How do data augmentation and autoencoder regularisation interact on 3D subcortical brain segmentation tasks?

Deep Learning in Healthcare mini-project − Candidate #1044704

FHS Part C Mathematics and Computer Science, University of Oxford

Abstract. Deep learning for medical image segmentation typically requires large quantities of training data to avoid over-fitting. However, labelled segmentation data is scarce and so data availability often constrains out-of-sample performance. One approach to mitigate this is expanding a dataset via data augmentation; another is to explicitly regularise a model's learned representations via the auxiliary task of image reconstruction, a method termed autoencoder regularisation. In this work, I perform ablation studies on Segresnet-VAE, a sophisticated deep learning segmentation model, to investigate the relationship between data augmentation and autoencoder regularisation. I find both techniques lead to improvements in generalisation error, with data augmentation affecting accuracy and training behaviour more significantly. Across the board, the combination of both techniques yields the strongest performance. Some results were obtained which support the conclusion that data augmentation is a requirement for autoencoder regularisation to have a positive effect on generalisation error.

1 Introduction

The automated segmentation of brain structures from MRI data can speed up disease diagnosis as well as reducing costs and improving accuracy. Medical image datasets are typically characterised by large individual samples and few samples per dataset. This holds particularly true for segmentation tasks, because labelling data is a time-consuming task requiring expert clinicians. As a result, it is difficult to construct deep learning models that generalise effectively beyond their (limited) training data. Investigating strong regularisation techniques is an essential step in solving this problem and overcoming the curse of dimensionality.

Regularisation can be derived from model architecture. Modern medical segmentation models use convolutions and down-sampling to encode strong geometric priors such as translational and scale invariance that constrain the model class. Deep models allow hierarchical feature representation and the distillation of semantic information. Furthermore, practitioners have a large regularisation bag-of-tricks including weight decay, normalisation layers and dropout.

A relatively novel technique is autoencoder regularisation [1], which can be used in conjunction with U-Net-style architectures to explicitly regularise

a model's learned semantic information. A model is trained to reconstruct input images concurrently with the primary segmentation task using an shared encoder. This auxiliary task implicitly exposes the encoder to more patterns and forces it to learn more representative features [2].

A ubiquitous, implicit form of regularisation is data augmentation, where new data is generated by applying transformations to existing data. An informed choice of transformations can train a model to become invariant to common imaging artefacts, variations in imaging protocol and even subject phenotype.

Data augmentation shares a number of similarities with autoencoder regularisation, including adding prior domain knowledge to a model and increasing the complexity of the learning task. This discussion motivates the following interesting and as-yet-unexplored research question: **how do data augmentation and autoencoder regularisation interact in a brain segmentation context?**

2 Related work

The first use of autoencoder regularisation in the medical imaging domain was by Myronenko in 2018 [1], in a model termed Segresnet-VAE (by [3]) or VAE-U-Net (by [4]) in the literature (Segresnet-VAE is used in this paper). Segresnet-VAE is a variant of U-Net [5] using ResNet [6] blocks for the encoder and decoder paths. Furthermore, an additional "VAE" path from the encoder (without skip-connections) is added to perform autoencoder regularisation via a Variational Autoencoder (VAE) [7]. Figure 1 shows a summary of the architecture. Segresnet-VAE proved highly effective, winning the brain tumour segmentation competition BraTS 2018 [8].

Following this success, Frey et al. explored a similar architecture with an emphasis on memory efficiency, preprocessing data to small patches for compatibility with GPU constraints [9]. Li et al. subsequently extended the original work by adding more sophisticated skip-connections between the VAE path and segmentation path [4]. These connections were designed to fuse more spatial information into the predicted segmentation mask.

Myronenko, Frey and Li all applied autoencoder regularisation to brain tumour segmentation tasks. Jin et al. instead tackled the task of retinal vessel segmentation, applying a Vector Quantized VAE (VQ-VAE) [10] in place of a VAE for autoencoder regularisation [2]. This modification to the autoencoder framework enabled higher coherence and higher quality reconstruction images, further improving generalisation performance.

In 2022, Pham et al. proposed a transformer extension of Segresnet-VAE called SegTransVAE [3], and validated its effectiveness on brain tumour MRI and kidney CT segmentation datasets. Inspired by the success of the transformer-based model UNETR [11], SegTransVAE extended Segresnet-VAE by integrating a transformer block into the encoder path. Exploiting the global receptive field of transformers yielded performance improvements over the Segresnet-VAE model, at the cost of a large increase in model parameters (from 7.5M to 45M).

