textheight has been altered.
paperwidth has been altered.
textwidth has been altered.

The page layout violates the ICML style.

Please do not change the page layout, or include packages like geometry, savetrees, or fullpage, which change it for you.

We're not able to reliably undo arbitrary changes to the style. Please remove the offending package(s), or layout-changing commands and try again.

Spine vertebrae segmentation in 3D CT scan images

Jan Alexander [1] **Joris Roels** [2] **Bert Vankeirsbilck** [3]

Abstract

Medical professionals use MRI or CT scans as essential components for diagnosis and planning of procedures. There is a trend towards machine vision to support medical professionals to interpret these images. This research investigates techniques to reduce the dataset labelling cost for this application by working with point annotation instead of full annotation. Based on publicly available datasets this work demonstrate two new loss components and a combination technique of different model results to generate pseudo masks. As a final result, one can obtain 72 % of the inversely weighted dice score performance of a fully annotated model at an estimated 12 % of the labelling cost.

1. Objective & Motivation

The use of radiological images is a crucial element in modern medical practice. MRI or CT scans are essential components for pre-operative and post-operative diagnosis, following the course of medical conditions and the planning of medical procedures.

Machine vision - deep learning in general - tends to be very *data-hungry*, requiring large, labelled datasets. Acquiring these datasets and the corresponding labels is time-consuming and expensive. Maximisation of the return of a given data and labelling budget is important. The use of weak labels (sometimes called *hints*) is one approach to attempt this. This approach aims to train a model capable of inferring more informative results than the information level explicitly available in the labelling.

This project presents a model for the automated segmentation of the five lumbar vertebrae of the human spine based on point level annotated medical scans, which is faster and cheaper than providing a complete label mask (estimated at 12% of cost(Bearman et al., 2016)). The labels only contain the true class of a handful of voxels.

2. Data sets and data preprocessing

All datasets used in this work are publicly available. Both CT and MRI scans are used. In 86 of these scans, complete volume masks of the vertebrae are available. In 22 volumes, only semantic labels are available. The complete dataset contains various patients with different pathologies, such as scoliosis and crushed or wedged vertebrae.

[1]Master Statistical Data Analysis [2]UGent, VIB [3]UGent, IMEC. Correspondence to: Jan Alexander <jan.alexander@ugent.be>.

Abstract master thesis Jan Alexander

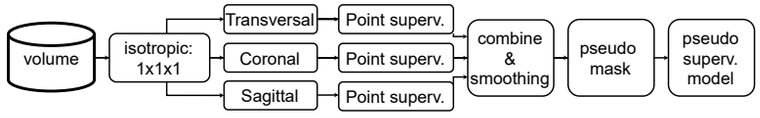

Figure 1. Model training approach

Data preprocessing starts with homogenising the scan resolution by resampling the image on an $1mm \times 1mm \times 1mm$ grid. Next, the image is sliced along one of the three principal axes. The contrast of the 2D image slices is first enhanced with the CLAHE algorithm and cropped (or padded, if needed) to form 352×352 slices. All models are built with this image size, sufficient to contain all 5 lumbar vertebrae $L1$ to $L5$ in one image.

3. Methodology

The performances of different models are compared based on the class-weighted dice score. This metric takes into account both the model precision and recall and the class imbalance. The performance of a fully supervised model is taken as benchmark.

3.1. Weakly supervised models

The model backbond is the VGG16-FCN8 network, pre-trained on a large classification dataset. The model estimates 6 segmentation classes (5 lumbar vertebrae and the background class). By training three different weakly supervised models on sets of 2D images sliced along the 3 main volume dimensions, three sets of segmentation masks are obtained. The combination of these different segmentation masks is used as an *pseudo* label set to train a fully supervised model on anatomic plane.

3.1.1. Loss function

The loss combines supervised and unsupervised loss components. The model loss to train three point-supervised models in the first step of the procedure consists of 4 components: the point loss $\mathcal{L}_P$ and the consistency loss $\mathcal{L}_C$ were defined in (Laradji et al., 2021) by I. Laradji, while this work introduced the prior extend and separation loss components $\mathcal{L}_E$ and $\mathcal{L}_S$.

The weighted cross-entropy loss is optimised for the fully supervised reference model, a classic choice for this problem. It is also the point loss $\mathcal{L}_P$ component of the weakly supervised model. Then it is only evaluated on the set of available point labels $\mathcal{I}_i$.

