A Transformer-Based Network for Anisotropic 3D Medical Image Segmentation

Danfeng Guo[1], Demetri Terzopoulos[1,2]
[1]University of California, Los Angeles, California, USA
[2]VoxelCloud, Inc., Los Angeles, California, USA

Image Anisotropy

- Common problem in 3D medical image modalities (CT, MRI)
- Voxel spacings along x, y, z dimensions are different. Usually spacing along z is different from that along xy.
- Different spacing indicates different level of voxel correlations.
- How to encode information of anisotropic images

Solutions

- Re-slice to isotropic spacing
  - Introduce noise
  - May cause information loss
- Hybrid 2D/3D convolution, CNN-RNN structure
  - Use the same kernel for all cases
  - May not adapt to variable slice spacing

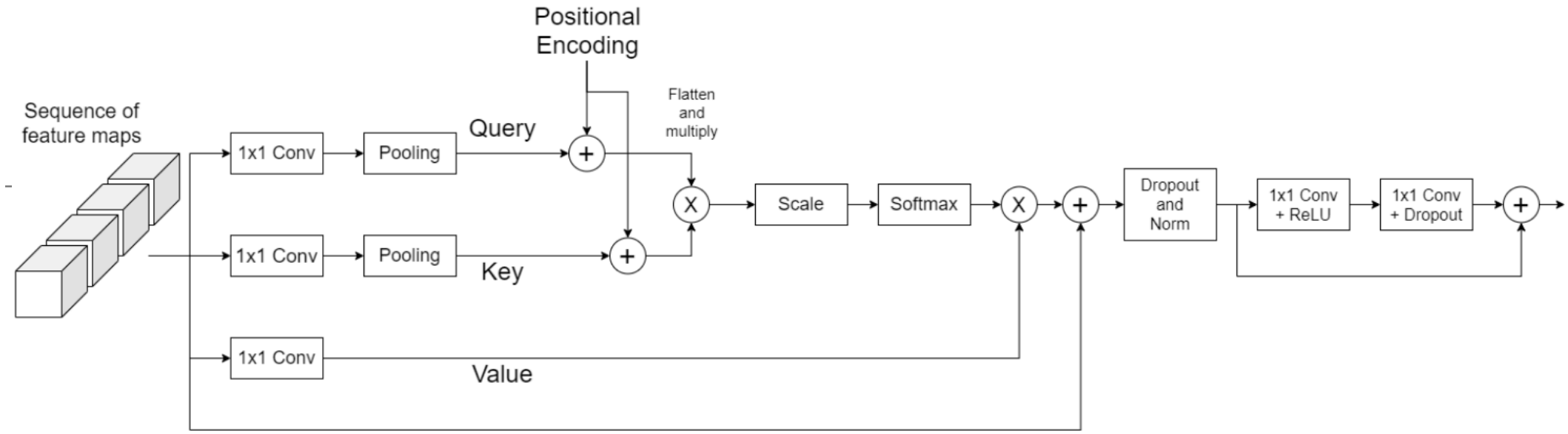
Transfomer-based Network

- Adapt to variable slice spacing
- Computationally efficient
- Consume fewer resources

