Attention-Gated UNETR: Investigating Skip-Connection Attention Mechanisms for 3D Medical Image Segmentation

Submitted by: Ori Zarchi

Selected paper: UNETR: Transformers for 3D Medical Image Segmentation

Proposed Extension: Integrating gated attention mechanisms into UNETR's skip connections to evaluate potential segmentation performance improvements on 3D medical imaging tasks.

Introduction

Segmentation of medical images is a fundamental task in image analysis, often serving as a first step for studying anatomical structures. Traditionally, convolutional neural network (CNN)–based architectures have been widely used for segmentation tasks. However, CNNs have a major limitation: they struggle to capture long-range spatial dependencies.

The original UNETR paper addresses this limitation by introducing a novel architecture that combines a Vision Transformer (ViT) encoder with a CNN-based decoder, reformulating 3D medical image segmentation as a sequence-to-sequence prediction problem. The transformer encoder processes 3D volumes as sequences of patches, enabling better learning of global context, while skip connections merge multi-resolution features with the decoder. UNETR achieved state-of-the-art performance on several benchmarks, including the BTCV dataset for multi-organ segmentation and the MSD dataset for brain tumor and spleen segmentation, demonstrating the effectiveness of transformer-based approaches for volumetric medical image analysis.

Although UNETR successfully integrates transformers in the encoder, the decoder remains purely CNN-based, leaving room to further address the challenge of modeling long-range dependencies. This project aims to investigate whether incorporating attention mechanisms into the decoder can further improve segmentation performance.

We selected UNETR because it represents a breakthrough in applying transformers to medical image segmentation, yet its decoder architecture presents a clear opportunity for improvement. Its modular design allows us to implement and test decoder modifications within our project constraints while exploring a meaningful research question regarding optimal attention placement in hybrid transformer–CNN architectures.

Proposed Extension

We propose integrating attention mechanisms into UNETR's decoder to enhance 3D medical image segmentation. Specifically, we implemented two gated attention variants applied to skip connections, transforming the standard UNETR into attention-augmented versions:

1. Attention Gate: This approach modulates the enc2 skip connection using a gating signal from dec2. The gating signal is upsampled, an attention map is computed via convolutions and a sigmoid, and it is multiplied with the skip. This simple design provides a first attempt at selective feature enhancement.
2. Improved Attention Gate: This version applies gates to both enc1 and enc2. It downsamples the skip to match the gating signal's scale, computes attention, and upsamples the resulting mask. Residual gating ($1 + \gamma * \text{mask}$) with identity initialization is included, following the selective enhancement principle from Attention U-Net.

Motivation:

UNETR's asymmetric design captures global context in the transformer encoder but leaves the decoder without sophisticated feature selection. Skip connections are critical for recovering spatial information and could benefit from attention-guided filtering to suppress irrelevant regions and emphasize salient features. Our hypothesis is that adding attention mechanisms to the decoder's skip connections will improve the model's ability to selectively leverage multi-scale features, enhancing segmentation performance.

Implementation & Experimental Design:

Experiments are designed on the MSD Spleen dataset in two phases:

- v1 (Low data regime): Limited data/epochs (6 train/4 val, 100 epochs).
- v2 (Comprehensive evaluation): More data/epochs (24 train/7 val, 200 epochs), lower learning rate.

Alternative Approaches:

Other possibilities included attention on decoder feature maps, or full transformer decoders. Skip connection gating was chosen for efficiency, its key role in U-Net architectures, and proven effectiveness in Attention U-Net. Full transformer decoders were avoided due to memory limits and the potential loss of CNN inductive biases.

Hypothesis:

Attention-gated skip connections will improve the selective utilization of multi-scale features in the decoder, leading to enhanced segmentation performance, particularly in regions with complex anatomical structures.

Methodology

This study implemented and evaluated two attention-augmented variants of the UNETR architecture on the MSD Spleen dataset, focusing on the proposed extension of integrating gated attention mechanisms into the decoder. The methodology outlines the model modifications, dataset preparation, training details, and tools used.

