Improving Generalizability in X-Ray Segmentation of the femur

Evaluating the Impact of Traditional Data Augmentation Techniques on the generalizability across Datasets

Roland Bockholt

Supervisors: M.A. van den Berg, G. van Tulder Responsible Professor: J.H. Krijthe Examinor: Xucong Zhang

TU Delft

Background

Osteoarthritis is a common age related medical condition in joints. The diagnnosis currently relies on the manual measurement of the Joint Space width (JSW). This procedure is time-consuming and prone to errors, therefore an accurate automated measurement of JSW can improve diagnosis. For this an accurate segmentation of the hip joint components are necessary.

Problem for generalisation can arise from differences in data sets, so called domain differences. They can be caused by factors like differences in the equipment used (e.g., different X-ray machines or settings), lighting conditions, and image resolution.

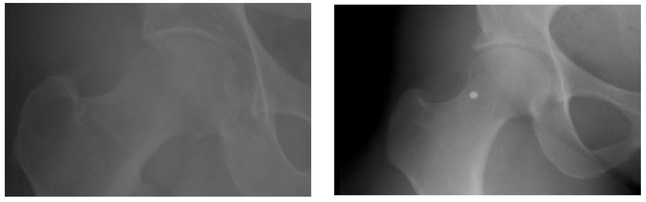

Figure 1. Left: Example from OAI, Right: Example from CHECK

Research Question

Can traditional data augmentation techniques
like flipping, brightness scaling,rotation, etc.
improve the generalisationof a UNET segmentation model
of femurs for X-ray images across different datasets?

Methods

The general approach is as follows:

We train a segmentation model on one data set and evaluate the segmentation accuracy on both data sets. If we see differences in performance, we propose data augmentations. For each data augmentation, we retrain the segmentation model and reevaluate the model

The differences between the X-ray images can arise from different sources like equipment, calibration or patient population. We will try out the following data augmentations.

- **Image Flipping:** Horizontal flipping of images is a common augmentation technique that has demonstrated effectiveness in numerous studies.
- **Random Rotation:** Introducing random rotations to the images can simulate the variability in patient positioning during X-ray capture.
- **Random Blur:** The sharpness of X-ray images can vary due to differences in X-ray machines, machine settings, and other factors.
- **Random Contrast and Brightness Adjustments:** X-ray images can exhibit a wide range of intensities due to differences in exposure settings and patient characteristics.

For evaluation we use the Jaccard index (IoU), an overlap measure and Hausdorff distance, a boundary accuracy measure. We chose a UNET[3] architecture for our segmentation model, a widely use model for medical image semgentation.

Experiment

Data: We use the CHECK[4] data set with 3707 images and 1002 participants and the OAI[2] data set with 12294 images and 4796 participants.

Preprocessing: We crop in each image to the general region of both femurs. We then use the Bonefinder[1] data to create a binary mask and resize both the mask and image to 256x256 pixels.

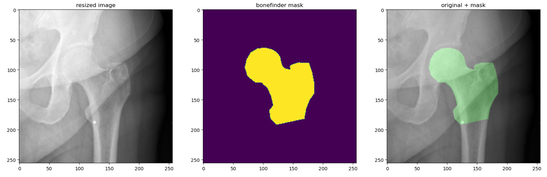

Figure 2. Left:cropped and resized image, Middle: cropped and re- sized binary mask, Right: mask overlaid on the image

Data Augmentations: We apply the data augmentations rotation, blur, brightness adjustment and contrast adjustment in 3 different intensities: high, medium and low.

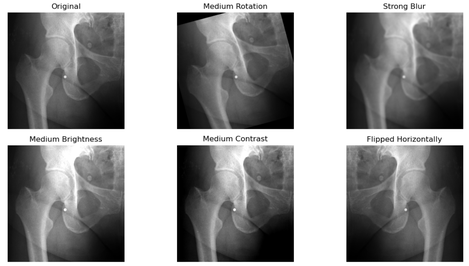

Figure 3. Data augmentation with different intensity levels

Model training: We divide the data set into 70% training 10% validation and 20% testing. We use the negative log liklihood as loss function and train for 30 epochs.

