IMPROVING MULTI-SCALE ATTENTION NETWORKS: BAYESIAN OPTIMIZATION FOR SEGMENTING MEDICAL IMAGES

A PREPRINT

Jimut Bahan Pal*
Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Howrah, India
22d1594@iitb.ac.in

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah, India
dripta.academic@gmail.com

October 24, 2025

ABSTRACT

Current deep learning based image segmentation methods are notable for their use of large number of parameters and extensive computational resources in training. There is a persistent need for more efficient flexible systems without compromising on precision. This work proposes a novel model that combines the best of deep learning and probabilistic machine learning to segment a wide variety of medical image datasets with state-of-the-art accuracy and limited resources. The approach benefits from the introduction of new diverse attention modules that serve multiple purposes including capturing of relevant information at different scales. These proposed attention modules are generic and can potentially be used with other architectures to boost performance. In addition, Bayesian optimization is employed to tune multi-scale weight hyperparameters of the model. The architecture combined with one of the proposed novel attention modules and tuned hyperparameters achieves the best results in segmenting ISIC 2017, LUNGS, NERVE, Skin Lesion, and CHEST datasets. Finally, the explainability of the network is analyzed by visualizing the feature map learned from the attention modules.

Keywords Attention Network · Bayesian Optimisation · Medical Image Segmentation · Min Pooling · Multi-scale Losses · Residual Networks

1 Introduction

Image segmentation methods are employed in a wide variety of domains in our day-to-day life, e.g., in self-driving cars [1] for decision making, amidst industrial robotics [2] for aiding interactions with the dynamic world, in security [3], as well as in unmanned aerial vehicles [3] for vigilance. Another important area of application is medical image analysis [4, 5] where it is used to infer critical information about the shapes and sizes of different organs for automatic generation of segmentation masks. Biomedical image segmentation often relies on scarce and imbalanced data, making it more challenging to develop data-hungry deep learning models. Several architectures have been proposed in the recent literature, each with its own capabilities and limitations.

*Review copy, donot distribute. Codes and models will be available only after publication at: https://github.com/Jimut123/bmsan. Current affliation of author is at Centre for Machine Intelligence and Data Science, Indian Institute of Technology, Bombay, Powai, Mumbai, Maharashtra. This work was a part of the Machine Learning 2 course project while the first author was at RKMVERI.

The medical-imaging community is in pressing need of segmentation architectures with low resource requirements and high precision performance capability. This work is driven by the motivation to develop single-class segmentation models that can operate using fewer resources in terms of model parameters and at the same time improve on the reported metrics without data augmentation or transfer learning. We have created two novel architectures for this purpose and used them to perform benchmarks on the already available state-of-the-art (SOTA) architectures in a wide variety of datasets. This study also investigates the explainability of deep learning architectures that reveals the superior performance of the proposed attention module in terms of incorporating prominent features toward the final segmentation results. The main contributions of this work may be summarized as follows:

- Two models are developed, i.e., Modified U-Net based on a fine-tuned and heavily modified version of MobileNetV2 [6] in the form of encoder-decoder architecture. The other is Multi-Scaled Attention Networks (MSAN) by taking motivation from the popular MultiResUNet [7] and additionally passing different sized (multi-scaled) inputs to the model, which shows significant improvements in terms of previous segmentation models.
- Three types of attention modules have been proposed that can potentially be used with other architectures to boost performance.
- Bayesian Optimization (BO) [8] is used for obtaining the optimized weights associated with the multi-scaled mask's features for improving segmentation by using a multi-scaled weighted loss function. The new model, with the weight hyperparameters tuned using BO, is called Bayesian Multi-Scaled Attention Network (BMSAN).
- The proposed model, with the attention modules incorporated, has lesser number of parameters compared to U-Net, while surpassing the current SOTA models with 4x more parameters in most of the datasets.

The paper is organized as follows: First, the related literature in the field of medical image segmentation is reviewed, followed by discussions on the datsets used for comparison, formulation of the model's multi-scale loss function, description of the proposed novel attention modules, and implementation of BO for performance enhancement. We then present the results of the benchmarking evaluations of the proposed architecture with respect to several SOTA models on diverse datasets. An ablation study is also undertaken and finally, the paper is concluded with discussions on the future scope and limitations of the work.

2 Literature Review

The U-Net architecture [9] and its variants have been successfully employed for segmenting biomedical image datasets. Some alterations of the U-Net [10] use contractive paths to capture context and symmetric expanding path for precise localization of segmentation masks. The addition of attention gates to U-Nets assists in capturing salient features, of varying shapes and sizes, that have been found to be useful in certain specific tasks [11, 12, 13]. Attention gates also help in suppressing irrelevant regions, thereby increasing the prediction accuracy of models. Some other works have incorporated Recurrent Convolutional Neural Networks (RCNN) as well as Recurrent Residual Convolutional Neural Network (RRCNN) [14] in the U-Net framework to achieve better segmentation results. The residual part of the network generally helps in training deep architectures without vanishing gradients, while the recurrent residual convolutional layers help in ensuring better feature representation for segmentation tasks. Bidirectional Convolutional Long Short Term Memory (LSTM) [15] models mostly helps in increasing performance and feature reuse. These methods are commonly used for elevating the capabilities of U-Net with less computation. Other networks [16] can exploit the capabilities of U-Net to find delineations in medical images.

An alternative modification to the U-Net is One-pass Multi-task Network (OM-Net) [17], based on a philosophy that humans tend to learn concepts much better when they are presented in the increasing order of difficulty. Residual blocks in U-Net architectures predominantly help in boosting performance, and skip connections that primarily reduce the distance between feature maps of the encoder and the decoder. This formation is further aided by the attention mechanism in the dense architecture that helps to focus on the most relevant information in OM-Net. Similarly, FocusNet [18] yields highly competitive performance over U-Net and its residual variants, incorporating attention within the Convolutional Neural Network (CNN) architecture, accompanying a separate convolutional autoencoder to generate feature maps to boost overall performance. Recent studies, e.g. [7], have shown that there are certain shortcomings in that U-Net architecture which can be improved. In addition, they found that batch normalization can sometimes limit the performance of the U-Net architecture. Some other works show that the MultiResUNet [7] architecture, inspired by Inception [19] module, ameliorates the performance, ensures faster convergence, and delineates faint boundaries better. Furthermore, it is immune to perturbations and outliers in terms of segmenting variants of challenging data. Likewise, pretrained ResNet block along with spatial feature extractors can be used to derive better

spatial features in Context Encoder Network (CE-Net) [20]. In a similar fashion, skip connections in U-Net++ [21] can contribute to the reduction of semantic gap between the encoder and decoder resulting in faster optimization.

Some of the key features of the proposed model are briefly outlined here: Firstly, it uses fewer resources in terms of parameters and at the same time performs better than other models in a wide variety of datasets, without the use of transfer learning. One important aspect of the architecture is the shared processing of images at multiple scales–this usually adds more gradients, thus helping in training deeper networks. Min pooling [22] is employed to extract those features that are not present in the deeper layers due to the use of Rectified Linear Unit (ReLU) [23] and max pooling. Further, the creation of segmentation masks and minimization of losses at different scales contribute toward faster reconstruction of actual segmentation masks. The novel self-attention modules proposed in this work delimit the attention to crucial information in the data, eventually in aiding in the generation of better segmentation masks. In addition, masks are added at different scales and then passed through a SoftMax function to derive appropriate information from deeper layers. Finally, the segmentation standards are further improved by tuning weights of different scales using BO. The use of BO in the development of optimized deep learning architectures has been proposed in some recent works. Researchers have conducted asynchronous parallel hyper-parameter optimization via a supercomputer [24] using BO. Others have tuned hyper-parameters of a transfer learning model [25] using the probabilistic framework. The method is also used in the identification of best task-specific deep learning architectures using a process known as Neural Architecture Search [26, 27]. Here we confine the use of the probabilistic optimization technique to the very specific task of fine-tuning the weight hyperparameters associated with the different scales of the model– such limited and well-defined use of BO helps in overcoming the challenges of computational requirements involved in the case of optimization of entire deep networks. The incorporation of BO generates results that surpass those obtained using MSAN. A comparative evaluation of the proposed model with other SOTA models, via 5-fold cross-validation under restrained conditions, demonstrates the overall efficacy of our model in a wide variety of datasets.

3 Materials and Methods

In this section, we present the various datasets used in this work, the general aspects of the model including the deep learning architecture, the novel attention modules, and the BO framework employed for attaining superior results compared to the initial MSAN network.

3.1 Datasets Summary

A wide variety of datasets have been selected to compare the performance of the proposed model with other SOTA models. Standard image pre-processing techniques were used to normalize the image intensities (in the range of 0 to 1) before they were fed to the models. For the MSAN model, we pass the inputs at different scales, i.e., I, I/2, I/4 and I/8, where I is the original size of the image.

Brain Magnetic Resonance Imaging (MRI) Dataset [28] - The Brain MRI dataset, as shown in Figure 6II and Figure 6III contains Brain MRI images together with manual abnormality segmentation masks. This dataset is originally retrieved from The Cancer Genome Atlas Low Grade Glioma (TCGA-LGG). Glioma is a type of tumor that starts in the glial cells of the brain or the spine. The selected samples highlight the variety of colours and image-sizes present in the dataset. It is arduous even for humans to exactly pinpoint or locate the origin of glioma in such images. On top of that many images do not have any segmentation masks, making the training of deep learning algorithms more challenging. The dataset contains about 3929 image mask pairs of dimensions 256x256.

International Skin Imaging Collaboration (ISIC) 2017 dataset [29] - The ISIC (International Skin Imaging Collaboration) 2017 dataset as shown in Figure 6V and Figure 6VII comprise images pertaining to skin lesion. A lesion is any damage or abnormal change in the tissue of organism [30]. Such images are routinely analyzed by the medical community for identifying the exact location of skin lesions from calibrated images. The challenges include occlusion of images by hairs and difficulties in locating lesions due to different textures and tones of the skin. The dataset contains about 2000 image mask pairs of dimensions 767x1022. The images are resized to 192x256 for preserving the aspect ratio before feeding them to segmentation models for training and other purposes.

Lung's dataset a.k.a. LUNGS dataset [31] – This dataset contains images on the horizontal cross-section of lungs and their segmented masks using Computed Tomography (CT) scan as shown in Figure 6IX and Figure 6XI. This work examines a two-dimensional version of this dataset to segment the region of interest. Images with occluded lung parts can lead to processing difficulties. The dataset has about 267 images mask pairs of dimensions 512x512. The images are resized to 256x256 to preserve the aspect ratio and minimize the computations when passed to segmentation models.

