Suggestive Annotation of Brain MR Images with Gradient-guided Sampling

Chengliang Dai[a,*], Shuo Wang[a], Yuanhan Mo[a], Elsa Angelini[b], Yike Guo[a], Wenjia Bai[a,c]

[a]*Data Science Institute, Imperial College London, United Kingdom*
[b]*NIHR Imperial Biomedical Research Centre, Imperial College London, United Kingdom*
[c]*Department of Brain Sciences, Imperial College London, United Kingdom*

Abstract

Machine learning has been widely adopted for medical image analysis in recent years given its promising performance in image segmentation and classification tasks. The success of machine learning, in particular supervised learning, depends on the availability of manually annotated datasets. For medical imaging applications, such annotated datasets are not easy to acquire, it takes a substantial amount of time and resource to curate an annotated medical image set. In this paper, we propose an efficient annotation framework for brain MR images that can suggest informative sample images for human experts to annotate. We evaluate the framework on two different brain image analysis tasks, namely brain tumour segmentation and whole brain segmentation. Experiments show that for brain tumour segmentation task on the BraTS 2019 dataset, training a segmentation model with only 7% suggestively annotated image samples can achieve a performance comparable to that of training on the full dataset. For whole brain segmentation on the MALC dataset, training with 42% suggestively annotated image samples can achieve a comparable performance to training on the full dataset. The proposed framework demonstrates a promising way to save manual annotation cost and improve data efficiency in medical imaging applications.

Keywords: Brain MRI, Suggestive annotation, Image segmentation, Active learning

1. Introduction

Medical imaging plays an significant role in many clinical applications such as: disease diagnosis, clinical decision making, treatment planning and patient management. There are hundreds of millions of medical imaging scans conducted every year around the world creating a huge amount of available data. In England, there were 43.5 million imaging tests reported in the 12 months from March 2018 to March 2019 (NHS England, 2019). The increasing number of medical images inevitably makes the number of experienced radiologists to interpret these complex images fall short of demand. The recent development of deep learning techniques can help to alleviate the shortage of qualified radiologists. Extensive research has shown that deep learning techniques can achieve performance comparable to or even surpassing human-level performance in a variety of tasks (Liang et al., 2019), particularly in medical imaging applications such as segmentation (Liu et al., 2020b), registration (Haskins et al., 2020) and reconstruction (Knoll et al., 2020).

Despite the growing popularity of deep learning, there are several major challenges; one of them being the scarcity of data required for training deep learning models. It is well-known that the good performance of deep learning models often relies on the large amount of data used for training. The real challenge in medical imaging applications is not just the availability of imaging data itself, but also the scarcity of expert annotations for these images. Medical image annotation can be expensive in terms of both time and required expertise. For example,

*Corresponding author: Chengliang Dai e-mail:c.dai@imperial.ac.uk

annotating a 3D brain tumour image for a single patient may take several hours even for a skilled image analyst (Fiez et al., 2000).

Several methods have shown promising results in alleviating the data scarcity challenge in medical imaging by training deep learning models from limited annotations. For instance, unsupervised and semi-supervised learning methods aim to leverage information from abundant unannotated samples and to improve the accuracy and robustness of the model using these samples (Cheplygina et al., 2019). Generative adversarial network (GAN) based approaches (Zhao et al., 2019; Yi et al., 2019; Ravanbakhsh et al., 2020) are widely adopted for generating synthetic images and pseudo annotations; images that can be used for model training. The accuracy and robustness of the models are improved after training synthetic datasets compared to training on the original dataset alone (Zhao et al., 2019; Yi et al., 2019; Ravanbakhsh et al., 2020). Some methods such as dataset distillation (Wang et al., 2018) aim to reduce the size of training dataset by creating very few synthetic samples that carry as much information as the original large dataset. Moreover, transfer learning and domain adaptation methods have been proposed to learn features from one domain where a large amount of annotated data is available and then transfer these features to a new domain where annotated data is limited (Tajbakhsh et al., 2016; Kamnitsas et al., 2017).

Most methods mentioned previously are developed for harvesting as many features or samples as possible from existing datasets to enhance the model performance without acquiring more annotations. On the other hand, an active learning (AL) approach tackles the problem by suggesting an optimal subset of an unannotated dataset that is effective and beneficial for model training. A typical AL framework is demonstrated in Fig. 1.

In this paper, we propose a gradient-guided sampling mechanism that can be used directly by a general AL framework to utilize the gradient of the loss term from the deep learning model for suggesting informative samples for manual annotation. We explored two brain image segmentation tasks, brain tumour segmentation and whole brain segmentation. In both tasks, our method is capable of suggesting samples that are more helpful to training a segmentation model compared to state-of-the-art methods, which allow the model to produce comparable performance to a model trained on the full dataset.

1.1. Related work

Selecting informative data samples is the essence of any AL framework. There are many approaches used in the area of AL for quantifying the informativeness of a data sample from the underlying data distribution for a given task. Depending on how informativeness is evaluated, these approaches can be mostly categorised into two groups: sampling by uncertainty and sampling by representativeness (Yang et al., 2015, 2017; Budd et al., 2019).

1.1.1. Uncertainty Sampling

The simplest and most intuitive way is to suggest samples about which it is uncertain how to annotate. By asking human experts to annotate these samples and including these annotations in the training set, more information can be learnt by the model. This is known as the uncertainty method (Settles, 2009). The uncertainty of a sample can be quantified in different ways; one way is to quantify uncertainty by the class probability of its most likely label. Samples whose most likely label have the lowest probability are most uncertain, these are referred to as the least confident method in (Settles, 2009). For image segmentation tasks, this method can be extended to measuring the sum of lowest class probability for each pixel in a given image segmentation (Budd et al., 2019). More certain predictions will have higher pixel-wise class probabilities. Therefore, a lower sum of the maximal class probability over all pixels in an image would indicate a higher uncertainty of prediction. However, the least confident method suffers the problem of under-representing the distribution of less likely labels, as only information about the most likely labels are considered. Marginal sampling was proposed as an improved version of the least confident sampling method (Settles, 2009), this considers the first and second most probable labels for multi-class scenarios and calculates the difference between them. Another commonly adopted approach is to use the entropy of the probability over all labels for quantifying the uncertainty (Settles, 2009).

Ensemble-based uncertainty measurement (Yang et al., 2017; Jungo and Reyes, 2019) is another direction for uncertainty estimation, it requires training multiple models or applying test-time data augmentation such as Monte-Carlo (MC) dropout for estimation. Ensemble-based methods usually measure the agreement between different

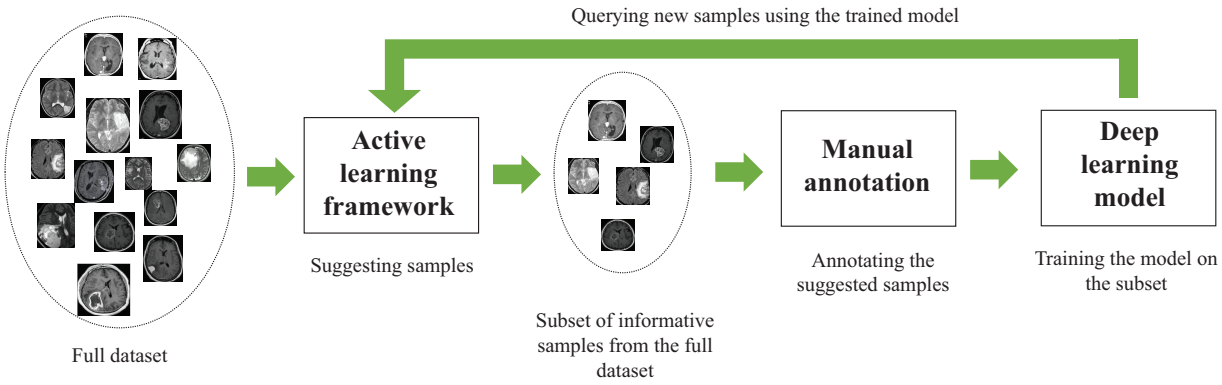


Figure 1: A typical framework of active learning. A full dataset without annotation is fed to an active learning framework, which suggests a subset of the unannotated dataset to be annotated by the experts. The annotated subset can then be used to train a deep learning model. The suggestion can be done iteratively until the deep learning model reaches a satisfactory performance for the downstream task.

models that perform the same task; assuming data samples that cause more disagreement between models imply a higher level of uncertainty and informativeness.

