A convolution free network for 3D medical image segmentation

Siva Prasad Patnayakuni , Data engineer, HEB,

Hyderabad,India

***Abstract* - Profound learning models address the cutting edge in clinical picture division. The vast majority of these models are fully convolutional networks (FCNs), in particular each layer processes the result of the former layer with convolution tasks. The convolution activity partakes in a few significant properties, for example, scanty connections, boundary sharing, and interpretation equivariance. Due to these properties, FCNs have areas of strength for a helpful inductive predisposition for picture demonstrating and examination. In any case, they likewise have specific significant deficiencies, for example, playing out a fixed and pre-decided procedure on a test picture no matter what its substance and trouble in demonstrating long-range cooperation's. In this work we show that an alternate profound brain network design, dependent completely upon self-consideration between adjoining picture patches and with practically no convolution tasks, can accomplish more exact divisions than FCNs. Our proposed model depends straightforwardly on the transformer network engineering. Given a 3D picture block, our organization separates it into non-covering 3D fixes and processes a 1D installing for each fix. The organization predicts the division map for the block in view of the self-consideration between these fix embeddings. Moreover, to resolve the normal issue of shortage of named clinical pictures, we propose techniques for pre-preparing this model on enormous corpora of unlabeled pictures. Our examinations demonstrate the way that the proposed model can accomplish division correctness's that are superior to a few best in class FCN designs on two datasets. Our proposed organization can be prepared utilizing just several marked pictures. Besides, with the proposed pre-preparing systems, our organization beats FCNs while named preparing information is little.**

***IndexTerms* - Fully convolutional networks (FCNs), 3D picture block, datasets, segmentation, medical image.**

I. INTRODUCTION

Image segmentation is required for evaluating the size and state of the volume/organ of interest, populace studies, infection measurement, and PC helped treatment and careful preparation. Given the significance and the trouble of this undertaking in clinical applications, manual division by a clinical master is viewed as the ground truth. Notwithstanding, manual division is expensive, tedious, and likely to entomb and intra spectator conflict. Programmed division techniques, then again, can possibly offer a lot quicker, less expensive, and more reproducible outcomes.

Traditional strategies for clinical picture division incorporate area developing [1], deformable models [2], diagram cuts [3], bunching techniques [4], and Bayesian methodologies [5]. Map book based strategies are one more extremely famous and strong arrangement of procedures [6]. With the presentation of profound learning strategies for picture division [7], [8], these techniques were immediately taken on for clinical picture division.

Profound learning strategies have accomplished extraordinary degrees of execution on a scope of clinical picture division responsibilities [9]-[14]. One can contend that profound learning techniques have to a great extent substituted the traditional strategies for clinical picture division. Late audits of the primary lines of examination and ongoing progressions on the utilization of profound learning for clinical picture division can be found here [15], [16].

Latest investigations have pointed toward further developing the organization design, misfortune capability, and preparing strategies. Late works have shown that standard profound learning models can be prepared utilizing little quantities of named preparing pictures [17], [18]. In spite of the enormous changeability in the proposed network designs, the one normal component in these works is that they all utilization the convolution activity as the primary structure block. The proposed models contrast as to the plan of the convolutional activities, yet they all depend on a similar essential convolution activity.

II. LITERATURE SURVEY

A couple of studies have proposed elective organization designs in light of repetitive brain networks [19], [20] and consideration components [21]. There have additionally been endeavors to work on the exactness and vigor of these strategies by displaying the measurable variety looking like the organ of interest and consolidating this shape data in the profound learning strategy [12], [22], [23]. Be that as it may, those models actually expand upon the convolution activity. A few ongoing investigations have recommended that a fundamental encoder-decoder-type completely convolutional network (FCN) can deal with different division undertakings and be basically as exact as more intricate organization designs [42]. The convolution activity is additionally the principal building block of the organization structures that have effectively tended to other focal PC vision undertakings, for example, picture grouping and article discovery [25], [26]. These outcomes verify the adequacy of the convolution activity for displaying and breaking down pictures. This viability has been credited to various key properties, including: 1) nearby (inadequate) associations, 2) boundary sharing, and 3) interpretation equivariance [27], [28]. Truth be told, a convolutional layer can be viewed as a completely associated layer with an "boundlessly solid earlier" over its boundaries [29].

The properties of the convolution activity that we referenced above are, to some extent, enlivened by neuroscience of the mammalian essential visual cortex [30]. They give convolutional brain organizations (CNNs), including FCNs, a solid and valuable inductive predisposition, which makes them profoundly compelling and effective in handling different vision tasks. Be that as it may, these equivalent properties likewise put CNNs in a tough spot. For instance, the organization not entirely set in stone at preparing time and consequently they are fixed. In this manner, these organizations treat various pictures and various pieces of a picture similarly. At the end of the day, they miss the mark on component to change their loads relying upon the

picture content. Moreover, because of the nearby idea of convolution activities with little portion sizes, CNNs can only with significant effort learn long-range cooperations between far off pieces of a picture. Consideration based brain network models can possibly address a portion of the limits of convolution-based models. To put it plainly, these models target learning the connection between various pieces of a succession [31]. Above all, in contrast to CNNs, in attention based networks not all organization loads are fixed after preparing. Rather, just a part of the organization loads are gained from preparing information and the other not set in stone at test time in view of the substance of the info. Consideration based networks have turned into the predominant brain network models in normal language handling (NLP) applications. Transformers are the most widely recognized consideration based models in NLP [31]. Contrasted and intermittent brain organizations, transformers can learn more complicated and longer range collaborations substantially more really.

