Karlsruher Institut für Technologie

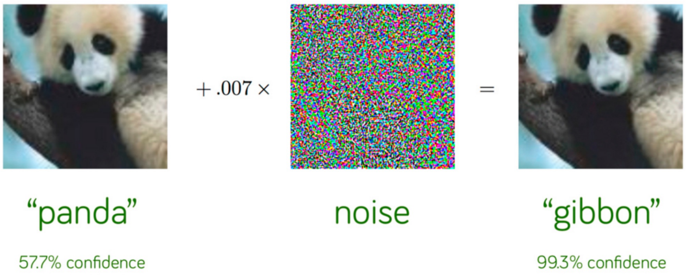
Measuring Robustness of Deep Learning Model for Head and Neck Tumor Volume Delineation

Chloe Griffin | August 2nd, 2023

Overview

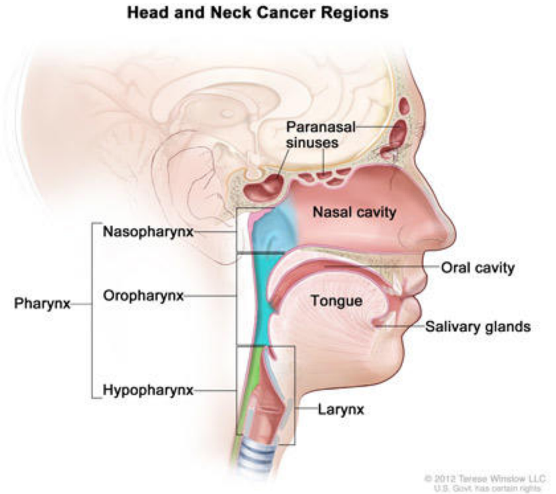
Summer Project

- Visiting from Brown University
- Two and a half month stay
- Advised by Alexandra Walter
- Work with trained nnU-Net model
  - Deep-learning biomedical segmentation method
  - Self-configuration
  - Trained for head and neck tumor volume delineation
- Measure neural network robustness (funny on right) [1]
- Develop procedure and analyze results



Tumor Volume Delineation

- Vital for head and neck cancer (HNC) radiotherapy
- Includes several sensitive regions (right) and near organs at risk [2]
- 800,000 new cases of HNC globally each year [3]
- Ionizing radiation damages DNA and destroys malignant cells
- Life-threatening postradiation changes [4]



Automating the Process

- Computer tomography (CT) scans determine tissue density and gross tumor volumes [5]
- Oncologists segment scans by hand
- Average of three hours per patient
- Results are highly subjective
- Vary from expert to expert
- Single clinician results not always consistent [6]

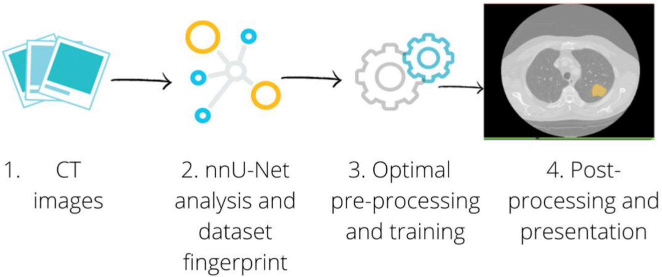
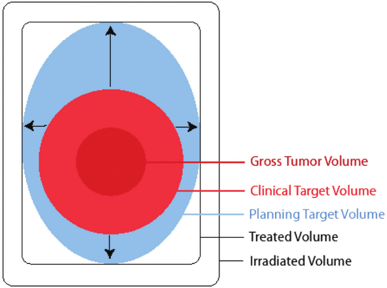


Figure: Example of automated tumor delineation pipeline [7]

Clinical Target Volume Delineation

- Gross Target Volume (GTV): Initial tumour volume
- Clinical Target Volume (CTV): GTV + volume with high probability of microscopic disease



