Segmentasi Organ-at-risk Otomatis pada Gambar CT Scan 3 Dimensi Bagian Toraks menggunakan Metode 3D Convolutional Neural Network = Automatic Organ-at-risk Segmentation for Thoracic 3 Dimensional CT Scan Images using 3D Convolutional Neural Network Method

Wahyu Hutomo Nugroho, author

Abstrak

Proses segmentasi organ secara manual memakan waktu dan hasilnya subyektif terhadap definisi batas-batas kontur. Pemanfaatan teknologi Machine Learning (ML) berjenis 3D convolutional neural network (3D CNN) untuk mensegmentasi organ secara otomatis dapat mempercepat dan menstandarisasi hasil segmentasi organ. Penelitian ini mengimplementasilan network ML berbasis VoxResNet dan memanfaatkan 60 dataset CT Scan toraks dari Grand Callenge AAPM 2017 untuk melatih, memvalidasi, dan menguji model-model ML dengan berbagai variasi hyperparameter. Pengaruh variasi hyperparameter terhadap hasil segmentasi model juga dipelajari. Dataset dibagi menjadi 3 yaitu, 36 untuk perlatihan, 12 untuk validasi, dan 12 untuk pengujian. Dalam penelitian ini paru-paru kiri dan paru-paru kanan dijadikan satu jenis OAR bernama paru-paru, esophagus dan spinal cord dijadikan satu OAR bernama ESP, sedangkan jantung tetap OAR tersendiri. Variasi hyperparameter adalah variasi ukuran patch, jumlah batch, dan weight class. Hasil segmentasi model-model dievaluasi dan diperbandingkan untuk mencari model terbaik dengan hyperparameter-nya yang mampu menghasilkan kualitas hasil segmentasi organ terbaik. Kemampuan network dalam proses perlatihan dan validasi dievaluasi menggunakan kurva pembelajaran. Kualitas hasil segmentasi model organ dievaluasi menggunakan boxplot distribusi populasi nilai metrik Dice Similiarity Coefficient (DSC) dan Housdorf Distance (HD) setiap slice. Peningkatan atau penurunan kinerja model akibat variasi hyperparameter dinilai menggunakan skor peningkatan metrik. Terakhir, metrik DSC dan HD95 secara 3D hasil segmentasi model terbaik dibandingkan dengan hasil segmentasi oleh interrater variability AAPM 2017 dan hasil segmentasi team virginia. Hasil kurva pembelajaran tidak mengalami underfitting menunjukkan bahwa network mampu mempelajari data perlatihan dengan baik. Overfitting terjadi pada model organ jantung dan ESP. Hasil eksperimen variasi ukuran patch menunjukkan bahwa besar ukuran patch tidak selalu linier dengan kinerja moukuran patch menunjukkan bahwa besar ukuran patch tidak selalu linier dengan kinerja model. Model ukuran patch tengah memberikan kualitas distribusi metrik dan skor paling baik dibandingkan model ukuran patch terkecil dan terbesar pada semua OAR dengan skor 11, 13, dan 13 dari 16. Hasil eksperimen variasi jumlah batch menunjukkan bahwa peningkatan jumlah batch tidak selalu berdampak positif terhadap kinerja model. Untuk model jantung ukuran patch terbesar, peningkatan batch dapat meningkatkan skor dari 2 menjadi 12. Untuk model ESP ukuran patch terbesar, peningkatan batch menurunkan skor dari 13 menjadi 2. Hasil eksperimen variasi weight class (W) menunjukkan bahwa baik model jantung maupun ESP cenderung memberikan distribusi metrik dan skor terbaik di sekitar W = [1,3.67] atau W = [1, C1 < 11]. Dibandingkan dengan interrater variability AAPM, model jantung terbaik menghasilkan nilai metrik yang comparable, yaitu untuk DSC 3D 0.932 ± 0.016 = 0.931 ± 0.015 dan untuk HD95 4.00 ± 0.25 < 6.42 ± 1.82. Sedangkan untuk model paru-paru memberikan metrik lebih baik, yaitu 0.964 ± 0.025 > 0.956 ± 0,019 dan 4,72± 0,21 < 6.71 ± 3,91. Dibandingkan dengan team virginia, model

