Does the Choice of Deep Learning Architecture Matter?

Experience from a Radiotherapy Case Study

Skylar S. Gay[1], Anuja Jhingran[1], Brian M. Anderson[1], Lifei Zhang[1], Dong Joo Rhee[1], Callistus Nguyen[1], Tucker Netherton[1], Jinzhong Yang[1], Kristy Brock[1], Hanna Simonds[2], Ann Klopp[1], Beth Beadle[3], Kelly D. Kisling[4], Laurence E. Court[1], Carlos E. Cardenas[1]

[1] Department of Radiation Physics, Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston TX
[2] Stellenbosch University, Stellenbosch, ZA
[3] Stanford University, Stanford, CA
[4] UC San Diego, La Jolla, CA

INTRODUCTION

- Deep Learning-based models are becoming widely used for target segmentation
- However there is little consensus on optimal training parameters

AIMS

- Extensively evaluate deep learning architectures commonly used for medical image segmentation
- Determine appropriate model choice for target segmentation tasks
- Identify influence of a wide range of hyperparameter and the optimal choices

METHODS

Data Curation

- Four-field box female pelvis cases (n=310)
- Models trained to delineate radiotherapy field apertures
- 2D anterior-posterior (AP), posterior-anterior (PA) and lateral DRRs
- 229 training and 26 validation cases
- 55 cases never seen by models during training and reserved for final evaluations

METHODS (continued)

Models

- Five commonly-used architectures selected as base models
  - DeepLab v3+[1]
  - D-LinkNet[2]
  - VGG-19 + U-Net style decoder[3]
  - U-Net[4]
  - Residual U-Net[5]
- Hyperparameter variations:
  - Learning rate: 0.01, 0.001, 0.0001
- Input normalizations
  - Z-score
  - Minimum-maximum value cropping
  - L-2
- Additional variations for U-Net and Residual U-Net
  - Network depth: 3, 4, 5, and 6 levels
  - Convolution kernel size: 3x3 and 5x5
  - First-level features: 16, 32, 48, and 64

Training

- Total of 1295 unique models trained and evaluated
- 23,000 computing hours

Evaluation Metrics

- Dice similarity coefficient (DSC)
- Composite of 5 overlap and distance scores[6]

RESULTS

- All models except VGG-19 achieved similar top DSC and composite scores
- Learning rate of 0.001 or lower was observed to be the most important hyperparameter contributed to good model convergence and performance
- Z-score intensity normalization similarly contributed to best model performance
- When evaluating model robustness using 25th percentile DSC and composite scores, these models performed best:
  - DeepLab v3+
  - Residual U-Net
- Model training time before convergence varied greatly for best models
  - > 20 hours for Deeplab v3+
  - 24 - 60 hours for Residual U-Net

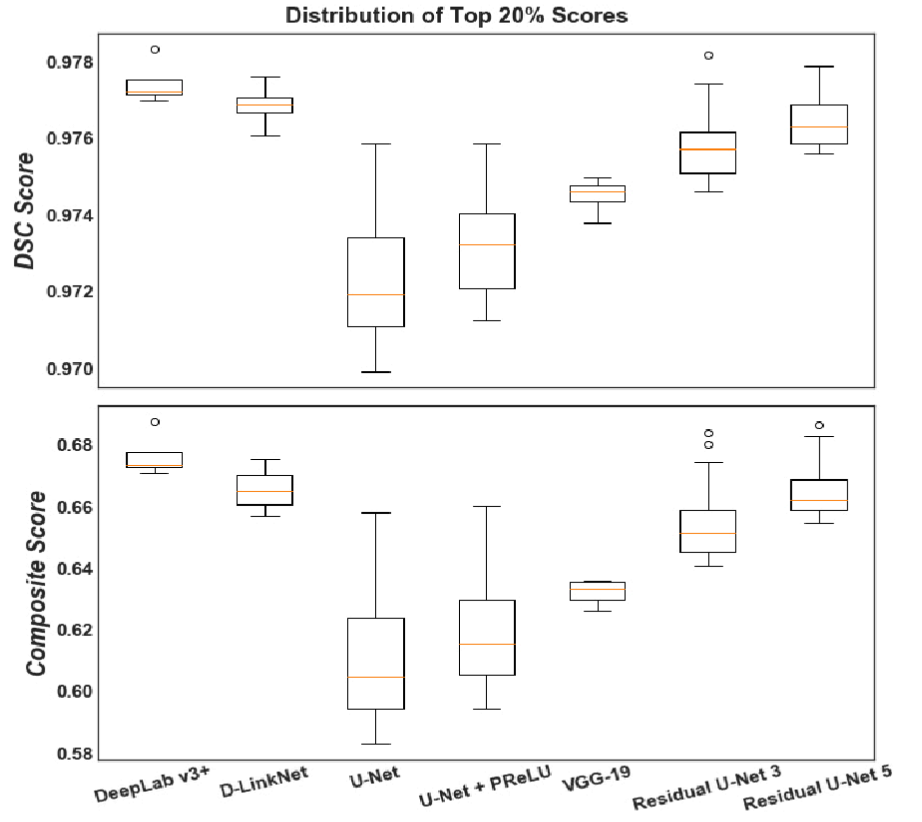

Figure 1: Distribution of 20% highest scores (DSC and composite) for all architecture variations evaluated. While all architectures were able to achieve similar best values, DeepLabv3+ and Residual U-Net with 5x5 kernel were most robust to initial hyperparameter selection.

Table 1: Model variations and performance. Maximum score represents best performance for each architecture. 25th percentile represents relative sensitivity to initial hyperparameters

Model (kernel size)	Variations	Top DSC	Top Composite	25th DSC	25th Composite
DeepLab v3+	21	0.978	0.687	0.97	0.571
D-LinkNet	21	0.978	0.675	0.968	0.524
U-Net	336	0.976	0.658	0.969	0.532
U-Net + PReLU	224	0.976	0.660	0.968	0.540
VGG-19 + U-Net	21	0.975	0.636	0.966	0.494
Residual U-Net (3x3)	336	0.978	0.684	0.971	0.578
Residual U-Net (5x5)	336	0.978	0.687	0.972	0.582

CONCLUSIONS

- Given appropriate, model-specific hyperparameters, most commonly-used models can approach acceptable convergence
- Too-high learning rate was the single largest contributor to poor model performance
- Residual U-Net was overall best for our dataset but much slower to training than similar-performing Deeplab v3+

REFERENCES

1. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11211 LNCS*, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
2. Zhou, L., Zhang, C., & Wu, M. (2018). D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2018-June*, 192–196. https://doi.org/10.1109/CVPRW.2018.00034
3. Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. http://arxiv.org/abs/1409.1556
4. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9351*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
5. Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, 565–571. https://doi.org/10.1109/3DV.2016.79
6. Yang, J., Veeraraghavan, H., Armato, S. G., Farahani, K., Kirby, J. S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., Aljabar, P., Oliveira, B., van der Heyden, B., Zamdborg, L., Lam, D., Gooding, M., & Sharp, G. C. (2018). Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Medical Physics, 45*(10), 4568–4581. https://doi.org/10.1002/mp.13141

Disclosures

Our research group receives funding from the NCI and Varian Medical Systems.

Contact information SGay1@mdanderson.org

Acknowledgements: The Courtyard Research Group, MD Anderson High Performance Computing Center Team