3D U-Net Architectures for Multimodal Brain Tumor Segmentation

Pranav Gurusankar, Jimming He, Alexander Kwon

Department of Computer Science, Stanford University

Introduction

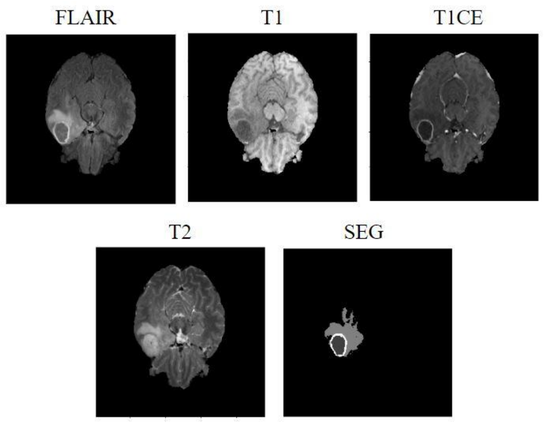
- Gliomas: brain tumor occurring in the central nervous system
  - Glioblastoma: most common malignant brain tumor, 5-year survival is < 7%
- Identification involves magnetic resonance imaging (MRI) but has limitations
  - Gliomas are heterogenous with irregular shape, size, location, histology → difficult to segment glioma subregions
  - Manual segmentation by radiologists is time-consuming and costly
  - Solution: automated segmentation using deep learning!

Problem Statement

- Inspired by the Brain Tumor Segmentation (BraTS) 2020 challenge
- **We will evaluate three different 3D U-Net-based architectures to automate segmentation of glioma subregions in multimodal brain tumor MRI images:**
  - Enhancing tumor (ET): hyperintensity compared to healthy white matter
  - Tumor core (TC): ET + necrotic and non-enhancing tumor (NCR/NET)
  - Whole tumor (WT): TC + peritumoral edema (ED)
- **Input**: MRI regions in NIfTI format (neuroimaging file format)
- **Output**: Segmentation masks labelling each glioma subregion
- **Evaluation**: Dice Similarity Coefficient (DSC) and Hausdorff distance (95%)

Dataset

- BraTS 2020 challenge dataset: publicly available
  - Training set: 369 glioma cases [269 high-grade GBM, 76 low-grade GBM]
- BraTS-preprocessing
  - Image size: 240 x 240 x 155
  - Skull-stripped and resampled to 1 mm^3 resolution (voxel size)
- Our preprocessing
  - Normalized four modalities to zero mean and unit standard deviation
  - Minimal disruptions
    - Random flip on three axes with 50% probability each
    - Random 90º rotation on two axes with 50% probability each
    - Random intensity between (-0.1, 0.1) for standard deviation
    - Random intensity on all input channels between (0.9, 1.1)

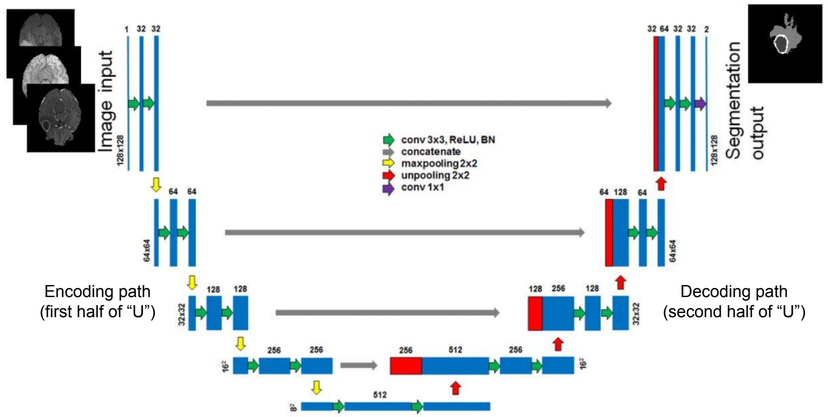


Each case had five total image types (4 'modalities' + seg. mask):

1. Fluid attenuation inversion recovery (FLAIR)
2. T1 weighting (T1)
3. T1-weighted contrast-enhanced (T1-CE)
4. T2 weighting (T2)
5. Manually-labeled segmentation mask labeling all glioma subregions

Methods

- Considered all possible approaches for a semantic segmentation task: sliding window approach, convolutional layers with no downsampling, etc.
- Needed both downsampling and upsampling, therefore: **3D U-Net**
  - *Encoding path:* convolutional layers with downsampling operators
    - Extract features, reduce spatial resolution
  - *Bottleneck layer:* convolutional layers for most abstract, high-level features
  - *Decoding path:* convolutional layers with upsampling operators
    - Combine high-level features with local ones, recover spatial resolution
  - *Skip connections:* concatenate encoding and decoding path information
  - *Non-linearity:* probability map for each class



General 3D-UNet architecture. Image source: https://arxiv.org/abs/1505.04597v1.

