Article

Investigation of loss functions for improving deep segmentation of abdominal organs from MRI

Pedro Furtado [1]

[1] University of Coimbra, DEI/CISUC; pnf@dei.uc.pt
* Correspondence: pnf@dei.uc.pt

Abstract: Segmentation of Magnetic Resonance Images (MRI) of abdominal organs is useful for analysis prior to surgical procedures and for further processing. Deep Learning (DL) has become the standard, researchers have proposed improvements that include multiple views, ensembles and voting. Loss function alternatives, while being crucial to guide automated learning, have not been compared in detail. In this work we analyze limitations of popular metrics and their use as loss, study alternative loss variations based on those and other modifications and search for the best approach. An experimental setup was necessary to assess the alternatives. Results for the top scoring network and top scoring loss show improvements between 2 and 11 percentage points (pp) in Jaccard Index (JI), depending on organ and patient (sequence), for a total of 22 pp over 4 organs, all this being obtained just by choosing the best performing loss function instead of cross-entropy or dice. Our results apply directly to MRI of abdominal organs, with important practical implications for other architectures, as they can be applied easily to any of them. They also show the worth of variants of loss function and loss tuning, with future work needed to generalize and test in other contexts.

Keywords: Computers in Medicine; Segmentation; Machine Learning; Deep Learning; MRI.

1. Introduction

Magnetic Resonance Imaging (MRI) is a medical imaging technique that forms images of the anatomy and physiological processes of some part of the body. MRI is obtained using strong magnetic fields, gradients and radio waves. For acquisition, radiofrequent pulses are triggered that will make the protons spin change their orientations, and receiver coils of the MRI equipment capture corresponding magnetic signal changes. Further processing of those signals results in MRI images that can then be used to gain insight and further knowledge concerning the structures and organs. In that context, precise segmentation is useful for advanced computer-aided analysis, measurements and visualizations related to medical procedures.

In recent years, Deep learning (DL) has quickly revolutionized the quality and flexibility of segmentation. The segmentation network is an evolution of the (deep) convolution neural network (DCNN) used for classification. The classification DCNN learns to classify images, and its structure is a sequence of encoder convolution stages that extract and compress features from the image directly into feature vectors, followed by a fully-connected neural network that classifies the image based on that feature vector. The segmentation network replaces the fully connected part by a decoder that restores the full image-size using de-convolution layers, and outputs segment labels for each pixel. The capacity to learn automatically in DL networks comes from a large number of iterations adjusting thousands of weights of convolution and de-convolution filters automatically based on back-propagation of the loss (a.k.a. error), a fundamental measure of the distance between the current quality of segmentation of training images and the groundtruth segmentation of those training images. Computation of the loss is a crucial step in this procedure: a loss function that fails to reveal deficiencies in segmentation of specific organs will not learn to segment those organs well. That is the reason why validation loss can be very low (e.g. below 1%) during training and yet the resulting segmentation network may still fail to segment some organs and some slices with the quality that would be expected just by looking at the loss. For instance, figure 1 shows segmentations of two test slices (the groundtruth segments are the left images shown on black background), with the one in (b) in particular being far from perfect in spite of the fact that final validation loss reported a very low cross entropy loss (0.3%), and both class re-balancing and data augmentation were used during training.

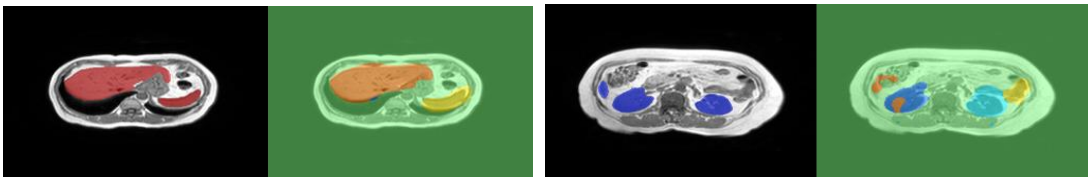


Figure 1. Example MRI segmentation of independent test images using DeepLabv3 segmentation network. The left of each image is the groundtruth on a black background, the right is the segmentation: (a) is a slice showing the liver and spleen; (b) is another slice showing the kidneys and a small extremity of the liver

There is no magic solution to the limitations of metrics used as loss function, and there already exist sufficient metrics that characterize segmentation quality. Our focus is not on devising completely new elaborate loss functions for the sake of novelty, but rather to engage in a careful consideration of metrics and experimental evaluation to help discover the best performing alternative. Intuitively, the best metric and loss function would be one that would best reveal deficiencies segmenting individual organs, but it is not clear how best to translate this into a single loss value. Besides comparing loss functions and variations in loss formula, such as different weights to false positives and negatives, we compare the alternative of not considering the background at all in the loss function, and also the alternative of simply replacing multiclass by uniclass segmentations. Based on the study of the alternatives we were able to improve 22 percentage points (pp) on the sum of pp improvement over the 4 organs tested.