Model Description:

The baseline UNETR model features a Vision Transformer (ViT) encoder with 12 layers, a hidden size of 768, and a patch size of 16x16x16, coupled with a CNN-based decoder utilizing skip connections from four encoder levels (feature sizes 16, 32, 64, 128). The first attention-augmented variant (v1) introduced a single attention gate on the enc2 skip connection, modulating it with a gating signal from dec2 via upsampling, convolution, and a sigmoid multiplication. The improved variant (v2) extended this by applying gates to both enc1 and enc2, downsampling the skip to the gating signal's scale, computing attention, upsampling the mask, and incorporating residual gating ($1 + \gamma * \text{mask}$) with

identity initialization (bias ~2.0). Both models maintained the original input channels (1), output channels (2 for binary spleen segmentation), and image size (96x96x96).

Dataset and Preprocessing:

The MSD Spleen dataset comprises 41 CT volumes with resolutions ranging from 0.613x0.613x1.50 mm^3 to 0.977x0.977x8.0 mm^3, re-sampled to an isotropic voxel spacing of 1.0 mm. Voxel intensities were normalized to [0,1] using the 5th and 95th percentiles of foreground intensities, implemented with a range of [-175, 250] and clipping. For v1, a subset of 6 training, 4 validation, and 10 test samples was used due to initial constraints. For v2, the training set was expanded to 24 samples, validation to 7, and test remained 10, utilizing the full 41-volume dataset. Random patches of [96,96,96] were sampled with a 1:1 foreground/background ratio.

Training Details:

Training employed the AdamW optimizer with an initial learning rate of 1e-4 for v1 and 5e-5 for v2, using a warmup cosine learning rate schedule over 100 epochs for v1 and 200 epochs for v2. The batch size was 1 due to Colab L4 GPU memory limits (24GB VRAM), with a sliding window batch size of 1 and an inference overlap of 0.5. Data augmentation included random flips (probability 0.2 for v1, 0.5 for v2), random 90-degree rotations (probability 0.2 for v1, 0.5 for v2), and random intensity scaling/shifting (probability 0.1). The Dice-CELoss function was used, evaluated with the Dice metric, and checkpoints were saved based on the best validation accuracy.

Tools and Frameworks:

The implementation leveraged PyTorch and the MONAI framework (version 0.8.0), with NumPy 1.23.5 for data handling and Nibabel for NIfTI file processing. Training and inference were conducted on a Colab L4 GPU, with TensorBoard for logging and visualization.

Results and Analysis

v1 Results: With 6 training samples and 100 epochs, the baseline UNETR achieved a test mean Dice score of 0.1833. Both of the attention-gated variants showed worse results than the base model, indicating the limited data and epochs hindered the attention mechanism's effectiveness.

v2 Results: Expanding to 24 training samples and 200 epochs, the baseline UNETR v2 improved to a test mean Dice of 0.6471, reflecting the benefit of increased data. The attention-gated v2 model achieved a test mean Dice of 0.6464, showing near parity with the baseline with improvement in 4/10 cases (notably *spleen_8*: 0.4172 → 0.7004, and *spleen_12*: 0.6602 → 0.7093), and regression in others. Overall mean Dice was essentially a tie ($\Delta = -0.0007$).

Despite a higher final training loss for the attention-gated model (0.456 vs 0.364), it achieved higher validation Dice. The best validation Dice scores were 0.8143 for Base UNETR v2, and 0.8336 for Improved Attention-Gated UNETR v2, suggesting less overfitting or better generalization under identical settings.