Additionally, they beat a portion of the focal constraints of intermittent brain organizations like disappearing slopes. They additionally take into consideration equal handling of sources of info, which can prompt essentially more limited preparing time on present day equipment. In spite of the possible benefits of TO (transformer organization s), up to this point they have not been generally taken on in PC vision applications. A new overview of the important deals with this subject can be found in [32]. Use of consideration based brain networks for PC vision applications faces a few significant difficulties. The quantity of pixels in a common picture is a lot bigger than the length of a sign succession (e.g., number of words) in ordinary NLP applications. This makes it difficult to straightforwardly apply standard consideration models to pictures. The subsequent principal reason has been the preparation trouble.

The solid inductive inclination of CNNs that we have referenced above makes them exceptionally information proficient. TOs, then again, require substantially more preparation information since they integrate insignificant inductive inclination. Ongoing examinations have proposed pragmatic answers for these two difficulties. To address the primary test, vision transformer (ViT) proposed considering picture patches, as opposed to pixels, as the units of data in a picture [33]. ViT inserts picture patches into a common space and learns the connection between these embeddings utilizing self-consideration modules. It was shown that, given enormous datasets of marked pictures and tremendous computational assets, ViT could outperform CNNs in picture grouping precision. One potential answer for the subsequent test was proposed in [34], where the creators utilized information refining from a CNN educator to prepare a TO. It was shown that with this preparing system, TOs could accomplish picture grouping precision levels comparable to CNNs utilizing similar measure of named preparing in series [34].

In this work, we propose a self-consideration based profound brain network for 3D clinical picture division. Our proposed network depends on self-consideration between straight embeddings of 3D picture patches, with no convolution activities. Offered the way that self-consideration models for the most part require huge named preparing datasets, we additionally propose unaided pre-preparing strategies that can take advantage of enormous unlabeled clinical picture datasets. We contrast our proposed model and a few cutting edge FCNs on two clinical picture division datasets.

The particular commitments of this work are as per the following:

1. We propose the main without convolution profound brain network engineering for division of 3D clinical pictures.
2. We demonstrate the way that our proposed organization can accomplish division execution levels that are better compared to or if nothing else comparable to the cutting edge FCNs. Despite the fact that earlier works have recommended that monstrous named preparing datasets are expected to actually prepare transformer networks for NLP and vision applications, we tentatively demonstrate the way that our organization can be prepared utilizing datasets of just ~ 20 - 200 marked pictures.
3. We propose strategies for pre-preparing our organization on enormous corpora of unlabeled pictures. We show that while marked preparing pictures are less in number, with these pre-preparing methodologies, our organization performs better compared to a cutting edge FCN with pre-preparing.

III. PROPOSED METHOD, IMPLEMENTATION AND TRAINING

Our proposed transformer network for 3D medical image segmentation is shown in Fig 1. The input to our network is a 3D image block the extent of the block (in voxels) in each dimension and c denotes the number of image channels. functioning with picture sub-blocks is a general come up to processing large volumetric pictures. It enables dispensation of big images of random size on imperfect GPU memory. In addition, it functions as an hidden data growth scheme since throughout training sub-blocks are sampled from arbitrary locations in the training Pictures.

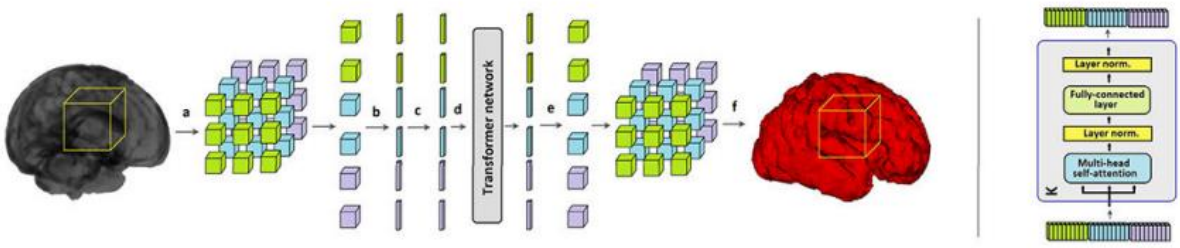


Figure 1. The proposed convolution-free network for 3D medical image segmentation.

As shown in Fig.1, our proposed network includes only the encoder section of the original transformer network proposed in [31]. The set-up has K matching stages, every consisting of a multi-head self-attention (MSA) and a successive two-layer fully-connected feed-forward network (FFN). All MSA and FFN modules include residual connections, ReLU activations, and layer state [35].Firstly with the input succession of rooted and position-encoded patches, X^0 described on top of, the kth step of the system performs the subsequent operations to map X^k to $^{Xk+1}$:

1. X^k goes through nh independent heads in MSA. The i^{th} head:

a. Computes the query, key, and value sequences from the input sequence using linear operations.
b. Computes the self-attention template and then the changed values.

2. Outputs of the n_h self-attention heads are stacked jointly and re-projected reverse against IRD.
3. The output of the modern multi-head self-attention unit is computed by a residual function.
4. Get the output of the k^{th} step of the system.

The estimated decomposition map for the block (as shown in Fig.1). Because our system predicts decomposition maps for pictures sub-blocks, in order to development a test image of uninformed size, we appropriates the system in a sliding porthole way on the picture.

We implemented the network in TensorFlow 1.16 and trained it on an NVIDIA GeForce GTX 1080 GPU on a Windows machine with 256 GB of memory and 16 CPU cores.