Skin Lesion dataset [29] - This dataset is created from ISIC dataset with about 200 image masks pairs of dimensions 192x256. The main purpose of selecting this dataset is to test the robustness of SOTA models without data augmentations

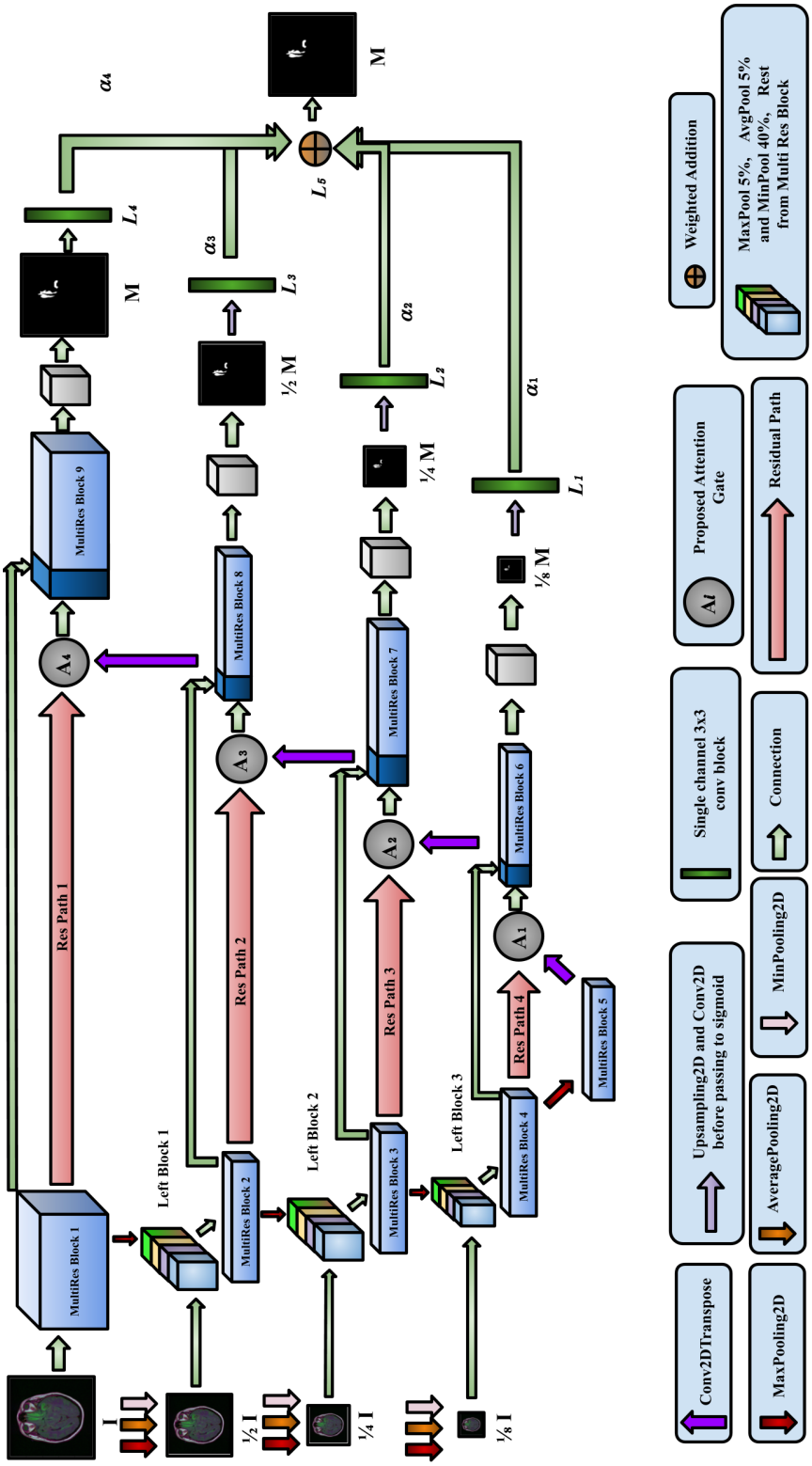

Figure 1: Our proposed MSAN Model. Images are passed at different scales and the places marked with A_i are replaced by a single attention module (i.e., either 1, 2 or 3 during training). Losses are separately computed at the different scales, which possibly contributes toward faster optimization. After up-sampling, segmentation masks are added in different proportions to generate the final segmentation mask.

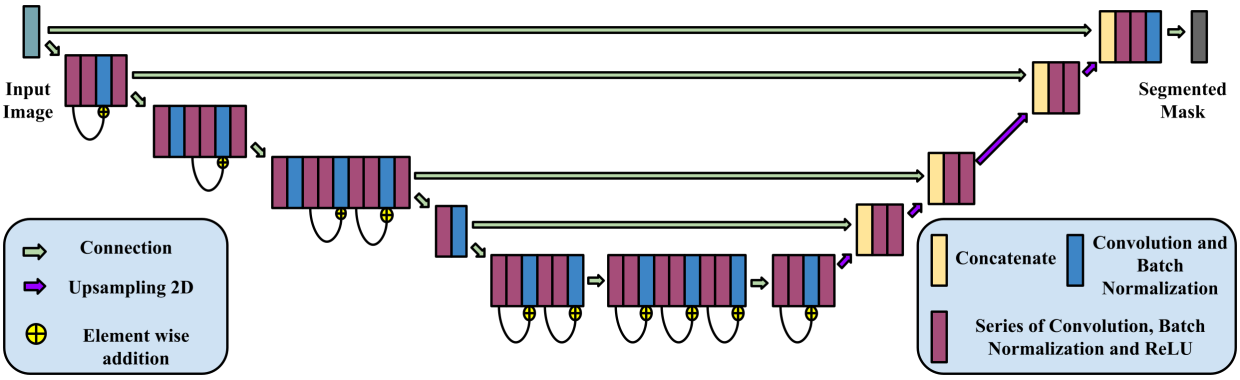

Figure 2: Modified U-Net model, created by using MobileNetV2's encoder architecture.

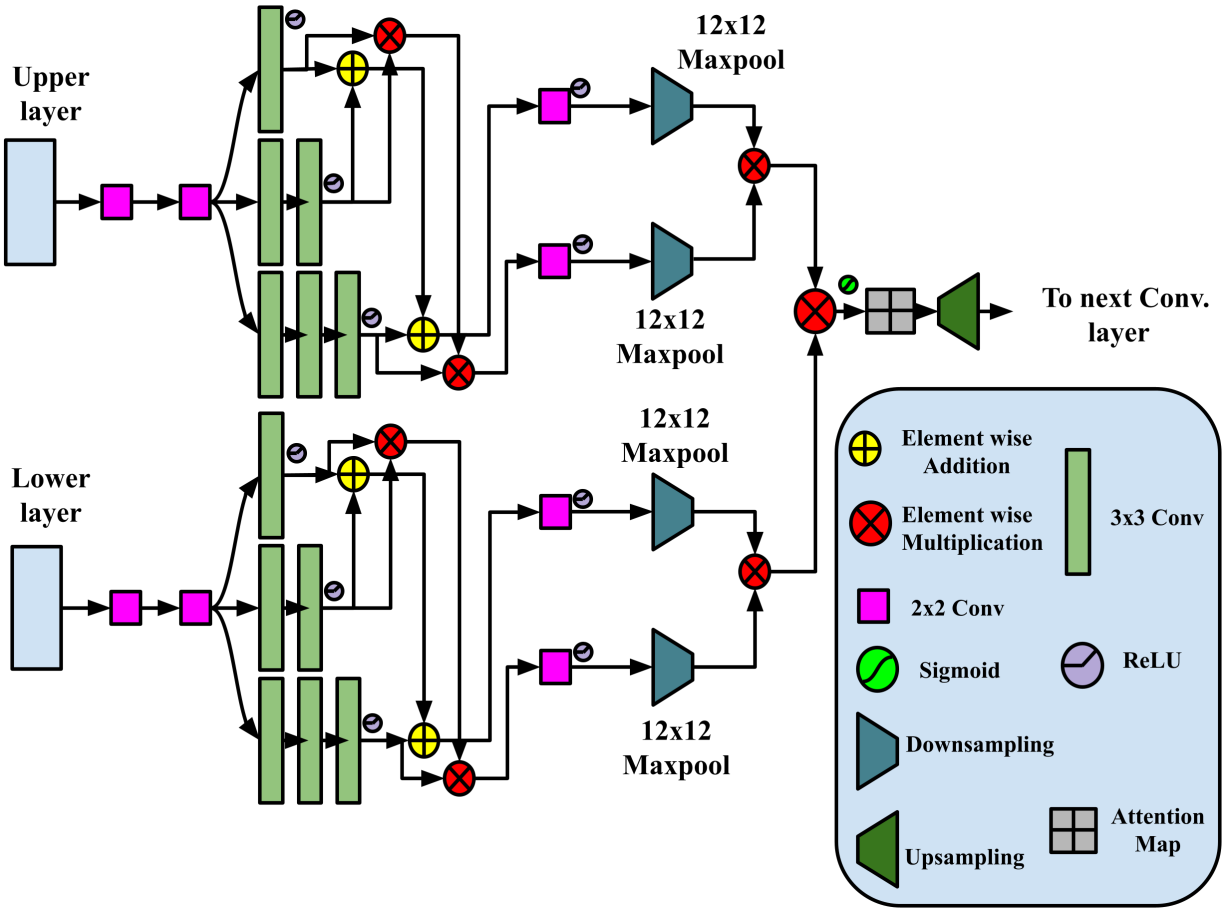

Figure 3: The Proposed Attention Module. Here, the Upper layer refers to the Residual Path and the Lower Layer refers to the Multi Res Block from Figure 1.

