Loss function analysis and optimization for spinal segmentation in Magnetic Resonance Imaging

Manuel Hun

mhun@uni-koblenz.de

Institute for Computational Visualistics
University of Koblenz and Landau

11. Mai 2021

Overview

Motivation

Motivation:

- finding an appropriate loss function is non-trivial
- choice is based on the data analysis, however no general heuristics exist
- many specialized loss functions were proposed, generalization is missing

Research questions

Primary research question:

- With the given MRI data of cervical and lumbar vertebrae, which loss function with the UNet 3D architecture yields the best segmentation results?

Secondary research questions:

- Is it possible to directly learn a loss function for segmentation tasks only based on the provided data?
- Is it possible to optimize existing loss functions for segmentation tasks?

Definition

Loss function

A loss function $\mathcal{L}$ describes the mapping of the prediction set $\hat{Y}$, which is generated by the parameterized function h_θ, and the ground truth set Y onto a non negative real number.

$$\mathcal{L} : h_\theta \times Y \to \mathbb{R}_0^+. \tag{1}$$

Aim

Empirical Risk Minimization

To find the optimal function h_θ the expected loss over a random sample set of X is minimized, a paradigm known as the Empirical Risk Minimization (ERM) [Vap92].

$$\underset{\theta}{\operatorname{argmin}} \, \mathcal{L}(\hat{Y}, Y). \tag{2}$$

Categories of loss functions

Distribution:

- Cross Entropy, Weighted Cross Entropy
- TopK, Focal

Region:

- Dice, Intersection over Union, Tversky
- Asymmetric,Effectiveness

Distance:

- Hausdorff, Boundary

Compound:

- Dice-Focal, Tversky-Focal, Cross Entropy + Dice

Distribution

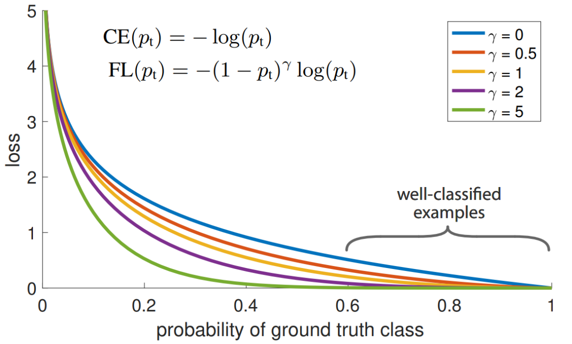


Abbildung: Visualization of the impact of different γ values of the Focal loss on examples compared to the standard Cross Entropy (CE) [LGG+18].

Region

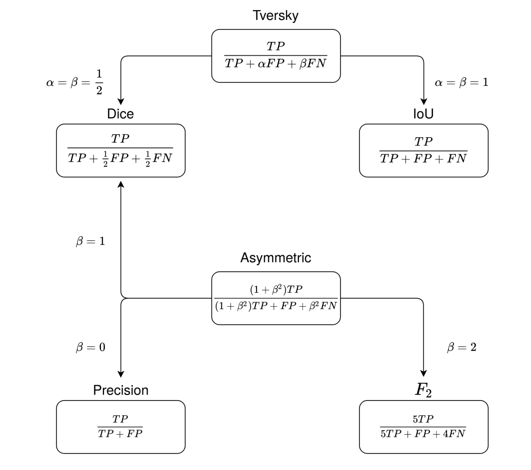


Abbildung: Overview of the relationship between different regional loss functions.

Distance

- complementary information for distribution and regional losses
- boundary measure
- makes use of distance maps
- Hausdorff: double-sided, Boundary: one-sided

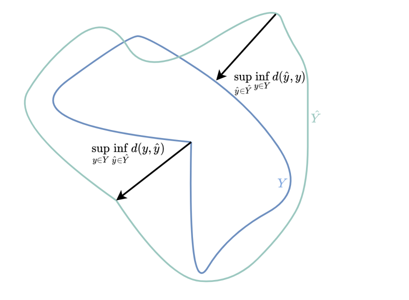


Abbildung: Visualization of the estimation of the Hausdorff distance.

