Multi-Scale ConvLSTM Attention-Based Brain Tumor Segmentation

Brahim AIT SKOURT
Laboratory of Intelligent Systems
and Applications
University of Sidi Mohammed Ben Abdellah
Fez, Morocco

Nikola S. Nikolov
Computer Science and
Information Systems Department
University of Limerick
Limerick, Ireland

Aicha MAJDA
University of Moulay Ismail
Networks and Computer Systems
research team
Meknes, Morocco

Ahlame BEGDOURI
Laboratory of Intelligent Systems and Applications
University of Sidi Mohammed Ben Abdellah
Fez, Morocco

***Abstract*—In computer vision, there are various machine learning algorithms that have proven to be very effective. Convolutional Neural Networks (CNNs) are a kind of deep learning algorithms that became mostly used in image processing with a remarkable success rate compared to conventional machine learning algorithms. CNNs are widely used in different computer vision fields, especially in the medical domain. In this study, we perform a semantic brain tumor segmentation using a novel deep learning architecture we called multi-scale ConvLSTM Attention Neural Network, that resides in Convolutional Long-Short-Term-Memory (ConvLSTM) and Attention units with the use of multiple feature extraction blocks such as Inception, Squeeze-Excitation and Residual Network block. The use of such blocks separately is known to boost the performance of the model, in our case we show that their combination has also a beneficial effect on the accuracy. Experimental results show that our model performs brain tumor segmentation effectively compared to standard U-Net, Attention U-net and Fully Connected Network (FCN), with 79.78 Dice score using our method compared to 78.61, 73.65 and 72.89 using Attention U-net, standard U-net and FCN respectively.**

***Keywords*—Convolutional neural networks; image processing; semantic brain tumor segmentation; convolutional long short term memory; inception; squeeze-excitation; residual-network; attention units**

I. Introduction

In recent years, there have been a large number of contributions to the field of deep learning. Year after year, deep learning proves its superiority by surpassing state-of-the-art solutions in various domains such as computer vision [1], natural language processing [2], speech recognition [3] and many other application domains. In particular, in the field of computer vision, which is the main focus of this paper, deep learning has been enormously successful for tasks such as image classification [4], face recognition [5], object detection [1] and image segmentation [6]. The significant improvement in the field of deep learning is due to multiple factors. High-performance computational resources (GPU, TPU) have become more easily available. At the same time, the investments made in research as well as the amounts of collected data have also increased.

Deep learning algorithms, like other machine learning algorithms are categorized into two main categories: supervised and unsupervised algorithms. Unsupervised algorithms, an example of which is Deep Belief Network (DBN), work with unlabeled data. They use a greedy layer-wise learning strategy to fine-tune the network's parameters. This learning strategy, which is based on a contrastive version of the wake-sleep algorithm [7], performs quickly and can find a good set of parameters, even with relatively very deep architectures. Supervised algorithms, an example of which is Convolutional Neural Network (CNN), work with labeled data. CNNs have been particularly successful for solving computer vision tasks. The fine-tuning phase of a CNN is composed of consecutive convolution and pooling operations for extracting fine-tuned features, which are then used in the discriminative phase of the training process. This automated process of extracting features is what made deep learning algorithms powerful, as opposed to conventional machine learning algorithms that use hand-crafted features.

Nowadays, the use of CNNs is widespread across industries and businesses. In healthcare, CNNs achieve very promising results due to their robust feature extraction capabilities. For example, in medical image segmentation, they have achieved state-of-the-art performance [8] with a significant margin compared to conventional machine learning models, which makes them the most popular choice in different medical imaging fields. They also dominate the health informatics literature on brain [9], lung nodule [10], spleen [11], and cardiac [12] medical imaging issues, to mention a few.