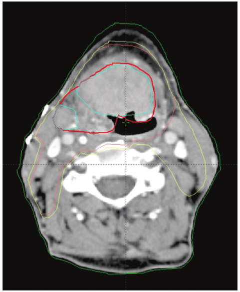


Figure: Schematic diagram of CTV (on left) and labeled slice with GTV in red and CTV in yellow (on right) [8]

nnU-Net

- “No new net:"
  - Primary advancement: automatically configures many aspects of the training process
  - Includes preprocessing, post-processing, and architecture structure
  - Methods and details can be found in [9]
- General segmentation tool
  - Divide an image key segments or isolate objects of interest
  - 2020, outperform other models due to automatic configuration
  - Task specific training may improve performance
  - Trained with Dice coefficent [9]

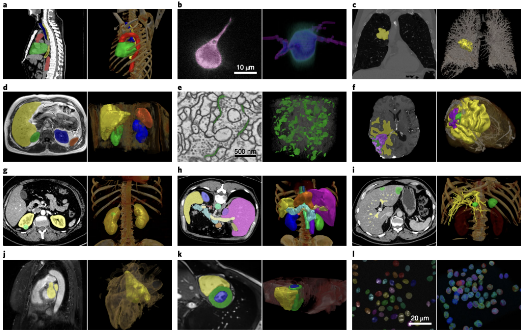


Figure: Examples of nnU-net applications

Data and Training

- Training and testing data was obtained from four cohorts
- 104 patient CT data sets for the model
- Split into 86 for training and 18 for testing
- Trained with original CT scan and manual CTV labels
- Scans were manually delineated at the German Cancer Research Center (DKFZ)

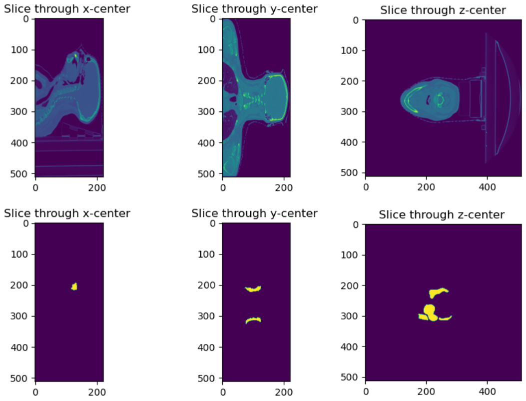


Figure: Example slices from CT scans and manual labels

Dice and sDice Metrics

- Dice:
  - Measures volumetric overlap
  - Ranges from 0 to 1
  - Insensitive to small deviations with large structures
  - Sensitive to image processing changes with small structures

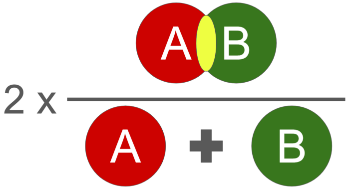


Figure: Formula for computing Dice metric [10]

- sDice:
  - Measures surface overlap
  - Ranges from 0 to 1
  - Penalizes border placement outside tolerance
  - Clinically significant for small deviations [4]

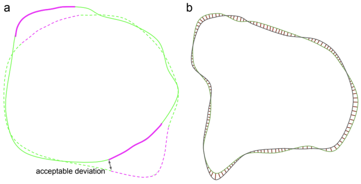


Figure: (a) Visual representation of sDice (b) Obtaining tolerance from oncologist manual labels [4]

Measuring Robustness

Karlsruher Institut für Technologie

Goal: Measure how output metrics change given input perturbations

General Approach:

1. Systematically rotate CT scans and labels by varying degrees
2. Feed rotated CT scan into the trained nn-Unet model
3. Compare prediction metrics after rotation with prediction metrics prior to rotation
4. Obtain measure of robustness using existing standards

Built Python pipeline:

1. Rotates CT scans and labels by desired degree,
2. Feeds rotated CT scan into the trained nn-Unet model,
3. Scores prediction with Dice and sDice metrics