Another option for measuring uncertainty is by predicting the training loss (Yoo and Kweon, 2019). The method adds a separate task head to the backbone network, which is trained for predicting the training loss of the downstream task. This allows the training loss to be estimated on the unannotated data and data samples with higher predicted loss can be sampled for annotation (Yoo and Kweon, 2019). Furthermore, Bayesian networks have been explored to model the uncertainty of predictions either directly (Gal et al., 2017) or by incorporating geometric smoothness priors (Konyushkova et al., 2019).

1.1.2. Representativeness Sampling

Representativeness sampling strategies are mostly used in addition to uncertainty-based methods (Yang et al., 2017, 2015; Budd et al., 2019) to avoid sampling from the same area of the data distribution and creating a biased model towards this particular area.

(Yang et al., 2017) proposed to sample uncertain data samples that better represent the full dataset by solving a maximum set-cover (max-cover) problem on image feature vectors. In (Zheng et al., 2019), the authors adopted a hierarchical clustering method in addition to the max-cover method, in which sampling was performed in a latent space learnt from generative models such as the generative adversarial network (GAN) (Goodfellow et al., 2014a) or the variational autoencoder (VAE) (Kingma and Welling, 2014). In (Zheng et al., 2019), uncertainty is not measured in order to decouple the sampling and segmentation processes. In (Smailagic et al., 2018), the distance between feature descriptors are calculated to make sure that the sampled images are as dissimilar to each other as possible so that a large area of the data distribution can be covered. Similarly in (Zhou et al., 2017), a patch-level diversity score is calculated in order to select a sample that is different from the previous one.

1.2. Contributions

Inspired by the success of adversarial learning (Goodfellow et al., 2014b; Kurakin et al., 2017) using adversarial samples to improve the efficiency of model training, we propose a gradient-guided sampling method to suggest informative data samples for manual annotation on a data manifold learnt by the generative model.

Major contributions of this work include:

- We incorporate the loss gradient into the suggestive annotation process, an area that has not been investigated in depth for active learning in medical image segmentation tasks.

- We utilise a generative model, the VAE, to model the data distribution on a latent space. This al-

lows the loss gradient to be projected from a high-dimensional image space into a low-dimensional latent space so sampling can be performed more efficiently.

- The proposed method suggests a batch of unannotated data samples based on the training loss of the samples suggested in the previous iteration and does not require ranking for the unannotated data samples based on their levels of uncertainty or representativeness.

- We re-implement and evaluate some of the state-of-the-art active learning methods published previously for different tasks on brain image segmentation task.

A preliminary version of this work was presented at MICCAI 2020 (Dai et al., 2020). Here we have substantially extended the conference paper, including:

- Previously, we only investigated the scenario of suggesting one batch of samples for manual annotation. To explore the full potential of the proposed framework, here we extend to an iterative sampling and annotation scenario.

- Previously, the framework was tested on a single-class brain tumour segmentation task on the BraTS dataset. Here we extend to the more challenging multi-class segmentation task.

- Apart from brain tumour segmentation task, we further include a new dataset and evaluate the method performance for whole brain segmentation task.

- Furthermore, we add other previously published state-of-the-art methods for comparison.

2. Methods

In our framework, we assume that there is a large pool of unannotated medical images $U = \{x_1, x_2, ..., x_n\}$ and there is one or multiple experts from whom we can request manual annotation for a data sample $x_U \in U$. A labelled training set L normally contains all the data samples from U and their annotations. The proposed framework aims to find an optimal subset $L' \subseteq L$ so that a model trained on L' can achieve comparable performance to the model trained on L.

An overview of the proposed framework is illustrated in Fig. 2. To utilise the gradient of loss from the segmentation model for sampling, we first train a VAE on the unannotated dataset U to learn a data manifold. The encoder of the VAE enables the projection of image samples onto the manifold. Then the sampling process consists of two steps: (1) Train an image segmentation model $f(x|L')$ in an active learning manner, where L' denotes the set of annotated samples suggested by the framework. (2) Suggest unannotated samples by exploring the latent space. After training for several epochs, the gradient of the segmentation training loss is backpropagated to the image space, then projected to the latent space using the VAE encoder. Unannotated samples are explored in the latent space and suggested based on the loss gradient. They are referred to experts for annotation and added to L'. The framework iterates between Step (1) and Step (2), gradually incrementing the labelled set for training the model.

Step (1) and (2) are performed iteratively to increment the labelled subset L' that contains the most informative data samples given a intermediate model $f(x|L')$ and an unannotated dataset U. Both the model and the annotated dataset L' will evolve over time within this framework.

2.1. Learning the Data Manifold

A data manifold that reflects the structure of the unannotated dataset is fundamental to the proposed framework. In this work, the VAE is used for learning the data manifold given its good potential shown in other works (Liu et al., 2017). The VAE is trained on U with the loss function formulated as

$$\begin{aligned}\mathcal{L}_{vae}(\theta, \phi; x_i) = MSE(f_\theta(g_\phi(x_i)), x_i) \\ + D_{KL}(q_\phi(z|x_i)||p_\theta(z)),\end{aligned} \tag{1}$$

where $g_\phi(\cdot)$ and $f_\theta(\cdot)$ denote the encoder and decoder, which typically consist of a number of convolutional layers (Kingma and Welling, 2014). MSE denotes the mean square error function and D_{KL} denotes the KL-divergence, which regularises the optimisation problem by minimising the distance between the latent variable distribution and a Gaussian distribution (Kingma and Welling, 2014). Once trained, the VAE can be used to obtain the latent representations of samples $Z = \{z_1, z_2, ..., z_n\}$ given U, which will be used for the main framework explained below.