In this work, we perform brain tumor semantic segmentation using a novel deep learning architecture. Brain tumors are considered one of the deadliest cancers in the world. There are various brain tumor types, but gliomas are the most common ones among adults. Furthermore, gliomas can be present with different degrees of aggressiveness with an average survival time for patients diagnosed with glioma lesser

than 14 months [13]. Therefore, time is a critical factor for doctors to act regarding gliomas. To diagnose a brain tumor, there are different types of medical image acquisition involved, such as MRI, CT scans and X-Ray, each having its pros and cons. For example, CT scans have the advantage of speed of tissue acquisition at the cost of lower quality of tissue contrast and higher radiation risk. On the other hand, MRIs are slow compared to CT scans but they are best suitable for capturing abnormal tissues with more details due to their accuracy in acquiring different types of contrasts. After the acquisition of the brain region, radiologists perform a manual segmentation of brain tumors from MRI images, which is time-consuming. Therefore, designing an automatic brain tumor segmentation is mostly desirable.

In this paper, we propose a novel deep neural network, called Multi-Scale ConvLSTM Attention Neural Network (MSConvLSTM-Att), to automatize brain tumor semantic segmentation. Our architecture is multi-scale-attention based with each level using Convolutional Long Short Term Memory (ConvLSTM) [14], Squeeze and Excitation-inception (SE-inception) [15] and Squeeze and Excitation-Residual-Network (SE-ResNet) [15]. The motivation behind using such architecture is to gather state-of-the-art feature extraction methods, the LSTM and the attention mechanism in one multi-scale architecture and perform brain tumor semantic segmentation efficiently compared to conventional deep learning based architectures.

The use of such multi-scale architecture, which is composed of multiple stages, is to generate multiple versions of the same image with different resolutions, each containing diverse semantics. The first low-level stage serves to model the spatially sequential relationship between different parts of each MRI modality (FLAIR, T1w,T1gd,T2w)[1], while the next stage manages the extraction of local features in addition to decreasing the size of the images for computational optimization. Finally, the third high-level stage captures the global representations. Thereafter, at each level, we introduce a stack of attention modules to gradually emphasize the regions that contain a large number of semantic features.

The integration of attention mechanism in the image segmentation of natural scenes has been widely adopted [16], [17], [18], [19]. However, in medical imaging, the inclusion of attention mechanism is rare [20], [21], [22], [23]. For this reason, we investigate the impact of a simple attention module in boosting the performance of standard deep networks for brain tumor semantic segmentation. Experimental results show that our proposed method improves the segmentation performance by modeling a combination of rich contextual features with local features.

The remainder of this paper is organized as follows. Next section presents related works. In Section III we introduce our proposed method in detail. Thereafter, we present and discuss the obtained results in section IV. Finally, we conclude our paper in section V.

II. Related Work

Most of the state-of-the-art deep learning architectures used for automatic medical image segmentation are inspired from Fully Convolutional Networks (FCN) [24] or U-Net [25]. Many variants of these architectures have been proposed to perform semantic segmentation in different application domains [26], [27], [28], [29].

FCN is an architecture in which fully connected layers are replaced by deconvolution layers to generate segmentation masks [24]. Jesson *et al.* [30] proposed a variant of the standard FCN with a multi-scale loss function. With this approach it is possible to model the context in both the input and output domains. A limitation of this approach is that FCN is not able not explicitly model the context in the label domain. Compared to U-Net, FCN does not use skip connections between the contracting (i.e feature extraction path) and the expanding paths (i.e data reconstruction path).

The U-Net architecture was introduced by Ronneberger *et al.* in 2015 [25]. It overcomes the limitations of FCN by including features from the contracting path. In order to obtain the missing feature-contexts, multi-scale features are concatenated in a mirroring way. Many works have adopted this architecture to perform medical image segmentation over different parts of the human body. In a previous work of ours [26], we also adopted the U-Net architecture to perform lung CT image segmentation.

A limitation of both FCN and U-Net is that they both do not perform very well in multi-class segmentation tasks [31]. To overcome this issue, cascaded architectures can be used. They have the beneficial effect of decomposing a multi-class segmentation problem into multiple binary segmentation problems. This approach is also used in various medical image segmentation works. For example, Chen *et al.* [32] adopted a cascaded classifier to perform a multi-class segmentation. Furthermore, in [33] authors proposed a cascaded architecture to merge different feature extraction methods. Nonetheless, these models still face a problem of focusing on pixel level classification while ignoring adjacent pixels' connections. To overcome this issue, Generative models were adopted. A widely used variant of generative models is Generative Adversarial Network (GAN) [34]. GANs are employed for semantic segmentation in the following way: a convolutional semantic segmentation network is trained along with an adversarial network to discriminate segmentation maps [35]. That is, two models are trained; the first captures data distribution, while the second is used for a discriminative purpose.

