International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 07 | July - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Comparative Study on Attention Unet for Image Segmentation and Propose Inception-Attention based Modification to the Architecture

Joshua Ryan Lawrence Dsouza
Student of MTech DSML, PES University Bengaluru, India joshuarld@yahoo.com

Ruchita Singhania
Mentor,
Great Learning,
Bengaluru, India
ruchita.agarwal@gmail.com

***Abstract*—The advent of Unet in 2015, by Ronneberger et al, the same brought about a uproar in the field of image segmentation and its application in Medical sector. Since then, Numerous efforts have been contributed to by researchers to enhance base architecture and bring about better yield from the algorithm. The main goal for the research being, getting accurate segmentation, yet maintaining the integrity of the result, this has been contributed to by the researchers in the field. This study is undertaken to understand the variations and improvements proposed understand the complexities involved, Implement the base architecture and few of the improvements find the best combination that achieves the desired outcome. A wide known application of Image segmentation algorithms is in Medical Sector. Mainly in identification and isolation of tumors and unknown growths in the human anatomy. The main concern here being, the time constraint, mainly from the first report of the problem to its diagnosis and then treatment. The problem towards which the proposed system adds an edge is to reduce the time needed to check and identify the unknown growth in the regions, using the MR Images of the area Based on the results of the study, the paper discusses about the proposal is to incorporate enhancement, that would yield a better yet accurate segmented result from the base algorithm. From the Research contributed to thus far we have seen attention gates, inception networks with dense inception and other variants, and swin transformer blocks come in as improvements to the base architecture of Unet. Current considerations for the enhancements are Attention Gating mechanism and Inception Mechanism to incorporated into the Base of Unet. The same would be gauged on its performance over the MRI dataset as a Base and later on check its performance other Datasets, the metric for the gauging under consideration would be DICE Scores and or IoU**

***Index Terms*—UNet, Attention Gates, Inception Blocks, Inception Network, Inception V1 Convolution Neural Network, Attention-UNet, Inception UNet.**

I. Introduction

Semantic segmentation is a computer vision method that assigns markers to each pixel in an image, aiming to member it into meaningful regions grounded on visual appearance. Unlike other segmentation methods, semantic segmentation classifies every pixel, capturing fine details and accurate object boundaries. It provides context and scene understanding by considering object relationships, helping differentiate visually similar but functionally different objects. Various approaches, such as CNNs, FCNs, and encoder-decoder architectures, are used for semantic segmentation, trained on annotated image datasets to learn pixel-label relationships.

Semantic segmentation finds applications in various fields, including medical imaging. It aids in identifying and classifying structures like organs, tumors, and blood vessels. However, current medical image segmentation systems face challenges in time efficiency, diagnosis accuracy, and handling occlusions and object boundaries. CNNs have made significant strides in medical image analysis but face limitations in terms of speed and handling complex shapes. Labeled training data requirements pose another challenge, as acquiring such data for all scenarios is often challenging and expensive.

To address these challenges, researchers have contributed to enhancing the base architecture, such as the widely adopted Unet model. Attention gates have also been introduced, sup- pressing irrelevant regions and improving model sensitivity and prediction accuracy. Inception blocks, as used in the Inception network, capture information at different scales by using parallel convolutional layers with varying filter sizes. These blocks enable the extraction of relevant features.

This paper focuses on implementing a base architecture of Unet with attention gates and explores the potential improvements achieved by incorporating inception blocks or the inception network. The goal is to enhance image segmentation performance beyond the baseline seen during the implementation of the same.

II. Related Works - Literature Survey

U-Net [12] is one of the most widely utilized scalable deep learning models for segmenting biological images. This architecture has been continually developed by several scholars. The issue relating to imaging modalities and the diversity of many organs has still not been resolved. The complexity of the photos makes it difficult for the physicians to jointly diagnose the patient, which takes time. Additionally, bias is introduced by manual segmentation if it incorporates medical professionals' own judgements. The use of deep learning

techniques in medical image analysis has attracted greater attention since it makes it possible for organs, lesions, and tumors to be automatically identified and segmented.