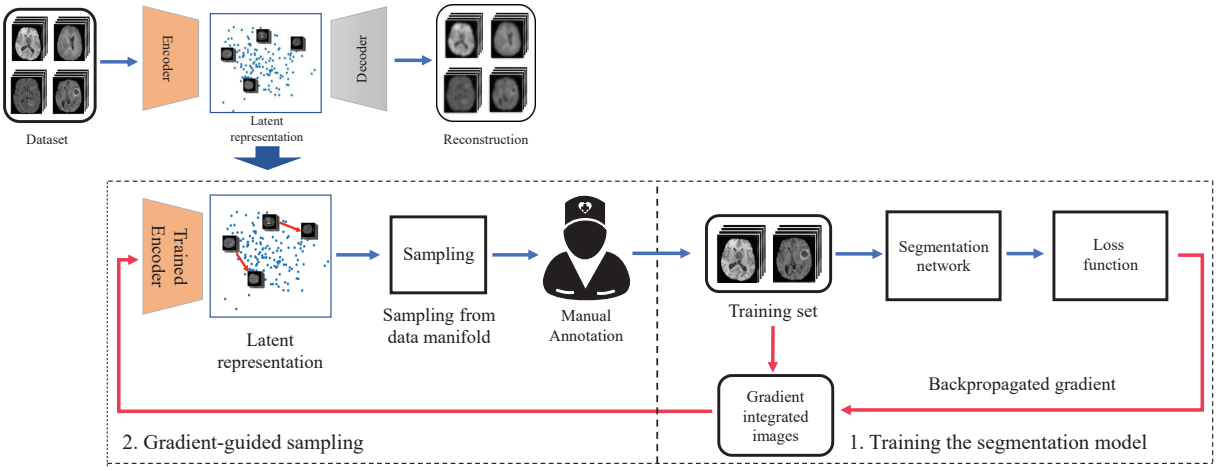


Figure 2: An overview of the gradient-guided suggestive annotation framework. The top panel illustrates the learning of a data manifold for the full dataset using a VAE, which enables the projection of samples from image space to latent space. The bottom panel illustrates the iterative steps of the framework: (1) The segmentation network is trained using suggested samples. (2) Explore the latent space to suggest samples for annotation, guided by the gradient of loss from the segmentation network. The loss gradient is integrated with samples in the image space and then projected to the latent space for exploration.

2.2. Step (1): Training the Segmentation Model

A segmentation model will be trained, using the labelled set L'. For the first iteration, a set of m random samples will be annotated and used to initialise the model training. For all subsequent iterations, informative samples will suggested, annotated and added to L' to re-train the model. This is consistent with the observations in (Kumar et al., 2010; Bengio et al., 2009) that less *informative* samples are more important at the early stage of model training, while *harder* samples are more important at the later stage.

An annotated dataset containing m initial samples $L' = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ is thus constructed, where y denotes the annotation by the expert. Unlike the single Dice loss function we used in (Dai et al., 2020), the segmentation model is trained on L' using the multi-class Dice loss function defined by,

$$\mathcal{L}_{Dice}(y_i, \hat{y}_i) = -\frac{2}{|K|}\sum_{k\in K}\frac{\hat{y}_{ik}y_{ik}}{\hat{y}_{ik}^2 + y_{ik}^2}, (x_i, y_i) \in L'. \quad (2)$$

where $\hat{y}_i$ denotes the segmentation probability map for image x_i and y_i denotes the ground truth annotation by the expert. $\hat{y}_i$ and y_i contain K channels of probabilities for K classes. $\hat{y}_{ik}$ denotes the probability map for class $k \in K$ and y_{ik} denotes the probability for class k from the one-hot encoding of ground truth.

2.3. Step (2): Gradient-guided Sampling

After the segmentaion model is trained, for each of the m samples $x_i \in L'$, its gradient of loss is backpropagated to the image space according to,

$$x_i' = x_i + \alpha \cdot \frac{\partial L_{Dice}}{\partial x_i}, \quad (3)$$

where the gradient of the loss function informs the direction to more informative samples and α denotes the step length along the gradient.

Then we sample informative data samples on the data manifold and strikes a balance between exploring uncertain samples and representative samples. Using the VAE encoder $g_\phi(\cdot)$, the hard sample x_i' can be projected to the latent space by,

$$z_i' = g_\phi(x_i'). \quad (4)$$

Fig. 3 illustrates the sampling process in the latent space. The blue dots represent samples from the unannotated dataset U. The red dots are the latent representation

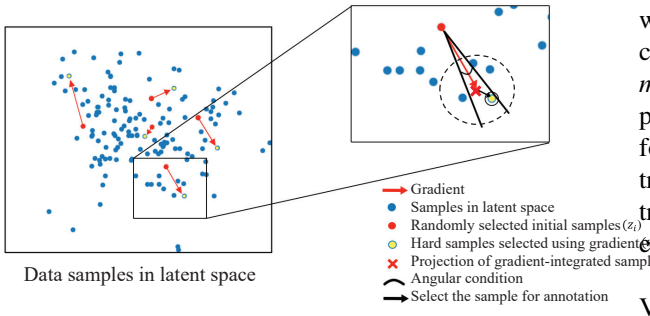


Figure 3: Sampling process in the latent space. The zoomed view shows an example of a newly suggested sample. The red dot denotes a random initial sample z_i. The red cross denotes the gradient-integrated sample z_i'. The proposed method searches along the gradient direction (red arrow) within a region defined by the angular condition and suggests the new sample z_j (yellow dot) for manual annotation.

of the samples suggested in the previous iteration. After training the segmentation model with these samples in the previous iteration, for any x_i from L', the loss gradient is backpropagated to the image space and integrated to x_i (Eq. 3), then projected to z_i' in the latent space (Eq. 4 and red cross in Fig. 3). We empirically define two criteria for suggestion. First, an existing real image should be close to z_i' in the latent space, which is to find a $z_j \in Z$ (yellow dot) that has the shortest Euclidean distance to z_i' using

$$d(z_j, z_i') = |z_j - z_i'|. \tag{5}$$

Second, an angular condition is introduced to limit the search angle (black angle in Fig. 3). If the selected sample does not satisfy the angular condition, we check the sample that has the second shortest Euclidean distance. The angular condition is defined using the cosine distance widely used in machine learning:

$$cos(\theta) = \frac{(z_j - z_i') \cdot (z_i' - z_i)}{\|z_j - z_i'\| \cdot \|z_i' - z_i\|}. \tag{6}$$

The angular condition was used to make sure the search would not deviate from the gradient direction.

Using the gradient-guided sampling method, one or more samples can be found for each z_i'. We only select one sample $z_j \in Z$ in the neighbourhood of z_i' in the latent space, which corresponds to image $x_j \in X$ in the image space. In this way, we sample m informative data samples in the first iteration and suggest them to the expert for manual annotation. An updated training dataset with m more samples $L' = \{(x_1, y_1), (x_2, y_2), ..., (x_{2m}, y_{2m})\}$ can be constructed for second iteration, which consists of m initial samples and m new samples suggested by the proposed method. The enlarged training set can be used for fine-tuning the segmentation model. If we repeat the training and suggesting steps iteratively for n times, the training set would contain $(n + 1) * m$ data samples in the end.

Theoretically, images can be synthesised from z_i' via the VAE decoder $f_\theta(\cdot)$ and suggested to the expert for annotation. However, the synthetic image may not be of high quality in reality, which would prevent the expert from producing reliable annotation. To mitigate this issue, here we sample in the real image space, searching for existing real images that are most similar to the synthesised image in the latent space. In this way, the expert would be able to annotate on high-quality real images.

3. Experiments

3.1. Datasets

We evaluated the proposed method on two public brain MRI datasets.

A. Brain Tumour Segmentation Challenge (BraTS) 2019: This dataset contains T1, T1 gadolinium (Gd)-enhanced, T2 and T2-FLAIR brain MRI volumes of 335 patients diagnosed with high-grade gliomas (HGG) or low-grade gliomas (LGG), acquired with different clinical protocols from multiple institutions (Bakas et al., 2017). The dataset was pre-processed with skull-striping, interpolation to a uniform isotropic resolution of 1×1×1 mm^3 and registered to the SRI24 template space (Rohlfing et al., 2010) with a dimension of 240 × 240 × 155. During pre-processing, the first and last few image slices (normally blank) were discarded the resulting dimension of the image is 256 × 256 × 150. The images were manually annotated with four labels: background, the peritumoural edema, the Gd-enhanced tumour (ET) and the necrotic and non-enhancing tumour core. The latter two labels were combined to form the tumour core (TC) label and latter three labels were combined to form the whole tumour (WT) label, which were the structures for evaluating segmentation accuracy in the BraTS challenge (Bakas et al., 2018). The dataset was randomly split into 260/75 for training and test, respectively.