To capture sequence patterns in medical imaging, Recurrent Neural Networks (RNNs) are typically used as they they are well suited for handling sequential data. Specifically in medical image segmentation, RNNs are used to keep track of features in previous image slices in order to better generate the corresponding segmentation maps. There are various RNN architectures mentioned in the literature, and amongst them Gated Recurrent Units (GRU) [36] and Long-Short Term Memory (LSTM) [37] are likely the most robust and widely used. GRU is memory efficient, nonetheless not very suitable for keeping track of long-term features. LSTM is better adapted to such tasks due to the forget gate that preserves features from previous sequences to use in upcoming sequences. [38], [39], [40] are some examples of employing RNNs for performing image segmentation for sclerosis lesions and brain tumors respectively.

[1]https://case.edu/med/neurology/NR/MRI Basics.htm

In the last few years, a new concept called *attention mechanism* was introduced into computer vision tasks. Attention mechanism was introduced first in neural machine translation [41] to help remember long range context from long source sentences. The added value brought by attention modules is the creation of shortcuts between the input sentence and the context vector. Attention in deep learning can be interpreted as a vector of weights that represent the importance of an element within a context. The attention vector is used to estimate how strongly is an element related to other elements (elements in this context are image pixels), it takes the sum of these elements' values weighted by the attention vector as the approximation of the target context.

The success of the attention mechanism for neural machine translation has encouraged its application to computer vision immediately [42]. In medical image segmentation, the attention mechanism was adopted in many works and various variants of attention modules have been introduced. In [43], authors propose a combination of FCN with a Squeeze and Excitation (SE) attention-based module to perform whole-brain and whole-body segmentation. They integrate the SE block in three ways: channel SE (cSE), spatial SE (sSE) and concurrent spatial-channel SE (csSE). In [20], Wang et al. perform prostate segmentation in ultrasound images using deep attentional features. They use an attention module to extract refined features at each layer, eliminate non-prostate noise and focus on more prostate details at deep layers. Furthermore, Li et al. propose an auto-encoder CNN-based architecture, called hierarchical aggregation network (HAANet) [21], which combines the attention mechanism and hierarchical aggregation to perform 3D left atrial segmentation. In another work, Oktay *et al.* propose an attention U-net [44] which extends the U-Net architecture by incorporating an attention gate in the expanding path in order to accurately segment the pancreas area.

III. Method

In this section, we describe our proposed architecture for brain tumor segmentation. Our method combines different techniques in order to extract relevant features and keep track of them during the entire process of segmentation.

We combine Inception, ResNet and Squeeze-Excitation blocks in one part of our architecture for relevant feature extraction, and attention modules in another part to perform brain tumor segmentation. The combination of Inception, ResNet and Squeeze-Excitation is known as the most successful architecture in the ImageNet challenge. With this combination, the team Trimps-Soushen achieved 2.99% error rate in object classification in the ImageNet challenge[2].

We first feed our network different modalities of brain MRI images (FLAIR, T1w, T1gd, T2w) to include various intensities and to better perform the semantic segmentation. Each modality is split into four patches, then for each modality, three scales of feature extraction are performed. The motivation behind this multi-scale mixture is to best separate each tumor label (enhancing tumor, tumor core, whole tumor and background).

[2]https://image-net.org/challenges/beyond_ilsvrc

At the first scale, ConvLSTM is used over each of the four patches to preserve the correlation among features. ConvLSTM are best suitable for catching spatiotemporal information without any much redundancy [14]. At the second scale, an SE-inception [45] module is used over the output of the first scale to extract low level features and decrease the computation cost. Fig. 1 shows the inception module [46] and Fig. 2 shows the SE-inception block.