Manual segmentation of complicated systems including the mind cortical plate can take numerous hours of a scientific expert`s time for a unmarried 3-D photo. Therefore, techniques that may obtain excessive overall performance with fewer classified schooling pix are enormously advantageous. This is particularly critical for transformer networks. As we referred to above, transformer networks lack an awful lot of the integrated inductive bias that many different networks including CNNs experience simply through the distinctive feature in their architectural design. Therefore, in comparison with the ones architectures, transformers generally want an awful lot large classified schooling datasets so that it will study the underlying styles immediately from facts. In NLP applications, a completely not unusualplace method is to pre-teach the community the usage of unsupervised schooling on large unlabeled datasets [45]. In the identical spirit, we suggest pretext responsibilities that may be used to teach our community on unlabeled 3-D scientific photo datasets.

For version pre-schooling with every of the above strategies, we use a distinctive output layer (with out the softmax operation). In order to fine-track the pre-skilled community for the segmentation assignment, we introduce a brand new output layer with the softmax activation and teach the community at the classified facts as defined above. We fine-track the complete community, in preference to simplest the output layer, at the classified pix due to the fact we've got located that fine-tuning the complete community for the segmentation assignment ends in an awful lot higher results.

Pre-schooling techniques also are typically used for FCNs. Prior research have proven that pre-schooling would possibly result in enormous enhancements in segmentation overall performance of FCNs, particularly while the segmentation assignment is tough and the scale of classified schooling facts is small [17], [46]. Therefore, we are able to use the identical denoising and inpainting responsibilities defined above to pre-teach the FCNs. Moreover, we are able to additionally use the semi-supervised FCN schooling approach proposed in [47]. The approach of [47] is primarily based totally on an alternating optimization strategy. It alternately updates the community parameters and the anticipated labels for the unlabeled pix in parallel.

Tbl.1 suggests the datasets used for version schooling and assessment on this work. The pix have been randomly break up into schooling and take a look at sets, and not using a affected person facts acting in each schooling and take a look at sets. The identical schooling/take a look at splits have been used for all networks. For every dataset, we used about 20% of the schooling pix for preliminary validation experiments to determine on schooling settings including the getting to know fee for every community. After deciding on the schooling settings, every community turned into skilled on all schooling pix. The simplest facts augmentation turned into the implicit augmentation thru sampling of photo blocks from random places withinside the schooling pix. Voxel intensities of all pix have been normalized to have a 0 imply and unit fashionable deviation. Moreover, all pix have been interpolated the usage of 3-D spline interpolation into isotropic voxel sizes proven withinside the Tbl.. The corresponding floor reality segmentations have been interpolated the usage of nearest neighbor interpolation. We evaluate our proposed approach with the competing networks in phrases of DSC, the ninety five percentile of the Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD).

IV. RESULTS AND DISCUSSION

Tbl.2 seems on the department execution of the proposed method with the contending FCNs at the thoughts cortical plate and hippocampus datasets. As depicted in Section III , the proposed community includes some hyper-obstacles that could effect the department outcomes. The consequences added in Tbl. 2 have been obtained with: K = 5, W = 24, n = three, D = 1024, D_h = 256, n_h = 4. These are our default settings for community hyper-obstacles that we've got applied in all examinations introduced withinside the rest of the paper, besides if commonly determined. We confirmed up at those obstacles utilising cross-approval probes the practice images withinside the cerebrum cortical plate and hippocampus datasets in addition to different datasets now no longer added withinside the paper. We gift trial outcomes at the affects of numerous hyper-obstacles at the department execution beneath.

Table.1. Datasets used for experiments in this work.

target organ	image modality	[ntrain, ntest]	image resolution (mm)	source
Brain cortical plate	T2 MRI	[24,8]	0.83	In-house (Boston Children's Hospital
Hippocampus	MRI	[212, 24]	0.78	https://decathlon-10.grand-challenge.org/

