Domain specific cues improve robustness of deep learning based segmentation of ct volumes

Marie Kloenne, Sebastian Niehaus, Leonie Lampe, Alberto Merola, Janis Reinelt, Ingo Roeder, and Nico Scherf

Scientific reports, 01 July 2020

proposed a framework that combines domain-specific data preprocessing and augmentation with state-of-the-art CNN architectures

The focus is not limited to optimise the score, but also to stabilise the prediction performance.

Why they do this?

- Although data-driven approaches are intrinsically adaptive and thus, generic, they often do not perform the same way on data from different imaging modalities.
- Computed tomography (CT) data poses many challenges, mostly due to the broad dynamic range of intensities and the varying number of recorded slices of CT volumes.

Preprocessing and Augmentation

To reduce the complexity and optimise the dynamic range, they apply a windowing to each volume by clipping the voxels grey value range to a (0.6, 0.99) percentile range that corresponds to the window a radiologist would use for decision-making.

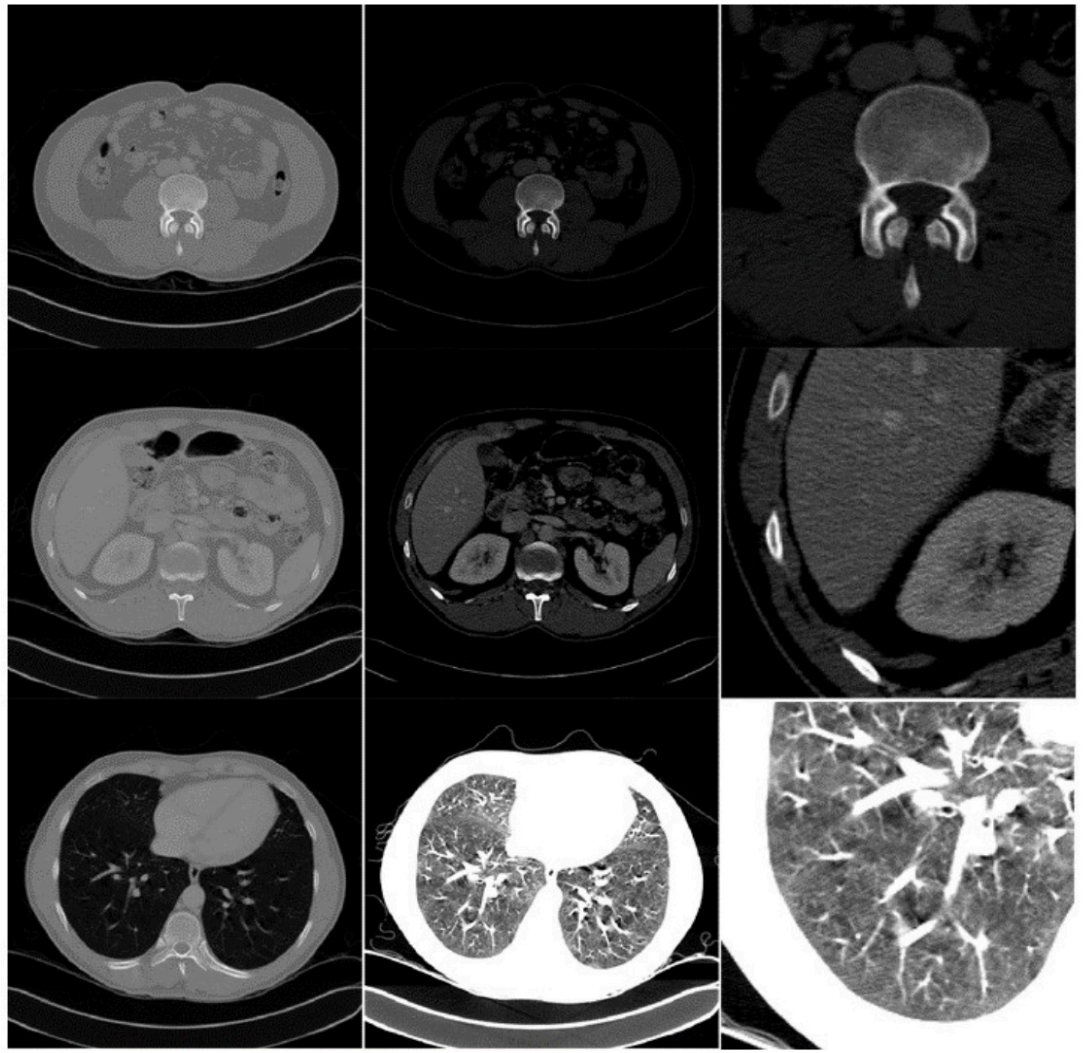

Fig. 1. Three examples for the use case oriented windowing ((A) Bone oriented windowing, (B) Organ oriented windowing, (C) Lung oriented windowing). The organ oriented windowing is applied in this work, while the other two examples would be used for the analysis of abnormalities in lung or bony structures in CT.