Related work

Next, we briefly review works on segmentation of MRI and CT scans, plus works modifying loss and also works discussing metrics. Most works on segmentation of MRI and CT actually propose modifications of architectural details, only a few test a different loss function, and none compares the effect of loss function variations that we compare in this work. Prior to the use of deep learning (DL), segmentation would most frequently be based multi-atlas approaches. [1] uses 3D models of the liver and probability maps, [2] is based on histograms to segment the liver, followed by active contours for refinement, [3] applies watershed together with active contours. Then deep learning-based segmentation revolutionized the field. Zhou [4] achieved top scores using a fully convolutional networks (FCN) by taking 3-D CT images and applying a majority voting scheme on the output of segmentation of 2D slices taken from different image orientations. [5] applied a similar approach to abdomen segmentation from MRI sequences, scoring (dice similarity coefficient=DSC) 0.93, 0.73, 0.78, 0.91, 0.56 for spleen, left kidney, right kidney, liver and stomach. Larsson [6] proposed SeepSeg which segments abdominal organs using 3 steps, regions localization by multi-atlas approach, CNN for pixel binary classification and post-process using thresholding to remove positive samples except those of the largest connected region. The proposed approach scored (Jaccard Index=JI) 0.9; 0.87; 0.76; 0.84 for liver, spleen, right and left kidney. [7] proposed multi-slice 2D neural network designed in a way that considers information of subsequent slices, plus augmented data and multiview training. Groza [8] presents an ensemble of DL networks with voting to achieve improved segmentation scores for MRI scans, proposing five networks and a voting-based ensemble mechanism. In another work, [9] tests different architectures (the basic U-Net, a deeper U-Net with VGG-19 layers, a cascade of two networks and a Generative Adversarial Network (cGAN). Only a few works modified the loss function to try to improve segmentation scores, and usually in slightly different contexts: [10]

proposed improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and "direct" loss function. They propose a Jaccard Loss (JACLoss) for training the neural network. As explained by the authors, "to optimize the Jackard Index (JI) (a main segmentation metric) directly in network training makes the learning and inference procedures consistent. It empirically works better than the cross-entropy loss or the class-balanced cross-entropy loss when segmenting small objects, such as pancreas in CT/MRI images". [9] also replaced cross-entropy but by the dice function to better deal with class imbalance. In a different context, [13] also investigated a modified loss function that is useful in our work. These works touch the surface of the relevance of the loss function in the quality of segmentation, but they do not study and compare the various alternatives that we do, and we arrive at best performing solution that extends their results. Investigating limitations of metrics is also important for a thorough investigation of loss, since loss is a metric itself. The authors in [12] identified a relevant limitation of metrics that we also discuss in this work. Regarding evaluation of segmentation of eye-fundus images, [12] states that "many scores are artificially high simply because the background is huge and hence the term TN (true negatives) is also huge, making specificity, ROC and AUC inviable as scores". Our analysis of metrics includes this problem, and the identification and analysis of those limitations is very important to define a suitable loss function.

2. Materials and Methods

In this section we first analyze metrics and their limitations, proposing loss function variations and alternatives based on that analysis. Then we describe an experimental setup that we use to evaluate the quality of segmentation using the varied loss alternatives.

2.1. Discussing metrics and loss

Both segmentation evaluation metrics and loss are expected to quantify the difference (error) between the groundtruth (GND), representing a correct segmentation of the image, and the segmentation output (SEG). The loss is f(SEG, GND), a quantity between 0 and 1, and the quality of segmentation is (quality=1-loss). SEG and GND are labelmaps, i.e. each position (pixel) in the labelmap is a class label. In most bibliography, metrics are defined considering a binary classification problem that classifies into two classes: positive (P), with the meaning "is", and negative (N), with the meaning "is not". The quantities TP, TN, FP and FN correspond to the number of pixels that are true positives, true negatives, false positives and false negatives, respectively. Given those quantities, some of the most frequent metrics are:

Accuracy (ac) = (TP+TN)//TP+TN+FP+FN); (1)
Sensitivity (se) = recall= True Positive Rate (TPR)=TP/(TP+FN) (2)
Specificity (sp) = TN/(TN+FP) (3)
Precision (p) = TP/(TP+FP) (4)
False Positive Rate (fpr) = FP/(FP+TN) (5)
ROC, a plot of TPR vs FPR, and AUC, the area under the curve of ROC (6)
IoU = JI = TP/(TP+FN+FP) (7)
Dice (dice) = DSC = 2TP/(2TP+FP+FN)=2JI/(JI+1), which is highly correlated with JI (8)

In multiclass problems we can apply the same formulas, but considering the following quantities: a TP pixel is a pixel that belongs to one class c different from background in groundtruth and also in the segmentation; a TN pixel is a pixel that belongs to background in both groundtruth and segmentation; an FP pixel is a pixel that belongs to background in groundtruth but is classified as some other class c in segmentation; an FN pixel is a pixel that belongs to some class c different from background in groundtruth but is then classified as background;

The following three observations are important reasons why the metrics defined in equations (1) to (8) can fail to evaluate segmentation correctly in many medical imaging contexts:

a) the number TP is always huge in all metrics, because TP of background pixels is huge. As a consequence, all metrics (1) to (8) report high scores regardless of the actual quality of segmentation of individual organs if evaluated over all pixels;
b) TN is also huge because it includes a huge number of background pixels that are well classified. It means that specificity (SP), FPR, ROC and AUC do not evaluate the quality of segmentation of organs well;
c) Sensitivity (a.k.a recall or TPR), although useful because it quantifies the fraction of organ pixels classified correctly as such, fails to capture very important possible deficiencies, because it does not include FP (background classified as organ) in the formula, a frequent occurrence.