B. Multi-Atlas Labelling Challenge (MALC): The dataset contains brain MRI scans from 30 subjects with manual annotations of the whole brain for 134 fine-grained regions including the background (Landman and Warfield, 2012). A coarse-grained segmentation map that contains 25 major brain structures (110 sub-regions are combined to create structure cortical gray matter left and right) was generated and evaluated in this work. The 25 major brain structures were previous chosen by multiple researchers (Wachinger et al., 2018; Roy et al., 2019) for investigating the performance of the segmentation algorithms. The dataset was split into 20/10 for training and test, respectively.

We further processed both BraTS and MALC datasets with zero padding and z-score intensity normalisation.

3.2. Experimental Design

For MR image segmentation tasks, the framework can be applied to either 3D or 2D networks. Due to the memory limit of standard graphics processing units (GPUs) that was available to us (NVIDIA Tesla P100, 16GB RAM), we utilised 2D networks in this work for image segmentation.

A VAE with the similar architecture from (Liu et al., 2017) was used for learning the data manifold. The VAE has four levels of blocks and each level contains one residual block with batch normalization and leaky ReLU (with a leakiness of 1e-2) activation function. It was trained for 50 epochs on both datasets with the Adam optimizer, a learning rate of 1e-4, a mini-batch size of 16, and the loss function defined in Eq. 1. A large z-dimension would result in a sparsely populated latent space, making gradient-guided suggestion difficult. Therefore, we empirically set the z-dimension to 3 for the BraTS dataset and 2 for the MALC dataset as the MALC dataset is much smaller than the BraTS dataset.

The vanilla U-Net architecture from (Ronneberger et al., 2015) was chosen for image segmentation with minor changes when training on the brain imaging dataset. The U-Net has four levels of blocks. Each block contains two convolutional layers with batch normalization and ReLU activation function, one max pooling layer in the encoding part and de-convolutional layers in the decoding part. The number of convolutional filters in each block is: 32, 64, 128, and 256; and the bottleneck layer has 512 convolutional filters. The U-Net was trained with the Adam optimizer using the loss function given by Eq. 2, a learning rate of 1e-3 and a mini-batch size of 32.

Data augmentation including flipping, random intensity and random elastic deformation were conducted on-the-fly during training. The step length α in Eq. 2 was set to 1e-3 empirically.

To initialise the segmentation network for the BraTS dataset, we randomly sampled 0.5% data samples from the BraTS dataset to train the network for 10 epochs. Then we applied the proposed sampling method to sample 0.5% more data samples from the dataset for manual annotation and added the annotated samples to the training set, making it 1% of the full dataset. The segmentation network was trained on the enlarged training set for 10 more epochs. We performed suggestive sampling and network training for 8 more iterations, each time adding 1% data samples with manual annotations to the training set, followed by network training.

For the experiment on the MALC dataset, the segmentation network for the MALC dataset was initialised and trained in a similar fashion. Due to the relatively small size of the MALC dataset, each iteration we added 6% data samples suggested by the method to the training set and performed network training for 10 epoch. We repeated the same training and suggestion processes for 8 more iterations on MALC datasets, adding 6% more data samples to the MALC training dataset each time.

We compared five different suggestive annotation methods: random suggestion, the deep active learning framework (DAL) (Yang et al., 2017), the clustering and max-cover based method (ClsMC) (Zheng et al., 2019), the learning loss method (LL) (Yoo and Kweon, 2019) and the oracle method. Random suggestion is the baseline method, which randomly selects data samples for training the segmentation model each time. DAL suggests both uncertain and representative samples. In each iteration, uncertain samples are selected using bootstrapping. Representative samples are selected by suggesting samples that bear useful features for as many unannotated samples as possible based on the cosine distance between high level features of candidate samples. (Yang et al., 2017). ClsMC suggests samples based on the representativeness qualified by adopting hierarchical clustering and max-cover algorithms in the latent space. The suggestion is made one-time only (Zheng et al., 2019). LL method selects samples that have the highest training loss pre-

dicted by a network using U-Net as the backbone and a separate head for predicting the training loss. For the LL method (Yoo and Kweon, 2019), samples are suggested iteratively in the same fashion as the proposed method. DAL, ClsMC and LL methods were originally proposed for different applications or on different datasets from this work. Where the original code was not available, we re-implemented DAL, ClsMC and LL methods and adopted the same experiment setting to enable a fair comparison of method performance.

For the DAL re-implementation, we made some modifications to cope with the experimental design in this paper. We replaced the FCN in the original framework with the same U-Net used by our framework. We trained 4 U-Nets to measure the uncertainty score. The High level features were extracted from the bottleneck layers of the U-Nets for measuring the cosine similarity. Due to the large number of samples suggested in each iteration, we had to empirically choose the number of candidate samples and perform max-cover algorithm multiple times in each iteration to reduce the computational complexity (e.g., if 200 samples are needed in an iteration, we select 20 samples with the highest uncertainty scores from the unannotated dataset and perform max-cover algorithm to suggest 10 samples from 20 candidate samples. The suggestion process is done 20 times in order to suggest 200 samples from 400 candidate samples.). We trained the U-Nets following the same procedure of the proposed method. Instead of selecting a best model using validation set, the result of the ensemble of 4 U-Nets was used for evaluation.

For the ClsMC re-implementation, we followed the steps described in (Zheng et al., 2019) by first conducting agglomerative clustering on the latent representations, for deciding how many clusters we would select images from. We then selected images from clusters using max-cover strategy and further select most representative images from previously selected images. We set the size of the candidate set in each cluster following a similar approach described in (Zheng et al., 2019) taking the computational complexity into account. The pixel ratio defined in (Zheng et al., 2019) is replaced with the percentage of full dataset used. The suggestion process stops when the size of the training set reaches a certain percentage of the full dataset. The U-Net trained on the samples selected by ClsMC and the training configuration are the same as the proposed method.

For the LL re-implementation, we attached a similar network structure for the loss prediction to the U-Net backbone and used the same loss prediction method as described in (Yoo and Kweon, 2019), which calculates the difference between a pair of loss predictions. The loss function for segmentation network is Dice loss. The margin and scaling constant for the loss function defined in (Yoo and Kweon, 2019) were set to 0.5 and 1 respectively. The whole LL network was trained for 50 epochs using the same learning rate, optimiser and batch size as the U-Net model used by other methods. We stopped propagating the gradient from the loss prediction module to the segmentation module after 30 epochs following a similar procedure mentioned in (Yoo and Kweon, 2019).

The oracle method serves as an empirical upper-bound for suggestive annotation. It assumes that the annotations are already known for all samples, and we choose the posteriori 'best' samples according to the downstream task. For segmentation tasks, we use the most challenging samples for the current segmentation model with lowest predicted Dice scores. To allow a fair comparison, each experiment was repeated 10 times and the average performance was evaluated.

