Presentation notes for poster presentation

Jan Alexander
31/03/2021

General intro

Jan Alexander, master student Master in Statistical Data analysis.

Master thesis subject: Machine vision for medical applications. An instance segmentation model for the lumbar vertebrae of the human spine in CT (*computed tomography*) images, based on weakly supervised data.

The lumbar vertebrae are the 5 spinal vertebrae in the lower back.

Instance segmentation indicates that each voxel is classified as one of the 5 lumbar vertebrae, indicated by L1 to L5, or as *background* everything that is not a vertebra is considered background in this problem. The objective is to obtain full segmentation masks for all 5 vertebrae.

Problem motivation:

First I would like to motivate the two subjects of my master thesis subject:

CT scan segmentation

Segmentation of the human spine from CT scan images is a useful support for spine pathology diagnosis and as a support for planning and performing surgical interventions on the spine. During surgical intervention, frequent imaging can be necessary. In this situation, it is useful to support the surgeon maximally in interpreting these images.

The illustration shows an artists impression of the start of a *micro-discectomy* to treat a *herniated spinal disc*. The first dilator for the *laparoscopy* is already inserted.

Weakly supervised learning

The classical approach for training a deep learning network requires a dataset of fully labelled data. In the case one wants to train an instance segmentation mask, this requires a labelled dataset with instance masks for all classes.

This means that an expert has to delineate annotation masks for all slices of all images in the dataset to generate per-pixel labels. Delineation of 1 CT-scan with 250 slices requires a budget of 400 minutes. This cost becomes prohibitive very fast.

One might therefore question whether this approach is the most efficient. The idea behind weak supervision is to leverage cheaper, less informative labels for the task at hand.

There are various weak supervision methods. Every label type that is *less informative* than the desired result of the model is considered weak supervision. Literature regarding weakly supervised deep learning

for segmentation has mostly focussed on optical camera images such as the *COCO* dataset or *PASCAL VOC 2012*. Several approaches have been investigated by various authors ranging:

- Image level labels: only the object classes present in the picture are provided.
- bounding box labels: When used to train a model that outputs bounding boxes, this is a strong label, but these can be leveraged to train a segmentation network.
- Point labels or squiggles: Very fast to mark. This is the annotation type used in my thesis.

According to Bearman, point level labels are 10 times less time consuming than full delineations for the PASCAL VOC 2012 dataset. For medical data, I would argue one could expect a comparable reduction in cost.

Previous work

Datasets

This work will be performed based on existing, publically available datasets. I managed to find 5 different datasets, all containing annotations for all 5 lumbar vertebrae.

The combination of these 5 datasets results in 359 volumes, scans of 242 different patients of both genders.
The average patient in this set is in his or her 50s.

The largest dataset, by Glocker and team at the University of Washington, is labelled with point labels. The other datasets, all four considerably smaller, are labelled with per pixel labels.

Architectures

As pre-processing, the orientation of the scans is made uniform such that the x and y-axis form the transverse plane and the z-axis is the craniocaudal axis. The scans are re-sampled on an isotropic grid of 1 mm x 1 mm x 1 mm. The encoding is converted from Hounsfield units to floating numbers in the interval [0, 1]. Hounsfield units are a measure of radiodensity commonly used in CT scans.

Fully supervised model

As a baseline, a fully supervised reference model will be used. The model on the illustration is first published by Nicolas Lessmann. It is reported to obtain a segmentation Dice score of 94% and a classification score, accuracy, for anatomic identification of 93%.

The system input is a combination of a scan patch of 128 x 128 x 128 voxels - so 128 mm in all three directions - and the instance memory, a volume with the same size that contains the masks of the vertebrae that have already been segmented. The objective of the system is to estimate three outputs:

- The segmentation mask for the next vertebra
- The classification of the next vertebra --> linear loss
- The completeness score of this vertebra --> cross-entropy

This is optimized based on an interesting loss function. The segmentation loss is the sum of the valse classifications (soft False positives and soft false negatives) but the weight of each voxel in this sum is

relative to its position in the vertebra. The loss function gives increasing weight to the voxels at the border of the vertebra volume to increase the separation.

Weakly supervised approach

The weakly supervised modelling approach starts from a model published by Issam Laradji for segmentation of regions of *pulmonary opacification* caused by Covid-19. For this approach, the volume masks are reduced to point annotations by sampling a point from the given mask as an annotation.

The network concept is illustrated on the poster. There are various options for the CNN encoder part. At the moment, I started the experiments with a VGG16 network, pre-trained on the *Imagenet* dataset (1.2 mil images)

This approach is based on the so-called consistency score. The idea is that the resulting segmentation mask should be consistent under transformation such as rotation.

The loss function is thus a combination of the point loss: *cross-entropy* loss on the annotation points and the *consistency loss*. The *consistency loss* measures the difference in the segmentation mask estimated on the image with the segmentation mask on the transformed image.

Approach

Both datasets with strong and datasets with weak annotation are available. The objective is to obtain the best dice score and classification accuracy.

Given the available datasets, I plan on following the following approach:

Train only on the fully supervised datasets

Train - Validation - Test split: 60 - 20 - 20, taking into account the grouping by patient. Scans of the same patient should not end up in different splits.

First step: Train the reference network on the fully supervised datasets. From the strong labels, a weak label can be sampled. From the complete mask, one can sample one or several points. Train several experiments of the weakly supervised model on the same set.

Some components that could be varied in the different experiments are:

- Different backbones: VGG11, VGG16, VGG19, ResNet, UNet
- Different loss functions: Consistency loss, WiSe (technique based on class activation maps)

The best experiment can, possibly, be improved with the large volume of weakly supervised scans.

Try to improve the result with the weakly annotated datasets:

Add the UW data to the training set.

The idea is to compare first the performance between fully supervised and weakly supervised model training and then investigate how much weakly supervised sets can contribute.

Additional idea

Recently, my promotors and I have discussed the idea of *Active Learning*. This would mean that the segmentation is performed in several iterations while prompting the input of the user. The uncertainty of the model output could be quantified to prompt the user for additional annotation in specific regions of specific slices where the highest gain in performance would be expected. This is a very interesting path that corresponds to the high-level objective of this Master thesis: **Increasing the efficiency of data labelling for medical image segmentation**