The majority of currently used fully convolutional neural network (FCNN) using U-shaped structure-based medical picture segmentation techniques. A symmetric encoder-decoder with skip links makes up the usual U-shaped network, U-Net [12]. To extract deep features with huge receptive fields in the encoder, a succession of convolutional layers and continuous down-sampling layers are utilized. The high-resolution features from the encoder's various scales are fused using skip connections to prevent the loss of spatial information caused by down-sampling before the decoder up-samples the recovered deep features to the input resolution for pixel-level semantic prediction. Thanks to its lovely structural architecture, U-Net has achieved amazing success in a variety of medical imaging applications. Following this technical path, several algorithms for image and volumetric segmentation of various medical imaging modalities have been developed, including 3D U-Net [2], residual-Unet [4], attention-UNet [10], and UNet3+ [6]. This fully convolutional neural net-works - based approach's superior performance in cardiac segmentation, organ segmentation, and lesion segmentation demonstrates convolutional neural network 's potent capacity to learn discriminating characteristics.

Since the introduction of the U-Net architecture [12], several variations and improvements have been proposed to enhance its performance in different applications.one of the variation of the U-Net architecture is the Attention U-Net proposed by Oktay et al. in 2018 [10], which introduces an attention mechanism to focus on relevant regions of the image for the segmentation task. The Attention Mechanisms came to about, where Bahdanau Attention [1] and Luong attention [9], where the concepts of additive and multiplicative attention come about. Additive attention, also known as "Bahdanau attention,"
[1] weights the input by learning a set of weights that are summed to produce the output. Multiplicative attention, also known as "Luong attention," [9] weights the input by learning a set of weights that are multiplied with the input to produce the output. Additive attention [1] and has been widely used in natural language processing tasks. It has the advantage of being able to handle alignment errors, which occur when the attention mechanism is not able to align the input and output correctly. However, additive attention is computationally expensive because it requires computing a weighted sum for each element in the input. Multiplicative attention [9] and has the advantage of being faster and more memory efficient than additive attention. However, multiplicative attention is sensitive to alignment errors and can struggle with long input sequences. In the case of Attention Unet [10], the attention mechanism is used to enhance the segmentation performance of the UNet model by allowing it to selectively focus on important regions of the input image and the Widely known Additive attention [1] is employed.

In addition to medical image segmentation, the U-Net architecture has been used in other domains such as building detection, hand action recognition, and character recognition, among others. With The Paper Rethinking the Inception Architecture for Computer Vision [15] proposed a number of upgrades which increase the accuracy and reduce the Computational complexity of a deep convolutional network. Delibaşoğlu and Cetin in 2020 proposed improved U-Nets with Inception blocks for building detection [3], achieving better performance than the original U-Net architecture. S. Rubin Bose and Kumar in 2020, proposed an efficient Inception V2-based deep convolutional neural network for real-time hand action recognition [14], which outperformed other hand recognition methods in terms of accuracy and speed. Guo et al. in 2022 suggested an enhanced neural network model for oracle bone inscription character identification based on Inception-v3, which attained state-of-the-art performance on the test [5].

In summary, the U-Net architecture and its variations have shown promising results in various image segmentation tasks in different domains. These architectures have enabled re- searchers to achieve state-of-the-art performance on different segmentation tasks, and they continue to be an active area of research.

III. PROPOSED WORK

Since its advent, Unet has become one of go to architectures for the image segmentations tasks. Over shadowing its predecessors, the CNN while they were good at the task, but when it came to understand from the images down to the level of pixels, with multi-class classification and training with multiple images can present and unique challenge that can question: the compute capacity needed. The training times raising questions If they could handle the loads. This is where Unet comes in, bring a Seq2Seq models by classification of its architecture. Is known to have encoder and decoder part, where in the encoder part the same is known as contracting path. As the same goes on convoluting the image, such as contracting it while maintaining the feature and the decoder part known as the expanding path, expands the images while maintaining the feature. The Unet make use of Skip connections that parse and pass the feature information from each encoder block to the decoder. which allows it to effectively combine high-level features from the contracting path with low-level features from the expansive path, resulting in more accurate segmentation. and low-level features make it a popular choice for medical image segmentation tasks. Using extremely small trade samples, U-net creates extremely precise segmentation maps, which is what makes it so helpful. As correctly labelled pictures are frequently scarce, this quality is crucial for the medical imaging community. This is accomplished by applying random elastic deformation to the training data, allowing the network to pick up these differences without needing fresh labels.