4. Results

4.1. Parameter settings

An optimal manifold representation of the dataset and a sampling mechanism that can maximise the effectiveness of using gradient of loss from the segmentation network are the two most essential elements in the proposed framework. Consequently, we experimented with several parameter settings for learning the data manifold and for sampling. For the data manifold, we tested different dimensions for the latent space z in training the VAE. For the sampling mechanism, we explored sampling new data samples on the data manifold using Euclidean distance only and with an additional angular condition. The results of using different parameter settings are reported below.

4.1.1. Dimensionality of the Latent Space

We found that a higher z-dimension number brings a better quality to the image reconstruction of the VAE, but a more challenging data manifold for the subsequent sampling process. The results of using different dimensions

Table 1: Influence of the latent space z dimension on the method performance. The performance is evaluated using the mean Dice scores of three tumour structures (whole tumour, enhancing tumour and tumour core) when training the segmentation model using samples suggested in the latent space. 3% or 6% of the full BraTS dataset are suggested. Figures in bold are the best results among using a z-dimension number of 3, 10 and 50.

	3% dataset			6% dataset		
z dimension	3	10	50	3	10	50
Whole Tumour	**0.8238**	0.8081	0.8013	**0.8437**	0.8259	0.8195
Enhancing Tumour	**0.6972**	0.6783	0.6692	**0.7182**	0.7049	0.6898
Tumour Core	**0.7619**	0.7542	0.7479	**0.7895**	0.7687	0.7582

of z are compared in Table 1. It shows that the performance of using a 3-dimensional latent space outperforms using a 10-dimensional and a 50-dimensional spaces. A 50-dimensional space leads to an exceptionally good image reconstruction quality for VAE, which is potentially caused by an overfitting latent space given the size of the training set. The segmentation result of using such a high dimensional space is close to random suggestion method on the BraTS dataset as shown in Table 1. The inferior performance of a higher z dimension number is possibly because that in a high-dimensional latent space, a limited amount of data samples will be sparsely distributed and searching samples along the gradient direction in such a space is challenging.

4.1.2. Ablation Study on Angular Condition

We investigated the effectiveness of angular condition for sampling. Samples on the data manifold are suggested with or without adding the angular condition. The segmentation performance is compared in Table 2. For suggesting with angular condition, an empirical threshold of $\pi/3$ is set for the angular condition. It shows that searching samples with angular condition outperforms that without angular condition. Our experiments indicate that sampling without the angular condition is highly likely to cause the direction of searching new samples on the data manifold to deviate from the gradient, undermining the effectiveness of using gradient in the framework. Therefore, angular condition is used throughout the rest of the experiments in this work.

4.2. Brain Tumour Segmentation on BraTS dataset

We first evaluated the proposed framework for brain tumour segmentation task. Fig. 4 compares the performance of proposed method to the random suggestion method and the oracle method (upper bound), in terms of the average Dice score across three tumour structures. It shows that with the increase in the size of labelled dataset, the performance of the segmentation model improves for all three methods and the proposed framework consistently outperforms the random suggestion. The performance increase becomes slower after the training set is expanded to 5% of the full dataset. The proposed method achieves slightly inferior performance than the oracle method.

We further compared the performance of the proposed method to other suggestive sampling methods, including DAL, ClsMC and LL. In this comparison, all methods suggested 3% and 7% of the full dataset. Tables 3 and 4 compare the mean Dice score and Hausdorff distance of the methods for each tumour type. We also verified the statistical significance of the results by performing the Wilcoxon signed-rank test between two top performing methods among random, DAL, ClsMC, LL and ours. The result of the Wilcoxon signed-rank tests are included in Tables 3 and 4. It shows that the proposed method yields slightly inferior performance as the oracle method and it outperforms random, ClsMC and LL methods in terms of Dice score. For Hausdorff distance, the proposed method outperforms random, ClsMC and LL methods for segmenting whole tumour and tumour core. Furthermore, by training using only 7% of the suggested samples, the segmentation performance is close to the performance of training on the full dataset.

The random method outperforms the ClsMC method in some cases as shown in Tables 3 and 4. This is most likely because a random method is quite a strong baseline. As discussed in (Kumar et al., 2010; Bengio et al., 2009), less informative samples are more important at the early stage of the model training. Nonetheless, ClsMC suggests samples only once based on the pre-defined similarity/coverage scores, therefore it is not feasible for ClsMC to keep improving the quality of suggested samples based on the performance of the segmentation model.

Fig. 5 illustrates the t-SNE plot of the latent space for images from the BraTS dataset. The right side of Fig. 5 shows three examples of the sampling process, including the initial images that provide the loss gradient and the suggested images after search in the latent space in the following two iterations. As it shows, the suggested images tend to present different intensity contrasts or differ-

Table 2: Influence of the angular condition on the method performance, evaluated using the mean Dice scores for three tumour structures when training the segmentation model using suggested samples. 3% or 6% of the full BraTS dataset are suggested. Figures in bold are the results of using random method, our method withut angular condition and our method with angular condition.

	3% dataset			6% dataset		
Methods	Random	W/o angular condition	With angular condition	Random	W/o angular condition	With angular condition
Whole Tumour	0.8124	0.8173	**0.8238**	0.8225	0.8243	**0.8437**
Enhancing Tumour	0.6857	0.6723	**0.6972**	0.7062	0.6985	**0.7182**
Tumour Core	0.7509	0.7431	**0.7619**	0.7703	0.7562	**0.7895**

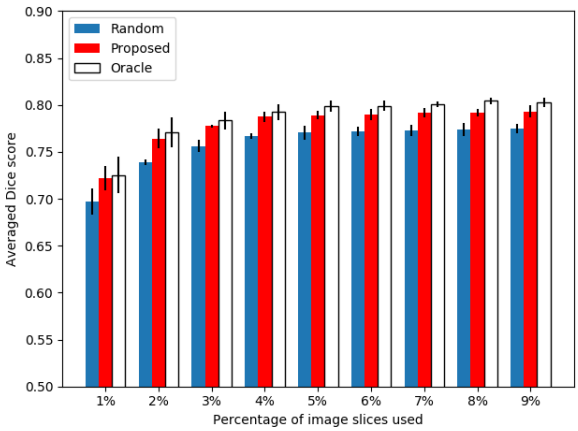


Figure 4: Comparison of the proposed method with the random suggestion method and the oracle method. For each iteration, data samples of 1% of the full BraTS dataset are suggested and added into the training set.

ent anatomical structures from the initial images, which are likely to provide more information for segmentation model training.

4.3. Whole Brain Segmentation on the MALC dataset

We then evaluated our method on the MALC dataset for whole brain segmentation task, in total segmenting 25 brain structures. Fig. 6 compares the performance of the proposed method to the random suggestion method and the oracle method, in terms of the averaged Dice score of 25 brain structures. The performance of all methods improves with the increase of the size of the training set. The proposed method constantly outperforms the random method during the experiment and achieves comparable performance to the oracle method when using 60% of the training set. When training with 12% of full MALC dataset, random suggestion yields superior performance to both proposed and oracle method. Similar to the good performance of the random method we observed when comparing with the ClsMC method for suggesting brain tumour images, one plausible reason for this was that less informative samples are more important at the early stage of the model training. Fig. 7 compares the performance of proposed method to the baseline, oracle methods for the Dice score of each individual brain structures when using 42% of training set. The result of training on the full training set of MALC dataset is also shown on Fig. 7. The proposed method outperforms the random method across all brain structures, achieving comparable performance to the oracle method and even close to the segmentation model trained on the full training set.

