Submission and Formatting Instructions for International Conference on Machine Learning (ICML 2021)

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

Abstract

Medical professionals use mri or ct scans as essential components for medical diagnosis, following the course of medical conditions and the planning of medical procedures. There is a trend towards machine vision to support medical professionals interpreting and using these images. Building these applications requires expensive labelled datasets. This research investigates techniques to reduce the dataset labelling cost by working with point annotation instead of full annotation. Experiments are conducted on publicly available datasets and demonstrate two new loss components and a combination technique of different model results to generate pseudo masks. As a final result, this work demonstrates that one can obtain 72 % of the performance of a fully annotated model at an estimated 12 % of the labelling cost.

1. Thesis 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. Automated interpretation of medical images can mean a gain in efficiency.

Machine vision - deep learning in general - tends to be very *data-hungry*. Constructing a new model requires 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 through is a goal shared by all ml practitioners. The use of weak labels, or 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 lumbar vertebrae of the human spine based on point level annotated medical scans. Point level annotation is faster and cheaper than providing a complete label mask (estimated at 12% of cost(**?**)), this technique provides a cost-benefit. The labels only contain the true class of a mere handful of voxels. This is a weak label to classify all voxels.

2. Data sets and data preprocessing

All datasets used in this work are publicly available (all datasets are listed on page **??**). These datasets contain both ct and mri scans. In 86 of these scans, complete volume masks of the vertebrae are available. In 20 volumes, only semantic labels are available. For 125 volumes, point level annotation is available.

The complete dataset of 231 patients consists of 112 women and 99 men. Of 23 people, no gender information is available. Since a medical professional does not order a medical scan unless there is a suspicion of a medical condition, the dataset contains various patients with different pathologies, such as patients with scoliosis and with crushed and wedged vertebrae.

Different datasets vary in data formats and different scan resolutions. 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. Then the images are cropped (or padded, if needed) to form $352px \times 352px$ slices. All models are built with this image size, sufficient to contain all 5 lumbar vertebrae L_1 to L_5 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 as well as the class imbalance.

For 86 scans, full annotation masks are available. As a performance benchmark, the performance of a fully supervised

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

model trained on these images ($Dice_w = 0,76$) is taken.

3.1. Weakly supervised models

The model backbond is the VGG16-FCN8 network, pretrained 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 one volumetric dimension.

3.1.1. Loss function

To train the weakly supervised network, several loss components, both supervised and unsupervised, are combined. The model loss to train three point-supervised models in the first step of the procedure presented in this work consistents of 4 components: the point loss $\mathcal{L}_P$ and the consistency loss $\mathcal{L}_C$ were defined in (**?**) by I. Laradji, while this work introduced the prior extend and separation loss components $\mathcal{L}_E$ and $\mathcal{L}_S$ are introduced in this work.

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 function combines the six network output channels with a softmax function σ, after which the negative log-loss function is calculated, weighted with factors w.

$$\mathcal{L}_P(X_i) = -\sum_{\vec{p}\in\mathcal{I}_i} w_{\mathcal{Y}_i(\vec{p})} . \log\left[\sigma_{\mathcal{Y}_i(\vec{p})}\left(z_i(\vec{p})\right)\right] \tag{1}$$

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. In this work, the chosen transformations are image rotations over $0°, 90°, 180°$ or $270°$, combined with an image flip.

$$\mathcal{L}_C(X_i) = \sum_{p\in\mathcal{P}_i}\left|t_k\left[f_\theta(X_i)\right]_p - f_\theta\left(t_k[X_i]\right)_p\right| \tag{2}$$

The second unsupervised loss term is the separation loss term. Due to the low volume of labelled voxels, the the model lacks the incentive to output differentiating expressions of the output channels $\vec{z}_i$. $\mathcal{L}_S$ forces the model to do this.

$$\mathcal{L}_S(X_i) = -\sum_{\vec{p}}\sum_{m\in\mathcal{K}}\sum_{n\in\mathcal{K},n>m}\mathbf{S}(z_i[m]) - \mathbf{S}(z_i[n]) \tag{3}$$

Finally, $\mathcal{L}_E$, the maximal extend supervised loss term, takes into account that a lumbar vertebra has a limited size ($r = 110mm$). The Euclidian distance field $\mathbf{d}$ from the annotation point is converted to a semi-mask for each class k:

$$\begin{aligned}\mathbf{d}_k(\vec{q}) &= \max_{\vec{p}:\mathcal{Y}_i(\vec{p})=k} \|\vec{q}-\vec{p}\| && (4)\\ \mathbf{m}_k(\vec{q}) &= \mathbf{I}\left((-\mathbf{d}(\vec{q})+r)>0\right) && (5)\end{aligned}$$

Now, $\mathbf{m}$ is 1 only for positions closer than distance r from the annotation points for class k. Where $\mathbf{m}_k = 0$, the model output should not indicate output class k. Where $\mathbf{m}_k = 1$, 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 $\{1, 0\}$.

$$\mathcal{L}_E(X_i) = \sum_{k\in\mathcal{K}}\sum_{\vec{q}\in X_i}(1-\mathbf{m}_k(\vec{q}))\log(\mathbf{S}(z_i(\vec{q})_k)) \tag{6}$$

3.1.2. Model metric

The model performances are compared based on the inversely weighted dice metric, defined as:

$$Dice_{wi} = \frac{\sum_{i=0}^{k-1}\left[Dice_i\left(\sum_{j=0}^{k-1}a_{i,j}\right)^{-1}\right]}{\sum_{i=0}^{k-1}\left(\sum_{j=0}^{k-1}a_{i,j}\right)^{-1}} \tag{7}$$

where $a_{i,j}$ indicates the count of observations with true class i, classified as class j.

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 performance of $dice_{wi} = 0,76$, evaluated on the test set. This model is trained on the same dataset as the weakly supervised model, but with full label masks. Thus, this is a reference that could be described as a *classic* approach. This model is trained based on the cross entropy loss.

The first step in producing the pseudo masks is the construction of 3 models, each trained on a different set of volume slices. Each volume can be sliced to obtain a stack of transverse, coronal or sagittal slices. The performance of each model is low, compared to the reference model. Yet, combining these models allows to obtain a pseudo mask. This pseudo mask is already a segmentation mask wich shows to have a higher performance than the segmentation masks of the combined models. Since the transverse slices do not provide sufficent context to identify the individual lumbar vertebrae, this model only segments the vertebrae semantically. It does not distinguish between $L1$ to $L5$. For this reason, the numerically higher $dice_{wi}$ score in the table is misleading.

Separate models		Combined model
Transverse	0.51	3*0.48
Coronal	0.41	
Sagittal	0.34	

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 inferrence time is identical (about 12s per volume). Its performance is an improvement of the pseudo mask performance. A performance of $dice_{wi} = 0.552$ could be obtained.

5. Conclusion

The work in (**?**) is extended with two new loss functions and a combination algorithm, allowing the model to approximate the performance of a fully supervised model at 12% of the labelling cost.