Fourier Neural Operator on HaN-Seg

Shantanu Kapoor

University of Florida

Abstract. In this study, we use FNOSeg3D, a 3D medical image segmentation model that addresses the challenges associated with conventional Convolutional Neural Networks (CNNs) such as high computational complexity, memory constraints, and sensitivity to spatial resolution. Traditional methods often train CNNs with downsampled images to mitigate out-of-memory errors during training, but this approach leads to suboptimal performance due to standard spatial convolution being sensitive to variations in image resolution. The solution leverages the Fourier Neural Operator (FNO), a deep learning framework designed for learning mappings between functions in partial differential equations, to achieve zero-shot super-resolution and global receptive field properties. The FNO is enhanced by decreasing its parameter requirement and improving its learning capability via residual connections and deep supervision. This model was tested on HaN-Seg dataset which resulted in a parameter-efficient and resolution-robust model. It was on par with state-of-the-art models in terms of robustness to training image resolution while utilizing less number of model parameters. The deployment of FNOSeg3D in remote areas with limited resources offers benefits including faster processing times and consistency across different imaging equipment. Overall, FNOSeg3D holds promise for enabling accurate and efficient medical image analysis, even in challenging environments where resources are constrained.

Keywords: Fourier Neural Operator · image segmentation · Deep Learning · Fourier Transform

1 Introduction

With the emergence of convolutional neural networks (CNNs), the precision and speed of medical image segmentation have experienced significant advancements. However, due to the computationally intensive nature of CNNs, encountering out-of-memory errors in GPU is prevalent during 3D image segmentation. Compared to 2D segmentation, 3D segmentation demands greater resources since the quantity of feature elements per layer can be several magnitudes larger.

To tackle out-of-memory errors, two commonly adopted techniques involve downsampling images and implementing patch-wise training to decrease input image dimensions. Nonetheless, both methods come with certain drawbacks. Spatial convolution exhibits heightened sensitivity towards variations in image resolution, causing CNNs trained with downsampled images to yield subpar outcomes upon application on original image resolution. While patch-wise training maintains the native image resolution, selecting small patch sizes may substantially limit the receptive field. Additionally, requiring post-processing to amalgamate patch-wise predictions introduces further complications. Thus, distinct compromises accompany varying strategies.

Despite efforts to minimize input sizes, the computational intricacy inherently impedes CNNs' capacity to discern lengthy range spatial interdependencies within 3D CNNs. Given that memory utilization scales proportionately with model breadth and depth, filter and layer quantities must remain considerably lower in 3D segmentation tasks. Such restrictions not only compromise the degree of abstraction but also hinder optimal receptive fields owing to locally confined reception fields of convolutional layers

As a potential solution, researchers have begun incorporating transformers into medical image segmentation to overcome these obstacles. By partitioning images into smaller patches or employing low-resolution feature pixel values, sequences are generated and fed into transformer's multi-headed attention modules to decipher complex spatial correlations. Despite promising initial findings, the substantial computational needs of transformers render size reduction indispensable.

This study aims to study a robust 3D segmentation model capable of producing precise results over images characterized by higher resolutions compared to those utilized during training. Dubbed FNOSeg3D, relies on the Fourier neural operator (FNO), a deep learning paradigm engineered to learn mappings among functions rooted in partial differential equations (PDEs). Built around continuous space concepts encapsulated by Green's function, FNO boasts impressive traits like zero-shot super-resolution and expansive receptive fields. This work encompass the following

- Adapting FNO for computationally expensive 3D medical image segmentation, accompanied by diminished parameter count. The presented FNOSeg3D demonstrates remarkable robust performance with a drastically reduced parameter set relative to existing deep learning 3D segmentation counterparts like nnUNet.
- Testing FNOSeg3D on the Head and Neck Organ-At-Risk CT and MR Segmentation challenge. Upon training at downsampled images, the suggested technique yields a mean Dice coefficient measuring 0.413, comparable to rival offerings of about .533 yet maintaining just 30k parameters.