The problems identified in a) and b) are a consequence not only of class imbalance, but most importantly of the fact that background pixels are much easier to segment (score much higher) than organ pixels because they are more constant across most slices and patients (since they include all pixels “framing” the image except the organs). The issue identified in a) means that it is necessary to use metrics that evaluate each class separately instead of computing them over all pixels, requiring modifications to how equations (1) to (8) were defined above. Additionally, since b) and c) discard many metrics that are inappropriate, the metrics that are left for use are JI, DSC and precision (which should be used together with recall). Given the observations in a), these need to be evaluated separately for each class. That means each quantity TP, TN, FP and FN must be replaced by TPc, TNc, FPc and FNc respectively, where c is a class, and the metrics should be obtained and reported separately for each class c.

But while we can report a different value of JI or DSC for each class when evaluating segmentation quality, the loss function needs to output a single value to be used as delta in backpropagation learning. Therefore, the final loss must be averaged over the loss of each class. This solution is still not perfect because the loss of class “background” is in practice always almost zero (due to a) and b)), contributing to push the average loss down even if specific organs are very well segmented. Based on these observations, we define the loss functions and variations to consider in the next sub-section.

2.2. Defining metrics for use as loss function

Based on the previous analysis we define a set of loss functions besides cross entropy and a set of variations and alternatives that may contribute to improve the quality of the learning process. Of course we also include the standard cross entropy as one of the options to compare to.

Cross entropy (crossE, the default to compare with): cross-entropy is well-known and the default loss function. Given the set of probabilities p of a single pixel of the segmentation output to be of each possible class, and the real probabilities (one-hot encoding of the class), cross entropy measures dissimilarity between p and q. If ti and si are the groundtruth and the CNN score of each pixel for each class i respectively,

$$crossE = -\sum_{i}^{C} t_i \log(s_i)$$

By applying a class frequency inverse weight to the value for each pixel we obtain class-weighted cross-entropy, which is the variant we use and denote as “crossE”.

Intersect over the Union (IoU): IoU is a convenient measure of the degree of overlap or match between segmentation-obtained regions and the corresponding groundtruth regions. Given the number of true positives (TP), false positives (FP) and false negatives (FN) in the classification of pixels, loss is (1-IoU),

$$IoU(loss) = 1 - IoU = 1 - \frac{TP}{TP + FP + FN}$$

But since this IoU averages over all pixels and we identified the problem with that measurement, IoU averaged over the classes is used instead,

$$IoU(loss) = 1 - \frac{\sum_{i=1}^{C}(IoU_i)}{C}, IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}$$

Dice (dice): The dice or Dice Similarity Coefficient (DSC) is a metric that is highly correlated and can be obtained from IoU directly. The loss formula for the dice is:

$$dice(loss) = 1 - DSC = 1 - \frac{2TP}{2TP + FP + FN}$$

As with IoU we use an average over classes,

$$dice(loss) = 1 - \frac{\sum_{i=1}^{C}(dice_i)}{C}, dice_i = \frac{2TP_i}{2TP_i + FP_i + FN_i}$$

Intersect over the Union with penalties (IoUxy): IoUxy is similar to IoU but penalizes differently FP and FN in the denominator of the formula. The resulting formula weighting over classes is:

$$IoU_{xy}(loss) = 1 - \frac{\sum_{i=1}^{C}(IoU_{xyi})}{C}, IoU_{xyi} = \frac{TP_i}{TP_i + \alpha FP_i + \beta FN_i}$$

In these formulas α and β are such that α+β=2, α,β>=0. The question to answer is whether giving different weights to FN and FP (the two types of unwanted errors) will allow the approach to better segment each organ, and what is the winning α and β combination. We evaluate this by means of experimentation.

Loss without considering the background (diceBK): Since the background is easier to segment than organs and huge, diceBK is an alternative that removes the background from the loss formula (i.e. it averages loss over all classes except the background). The objective is to try to emphasize the need to segment the organs well. An experimental approach is necessary to evaluate if this alternative improves the outcome.

Uniclass segmentation: instead of a single multi-class problem with a single segmentation network, we can have one specific segmentation network specializing in segmenting each organ. The potential advantage is that we will be replacing a difficult multi-objective optimization problem [14] (minimize loss of segmentation of each organ) by *n* easier to optimize single objective uniclass problems (each one optimizes segmentation of one organ). Note however that, on the other hand, in uniclass versions all organs are marked as background except the one being segmented, and since organs have similarities that can be confounded, it may be difficult for the network to classify well. An experimental approach is necessary to reach conclusions regarding which alternative scores best, either a single multiclass segmentation network or *n* uniclass segmentation networks, one for each class.

2.3. Experimental setup

The segmentation network architecture is a relevant factor for the quality of segmentation. For this work we pick some of the most well-known generic segmentation networks, the U-NET [15], FCN [16] and DeepLabV3 [17]. The U-Net uses a 58-layer segmentation network with VGG-16 (7 stages, corresponding to 41 layers) for feature extraction (encoding). The FCN also used VGG-16 as encoder, and its total network size is smaller than UNet (51 layers). The decoder stages of U-Net are symmetric to the encoder stages, while FCN uses simple interpolation in the decoder stages. Both networks also include forward connections feeding feature maps from encoder to decoder stages. The two networks (U-Net and FCN) are the most-frequently used ones in segmentation of medical images. The third network, DeepLabV3, is a well-known segmentation network oft used in object recognition applications that outperformed most competitors due to some innovations. It is the deepest network tested in this work, with 100 layers and uses Resnet-18 as feature extractor (8 stages, totaling 71 layers). DeepLabV3 incorporates two important segmentation quality enhancing improvements, the Atrous Spatial Pyramid Pooling (ASPP) (improving segmentation of objects at multiple scales) and fully connected Conditional Random Fields (CRF) for improved localization of object boundaries using mechanisms from probabilistic graphical models. All segmentation networks were pre-trained versions based on object recognition data.