The unsupervised rotation consistency loss $\mathcal{L}_C$ imposes that the model output f_θ should be consistent for a transformation t_k of the input image. The chosen transformations are image rotations over $0°, 90°, 180°$ or $270°$, combined with an image flip.

$$\mathcal{L}_C(X_i) = \sum_{p \in \mathcal{P}_i} \left| t_k \left[f_\theta(X_i) \right]_p - f_\theta \left(t_k[X_i] \right)_p \right| \tag{1}$$

Due to the low volume of labelled voxels, the the model lacks the incentive to output differentiating expressions of the output channels $\vec{z}_i$. $\mathcal{L}_S$, the separation loss, forces the model to ensure sigmoid $\mathbf{S}$ of channel expression n is different from channel m.

$$\mathcal{L}_S(X_i) = -\sum_{\vec{p}} \sum_{m \in \mathcal{K}} \sum_{n \in \mathcal{K}, n > m} \mathbf{S}(z_i[m]) - \mathbf{S}(z_i[n]) \tag{2}$$

Finally, $\mathcal{L}_E$, the maximal extend supervised loss term, takes into account that a lumbar vertebra has a limited size ($r = 110mm$). The distance field $\mathbf{d}$ from the annotation points $\vec{p}$ is converted to a semi-mask for each class k. At distance r from an annotation point ($\mathbf{m}_k = 0$), the model output should not indicate output class k. Else, the output class is unknown. The loss function is the binary cross-entropy between $\mathbf{m}_k$ and the sigmoid of the kth channel of the logits z_i with weight vector $\{1, 0\}$.

$$\mathbf{m}_k(\vec{q}) = \mathbf{I}\left(\left[- \max_{\vec{p}: \mathcal{Y}_i(\vec{p}) = k} ||\vec{q} - \vec{p}|| + r \right] > 0 \right) \tag{3}$$

$$\mathcal{L}_E(X_i) = \sum_{k \in \mathcal{K}} \sum_{\vec{q} \in X_i} (1 - \mathbf{m}_k(\vec{q})) \log(\mathbf{S}(z_i(\vec{q})_k)) \tag{4}$$

3.1.2. Model metric

The model performances are compared based on the dice metric weighted by the inverse of the class occurrence.

$$dice_{wi} = \frac{\sum_{m=0}^{k-1} \left[dice_m A_m^{-1} \right]}{\sum_{m=0}^{k-1} A_m^{-1}} \tag{5}$$

where A_m indicates the number of observations with true class m.

3.1.3. Model result combination

Combining the results of the three models trained on the three geometric axes (transverse, frontal & sagittal) is a pseudo-mask of higher quality than the results of the individual models. After morphological smoothing, the pseudo mask is used to train the final model on sagittal slices.

4. Results

The reference model was found to have a test performance of $dice_{wi} = 0,76$. This model is trained on the same dataset as the weakly supervised model, but with full label masks.

To produce the pseudo masks, first 3 models are constructed, each trained on a different set of volume slices: transverse, coronal or sagittal. The performance of each model is low compared to the reference model.

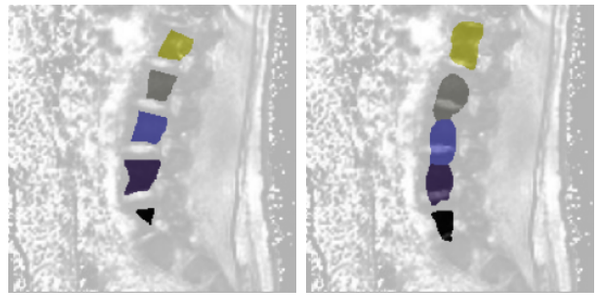

Figure 2. Ground truth (left) vs. weakly supervised method result (right)

Yet, combining these models allows to obtain a pseudo mask which is already a segmentation mask with a higher performance than the segmentation masks of the combined models. The transverse slices do not provide sufficient context to identify the individual lumbar vertebrae, this model does not distinguish between $L1$ to $L5$. Thus, the numerically higher inversely weighted dice score in the table is misleading. Finally, this pseudo mask is used to train the resulting model. This last model has the exact same architecture as the reference model, therefor the inference time is identical , about 12s per volume. Its performance is an improvement of the pseudo mask performance.

Separate		Pseudo mask	Procedure result	Reference
Transverse	0.51	0.48	0.55	0.76
Coronal	0.41			
Sagittal	0.34			

Table 1. Inversely weighted dice score comparison

5. Conclusion

The work in (Laradji et al., 2021) is extended with two new loss functions and a combination algorithm to combine models trained on different slice stacks from the same volume into pseudo masks to train a final model. This procedure allows the model to approximate the performance of a fully supervised model at 12% of the labelling cost.

References

Bearman, A., Russakovsky, O., Ferrari, V., and Fei-Fei, L. What's the point: Semantic segmentation with point supervision. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9911 LNCS, pp. 549–565. Springer, Cham, jun 2016. ISBN 9783319464770. doi: 10.1007/978-3-319-46478-7_34. URL http://arxiv.org/abs/1506.02106.

Laradji, I., Rodriguez, P., Mañas, O., Lensink, K., Law, M., Kurzman, L., Parker, W., Vazquez, D., and Nowrouzezahrai, D. A Weakly Supervised Consistency-based Learning Method for COVID-19 Segmentation in CT Images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, jan 2021. URL http://arxiv.org/abs/2007.02180.