2 Methodology

2.1 Operator Learning

Operator learning establishes mappings between function spaces. It discovers a non-linear relationship between two infinite-dimensional spaces using finite collections of input-output (I/O) pairs. Crucial for resolving Partial Differential Equations (PDEs), operator learning seeks to identify an inverse operator for a provided PDE, Lu = v, where v and u represent functions. Applying machine learning algorithms, we approximate the elusive inverse operator of L linking input and output spaces, revealing hidden relationships governing the transformation of functions.

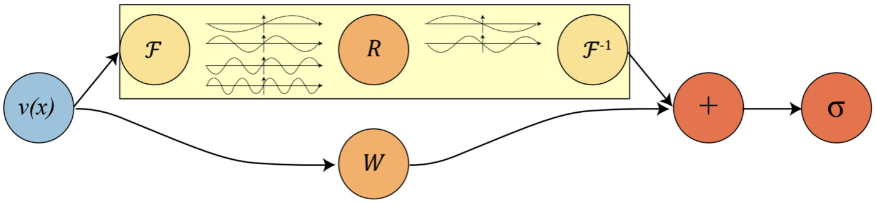


Fig. 1. The fourier layer consists of three steps: Fourier Transform F , Linear Transform on the lower Fourier Modes $\mathcal{R}$ and Inverse Fourier Transform $\mathcal{F}^{-1}$

2.2 Fourier Neural Operators

Fourier Neural Operators leverage the Fourier Transform to execute convolutions in the frequency domain by element wise products, reducing computational complexity. They use an iterative layer-wise propagation mechanism, updating functions v(x) through kernel integral operators K, point-wise non-linear activation functions σ, and learn-able linear transformations W. Functions u and v denote state mappings on variable grid resolutions, highlighting FNO's capability to manage non-uniform grids.

More specifically, the kernel integral operators apply a convolutional transformation defined by $(Kv)(x) = \int Dk(x-y)v(y)dy$. Introducing a learnable kernel function k boosts the model's expressiveness. Often operating on subsets of $\mathbf{R}^3$ in 3D imaging, the Fourier Transform facilitates efficient calculation in the frequency domain as $\mathcal{F}^{-1}\left(\mathcal{F}(k)\cdot\mathcal{F}(\nu)\right)(x)$, where the model learns the transformation $\mathcal{F}(k)$ on stable lower frequency components.

2.3 FNOSeg3D

The network architecture of FNOSeg3D, as depicted in Fig2, is a multi-layered construct wherein each Fourier layer is implemented using the fast Fourier transform (FFT) to transition input data to the frequency domain, apply convolution, and then revert to the spatial domain via the inverse FFT. This process enables the model to capture a global receptive field, eliminating the need for pooling operations.

The FNOSeg3D leverages a uniform transformation function R(k) across all the fourier layer, contrary to the original FNO which utilized a distinct R(k). This shared approach not only results in a significant reduction in the number of model parameters but also addresses the over-parameterization issue prevalent