The Magnetic Resonance Imaging data used in our experimentation is a set of scans available in [18]. It consists of (MRI) acquisitions of 120 MRI sequences capturing abdominal organs (liver, kidneys and spleen) obtained using T1-DUAL fat suppression protocol. The sequences were acquired by a 1.5T Philips MRI, which produces 12-bit DICOM images with a resolution of 256 x 256. The ISDs varies between 5.5-9 mm (average 7.84 mm), x-y spacing is between 1.36 - 1.89 mm (average 1.61 mm) and the number of slices is between 26 and 50 (average 36). In total there are 1594 slices (532 slice per sequence) used for training and testing, with the testing sequences being chosen randomly to include 20% of all sequences in 5-fold cross-validation runs. Given the relatively limited size of the dataset, data augmentation was added after we verified that it contributes to improved scores, by increasing diversity and size of the dataset. Data augmentation was defined based on random translations of up to 10 pixels, random rotations up to 10 degrees, shearing up to 10 pixels and scaling up to 10%.

The experiments reported in this work were preceded by a set of iterations tuning configurations to the best possible results. Those iterations were based on setting configuration parameters, running training and testing data and interpreting output metrics. With this process we arrived at a set of configurations that would result in improved performance. The final network training parameters, used in our experimental work, were: learning function SGDM, with an initial learning rate=0.005, piecewise learning rate with drop period of 20 and learn rate drop factor of 0.9 (i.e. the learn rate would decrease to 90% every 20 epochs). Class balancing was applied in the pixel classification layer; training iterations were 500 epochs; minibatch sz=32; momentum= 0.9. But the factor that most improved performance was data augmentation, which we described before. A machine with a GPU NVIDEA G Force GTX1070 was used for the experiments.

The experiments were divided into two phases. The first phase chose the best performing segmentation network among the three candidates, using the default cross entropy loss function. Using the best performing chosen network, we then tested the various loss functions (we actually also tested the different loss function alternatives on FCN, but the results were worse than DeepLabV3, as expected). The loss functions used are cross entropy (crossE), IoU (IoU11) and IoUxy with different configurations of x and y, dice and dice without considering the background (diceBK). In the case of IoUxy we test the following options, besides IoU11: IoU1505 = IoUxy with α=1.5, β=0.5, IoU0515= IoUxy with α=0.5, β=1.5. Finally, we also compared multiclass versus n uniclass segmentations, the last alternative that we discussed in our proposals. In what concerns metrics used to evaluate the quality of the resulting segmentations, we focused mostly our analysis on per-class IoU (JI), since it allows us to assess the quality of segmentation of each organ separately. We also include initially an evaluation using different "global" metrics (metrics with scope over all pixels and averaged over classes), where it will be visible that some of the metrics do not evaluate well because they always score very high, due to the factors a), b) and c) that we studied before. In those results, besides mean and weighted IoU and mean and weighted accuracy, we also report mean BFScore. The BFScore measures the degree of matching between boundaries of the found segments and those of the corresponding groundtruth segments. Although BFScore is interesting because it applies a different perspective based on degree of match between boundaries, its scores depend on the settings of a boundary distance threshold (we used a default value is 0.75% of the length of the image diagonal) to detect a match or not between the two boundaries.

3. Results

3.1. Choose best-performing network

Table 1 shows the IoU (JI) of three segmentation network architectures (the default cross-entropy crossE loss function is used in this experiment). The best-performing network was deepLabV3 (2 percentage points (pp) better than FCN, FCN being 3 pp better than UNET). For that reason, we concentrate on reporting the results we obtained with deepLabv3 for the remainder of this work.

Table 1. IoU of segmentation networks with base crossE loss

class	DeepLabV3	FCN	UNET
Background	99%	99%	98%
Liver	86%	86%	74%
Spleen	82%	74%	73%
rKidney	77%	78%	75%
lKidney	81%	77%	78%
Avg IoU	**85%**	**83%**	**80%**

3.2. Comparison of loss function variations

Table 2 shows the global scores of the different loss functions, and table 3 details the results further by displaying loss scores for each organ measured as IoU. First of all, note that accuracy and weighted IoU always scored very high, they fail to detect deficiencies that exist in the segmentation for reasons we have discussed in section 2 of this work. The remaining metrics show different absolute values but rank different loss functions similarly to each other. We will focus the rest of the analysis on (mean) IoU because, contrary to sensitivity and BFScore, it does not depend on the setting of a distance threshold (BFScore) and accounts for FP (which sensitivity does not).

Based on the scores of mean IoU from table 2 we conclude that the best performing loss function was IoU0515, scoring 0.9 and improving 5 percentage points (pp) when compared with the default cross entropy (crossE) loss function (and 6 pp on sensitivity). Apart from IoU0515, all other alternatives tested except IoU1505 (i.e. crossE, IoU11, dice, dice no BK) had similar average scores (0.85); IoU (i.e. IoU11) and dice are highly correlated, which is confirmed by their close results (and those results were also similar to crossE).