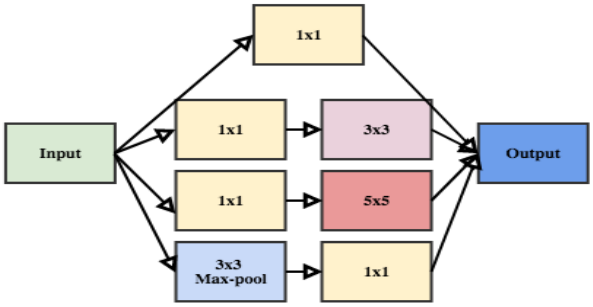

Fig. 1. Inception Block.

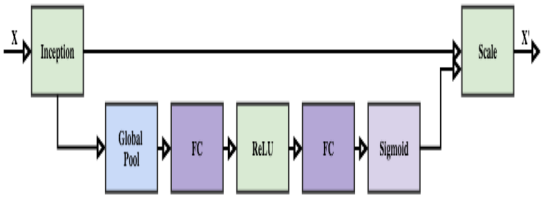

Fig. 2. SE-Inception Block.

At the third scale, we extract high level features by integrating an SE-ResNet module [45]. The use of such block increases computational complexity with a thin margin but in exchange of increasing the accuracy [45]. The ResNet block [47] and SE-ResNet are described in Fig. 3 and Fig. 4 respectively.

At each scale (different scales are highlighted by green color in Fig. 6), we combine the four outputs to form what we call *single-scale features* as stated in Fig. 6. These three *single-scale features* are then concatenated and convolved to form *multi-scale features* as mentioned in the same figure. We then take the *multi-scale features* and we combine them with each single-scale feature.

At this stage, our model holds general context feature-maps that contain different levels of features, from low to high level features. Thereafter, we add a convolution layer to refine these features.

Furthermore, in order to explore more global contextual characteristics by building connections among features, we

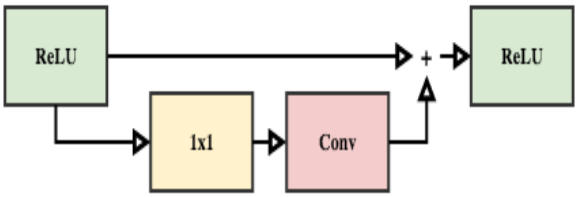

Fig. 3. ResNet Block.

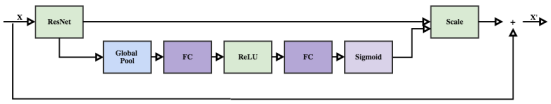

Fig. 4. SE-ResNet Block.

include attention mechanism in the form of a location-based attention module, we call it Spatial Attention Module (SAM). The attention mechanism is presented in the Fig. 5.

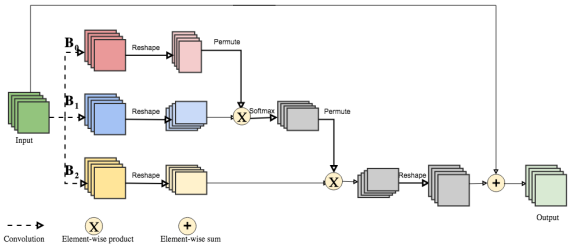

Fig. 5. Spatial Attention Module.

In Fig. 5, we assume the input to the SAM module is V, which is a 3D shaped input (W, H, C), here W, H and C represent the width, height and depth respectively. In the red branch, we perform a convolution operation, resulting in a feature map B_0 with same width and height but with depth equals to $C/8$. B_0 is then reshaped to (W, H, C). The same operation is applied to the blue branch B_1. Thereafter, we perform a matrix multiplication $B_0 * B_1^T$ and apply a softmax operation to calculate the spatial attention map following the formula in (1), where $S_{i,j}$ represents the impact of the pixel in the i^{th} position on the pixel in the j^{th} position.

$$S_{i,j} = \frac{exp(B_0 * B_1)}{\sum_{i=1}^{W*H} exp(B_0 * B_1)} \quad (1)$$

The yellow branch performs a convolution and results in B_2 with the same shape as V. B_2 is then reshaped to (C, W, H) then it is multiplied by the transpose of the spatial attention map S. Furthermore, the output R is reshaped to $C \times (W \times H)$ and multiplied by a parameter λ and then an element-wise sum with input V is performed to obtain the output O as expressed in (2).

$$O = \lambda \sum_{i=1}^{W*H} exp(S_{i,j} * B_2) + V \quad (2)$$

In (2), λ is initialized by 0 and gradually updated to give more weight to the spatial attention map, as adopted in [17].