They then normalise the windowed data using the z-score using the intensity statistics (mean, standard deviation) from only a random sample of the data set. Using the statistical information from the full dataset would be better but does not reflect the real conditions in a clinical environment.

Thus, in a clinical setting, the number of acquired slices in a CT volume considerably varies. This poses a challenge to the application of standard CNN pipelines which often assume a regular data sampling. To standardise the data, they decided to reduce each volume to 16 slices.

They selects slices at random from each volume, and by repeating the sampling process per volume, they also get a simultaneous data augmentation effect. They exclude background slices during the training phase since these are also not considered in the test phase.

They observed that increasing the number of slices did not yield better results, which is consistent with the observation that most CNNs only use a small semantic context for decision making.

In order to save GPU memory, they downsampled each slice from 512 * 512 voxels to 128 * 128 voxels.

As additional augmentation steps they used image noising with a normally distributed noise map, slice skipping, slice interpolation and a range shift to address potential variation in the CT acquisition process.

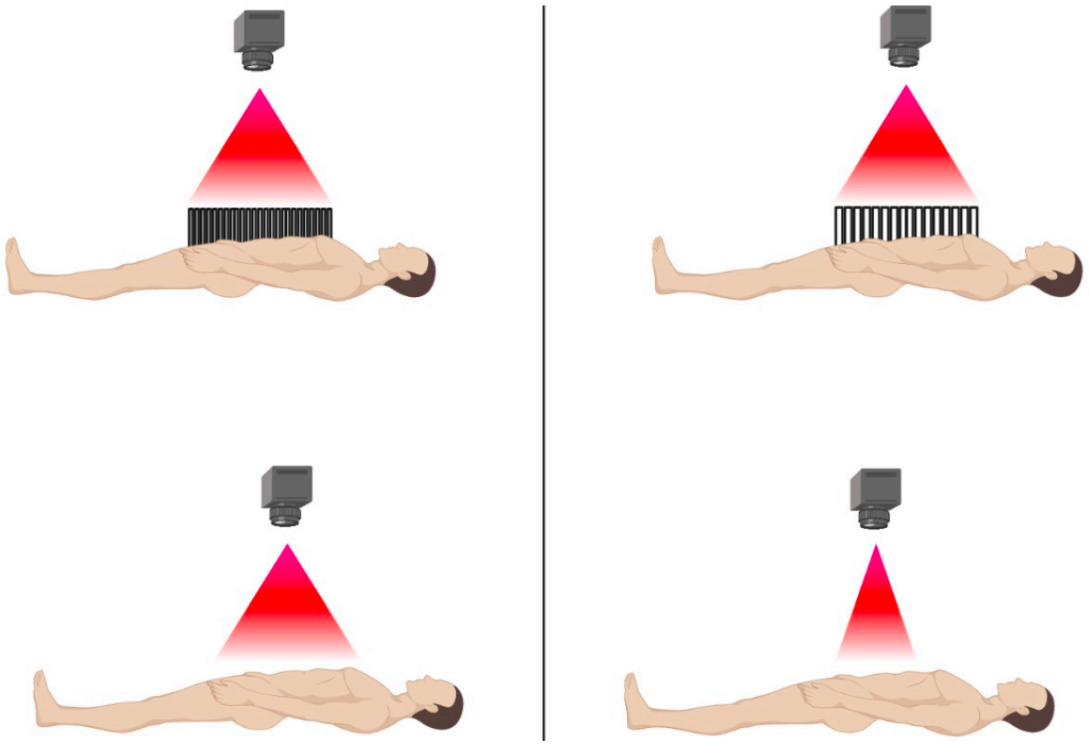

Fig. 2. CT scanning configuration, which poses challenges to the application of CNNs. The representation above presents the varying slice thickness, which allows mapping the same region of interest to a different number of slices. The representation below shows the varying size of volumes depending on the chosen region of interest.