Table 3 shows the detail for each organ. There we can see that IoU0515 achieved scores between 0.85 and 1 for the different classes, with the lowest scoring organ being the left kidney with a score of 0.85.

We conclude that IoU loss with higher weights on FN than on FP can improve segmentation quality. Recall that IoU0515 corresponds to the coefficient for mean false positives (FP) being assigned a weight of 25% (0.5/2) and that of false negatives (FN) being given 75% weight (1.5/2). Hence, quality of segmentation was improved by penalizing more the existence of false negatives than of false positives. False negatives correspond to an organ being classified as another organ or the background, so this alternative is focusing especially in optimizing the network to avoid this possibility.

Table 2. Global metrics for segmentation network DeepLabV3 with base crossE loss

	Accuracy	Mean Sensitivity	Mean IoU	Weighted IoU	Mean BFScore
crossE	0.99	0.88	0.85	0.99	0.90
iou11	0.99	0.88	0.85	0.99	0.90
iou1505	0.99	0.83	0.79	0.98	0.85

iou0515	**1.00**	**0.94**	**0.90**	**0.99**	**0.92**
dice	0.99	0.87	0.85	0.99	0.91
dice noBK	0.99	0.89	0.85	0.99	0.90

Table 3. IoU of segmentation network DeepLabV3 with diff. loss functions

IoU	**crossE**	**IoU**	**IoU**	**Iou**	**dice**	**dice BK**
α β	-	1 1	1.5 0.5	0.5 1.5	-	-
BackGround	0.99	0.99	0.99	**1.00**	0.99	0.99
liver	0.86	0.84	0.68	**0.88**	0.87	0.84
spleen	0.82	0.84	0.79	**0.87**	0.80	0.81
rkidney	0.77	0.82	0.76	**0.88**	0.81	0.82
lkidney	0.81	0.74	0.71	**0.85**	0.76	0.79
avg	**0.85**	**0.85**	**0.79**	**0.90**	**0.85**	**0.85**
rank	3	2	9	1	3	3

Table 4 tests another alternative: it compares the scores of the multiclass problem with those obtained for n uniclass problem (n=4, one for each organ). As discussed in section 2, the objective of this alternative is to evaluate whether replacing one multi-objective loss minimization problem by n single objective minimization tasks (i.e. training to segment a single organ in each separate network) would improve or damage segmentation quality. To compare the two options, the next experiment reports results of two runs: in the first run we used the original multiclass groundtruths provided by the dataset and configured different loss functions in order to run the multiclass problem segmenting into a set of organs simultaneously (liver, spleen, right and left kidney). For the second run we first transformed the multiclass groundtruth dataset into 4 uniclass groundtruth datasets, one for each organ, trained one segmentation network for each of the uniclass problems and used those networks to classify organs of the test images.

The results reported in table 4 show that the multiclass alternative obtained higher scores for any of three reported loss functions. Taking the average IoU over all classes, which is reported as the last row of table 4, crossE, dice and IoU improved from (0.77,0.73,0.79) to (0.85,0.85,0.85) when the multiclass version is ran instead of the uniclass networks. Looking at the details per organ, we can see that the liver actually scores the same (crossE, dice) or better (IoU11) using the uniclass alternative, but the other organs have worse scores in general. Note that, as we already hypothesized in section 2, the reason why uniclass fails more should be related to the increased chance of confounding background with the target organ, because in this formulation of the problem all other organs except the target organ are part of the background in the segmentation network of a specific organ.

Table 4. IoU achieved with multiclass vs uniclass

	multiclass			**uniclass**		
IoU	**crossE**	**IoU**	**dice**	**crossE**	**IoU**	**dice**
α β	-	1 1	-		1 1	

BackGround	0.99	0.99	0.99	-	-	-
liver	0.86	0.84	0.87	0.86	0.89	0.87
spleen	0.82	0.84	0.80	0.58	0.62	0.52
rkidney	0.77	0.82	0.81	0.72	0.50	0.79
lkidney	0.81	0.74	0.76	0.70	0.67	0.79
avg	**0.85**	**0.85**	**0.85**	**0.77**	**0.73**	**0.79**

4. Discussion, comparisons and illustration of results

In this section we first analyze and conclude based on the results reported in the previous section. Then we compare to results obtained in related works, and finally we illustrate MRI segmentations and 3D visualizations.

4.1. Discussion

From the previous experiments we conclude that IoU with modified FP and FN weights to weight FN more heavily was the best performing loss function variant in our experimental setup, and improved segmentation quality (measured as JI) significantly. We also conclude that cross entropy, dice and "balanced" IoU (IoU without modification to FP and FN weights) scored very similarly, but lower than the best variation. Note that we were already expecting dice and IoU to score very similarly because they are highly correlated, in fact one can be obtained from the other using as simple formula. Also interestingly, dice without background scored similarly to dice with background and most other options, we couldn't notice any sensible improvement or degradation with that choice. This is most probably related to the fact that the organs are "immersed" in the middle of the background, and the dice formula for each organ, with its factors (FP, FN, TP, TN), already accounts for the factors that are accounted by the dice of the background class. The conclusion is that there was no sensible advantage of excluding background from the loss function. Finally, we have seen that the single multiclass problem is preferable to *n* uniclass problems, as JI of organs in the uniclass problems were consistently lower than in the multiclass problem.