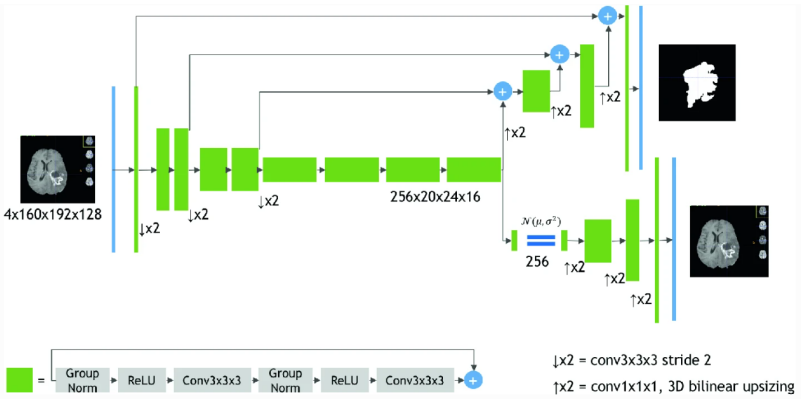


Fig. 1: A summary of Segresnet-VAE's architecture (reproduced from [1]). The input image is first downsampled to a shared encoder representation. The segmentation (upper) and "VAE"" (lower) paths then diverge to perform their respective tasks.

Data augmentation is a frequently-employed technique in the literature. For a comprehensive recent overview see Garcea et al. [12] or Chlap et al. [13]. Common operations include affine and pixel-level transforms to increase robustness to different imaging protocols. More sophisticated approaches go beyond transforming original data to generate artificial data via generative deep learning approaches. Garcea et al. report a median increase in performance when using data augmentation of around 10 %.

This study focuses on the original Segresnet-VAE model proposed by Myronenko [1]. I limited the data augmentations used to simple spatial and intensity transforms. While several papers in this section utilise data augmentation and autoencoder regularisation together, none describe the relative effect of these model components; this study aims to fill this gap in the literature.

3 Proposed approach

3.1 Overview

To investigate the research question, I chose to perform an ablation study using Myronenko's original Segresnet-VAE model [1]. The two components under investigation were the model's VAE path for autoencoder regularisation and the use of on-the-fly data augmentation during training.

Goals The goals of the study were to investigate how the model's convergence behaviour and segmentation performance (including on out-of-sample data) are affected by each component. It was hypothesised that the combination of components would have the slowest convergence but highest accuracy, with each component individually contributing to this effect.

Task I chose the challenging segmentation task of multi-label subcortical segmentation to best evaluate the model's performance. I decided to process

data as 3D voxels; this has the benefit of maximising contextual information when segmenting, but requires more memory than processing 2D or 2.5D slices. Data from Guy's Hospital was used as training, validation and test data, with the Institute of Psychiatry (IOP) and Hammersmith Hospital (HH) datasets used exclusively for testing. This train/val/test split was selected to mimic the realistic healthcare scenario of training on a single data acquisition method, and to investigate model performance on images taken with different MRI scanners.

Summary I implemented and trained the Segresnet-VAE with and without the VAE path and/or data augmentation (often abbreviated to DA), and evaluated its 3D subcortical segmentation accuracy.

3.2 Architecture

The Segresnet-VAE architecture used by Myronenko [1] is summarised in Figure 1. Key components include:

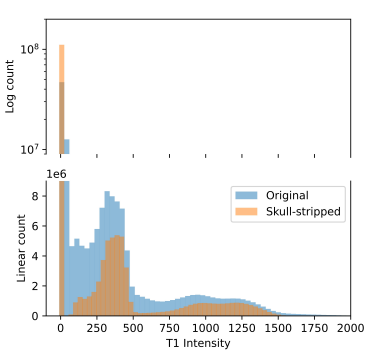
- GroupNorm: Normalisation layers were used to mitigate the issue of covariate shift. GroupNorm was chosen as it performs well with small batch size.
- ResNet blocks: Contextual information was aggregated by ResNet blocks (shown in green in Figure 1) to extract relevant features.
- Downsampling: Strided convolutions reduce the spatial dimensions of representations while increasing the number of features.
- Upsampling: Trilinear upsizing and 1x1x1 convolutions increase the spatial dimensions of representations while decreasing the number of features.