Sample	Base UNETR v2 Dice	Improved Attention-Gated UNETR v2 Dice	Difference (AG – Base)
spleen_2.nii.gz	0.6919	0.5705	-0.1214
spleen_6.nii.gz	0.7208	0.6636	-0.0572
spleen_8.nii.gz	0.4172	0.7004	**+0.2832**
spleen_10.nii.gz	0.6123	0.5233	-0.0891
spleen_12.nii.gz	0.6602	0.7093	+0.0491
spleen_13.nii.gz	0.7003	0.7006	+0.0003
spleen_14.nii.gz	0.7752	0.7643	-0.0109
spleen_16.nii.gz	0.4059	0.4024	-0.0035
spleen_18.nii.gz	0.6792	0.6198	-0.0594
spleen_20.nii.gz	0.8078	0.8100	+0.0022
Overall Mean	**0.6471**	**0.6464**	**-0.0007**

Table 1: Dice Scores for test samples, comparing between "Base UNETR v2" and the "Improved Attention-Gated" v2 model

Figure 1 presents visualization of a single slice from the spleen_8.nii.gz sample, which showed the highest improvement in Dice score (from 0.4172 in the Base UNETR v2 to 0.7004 in the Attention-Gated v2 model). The visualization displays the original scan, ground truth, base UNETR v2 prediction, and attention-gated v2 prediction. Despite the quantitative improvement in Dice score, the visual results reveal significant limitations: both models produce noisy or incomplete segmentations, with the spleen appearing fragmented or missing, even in the improved model. This discrepancy highlights that while the Dice metric captures overall overlap, the absolute quality of segmentation remains poor, with evident artifacts and inaccuracies.

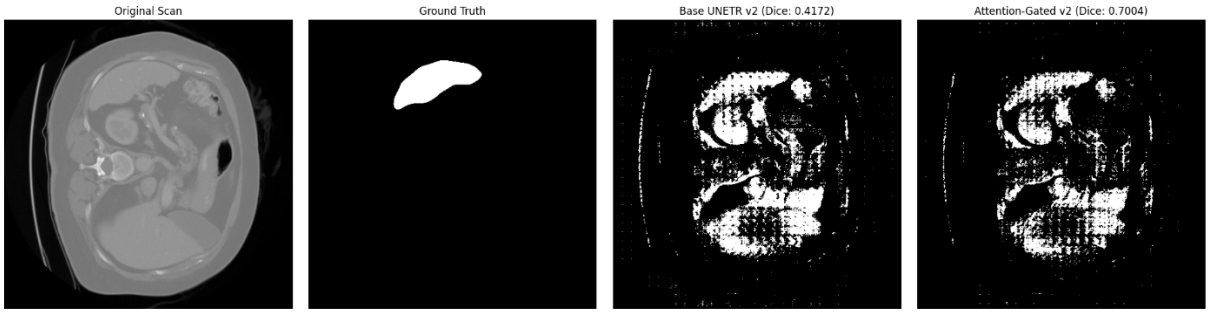

Figure 1: Visualization of a slice from spleen_8.nii.gz showing original scan, ground truth, "Base UNETR v2", and "Improved Attention-Gated v2 predictions.

The training and validation curves shown in figures 2-3 highlight the effect of scaling up data and epochs. In v2, both models show consistently lower training loss compared to v1, reflecting improved optimization. Validation Dice scores are also higher overall, with the Improved Attention-Gated UNETR v2 model surpassing the Base model for much of the training after roughly 110 epochs and reaching a higher peak. However, the advantage

is not uniform, as fluctuations indicate that the gains are case- and epoch-dependent rather than consistent across the entire run.

Figure 2: Training loss vs. epoch for base UNETR and attention-gated UNETR in v1 and v2. Both models achieve lower loss in v2, reflecting improved optimization with more data and epochs.