In what concerns generalization of our results to other data and techniques: the purpose of our experimentation was to use a statistically relevant experimental setup that would allow us to compare and validate loss function alternatives and variations. Hence, the results obtained should be generalizable to other segmentation networks segmenting MRIs (or CTs) of abdominal organs. That means it is preferable to apply IoUxy with higher weight on FN to improve segmentation quality. The remaining conclusions also apply to other techniques segmenting MRI of abdominal organs. On the other hand, we believe more work would be necessary to reach comparable conclusions in other medical imaging segmentation tasks, e.g. segmenting tumours or eye lesions or other body parts, since important factors may vary in those cases. Additionally, given our focus on testing alternatives to loss functions, we chose to test only three discrete choices of x and y in the best scoring IoUxy (11, 1.5/0.5, 0.5/1.5). One obvious generalization is to experiment with a more complete range of values and choose the combination yielding top JI. However, the precise combination that optimizes quality (JI) will depend on the dataset, therefore the most important conclusion is that it is worth tuning the weights on FN and FP in the loss formula with higher FN weights. In a practical system implementing these recommendations in some segmentation task, a calibration run of that task could be done to tune the best FN/FP weight combination (with FN weight higher than FP).

4.2. Comparison to other works

In this section we put our results into perspective by comparing with what was reported in other works regarding segmentation of abdominal organs, all of them proposing modifications of network architecture. Note also that our conclusions could be applied to improve any of those approaches further. Tables 5 and 6 show the IoU reported by those related approaches, including both MRI and CT sequences (computer tomography) (in fact, most works segment CT scans).

First, Table 5 compares our scores with those of a few other approaches running on the same MRI dataset as ours (therefore directly comparable). We can see that our best performing approach was superior to those compared. Both V19pUnet1-1 and V19pUnet can also be improve further if ran with our best found IoU0.5/1.5 loss or some tuned IoUxy alternative.

Table 6 shows a broader picture of scores reported in other works which implemented enhanced networks with architectural modifications to improve segmentation quality of CT and MRI scans of abdominal organs. These works use different datasets from ours, and many of them segment CT instead of MRI, therefore they are not directly comparable to our results, however it is interesting to analyze their scores. In those results, [19] and [20] achieved highest scores in segmentation of MRI images, and Hu et al. [21] and [22] obtained the best scores for CT. The results we obtained in this work, in spite of using only a general-purpose segmentation network and not testing other architectural modifications that were proposed in each of the works referenced in table 6, are still "competitive". Most importantly, they could be applied to any of those works to improve segmentation quality further. Note also that, in general, in table 6 segmentation of CT scans achieved better top scores than segmentation of MRI scans.

Table 5. Comparing to IoU of related approaches (CHAOS dataset)

MRI JI=IoU	Liver	spleen	R Kidney	L kidney
[9] teamPK				
U-Net	0.73	0.76	0.79	0.83
V19UNet	0.76	0.79	0.84	0.85
V19pUNet	0.85	0.83	0.85	0.86
V19pUnet1-1	0.86	0.83	0.86	**0.87**
deeplabV3 iou 0.5/1.5	**0.88**	**0.87**	**0.88**	0.85

Table 6. IoU as reported in some related approaches (MRI and CT)

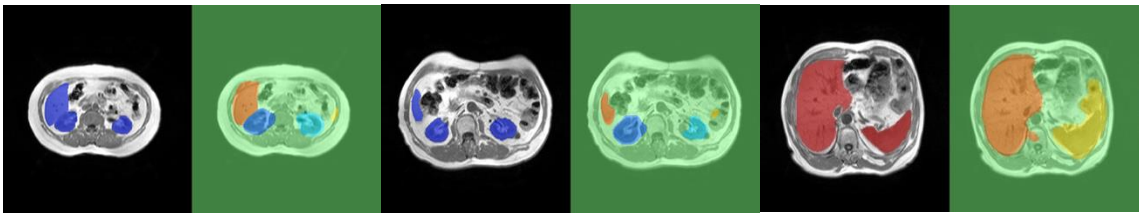
MRI JI=IoU	Liver	spleen	R Kidney	L kidney
[5]	0.84	0.87	0.64	0.57
[20]	0.90(LiverNet)	-	-	-
[19]	0.91	-	0.87	0.87
CT JI=IoU	**Liver**	**spleen**	**R Kidney**	**L kidney**
[23]	0.938	0.945		
[24]	0.85	-		
[4]	0.88	0.77		
[21]	0.92	0.89		
[22]	0.96	0.94	0.96	0.94
[25]	0.9	-	0.84	0.80
[8]				
F-net	0.86	0.79	0.79	0.80
BRIEF	0.74	0.60	0.60	0.60