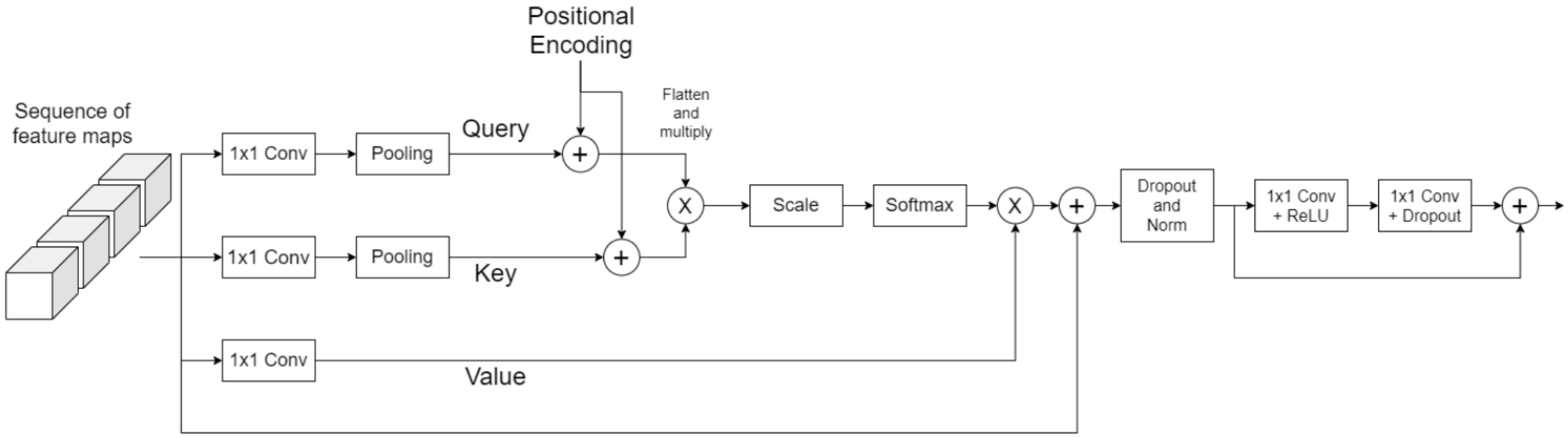
Transformer

- A self-attention mechanism developed by Vaswani et al [1] in 2017
- The basic structure of many state-of-art natural language processing models (eg: BERT)

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008. 1, 2, 3



- Use queries and keys to compute weight vector that represents the slice correlations
- Represent the new feature map as a weighted sum of itself and the feature maps of neighboring slices
- Use Positional Encoding (PE) to inject information about the sequence order



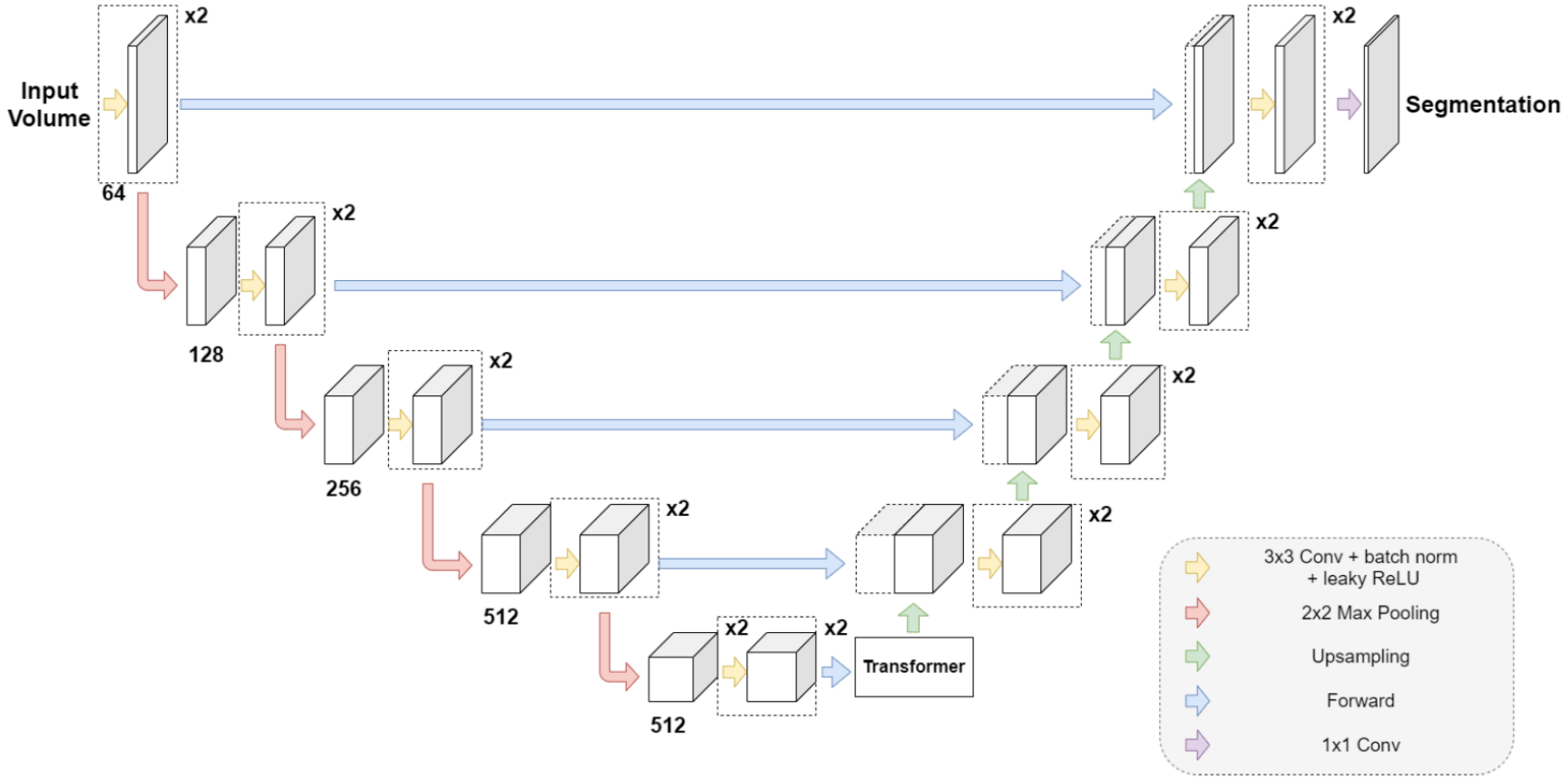
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

