THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

AAPM 2022
JULY 10–14 | WASHINGTON, DC
64TH ANNUAL MEETING & EXHIBITION

AI AUTO-SEGMENTATION: CLINICAL IMPLEMENTATION CONCERNS AND COMMISSIONING GUIDELINES

Jinzhong Yang, PhD
Department of Radiation Physics
MD Anderson Cancer Center

1

Disclosure

- None related to this presentation

2

Learning Objectives

- To understand the limitation factors for clinical implementation of auto-segmentation tools
- To understand the concerns in clinical implementation of AI segmentation tools
- To learn the proper commissioning procedures of auto-segmentation tools

3

Outline

- What is AI Segmentation?
- What are the concerns in clinical implementation of AI segmentation tools?
- What are the guidelines in commissioning AI segmentation tools?

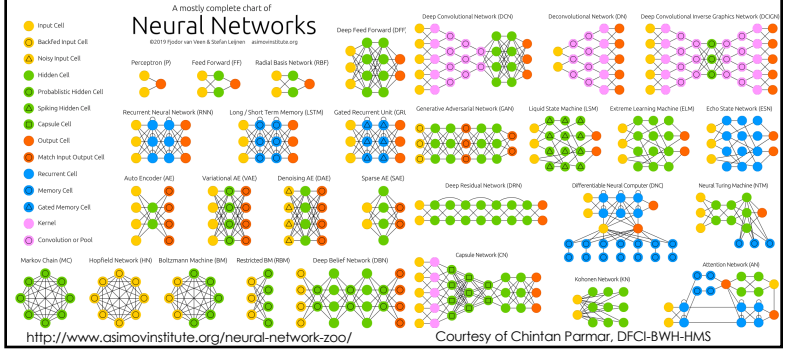
4

What is AI Segmentation?

5

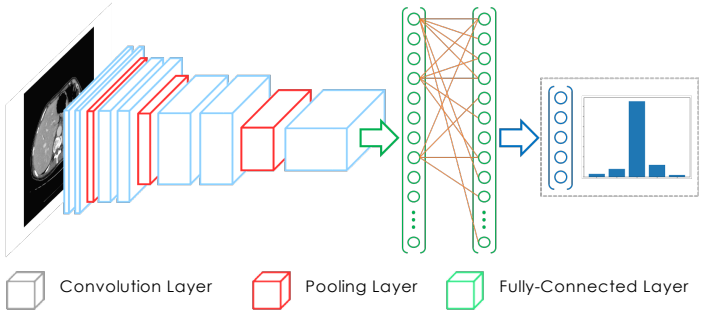
Neural Networks

A set of algorithms modeled loosely after the human brain to recognize patterns



6

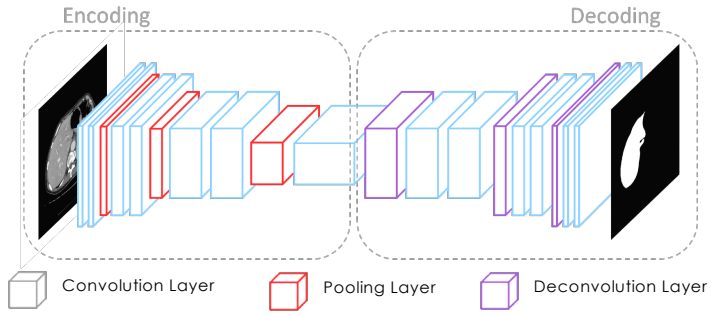
Convolutional Neural Networks (CNN)



CE Cardenas, et al. Advances in auto-segmentation, Semin Radiat Oncol, 2019

7

Fully Convolutional Networks (FCN)

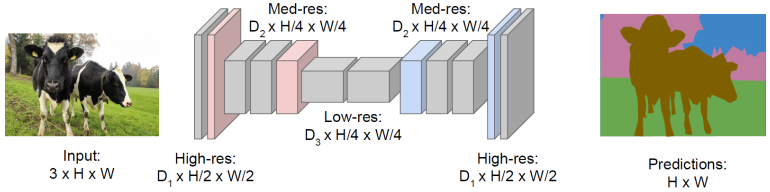


CE Cardenas, et al. Advances in auto-segmentation, Semin Radiat Oncol, 2019

8

Segmentation Using Fully Conventional Networks

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!