U-Net	0.89	0.80	0.77	0.78
[6]	0.90	0.87	0.76	0.84

4.3. Illustrative MRI segmentations and 3D visualization

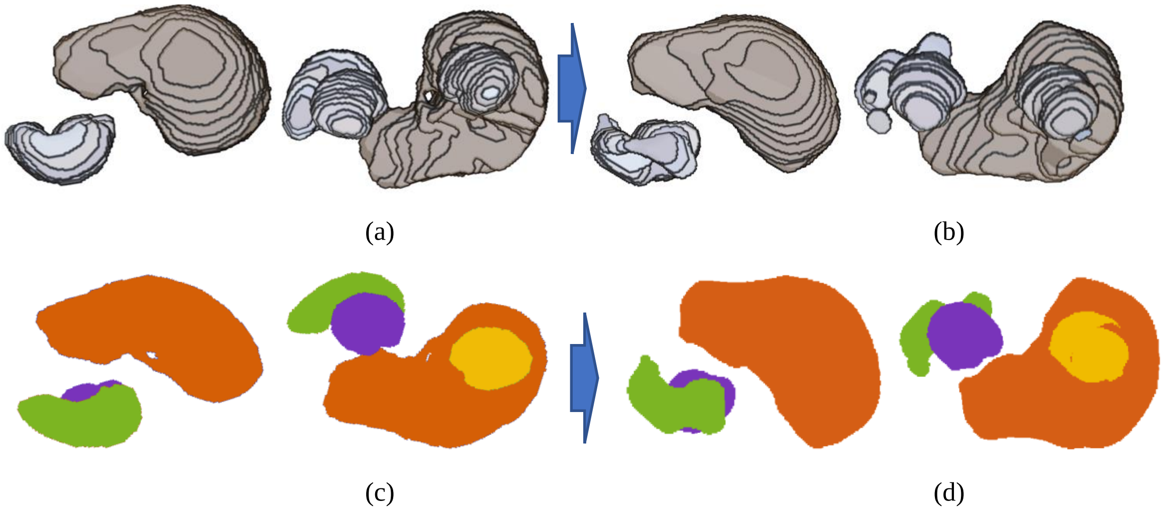
We end this work with a few examples of segmentation outputs obtained using the DeepLabv3 segmentation network. These serve an illustrative purpose. Figure 2 shows three randomly picked pairs of test slices, showing the segmentation groundtruths and the corresponding segmentation outputs. Figure 3 shows a 3-dimensional view of the stacked slices of one MRI sequence. We also provide a video with an animation of the sequence of slices as supplementary material.

In what concerns Figure 2, by comparing groundtruths with segmentations, we can see that the segmentations succeed at finding the organs and filling the organs areas, although with some imperfections, including spilling into the neighborhoods. In Figure 3 we can see both a 3D slices stacked model and a coloured model that distinguishes organs by colour. On the left we have the groundtruths and on the right we have the segmented outputs. It is possible to see some imperfections in this specific sequence as well, but in both illustrative examples the results are quite good.



(a) (b) (c)

Figure 2. Three slices with groundtruth and segmentation output (DeepLabV3). The left of each image is the groundtruth shown on a black background, the right is the segmentation. The segmentation succeeds at correctly finding the organs areas (liver and kidneys in (a) and (b), liver and spleen in (c)), with slight imperfections. In (a) both the liver and kidneys overflow slightly to neighborhood regions and there is a very small spurious spleen marking as well. In (c) the spleen overflows more seriously and the liver also overflows slightly in the segmentation output.



(a) (b)

(c) (d)

Figure 3. 3D model with slices of Abdomen. (a) is the groundtruth shown as a 3d model built with stacked slices, top and bottom view; (b) is the corresponding segmentation output 3D stacked model, showing some slight imperfections when compared with the groundtruth model; (c) is a coloured sketch of the 3D stacked model, showing each organ in a different colour, top and bottom view. (d) is the corresponding segmentation output, where a few imperfections are also visible.

5. Conclusions and future work

In this work we proposed and evaluated loss functions and loss function-related variations to improve quality of segmentation of MRI of abdominal organs using deep learning. Our work was driven by the observation that metrics and how metrics are evaluated is a relevant detail in scoring segmentation quality, and since loss is also a metric, one should explore the best possible loss variations and alternatives to improve segmentation outcome. We defined a set of loss functions and further variations, including multiple loss functions, uneven weighting of false positives and false negatives, not considering the background in the loss formula and also comparing multiclass with uniclass problem. We picked three segmentation networks that are most popular and frequently used in medical imaging to test the solutions. Our experiments allowed us to conclude that a certain variant of IoU (JI) that weights FN more than FP achieved the best scores on the best performing DeepLabV3 network, other alternatives tested did not improve the quality of segmentation as measured by JI. Since we studied a common theme to all deep learning segmentation networks, our conclusions can be applied further in the future to improve quality of segmentation in any advanced segmentation network architecture. Our future work on this issue will focus on investigating per-organ loss functions in some form of modified network architecture to allow multiple loss functions, applying the approaches to advanced segmentation architectures and an FN/FP weights calibration step for optimal loss function in any segmentation network.

Supplementary Materials: The following are available online at eden.dei.uc.pt/~pnf/MRIexample.mp4, a short movie illustrating segmentation output.

Funding: This research received no external funding.