Post its inception, lot of variations and modifications have been seen, one specific one being about the integration of attention gating mechanism, devised by Ozan Oktay et al. the method/procedure made use of the attention gating mechanism

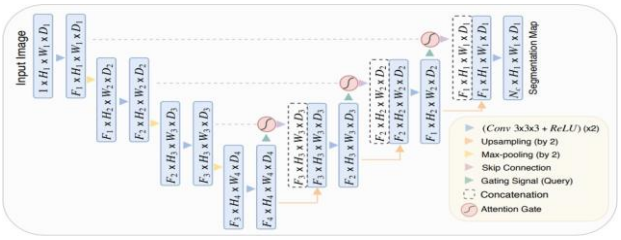

Fig. 1. [10] The architecture of Attention-Unet, which is composed of encoder, decoder and skip connections and the attention gating mechanism at the decoder blocks

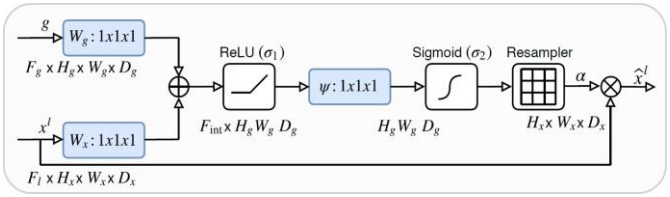

Fig. 2. [10] The AG Mechanism employed.

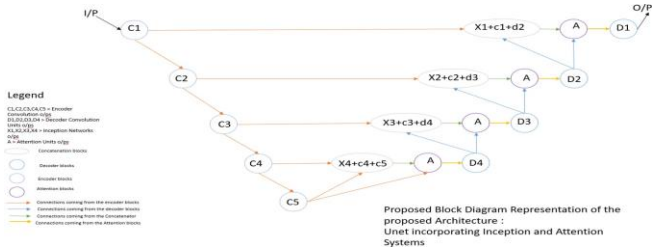

Fig. 4. The layout of the architecture of the same described as Mod 1

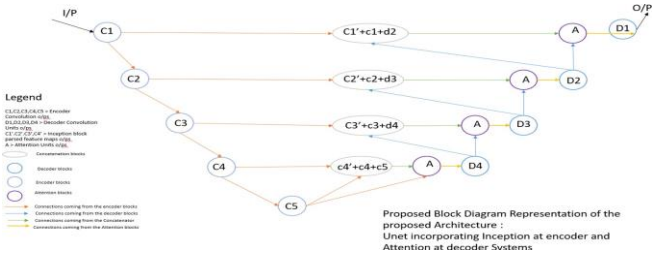

Fig. 5. The layout of the architecture of the same described as Mod 2

and produced yet accurate results, in this method, the attention gating mechanism is used on the feature maps that are parsed on the skip connections between the encoder – decoder blocks. Figure (1). Shows the layout of the architecture same.

And the other being about UNet's integration with inception blocks, proposed by Punn and Agarwal, the method/procedure made use of the network that can collect information at various scales and generate reliable results because to the Inception module's/block's many parallel convolutional layers with various filter sizes. Figure (3). Shows the layout of the architecture of the inception Blocks used in the architecture.

A. Network Architecture Overview

To improve the architecture further, I propose two layouts of the Unet. Namely Mod 1 and Mod 2: collectively termed as Inception-Attention Unet.

The Mod 1 the inception network used the same way as that of the Inception Unet, with underlying architecture being

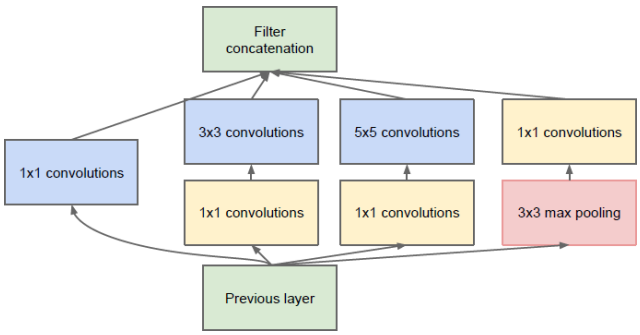

Fig. 3. [11] The layout of the architecture of the same inception block

Attention Unet, where at the decoder end: staged inception output, Feature map parted from the encoder are combined with previous stage decoder outcomes prior to undergoing attention mechanism. and being parsed to the decoder.