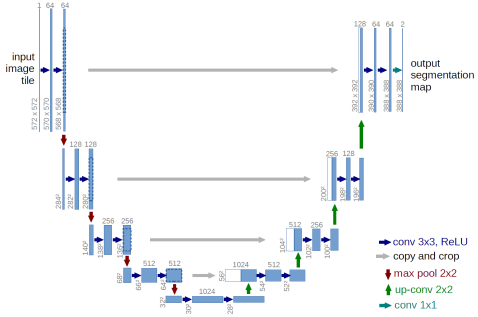
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Slide adapted from Fei-Fei Li & Justin Johnson & Serena Yeung, CS231n

9

Segmentation using Unet

- Training in an end-to-end setting
- Need very few annotated images

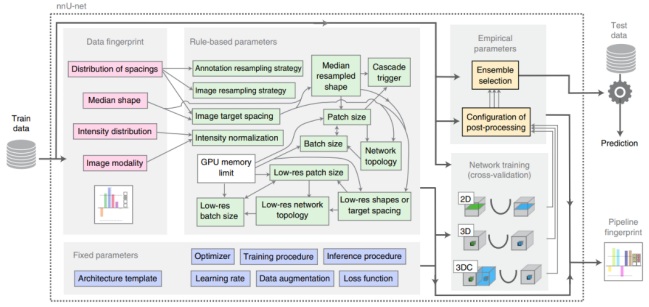


HTTPS://LMB.INFORMATIK.UNI-FREIBURG.DE/PEOPLE/RONNEBER/U-NET/

10

Segmentation using nnUnet

Self-configuration including preprocessing, data augmentation, network architecture, training and post-processing



Isensee et al. Nat. Methods, 18, 2021

11

AI Segmentation

- Network training
  - Data curation: a lot of training data may need to be curated for training
  - Data augmentation: to create variations from limited data for network training
- Network architecture
  - Number of layers
  - Type of layers
- Hyperparameter settings
- Pre- and post-processing of segmentation

12

WHAT ARE THE CONCERNS IN CLINICAL IMPLEMENTATION OF AI SEGMENTATION?

13

Limitation Factors in Clinical Use of Auto-Segmentation

- Poor quality of auto-segmentation results
  - The time needs to review auto-segmentation results
  - The time needs to modify inaccurate contours
- Poor workflow integration
  - Multiple steps to run auto-segmentation
  - Direct access in a treatment planning system
  - The time needs to wait for segmentation results

J Yang, et al. Med. Phys. 45(10), 2018

14

AI Segmentation Commissioning Concerns

- Segmentation accuracy
  - Benchmark datasets
  - Inter-institutional variations
  - Evaluation Metrics
- Workflow integration
  - Seamless integration with local clinical practice

15

AI Segmentation Model

- Vendor supplied model
  - Understand the variations between training data and local institution data
- Vendor supports local model training (or in-house AI segmentation models)
  - Curate local institutional data for model training
  - Perform data augmentation if needed
  - Understand network architecture
  - Understand the impact of hyperparameters in model training

16

Benchmark Datasets

- Datasets with ground-truth contours that can be used for evaluating auto-segmentation tools
- Datasets for different anatomical sites
- Image type of the datasets
- Number of datasets for each anatomical site
- Source of benchmark datasets
  - Public benchmark datasets
    - Understand the inter-institutional variations
  - Local institutional benchmark datasets
    - Time for data curation

17

Common Evaluation Metrics

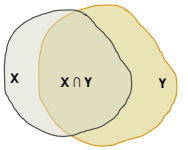
- Dice coefficient
  - A measure of relative overlap $D = \frac{2|X \cap Y|}{|X| + |Y|}$
- Mean surface distance
  - Measure average distance of two contours in millimeters

$$\vec{d}_{avg}(X,Y) = \frac{1}{|X|}\sum_{x \in X} \min_{y \in Y} d(x,y) \qquad d_{avg}(X,Y) = \frac{\vec{d}_{avg}(X,Y) + \vec{d}_{avg}(Y,X)}{2}$$