Acknowledgments: We used a publicly available MRI dataset for our experiments. The dataset is presented in [18]. We acknowledge the organizers for allowing researchers to use this data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bereciartua, A., Picon, A., Galdran, A. and Iriondo P. (2015). Automatic 3D model-based method for liver segmentation in MRI based on active contours and total variation minimization. Biomedical Signal Processing and Control. 2015; 20:71–77. https://doi.org/10.1016/j.bspc.2015.04.005.
2. Le N., Bao P. and Huynh H. (2015). Fully automatic scheme for measuring liver volume in 3D MR images. Bio-medical materials and engineering. 2015; 26(s1):1361–1369. https://doi.org/10.3233/BME-151434.
3. Huynh H., Le N., Bao P., Oto A. and Suzuki K. (2017). Fully automated MR liver volumetry using watershed segmentation coupled with active contouring (2018). International journal of computer assisted radiology and surgery. 2017; 12(2):235–243. https://doi.org/10.1007/s11548-016-1498-9 PMID: 27873147.
4. Zhou, X., Takayama, R., Wang, S., Zhou, X., Hara, T. and Fujita, H. (2017). Automated segmentation of 3D anatomical structures on CT images by using a deep convolutional network based on end-to-end learning approach. In Medical imaging 2017: image processing (Vol. 10133, p. 1013324). International Society for Optics and Photonics.
5. Bobo M., Bao S., Huo Y., Yao Y., Virostko J., Plassard A. and Landman B. (2018). Fully convolutional neural networks improve abdominal organ segmentation. In Medical Imaging 2018: Image Processing (Vol. 10574, p. 105742V). International Society for Optics and Photonics.
6. Larsson, M., Zhang, Y. and Kahl, F. (2016). Deepseg: Abdominal organ segmentation using deep convolutional neural networks. In Swedish Symposium on Image Analysis 2016.

7. Chen, Y., Ruan, D., Xiao, J., Wang, L., Sun, B., Saouaf, R., Yang W., Li D. and Fan, Z. (2019). Fully Automated Multi-Organ Segmentation in Abdominal Magnetic Resonance Imaging with Deep Neural Networks. arXiv preprint arXiv:1912.11000.
8. Groza, V., Brosch, T., Eschweiler D., Schulz, H., Renisch, S. and Nickisch, H. (2018). "Comparison of deep learning-based techniques for organ segmentation in abdominal CT images," in 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands, pp. 1–3, 2018. → pages 15, 16.
9. Conze, P., Kavur, A., Gall, E., Gezer, N., Meur, Y., Selver, M. and Rousseau, F. (2020). Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks. arXiv preprint arXiv:2001.09521.
10. Cai, J., Lu, L., Zhang, Z., Xing, F., Yang, L., Yin, Q.: Pancreas segmentation in MRI using graph-based decision fusion on convolutional neural networks. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MIC- CAI 2016. LNCS, vol. 9901, pp. 442–450. Springer, Cham (2016). doi:10.1007/ 978-3-319-46723-8 51.
11. Sørensen, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons". Kongelige Danske Videnskabernes Selskab. 5 (4): 1–34.
12. Zhang, X., Thibault, G., Decencière, E., Marcotegui, B., Laÿ, B., Danno, R. & Chabouis, A. et al. (2014). Exudate detection in color retinal images for mass screening of diabetic retinopathy. Medical image analysis, 18(7), 1026-1043.
13. Salehi, Seyed Sadegh Mohseni, Deniz Erdogmus, and Ali Gholipour. "Tversky loss function for image segmentation using 3D fully convolutional deep networks." International Workshop on Machine Learning in Medical Imaging. Springer, Cham, 2017.
14. Deb, K. (2014). Multi-objective optimization. In Search methodologies (pp. 403-449). Springer, Boston, MA.
15. Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
16. Long, J., Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).
17. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.
18. Kavur A., Sinem N., Barıs M., Conze P., Groza V., Pham D.,Chatterjee S., Ernst P., Ozkan S., Baydar B., Lachinov D., Han S., Pauli J., Isensee F., Perkonigg M., Sathish R., Rajan R., Aslan S., Sheet D., Dovletov G., Speck O., Nurnberger A., Maier-Hein K., Akar B.,Unal G., Dicle O. and Selver M, (2020). CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation. In arXiv pre-print, Jan. 2020. https://arxiv.org/abs/2001.06535CHAOS data a DOI number: https://doi.org/10.5281/zenodo.3362845.
19. Fu Y., Mazur, T., Wu, X., Liu, S., Chang, X., Lu, Y., Harold, H., Kim, H., Roach, M., Henke, L. and Yang, D. (2018). A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. Medical physics, 2018. 45(11): p. 5129-5137.
20. Chlebus, G., Meine, H., Thoduka, S., Abolmaali, N., van Ginneken, B., Hahn, H., and Schenk, A. (2019). Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. PloS one, 14(5), e0217228.
21. Hu, P., Wu, F., Peng, J., Bao, Y., Chen, F. and Kong, D. (2017). Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. International journal of computer assisted radiology and surgery 12(3) (2017) 399–411.
22. Wang Y., Zhou Y., Shen W., Park S., Fishman E. and Yuille A. (2019). Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical image analysis, 2019. 55: p. 88-102.
23. Roth, R., Shen C., Oda H., Sugino T., Oda M., Hayashi H., Misawa K. and Mori K. (2018). A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 417–425.
24. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S., Clarkson, M. and Barratt, D. (2017). Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal ct with dense dilated networks. In: MICCAI, Springer (2017) 728–736.

25. Kim, J. and Lee, J. (2019). Deep-learning-based fast and fully automated segmentation on abdominal multiple organs from CT. In International Forum on Medical Imaging in Asia 2019 (Vol. 11050, p. 110500K). International Society for Optics and Photonics.