Differences to the original model The model I implemented for this study differs from Segresnet-VAE in a few minor ways. Firstly, fewer parameters (8 vs 32) were used as the initial convolution kernel feature size, to satisfy computational constraints. Secondly, the different input dataset necessitated a different number of input (1 vs 4) and output channels (4 vs 3).

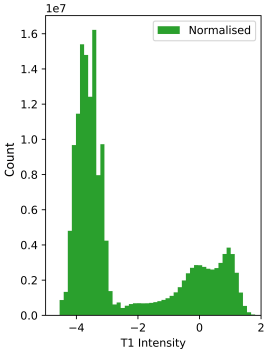
To deal with class imbalance between non-subcortical regions and the subcortical regions, I diverged from Myronenko's methodology for assigning labels given prediction values. One output channel was used for each of the four subcortical regions, with no channel for "non-subcortical region". Predicted class labels were then determined voxel-wise as the class with the largest prediction; if no class had a prediction value of ≥ 0.5 the voxel was left unlabelled.

VAE path Feature representations computed by the encoder were downsampled before being flattened and passed through a dense layer. Two further dense layers computed a latent mean and standard deviation. Next, a vector was sampled from a Gaussian with the computed parameters. This sampled vector was then progressively upsampled to generate a reconstruction image.

Loss Soft Dice loss was chosen as the segmentation loss [14]. To counteract class imbalance, this was calculated per output channel and summed. Loss for the VAE path was the sum of a KL-divergence term regularising the latent space and standard L2 reconstruction loss on the output. The VAE loss was weighted by $\times 0.1$ before summation with the segmentation loss to calculate total loss.



(a) Intensity distributions pre- and post-skull-stripping



(b) Post-normalisation intensity distribution

Region	%
Thalamus	0.65
Caudate	0.30
Putamen	0.29
Amygdala	0.09
Other	98.67

(c) Average volume per subcortical region

Fig. 2: Data pre-processing summary statistics for training data

4 Data

Pre-processing Labels were first converted to a one-hotted representation for each of the four subcortical regions. Prior to data augmentation, all images were skull-stripped by applying the provided brain extraction masks. After data augmentation, individual input images were normalised to have zero mean and unit standard deviation (based on non-zero voxels only [1]). Data from Guy's was subject to a train/val/test split of 80/10/10; data from the IOP and HH were used exclusively for testing [1]. Figures 2a and 2b show the effect of data normalisation and Figure 2c details the severe class imbalance in the dataset.

Data augmentation The Torchio package [12] was used to generate on-the-fly data augmentations during training. Firstly, equal probability random axis flip on all 3 axes was performed. The following transforms were then applied with independent probability 0.25 (with default parameters except where specified): affine transformation (max. 5 deg./5 voxels); random bias field; random MRI motion artefact (max. 5 deg./5 voxels); random gamma contrast shift.

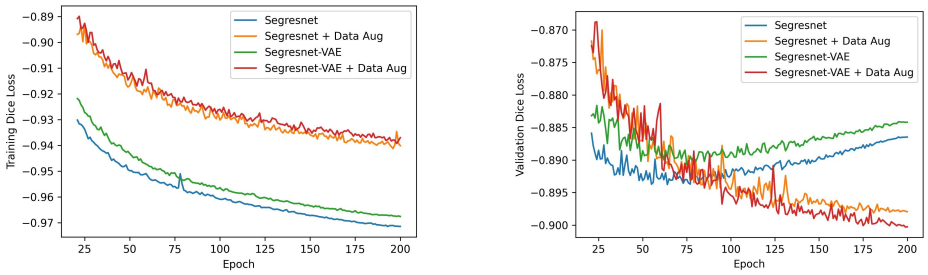
5 Experiments

5.1 Implementation

Models were trained for 200 epochs using the Adam optimiser, with a weight decay of $1e-5$ and initial learning rate of $\alpha_0 = 1e-4$. Learning rate annealing was employed: at epoch n learning rate was $\alpha_n = \alpha_0(1-\frac{n}{200})^{0.9}$ (following [1]).

The implementation was written in PyTorch 1.9.0. Segmentation metrics were calculated with the seg-metrics package [15]. Experiments were run using

[1] Two corrupted label maps in the IOP test dataset were not used.