Compound

Combine properties:

- Tversky-Focal
- Dice-Focal

Multi-loss functions:

- Cross Entropy + Dice
- Distance loss + Regional loss

Analysis - structure size

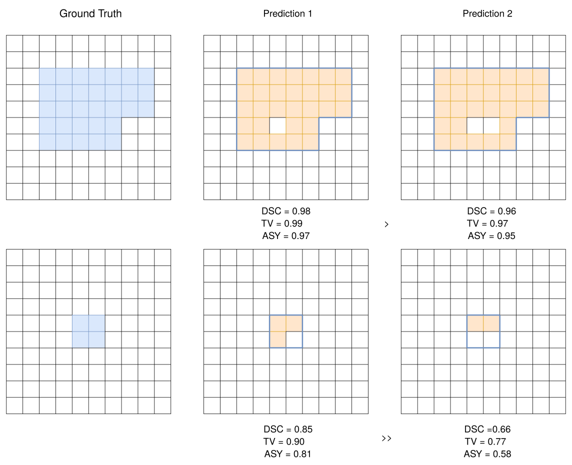


Abbildung: Behavior of the loss functions with different sizes of prediction.

Analysis - shape awareness

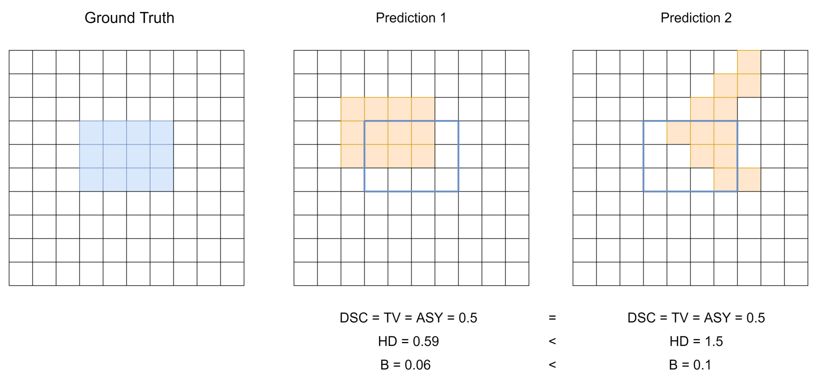


Abbildung: Behavior of the loss functions with different shapes of prediction.

Analysis - overfitting and underfitting

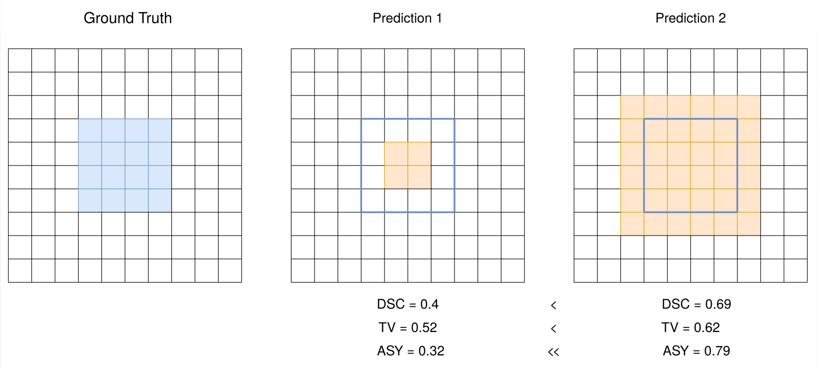


Abbildung: Behavior of the loss functions with predictions, that overfit and underfit.

Methods

How to learn loss functions directly from data?

- genetic algorithms
- meta-learning

How to optimize existing loss functions?

- adjust learnable parameters of loss function during training
- calculate adaptive weightings of compound loss functions

Learn loss - genetic algorithms

- proposed by Gonzalez [GM20]
- only tested on 1-dimensional target domain (regression, classification)

Idea:

- represents loss functions as expression tree
- each tree consists of computational rules (primitives) and constant values (terminals)
- individuals get modified through genetic operations like mutation, crossover and selection
- loss function are learned before the actual training

Learn loss - genetic algorithm overview

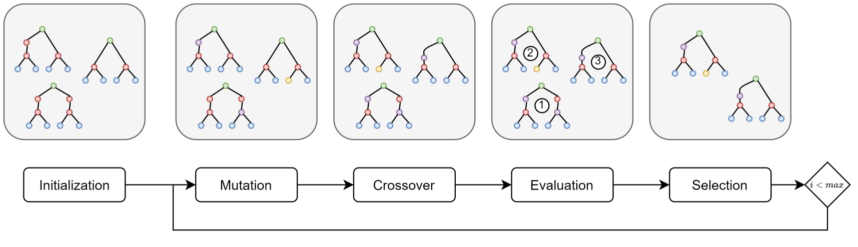


Abbildung: Overview of the genetic algorithm approach.