The Mod 2 every feature map parsed over the skip connection also under goes through the inception block, with underlying architecture being Attention Unet, where at the decoder end: inception output of the parsed feature map, Feature map parted from the encoder are combined with previous stage decoder outcomes prior to undergoing attention mechanism. and being parsed to the decoder.

With the Mod 2 is theorized to pay more detail during the learning stage of the Model,

B. Data Augmentation

For Deep Learning models to be trained on tasks like segmenting medical images, data addition is a pivotal strategy. Since there are constantly many annotated prints available for these jobs, it might be delicate to make a strong model that's good at generalizing to new data. In order to instinctively expand the size of the training set, several changes are applied to the given and as part of data addition. Common addition ways used in medical picture segmentation tasks include arbitrary reels, flips, elastic distortions, and intensity shifts. The model may learn to come steady to similar changes thanks to these variations, which reflect the natural diversity of medical. In order to train the model to member the objects of interest under colorful distortion situations, elastic distortions, for illustration, imitate genuine towel distortions that are constantly seen in medical imaging. analogous to illumination changes, intensity shifts allow the model to develop object segmentation chops in a variety of lighting circumstances. By applying data addition, the model is exposed to a larger variety

of training samples, making it more robust to new images. As a result, the model is better equipped to handle the variability and query frequently present in medical images, leading to bettered segmentation performance.

C. Training

The input images and their corresponding segmentation maps are used to train the network with Stochastic Gradient Descent as the optimizer. A momentum of 0.05 with dampening of 0.003 and weight decay of 0.0001 is used to achieve a smoother learning curve where a learning rate of 0.01 is set. these set of parameters were found post rigorous experiments to achieve a smoother curve that could achieve the model to learn better all the while ensure the same did not over-fit.

The initialization of weights is a vital step in deep networks with several convolution layers and many routes. Certain network components may contribute excessive activations as a result of poor weight initialization, whereas other components may not participate at all. The initial weights should ideally be changed to give each feature map in the network a variance of about one. This may be done by choosing the starting weights for a network using the U-Net design, which alternates convolution and ReLU layers.

The models taken under for comparison have been trained using the Dice loss function. The gap between the anticipated segmentation mask and the ground truth segmentation mask is measured by the Dice loss, a frequently used loss function for image segmentation applications. The Dice loss penalizes false negatives and false positives while encouraging the projected segmentation mask to have high overlap with the actual segmentation mask.

For evaluating the performance of segmentation algorithms, the DICE score, also known as the Sorensen-Dice score (1), is a commonly adopted metric.

$$DICE = (2 * |A \cap B|)/(|A| + |B|) \quad (1)$$

where —A n B— is the count of pixels in both the predicted and actual masks, —A— is the count of pixels in the predicted mask, and —B— is the number of pixels in the ground truth mask.

The advantage of the DICE score (1) is that it is sensitive to both the true positives (pixels correctly identified as belonging to the target region) and the false positives and negatives (pixels incorrectly classified). It provides a balanced mea- sure taking into account the class imbalance errors, affecting the segmentation accuracy, and accounting for both under-segmentation and over-segmentation errors.

IV. Implementation

A. Dataset: LGG segmentation dataset

The Dataset used to test initial implementation is the Kaggle Brain MRI Segmentation Dataset the dataset was sourced from TCIA (the cancer imaging archive) under the TCGA program (the cancer genome atlas program) based as LGG Segmentation Dataset. The same is avail- able here https://www.kaggle.com/datasets/mateuszbuda/lgg- mri-segmentation. The early diagnosis of Brain tumors, and irregularities such as aneurysms are vital to avoid fatalities and or effective treatments plans and recovers, as it is one of the life threatening aliments that affect people worldwide. Deep learning has emerged as a powerful tool for fast and accurate pattern recognition in medical applications, including diagnosing the irregularities in the Brain, with its ability to analyze large amounts of data, deep learning can provide valuable assistance to medical specialists during the diagnosis stage. To evaluate the performance of the model, MR Images of the brain were provided by the TCIA gathered under the TCGA program for the LGG segmentation challenge on the widely known assessment platform. The training dataset contains 3000+ MR images. In this dataset, images and the masks approved by board certified Radiologist from Duke university. the images are varied sizes hence they are resized to 128 by 128.