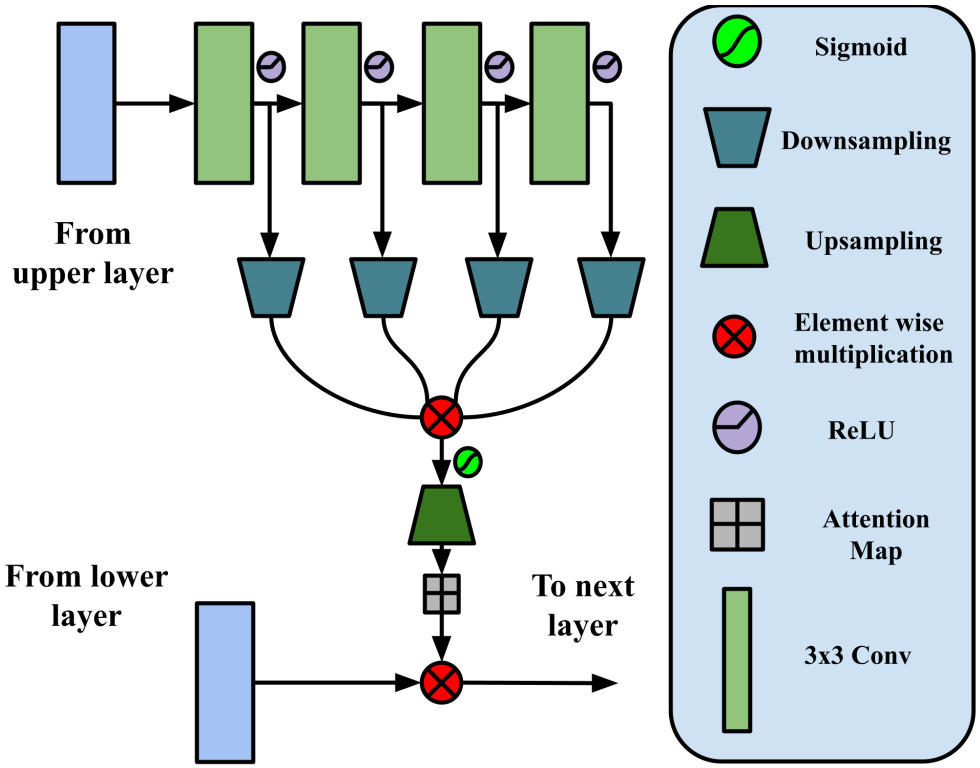

Figure 4: The Proposed Attention Module - 2. Here, the Upper layer refers to the Residual Path and the Lower Layer refers to the Multi Res Block from Figure 1.

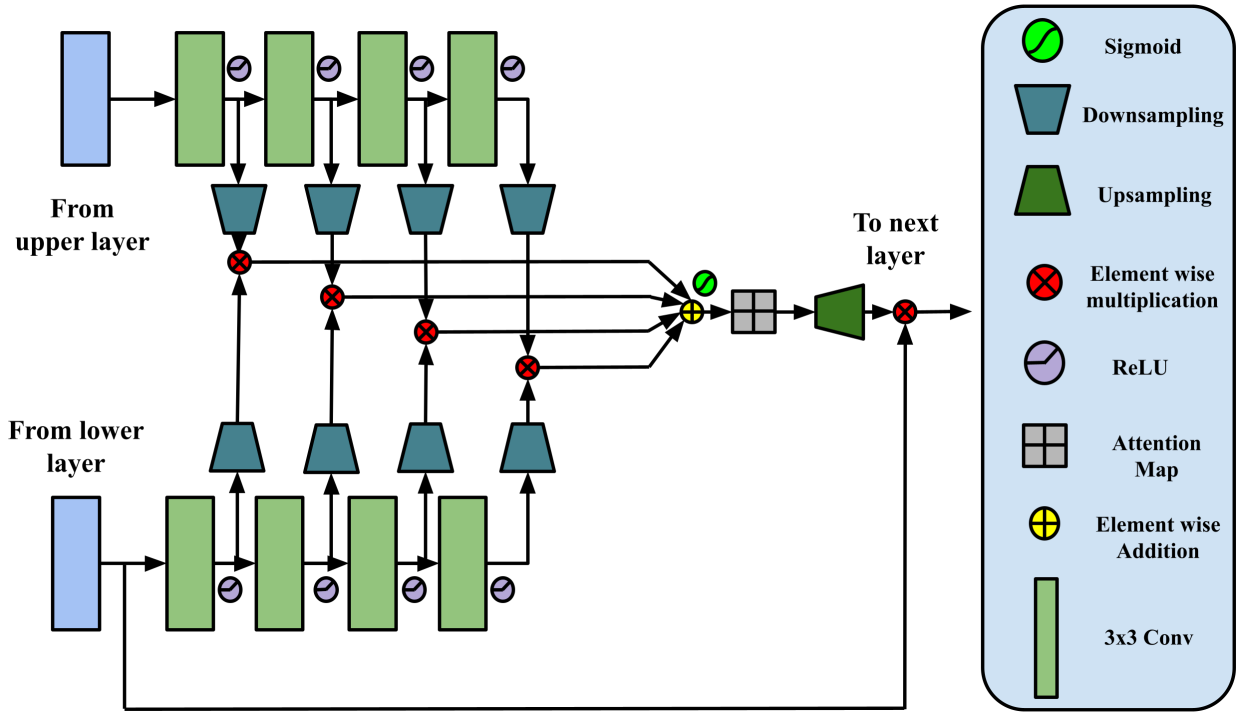

Figure 5: The Proposed Attention Module - 3. Here, the Upper layer refers to the Residual Path and the Lower Layer refers to the Multi Res Block from Figure 1.

and other forms of regularizations. Usually, such kinds of datasets are used with intensive data augmentations and transfer learning. Samples from the dataset are shown in Figure 6XIII and Figure 6XV.

KVASIR-SEG dataset [32] - This is an open-access dataset which comprise gastrointestinal polyp images that are manually annotated and verified by experienced gastroenterologists as shown in Figure 6XVII and Figure 6XIX. The dataset has about 1000 image mask pairs. The sizes of images vary from 332x487 to 1920x1072 pixels. The images are resized to 256x256 before passing to the segmentation models. The polyps are camouflaged into the background and it will be challenging for the model to find the regions of interest in different colours of the intestine.

Chest X-ray Collection a.k.a CHEST dataset [12] - This dataset is provided to the scholars upon request from Indiana University. It has about 138 grayscale images of chest radiographs and their corresponding masks in 4020x4892 resolution. The images are rescaled to 256x256 before passing to the models as shown in Figure 6XXI and Figure 6XXIII.

Nerve dataset a.k.a. NERVE dataset [33] - This dataset comprises of 5635 Brachial Plexus Nerve images as shown in Figure 6XXV and Figure 6XXVII. The images and segmentation masks are of dimensions 420x580. They are resized to 256x256 before passing through the model.

3.2 Proposed Model

Biomedical datasets have intra and inter-class diversity [34] when it comes to segmenting 2D imagery. In traditional convolutional neural network (CNN) architectures, these diversities are captured locally, which can potentially degrade the accuracy. To mitigate such shortcomings, we pursue a multi-scale approach [35] that helps to focus on relevant regions irrespective of scale. We feed the inputs at different scales to the left blocks, i.e., I, I/2, I/4 and I/8 , where I is the original size of the image. The use of Min pooling [22] assists in capturing information that is lost due to the use of Max pooling and ReLU. Min pooling may also seem to regularize the feature space. 40 $\%$ min pooling is used from each of the multi-scaled input images, which helps to capture those information that are lost in deeper layers. The combination of Min pooling, Max pooling and Average pooling appears to keep the actual feature distribution intact down the deeper layers. To help the model focus on relevant information down the deeper layer, attention mechanism is introduced to build association between different features. At a time a single attention module is replaced by all the four place holders represented by A_i in Figure 1. The Multi Res block [7] and the Residual blocks [36] helps to reduce the semantic gap between the encoder and the decoder, successively helping in faster optimization. Losses are computed at individual scales by rescaling them to the original mask size and taking a weighted addition, which eventually aids in the faster incremental reconstruction of masks. The segmentation masks at the individual scales are combined using optimized weight hyperperameters, which enable assigning appropriate importance to different feature space. The use of BO further helps in improving the segmentation masks for each dataset. The proposed architecture is shown in Figure 1. The code will be publicly available on https://github.com/Jimut123/bmsan.

3.3 Modified U-Net

To compare the efficiency of related models with a pre-trained transfer learning model, we have selected MobileNetV2 [6] as the encoder. We name this model as Modified U-Net and the same is shown in Figure 2. The latter is trained on ImageNet [37] dataset with some minor modifications that involve the introduction of skip connections as shown in Figure 2. This model is fine-tuned by removing fully connected layers and adding a series of batch normalization convolution ReLU to preserve the gradients at the deeper layers. The outputs from the encoder are concatenated with the decoder after up-sampling to preserve the sizes. The loss function used is Binary Cross entropy.

3.4 Proposed Attention Modules

At a time, a single attention module is put in all the places marked by A_i in Figure 1 during training. Here we describe the underpinnings of all the proposed attention modules.

3.4.1 Attention Module 1

A spatial soft attention module has been proposed as shown in Figure 3, which can capture spatial context from different kernel spaces. This idea is inspired by the attention modules proposed in [34, 11, 38, 39] and [40]. The main objective behind selecting the attention module is to get a compressed form of feature volume, which when added to the incoming volume of feature space makes the important features more prominent than the less significant features. The successive application of two 2x2 convolutions supports capturing the receptive field of the corresponding 3x3 kernels; this eventually enables learning of the important features with fewer parameters. Also, the use of a relatively smaller number of parameters in the kernels helps to capture the appropriate features only. The resulting volume is passed through a