Learn loss - meta learning

- proposed by Bechtle *et al.* (Facebook Research) [BMC$^+$21]
- only tested on 1-dimensional target domain (regression, classification)

Idea:

- learns a loss function during training
- loss function is represented by another neural network

Learn loss - meta learning

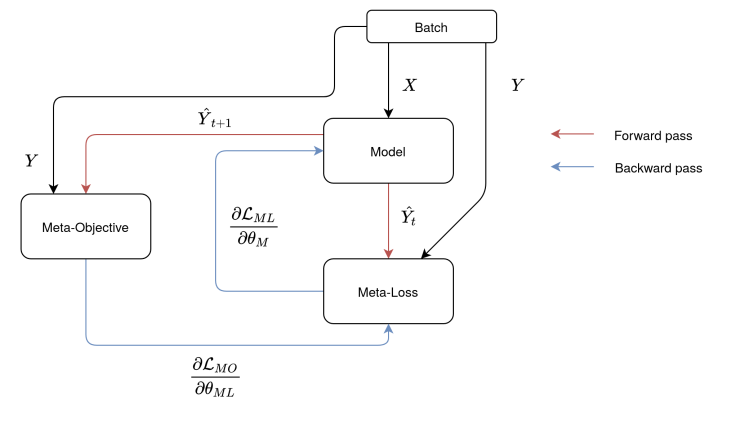


Abbildung: Overview of the meta-learning approach.

Optimize loss - greedy approach

- proposed by Seo [SBX20]

Idea:

- take discrete actions A at each training step
- each action operates on the learnable parameters p by reducing, increasing or leaving
- cartesian product delivers all possible combinations (A^p)
- test all possible combinations A^p
- choose optimal A^*, that provides minimal L^2
- strongly explorative

Optimize loss - greedy approach

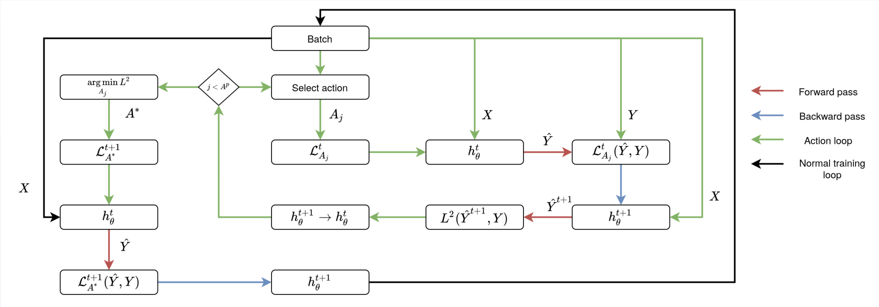


Abbildung: Overview of the approach with the greedy algorithm.

Optimize loss - adaptive weighting

Adaptive weighting

Adaptive weighting methods aim to change the weightings of each individual loss function of the multi loss function depending on a basic heuristic during the training process.

$$\mathcal{L}_{AW} = \sum_{i}^{M} \alpha_i \mathcal{L}_i. \tag{3}$$

Optimize loss - rebalance

Idea:

- increase α_t^i per training step with a constant value
- constant value is calculated over the maximum number of training steps
- only suitable for multi-loss functions consisting of 2 loss functions

Optimize loss - Soft Adapt

Idea:

- proposed by Hedari [HTM19]
- uses information of prior loss values
- approximation of the forward finite difference s_t^i
- softmax operation normalizes all weights

$$\alpha_t^i = \frac{e^{\beta s_t^i}}{\sum_l^n e^{\beta s_t^l}}. \tag{4}$$

Optimize loss - Coefficient of Variations

Idea:

- proposed by Groenendijk et al. [HTM19]
- instead of the direct loss values a ratio-scale ℓ_t is measured
- Welford algorithm [WW62] tracks $\sigma_{\ell_t^i}$ and $\mu_{\ell_t^i}$
- α_t^i is determined via the relative standard deviation

$$\alpha_t^i = \frac{1}{z_t} c_{\ell_t^i} = \frac{1}{z_t} \frac{\sigma_{\ell_t^i}}{\mu_{\ell_t^i}}. \tag{5}$$

Datasets

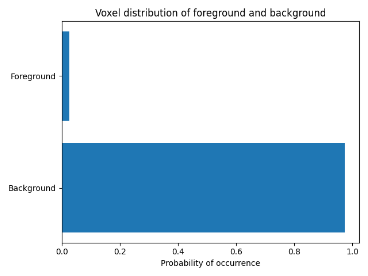
Cervical:

- provided by VisSim research group
- 12 subjects, distribution between men and women unknown
- C1 to C7 cervical vertebral bodies
- Segmentation time per volume: 2:30h

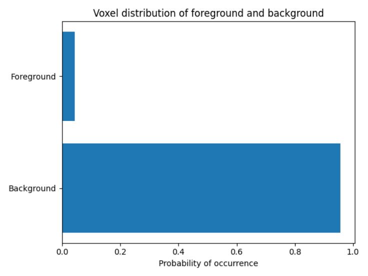
Lumbar:

- provided by MyoSegmenTum database
- 51 subjects (37 female, 14 male)
- L1 to L5 lumbar vertebral bodies
- Segmentation time per volume: 2h

Datasets - voxel distribution



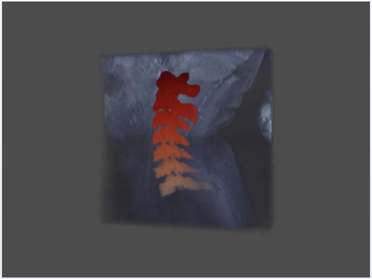
(a)



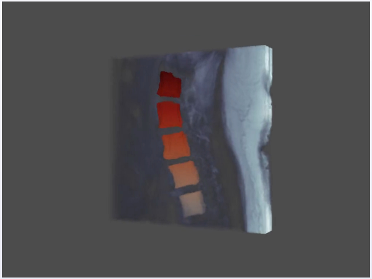
(b)

Abbildung: Voxel distribution between foreground and background in the cervical (a) and lumbar dataset (b).

Datasets - visualization



(a)



(b)

Abbildung: 3D rotated sagittal view of the ground truth cervical vertebrae (C1-C7) (a) and lumbar vertebrae (L1-L5) (b) embedded in the normalized volume data.

Augmentation and Preprocessing

Augmentation:

- Affine transformations
- Gaussian noise
- Interpolate random element from batch with entire batch (MixUp[1])

Preprocessing:

- Percentile normalization (0.5,99.5)
- Z-Normalization, zero mean and variance of one

[1] [ZCDLP18]

3D UNet

- Model architecture follows the generalization of UNet [RFB15] architecture for 3D [L+16]
- Batch normalization [IS15] replaced by Instance normalization [UVL17]
- ReLU replaced by Leaky ReLU
- Initiliasation method follows the proposal of Kaimin [HZRS15]

Training

- mixed-precision training and validation
- training is done in FP16, validation is done in FP32
- k-fold cross validation with predefined seed

K-fold cross validation

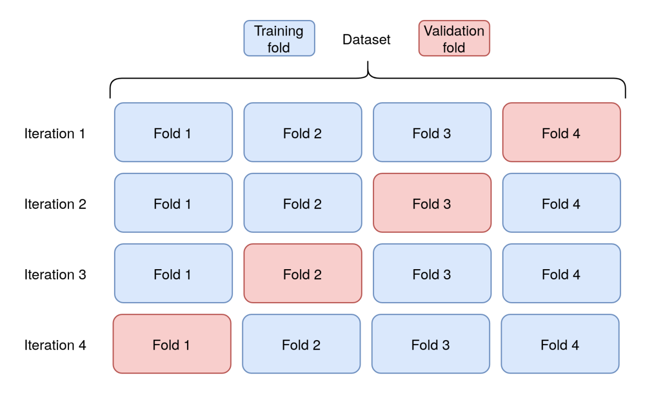


Abbildung: Example k-fold cross validation with k=4, where a blue rectangle indicates a training fold and a red rectangle a validation fold.

Results - cervical baseline

	Loss	Acc	Dice	F_2	IoU	Precision	Recall	Hausdorff
D	$\mathcal{L}_{BC}$	98.15 ± (0.17)	70.99 ± (3.17)	82.79 ± (1.37)	55.61 ± (3.70)	57.97 ± (4.44)	**95.11** ± **(1.92)**	27.94 ± (8.12)
	$\mathcal{L}_{F}$	98.35 ± (0.28)	62.02 ± (1.77)	58.86 ± (4.06)	45.67 ± (2.03)	79.11 ± (1.46)	57.24 ± (5.42)	29.59 ± (2.93)
R	$\mathcal{L}_{D}$	99.00 ± (0.06)	80.55 ± (2.87)	84.39 ± (0.80)	67.80 ± (3.69)	77.65 ± (7.40)	89.30 ± (2.75)	24.65 ± (6.51)
	$\mathcal{L}_{T}$	99.03 ± (0.04)	79.69 ± (1.56)	79.34 ± (2.36)	66.59 ± (1.76)	83.62 ± (7.16)	79.91 ± (4.86)	31.40 ± (1.73)
	$\mathcal{L}_{A}$	98.79 ± (0.01)	78.50 ± (4.50)	**86.46** ± **(2.27)**	65.08 ± (5.75)	69.24 ±(6.75)	94.63 ± (0.75)	20.13 ± (4.30)
Ds	$\mathcal{L}_{B}$	98.96 ± (0.17)	77.95 ± (7.52)	78.48 ± (7.43)	64.84 ± (8.93)	82.04 ± (6.19)	79.30 ± (7.23)	**19.02** ± **(6.42)**
	$\mathcal{L}_{HD}$	98.39 ± (0.16)	65.17 ± (3.93)	63.48 ± (5.43)	48.97 ± (4.11)	75.66 ± (5.85)	62.84 ± (6.36)	25.34 ± (1.27)
C	$\mathcal{L}_{DCE}$	98.93 ± (0.02)	79.35 ± (2.22)	83.40 ± (2.26)	66.04 ± (2.72)	75.82 ± (5.97)	87.74 ± (4.58)	22.06 ± (1.87)
	$\mathcal{L}_{TF}$	**99.09** ± **(0.01)**	**81.15** ± **(2.53)**	80.83 ± (1.45)	**68.65** ± **(3.16)**	**84.78** ± **(6.21)**	81.77 ± (3.44)	23.95 ± (5.73)
	$\mathcal{L}_{DF}$	98.99 ± (0.01)	80.53 ± (2.81)	84.72 ± (1.56)	67.83 ± (3.43)	76.63 ± (4.92)	89.27 ± (1.85)	27.51 ± (3.84)

