Large-kernel Attention U-Net for Lesion Segmentation

Liam Chalcroft[1] and Ioannis Pappas[2]

[1] Wellcome Centre for Human Neuroimaging, University College London
[2] Stevens Institute for Neuroimaging and Informatics, University of Southern California
l.chalcroft@cs.ucl.ac.uk, ipappas@usc.edu

1 Introduction

Recent developments in computer vision have proposed the use of attention-based transformer models for a variety of tasks including classification and segmentation, by separating images patch-wise into a series of input tokens [1]. Whilst such models have shown excellent performance in both natural images and in the medical domain due to a supposed ability to better model long-range interactions over a standard convolutional neural network (CNN), they are often hampered by expensive scaling laws in their attention mechanism. Further, the lack of any inductive bias towards local interactions, and the flattening of an image to a 1D vector of patches, has potential to reduce efficiency in their ability to learn on data in the vision domain. To combat this, a number of window-based mechanisms have been proposed [9]. These have been shown to improve over standard attention in medical image segmentation [5]. However, such methods still add a large amount of computational overhead compared to existing CNNs.

A recently proposed alternative to windowed attention is large-kernel attention (LKA) [4], where a large-kernel convolution is made feasible by decomposing the transform into a series of convolutions. As shown in Figure 1, this is achieved by a series of a depth-wise convolution (DWConv), a dilated depth-wise convolution (dDWConv) and a pointwise convolution (PConv). Results on natural images have indicated that this method both attains greater accuracy than other vision models, and has a favourable accuracy-efficiency tradeoff when compared to the Swin architecture.

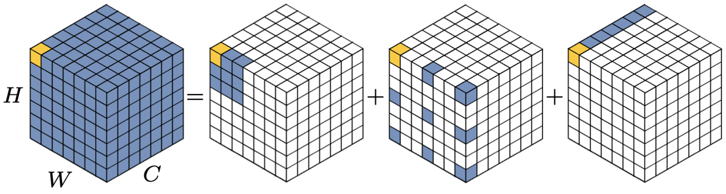

Fig. 1. Large-kernel attention mechanism, showing a 7^2 kernel decomposed into a 3^2 depth-wise convolution (DW-Conv), a 3^2 depth-wise convolution with dilation= 2 (dDWConv) and a pointwise convolution (PConv). Figure reprinted from [4].

2 Methods

In our method, the LKA module was incorporated into the encoder of a U-Net architecture, with an encoder containing 6 blocks of channels (32, 64, 128, 256, 320, 320). Each layer's LKA unit is used in a similar method to the Visual Attention Network (VAN) architecture [4], with the LKA output multiplied by the input to complete the attention unit, a GELU activation [6] applied prior, and PConv layers applied at the beginning and end of the unit. Each block contains a sequence of ×2 downsampling, LKA unit and convolutional feed-forward unit [12], with a residual connection after and a batch normalization before both the LKA and feed-forward units. Each LKA unit uses an equivalent kernel size of 21^3, comprising a 5^3 DWConv, a 7^3 dDWConv with dilation= 3 and a standard PConv. All U-Net stages contained a single LKA block, except for the second lowest which contained two blocks. The decoder follows the standard U-Net layout, with a kernel size of 3^3 at each stage.

3 Experiments

For each subject, the raw image was resliced to 1 mm^3 resolution, skull-stripped using Robex [7] and bias-corrected using the SimpleITK [10] implementation of N4 [11]. Data was resliced to $1mm^3$ and cropped to the brain foreground using MONAI*. In training, the DALI library† was used for loading with augmentations following the winner of the BraTS 2021 competition [2] to produce an augmented crop of size 128^3. An additional augmentation was used, with 50% of batches having a copy-paste [3] augmentation applied to add additional lesions, with the foreground and background blended using a gaussian filter. 5 models were trained using 5-fold cross validation in a split of 524 training and 131 validation samples. A combination of Dice and cross entropy loss were used, and optimized using the Apex‡ implementation of Adam with a learning rate of 0.0002 and a batch size of 3 for $1,000$ epochs. Deep supervision was used to calculate additional loss components for the low-resolution segmentation predictions at the decoder layers of channels 32 and 64, with weightings of $\frac{1}{2}$ and $\frac{1}{4}$ respectively. All training was performed with Auto-Mixed Precision in Pytorch Lightning§ using an NVIDIA RTX A6000 GPU with 48GB VRAM.

Final inference was performed via an ensemble of the 5 trained models, with test-time augmentation used for each model's predictions to average logits over all possible orientations. Inference of the model in all orientations was performed using a sliding window of size 128^3, with an overlap of 0.5 and a gaussian weighting to merge windowed predictions. Output soft probabilities were then averaged across all models. The resulting likelihood map was resliced to the original DWI image's resolution and then further refined using a Conditional Random Field (CRF)[8], hole filling and removal of small objects.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020). https://doi.org/10.48550/ARXIV.2010.11929, https://arxiv.org/abs/2010.11929
2. Futrega, M., Milesi, A., Marcinkiewicz, M., Ribalta, P.: Optimized u-net for brain tumor segmentation (2021). https://doi.org/10.48550/ARXIV.2110.03352, https://arxiv.org/abs/2110.03352
3. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation (2020). https://doi.org/10.48550/ARXIV.2012.07177, https://arxiv.org/abs/2012.07177
4. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network (2022). https://doi.org/10.48550/ARXIV.2202.09741, https://arxiv.org/abs/2202.09741
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images (2022). https://doi.org/10.48550/ARXIV.2201.01266, https://arxiv.org/abs/2201.01266
6. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus) (2016). https://doi.org/10.48550/ARXIV.1606.08415, https://arxiv.org/abs/1606.08415
7. Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z.: Robust brain extraction across datasets and comparison with publicly available methods. IEEE transactions on medical imaging **30**(9), 1617–1634 (2011)
8. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials (2012). https://doi.org/10.48550/ARXIV.1210.5644, https://arxiv.org/abs/1210.5644
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021). https://doi.org/10.48550/ARXIV.2103.14030, https://arxiv.org/abs/2103.14030
10. Lowekamp, B.C., Chen, D.T., Ibáñez, L., Blezek, D.: The design of SimpleITK. Front. Neuroinform. **7**, 45 (Dec 2013)
11. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: Improved n3 bias correction. IEEE Transactions on Medical Imaging **29**(6), 1310–1320 (Jun 2010). https://doi.org/10.1109/tmi.2010.2046908, https://doi.org/10.1109/tmi.2010.2046908
12. Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvtv2: Improved baselines with pyramid vision transformer. CoRR **abs/2106.13797** (2021), https://arxiv.org/abs/2106.13797

*monai.io

†docs.nvidia.com/deeplearning/dali/

‡nvidia.github.io/apex/

§pytorchlightning.ai