At the last level, we perform a convolution operation to generate the final prediction map for each scale and then average all these maps to output the segmentation map. Fig. 6 presents an overview of our proposed architecture.

IV. Experiments and Results

To evaluate our architecture, we are using BRATS'18 data set for brain tumor segmentation, provided in the Medical Segmentation Decathlon Challenge[3]. This data set contains multimodal MRI data (FLAIR, T1w,T1gd,T2w)[4]. Furthermore, it contains 210 High Grade Glioma (HGG) scans and 75 Low Grade Glioma (LGG) scans. In this data set, the focus is mainly on the segmentation of different sub-regions of the glioma. First, the enhancing tumor (ET), the tumor core (TC) and finally the whole tumor (WT) as can be seen in Fig. 7. Each one of these sub-regions have some specific characteristics regarding their intensities, hence different modalities are responsible for capturing different characteristics. For example, the ET is described by areas that are hyper-intense in T1gd. The appearance of the non-enhancing tumor (NET) (solid parts) and the necrotic (NCR) (fluid-filled) is represented by areas that show hypo-intensity in T1gd when compared to T1. The WT describes the whole disease and it contains the TC and the peritumoral edema (ED), which is characterized by hyper-intensity in the FLAIR modality. The provided labels in this data set are as follows: 1 for NCR and NET, 2 for ED, 3 for ET and finally 0 for other parts of the brain. The annotations were created by domain experts and approved by other domain experts as described in [48].

Given the presence of different features related to gliomas in different modalities, we feed the four modalities as input to our architecture, then we get the semantic segmentation that belongs to these inputs. The loss function we use is the dice loss optimized using the Adam optimizer [49]. The learning rate is initially set to 0.001 and then multiplied by 0.5 after each 30 epochs. We used 500 epochs to train our network. Due to limitations in computational resources, we reduced the input size to 190×190 by cropping some of the background area and we only took from the 30^{th} slice to the 120^{th} given that most of the brain information is present in that interval. Furthermore, we normalize the inputs to have zero mean and unit standard deviation.

In addition, given that each session of the notebook used for training has 12 hours lifetime, we use the following strategy to train our network. We save our model and its weights after each 50 epochs and we reload it and continue training with

[3]http://medicaldecathlon.com
[4]https://case.edu/med/neurology/NR/MRI Basics.htm