Tabelle: Baseline table for the cervical dataset.

Results - cervical visualization

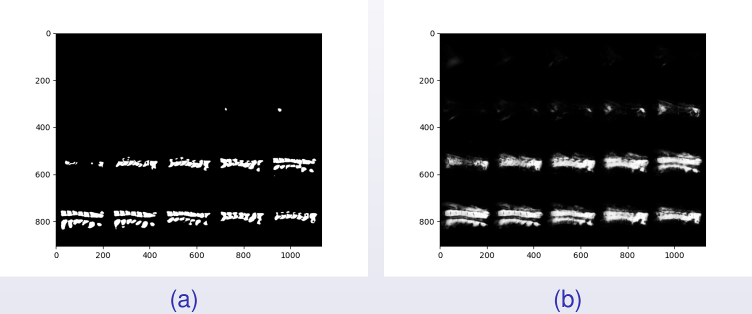
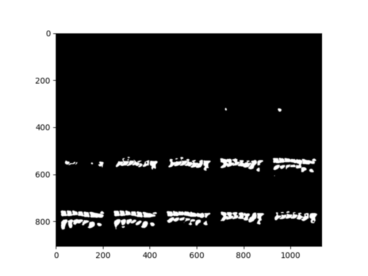
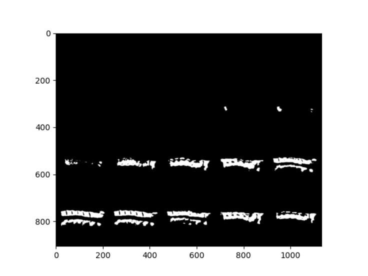


Abbildung: Example visualization of ground truth (a) and prediction of the model (b) with the Cross Entropy loss function.

Results - cervical visualization



(a)



(b)

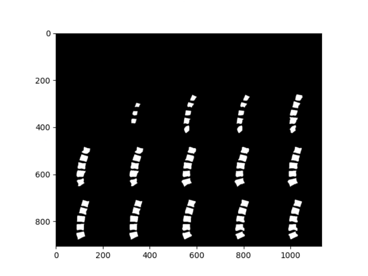
Abbildung: Example visualization of ground truth (a) and prediction of the model (b) with the TverskyFocal loss function.