The training loss curves of the proposed method, baseline method and oracle method are illustrated in Fig. 8. As shown in the figure, in the beginning of the experiment, the training losses of the proposed method and the baseline method decrease at a similar rate. However, with more suggested images being added into the training set, the training loss curve of the proposed method decreases further compared to the baseline method and becomes closer to the oracle method.

Table 5 compares the performance of the proposed method and other competitive method when training using 24% and 42% of MALC dataset. We also performed the Wilcoxon signed-rank test between two top performing methods among random, DAL, ClsMC, LL and ours for Dice score and Hausdorff distance. It shows that the proposed method achieves better performance than random, DAL, ClsMC and LL methods in terms of the mean Dice score of 25 brain structures for using 42% of MALC dataset. In terms of the Hausdorff distance, the ClsMC

Table 3: Comparison of the proposed method with other suggestion methods in terms of average Dice score and standard deviation, using 3% and 7% of full training set of BraTS dataset. The figures in bold are the best results among random, DAL, ClsMC, LL and our method. The results of training on the full training set and the Wilcoxon signed-rank test between two top performing methods among random, DAL, ClsMC, LL and ours are also given in this table.

Methods	3% dataset			7% dataset		
	Whole tumour	Enhancing tumour	Tumour core	Whole tumour	Enhancing tumour	Tumour core
Random	0.8124±0.0172	0.6857±0.0237	0.7509±0.0189	0.8402±0.0197	0.7035±0.0201	0.7701±0.0177
DAL (Yang et al., 2017)	0.8156±0.0146	0.6892±0.0193	0.7535±0.0211	0.8457±0.0133	0.7093±0.0181	0.7795±0.0134
ClsMC (Zheng et al., 2019)	0.8113±0.011	0.6823±0.0122	0.7587±0.0131	0.8369±0.0154	0.7018±0.0186	0.7801±0.0164
LL (Yoo and Kweon, 2019)	0.8207±0.0125	0.6874±0.0171	0.7562±0.014	0.8491±0.0106	0.7103±0.0193	0.7746±0.0129
Proposed	**0.8238±0.0101**	**0.6972±0.0103**	**0.7619±0.0137**	**0.8589±0.0093**	**0.7291±0.0127**	**0.7947±0.0105**
Oracle	0.8297±0.0141	0.7062±0.0149	0.7649±0.017	0.8614±0.0101	0.7396±0.0116	0.7998±0.0112
Full dataset	-	-	-	0.8663±0.0081	0.7457±0.0104	0.8026±0.0093
p-value	≈0.27	≈0.05	≈0.04	<0.01	<0.01	<0.01

Table 4: Comparison of the proposed method with other methods in terms of averaged Hausdorff distance (unit: mm) and standard deviation, using 3% and 7% of full dataset. The figures in bold are the best results among random, DAL, ClsMC, LL and our method. The results of training on the full training set and the Wilcoxon signed-rank test between two top performing methods among random, DAL, ClsMC, LL and ours are also given in this table.

Methods	3% dataset			7% dataset		
	Whole tumour	Enhancing tumour	Tumour core	Whole tumour	Enhancing tumour	Tumour core
Random	12.35±1.49	**16.05±2.37**	20.73±2.89	10.29±1.26	**14.75±1.79**	19.54±1.87
DAL (Yang et al., 2017)	12.01±1.23	16.56±2.1	20.03±2.95	10.27±1.04	15.39±1.8	19.23±2.13
ClsMC (Zheng et al., 2019)	11.97±1.27	16.43±2.05	**19.81±2.4**	10.58±1.12	15.24±1.92	19.06±1.95
LL (Yoo and Kweon, 2019)	12.21±1.2	16.37±2.17	19.97±2.67	10.96±0.97	15.43±1.67	19.12±2.09
Proposed	**11.42±1.15**	16.13±2.03	19.89±2.53	**9.03±1.18**	15.04±1.71	**18.51±1.93**
Oracle	11.03±1.09	15.82±1.9	19.42±2.16	9.43±0.95	14.89±1.32	18.11±1.77
Full dataset	-	-	-	8.99±0.87	14.58±1.19	17.54±1.41
p-value	≈0.01	≈0.47	≈0.59	≈0.02	≈0.53	≈0.02

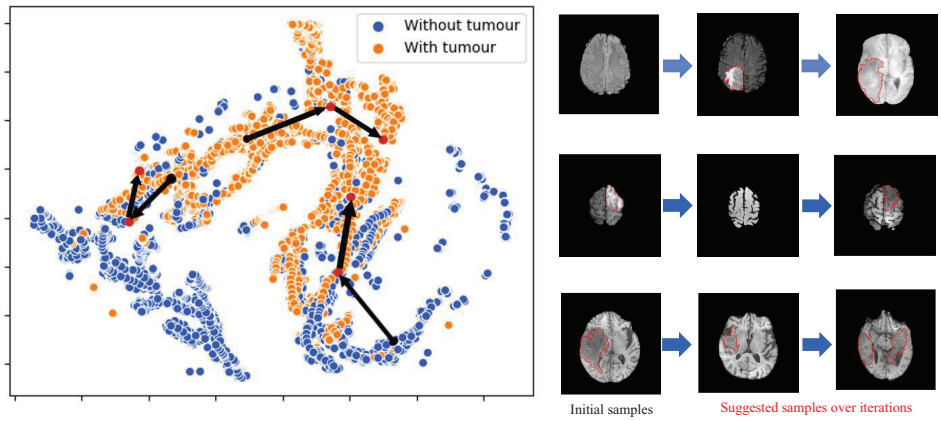


Figure 5: Qualitative illustration of the proposed method. The left panel shows the t-SNE plot of the latent space for images from the BraTS dataset. The black arrow denotes the gradient direction, starting from an initial sample (black point) and pointing to a suggested sample (red point). The right panel shows the images corresponding to the starting point and ending point of the black arrow.

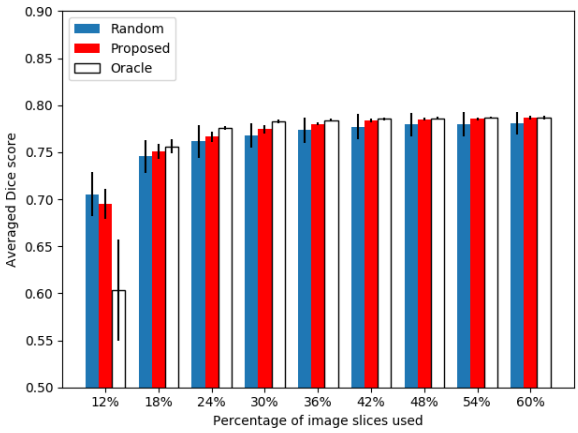


Figure 6: Comparison of the proposed suggestive annotation method with the random suggestion method and the oracle method. For each iteration, data samples of 6% of the full MALC dataset are suggested and added into the training set.

method achieves the best performance but not statistically significant.

5. Discussion

The proposed method achieved close segmentation results as training on the full dataset. For BraTS dataset, training on 7% of total image slices suggested from 335 MRI volumes by the proposed method achieved very good result. The significant reduction in the size of the training set is because the MRI volumes often contain similar adjacent image slices, which leaves us a large space for reducing the data redundancy. For a smaller MALC dataset, 42% of the of total image slices were suggested to achieve a good segmentation performance. Overall, the experimental results suggest that the proposed framework has a strong potential in reducing the annotation cost for medical imaging data, which will be beneficial to the medical AI community, especially to machine learning models (Razzak et al., 2018; van der Laak et al., 2021) that rely on training on a large amount of annotated data.