(a) Training Dice loss (epoch ≥ 20)

(b) Validation Dice loss (epoch ≥ 20)

Fig. 3: Training and validation Dice loss curves for the four models trained.

default GPUs (with 16GB RAM) in Google Colab and Paperspace. Practicals 2 and 3 from the Deep Learning in Healthcare course were used as an implementation base [1]. Where technical details were missing from Myronenko's paper I consulted an open-source TensorFlow implementation of Segresnet-VAE [2],

5.2 Convergence behaviour

Figure 3 shows learning curves for the trained models, which support the initial hypothesis. It can be observed that without data augmentation the models overfit more severely; the effect of autoencoder regularisation is much less pronounced. Data augmentation clearly also decreases the smoothness of convergence.

With/out the VAE path, the model had 2.2M/1.1M parameters. In Colab, the base model took 53s per training epoch (253 images). On-the-fly data augmentation added 6s/epoch, while the VAE path added 18s/epoch.

5.3 Quantitative measures of segmentation

Mean surface distance (MSD) and Dice score were chosen as complementary metrics to evaluate both distance and overlap between segmentation predictions. Table 1 demonstrates the effectiveness of regularisation, particularly for the out-of-sample IOP and HH datasets. Data augmentation can be observed to have the greatest contribution, although autoencoder regularisation also improves segmentation accuracy. As expected, both together yields the strongest accuracy. Table 2 shows the per segmentation class results, which validate this finding; this table also shows the effect of class imbalance (cf. Figure 2c).

5.4 Qualitative comparison

Figure 4 shows predicted segmentations for the image in the test sets with greatest accuracy disparity between methods; this image has been poorly skull-

[1] https://courses.cs.ox.ac.uk/mod/folder/view.php?id=624 ; by Dr Nicola Dinsdale

[2] https://github.com/IAmSuyogJadhav/3d-mri-brain-tumor-segmentation-using-autoencoder-regularization

Table 1: Mean Dice score (%) and Mean Surface Distance (MSD) (in unit voxels) averaged over all four segmentation classes for three datasets. The best values per dataset are in bold. (DA = Data Augmentation)

Test dataset	Guys (n=33)		IOP (n=69)		HH (n=176)	
	Dice	MSD	Dice	MSD	Dice	MSD
Segresnet	87.3	0.414	81.6	0.469	86.1	0.275
Segresnet-VAE	87.3	0.421	82.7	0.433	86.1	0.284
Segresnet + DA	87.9	0.402	**84.2**	0.446	87.1	0.262
Segresnet-VAE + DA	**88.1**	**0.396**	84.1	**0.393**	**87.8**	**0.242**

Table 2: Mean Dice score (%) and Mean Surface Distance (MSD) (in unit voxels) per segmentation class averaged over all test images. The best values per class are in bold.

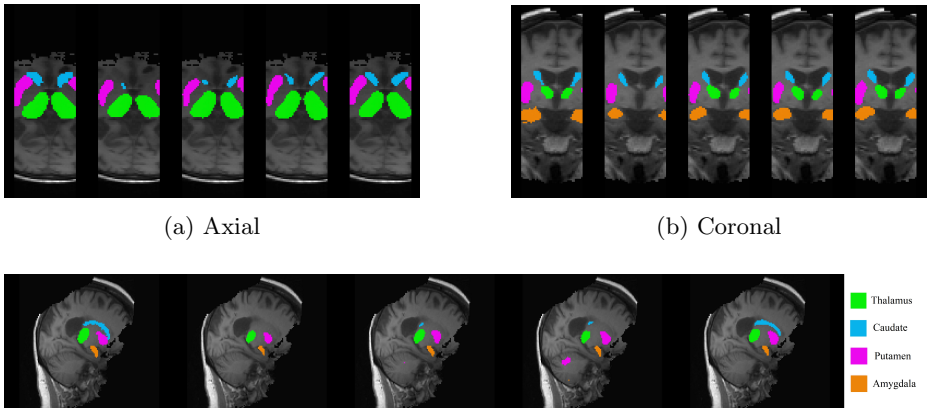
Region	Thalamus		Caudate		Putamen		Amygdala	
	Dice	MSD	Dice	MSD	Dice	MSD	Dice	MSD
Segresnet	92.3	0.276	84.6	0.295	87.5	0.311	75.9	0.476
Segresnet-VAE	92.6	0.283	84.9	0.284	87.3	0.321	76.7	0.460
Segresnet + DA	93.0	0.259	85.9	0.272	88.6	0.316	78.4	0.449
Segresnet-VAE + DA	**93.2**	**0.247**	**86.2**	**0.263**	**89.1**	**0.266**	**79.1**	**0.415**

stripped, which may explain this disparity. It may be visually confirmed that no regularisation performs worst, and the full model performs best. Figure 4c is representative of how Segresnet + Data Augmentation (DA) can introduce artefacts to segmentation maps, and how autoencoder regularisation eliminates this phenomenon. Further visualisations are available in the Supplementary Material.