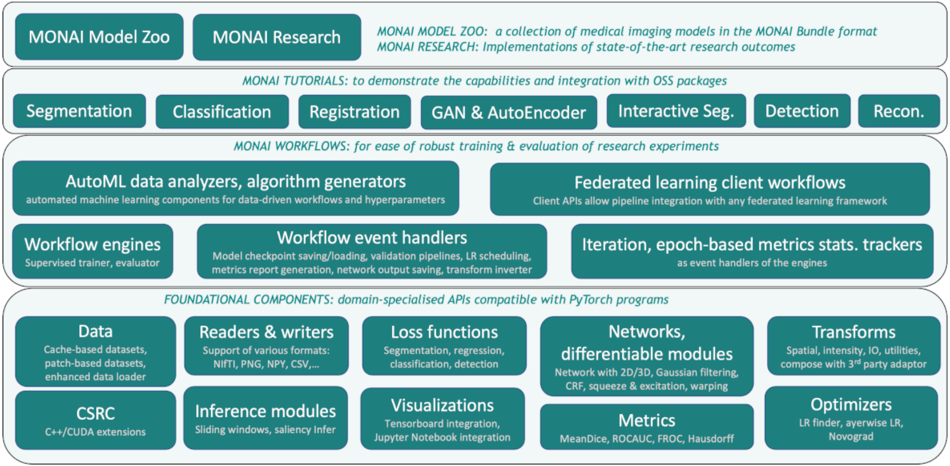
To-do:

1. Determine best approach for sampling rotations
2. Decide on most relevant standard for robustness

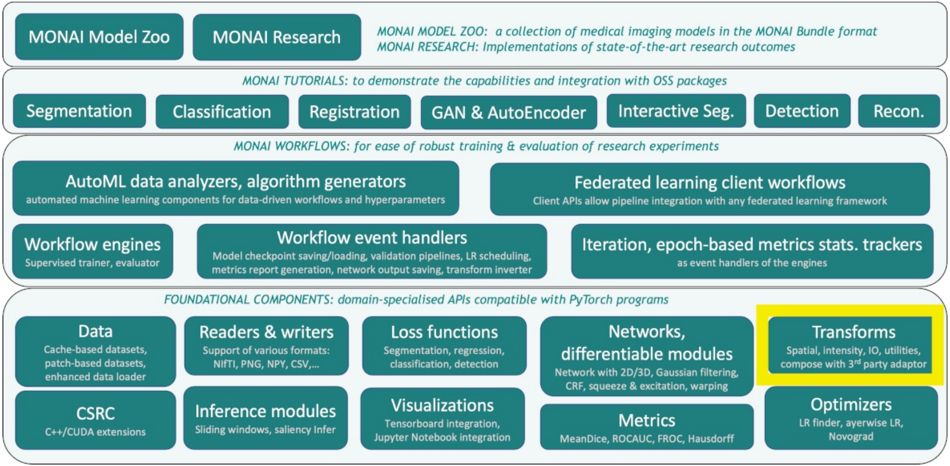
Introducing Monai

- "Medical Open Network for AI"
- Pytorch-based
- Open-source
- Deep learning Framework
- Emphasis on healthcare imaging
- Pytorch Ecosystem
- Started by NVIDIA and King’s College London
- 167 contributors [11]

Monai Applications and Use Cases



Monai Applications and Use Cases



Monai Transformations

- Many transformations found in Monai documentation
- Dictionary transformations
  - Create list of dictionaries with associated paths to stored image and label data
  - Use Compose function to load, transform, and save images
  - Create Monai dataset as input of PyTorch DataLoader and continue with training, inference, etc.

