$

The first loss term represents the mean-squared error associated with the accuracy of the reconstruction of the final image. The second term is used to reward smoothness in the deformation field, to ensure continuity, where $\nabla\phi$ is approximated using finite differences. The two terms are weighted according a tuning parameter λ_{smooth}.

Parameters were learned using an Adam optimizer with L2 penalty λ = 0.1. A grid search was performed to identify the optimal learning rate and λ_{smooth}. For hyperparameter tuning, all models were run for 40 epochs. Final model training was run for a minimum of 5 epochs, stopping either after 40 epochs or when the validation error for one epoch exceeded that of the previous.

Models were trained using a subset of the data that only included the number 7. The test loss was then generated for two scenarios, one for mapping instances of the number 7 to the number 7, and one for mapping the number 4 to the number 4 in order to determine how generalizable the model is beyond its initial training set.

3 Results

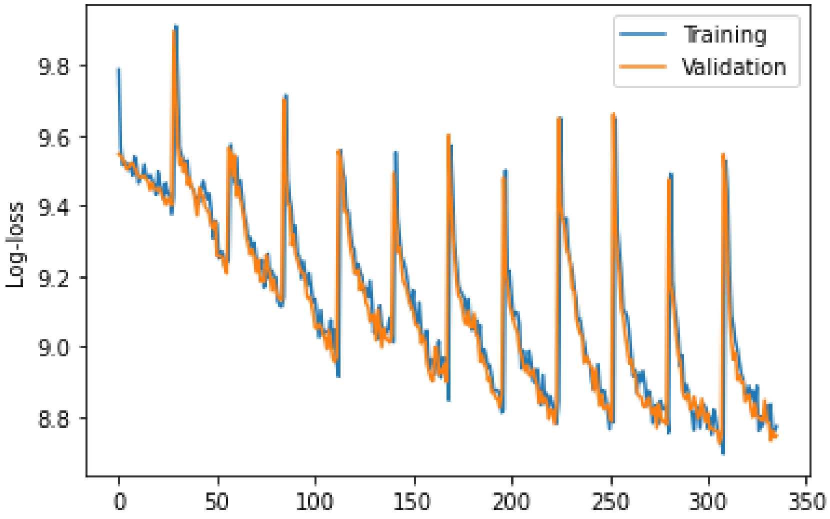

Figure 1: Training and validation loss, by batch.

3.1 Model Training

Hyperparameter tuning identified an optimal $\lambda_{smooth} = 0.1$ as its tuning parameter, and an optimal learning rate of $\eta = 0.1$. On final training, the model was run for 12 epochs before the validation converged, and the training was stopped. The training and validation loss are presented in Figure 1.

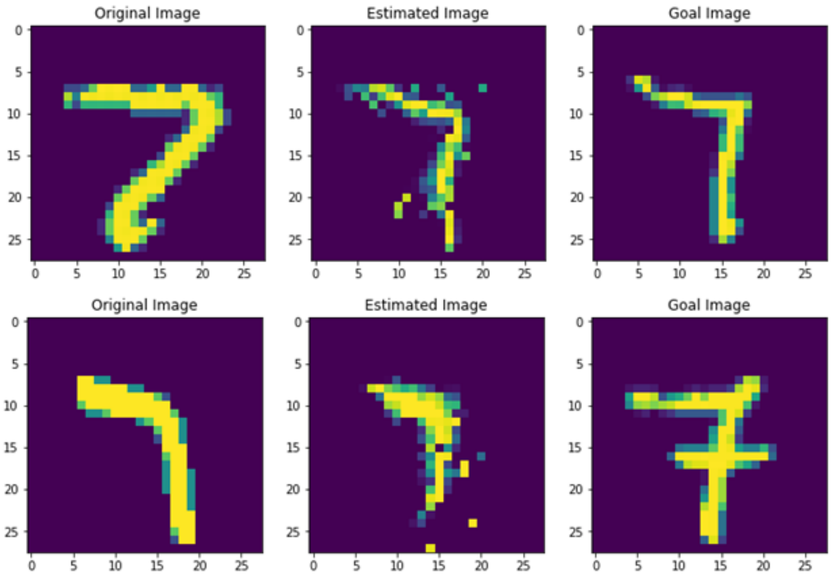

Figure 2: Example reconstructions of goal (target) image from original (fixed) image.

3.2 Generalization

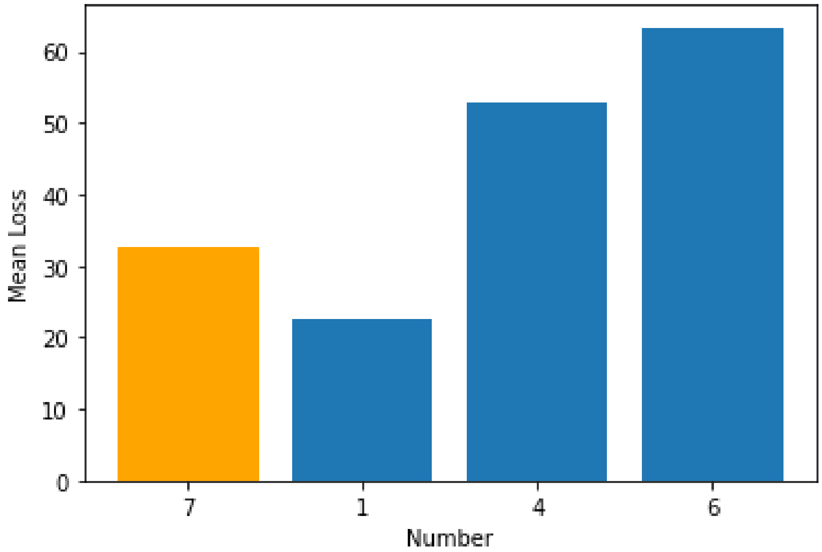

Figure 3: Mean test loss by number.