- **Loss function:** Generalized Dice Loss (GDL), multi-class version of Dice Score

$$L_{GDL} = 1 - 2\frac{\sum_{l=1}^{L} w_l \sum_{i=1}^{N} p_{li} g_{li}}{\sum_{l=1}^{L} w_l \sum_{i=1}^{N} p_{li} + g_{li} + \epsilon}$$

L is number of classes, w_l is weight for each class, N is voxels per image, p_{li} and g_{li} are voxel's predicted and ground truth labels, respectively, and ϵ prevents division by zero.

Experiments

- Three models: **baseline**, **residual**, and **ensemble** (Adam optimizer, 100 epochs)
- **Baseline**: build from basic 3D-UNet architecture
  - Group Normalization instead of Batch Normalization
  - 32 feature maps at finest resolution
  - Patch size of 112^3 and batch size of 2
- **Residual**: residual block with a skip connection to learn residual mappings
  - Five layers (32 channels, 160 x 160 x 160 voxels), two res. blocks each
  - 15% channel dropout b/n res. blocks [which contain Conv blocks]
  - Residual output added to convolved inputs, concatenation, σ activation
- **Ensemble**: averages multiple 3D-UNets' outputs to yield more robust predictions
  - 10-fold cross-validation strategy using mean to predict final output
  - Weighted contours and randomized input volumes
  - Preserved best-performing cross-validation models using GDL
  - Averaged ensemble predictions compared to best cross-validation models

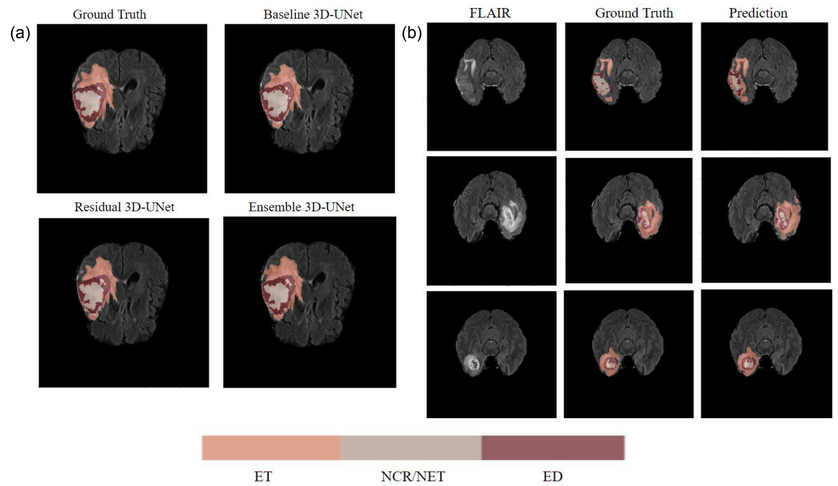
Results, Analysis, and Discussion

(a) $DSC = 2 \cdot \frac{X \cap Y}{|X| + |Y|}$ (b) $D_H(X,Y) = \max(\sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y))$

Let X be the predicted segmentation mask and Y be the ground-truth segmentation mask. (a) DSC measures the proportion of overlap between X and Y. (b) Hausdorff distance (95%) measures dissimilarity between a set of points in X and Y. It takes the maximum distance from each point in one set to the closest point in the other set. Considering the 95th percentile discounts outliers. Note: we use mm as our unit of measure.

Method	Dice			Hausdorff (mm)		
	WT	TC	ET	WT	TC	ET
Baseline 3D-UNet	0.77	0.74	0.63	12.75	18.14	43.02
Residual 3D-UNet	0.77	0.74	0.67	11.45	13.40	34.18
Ensemble 3D-UNet	0.80	0.75	0.68	10.55	11.88	37.24

Metrics on the test dataset with 166 data points. We want large DSC and small Hausdorff distance.



(a) Representative masks for ground truth and three 3D-UNets. (b) Sample test results for ensemble 3D-UNet.

- Residual 3D-UNet learns long-range dependencies to outperform the baseline
- Ensemble performed the best based on DSC and Hausdorff distance
  - DSC and Hausdorff scores worst on ET because it's the most precise region
- Baseline and residual masks are more ragged/cut-off than that of the ensemble
  - Ensemble averages predictions to smooth differences but may over-smooth
- Fringe "islands" on the edges of ground truth scans could be a failure mode

Conclusions and Future Work

- Results not near the state-of-the-art or what is acceptable in the medical realm
  - Deep learning will be auxiliary, for now, and will continue to improve
- Ensemble 3D-UNet leverages the best of multiple 3D-UNets → robustness
- Future work will explore greater diversity of models
  - Attention U-Net: can attend to target structures of varying shapes and sizes
  - V-Net: like U-Net, but uses volumetric rather than standard convolution
  - Generative models (VAEs, GANs): capture complex distribution of MRI data