They further rotated the images by a random angle (maximum of 16 degrees) to simulate the inevitable variability in patient positioning, that occurs in clinical routine despite fixation.

Architecture

To demonstrate the independence of the preprocessing and augmentation framework from the concrete underlying neural network architecture, they compared two conceptually different CNN models.

- nnU-Net
- Mixed-scale dense convolutional network(MS-D net)

Same normalization(instance normalization)
Same activation function(LeakyReLu units of slope 1e-2)

chose these two rather extreme variants of CNNs to compare the traditional down- and upscaling flow with the parallel multi-scale approach using dilated convolutions.

Architecture

For the kidney-tumour segmentation they stacked a set of 3D MS-D Nets trained to classify voxels into kidney and background (without a distinction between the healthy kidney tissue and the tumour tissue), and a set of 2D nnU-Nets trained to perform classification into healthy tissue, tumour and background. For the liver segmentation, both models perform binary classification of voxels into liver and background.

Training

All networks are trained independently from scratch.

Algorithm 1 Training procedure

Initialize network f with random weights θ_0
Initialize validation data $V_{validate}$
Initialize batch size n
Assume standard deviation σ
Select windowing percentile P
repeat
    repeat
        Select random volume v
        Windowing(v, P_v)
        Normalization(v, σ)
        Augmentation of v
        Downsampling and slide reduction of v
        $V_{batch} \leftarrow v$
    until Number of v in $V_{batch} = n$
    $V_{batch,\hat{y}} = f(V_{batch,x}; \theta_i)$
    $L_i = L_{Tanimoto}(V_{batch,\hat{y}}, V_{batch,y})^{\alpha} + L_{CE}(V_{batch,\hat{y}}, V_{batch,y})^{\beta}$
    $\theta_{i+1} = \text{ADAM}(L_i, \theta_i)$
    $L_{validation} = \text{Validate}(f(V_{validate,x}; \theta_{i+1}, V_{validate,y})$
until Convergence of $L_{validation}$

Training

To update the weights θ_i of the neural network function f, they used ADAM optimisation. The loss function L (line 16) is a combination of the Tanimoto loss $L_{Tanimoto}$ and the categorical crossentropy L_{CE}, weighted by $\alpha = 0.6$ and $\beta = 0.4$ respectively. The Tanimoto loss is implemented as shown in the following equation.

$$L_{Tanimoto}(\hat{Y}, Y) = 1 - \frac{\hat{Y}Y + smooth}{|\hat{Y}|^2 + |Y|^2 - \hat{Y}Y + smooth}$$

Where $\hat{y} \in \hat{Y}$ denotes the set of predicted voxel-wise annotations and $y \in Y$ denotes the set of ground truth voxel-wise annotations.

Evaluation

They compared the augmentation of their framework to the multidimensional image augmentation method as shown in following Figure 3.