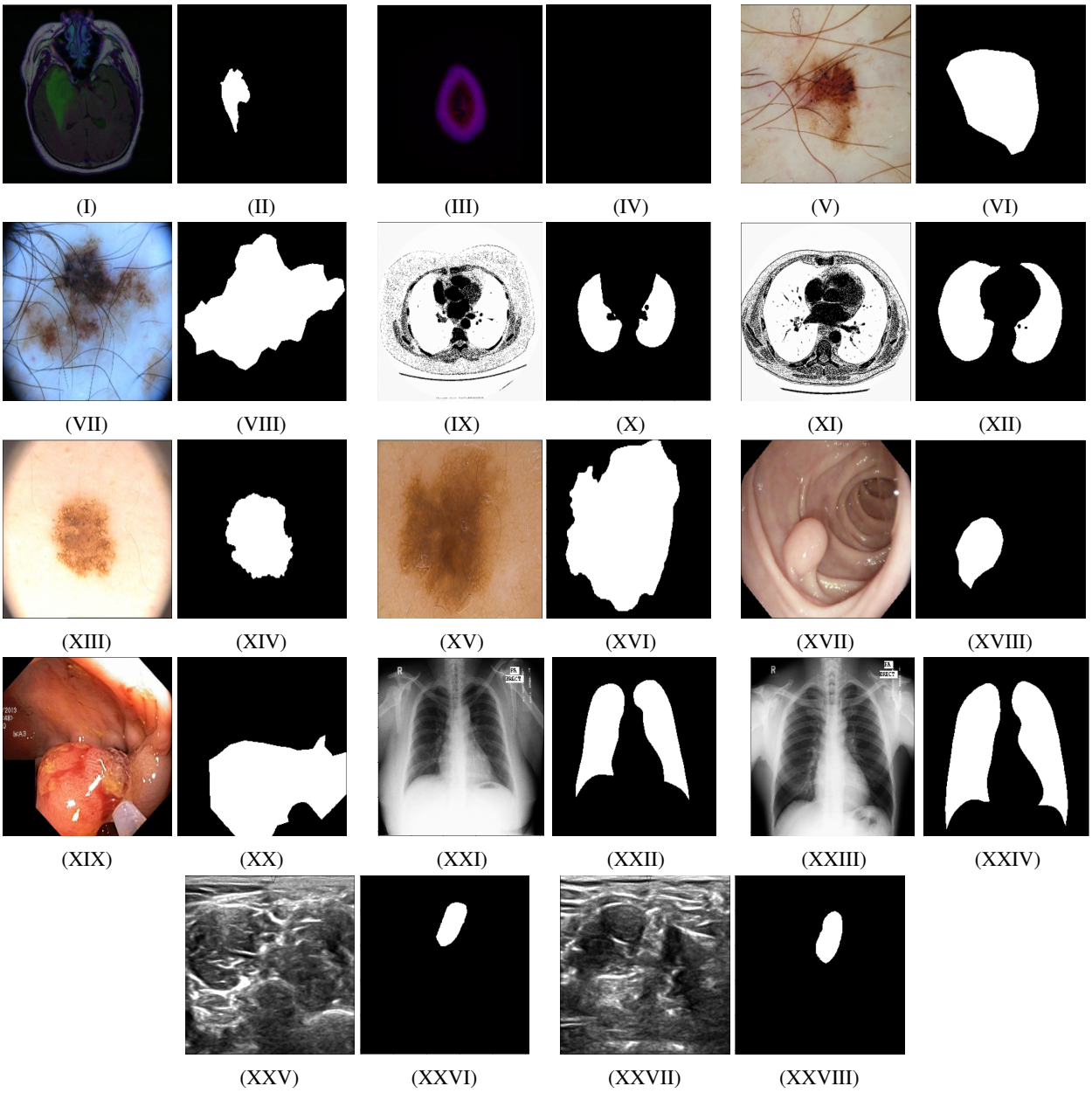

Figure 6: Samples of dataset used (from top), Brain MRI dataset, ISIC 2017 dataset, Lung dataset, Skin Lesion dataset, Kvasir-SEG dataset, Chest X-ray Collection and NERVE Ultrasound segmentation dataset. Some of the image samples from Brain MRI and NERVE dataset might not have any segmentation mask.

series of two 3x3 convolutional layers to capture the receptive field as that of a 5x5 kernel. Similarly, a series of three 3x3 convolutional layers culminate in a receptive field of 7x7 kernel with a comparatively lesser number of parameters. This attention module was partly motivated by the Inception module [19]. Each of the volumes has ReLU activations that capture positive non-linearity of the feature space. Adding the feature spaces captured by different types of kernels assists in amplifying the important features, while simultaneously reducing the effect of less important features. The volumes are multiplied in a similar architectural fashion, and the less weighted feature values will tend to diminish, giving importance to only the most important features, which will help to converge to a local minimum faster, in turn creating better attention maps.

The resultant feature maps are then passed through a max pooling layer, which captures the most important feature by creating an attention map and confining the convolutional feature map to a restricted space. The intention behind

multiplying the weights with the additive upper block for the upper layer (i.e., Res Path in Figure 1) is to retain the important features. The less important features receive a lower value in the feature space, thus facilitating the automatic assimilation of the relevant features. A similar approach is taken for the lower convolutional block. This is followed by the multiplication of resultant volumes from the lower and upper convolutional blocks to derive the important features common to both blocks. The resultant attention maps are then upsampled to revert to the original dimension. The mechanism can aid in the preservation of gradients and features from the upper layer while retaining the influence of the lower convolutional block. In this process, we use the sigmoid layer as activation to restrict values in the range of 0 to 1. The resulting volume is concatenated with the residual block to fasten the maximization of relevant features.

3.4.2 Attention Module 2

Another spatial self-attention module is proposed as shown in Figure 4. It takes features from the less compressed residual path and compresses it via 2x2 max pooling. It extends four convolutional layers, since this is the optimum, as investigated in the ablation studies in section 4.1 of the paper later. It takes the compressed feature maps and multiplies them with each other. It then uncompresses the information captured via upsampling, generating the attention map. The latter is subsequently multiplied with the features captured by the deeper multi-residual block. The main aim of this module is to multiply the less compressed features over a succession of four layers and accumulate the information to get a more meaningful and compact representation of the attention map, this will act as a factor multiplied to the information of the more compressed layers.

3.4.3 Attention Module 3

The main purpose behind developing this attention module is to take information from two successive layers, i.e., one less compressed, which may be considered as the upper layer (i.e., Res Path in Figure 1), and another more compressed which may be considered as the lower layer (i.e., MultiRes Block in Figure 1). In U-Net [9], the lower layers have a compact representation that helps to capture more relevant information related to a dataset. Taking information from two blocks, as shown in Figure 5, and multiplying them with one another gives importance to those features which are important to both blocks. Combining a series of such processing mechanism creates an attention map that helps to boost the compact information from the lower layer. Ablation studies further explain the rationale behind the choice of such type of architecture.

3.5 Formulation of Loss Function

The standard binary cross entropy [41] loss function was used for optimizing all the models. Before evaluating the overall multi-scaled weighted loss function, all the images were rescaled to the original segmentation mask size to ensure unbiased calculation of losses at individual scales. Suppose $L_k(y', y)$ is the loss associated with the k-th scale, where y' denotes the original value, and y, the prediction. The autoencoder has four levels of encoder-decoder structure and a separate loss is computed at each scale. The overall loss $L_5(y', y)$ is evaluated using the weighted average of the losses at the individual scales: $L_5(y', y) = \sum_{k=1}^{4} \alpha_k L_k(y', y)$, where α_k is the weight associated with the k-th scale. The weight hyperparameters are normalized by employing the following constraints: $\sum_{k=1}^{4} \alpha_k = 1$ and $0 \leq \alpha_k \leq 1$, where $k = 1\ to\ 4$. These constraints were also applied to the BO solver to get the set of α_i's which maximizes the Dice Coefficient $\times$ Jaccard i.e., $f(x)$ for each of the datasets as shown in Equation 1.

3.6 Bayesian Optimisation

The overall loss function of the proposed model encompasses unknown hyperparameters α_1, α_2, α_3 and α_4 that are associated with the losses at the different scales. Simple methods for hyperparameter tuning include grid and random search-based algorithms. We employ here a more sophisticated BO approach to determine the optimal configuration of the loss function hyperparameters. Such a methodology is preferred in settings involving expensive objective function, like in the present instance. BO [8] uses Gaussian process as surrogate model to approximate the costly objective function (see e.g. [42]), and optimizes an acquisition function, which is defined based on the posterior mean and variance of the proxy, for identifying the next input location for evaluation. The process typically involves making a trade-off between exploration and exploitation [8].

The objective function $f(x)$ for BO is taken to be the product of Dice coefficient and Jaccard, i.e.

$$f(x) := f(\alpha_1, \alpha_2, \alpha_3, \alpha_4) := \text{Dice Coefficient} \times \text{Jaccard} \tag{1}$$

Datasets	Model Name	Dice Coefficient (%)	IoU (%)	Precision (%)
Brain MRI [28]	**U-Net** [9]	66.45 ± 5.48	57.39 ± 4.55	56.75 ± 4.36
	MultiResUNet [7]	62.10 ± 5.63	53.16 ± 4.50	52.59 ± 4.38
	Modified U-Net	62.39 ± 5.07	53.42 ± 4.01	52.70 ± 3.93
	R2U-Net [14]	70.32 ± 5.03	61.43 ± 4.68	60.76 ± 4.64
	Attention R2U-Net	68.25 ± 2.49	59.69 ± 2.83	59.07 ± 2.86
	Attention U-Net [11]	67.31 ± 4.98	58.32 ± 4.02	57.66 ± 3.87
ISIC 2017 [29]	**U-Net**	82.36 ± 4.48	73.95 ± 5.33	75.21 ± 2.50
	MultiResUNet	81.67 ± 4.10	72.76 ± 4.76	76.10 ± 4.01
	Modified U-Net	84.86 ± 4.48	76.96 ± 4.99	82.73 ± 2.11
	R2U-Net	86.74 ± 2.63	79.19 ± 3.49	79.85 ± 1.16
	Attention R2U-Net	86.49 ± 3.90	78.89 ± 4.78	82.29 ± 3.91
	Attention U-Net	82.64 ± 4.24	74.19 ± 4.97	75.06 ± 1.61
LUNGS [31]	**U-Net**	95.65 ± 0.57	92.18 ± 0.88	93.74 ± 0.45
	MultiResUNet	97.07 ± 0.22	94.65 ± 0.25	95.25 ± 0.30
	Modified U-Net	97.27 ± 0.35	95.04 ± 0.49	95.50 ± 0.56
	R2U-Net	97.44 ± 0.34	95.37 ± 0.49	95.86 ± 0.47
	Attention R2U-Net	97.34 ± 0.35	95.19 ± 0.50	95.75 ± 0.45
	Attention U-Net	96.56 ± 0.31	93.76 ± 0.56	94.84 ± 0.50
Skin Lesion[29]	**U-Net**	86.93 ± 5.07	78.81 ± 6.15	83.33 ± 2.51
	MultiResUNet	88.50 ± 1.70	81.06 ± 2.35	84.91 ± 2.52
	Modified U-Net	92.76 ± 1.22	87.03 ± 1.87	88.35 ± 1.64
	R2U-Net	92.92 ± 1.45	87.43 ± 2.18	89.05 ± 1.70
	Attention R2U-Net	93.02 ± 1.31	87.46 ± 2.01	88.70 ± 1.76
	Attention U-Net	87.85 ± 1.67	79.82 ± 2.40	83.28 ± 2.00
KVASIR-SEG [32]	**U-Net**	64.59 ± 2.91	54.00 ± 2.43	56.12 ± 5.66
	MultiResUNet	60.36 ± 2.57	49.72 ± 2.04	56.43 ± 0.00
	Modified U-Net	79.34 ± 1.21	70.42 ± 1.43	71.63 ± 1.79
	R2U-Net	77.78 ± 1.80	69.41 ± 1.92	71.43 ± 0.00
	Attention R2U-Net	75.42 ± 2.65	66.62 ± 3.40	66.23 ± 0.00
	Attention U-Net	49.29 ± 4.13	38.91 ± 3.23	37.21 ± 0.05
CHEST [12]	**U-Net**	95.15 ± 0.77	90.87 ± 1.32	91.05 ± 1.36
	MultiResUNet	96.67 ± 0.55	93.64 ± 0.97	94.41 ± 0.98
	Modified U-Net	97.39 ± 0.57	94.98 ± 1.03	95.51 ± 1.05
	R2U-Net	97.19 ± 0.62	94.61 ± 1.11	95.20 ± 1.11
	Attention R2U-Net	97.15 ± 0.68	94.54 ± 1.21	95.05 ± 1.19
	Attention U-Net	94.67 ± 1.04	90.06 ± 1.78	90.24 ± 1.82
NERVE [33]	**U-Net**	52.96 ± 1.51	44.80 ± 1.49	44.54 ± 1.50
	MultiResUNet	43.26 ± 4.65	36.17 ± 3.82	35.84 ± 3.69
	Modified U-Net	39.32 ± 2.67	32.29 ± 2.07	31.99 ± 1.98
	R2U-Net	54.98 ± 2.15	46.42 ± 2.03	46.02 ± 2.03
	Attention R2U-Net	54.60 ± 3.01	46.12 ± 2.59	45.79 ± 2.46
	Attention U-Net	52.84 ± 1.27	44.76 ± 1.17	44.54 ± 1.17

Table 1: Results obtained from 5 fold cross validation (presented as mean ± standard deviation, $\mu \pm \sigma$) by keeping a batch size of 2, using Adam optimizer with a learning rate of 1e-05 for different models with no data augmentation. The brown colour ranks third, blue colour ranks second and red colour ranks first.