Results - lumbar baseline

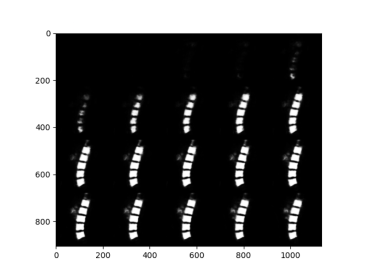
	Loss	Acc	Dice	F_2	IoU	Precision	Recall	Hausdorff
D	$\mathcal{L}_{BC}$	98.09 ± (0.06)	83.27 ± (1.20)	89.16 ± (0.13)	72.09 ± (2.13)	75.80 ± (2.81)	**95.36** ± **(1.64)**	19.01 ± (3.02)
	$\mathcal{L}_F$	98.34 ± (0.29)	83.07 ± (2.31)	80.90 ± (4.18)	71.82 ± (3.57)	90.55 ± (0.65)	79.99 ± (5.63)	13.80 ± (1.70)
R	$\mathcal{L}_D$	**98.79** ± **(0.23)**	**88.63** ± **(1.08)**	89.41 ± (1.35)	**80.35** ± **(0.73)**	89.44 ± (1.45)	90.32 ± (2.04)	11.44 ± (1.74)
	$\mathcal{L}_T$	98.79 ± (0.24)	88.04 ± (1.14)	85.88 ± (1.06)	79.46 ± (0.92)	93.55 ± (2.15)	84.65 ± (1.06)	**10.89** ± **(1.50)**
	$\mathcal{L}_A$	98.63 ± (0.17)	87.55 ± (0.66)	**90.61** ± **(1.22)**	78.64 ± (0.32)	84.03 ± (1.75)	94.03 ± (2.03)	12.65 ± (3.47)
Ds	$\mathcal{L}_B$	98.64 ± (0.23)	86.79 ± (0.83)	86.42 ± (1.12)	77.40 ± (0.82)	90.43 ± (0.49)	86.26 ± (1.46)	12.78 ± (2.39)
	$\mathcal{L}_{HD}$	97.93 ± (0.31)	78.64 ± (2.72)	75.72 ± (5.29)	65.48 ± (3.53)	87.46 ± (1.47)	74.11 ± (6.89)	15.57 ± (0.37)
C	$\mathcal{L}_{DCE}$	98.76 ± (0.26)	88.35 ± (1.24)	88.87 ± (1.41)	79.88 ± (1.34)	89.12 ± (1.36)	90.09 ± (1.72)	12.45 ± (1.76)
	$\mathcal{L}_{TF}$	**98.79** ± **(0.23)**	88.05 ± (0.87)	85.94 ± (0.59)	79.43 ± (0.48)	**93.56** ± **(2.04)**	84.82 ± (0.85)	11.61 ± (0.70)
	$\mathcal{L}_{DF}$	98.78 ± (0.27)	88.46 ± (1.36)	89.35 ± (1.15)	80.06 ± (1.19)	89.88 ± (1.80)	90.95 ± (1.45)	12.57 ± (0.87)

Tabelle: Baseline table for the lumbar dataset.

Results - lumbar visualization



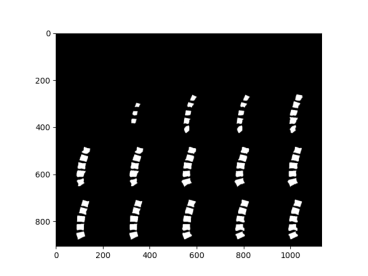
(a)



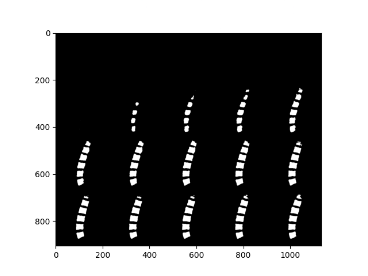
(b)

Abbildung: Example visualization of ground truth (a) and prediction of the model (b) with the Cross Entropy loss function.

Results - lumbar visualization



(a)



(b)

Abbildung: Example visualization of ground truth (a) and prediction of the model (b) with the TverskyFocal loss function.

Results - genetic approach

Genetic approach:

- could not yield any competitive results
- Possible causes:
  - large search space
  - insufficient time given (5 days)
  - input data dimension too large compared to authors applied data (regression, classification)

Results - meta learning

Meta learning:

- could not yield any competitive results
- Possible causes:
  - many sensitive decisions regarding the meta loss network:
    - model architecture
    - optimizer
    - learning rate
  - input data dimension too large compared to authors applied data (regression, classification)

Results - greedy algorithm

Loss	Acc	Dice	F_2	IoU	Precision	Recall	Hausdorff
$\mathcal{L}_D$	99.00 ± (0.06)	80.55 ± (2.87)	84.39 ± (0.80)	67.80 ± (3.69)	77.65 ± (7.40)	89.30 ± (2.75)	24.65 ± (6.51)
$\mathcal{L}_T$	99.03 ± (0.04)	79.69 ± (1.56)	79.34 ± (2.36)	66.59 ± (1.76)	**83.62** ± **(7.16)**	79.91 ± (4.86)	31.40 ± (1.73)
$\mathcal{L}_A$	98.79 ± (0.01)	78.50 ± (4.50)	**86.46** ± **(2.27)**	65.08 ± (5.75)	69.24 ±(6.75)	**94.63** ± **(0.75)**	20.13 ± (4.30)
$\mathcal{L}_T/G$	**99.11** ± **(0.05)**	**82.16** ± **(1.96)**	83.81 ± (0.82)	**70.06** ± **(2.47)**	83.22 ± (3.85)	86.74 ± (1.86)	**20.06** ± **(0.37)**
$\mathcal{L}_E/G$	99.07 ± (0.06)	81.36 ± (1.61)	84.28 ± (0.66)	68.97 ± (1.75)	81.78 ± (5.60)	88.39 ± (2.31)	22.29 ± (2.51)

Tabelle: Greedy algorithms results for the cervical dataset.