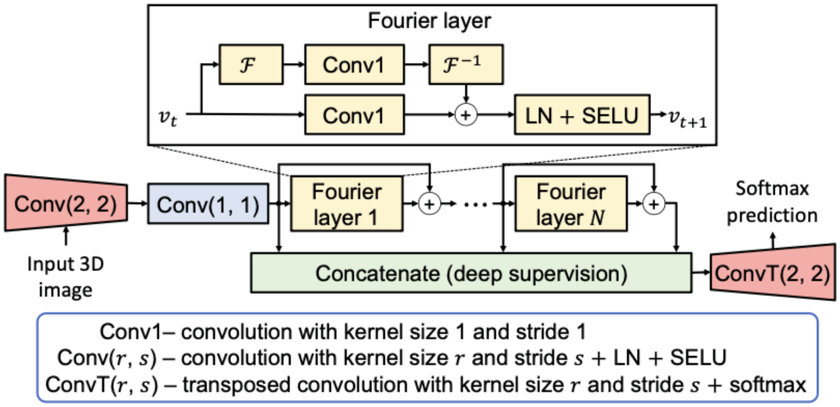

Fourier layer
$\mathcal{F}$
Conv1
$\mathcal{F}^{-1}$
v_t
Conv1
+
LN + SELU
v_{t+1}
Conv(2, 2)
Conv(1, 1)
Fourier layer 1
+
Fourier layer N
+
Softmax prediction
Input 3D image
Concatenate (deep supervision)
ConvT(2, 2)
Conv1– convolution with kernel size 1 and stride 1
Conv(r, s) – convolution with kernel size r and stride s + LN + SELU
ConvT(r, s) – transposed convolution with kernel size r and stride s + softmax

Fig. 2.

in the original FNO. This corresponds to carrying out 3D convolutions with unit-size kernels in the Fourier domain.

To bolster training stability and accuracy, FNOSeg3D incorporates residual connections and deep supervision mechanisms. Layer Normalization (LN) and Scaled Exponential Linear Units (SELUs) are employed as activation functions, with softmax prediction for the final layer.

Lastly, to replace the traditional image resampling method, the model uses an initial convolution layer with a small kerenel size, followed by a transpose convolution layer at the output , ehich effectively learns the optimal downsampling and upsampling required for 3D image segmentation.

2.4 Training

For the training strategy, the initial step involves aggregating segments from all OARs to form a comprehensive dataset. Subsequent to this, image files are converted from the NRRD format to the more widely compatible NIfTI-GZ format, which is better suited for our processing pipeline. A crucial preprocessing step involves reorienting the imaging data from the left-posterior-superior (LPS) coordinate system to the right-anterior-superior (RAS) system, ensuring consistency with standard imaging orientations.

Before feeding the data into the neural network, it undergoes a CLIP normalization process, which adjusts the Hounsfield Unit (HU) range to the 0.5 to 0.95 percentile, effectively enhancing the contrast of important features while mitigating the influence of outliers. This is followed by a global normalization step, where the mean and standard deviation of the dataset are used to standardize the images, thus facilitating more stable and faster convergence during training.

Spatial resolution is another key consideration; images are downsampled to an approximate resolution of 290x290, with careful attention to maintaining average spacing, which preserves the structural integrity of the images during resampling. To increase the robustness of the model against variations in new data, we employ image augmentation techniques such as rotation (axial, ±30 degrees), shifting (±20%), and scaling (ranging from 0.8 to 1.2).

For the optimization process, we utilize the ADAMAX optimizer, an extension of the ADAM optimizer, with a learning rate set between 10e-2 and 10e-3, to adaptively adjust the learning steps based on the training dataset. This strategy is executed on a high-performance NVIDIA RTX QUATRO 6000 GPU, with 24 GB of VRAM, training with a batch size of one and 50 epochs.

3 Experiment

3.1 Dataset

The dataset utilized for model training and evaluation comprises HaN-seg imaging data, a cohort of 42 patients, with a gender distribution of 30 males and 12 females. The average age of the participants is 60 years, with a standard deviation of 14, indicating a middle-aged to senior demographic. Each patient's data is presented in the form of CT scans and MR scans with a resolution of 1024 x 1024 and 512 x 512 respectively(models evaluated on CT). On average, each scan consists of 201 slices, with a range between 174 and 323 slices.

The average spacing of the slices is [4.26, 2.14, 2.14] millimeters, offering fine detail in two dimensions and a slightly coarser slice thickness. The dataset targets 30 different Organs At Risk (OAR), 22 of which are unpaired organs that could be affected by medical interventions.