d_k is the dimension of queries and keys

$$\text{PE}(i, j) = \begin{cases} \sin(j/w^{\frac{i}{d_k}}), & i \text{ even}, \\ \cos(j/w^{\frac{i}{d_k}}), & i \text{ odd}, \end{cases}$$

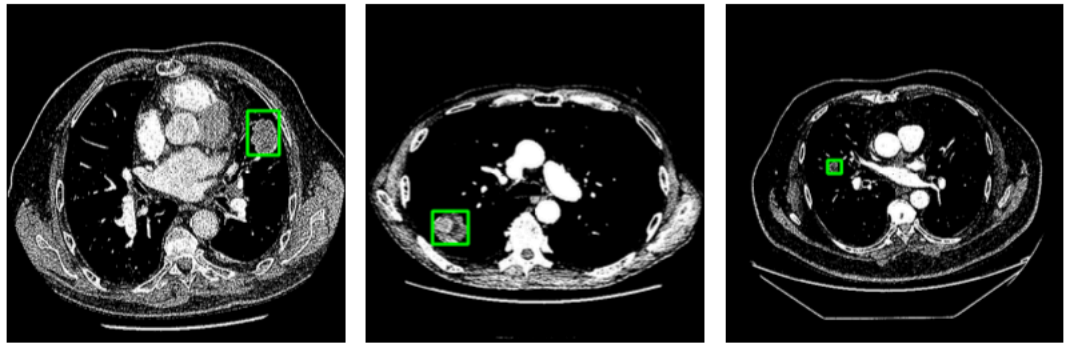
i is from 0 to d_k , j is slice number.
PE at j+n has linear relationship with PE at j

Network


Input Volume
64
128
256
512
512
x2
Transformer
Segmentation
3x3 Conv + batch norm + leaky ReLU
2x2 Max Pooling
Upsampling
Forward
1x1 Conv

Experiments

- Lung cancer segmentation dataset from the Medical Segmentation Decathlon [2]
- 20,207 lung CT slices from 63 subjects.
- Selected 2,027 positive slices and 1,890 negative slices.