jantung terbaik berhasil memberikan nilai metrik yang lebih baik. yaitu 0.932 ± 0.016 > 0.925 ± 0.015 dan 4.00 ± 0.25 < 6.57 ± 1.50, sedangkan model ESP terbaik menghasilkan metrik yang comparable, yaitu 0.815 ± 0.049 = 0,810 ± 0,069 dan 4,68 ± 0,17 < 8,71 ± 10,59. Dari hasil-hasil ini memberikan potensi adanya perpaduan ukuran patch, jumlah batch, dan weight class tertentu yang dapat menyebabkan hasil segmentasi model ukuran patch lebih kecil dapat mengimbangi hasil segmentasi model ukuran patch lebih besar sehingga tuntutan akan perangkat dengan spesifikasi tinggi dan mahal dapat berkurang.
.....The process of manual organ segmentation is time consuming and the results are subjective in term of definition of contour boundaries. The utilization of Machine Learning (ML) technology using 3D convolutional neural network (3D CNN) to segment organs automatically can speed up the procces as well as standardizing the results of organ segmentation. This study implements a VoxResNet-based ML network and utilizes 60 thoracic CT scan datasets obtained from Grand Callenge AAPM 2017 to train, validate, and test ML models with various hyperparameter variations. The effects of hyperparameter variations on the segmentation results of models are also studied. The dataset is divided into 3 parts, namely 36 for training, 12 for validation, and 12 for testing. In this study the left lung and right lung were combined into one type of OAR called the lung, the esophagus and spinal cord were combined into one OAR called ESP, while the heart remained a separate OAR. Hyperparameter variations are variations in patch size, number of batches, and weight loss. The segmentation results of the models are evaluated and compared each other to find the best model and it’s hyperparameters which is able to produce the best segmentation’s quality. The ability of the network in training and validation procceses is evaluated using learning curve. The quality of the organ model’s segmentation results is evaluated using boxplot of population’s distribution of the Dice Similiarity Coefficient (DSC) and Housdorf Distance (HD) metrics for each slice. The increases or decreases in model performance due to variations in hyperparameters are assessed using the metric improvement score. Finally, the 3D DSC and HD95 metrics of the best model’s segmentation results are compared to the results of segmentation by the AAPM 2017’s interrater variability and to the segmentation results by team virginia. There is no underfitting of learning curve indicates that the network is able to learn the training data. Overfitting occurs in the heart and ESP models. The experimental results from patch size variations show that the size of the patch is not always linear with the performance of the model. The middle patch sized models give the best metric distribution’s quality as well as scores compared to the smallest and largest patch sized models for all OARs with scores of 11, 13, and 13 out of 16. The experimental results from batch number variations show that an increase in batch does not always have a positive impact on model performance. For the largest patch sized heart’s model, the increase increases the score from 2 to 12. For the largest patch sized ESP's model, the increase reduces the score from 13 to 2. The results from variations in weight loss (W) experiment show that both heart’s and ESP's models tend to provide the best distributions in term of metrics and scores around W = [1, 3.67] or W = [1, C1 < 11]. By comparing with AAPM's interrater’s variability, the best heart model produces comparable metric's result, that is 0.932 ± 0.016 = 0.931 ± 0.015 for DSC 3D and 4.00 ± 0.25 < 6.42 ± 1.82 for HD95. The best lungs model produces better metrics, that is 0.964 ± 0.025 > 0.956 ± 0,019 and 4,72 ± 0,21 < 6.71 ± 3,91. By comparing with team virginia's results, the best heart model produces better results that is 0.932 ± 0.016 > 0.925 ± 0.015 and 4.00 ± 0.25 < 6.57 ± 1.50. Meanwhile the best ESP model produces comparable results that is 0.815 ± 0.049 = 0,810 ± 0,069 and 4,68 ± 0,17 < 8,71 ± 10,59. The results of this study suggests that there is a certain combination of patch size, batch, and weight class by which enables smaller patch sized model to produce comparable metric's result produced by larger patch sized model thus decreasing the need to use higher

specificationed and expensive computer.