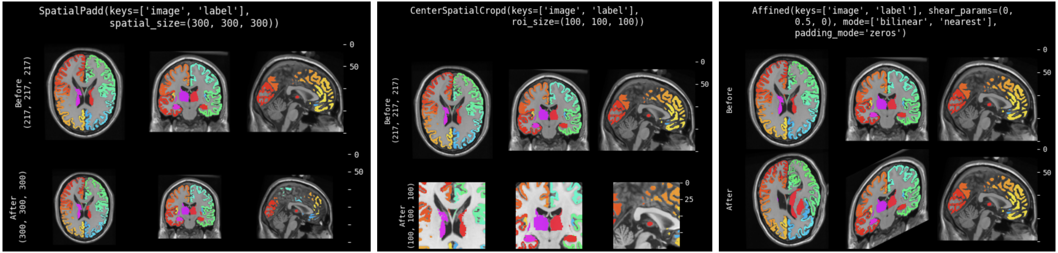


Figure: Examples of Monai Dictionary Transformations

Rotating 3D Images with Monai

1. Load image and label
2. Ensure Channel First
3. Use BorderPadd to ensure correct fill values for rotation
4. Affined rotation, padded with maximal spatial size for any rotation
5. Save image

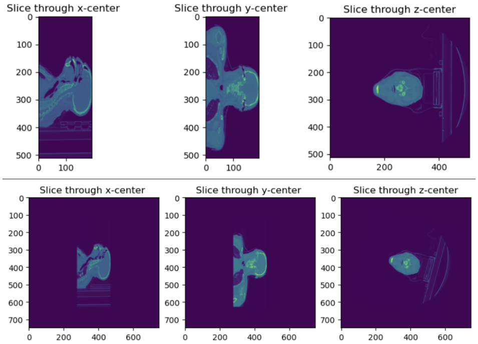


Figure: Example image before and after rotation with 15 degree rotation about z-axis

Defining Robustness

“Robustness measures the resilience of a system towards perturbations in any of its components"[12]

- Local Robustness
  - For a given input x, model provides same result for x and all inputs x' within ball of radius δ centered at x
- Global Robustness
  - Model is locally robust for all inputs in the input space [13]
- Adversarial Robustness
  - Worst-case scenario, model's ability to resist being fooled
- Probabalistic/Physically Real Robustness
  - Probability that difference in δ close inputs satisfy Lipschitz property is greater than $(1-\epsilon)$

$$\Pr_{x,x'\sim D}(||f(x') - f(x)|| \leq k * ||x' - x||) \mid ||x' - x|| \leq \delta) \geq 1 - \epsilon$$

  - Still difficult to verify, but only have to show true with probability of at least $(1-\epsilon)$ with respect to realistic input distribution

Proposed Inference Protocols

- Mean Absolute Error of Dice and sDice as robustness metric

$$DiceMAE(\epsilon) = \frac{\sum_{i=1}^{n}(|O^i_{Dice} - AR^i_{Dice}(\epsilon)^i|)}{n}$$

$$sDiceMAE(\epsilon) = \frac{\sum_{i=1}^{n}(|O^i_{sDice} - AR^i_{sDice}(\epsilon)|)}{n}$$

  where n is number of patient samples, O^i_{Dice} is score of a given sample prior to rotation, AR^i is score of the given sample after rotation, and $\epsilon = \begin{bmatrix} a & b & c \end{bmatrix}$ is the rotation array about (x, y, z)
- Plot with respect to increasing rotation around each axis
- Use Wilcoxon signed-rank tests (non-parametric alternative to t-test) to see if selected pertubations lead to significantly different Dice scores across samples as in [14].

Open Questions

1. Should we focus on realistic cases for robustness test? Or take adversarial approach?
2. Is there an exact criterion to answer “is this neural network robust"? Is this too problem dependent?
3. If adversarial, should we stick with rotations across one axis at a time? Or try rotating across several dimensions at once since we have the capacity?

Questions?