Compared to most of other AL frameworks (Yang et al., 2017; Yoo and Kweon, 2019; Liu et al., 2020a; Shen et al., 2020; Aghdam et al., 2019) which need to iterate the full unannotated dataset for each time of the suggestion to sample most uncertain and/or most representative ones, the proposed method directly suggests the informative samples by using the gradient of loss acquired from the segmentation model. Although the proposed method requires training a VAE, the learnt data manifold can be re-used for different downstream tasks. Furthermore, selecting uncertain samples and representative samples are usually done in separate steps in many active learning frameworks, but the proposed method is designed to combine these steps and strike a balance between the uncertainty and the representativeness of the selected samples to suggest the samples that are beneficial to model training. A well-learnt data manifold is a good latent representation of the original dataset. The uncertainty is taken into account by leveraging the gradient of loss from the segmentation network. The experimental results indicate that the proposed method outperforms both random suggestion and other suggestion methods that utilise the training loss (Yoo and Kweon, 2019) or the latent representation (Zheng et al., 2019). We have successfully shown the proposed method achieves a performance close to the oracle method.

Comparing to other methods that are proposed to reduce the annotation cost and to alleviate the data scarcity challenge, such as semi-supervised learning and data augmentation approaches, our proposed method encourages the idea of human-in-the-loop when training the segmentation network. We believe the experts play an irreplaceable role in providing high quality annotation for medical imaging data and efficiently keeping them involved in the process of training and tuning the segmentation network is beneficial to improving the performance of the segmentation network. On the other hand, experts are susceptible to the quality of the MR images, therefore it may be difficult to implement methods like dataset distillation for MRI segmentation task.

There are a few limitations with the proposed framework and our research. Firstly, the performance of the proposed framework relies on the quality of the data manifold learnt by the VAE. Our experiments suggest that the proposed method is more effective with a data manifold learnt with a relatively low z-dimension number. This issue may be alleviated by including more unannoated data samples for training the VAE or by generating multiple gradient-integrated images for the same image in each iteration so that the proposed method can search multiple gradient directions and multiple samples can be sug-

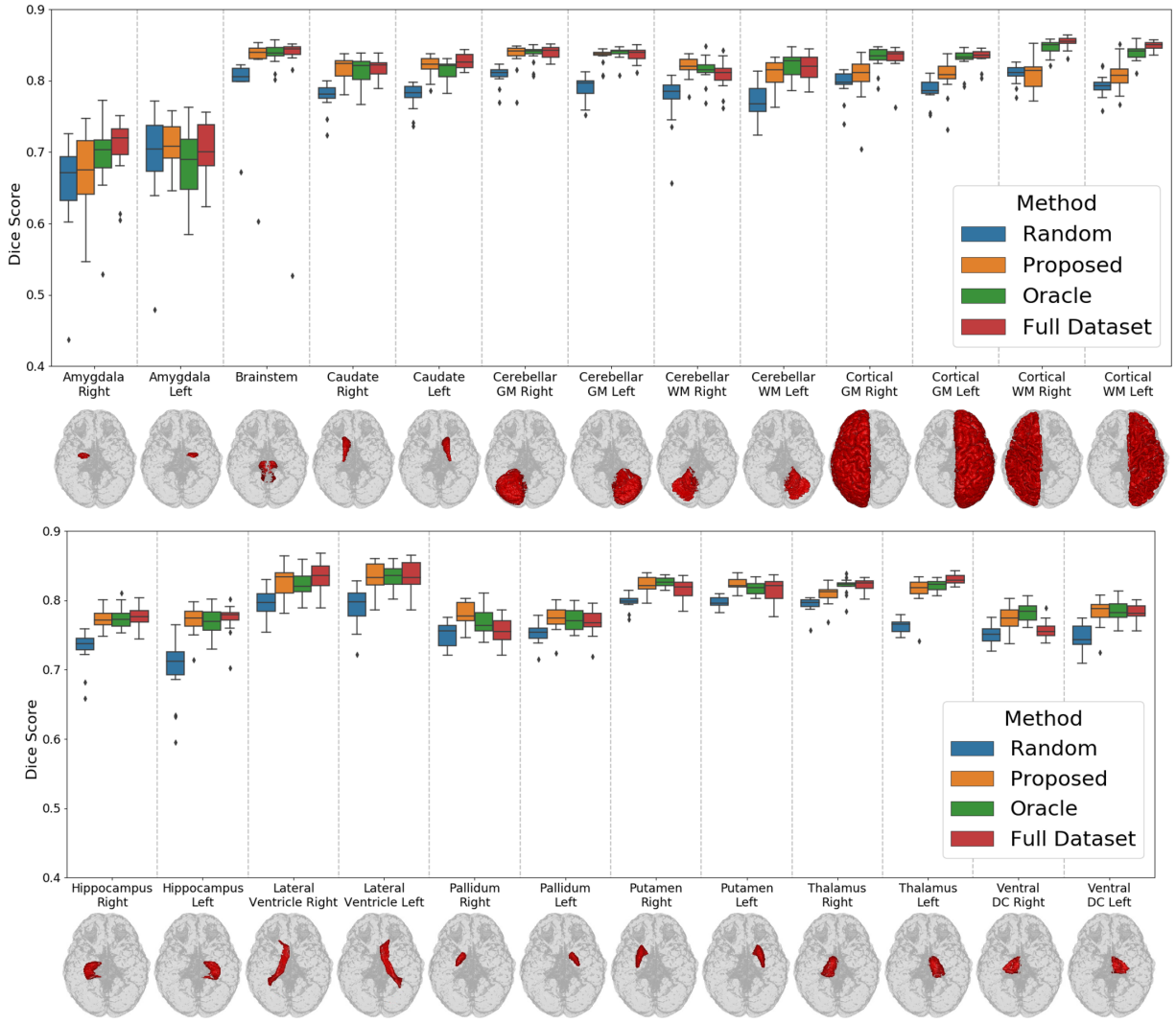


Figure 7: Boxplots of the Dice scores for 25 brain structures on MALC test dataset, comparing the proposed method training on 42% of training set with the random method, oracle method and training on full dataset.

Table 5: Comparison of the proposed method with other methods in terms of the averaged Dice score, the averaged Hausdorff distance (unit: mm) and standard deviations, using 24% and 42% of full dataset. The figures in bold are the best results among random, DAL, ClsMC, LL and our method. The results of training on the full training set and the Wilcoxon signed-rank test between two top performing methods among random, DAL, ClsMC, LL and ours are also given in this table.