B. Experimental setup

The current base code was working with the Image size as 128x128, having 1 class. The Batch Size of 10 for 30,60,100 epochs. A wrapper around the base code was introduced for more efficient check pointing. Re-adaptation of the model was done as well where the difference in layers was checked Vs Optimizers. Model training is conducted on CPU run time on Intel 12900K with 64GB memory.

C. Implementation details

The Models are achieved based on Python 3.9 and PyTorch 2.0. For all training cases, data augmentations such as flips and rotations are used to increase data diversity. The input image size and patch size are set 512x512 and 128x128 respectively. The model training was done on a CPU Compute loads on i9-12900k with 64 GB Memory. During the training phase, a batch of 10 is created, and our model is optimized using the SGD [13] optimizer with momentum 0.05 and weight decay 0.0001. This results in a better learning curve. There were also different parameters tried on deriving variations of five-layered architecture and three-layered architecture for the attention UNet base. Post that as the observations derived that the five layered variant was better, rest of models that were derived were based off of the 5-layer variation. Where both were tested with Adamax and SGD [7] Optimizers with various parameters to understand which give a better learning curve. with Adamax [8] parameters worked with were: learning rates of 0.01, 0.001, 0.1, 0, 1, 10. With SGD [13] parameters worked with were: learning rates of 0.01 momentum 0.05 and weight decay 0.0001 and dampening=0.003. there were two rounds of training's done: in round 1: with SGD and model activation function being ReLU in round 2 : was with Adamax with learning rate as 0.01. and model activation function being Leaky ReLU.

V. Results and Discussion

The models Mod 1 and Mod 2 have a total of 33,609,421 trainable parameters, which is optimized compared to Attention Unet : 34878573, however Slightly higher when compared to Inception UNet : 32042985 . As a result, a lower batch size of 10 is employed during training.

For the evaluation of the segmentation models, a commonly used segmentation metric called Dice Score Eq.1 is used. The loss function used is Dice Loss.

There below are the metrics used to gauge the results:

- Dice Loss
- Train Dice Score
- Test Dice Score

the Metrics are recorded and later plotted to capture the trace of the learning curves and gauge the loss seen on training.

A. Over LGG segmentation dataset

Here is a comparative analysis between the Models: Attention-Unet and Inception UNet with Mod 1 and Mod 2 over on LGG segmentation dataset is presented below. focusing on the evaluation results obtained:

Here in the test runs as there were various variations used with various optimizers, below are tabulated metrics

TABLE I
Mean DICE Score range on Train Set

Model	DICE Loss	Train DICE	Test DICE
Att-UNet 3L 30 ep	0.3052	0.6944	0.15-0.64
Att-UNet 5L 60 ep	0.2346	0.7902	0.82
Att-UNet 5L 100 ep	0.16	0.8566	0.8399
Incep-UNet 5L 150 ep	0.13	0.8790	0.8886
Mod 1 5L 150 ep	0.13	0.8917	0.8961
Mod 2 5L 300 ep	0.14	0.8674	0.8897
Mod 2 LRM 5L 60 ep	0.15	0.8492	0.9371

In initial stages of the research the differences with Optimizers and model modification are tested with variation of epochs to learn:

- to check if addition of more layers would have an impact or not
- With more layers would there be an impact on the training time
- with different optimizers would the same contribute to the learning curve

The Modifications done here are:

- test of three layered Attention Unet Vs Five Layered Attention Unet
- Variation of Optimizers and the learning parameters
- variation in number of Epochs Vs Number of Layers and the learning curve observed

Experimental results show that the model:

- comparing layers: more layers. Less epochs [30-80 range] algorithm can learn better against the one with less layers.
- with finer parameters set the learning curve of the model gets evened out and better results are seen.