Differences between Data sets

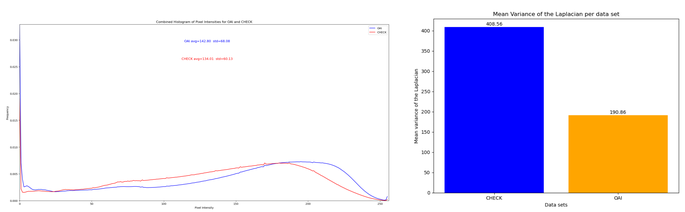

Figure 4. Left: Histograms of OAI and , Right: Example from CHECK

Between the data sets we can measure differences in brightness distribution and blur. We can use the variance of the Laplacian as an approximate measure of blur.

Results

We trained models on both data sets and evaluated them. The model trained on OAI performed much better and had little difference between its dataset and CHECK. CHECK performed worse and had a larger difference in performance.

Trained on	Tested on	Jaccard	Hausdorff
CHECK	CHECK	0.967042	6.83842
CHECK	OAI	0.955002	10.9427
OAI	OAI	0.97302	4.91234
OAI	CHECK	0.97319	5.13456

Table 1. Results of Jaccard and Hausdorff distances for different training and testing data sets.

In this table we see the performance of the model trained on CHECK with various data augmentations and evaluated on OAI.

Augmentation	Jaccard	Std Jaccard	Hausdorff	Std Hausdorff
none	0.9550	0.0456	11.0472	21.6011
Augmentation	**Δ Jaccard**	**Δ Std Jaccard**	**Δ Hausdorff**	**Δ Std Hausdorff**
flipped	+0.0059	-0.0060	-4.9325	-11.9413
rotation high	+0.0046	-0.0067	-5.0488	-12.5593
rotation medium	+0.0059	-0.0075	-4.9767	-12.0267
rotation low	+0.0033	-0.0048	-4.3427	-10.5684
brightness high	+0.0026	-0.0025	-5.0405	-13.9138
brightness medium	+0.0034	-0.0037	-5.1056	-13.5936
brightness low	+0.0028	-0.0061	-4.1101	-9.3770
contrast high	+0.0033	-0.0037	-4.8256	-12.5132
contrast medium	+0.0005	+0.0029	-4.2099	-11.1701
contrast low	+0.0034	-0.0054	-4.8653	-12.5176
blur high	+0.0031	-0.0051	-4.8749	-13.0071
blur medium	+0.0038	-0.0046	-5.1504	-13.7956
blur low	+0.0035	-0.0039	-5.1933	-13.6012

Table 2. Performance metrics for different augmentations in the CHECK dataset tested on OAI

For the generalisability within the CHECK data set we see that data augmentations except for rotation and flipping have only minimal impact on the Jaccard index and lead to a higher standard deviation. All augmentations reduced the Hausdorff distance.

Conclusion

We can conclude that traditional data augmentations effectively enhance the generalizability of a UNET segmentation model to different datasets, improving segmentation accuracy and reducing variability. Among the tested augmentations, random rotations of 15 degrees or more and horizontal flipping were the most effective, followed by medium blur, medium brightness, and low contrast adjustments. The effectiveness of brightness, contrast, and blur adjustments depends on the strength of the modifications and the target dataset. Within the training dataset, all augmentations reduced the Hausdorff distance but had a smaller impact on the Jaccard index, leading to increased variability in segmentation accuracy. Flipping and rotation were the best augmentation method for generalizing within CHECK and to OAI.

References

[1] C. Lindner, S. Thiagarajah, J. M. Wilkinson, The arcOGEN Consortium, G. A. Wallis, and T. F. Cootes. Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE Transactions on Medical Imaging*, 32(8):1462–1472, 2013.

[2] National Institutes of Health. `https://nda.nih.gov/oai`. Accessed: 2024-04-25.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[4] Janet Wesseling, Maarten Boers, Max A Viergever, Wim KHA Hilberdink, Floris PJG Lafeber, Joost Dekker, and Johannes WJ Bijlsma. Cohort profile: Cohort hip and cohort knee (check) study. *International Journal of Epidemiology*, 45(1):36–44, 2016.