Loss	Acc	Dice	F_2	IoU	Precision	Recall	Hausdorff
$\mathcal{L}_D$	98.79 ± (0.23)	88.63 ± (1.08)	89.41 ± (1.35)	80.35 ± (0.73)	89.44 ± (1.45)	90.32 ± (2.04)	11.44 ± (1.74)
$\mathcal{L}_T$	98.79 ± (0.24)	88.04 ± (1.14)	85.88 ± (1.06)	79.46 ± (0.92)	**93.55** ± **(2.15)**	84.65 ± (1.06)	**10.89** ± **(1.50)**
$\mathcal{L}_A$	98.63 ± (0.17)	87.55 ± (0.66)	**90.61** ± **(1.22)**	78.64 ± (0.32)	84.03 ± (1.75)	**94.03** ± **(2.03)**	12.65 ± (3.47)
$\mathcal{L}_T/G$	**98.82** ± **(0.20)**	**88.74** ± **(0.63)**	89.11 ± (0.59)	**80.65** ± **(0.18)**	91.32 ± (1.64)	89.47 ± (0.51)	11.39 ± (3.66)
$\mathcal{L}_E/G$	98.80 ± (0.26)	88.54 ± (1.21)	88.99 ± (0.97)	80.16 ± (0.99)	91.89 ± (1.13)	90.60 ± (2.30)	11.94 ± (2.69)

Tabelle: Greedy algorithms results for the lumbar dataset.

Results - adaptive weighting

Loss	Acc	Dice	F_2	IoU	Precision	Recall	Hausdorff
$\mathcal{L}_B$	98.96 ± (0.17)	77.95 ± (7.52)	78.48 ± (7.43)	64.84 ± (8.93)	**82.04** ± **(6.19)**	79.30 ± (7.23)	**19.02** ± **(6.42)**
$\mathcal{L}_B/R$	**98.97** ± **(0.13)**	**80.45** ± **(3.17)**	**84.59** ± **(2.38)**	**67.83** ± **(3.98)**	76.69 ± (4.35)	**89.68** ± **(1.56)**	29.37 ± (7.71)
$\mathcal{L}_B/COV$	98.62 ± (0.06)	74.68 ± (2.02)	80.31 ± (0.62)	59.95 ± (2.19)	69.26 ± (3.48)	87.09 ± (3.28)	30.15 ± (1.47)
$\mathcal{L}_{HD}$	98.39 ± (0.16)	65.17 ± (3.93)	63.48 ± (5.43)	48.97 ± (4.11)	75.66 ± (5.85)	62.84 ± (6.36)	**25.34** ± **(1.27)**
$\mathcal{L}_{HD}/R$	**98.40** ± **(0.33)**	**69.29** ± **(4.70)**	**74.47** ± **(4.20)**	**53.95** ± **(5.34)**	67.19 ± (0.75)	**79.67** ± **(5.72)**	28.67 ± (6.30)
$\mathcal{L}_{HD}/COV$	98.15 ± (0.15)	64.98 ± (4.99)	68.97 ± (4.05)	48.71 ± (5.47)	64.10 ± (4.03)	72.79 ± (5.39)	33.22 ± (9.12)
$\mathcal{L}_{DCE}$	**98.93** ± **(0.02)**	**79.35** ± **(2.22)**	**83.40** ± **(2.26)**	**66.04** ± **(2.72)**	**75.82** ± **(5.97)**	87.74 ± (4.58)	**22.06** ± **(1.87)**
$\mathcal{L}_{DCE}/S$	98.84 ± (0.06)	78.76 ± (2.09)	84.50 ± (2.73)	65.33 ± (2.13)	72.39 ± (3.30)	**90.12** ± **(4.06)**	36.50 ± (10.70)

Tabelle: Adaptive weighting results for the cervical dataset.