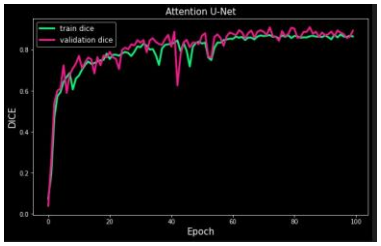

Fig. 6. Attention Unet train scores

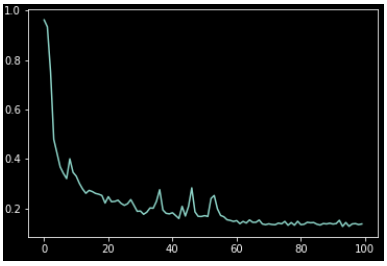

Fig. 7. Attention Unet training loss

- with the tests conducted on a given data set, the results are closely in line with ones mentioned in the base paper: where the algorithm is seen to derive the outcome of 82- 84 percent mean test IoU and observe a validation DICE in the range of 82-90m percent [on peaks]

In later stages of the research for Devised models, with Modifications based on the placements of the inception and attention block the study was undertaken with below objectives.

- understand if the models underwent parameter reduction
- would the modification yield a better result?
- what modification works bests for devised models The

Modifications done here are:

- Mod 1 is a modification over Attention Unet Inception Unet: where both features Attention gates and Inception network are incorporated.
- the modification of Mod 2 over Mod 1: was that the Mod 2 had inception block parsing the feature maps at each stage at the encoder instead of having inception network
- Variation of Optimizer and the learning parameters and activation functions

Experimental results show that the model:

- comparing the models, we do see Mod 2 has with Leaky ReLU trains faster compared with ReLU
- With tests done we see Mod 1 having a better outcome when compared with all models with ReLU
- with the tests conducted on a given data set, the algorithm [mod 1 and mod 2] is seen to derive the outcome of 90-95 percent mean test IoU over Base model that was tested on in earlier phase

