Comparison of Deep Learning-Based Auto-Segmentation Algorithms for Head and Neck cancer

Min Seo Choi, Jaehee Chun, Nalee Kim, Jee Suk Chang, Jin Sung Kim

Department of Radiation Oncology, Yonsei University College of Medicine, Seoul, Korea

Purpose: Accurate delineation of the organs at risk and targets is a crucial part of head and neck (H&N) cancer radiation therapy planning. Increased toxicity to the surrounding healthy tissues must be avoided in order to reduce the likelihood of side effects such as xerostomia and mucositis which can severely affect the patients' quality of life following the treatment [1]. However, it is well-known that manually segmenting normal and target structures in this site is especially challenging due to a large number of OARs near H&N tumors, making it very laborious and subject to inter-observer variations [2]. In order to overcome this issue and the need for manual segmentation, many previous studies on H&N auto-segmentation have been conducted [3]–[7]. The aim of our study is to compare and evaluate the state-of-the-art deep learning based auto-segmentation (DLS) approaches.

Materials and Methods: The CT datasets used in this study consisted of 115 simulation CT scans of head and neck cancer patients and manual delineations of ten organs at risk (OAR) drawn by a single expert: brainstem, spinal cord, parotids, oral cavity, submandibular glands, thyroid, mandible, and esophagus. The data was split into 80 training sets, 10 validation sets, and 25 test sets. An in-house segmentation software based on deep learning was developed. It is based on three-dimensional fully convolutional DenseNet comprised of localization and regions of interest-specific segmentation steps and made up of dense blocks each containing [3,4,4,5 and 7] layers. In this study, we also included two additional commercial DLS software packages: AccuContour (Manteia tech, Xiamen, China) and DLCExpert (Mirada Medical, Oxford, United Kingdom). These packages were pre-trained by the manufacturer and hence did not utilize our training datasets, and were only used for evaluation and comparison purposes. Lastly, the Dice similarity coefficient (DSC) was used to compare the similarity of the auto-segmentation to the manually expert generated contours.

Results: Our in-house software's DSCs ranged from 81% similarity in the esophagus and up to 95% similarity in the mandible and produced the highest average DSCs in every OARs except for the right submandibular gland and esophagus. The two commercial DLS packages showed an average DSC of over 70% across most structures. AccuContour outperformed DLCExpert in eight out of 10 structures. DLCExpert scored 60% and 22% in the spinal cord and esophagus respectively, but this is likely to be the result of being trained on an external dataset with different structure definitions, possibly with different starting and ending points.

Conclusion: In summary, we have, for the first time, compared multiple DLS algorithms in the head and neck region. Through similarity evaluation, we have found that our in-house model produced the highest similarity to the ground truth. This further implies our model mimicked the drawing style of the expert the best out of the three algorithms. Our results also highlight the importance of on-site training with institution-specific datasets to ensure better segmentation outcomes from a DLS algorithms.

Keyword: Auto-segmentation, Deep learning-based segmentation