Methods	24% dataset		42% dataset	
	Dice score	Hausdorff distance	Dice score	Hausdorff distance
Random	0.7427±0.0425	14.21±2.75	0.7679±0.0231	12.15±2.11
DAL (Yang et al., 2017)	0.7478±0.0262	12.9±2.24	0.7701±0.0103	10.85±1.65
ClsMC (Zheng et al., 2019)	0.7503±0.0192	**11.83±2.15**	0.7713±0.0127	**10.07±1.59**
LL (Yoo and Kweon, 2019)	0.7485±0.0233	13.41±1.99	0.7692±0.0148	11.3±1.8
Proposed	**0.7519±0.014**	12.56±2.77	**0.7782±0.0098**	10.21±1.92
Oracle	0.7742±0.0076	11.69±2.08	0.7802±0.0064	10.39±1.78
Full dataset	-	-	0.7841±0.0071	9.8±1.46
p-value	≈0.38	≈0.05	<0.01	≈0.4

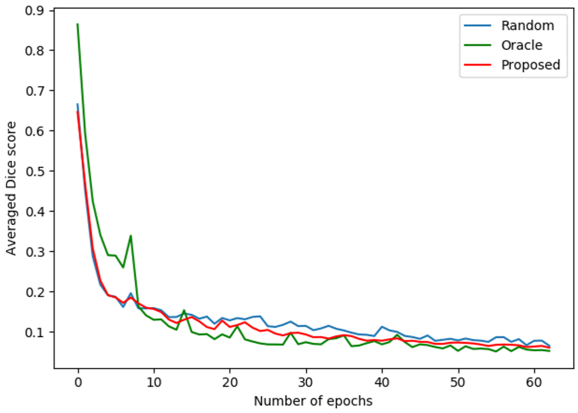


Figure 8: Comparison of training loss curves for the proposed suggestive annotation method with the random suggestion method and the oracle method.

gested based on the previous sample; and a new image that is most frequently suggested based on these gradient-integrated images will be given to the expert for annotation. Secondly, what each dimension of the latent space represents is not explainable, therefore it is difficult to understand the behaviour of the sample suggestion process clearly. This would be an interesting direction to explore in the future. Thirdly, the sampling and segmentation stages in our framework are coupled in the iterations, which means the experts for annotating the suggested images are expected to be involved after each iteration of model training. This requires close collaboration and interaction with experts in the process. Fourthly, some modifications had to be made to the original implementations of DAL, ClsMC and LL methods to adapt our experimental design, it is difficult to guarantee an optimal performance from these methods.

6. Conclusion

Here we propose a gradient-guided suggestive annotation method, which can be used to train a segmentation network efficiently with a small number of samples in an active learning manner. Using only 7% of the BraTS dataset or 42% of the MALC dataset, it achieves a high segmentation performance that is a comparable to that by training on the full dataset. The proposed framework is generic, and it has the potential to be extended to other medical image analysis tasks to reduce the annotation cost and improve learning efficiency.

7. Acknowledgement

This independent research was funded by the NIHR Imperial Biomedical Research Centre (BRC). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, NIHR or Department of Health. We appreciate the script for 3D renderings of brain structures provided by Prof. Bennett Landmann and Dr. Yuankai Huo (Huo et al., 2018).

References

Aghdam, H.H., Gonzalez-Garcia, A., Weijer, J.v.d., López, A.M., 2019. Active learning for deep detection neural networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3672–3680.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C., 2017. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. The Cancer Imaging Archive 286.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., et al., 2018. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BraTS Challenge. arXiv:1811.02629 .

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning, in: International Conference on Machine Learning, pp. 41–48.

Budd, S., Robinson, E.C., Kainz, B., 2019. A survey on active learning and human-in-the-loop deep learning for medical image analysis. arXiv preprint arXiv:1910.02923 .

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis 54, 280–296.

Dai, C., Wang, S., Mo, Y., Zhou, K., Angelini, E., Guo, Y., Bai, W., 2020. Suggestive annotation of brain tumour images with gradient-guided sampling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 156–165.

Fiez, J.A., Damasio, H., Grabowski, T.J., 2000. Lesion segmentation and manual warping to a reference brain: Intra-and interobserver reliability. Human brain mapping 9, 192–211.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep bayesian active learning with image data, in: International Conference on Machine Learning, pp. 1183–1192.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014a. Generative adversarial nets. Advances in neural information processing systems 27, 2672–2680.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2014b. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 .

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Machine Vision and Applications 31, 8.

Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A., 2018. Slant: Deep whole brain high resolution segmentation. URL: `https://github.com/MASILab/SLANTbrainSeg`.

Jungo, A., Reyes, M., 2019. Assessing reliability and challenges of uncertainty estimations for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 48–56.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International Conference on Information Processing in Medical Imaging, pp. 597–609.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes, in: International Conference on Learning Representations.

Knoll, F., Hammernik, K., Zhang, C., Moeller, S., Pock, T., Sodickson, D.K., Akcakaya, M., 2020. Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues. IEEE Signal Processing Magazine 37, 128–140.

Konyushkova, K., Sznitman, R., Fua, P., 2019. Geometry in active learning for binary and multi-class image segmentation. Computer Vision and Image Understanding 182, 1–16.

Kumar, M.P., Packer, B., Koller, D., 2010. Self-paced learning for latent variable models, in: Advances in Neural Information Processing Systems, pp. 1189–1197.

Kurakin, A., Goodfellow, I., Bengio, S., 2017. Adversarial machine learning at scale, in: International Conference on Learning Representations.

van der Laak, J., Litjens, G., Ciompi, F., 2021. Deep learning in histopathology: the path to the clinic. Nature medicine 27, 775–784.

Landman, B., Warfield, S., 2012. MICCAI 2012 Workshop on Multi-Atlas Labeling, in: Medical Image Computing and Computer Assisted Intervention Conference.

Liang, H., Tsui, B.Y., Ni, H., Valentim, C.C., Baxter, S.L., Liu, G., Cai, W., Kermany, D.S., Sun, X., Chen, J., et al., 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nature Medicine 25, 433–438.

Liu, J., Cao, L., Tian, Y., 2020a. Deep active learning for effective pulmonary nodule detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 609–618.

Liu, M.Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks, in: Advances in neural information processing systems, pp. 700–708.

Liu, Z., Chen, L., Tong, L., Zhou, F., Jiang, Z., Zhang, Q., Shan, C., Wang, Y., Zhang, X., Li, L., et al., 2020b. Deep learning based brain tumor segmentation: A survey. arXiv preprint arXiv:2007.09479 .

NHS England, 2019. Diagnostic imaging dataset statistical release. London: Department of Health .

Ravanbakhsh, M., Tschernezki, V., Last, F., Klein, T., Batmanghelich, K., Tresp, V., Nabi, M., 2020. Human-machine collaboration for medical image segmentation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1040–1044.

Razzak, M.I., Naz, S., Zaib, A., 2018. Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps , 323–350.

Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A., 2010. The sri24 multichannel atlas of normal adult human brain structure. Human brain mapping 31, 798–819.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al., 2019. Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage 186, 713–727.

Settles, B., 2009. Active learning literature survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

Shen, H., Tian, K., Dong, P., Zhang, J., Yan, K., Che, S., Yao, J., Luo, P., Han, X., 2020. Deep active learning for breast cancer segmentation on immunohistochemistry images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 509–518.

Smailagic, A., Noh, H.Y., Costa, P., Walawalkar, D., Khandelwal, K., Mirshekari, M., Fagert, J., Galdrán, A., Xu, S., 2018. Medal: Deep active learning sampling method for medical image analysis. arXiv preprint arXiv:1809.09287 .

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis:

Full training or fine tuning? IEEE Transactions on Medical Imaging 35, 1299–1312.

Wachinger, C., Reuter, M., Klein, T., 2018. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage 170, 434–445.

Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A., 2018. Dataset distillation. arXiv preprint arXiv:1811.10959 .

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 399–407.

Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G., 2015. Multi-class active learning by uncertainty sampling with diversity maximization. International Journal of Computer Vision 113, 113–127.

Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Medical Image Analysis 58, 101552–101552.

Yoo, D., Kweon, I.S., 2019. Learning loss for active learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 93–102.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019. Data augmentation using learned transformations for one-shot medical image segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8543–8553.

Zheng, H., Yang, L., Chen, J., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., Chen, D.Z., 2019. Biomedical image segmentation via representative annotation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5901–5908.

Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., Liang, J., 2017. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7340–7351.