- Hausdorff distance (95% Hausdorff distance)
  - Measure the 95% distance of all points in one contour to the other in millimeters

$$\vec{d}_{H95}(X,Y) = K_{95}\left(\min_{y \in Y} d(x,y)\right) \qquad d_{H95}(X,Y) = \frac{\vec{d}_{H95}(X,Y) + \vec{d}_{H95}(Y,X)}{2}$$

$K_{95}(\cdot)$ is the 95th percentile



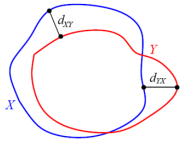


Image from wikipedia

18

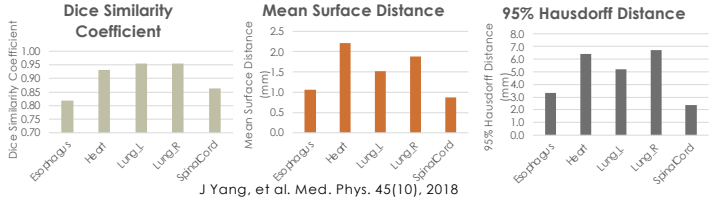
Pitfall of Evaluation Metrics

- Understand the meaning of evaluation metrics and their implementation details
- Understand that sensitivity of each evaluation metric to different organs
  - Dice value is more sensitive for small-volume structures than large-volume structures
  - Hausdorff distance is very sensitive to outliers
- Individual quantitative evaluation metrics may not directly relate to human perception

19

Account for Inter-Observer Variability

- Observer contouring variability
  - Manual contouring uncertainty could be as large as 2 cm (*IAEA Human Health Series No. 31, 2016*)
- To account for inter-observer variability, ground-truth can be manually contoured by multiple observers
- Mean score of these observers used as reference score (R)



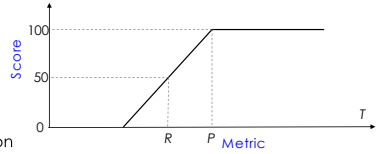
J Yang, et al. Med. Phys. 45(10), 2018

20

Account for Inter-Observer Variability

- Perfect measure (P): Dice = 1, MSD/HD95=0
- Normalization of the score (T) for each metric, each structure

$$Score = max\left(50 + \frac{T-R}{P-R} * 50\,, 0\right)$$



- Score = 100; perfect segmentation
- Score = 50; equivalent to average inter-observer reference
- Score =0; below the reference by more than the perfect score above the reference

- Final score is the average of normalized scores of all metrics structures

J Yang, et al. Med. Phys. 45(10), 2018

21

Workflow Integration Consideration

- Consider clinical applications
  - Treatment planning
  - Online adaptive planning
  - Outcome/toxicity analysis
- Consider integration with application software
  - Easy access the tool from application software, e.g., TPS
  - Easy retrieve segmentation results to application software
  - Parameter options for segmentation algorithms

22

Guidelines in Commissioning AI Segmentation Tools

23

Commissioning Guidelines

- Planning step
  - Read vendor supplied documents
  - Understand the AI segmentation tool
    - Support image types
    - Support anatomical sites
    - Does it require model training?
  - Assess the workflow integration with local clinical practice
    - Seamless integration with local treatment planning systems

24

Commissioning Guidelines

- Preparation step
  - Determine the anatomical sites and image types
  - Prepare local institutional benchmark datasets for each anatomical site and image type
    - Determine the number of test images (the more the better)
    - Collect manual contours as ground-truth from multiple experts if possible
    - Perform quality check of the manual contours
  - Choose/Implement quantitative evaluation tool
    - Select metrics to be used for the evaluation
    - Verify the proper functionality of the evaluation tool

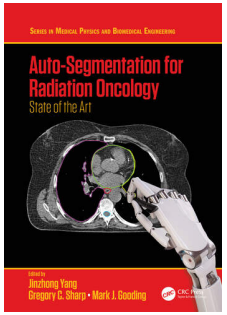
25

Commissioning Guidelines

- Testing and evaluation step
  - Perform segmentation for each image type and anatomical site in a clinical test environment
  - Generate quantitative evaluation metrics
  - Perform subjective evaluation
    - Score the auto-segmented contours by at least two experts in terms of the need for editing
  - Analyze the results to decide the organs/sites/images for clinical use
  - Clearly document the limitations on insufficient accuracy

26

Extended Reading



27