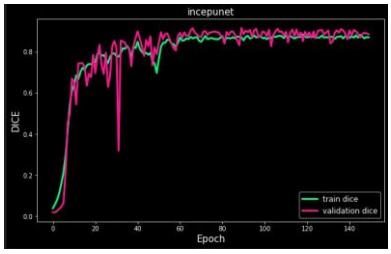

Fig. 8. Inception Unet train scores

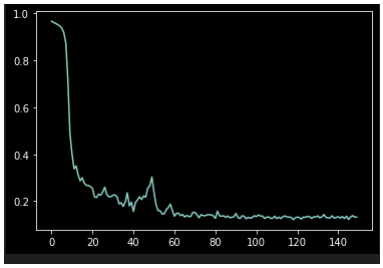

Fig. 9. Inception Unet training loss

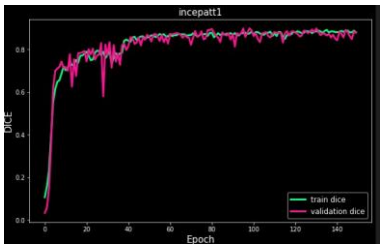

Fig. 10. Mod 1 train scores

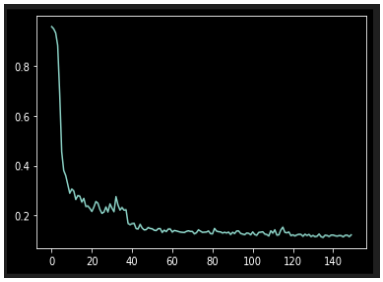

Fig. 11. Mod 1 training loss

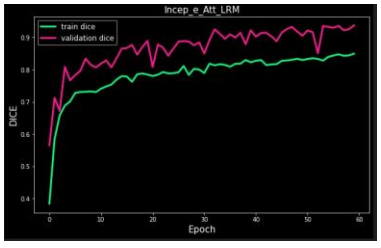

Fig. 12. Mod 2 train scores

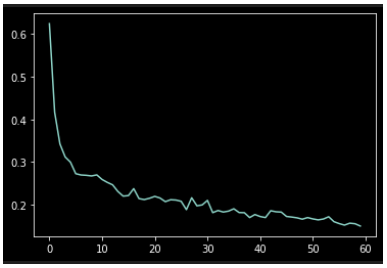

Fig. 13. Mod 2 training loss

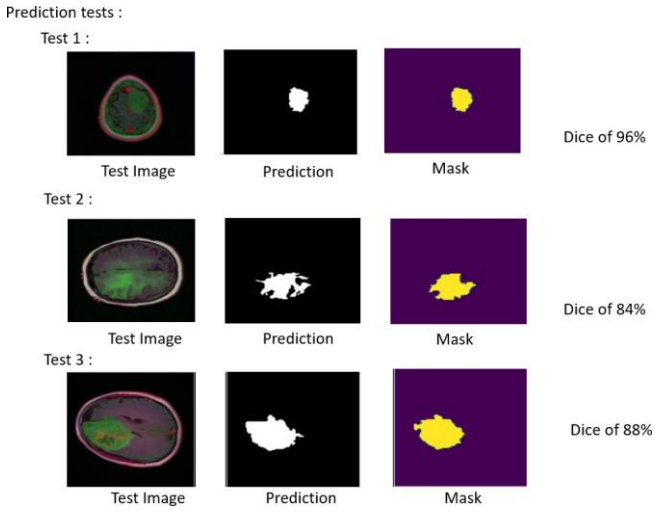

Fig. 14. Sample test predictions that were seen and their respective DICE Scores from the test runs under Attention Unet

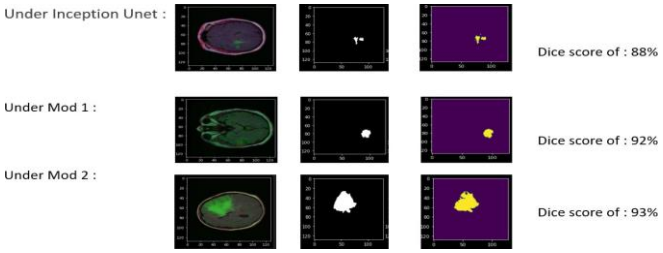

Fig. 15. Sample test predictions that were seen and their respective DICE Scores from the test runs in Under Inception Unet, Mod1 and Mod2

VI. CONCLUSIONS

As a consequence of this investigation, it is clear that the Attention Unet and Inception Unet architecture is useful for segmenting medical picture data. it is also seen that MOD 1 and MOD 2 Architectures also can contribute to the field as same as base principal architectures.

For the patient's prognosis, diagnosis, and therapy, the model does offer correct segmentation. It has also been shown that by adding variables, which have an impact on the model's learning curve during training, the variations may be con-

trolled, regularized, and smoothed. This in turn influences the model's learning, leading to improved but consistent results.

That said, however the same leaves of room for future work, where the same can be added to and enhanced further, where variations of layers, addition of modules and or combination of parallel working or running mechanisms can be added where the same can aid, contribute to and improve the give model. This is including but is not limited to the parameter optimization as well, such as batch size, hyper parameters etc. to name a few.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[2] Ozgun Cicek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.

[3] ˙Ibrahim Delibas¸o˘glu and Mufit Cetin. Improved u-nets with inception blocks for building detection, 11 2020.

[4] Foivos I. Diakogiannis, Franc¸ois Waldner, Peter Caccetta, and Chen Wu. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data, 2019.

[5] Ziyi Guo, Zihan Zhou, Bingshuai Liu, Longquan Li, Qingju Jiao, Chenxi Huang, and Jianwei Zhang. An improved neural network model based on inception-v3 for oracle bone inscription character recognition, 05 2022.

[6] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation, 2020.

[7] Satyen Kale, Ayush Sekhari, and Karthik Sridharan. Sgd: The role of implicit regularization, batch-size and multiple-epochs, 2021.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[9] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

[10] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[11] Narinder Punn and Sonali Agarwal. Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images, 03 2020.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[13] Sebastian Ruder. An overview of gradient descent optimization algorithms. 09 2016.

[14] Rubin Bose S. and Sathiesh Kumar. Efficient inception v2 based deep convolutional neural network for real-time hand action recognition, 03 2020.

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and ZB Wojna. Rethinking the inception architecture for computer vision, 06 2016.