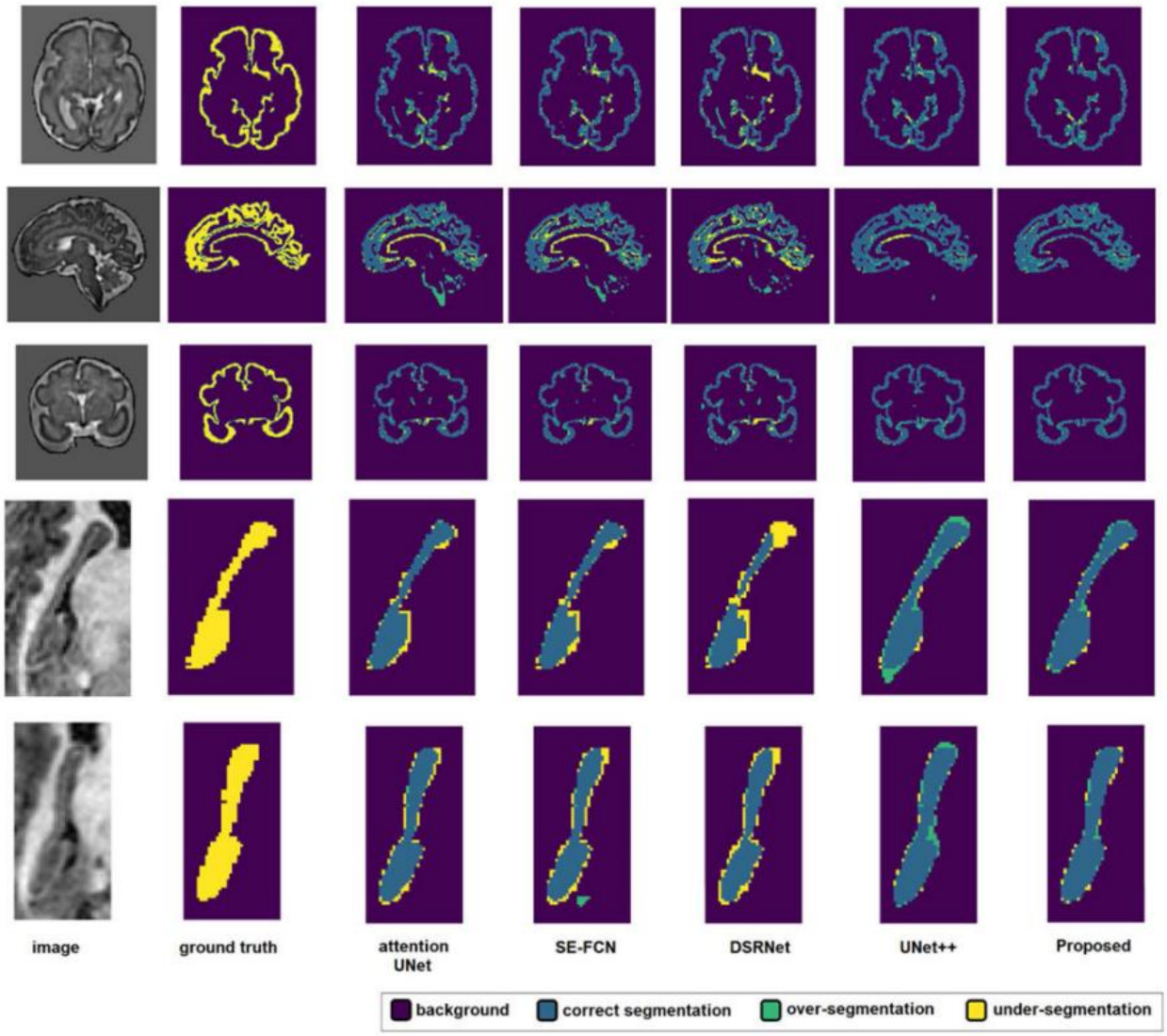


Figure 2. Example segmentations predicted by the proposed method and the four FCNs.

The consequences added in Tbl.2 display that the proposed community has done department execution ranges which are higher than the contending FCNs. For every dataset and each one of the 3 rules, we completed matched t-checks to test whether or not the differences have been measurably massive. As displayed withinside the Tbl., department execution of the proposed convolution unfastened company turned into basically higher in comparison to the 4 FCNs close to DSC, HD95, and ASSD at p consequences obtained with the proposed approach have been especially unmatched concerning the space measurements, i.e., HD95 and ASSD. Among the FCN structures, UNet++ completed appreciably higher in comparison to special designs on each datasets, but its department execution turned into basically sub-par in comparison to that of our proposed approach.

Fig.2 suggests version cuts from take a look at images in every dataset and the divisions expected through the proposed approach and the 4 FCNs. Visual exam of the consequences suggests that the proposed community is ready to do exactly sectioning first-class and thoughts boggling designs just like the cerebrum cortical plate. On each datasets, Attention UNet, DSRnet, and SEFCN regularly added approximately deceptive advantageous expectancies a long way farfar from the goal organ, that's the reason for his or her terrible displaying as a long way as the space measurements added in Tbl. 2.

Fig.3 suggests the aftereffects of this trial. The consequences display that with the proposed pretraining, our sans convolution community accomplishes basically greater specific divisions with much less named making ready images. True to form, on each datasets there has been a drop withinside the department execution as the amount of named making ready images turned into dwindled. Be that because it may, this drop turned into greater modest for the proposed community than for UNet++. We have observed very just like consequences with different FCN designs. For our company as nicely with admire to UNet++, the proposed inpainting pre-making ready activates rather stepped forward outcomes than the alternative pre-making ready strategies. Besides, through and large, pre-making ready on similar images activates desired department execution over pre-making ready on a dataset of numerous images.

Table.2. Comparison of the segmentation performance of the proposed method and several competing FCNs on the brain cortical plate and hippocampus datasets.

Dataset	Method	DSC	HD95 (mm)	ASSD (mm)
Brain cortical plate	Proposed	0.912 ± 0.023 *	0.862 ± 0.212 *	0.234 ± 0.053 *
	UNet++	0.798 ± 0.045	0.912± 0.327	0.245 ± 0.079

	Attention UNet	0.798 ± 0.059	1.234 ± 0.291	0.497 ± 0.131
	DSRnet	0.834 ± 0.048	0.976 ± 0.236	0.389 ± 0.129
Hippocampus	Proposed	0.901 ± 0.018 *	1.231 ± 0.199 *	0.426 ± 0.053 *
	UNet++	0.865 ± 0.026	1.543 ± 1.423	0.613 ± 0.187
	Attention UNet	0.830 ± 0.032	5.740 ± 5.231	1.154 ± 0.567
	DSRnet	0.721 ± 0.041	3.442 ± 3.055	1.485 ± 0.326
	SEFCN	0.746 ± 0.041	8.513 ± 5.320	1.865 ± 0.771

Better results for each dataset/criterion have been marked using bold type. We used paired t-tests to find statistically significant differences; asterisks denote significantly better results at p < 0.01.)

As displayed in Fig three, for each the proposed community and UNet++ the department execution is, rather but reliably, better at the same time as pre-making ready is completed on a similar dataset. This demonstrates that each the proposed community and UNet++ can get acquainted with the modern examples in unlabeled images and have an effect on this statistics to perform higher department outcomes. This is an exceedingly charming and promising notion because it demonstrates the manner that the proposed company may be organized regarding a modest bunch of marked images for dividing complicated designs in three-D medical images. This final results is a great deal greater crucial while we keep in mind the consequences distinctive through past due image association studies. As we made feel of in Section I, image grouping concentrates on that applied a comparative methodology (i.e., making use of a transformer community on restore embeddings) required giant named datasets [33] or trusted data refining from a CNN educator version [34]. Our consequences, then again, display that major a small bunch of marked making ready images are good enough to put together a comparative company for three-D medical image department.