We are now interested in determining the values of the hyperparameters for which this objective function is maximized. Our implementation works as follows: Firstly, the original dataset is split into training (80%) and test (20%) datasets. The model is then trained on 80% of the original training dataset (i.e., 64% of the whole dataset) for a particular setting of the hyperparameters. The $f(x)$ on the remaining portion (20%) of the original training dataset (this will be referred to as the validation dataset) is considered as the objective function for BO, using the best performing MSAN model.

The BO approach is initialized with 80 different settings of the optimization variables (αs), which are sampled from a Dirichlet distribution, and the corresponding objective function (i.e. Dice coefficient $\times$ Jaccard) values. For each such hyperparameter setting, the model is trained for 100 epochs with the Adam optimizer using a batch size of two, and then the corresponding $f(x)$ value on the validation dataset is recorded. The popularly used expected improvement [43]

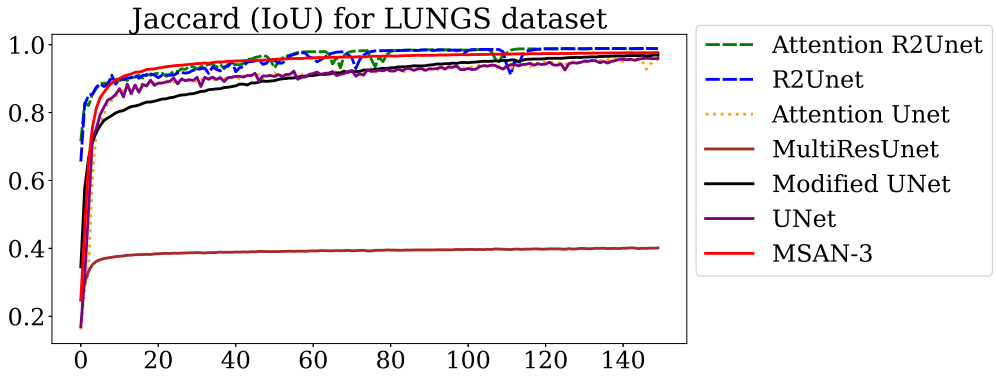

Figure 7: Comparison of Jaccard Index for LUNGS dataset obtained during training of all models.

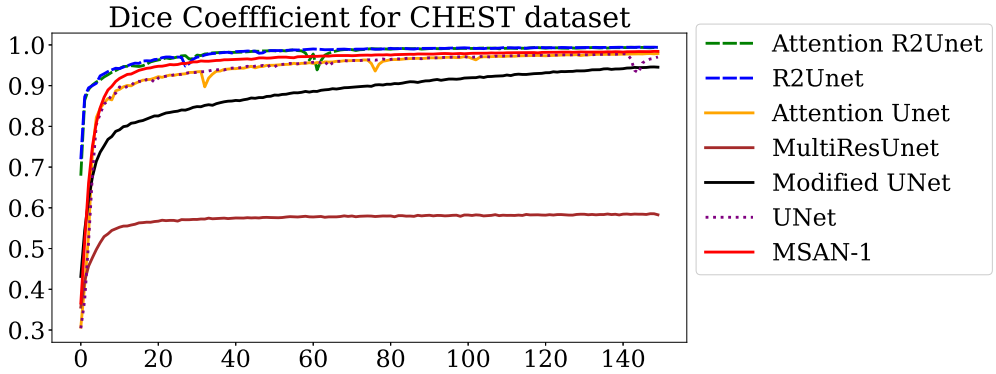

Figure 8: Comparison of Dice Coefficient for CHEST dataset obtained during training of all models.

algorithm is used as the acquisition function, and the *Matern* kernel is used in the Gaussian process surrogate model to capture structures in the objective function. We run another 40 iterations for the BO solver with 100 epochs in MSAN model to get the alpha combination, which adds up to 1 for BMSAN model. The alpha values obtained for each of the datasets are shown in Table 4. The combinations of αs which maximizes the function $f(x)$ is again used to train BMSAN model using k-fold Cross Validation for 150 epochs and the results obtained are shown in Table 3.

4 Results and Discussions

All the experiments were carried out in Asus RTX 2080 Ti (12 GB) and Quadro GV100 (32 GB) GPU machines with 64 GB RAM using TensorFlow framework [45] and Keras library: `https://github.com/keras-team/keras` written in Python3. Adam [46] optimizer was used in the training of the model with Binary Cross entropy [41] as the loss function. A small batch size, comprising two images, was used for testing the robustness of the model due to constraints. Increasing the batch size may marginally improve the overall performance of the models, but consideration of a small batch size reduces the computational overheads. A learning rate of 1e-05 was selected, which yields standard performance when compared with other recorded values in the related literature. As medical images are collected in a constrained environment, there is a need for data augmentation by innovative methods, e.g. polar transformations, as assessed in Skip-Link Attention Guidance Network (SLAG-CNN) [35] model. The issue is addressed by following the research methodology of MultiResUNet [7], and data augmentations were not used anywhere in this study. All models were trained for 150 epochs beyond which no further improvement was observed in either of the models. BO implementation is made using GPyOpt toolkit: `http://github.com/SheffieldML/GPyOpt`. Standard metrics were calculated and recorded throughout the experiments, i.e., Dice Coefficient (refer to Equation 2), Jaccard Index (refer to Equation 3), and Precision (refer to equation 4) which are widely used in medical image segmentation community.

The graph for Jaccard obtained during training of the Lungs dataset is shown in Figure 7. We can see that our model is superior to other models during training. The plot is also much smoother than the other models, especially AttentionR2U-Net and R2U-Net. This can be possibly due to the inherent property of the model to regularize itself, avoiding kinks in the graph that are potentially caused by exploding gradients [47] in the parameter space. We have

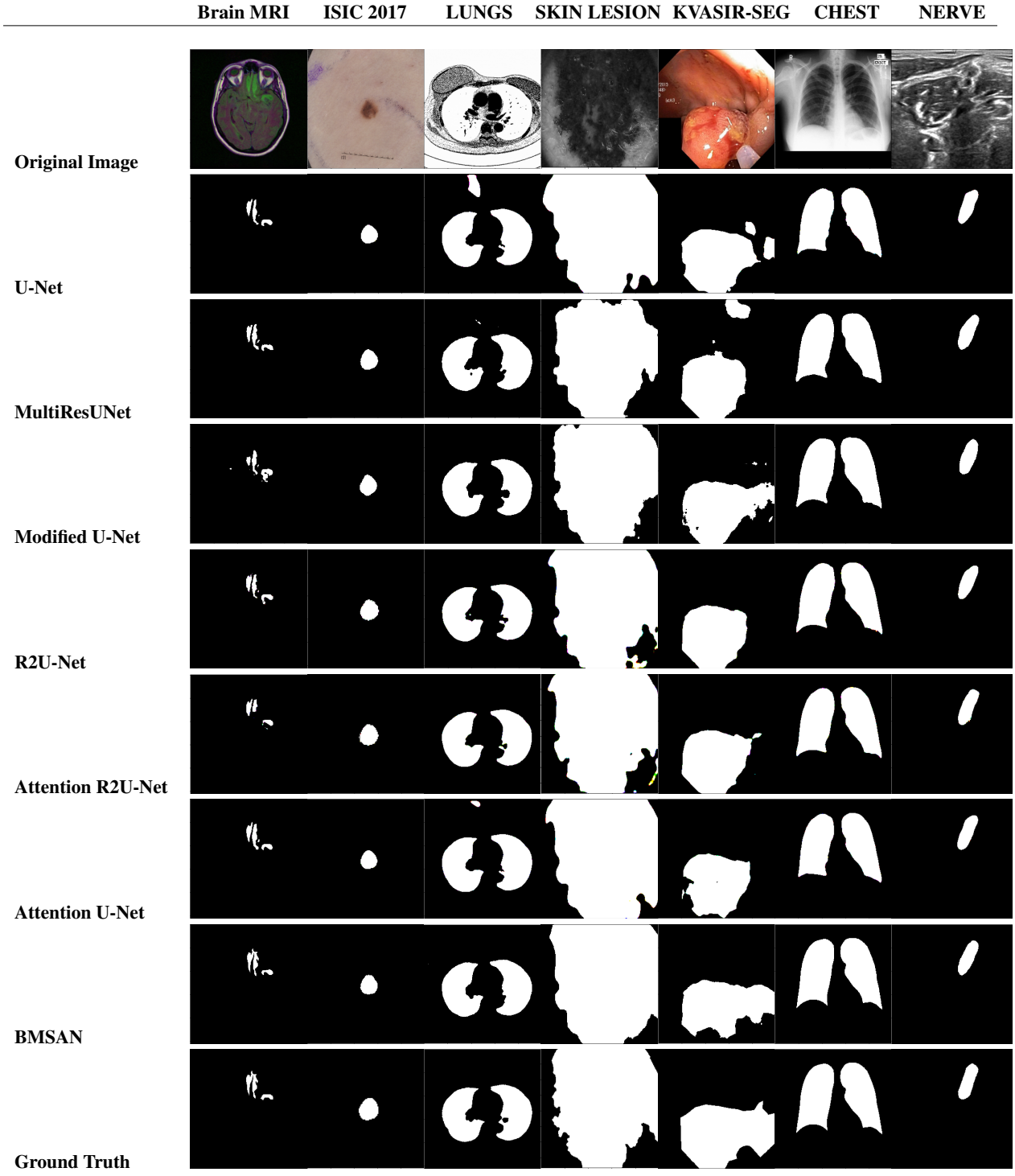

Table 2: Segmentation masks obtained using a batch size of 2, with Adam optimizer and a learning rate of 1e-05 for different models. Each dataset has a different level of difficulty. The masks were obtained from the test set with 80-20 train-test split.