Experiments

- We re-sliced the original training set such that the voxel spacings along the three dimensions are the same.
- Train models on this isotropic dataset.
- Test on the original dataset.
- Compare the models' ability to adapt to variable spacing.

Experiments

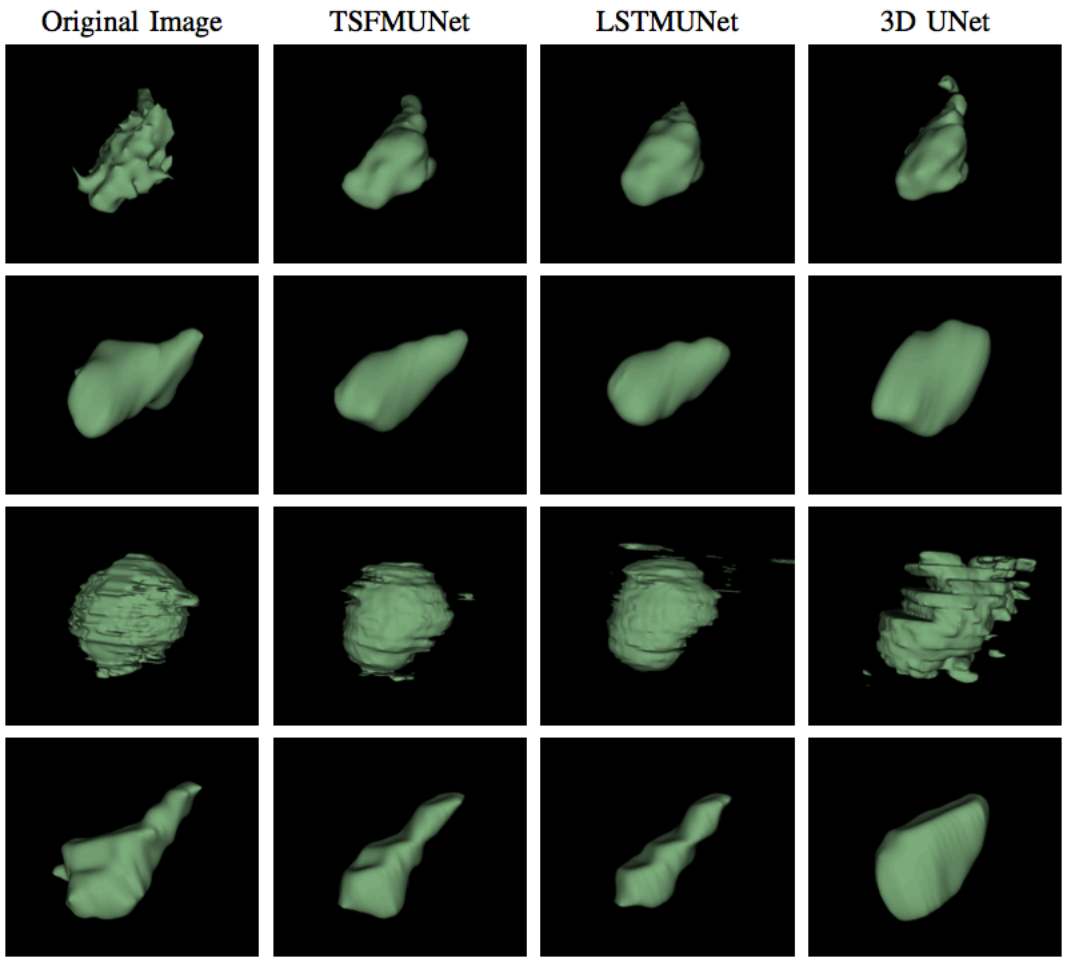
DICE SCORE COMPARISON (ORIGINAL DATA).