During model training, a selection of 35 patient scans was utilized whereas the independent test set consisted of 7 distinct patients' scans. Preprocessing and learnable resampling approach from 2.4 was applied. Model evaluation took place using the DICE score metric.

nnUNet	FNOSeg3D
0.43	0.76

Table 1. Inference time per image for 290 x 290 x 200 in seconds

3.2 Results

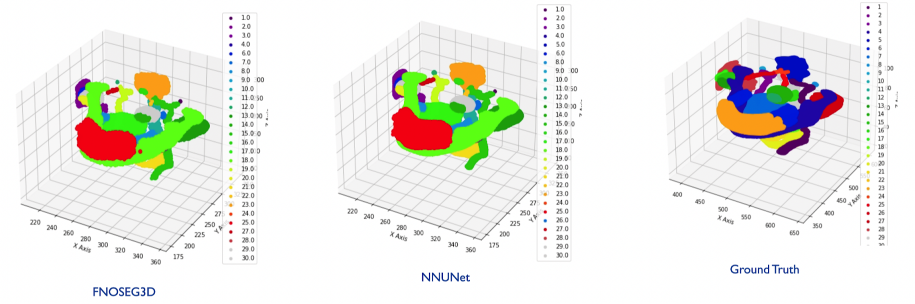
For FNOseg3D, the mean DICE score stands at 0.4129 with a standard deviation of 0.1752, indicating a moderate level of segmentation accuracy with some variability across different labels. It is worth noting that the model achieved this performance with an impressively low number of parameters around 30K.

In contrast, nnUNet exhibits a higher mean DICE score of 0.5334 with a greater standard deviation of 0.2403. This suggests that while nnUNet generally provides better segmentation accuracy than FNOseg3D, it also shows more variability in its performance across the dataset. This increased performance comes at the expense of a much larger number of parameters roughly 8.6 million.

Further analysis of the models' performance is provided by the inference times listed in Table 1. Here, FNOseg3D shows a slower inference time per image (0.76 seconds) compared to nnUNet (0.43 seconds) for images of size 290 x 290 x 200. This indicates that while FNOseg3D is parameter-efficient, it may require optimization to match the inference speed of nnUNet.

Structure	**FNOSeg3D**	**nnUNet**
A_Carotid_L	0.2900	0.1217
A_Carotid_R	0.5341	0.6966
Arytenoid	0.5383	0.7118
Bone_Mandible	0.4496	0.6489
Brainstem	0.0000	0.0000
BuccalMucosa	0.3177	0.4705
Cavity_Oral	0.3758	0.5632
Cochlea_L	0.3722	0.6279
Cochlea_R	0.4464	0.7099
Cricopharyngeus	0.5724	0.7083
Esophagus_S	0.5167	0.0000
Eye_AL	0.5429	0.7134
Eye_AR	0.5415	0.6441
Eye_PL	0.4888	0.6800
Eye_PR	0.4864	0.6538
Glnd_Lacrimal_L	0.2844	0.3801
Glnd_Lacrimal_R	0.5799	0.7276
Glnd_Submand_L	0.5333	0.6273
Glnd_Submand_R	0.4708	0.6146
Glnd_Thyroid	0.4574	0.7095
Glottis	0.6631	0.2139
Larynx_SG	0.0440	0.7319
Lips	0.5450	0.6676
OpticChiasm	0.3714	0.6125
OpticNrv_L	0.4937	0.0000
OpticNrv_R	0.0000	0.2472
Parotid_L	0.0855	0.5533
Parotid_R	0.3071	0.7081
Pituitary	0.5506	0.7254
SpinalCord	0.5268	0.7254

Table 2. Comparison of DICE Scores for FNOSeg3D and nnUNet



FNOSEG3D

NNUNet

Ground Truth

Fig. 3.

4 Conclusion

In conclusion, the FNOSeg3D is a significant advancement in the field of robust 3D image segmentation. By streamling the FNO to reduce parameter count and enhancing it with residual connections and deep supervision creates a model that has parameter efficiency with robust learning. Its global receptive field futher reinforce its resilience to different modalities and resolution. Comparitive analysis shows that FNOSeg3D was comparable to nnUNet with when utilizing low resoltion images. Thus FNOSeg3D is well suited for scenarios demanding portability and adaptibility. Meanwhile , nnUnet can be used as a gold standard for utmost accuracy with high compuatational capability.