This may be ascribed to three variables: 1) In image grouping there are important sorts in large image highlights (even amongst images which have an area with a comparable elegance). In the department assignments taken into consideration right here, then again, there's massive closeness throughout topics or even amongst numerous patches in a comparable image. 2) There are a long way much less elegance marks (simply two) withinside the department assignments taken into consideration right here contrasted and image order applications. three) Working with image sub-blocks is going approximately as extreme regions of electricity for an enlargement system and empowers best use of named making ready images. Subsequently, irrespective of their insignificant inductive inclination, transformer networks seem, through all accounts, to be suitable for medical image department tasks.

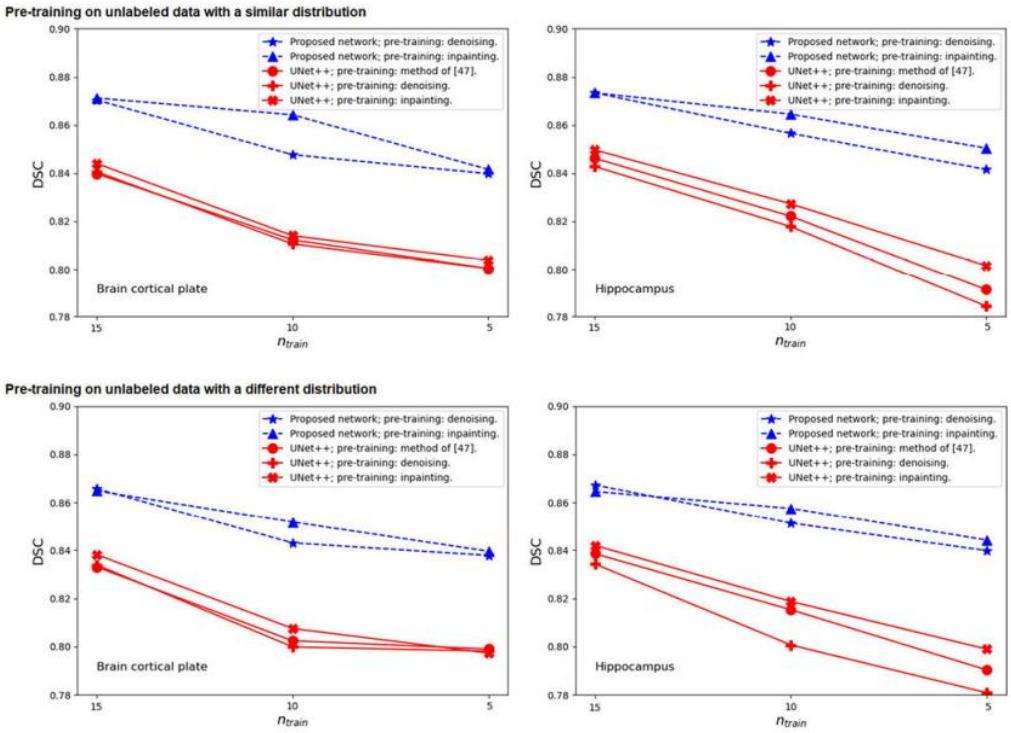


Figure 3. Segmentation performance (in terms of DSC) for the proposed network and UNet++ with reduced number of labeled training images on the brain cortical plate dataset (left) and the hippocampus dataset (right).

The trial outcomes added above exhibit the manner that the proposed method can accomplish department execution similar to or higher than FCNs with as now no longer many as 10-20 marked making ready images. This is a large and empowering bring about mild of the reality that withinside the medical imaging area guide names are hard to acquire. In any case, to assess the presentation of the proposed approach on larger datasets, we led one greater exam with the little one cerebrum filters withinside the growing Human Connectome Project (dHCP) dataset [14]. This dataset includes 558 T2 MRI cerebrum filters with cortical plate department. We arbitrarily selected fifty eight of those outputs as take a look at images. We then organized our version and UNet+ + on every of the five hundred leftover images in addition to subsets of one hundred and 10 images. Notwithstanding the

verifiable data enlargement added approximately through trying out patches from abnormal regions withinside the practise images, we carried out arbitrary turn and revolution and we delivered arbitrary Gaussian commotion to the images. We likewise attempted various things with abnormal down/up-scaling of the images and arbitrary bendy disFigment, but those expansions adversely affected department execution considering they dwindled the exactness of making ready marks for first-class and complicated cortical plate department.

In Fig.4 and 5, we've got proven version attention publications of the proposed community for 2 awesome datasets. As referenced above, to deal with a check image of erratic length, we follow our enterprise in a sliding window layout. To produce the attention maps for the whole image, at each region of the sliding window the attention grids (which can be of length IRN×N) are delivered alongside their segments to determine absolutely the attention paid to each one of the N patches via way of means of one of a kind patches withinside the block. Playing out this calculation in a sliding window layout and registering the voxel-savvy everyday gives us the attention maps displayed in those Figs. They display how a great deal attention is paid to all components of the image at some point of the time spent developing the department map.