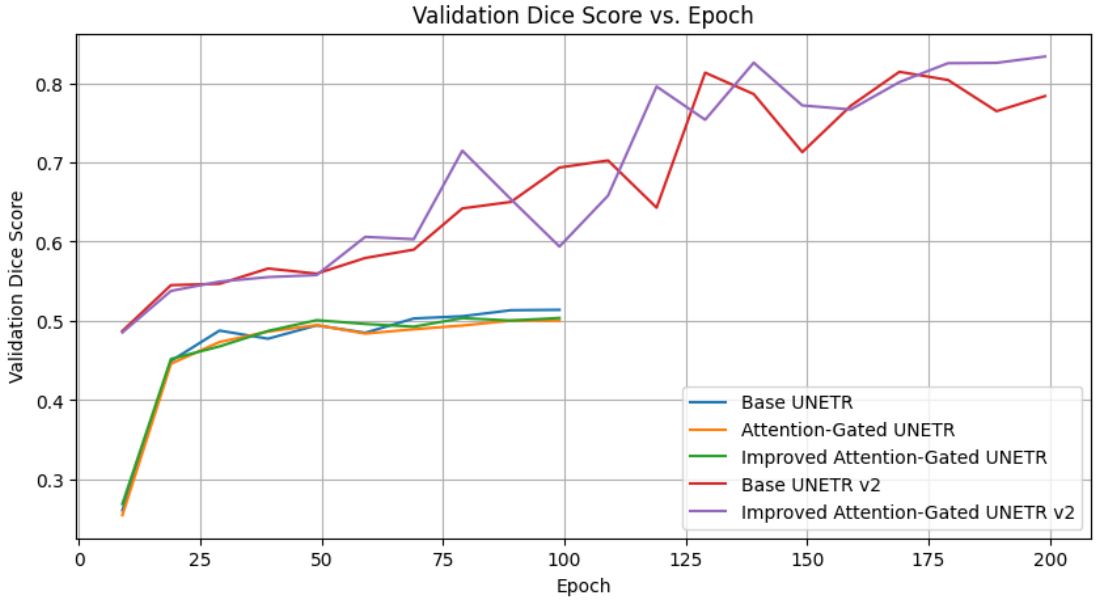

Figure 3: Validation Dice vs. epoch for base UNETR and attention-gated UNETR in v1 and v2. The v2 attention-gated model reaches a higher peak Dice and outperforms the base model for portions of training, though gains are not consistent across all epochs.

Comparison with Baseline:

The v1 attention gate failed to improve over the baseline, likely due to insufficient data and training, aligning with our hypothesis that attention requires adequate resources to be effective. In v2, the attention-gated model showed a first instance of improvement in specific test samples, supporting the idea that gated attention can enhance selective feature utilization. Taken together, the higher validation Dice for the attention-gated v2

model, the per-case improvements in several test samples, and the reduced tendency to overfit (higher training loss but better validation) indicate that the attention-gated modification provides a meaningful architectural benefit. While absolute segmentations remain weak due to limited data and training time, these trends support our change as proof-of-concept that would likely yield clearer test improvements with more data, longer training, and additional tuning.

A key challenge was the small dataset size (initially 6, later 24 samples), which limited the attention gates' ability to learn consistent patterns, as seen in poor segmentation in some spleen_8 layers. The 7-sample validation set may have biased the gates toward validation cases, reducing test generalization (validation 0.8336 vs. test 0.6464). Memory constraints on Colab prevented larger batch sizes or full self-attention, restricting the extension's scope.

Conclusion

We modified UNETR by adding attention gates on high-resolution skips (enc1/enc2) with residual gating and identity initialization. Under severe constraints (v1), attention underperformed. With more data/epochs and lower LR (v2), the attention-gated model surpassed the base on validation Dice, matched it on mean test Dice, and improved several hard cases, suggesting the architectural change helps once training conditions are reasonable.

Limitations in the project included a small training set, 200 epochs (vs. ~5,000 in the original paper), single split without cross-validation, and limited tuning of gate-related hyperparameters. Qualitative segmentations remain noisy, and the improvement was not consistent across all test cases (4/10 improved).

Future work can include scaling data and training length, performing k-fold cross-validation, tuning gate strength (γ) and biases, exploring addition of gates at deeper skips (enc3/enc4) or freezing gates early then unfreezing.