Model	Dice Score
TSFMUNet	**0.8717**
LSTMUNet	0.8573
3D U-Net	0.7744
2D U-Net	0.7309

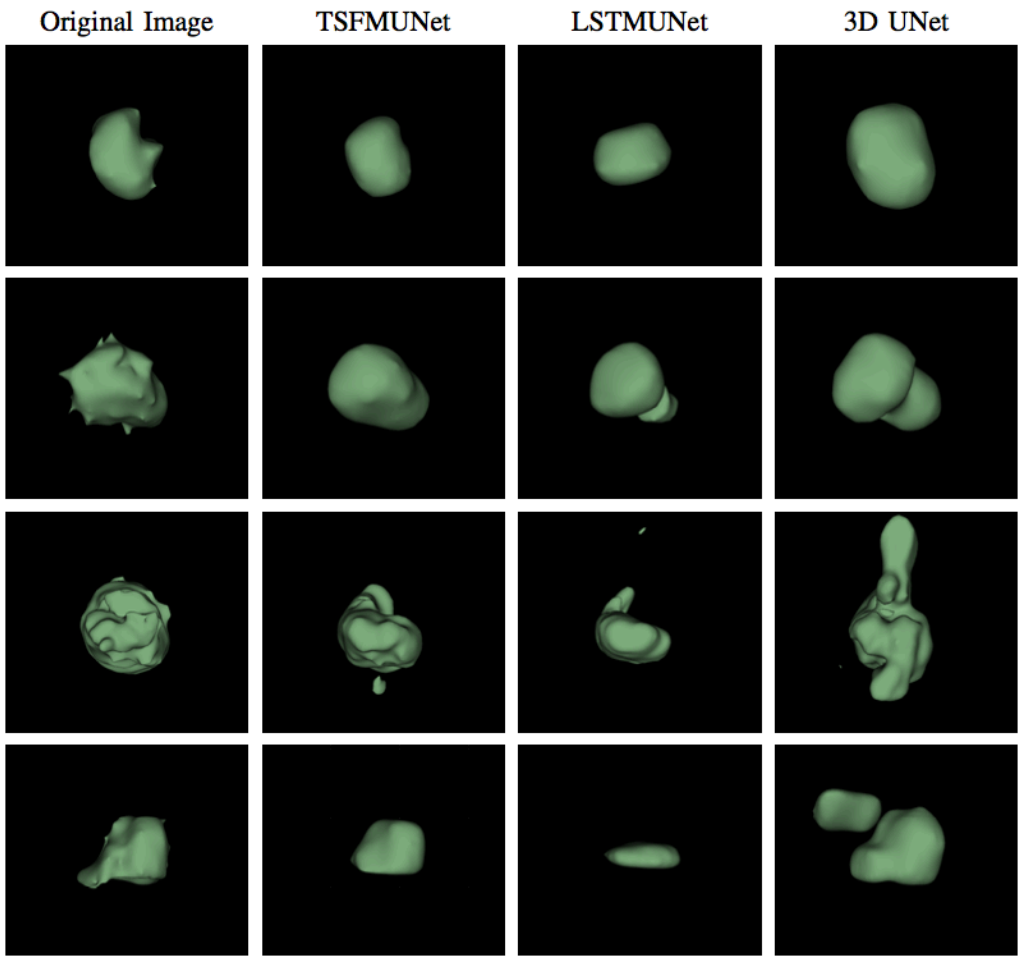
DICE SCORE COMPARISON (RE-SLICED DATA).

Model	Dice Score	Performance Drop
TSFMUNet	**0.8674**	**0.0043**
LSTMUNet	0.8217	0.0356
3D U-Net	0.7261	0.0483

Performance drop: 3D U-Net > LSTMUNet > TSFMUNet



Segmentation results on models trained on the original dataset



Segmentation results on models trained on re-sliced dataset

Conclusion

- We have proposed a transformer-based network to deal with the anisotropy problem in 3D medical image analysis.
  - Self-attention mechanism
  - Adapts to images with variable slice spacing
- Experimental results with a lung cancer segmentation task reveal that our architecture outperforms baseline models.

Thank you!