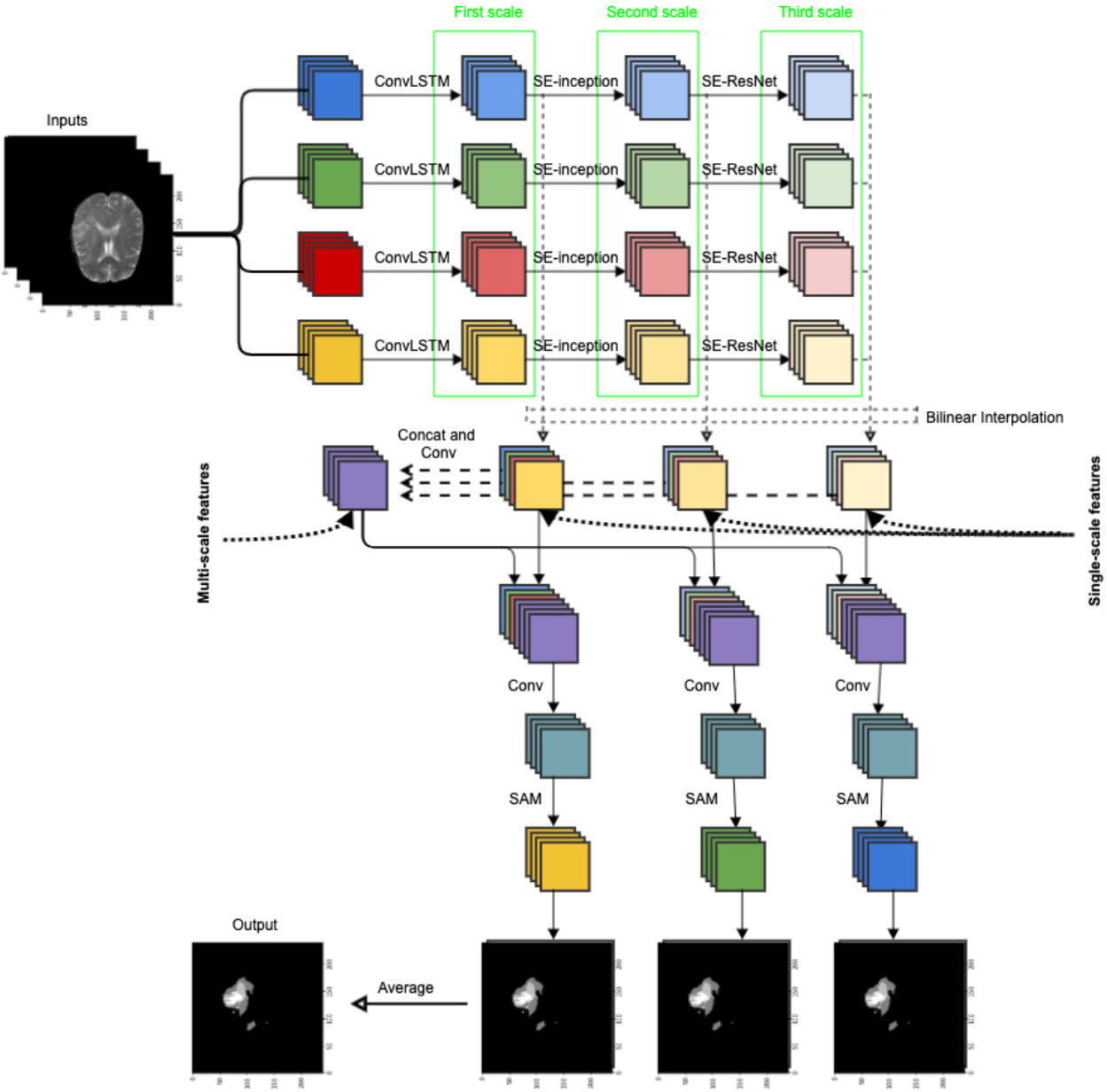

Fig. 6. Overview of our Proposed Architecture.

new data. For development, we shuffled and randomly split the images into training (225 patients), validation (30 patients), and test (30 patients). Experiments were performed in a server equipped with a single 12GB NVIDIA Tesla K80 GPU.

We compare our method with the standard UNet [25], standard FCN [24] and the Attention U-Net [44] architectures. And we evaluate their performance using the dice coefficient (DSC) as a comparison metric. Table I contains experimental results obtained using the different segmentation methods described above, and compared regarding their DSC score. Our proposed architecture achieved the best score with 64.95, 88.16 and 86.50 in ET, WT, and TC respectively and a mean score of 79.87.

It can be observed that both our method and AttUnet, which also includes attention modules, perform better than the other ones without attention modules. This proves that adding attention modules surely enhances the segmentation procedure by putting more attention into the tumor location. Oktay *et al.* have reported the same observation in their work with MSConvLSTM-Att [44].