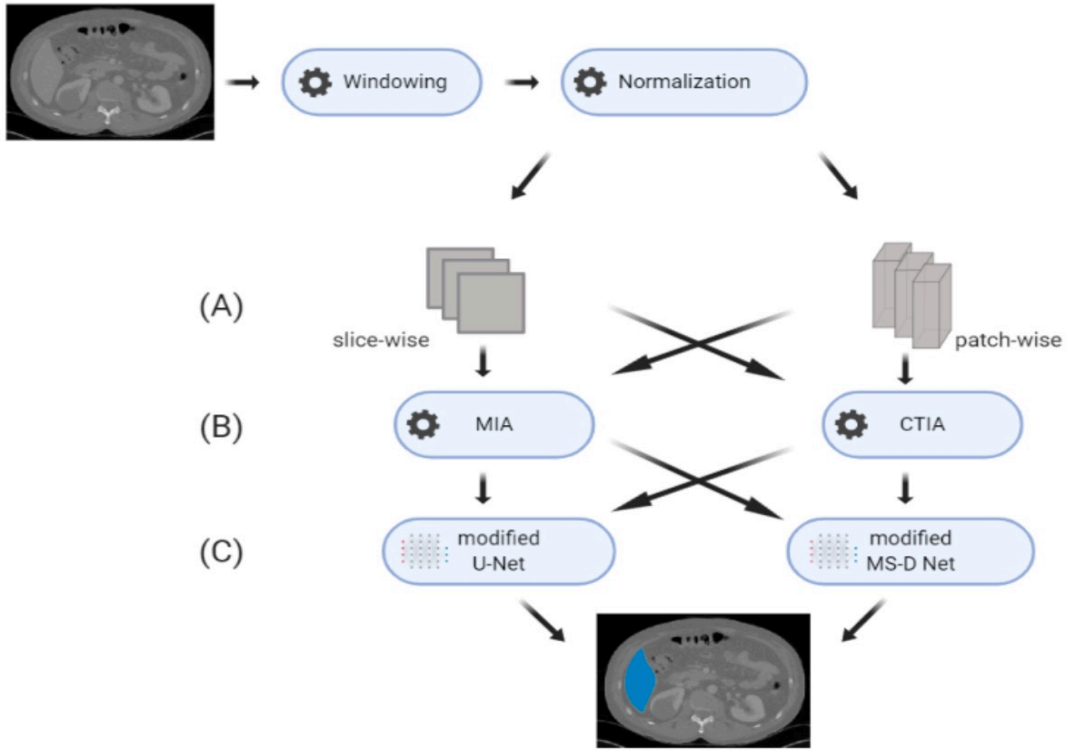

Fig. 3. Overview of the workflows considered in the experiments. We switch three parts of the workflow: (A) Input dimensionality, (B) Augmentation toolkit, (C) Convolutional Neural Network. This figure does not include the experiment with the ensemble model.

Evaluation

Table 1. Results for the kidney tumor segmentation: Total Dice scores are reported (mean ± stdv.) for each segmentation class, the different architectures and input dimensionalities (2D and 3D). Each approach is validated with the multidimensional image augmentation (MIA) for Tensorflow and with our CT-specific image augmentation (CTIA).

		Kidney	Tumor	Total
nnU-Net + MIA	2D	0.962 ± 0.006	0.840 ± 0.013	0.929 ± 0.009
nnU-Net + CTIA	2D	0.961 ± 0.001	0.844 ± 0.007	0.931 ± 0.002
nnU-Net + MIA	3D	0.960 ± 0.012	0.839 ± 0.021	0.929 ± 0.014
nnU-Net + CTIA	3D	0.960 ± 0.002	0.841 ± 0.008	0.925 ± 0.003
MS-D Net + MIA	2D	0.950 ± 0.011	0.774 ± 0.022	0.913 ± 0.014
MS-D Net + CTIA	2D	0.950 ± 0.001	0.779 ± 0.009	0.914 ± 0.003
MS-D Net + MIA	3D	0.947 ± 0.012	0.764 ± 0.024	0.906 ± 0.018
MS-D Net + CTIA	3D	0.948 ± 0.002	0.765 ± 0.009	0.907 ± 0.003
Stacked CNN		0.968 ± 0.001	0.845 ± 0.004	0.947 ± 0.002

$$s_{Dice}(\hat{Y}, Y) = \frac{2\hat{Y}Y}{|\hat{Y}|^2 + |Y|^2}$$

Evaluation

Table 2. Results for liver segmentation: Total Dice score (mean ± stdv.) for the different architectures and input dimensionalities (2D and 3D). Each approach is validated with the multidimensional image augmentation (MIA) for Tensorflow and with our CT-specific image augmentation (CTIA).