References I

Why Robustness is not Enough for Safety and Security in Machine Learning | by Christian Kästner | Towards Data Science.
URL `https://towardsdatascience.com/why-robustness-is-not-enough-for-safety-and-security-in-machine-learning-1a35f6706601`

L. Q. Chow, Head and Neck Cancer, New England Journal of Medicine 382 (1) (2020) 60–72.
`doi:10.1056/NEJMRA1715715`.
URL `http://www.nejm.org/doi/10.1056/NEJMra1715715`

D. Kawakita, I. Oze, S. Iwasaki, . Tomohiro Matsuda, . Keitaro Matsuo, H. Ito, Trends in the incidence of head and neck cancer by subsite between 1993 and 2015 in Japan, Cancer Medicine 11 (2022) 1553–1560.
`doi:10.1002/cam4.4539`.
URL `https://onlinelibrary.wiley.com/doi/10.1002/cam4.4539`

References II

S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, C. Kelly, A. Karthikesalingam, C. Chu, D. Carnell, C. Boon, D. D'Souza, S. A. Moinuddin, B. Garie, Y. McQuinlan, S. Ireland, K. Hampton, K. Fuller, H. Montgomery, G. Rees, M. Suleyman, T. Back, C. Hughes, J. R. Ledsam, O. Ronneberger, Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy (9 2018).
URL `http://arxiv.org/abs/1809.04430`

N. G. Burnet, S. J. Thomas, K. E. Burton, S. J. Jefferies, Defining the tumour and target volumes for radiotherapy, Cancer Imaging 4 (2) (2004) 153.
`doi:10.1102/1470-7330.2004.0054`.
URL `/pmc/articles/PMC1434601//pmc/articles/PMC1434601/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1434601/`

References III

B. Huang, Y. Ye, Z. Xu, Z. Cai, Y. He, Z. Zhong, L. Liu, X. Chen, H. Chen, B. Huang, 3D Lightweight Network for Simultaneous Registration and Segmentation of Organs-at-Risk in CT Images of Head and Neck Cancer, IEEE Transactions on Medical Imaging 41 (4) (2022) 951–964. doi:10.1109/TMI.2021.3128408.

M. Ferrante, L. Rinaldi, F. Botta, X. Hu, A. Dolp, M. Minotti, F. De Piano, G. Funicelli, S. Volpe, F. Bellerba, P. De Marco, S. Raimondi, S. Rizzo, K. Shi, M. Cremonesi, B. A. Jereczek-Fossa, L. Spaggiari, F. De Marinis, R. Orecchia, D. Origgi, Application of nnU-Net for Automatic Segmentation of Lung Lesions on CT Images and Its Implication for Radiomic Models, Journal of Clinical Medicine 11 (24) (2022) 7334. doi:10.3390/JCM11247334/S1.
URL https://www.mdpi.com/2077-0383/11/24/7334/htmhttps://www.mdpi.com/2077-0383/11/24/7334

References IV

- A. K. Berthelsen, J. Dobbs, E. Kjellén, T. Landberg, T. R. Möller, P. Nilsson, L. Specht, A. Wambersie, What's new in target volume definition for radiologists in ICRU Report 71? How can the ICRU volume definitions be integrated in clinical practice?, Cancer Imaging 7 (1) (2007) 104–116. `doi:10.1102/1470-7330.2007.0013`.
- F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211. `doi:10.1038/s41592-020-01008-z`.
- The Difference Between Dice and Dice Loss - PYCAD. URL `https://pycad.co/the-difference-between-dice-and-dice-loss/`
- Project MONAI — MONAI 0 Documentation. URL `https://docs.monai.io/en/latest/index.html`

References V

- R. Fabra-Boluda, C. Ferri, F. Martínez-Plumed, M. José Ramírez-Quintana, Robustness Testing of Machine Learning Families using Instance-level IRT-Difficulty (2022).
- R. Mangal, A. V. Nori, A. Orso, Robustness of Neural Networks: A Probabilistic and Practical Approach.
- Z. Liu, J. Zhang, V. Jog, Po, L. Loh, A. B. Mcmillan, Robustifying Deep Networks for Medical Image Segmentation, Journal of Digital Imaging 1 (2021) 3. `doi:10.1007/s10278-021-00507-5`. URL `https://doi.org/10.1007/s10278-021-00507-5`