Loss	Acc	Dice	F_2	IoU	Precision	Recall	Hausdorff
$\mathcal{L}_B$	**98.64** ± **(0.23)**	86.79 ± (0.83)	86.42 ± (1.12)	77.40 ± (0.82)	**90.43** ± **(0.49)**	86.26 ± (1.46)	12.78 ± (2.39)
$\mathcal{L}_B/R$	98.53 ± (0.01)	**87.26** ± **(0.19)**	87.29 ± (0.08)	**79.27** ± **(0.31)**	88.57 ± (0.46)	87.73 ± (0.21)	**9.56** ± **(0.62)**
$\mathcal{L}_B/COV$	98.63 ± (0.19)	86.89 ± (0.91)	**87.82** ± **(0.93)**	77.59 ± (0.92)	88.22 ± (2.24)	**89.22** ± **(1.75)**	14.14 ± (2.30)
$\mathcal{L}_{HD}$	97.93 ± (0.31)	78.64 ± (2.72)	75.72 ± (5.29)	65.48 ± (3.53)	87.46 ± (1.47)	74.11 ± (6.89)	15.57 ±(0.37)
$\mathcal{L}_{HD}/R$	98.42 ± (0.17)	84.23 ± (0.71)	83.58 ± (2.06)	73.61 ± (0.35)	89.20 ± (3.04)	83.57 ± (3.18)	14.08 ± (1.01)
$\mathcal{L}_{HD}/COV$	**98.51** ± **(0.23)**	**85.52** ± **(1.03)**	**85.10** ± **(1.46)**	**75.54** ± **(1.69)**	**89.73** ± **(1.94)**	**85.12** ± **(1.99)**	**11.98** ± **(1.24)**
$\mathcal{L}_{DCE}$	**98.76** ± **(0.26)**	**88.35** ± **(1.24)**	88.87 ± (1.41)	**79.88** ± **(1.34)**	**89.12** ± **(1.36)**	90.09 ± (1.72)	12.45 ± (1.76)
$\mathcal{L}_{DCE}/S$	98.65 ± (0.17)	87.46 ± (0.34)	**89.42** ± **(1.28)**	78.50 ± (0.98)	86.31 ± (0.79)	**92.26** ± **(1.93)**	**12.38** ± **(4.15)**

Tabelle: Adaptive weighting results for the lumbar dataset.

Summary

- the primary research objectives of segmenting cervical and lumbar vertebrae from MRI data were met
- region based and compound loss functions showed superior results over distribution based and distance based loss functions
- learning a loss could not be verified by genetic algorithms as well as by meta-learning for the given data
- greedy approach led to a better performance compared to the corresponding baseline
- rebalance strategy achieved a steady improvement

Thank you for your attention!

BECHTLE, Sarah ; MOLCHANOV, Artem ; CHEBOTAR, Yevgen ; GREFENSTETTE, Edward ; RIGHETTI, Ludovic ; SUKHATME, Gaurav ; MEIER, Franziska:
Meta-Learning via Learned Loss.
2021

GONZALEZ, Santiago ; MIIKKULAINEN, Risto:
Improved Training Speed, Accuracy, and Data Utilization Through Loss Function Optimization.
2020

HEYDARI, A. A. ; THOMPSON, Craig A. ; MEHMOOD, Asif:
SoftAdapt: Techniques for Adaptive Loss Weighting of Neural Networks with Multi-Part Loss Functions.
2019

- HE, Kaiming ; ZHANG, Xiangyu ; REN, Shaoqing ; SUN, Jian:
  Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.
  2015
- IOFFE, Sergey ; SZEGEDY, Christian:
  Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.
  2015
- LIN, Tsung-Yi ; GOYAL, Priya ; GIRSHICK, Ross ; HE, Kaiming ; DOLLÁR, Piotr:
  Focal Loss for Dense Object Detection.
  2018

- RONNEBERGER, Olaf ; FISCHER, Philipp ; BROX, Thomas: *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015
- SEO, Hyunseok ; BASSENNE, Maxime ; XING, Lei: Closing the Gap Between Deep Neural Network Modeling and Biomedical Decision Making Metrics in Segmentation via Adaptive Loss Functions. In: *IEEE Transactions on Medical Imaging* PP (2020), 10, S. 1–1. http://dx.doi.org/10.1109/TMI.2020.3031913. – DOI 10.1109/TMI.2020.3031913

ULYANOV, Dmitry ; VEDALDI, Andrea ; LEMPITSKY, Victor:
Instance Normalization: The Missing Ingredient for Fast Stylization.
2017

VAPNIK, V.:
Principles of Risk Minimization for Learning Theory.
In: MOODY, J. (Hrsg.) ; HANSON, S. (Hrsg.) ; LIPPMANN, R. P. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 4, Morgan-Kaufmann, 1992

WELFORD, Author(s) B. P. ; WELFORD, B. P.:
Note on a method for calculating corrected sums of squares and products.
In: *Technometrics* (1962), S. 419–420

- ZHANG, Hongyi ; CISSE, Moustapha ; DAUPHIN, Yann N. ; LOPEZ-PAZ, David:
  mixup: Beyond Empirical Risk Minimization.
  2018
- ÇIÇEK Özgün ; ABDULKADIR, Ahmed ; LIENKAMP, Soeren S. ; BROX, Thomas ; RONNEBERGER, Olaf:
  3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation.
  2016