Datasets	Model Name	Dice Coefficient (%)	IoU (%)	Precision (%)
Brain MRI [28]	**R2U-Net**	**70.32 ± 5.03**	**61.43 ± 4.68**	**60.76 ± 4.64**
	MSAN-1	57.62 ± 9.45	49.40 ± 8.37	49.59 ± 7.69
	MSAN-2	64.24 ± 5.51	55.13 ± 4.82	54.56 ± 4.68
	MSAN-3	56.71 ± 7.18	48.26 ± 5.79	47.60 ± 5.62
	BMSAN-2	66.35 ± 7.22	57.36 ± 6.85	54.73 ± 4.88
ISIC 2017 [29]	**R2U-Net**	86.74 ± 2.63	79.19 ± 3.49	79.85 ± 1.16
	MSAN-1	79.24 ± 5.65	70.14 ± 6.31	78.70 ± 0.17
	MSAN-2	83.83 ± 3.98	75.52 ± 5.10	76.24 ± 0.50
	MSAN-3	86.38 ± 2.41	78.65 ± 3.06	82.34 ± 3.70
	BMSAN-3	**86.86 ± 2.52**	**79.57 ± 2.97**	**79.86 ± 0.16**
LUNGS [31]	**R2U-Net**	97.44 ± 0.34	95.37 ± 0.49	95.86 ± 0.47
	MSAN-1	97.25 ± 0.30	94.99 ± 0.39	95.53 ± 0.34
	MSAN-2	97.32 ± 0.18	95.11 ± 0.17	95.61 ± 0.19
	MSAN-3	97.38 ± 0.26	95.23 ± 0.30	95.70 ± 0.27
	BMSAN-3	**97.49 ± 0.29**	**95.40 ± 0.44**	**95.86 ± 0.53**
Skin Lesion [29]	**AttentionR2U-Net**	93.02 ± 1.31	87.46 ± 2.01	88.70 ± 1.76
	MSAN-1	90.31 ± 1.74	83.65 ± 2.40	86.82 ± 1.43
	MSAN-2	92.58 ± 1.30	86.74 ± 2.05	88.33 ± 1.39
	MSAN-3	92.55 ± 1.12	86.74 ± 1.74	88.50 ± 1.34
	BMSAN-2	**93.22 ± 0.26**	**87.77 ± 0.93**	**88.75 ± 1.57**
KVASIR-SEG [32]	**Modified U-Net**	**79.34 ± 1.21**	**70.42 ± 1.43**	**71.63 ± 1.79**
	MSAN-1	62.29 ± 1.05	52.29 ± 0.86	52.89 ± 0.78
	MSAN-2	69.34 ± 3.39	59.30 ± 3.32	60.45 ± 4.23
	MSAN-3	69.98 ± 1.15	59.95 ± 1.02	61.23 ± 2.24
	BMSAN-3	70.37 ± 1.77	60.25 ± 1.63	62.43 ± 1.45
CHEST [12]	**Modified U-Net**	97.39 ± 0.57	94.98 ± 1.03	95.51 ± 1.05
	MSAN-1	96.92 ± 0.00	94.09 ± 0.01	94.77 ± 0.01
	MSAN-2	96.68 ± 0.58	93.67 ± 1.03	94.13 ± 1.00
	MSAN-3	96.49 ± 0.59	93.30 ± 1.04	93.98 ± 1.15
	BMSAN-1	**97.42 ± 0.10**	**95.07 ± 0.10**	**95.65 ± 0.11**
NERVE [33]	**R2U-Net**	54.98 ± 2.15	46.42 ± 2.03	46.02 ± 2.03
	MSAN-1	54.65 ± 3.92	45.97 ± 3.80	45.70 ± 3.80
	MSAN-2	53.59 ± 1.39	45.38 ± 1.38	45.15 ± 1.40
	MSAN-3	52.65 ± 3.82	44.38 ± 3.23	44.14 ± 3.06
	BMSAN-1	**55.66 ± 1.55**	**46.99 ± 1.25**	**46.75 ± 1.21**

Table 3: Results obtained from 5 fold cross validation (presented as mean ± standard deviation, $\mu \pm \sigma$) by keeping a batch size of 2, using Adam optimizer with a learning rate of 1e-05 for different models with no data augmentation. The results in **bold** indicate the best performance for the corresponding dataset. For a particular dataset, BMSAN corresponds to the best MSAN model optimized using BO.

also shown a comparison of the number of parameters of the model with other models (in million) as shown in Figure 9. The size of our model is relatively small in terms of parameters, making it easier to optimize. This is possibly the reason why the graph of dice coefficient is much smoother as shown in Figure 8 for the CHEST dataset.

We analyze the performance of different models in terms of Dice Coefficient, IoU (Intersection Over Union, known as Jaccard) and Precision using a 5 fold cross validation (presented as mean ± standard deviation, i.e. $\mu \pm \sigma$). A comparison of the performance of different models employed on the discussed datasets is shown in Table 1. Let us first discuss the Brain MRI dataset. It can be seen that R2U-Net produces better results in all the three metrics i.e., Dice, IoU and Precision. This is followed by Attention R2U-Net and Attention U-Net. The results for the Brain MRI dataset show that the models with the largest number of parameters achieve the best results. Our model i.e., MSAN-2 after performing BO could not surpass the SOTA R2U-Net for this dataset as shown in Table 3. There is a huge difference in MSAN-1 and MSAN-2's result, showing that different attention modules respond differently in segmenting different dataset. It can also be seen that even Modified U-Net fails to learn meaningful information in this context even when it is initialized by ImageNet weights.

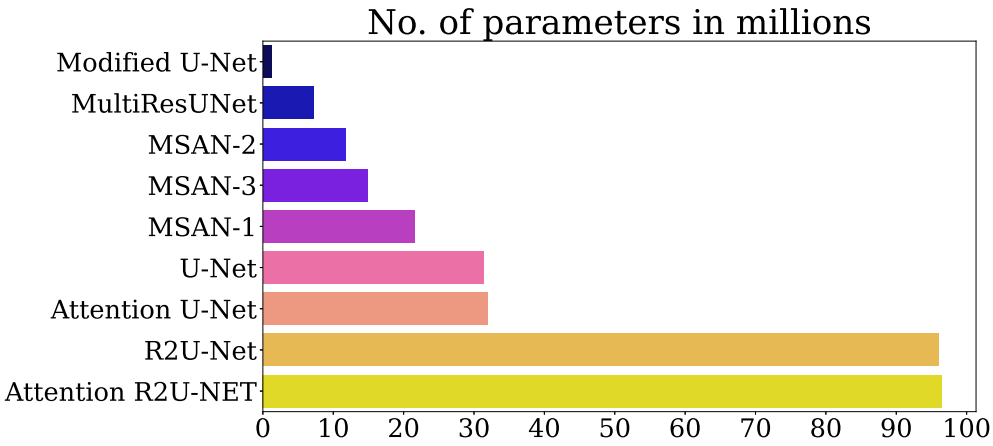

Figure 9: Comparison of parameters for all the models starting with Modified U-Net with 1.2 M parameters, MultiResUNet with 7.2 M parameters, MSAN-2 with 11.69 M parameters, MSAN-3 with 14.83 parameters, MSAN-1 with 21.55 M parameters, U-Net with 31.3 M parameters, Attention U-Net with 31.9 M parameters, R2U-Net with 95.9 M parameters and Attention R2 U-Net with 96.5 M parameters.

Datasets	**Model**	α_1	α_2	α_3	α_4	**Dice Coefficient**	**Jaccard**	$f(x)$
Brain MRI	MSAN-2	0.074502	0.223438	0.278507	0.421730	0.810567	0.724427	0.587197
ISIC 2017	MSAN-3	0.287012	0.026188	0.173405	0.511562	0.890383	0.822607	0.732435
LUNGS	MSAN-3	0.110000	0.000000	0.040000	0.840000	0.970680	0.948244	0.920442
Skin Lesion	MSAN-2	0.122950	0.173061	0.088320	0.596683	0.931593	0.876083	0.816153
KVASIR-SEG	MSAN-3	0.080000	0.320000	0.220000	0.370000	0.664659	0.558347	0.371110
CHEST	MSAN-1	0.110442	0.159336	0.395935	0.324284	0.965283	0.933954	0.901531
NERVE	MSAN-1	0.090000	0.160000	0.430000	0.310000	0.562848	0.477413	0.268711

Table 4: The values of αs obtained by training on 64 % of dataset and validating on 16 % of the dataset. The results presented here may deviate from those in Table 3 as the validation dataset may not capture the distribution of the whole dataset.

R2U-Net leads in two performance metrics in segmenting ISIC 2017 dataset, as shown in Table 1, followed by Attention R2U-Net and Modified U-Net. In comparison to Brain MRI dataset, the ImageNet initialization appears to have contributed towards the training of a better model for this dataset. It is found that the proposed MSAN model with Attention Module-3 performs significantly better than the same with Attention Modules 1 and 2, as shown in Table 3. The performance of MSAN-3 is close to the top-performing SOTA models (see Tables 1 and 3), but with significantly lesser number of parameters. However, on optimization using BO, the new BMSAN-3 outperforms all the SOTA models, producing a new standard.

In case of LUNGS dataset, R2U-Net attains the best performance in all three metrics (see Table 1), followed by Attention R2U-Net and Modified U-Net. MSAN-3 achieves comparable results in segmenting the dataset as shown in Table 3. However, MSAN-3 when combined with BO surpasses R2U-Net in performance metrics, thus creating a new SOTA benchmark in this category. The results demonstrate the efficacy of tuned multi-scale weight hyperparameters in boosting even the peak performance of models.

Attention R2U-Net surpasses all the models in segmenting the relatively smaller Skin Lesion dataset as shown in Table 1, followed by R2 U-Net and Modified U-Net. The MSAN segmentation model is found to benefit the most from the inclusion of the second attention module (see Table 3). Here again, the best-performing MSAN model when combined with BO outperforms all the SOTA models creating a new standard for the dataset.