Our architecture outperforms AttUNet with a significant margin, this is mainly due to the focus on location attention modules, besides the use of powerful feature extraction modules (ConvLSTM, SE-Inception and SE-ResNet) in the

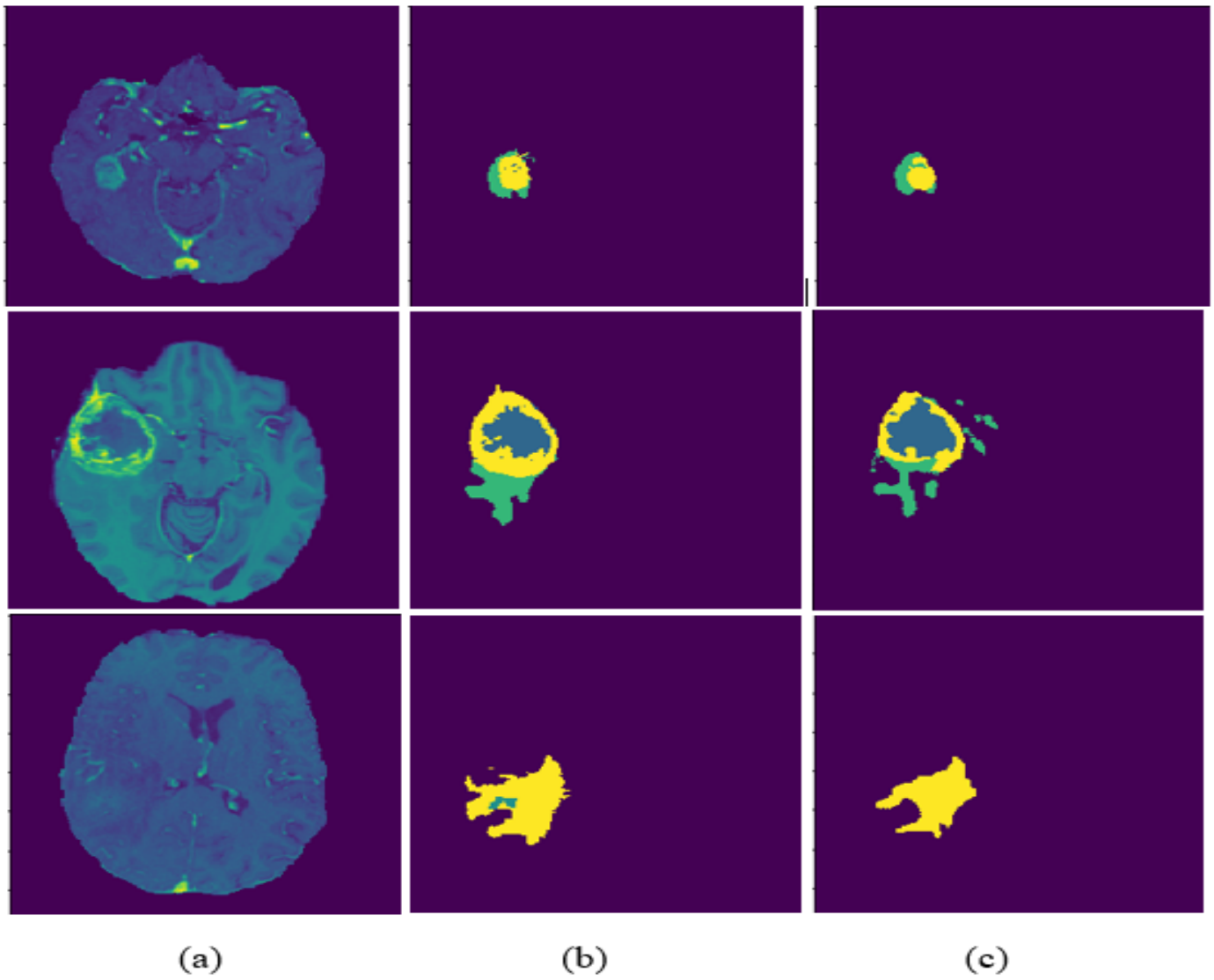

Fig. 7. Segmentation Results Sample: (a) is the Input MRI Images, (b) is the Ground Truth and (c) is the Segmentation Results using our Proposed Architecture.

TABLE I. PROPOSED METHOD'S DSC SCORE COMPARED TO THOSE OF U-NET, ATT-UNET AND FCN

Labels	ET	WT	CT	Mean
U-Net	0.563	0.848	0.797	**0.736**
Att-UNet	0.637	0.875	0.845	**0.786**
FCN	0.551	0.853	0.781	**0.728**
ours	0.649	0.881	0.865	**0.798**

first part of the architecture, which is beneficial in eliminating irrelevant features. Our proposed architecture can be implicitly considered as a cascaded architecture even though we do not explicitly use multiple cascaded architectures.