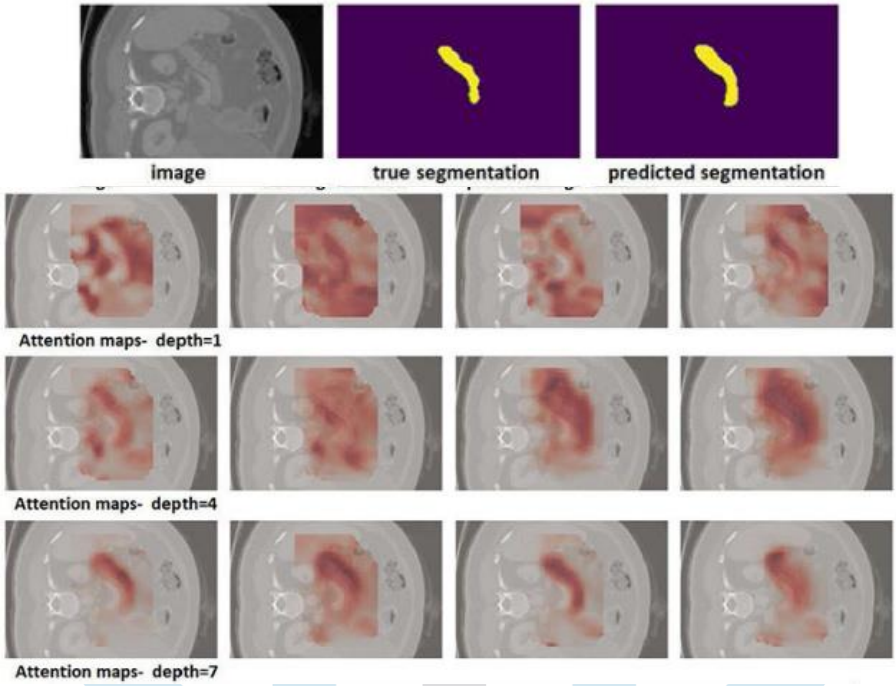
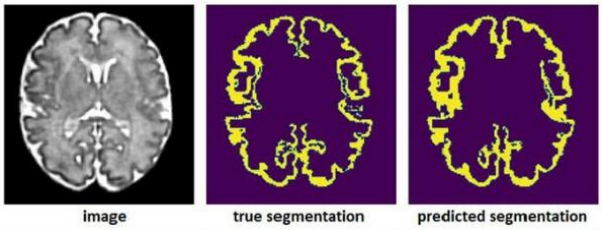


Figure 4. Example attention maps for two pancreas images. In this experiment, Attention maps for depths 1, 4, and 7 are shown.

Table. 3. The number of free parameters (n_{param}), number of floating point operations (FLOPS), and frames per second (FPS) for each of the FCNs and the proposed network. FLOPS and FPS are computed for processing patches of size 24^3 voxels.

Model	$n_{param} \times 10^6$	***FLOPS* $\times 10^9$**	***FPS***
SEFCN	2.58	4.21	191.44
DSRNet	3.77	3.43	174.1
Attention UNet	2.87	1.93	156.0
UNet++	3.81	4.13	159.1
Proposed	2.63	4.93	145.4

Tbl.3 indicates the amount of learnable obstacles, wide variety of drifting factor tasks (FLOPS), and casings every second (FPS). Failures and FPS are accounted for managing patches of length 243 voxels. We registered the FPS for all fashions on a NVIDIA RTX 2080TI GPU. In general, the fashions have reasonably similar wide variety of obstacles and computational expenses. Our proposed community has a relatively greater modest wide variety of obstacles than the checked out FCNs. Then again, the amount of FLOPS for the proposed community is higher, that is due to the large lattice duplications related to the attention modules. As a long way as getting ready time, our enterprise mixed in round 24 hours of GPU time, aleven though the FCNs joined in more or less four hours of getting ready. This can be due to the manner that transformer networks want more education time to assimilate the spatial examples withinside the image, aleven though FCNs` engineering makes this studying greater straightforward.



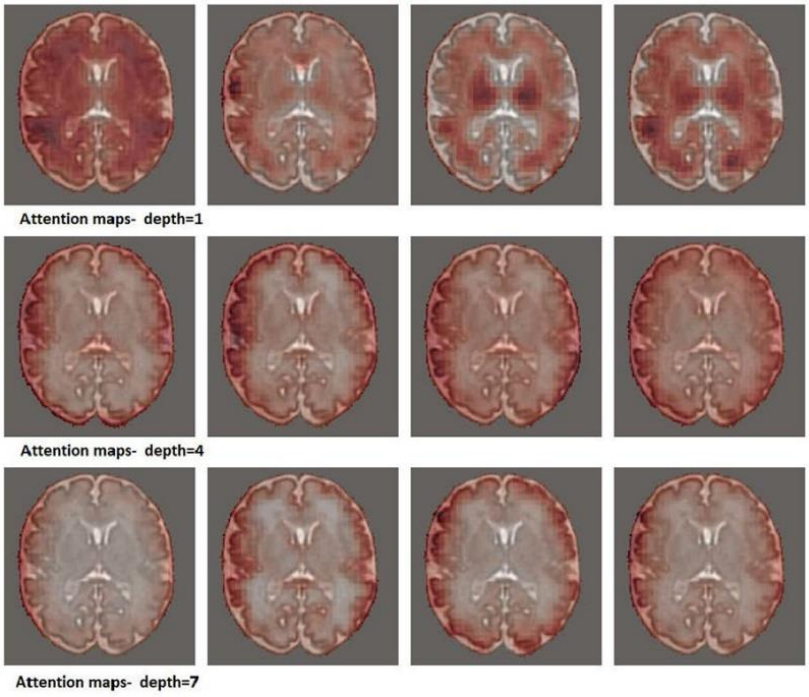


Figure 5. Example attention maps for a cortical plate image. In this experiment, Attention maps for depths 1, 4, and 7 are shown.