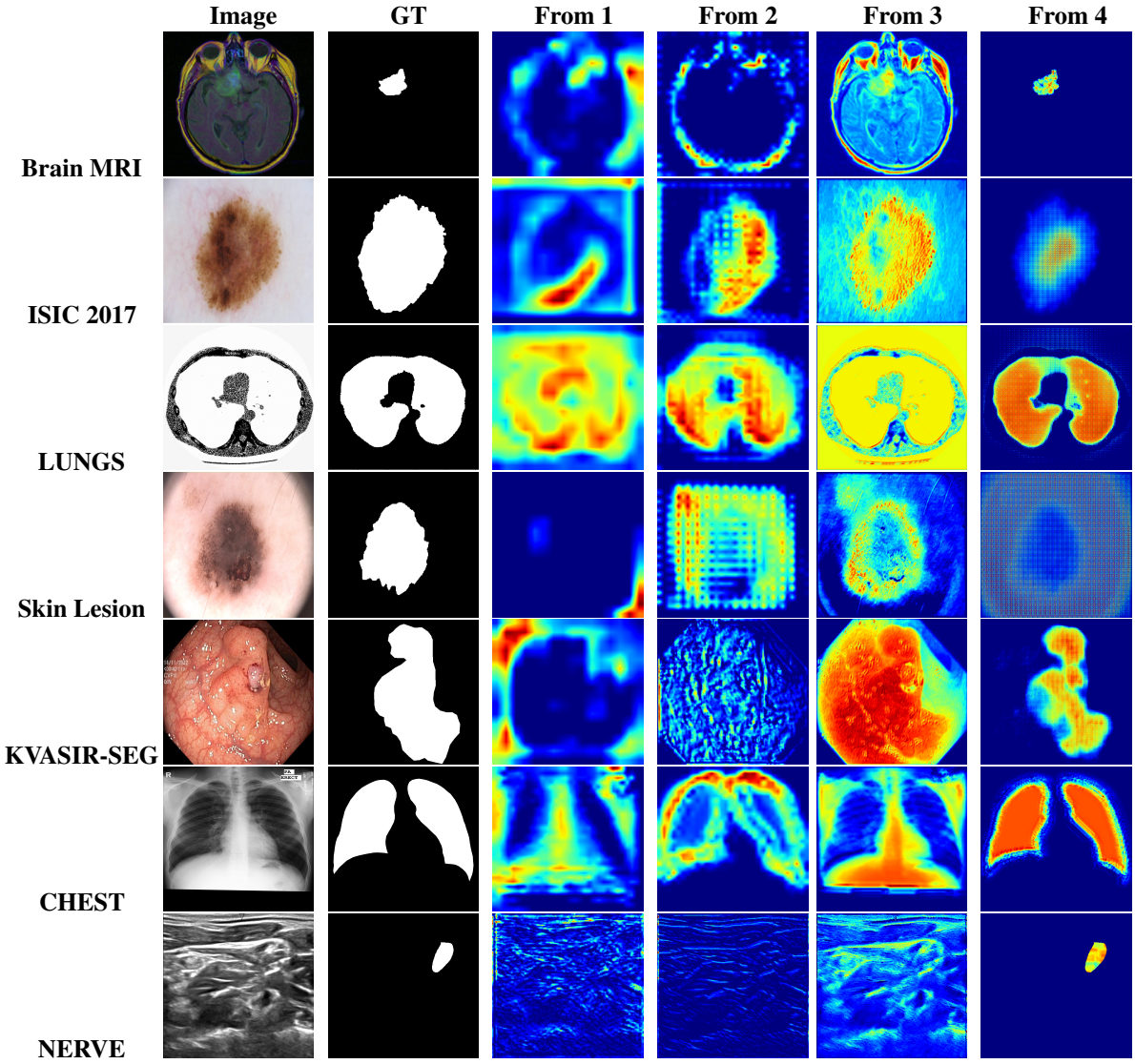

Figure 10: Attention maps generated from the four place-holders for attention modules from bottom to top (i.e., A_1, A_2, A_3 and A_4 from Figure 1) by use of GRAD-CAM [44] for different datasets. We select the best performing BMSAN model which have same attention module in all the four place holders. We collect the attention maps from each of the attention module.

In the case of KVASIR-SEG dataset the performance of our fine-tuned Modified U-Net model outclasses all the other models. The performance of MSAN models seems to have been affected by the camouflaging of polyps in the flesh, which possibly leads to confusion in segmentation. It appears that the ImageNet initialization supports the segmentation of camouflaged objects in comparison to training from scratch. Although the performances of MSAN models are below expectation, the optimized version (BMSAN-3) attains results better than that of U-Net, Attention U-Net, and MultiRes UNet (see Tables 1 and 3).

The segmentation results of the CHEST dataset show that our Modified U-Net excels all other models in the performance metrics (see Table 1). Interestingly, the first MSAN network on optimization creates a new SOTA performance standard on this dataset (see Table 3).

In the case of NERVE ultrasound dataset, R2 U-Net attains the top-most position in segmentation, performing slightly better than Attention R2U-Net. BO-based MSAN models again deliver the best results, creating a SOTA benchmark with the BMSAN-1 model (see Table 3). The final segmentation results from the test set for all the models are displayed in Table 2. The prediction of BMSAN is taken from the best MSAN model with its attention module (i.e., either 1, 2 or 3) and performing BO using the same. From the table we can see that BMSAN performs significantly better than other models in segmenting the datasets.

Remarks	**Dice Coefficient** (%)	**IoU** (%)	**Precision** (%)
Without any attention module, but with MI, with losses at different scales	91.84 ± 1.67	85.53 ± 2.70	87.25 ± 2.05
Without any attention module, without MI, but with losses at different scales	92.28 ± 1.41	86.37 ± 2.01	88.08 ± 1.25
Without any attention module, without MI, without losses at different scales	92.40 ± 1.27	86.54 ± 1.84	88.22 ± 1.42
A-3 attention module without MI, without losses at different scales	92.47 ± 1.37	86.53 ± 2.12	88.27 ± 1.46
A-2 attention module without MI, without losses at different scales	92.02 ± 1.96	86.03 ± 2.94	88.30 ± 2.11
A-1 attention module without MI, without losses at different scales	89.97 ± 2.98	82.87 ± 4.50	85.29 ± 4.28
Without min pool but everything constant with **A-3**	92.01 ± 0.97	85.91 ± 1.45	87.38 ± 1.17
Without min pool but everything constant with **A-2**	92.56 ± 1.32	86.71 ± 2.01	88.39 ± 1.05
Without min pool but everything constant with **A-1**	90.55 ± 2.46	83.88 ± 3.62	86.14 ± 3.24
A-2 with 2 blocks without maxpool	91.58 ± 1.22	85.37 ± 1.83	87.26 ± 1.39
A-2 with 2 blocks with (2,2) maxpool	92.29 ± 1.56	86.44 ± 2.26	87.83 ± 2.02
A-2 with 2 blocks with (3,3) maxpool	92.37 ± 1.03	86.38 ± 1.53	88.10 ± 1.32
A-2 with 2 blocks with (4,4) maxpool	91.75 ± 1.93	85.59 ± 2.85	87.55 ± 2.31
A-2 with 4 blocks with (3,3) maxpool	**92.58 ± 1.30**	**86.74 ± 2.05**	**88.33 ± 1.39**
A-2 with 6 blocks with (3,3) maxpool	91.83 ± 0.87	85.70 ± 1.19	87.79 ± 0.84
A-2 with 8 blocks with (2,2) maxpool	92.30 ± 1.21	86.43 ± 1.80	88.17 ± 1.28
A-3 with 2 blocks with (2,2) maxpool	91.64 ± 2.22	85.23 ± 3.58	87.10 ± 2.61
A-3 with 4 blocks with (2,2) maxpool	**92.55 ± 1.12**	**86.74 ± 1.74**	**88.50 ± 1.34**
A-3 with 5 blocks with (2,2) maxpool	92.16 ± 1.47	86.08 ± 2.19	87.62 ± 1.79
A-3 with 6 blocks with (2,2) maxpool	92.55 ± 1.51	86.79 ± 2.24	88.31 ± 1.83
A-3 with 8 blocks with (2,2) maxpool	92.45 ± 1.11	86.61 ± 1.66	88.05 ± 1.46
A-3 with 10 blocks with (2,2) maxpool	92.34 ± 1.50	86.64 ± 2.20	88.89 ± 1.50
A-3 with 4 blocks without maxpool	91.61 ± 2.14	85.26 ± 3.27	88.63 ± 1.39
A-3 with 4 blocks with (3,3) maxpool	91.55 ± 2.05	85.30 ± 2.75	87.47 ± 1.26
A-3 with 4 blocks with (2,2) maxpool multilpy with upper block	91.85 ± 1.11	85.53 ± 1.55	87.31 ± 0.76

Table 5: Ablation studies conducted via the Skin Lesion dataset with 5 fold cross validation, batch size of 2 images, and Adam Optimizer with learning rate 1e-05. The remarks and corresponding metrics are recorded to show the variations.

4.1 Ablation Studies

Ablation studies on the proposed MSAN model were undertaken using the Skin Lesion dataset due to its (relatively) small size and lower computational requirements. We also discuss the justifications behind certain modifications to the model. The experiments were performed with 5-Fold Cross Validation using Adam Optimizer and a learning rate of 1e-05 for 150 epochs. Here are some observations: Firstly, eliminating the attention module and keeping everything else intact results in a dice coefficient of 91.84 ± 1.67 as shown in Table 5. Secondly, removing the attention module and Multi-scaled inputs (MI), and keeping the losses at different scales (i.e., the multi-scaled losses), yielded a dice coefficient of 92.28 ± 1.41. We use the acronym MI (multi-scale) here to refer to the multi-scale inputs that were fed at at the different layers of the encoder (i.e., left blocks), excluding the original image of size I passed to the MultiRes

Block 1 as shown in Figure 1. It appears that the MI part is not productive for the model, but we will see the actual motivation for keeping them later. Thirdly, the model without MI part, attention module, and losses at different scales, gives 92.40 $\pm$ 1.27 as dice coefficient. Lastly, it can be seen from Table 5 that retaining the attention modules without any MI part and losses at different scales gives varied results for different models. For example, the use of the model with attention module 3 gives 92.47 $\pm$ 1.37 as dice coefficient. On the other hand, attention modules 2 and 1 yield dice coefficient values of 92.02 $\pm$ 1.96 and 89.97 $\pm$ 2.98, respectively. We now eliminate min pooling but retain everything else the same as in the MSAN model proposed earlier (see Figure 1). This modification gives dice coefficient values of 92.01 $\pm$ 0.97, 92.56 $\pm$ 1.32, and 90.55 $\pm$ 2.46 for attention modules 3, 2, and 1, respectively. The incorporation of max pooling improves all the results by a slight margin, in general. For example, an improved dice coefficient of 92.58 $\pm$ 1.30 is obtained in the case of attention module 2. Further, the use of max pooling in attention module 3 leads to a considerable improvement in the dice coefficient value (92.55 $\pm$ 1.12).