Figure 5 shows the output reconstruction from the autoencoder-regularised models for the same image. Interestingly, Segresnet-VAE predicts a visually more realistic brain than Segresnet-VAE + DA, but its reconstruction is less related to the input image. This may show how data augmentation has allowed the VAE to better capture the variation between brains.

6 Discussion and Conclusion

Section 5 shows a comprehensive validation of the hypothesis that the full model has the strongest performance in terms of segmentation accuracy as well as slowest convergence. Data augmentation is found to play a more significant role than autoencoder regularisation. The results on the Hammersmith Hospital dataset (a dataset collected using a more powerful scanner than that used for the training data) suggest that data augmentation can be a necessary step for autoencoder regularisation to improve out-of-sample generalisation. Figure 5 suggests a possible explanation: without sufficiently varied training data the autoencoder cannot adequately capture the sophisticated variations between images.



(a) Axial (b) Coronal

(c) Sagittal

Fig. 4: Predicted subcortical brain region segmentations shown as 2D MRI slices. From left: (1). Ground truth (2). Segresnet (3). Segresnet-VAE (4). Segresnet + DA (5). Segresnet-VAE + DA. The image in the test sets with the greatest disparity between Dice scores for different methods was chosen for this visualisation.

Fig. 5: VAE reconstructed image outputs using the same image as Figure 4. From left: (1). Original (2). Reconstruction without DA (3). Reconstruction with DA

The study's use of a complex and deep model with realistic training conditions gives it relevance to the state-of-the-art literature. Quantitative and qualitative agreement on three separate datasets and four separate segmentation classes increases the confidence in the results on out-of-sample performance.

However, the implementation lacks width, with 4× fewer parameters used relative to the literature. Furthermore, the study only evaluated a single choice of hyperparameters and data augmentation transformations (due to time constraints). For example, Figure 3a suggests that longer training could improve the models that used data augmentation, potentially altering accuracy results.

In addition, the provided computer-generated ground truth labels are fallible [16]; training with expert labels could pair best with autoencoder regularisation as more representative learned features work best without labelling errors.

These limitations naturally suggest replicating the study with expert labels, a wider network and detailed hyperparameter sweep. Furthermore, a more extensive ablation study could reveal the most significant data augmentations, and whether more extreme augmentations help or hinder the autoencoder regularisation. A further study could evaluate whether these conclusions remain consistent across the models discussed in Section 2.

References

1. A. Myronenko, "3D MRI Brain Tumor Segmentation Using Autoencoder Regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, (Cham), pp. 311–320, Springer, 2019.
2. G. Jin, X. Chen, and L. Ying, "Deep Multi-Task Learning for an Autoencoder-Regularized Semantic Segmentation of Fundus Retina Images," *Mathematics*, vol. 10, no. 24, p. 4798, 2022.
3. Q.-D. Pham, H. Nguyen-Truong, N. N. Phuong, and K. N. A. Nguyen, "SegTransVAE: Hybrid CNN – Transformer with Regularization for medical image segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2022.
4. K. Li, L. Kong, and Y. Zhang, "3D U-Net Brain Tumor Segmentation Using VAE Skip Connection," in *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*, pp. 97–101, IEEE, 2020.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *LNCS*, vol. 9351, pp. 234–241, 2015.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, 2016.
7. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations (ICLR)*, 2014.
8. B. H. Menze, A. Jakab, S. Bauer, *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, Oct. 2015.
9. M. Frey and M. Nau, "Memory Efficient Brain Tumor Segmentation Using an Autoencoder-Regularized U-Net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, (Cham), pp. 388–396, Springer International Publishing, 2020.
10. A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
11. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (Waikoloa, HI, USA), pp. 1748–1758, IEEE, 2022.
12. F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, 2023.
13. P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545–563, 2021.
14. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.
15. J. Jia, "A package to compute segmentation metrics: Seg-metrics," 2020.
16. N. Furuhashi, S. Okuhata, and T. Kobayashi, "A Robust and Accurate Deep-learning-based Method for the Segmentation of Subcortical Brain: Cross-dataset Evaluation of Generalization Performance," *Magnetic Resonance in Medical Sciences*, vol. 20, no. 2, pp. 166–174, 2021.