Generalization was tested by applying the model to images in the MNIST dataset outside the original scope. 1,028 images were used to determine the out-of-sample loss among images of the number 7 in the test dataset. In addition, the mean loss was calculated 1,135 images of the number 1, 982 images of the number 4, and 958 images of the number 6. Only images within each number-specific dataset were matched together. The mean loss for each category is shown in Figure 3.

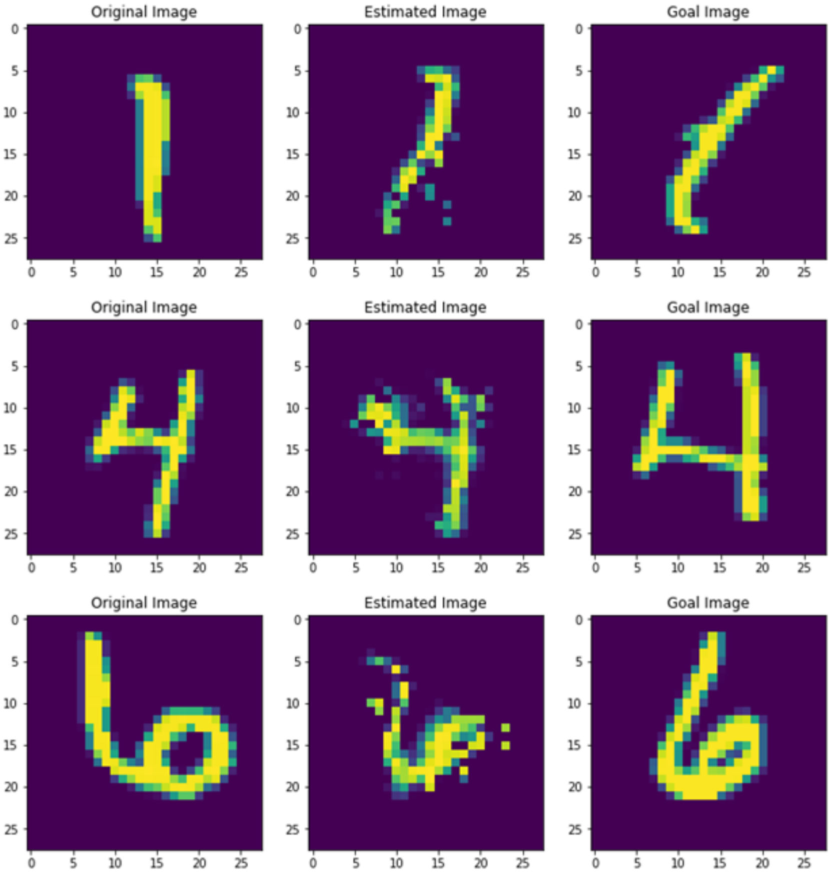

Figure 4: Example reconstructions of numbers 1, 4, and 6.

4 Discussion

The model trained here has demonstrated some success in the image registration task, although not to the extent that the generated and target images are indistinguishable. The close tracking of the validation loss with the training loss suggests that future iterations of the model may benefit from a deeper architecture, larger network size, or longer training period.

For the training case, one obstacle encountered was that there are two common variants of the number 7 – one with a crossbar, and one without. In all iterations of the training, the model struggled to generate a crossbar that was present in the target image, but not in the fixed image. This is to be expected by the nature of the problem setup – because the deformation process is continuous, it is difficult to invent new features in the final product that are not present in the original image.

Based on the average test loss, the model's ability to generalize to numbers other than 7 was also dependent on the similarity between the two numbers. For example, the model struggled the most to register the number 6, likely because it is significantly less linear than 7, 1, or 4. Interestingly, on inspection, certain features were successfully deformed in the 6-dataset that would not have been present in the training data, such as the loop on the right-hand side. This suggests that the model's ability to achieve a deformation was not limited to simple linear structures, and that the lack of generalizability has a more complex source.

A final limitation for this project was that the hardware used was not sufficient to train moderately-sized models in a practical way, limiting the scope of the hyperparameter tuning. The original

VoxelMorph model used a 12-layer CNN to estimate f, and was reported to train in “a few minutes”. For comparison, this model used a 6-layer network and was only able to train successfully over the course of several hours.

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, August 2019. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2019.2897538. URL `http://arxiv.org/abs/1809.05231`. arXiv:1809.05231 [cs].

Manolis Kellis and Adrian V. Dalca. Deep Learning Image Registration and Analysis - Lecture 21 - MIT ML in Life Sciences (Spring 2021), May 2021. URL `https://www.youtube.com/watch?v=c4dvyTBvysQ`.

Ameneh Sheikhjafari, Michelle Noga, Kumaradevan Punithakumar, and Nilanjan Ray. Unsupervised deformable image registration with fully connected generative neural network. July 2022. URL `https://openreview.net/forum?id=HkmkmW2jM`.