A sensitivity analysis of attention module hyperparameters was undertaken for the selection of the number of convolutional layers and the size of max pooling. First, let us consider attention module 2. It is found that increasing the max pooling from 2x2 to 3x3 enhances the dice coefficient, but on further increase, there appears to be a loss in information due to more feature compression. It is observed that the architecture with four convolutional blocks yields optimal results for this particular attention module, with a corresponding dice coefficient of 92.58 $\pm$ 1.30 as shown in Table 5. The improvement in performance could be attributed to the combined effect of tuned components of a well-designed attention module. The same process is repeated for attention module 3 and it is found that the design with four blocks with 2x2 max pool gives the optimal dice coefficient of 92.55 $\pm$ 1.12 (Dice Coefficient) as shown in Table 5.

The attention maps obtained from all four modules are shown in Figure 10 by the use of GRAD-CAM [44]. Since the performance on different datasets varied with disparate modules, the best BMSAN model (see Table 4) was selected for each case. The gradients are visualized in Figure 10. It is observed that the attention modules in the lower layer capture less semantic details, but on moving higher, they capture more details, which likely assists in creating better segmentation masks for corresponding datasets.

5 Conclusions and Future Scope

The work proposes a sophisticated deep learning based architecture for medical image segmentation that improves upon standard frameworks through the incorporation of novel attention modules and tuning hyperparameters at multiple scales using BO. Results show that the proposed BMSAN model, with fewer parameters, achieves similar or better performance than most of the SOTA models in segmenting a wide variety of medical datasets, including ISIC 2017, LUNGS, NERVE, Skin Lesion and CHEST datasets. The consideration of a multi-scaled loss function helps in faster convergence of the model and aiding in the incremental reconstruction of segmentation masks. The novel attention modules also contribute toward attaining superior results by assisting the model in focusing on the relevant regions of the images. The improved model also benefits from data-specific fine-tuned multi-scale coefficients of the architecture that are obtained using BO. The consideration of BO adds to the computational requirements, although the probabilistic model is notable for yielding optimized results with fewer evaluations of the objective function. It is anticipated that the incorporation and extension of ideas presented in this work can aid in the development of powerful machine learning models that can be employed in diverse settings. For example, the spatial attention modules can potentially boost the performance of other models too in disparate tasks, such as classification, detection, etc., in addition to the segmentation of medical images.

A key feature of this work is the optimization of complex deep learning architectures using a probabilistic approach. More work could be pursued in this direction. For example, the BO employed here tunes only the hyperparameters associated directly with the loss function of the model. The scope of the hyperparameters could be expanded to incorporate some key parameters associated with the architecture itself. Such processes can culminate in the development of more flexible models that can adapt to the nature and complexity of diverse datasets, unlike the current general trend of using data-specific models. The challenge from the deep learning side is in identifying a base architecture that is fluid enough to capture data-specific changes while retaining the capacity to extract and express generic properties of images (medical images in this case). BO, on the other hand, entails a well-informed selection of key model parameters for optimization as the probabilistic approach is ideally suited for a relatively smaller number of optimization variables. An appropriate synthesis of the two paradigms can create powerful models that can solve a wide variety of problems.

Acknowledgments

The authors are thankful to Swathy Prabhu Mj for arranging Asus RTX 2080 Ti (12 GB) and Quadro GV100 (32 GB) GPUs with 64 GB RAM, to hasten the research. The first author is thankful to Br. Tamal Maharaj and Dr. Jadab Kumar

Pal for their suggestions. The authors would also like to thank Github usernames: lixiaolei1982 and nibtehaz for their implementations of SOTA models.

Appendix A. (Performance Metrics)

For quantitative analysis the following performance metrics were used, including Dice Coefficient (**DC**), Precision (**PC**) and Jaccard Similarity (**JS**) or Intersection over Union (**IoU**). For calculating these we have to use the following variables, True Positive (**TP**), True Negative (**TN**), False Positive (**FP**), False Negative (**FN**), Ground truth (**GT**), and segmented result (**SR**).

Dice Coefficient, Jaccard Index (IoU) and Precision are calculated using the following equations,

$$DC = 2 \times \frac{GT \cap SR}{GT + SR} \tag{2}$$

$$IoU = \frac{GT \cap SR}{GT \cup SR} \tag{3}$$

$$PC = \frac{TP}{TP + FP} \tag{4}$$

The results of 5 Fold-Cross Validation are presented as $\mu \pm \sigma$, where μ is the mean and σ is the standard deviation of the five folds given by the following equation,

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \tag{5}$$

Statements and Declarations

Funding

There were no funding available for this study.

Conflict of interests

The authors declare no conflict of interests.

References

[1] M. Siam, S. Elkerdawy, M. Jagersand, and S. Yogamani. Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8, 2017.

[2] J. Bruce, T. Balch, and M. Veloso. Fast and inexpensive color image segmentation for interactive robots. In *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113)*, volume 3, pages 2061–2066 vol.3, 2000.

[3] J. S. Sevak, A. D. Kapadia, J. B. Chavda, A. Shah, and M. Rahevar. Survey on semantic image segmentation techniques. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 306–313, 2017.

[4] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul J. Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *J. Digit. Imaging*, 32(4):582–596, 2019.

[5] Jimut Bahan Pal and Nilayan Paul. Classifying chest x-ray covid-19 images via transfer learning. In *2021 Ethics and Explainability for Responsible Data Science (EE-RDS)*, pages 1–8, 2021.

[6] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[7] Nabil Ibtehaz and M. Sohel Rahman. Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74 – 87, 2020.

[8] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, and N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[10] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 287–297, Cham, 2018. Springer International Publishing.

[11] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.

[12] Junlong Cheng, Shengwei Tian, Long Yu, Hongchun Lu, and Xiaoyi Lv. Fully convolutional attention network for biomedical image segmentation. *Artificial Intelligence in Medicine*, 107:101899, 2020.

[13] Chen Li, Yusong Tan, Wei Chen, Xin Luo, Yulin He, Yuanming Gao, and Fei Li. Anu-net: Attention-based nested u-net to exploit full resolution features for medical image segmentation. *Computers and Graphics*, 90:11 – 20, 2020.

[14] M. Z. Alom, C. Yakopcic, T. Taha, and V. Asari. Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net). *NAECON 2018 - IEEE National Aerospace and Electronics Conference*, pages 228–233, 2018.

[15] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Bi-directional convlstm u-net with densley connected convolutions, 2019.

[16] Jimut Bahan Pal. Holistic network for quantifying uncertainties in medical images. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 560–569, Cham, 2022. Springer International Publishing.

[17] Chenhong Zhou, Shengcong Chen, Changxing Ding, and Dacheng Tao. Learning contextual and attentive information for brain tumor segmentation. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 497–507, Cham, 2019. Springer International Publishing.

[18] C. Kaul, S. Manandhar, and N. Pears. Focusnet: An attention-based fully convolutional network for medical image segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 455–458, 2019.

[19] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[20] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10):2281–2292, 2019.

[21] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing.

[22] Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 801–809, 2011.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[24] Yukihiro Nomura, Issei Sato, Toshihiro Hanawa, Shouhei Hanaoka, Takahiro Nakao, Tomomi Takenaga, Tetsuya Hoshino, Yuji Sekiya, Soichiro Miki, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. Development of training environment for deep learning with medical images on supercomputer system based on asynchronous parallel bayesian optimization. *J. Supercomput.*, 76(9):7315–7332, 2020.

[25] Rune Johan Borgli, Håkon Kvale Stensland, Michael Alexander Riegler, and Pål Halvorsen. Automatic hyperparameter optimization for transfer learning on medical image datasets using bayesian optimization. In *13th International Symposium on Medical Information and Communication Technology, ISMICT 2019, Oslo, Norway, May 8-10, 2019*, pages 1–6. IEEE, 2019.

[26] Kirthevasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabás Póczos, and Eric P. Xing. Neural architecture search with bayesian optimisation and optimal transport. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2020–2029, 2018.

[27] Colin White, Willie Neiswanger, and Yash Savani. BANANAS: bayesian optimization with neural architectures for neural architecture search. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10293–10301. AAAI Press, 2021.

[28] Mateusz Buda, Ashirbani Saha, and Maciej A. Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Comput. Biol. Medicine*, 109:218–225, 2019.

[29] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), 2018.

[30] M. A. Albahar. Skin lesion classification using convolutional neural network with novel regularizer. *IEEE Access*, 7:38306–38313, 2019.

[31] Keelin Murphy, Bram van Ginneken, Arnold M. R. Schilham, Bartjan de Hoop, H. A. Gietema, and Mathias Prokop. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Medical Image Anal.*, 13(5):757–770, 2009.

[32] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[33] Vidushi Vashishtha and D Aju. Nerve segmentation in ultrasound images. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–5, 2017.

[34] A. Sinha and J. Dolz. Multi-scale self-guided attention for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2020.

[35] Yun Jiang, Jing Gao, and Falin Wang. Joint optic disc and optic cup segmentation based on new skip-link attention guidance network and polar transformation. In Haiqin Yang, Kitsuchart Pasupa, Andrew Chi-Sing Leung, James T. Kwok, Jonathan H. Chan, and Irwin King, editors, *Neural Information Processing*, pages 399–410, Cham, 2020. Springer International Publishing.

[36] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[38] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017.

[39] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[40] H. Zhao, J. Jia, and V. Koltun. Exploring self-attention for image recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10073–10082, 2020.

[41] Y. Ho and S. Wookey. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, 8:4806–4813, 2020.

[42] Dripta Mj and Denys Dutykh. Learning extreme wave run-up conditions. *Applied Ocean Research*, 105:102400, 2020.

[43] J. Mockus, V. Tiesis, and A. Zilinskas. Toward global optimization, volume 2, chapter bayesian methods for seeking the extremum, 1978.

[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[45] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[47] Sekitoshi Kanai, Yasuhiro Fujiwara, and Sotetsu Iwamura. Preventing gradient explosions in gated recurrent units. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 435–444, Red Hook, NY, USA, 2017. Curran Associates Inc.