V. CONCLUSION

The convolution interest has critical regions of power for an withinside the layout of the mammalian vital visible cortex and it's far suitable for developing sturdy techniques for image demonstrating and image getting it. As of late, CNNs were established to be profoundly compelling in managing one of a kind PC imaginative and prescient issues. In any case, there's no first rate rationalization to anticipate that no different version can outflank CNNs on a selected imaginative and prescient task. Clinical image exam applications, specifically, act express problems such like 3-d nature of the photographs and modest wide variety of marked photographs. In such applications, one of a kind fashions can be greater a success than CNNs. In this paintings we brought any other version for 3-d scientific image department. Dissimilar to all new fashions that usage convolutions as their foremost constructing blocks, our version relies upon on self-attention among adjacent 3-d patches. Our consequences exhibit the manner that the proposed enterprise can beat slicing side FCNs on 3 scientific image department datasets. With pre-getting ready for denoising and in-portray obligations on unlabeled photographs, our enterprise likewise carried out higher in comparison to a FCN whilst simply 5-15 marked getting ready photographs had been accessible. We anticipate that the enterprise proposed on this paper have to be compelling for one of a kind undertakings in scientific image exam like abnormality identity and grouping.

REFERENCES:

[1]. Gibbs P, Buckley DL, Blackband SJ, and Horsman A, "Tumour volume determination from MR images by morphological segmentation," Phys. Med. Biol, vol. 41, no. 11, p. 2437, 1996. [PubMed: 8938037]

[2]. Wang Y, Guo Q, and Zhu Y, "Medical image segmentation based on deformable models and its applications," in Deformable Models. New York, NY, USA: Springer, 2007, pp. 209–260.

[3]. Mahapatra D and Buhmann JM, "Prostate MRI segmentation using learned semantic knowledge and graph cuts," IEEE Trans. Biomed. Eng, vol. 61, no. 3, pp. 756–764, Mar. 2014. [PubMed: 24235297]

[4]. Goldszal AF, Davatzikos C, Pham DL, Yan MXH, Bryan RN, and Resnick SM, "An imageprocessing system for qualitative and quantitative volumetric analysis of brain images," J. Comput. Assist. Tomogr, vol. 22, no. 5, pp. 827–837, Sep. 1998. [PubMed: 9754125]

[5]. Prince JL, Pham D, and Tan Q, "Optimization of MR pulse sequences for Bayesian image segmentation," Med. Phys, vol. 22, no. 10, pp. 1651–1656, Oct. 1995. [PubMed: 8551990]

[6]. Thompson PM and Toga AW, "Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations," Med. Image Anal, vol. 1, no. 4, pp. 271–294, Sep. 1997. [PubMed: 9873911]

[7]. Shelhamer E, Long J, and Darrell T, "Fully convolutional networks for semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell, vol. 39, no. 4, pp. 640–651, Apr. 2017. [PubMed: 27244717]

[8]. Chen L-C, Papandreou G, Kokkinos I, Murphy K, and Yuille AL, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell, vol. 40, no. 4, pp. 834–848, Apr. 2018. [PubMed: 28463186]

[9]. Bakas S et al. , "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, arXiv:1811.02629.