		Total
nnU-Net + MIA	2D	0.974 ± 0.031
nnU-Net + CTIA	2D	0.978 ± 0.001
nnU-Net + MIA	3D	0.941 ± 0.027
nnU-Net + CTIA	3D	0.944 ± 0.014
MS-D Net + MIA	2D	0.961 ± 0.032
MS-D Net + CTIA	2D	0.964 ± 0.002
MS-D Net + MIA	3D	0.942 ± 0.037
MS-D Net + CTIA	3D	0.942 ± 0.004
Stacked CNN		0.980 ± 0.001

Evaluation

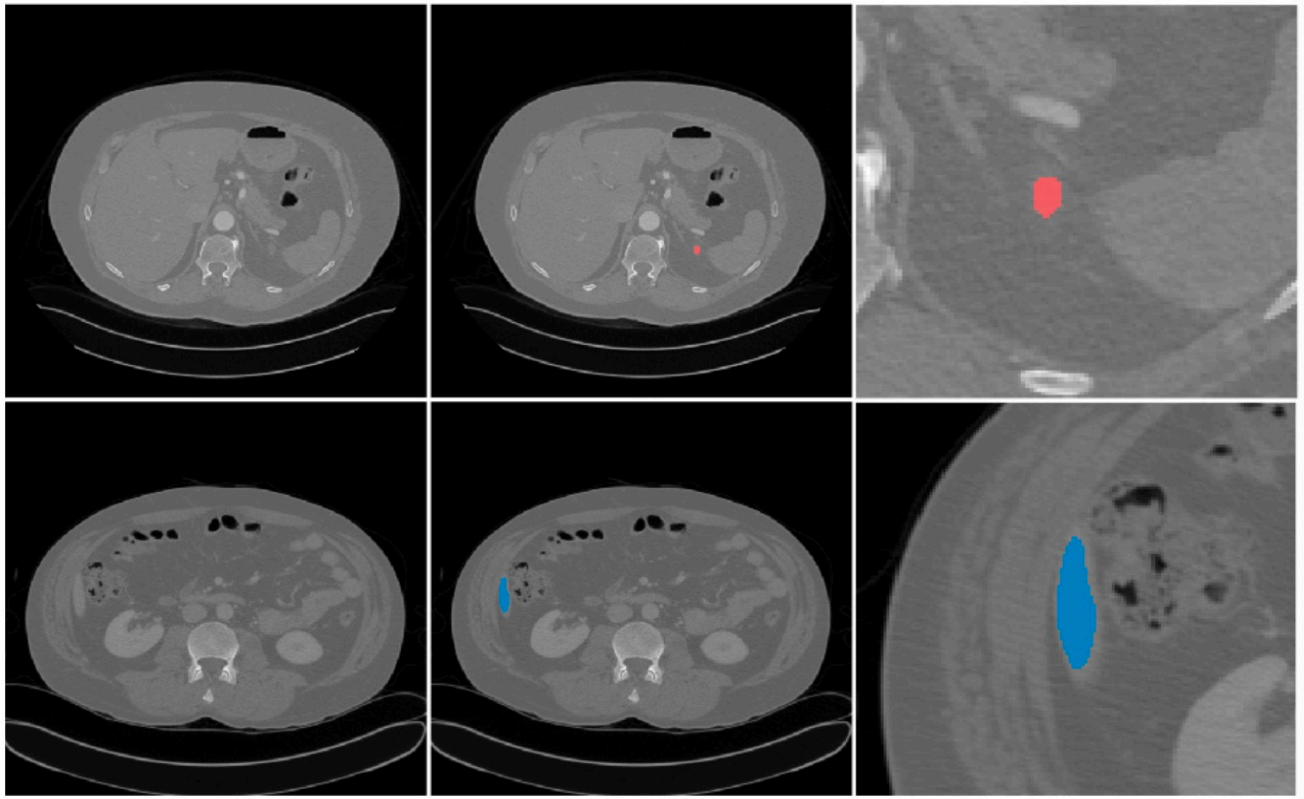

Fig. 4. Examples of challenging 2D segmentation cases for liver segmentation (top) and kidney tumor segmentation (bottom).

For liver segmentation, they found that the MS-D Net generally led to more segmentation errors. However, the MS-D Net errors are typically independent of the segmentation errors of the U-Net approach. In particular, slices with only small regions of interest (shown in Figure 4) pose a challenge.

Conclusion

Their analysis focused on the often neglected influence of preprocessing and data augmentation on segmentation accuracy and stability.

Their results show that 3D spatial information does not necessarily lead to better segmentation performance in particular concerning detailed, small-scale image structures.

Thanks !