[10]. Bernard O et al. , "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" IEEE Trans. Med. Imag, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.
[11]. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, and Glocker B, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Med. Image Anal, vol. 36, pp. 61–78, Feb. 2017. [PubMed: 27865153]
[12]. Karimi D, Zeng Q, Mathur P, Avinash A, Mahdavi S, Spadinger I, Abolmaesumi P, and Salcudean SE, "Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images," Med. Image Anal, vol. 57, pp. 186–196, Oct. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841519300623 [PubMed: 31325722]
[13]. Zeng Q, Karimi D, Pang EHT, Mohammed S, Schneider C, Honarvar M, and Salcudean SE, "Liver segmentation in magnetic resonance imaging via mean shape fitting with fully convolutional neural networks," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2019, pp. 246–254.
[14]. Bastiani M, Andersson JLR, Cordero-Grande L, Murgasova M, Hutter J, Price AN, Makropoulos A, Fitzgibbon SP, Hughes E, Rueckert D, Victor S, Rutherford M, Edwards AD, Smith SM, Tournier J-D, Hajnal JV, Jbabdi S, and Sotiropoulos SN, "Automated processing pipeline for neonatal diffusion MRI in the developing human connectome project," NeuroImage, vol. 185, pp. 750–763, Jan. 2019. [PubMed: 29852283]
[15]. Hesamian MH, Jia W, He X, and Kennedy P, "Deep learning techniques for medical image segmentation: Achievements and challenges," J. Digit. Imag, vol. 32, no. 4, pp. 582–596, Aug. 2019.
[16]. Taghanaki SA, Abhishek K, Cohen JP, Cohen-Adad J, and Hamarneh G, "Deep semantic segmentation of natural and medical images: A review," Artif Intell. Rev, vol. 54, no. 1, pp. 1–42, 2020. [PubMed: 32836651]
[17]. Ghafoorian M, Mehrtash A, Kapur T, Karssemeijer N, Marchiori E, Pesteie M, Guttmann CRG, de Leeuw F-E, Tempany CM, van Ginneken B, Fedorov A, Abolmaesumi P, Platel B, and Wells WM III, "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2017, pp. 516–524. [18]. Karimi D, Warfield SK, and Gholipour A, "Critical assessment of transfer learning for medical image segmentation with fully convolutional neural networks," 2020, arXiv:2006.00356.
[19]. Gao Y, Phillips JM, Zheng Y, Min R, Fletcher PT, and Gerig G, "Fully convolutional structured LSTM networks for joint 4D medical image segmentation," in Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), Apr. 2018, pp. 1104–1108.
[20]. Bai W, Suzuki H, Qin C, Tarroni G, Oktay O, Matthews PM, and Rueckert D, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2018, pp. 586–594.
[21]. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, and Zhou Y, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, arXiv:2102.04306.
[22]. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, Cook SA, De Marvao A, Dawes T, O'Regan DP, and Kainz B, "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," IEEE Trans. Med. Imag, vol. 37, no. 2, pp. 384–395, Feb. 2017. [23]. Karimi D, Samei G, Kesch C, Nir G, and Salcudean SE, "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models," Int. J. Comput. Assist. Radiol. Surg, vol. 13, no. 8, pp. 1211–1219, Aug. 2018, doi:10.1007/s11548-018-1785-8. [PubMed: 29766373]
[24]. Lee C-Y, Xie S, Gallagher P, Zhang Z, and Tu Z, "Deeply-supervised nets," in Proc. Artif. Intell. Statist, 2015, pp. 562–570.
[25]. Malle Raveendra and K Nagireddy "Inter frame Tampering Detection based on DWT-DCT Markov Features and Fine tuned AlexNet Model," Vol. 20 No. 12 pp. 1-12, http://paper.ijcsns.org/07_book/202012/20201201.pdf.
[26]. Ren S, He K, Girshick R, and Sun J, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. Adv. Neural Inf. Process. Syst, 2015, pp. 91–99.
[27]. LeCun Y, Bengio Y, and Hinton G, "Deep learning," Nature, vol. 521, no. 7553, p. 436, 2015. [PubMed: 26017442]
[28]. Le Cun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, and Jackel L, "Handwritten digit recognition with a back-propagation network," in Proc. 2nd Int. Conf. Neural Inf. Process. Syst., 1989, pp. 396–404.
[29]. Goodfellow I, Bengio Y, Courville A, and Bengio Y, Deep Learning, vol. 1. Cambridge, MA, USA: MIT Press, 2016.
[30]. Olshausen BA and Field DJ, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature, vol. 381, no. 6583, pp. 607–609, Jul. 1996. [PubMed: 8637596]
[31]. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I, "Attention is all you need," 2017, arXiv:1706.03762.
[32]. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, and Shah M, "Transformers in vision: A survey," 2021, arXiv:2101.01169.
[33]. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, and Houlsby N, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, arXiv:2010.11929.
[34]. Milletari F, Navab N, and Ahmadi S-A, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. 4th Int. Conf. 3D Vis. (3DV), Oct. 2016, pp. 565–571.
[35]. Lei Ba J, Kiros JR, and Hinton GE, "Layer normalization," 2016, arXiv:1607.06450.
[36]. Zhou Z, Siddiquee MMR, Tajbakhsh N, and Liang J, "UNet++: A nested U-Net architecture for medical image segmentation," in Deep Learning in Medicallmage Analysis and Multimodal Learning for Clinical Decision Support. Cham, Switzerland: Springer, 2018, pp. 3–11.

[37]. Ronneberger O, Fischer P, and Brox T, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2015, pp. 234–241. [38]. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, and Rueckert D, "Attention U-Net: Learning where to look for the pancreas," 2018, arXiv:1804.03999.
[39]. Roy AG, Navab N, and Wachinger C, "Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks," IEEE Trans. Med. Imag, vol. 38, no. 2, pp. 540– 549, Aug. 2018. [40]. Hu J, Shen L, and Sun G, "Squeeze-and-excitation networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.
[41]. Dou H, Karimi D, Rollins CK, Ortinau CM, Vasung L, Velasco-Annis C, Ouaalam A, Yang X, Ni D, and Gholipour A, "A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal MRI," IEEE Trans. Med. Imag, vol. 40, no. 4, pp. 1123–1133, Apr. 2021.
[42]. Isensee F, Kickingereder P, Wick W, Bendszus M, and Maier-Hein KH, "No new-net," in Proc. Int. MICCAI Brainlesion Workshop. Cham, Switzerland: Springer, 2018, pp. 234–244.
[43]. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, and Jégou H, "Training data-efficient image transformers & distillation through attention," 2020, arXiv:2012.12877.
[44]. Kingma DP and Ba J, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), 2014, pp. 1–13.
[45]. Devlin J, Chang M-W, Lee K, and Toutanova K, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805.
[46]. Karimi D, Warfield SK, and Gholipour A, "Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations," Artif. Intell. Med, vol. 116, Jun. 2021, Art. no. 102078. [Online]. Available: https:// www.sciencedirect.com/science/article/pii/S0933365721000713
[47]. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, Glocker B, King A, Matthews PM, and Rueckert D, "Semi-supervised learning for network-based cardiac mr image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2017, pp. 253–260. [48]. Bilic P et al. , "The liver tumor segmentation benchmark (LiTS)," 2019, arXiv:1901.04056.
[49]. Heller N, Sathianathen N, Kalapara A, Walczak E, Moore K, Kaluzniak H, Rosenberg J, Blake P, Rengel Z, Oestreich M, Dean J, Tradewell M, Shah A, Tejpaul R, Edgerton Z, Peterson M, Raza S, Regmi S, Papanikolopoulos N, and Weight C, "The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," 2019, arXiv:1904.00445.
[50]. Tolstikhin I, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J, Lucic M, and Dosovitskiy A, "MLP-mixer: An all-MLP architecture for vision," 2021, arXiv:2105.01601.
[51]. Voita E, Talbot D, Moiseev F, Sennrich R, and Titov I, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," 2019, arXiv:1905.09418.
[52]. Behnke M and Heafield K, "Losing heads in the lottery: Pruning transformer attention in neural machine translation," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 2664–2674.