Fig. 7 displays a sample of the input MRI images, ground truth and the segmentation results using our proposed architecture. As seen in Table I, the ET segmentation has the smallest DSC value. It can be seen also in Fig. 7, where the ET region is not well detected especially in the first and third row.

It has to be mentioned that our method is slightly slower compared to the other methods, which is normal given the fact that complex building blocks has been used in order to ensure a better segmentation result.

V. CONCLUSION

In this paper, we propose a novel deep learning architecture for brain tumor segmentation we call multi-scale ConvLSTM Attention Neural Network, and we compare its performance to various deep learning architectures that are tailored to such kind of tasks. Our proposed method is built as a multi-scale

architecture composed of different state-of-the-art feature extraction blocks such as Inception, Squeeze-Excitation, Residual Network, ConvLSTM and finally Attention units. We compare the performance of our architecture to standard U-net, AttU-net and FCN that have shown effective results in semantic segmentation. Experimental results show that our proposed model outperforms standard U-net, AttU-net and FCN in terms of dice score. Our model reached 79.78 as a mean dice score for the three parts of the brain tumor, while Attention U-net, standard U-net and FCN reached 78.61, 73.65 and 72.89 respectively. We observe that both our method and the AttU-net perform better than the other ones, which can be explained that the integration of attention modules enhances the segmentation procedure. Besides, our method outperforms the AttU-net, and this is due to the use ConvLSTM, SE-Inception and SE-ResNet.

REFERENCES

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[3] S. Hourri, N. S. Nikolov, and J. Kharroubi, "Convolutional neural network vectors for speaker recognition," *International Journal of Speech Technology*, pp. 1–12, 2021.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[5] J. Wei, "Video face recognition of virtual currency trading system based on deep learning algorithms," *IEEE Access*, vol. 9, pp. 32 760–32 773, 2021.

[6] Y. Jalali, M. Fateh, M. Rezvani, V. Abolghasemi, and M. H. Anisi, "Resbcdu-net: A deep learning framework for lung ct image segmentation," *Sensors*, vol. 21, no. 1, p. 268, 2021.

[7] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[8] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[10] A. E. HASSANI, B. A. SKOURT, and A. MAJDA, "Efficient lung nodule classification method using convolutional neural network and discrete cosine transform," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120296

[11] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori, "Hierarchical 3d fully convolutional networks for multi-organ segmentation," *arXiv preprint arXiv:1704.06382*, 2017.

[12] R.-R. Galea, L. Diosan, A. Andreica, L. Popa, S. Manole, and Z. Bálint, "Region-of-interest-based cardiac image segmentation with deep learning," *Applied Sciences*, vol. 11, no. 4, p. 1965, 2021.

[13] M. S. Walid, "Prognostic factors for long-term survival after glioblastoma," *The Permanente Journal*, vol. 12, no. 4, p. 45, 2008.

[14] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 2015, pp. 802–810, 2015.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[17] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[18] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[19] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267–283.

[20] Y. Wang, Z. Deng, X. Hu, L. Zhu, X. Yang, X. Xu, P.-A. Heng, and D. Ni, "Deep attentional features for prostate segmentation in ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 523–530.

[21] C. Li, Q. Tong, X. Liao, W. Si, Y. Sun, Q. Wang, and P.-A. Heng, "Attention based hierarchical aggregation network for 3d left atrial segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 255–264.

[22] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[23] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: attention based semi-supervised deep networks for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 370–378.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[26] B. A. Skourt, A. El Hassani, and A. Majda, "Lung ct image segmentation using deep neural networks," *Procedia Computer Science*, vol. 127, pp. 109–113, 2018.

[27] J. Dolz, X. Xu, J. Rony, J. Yuan, Y. Liu, E. Granger, C. Desrosiers, X. Zhang, I. Ben Ayed, and H. Lu, "Multiregion segmentation of bladder cancer structures in mri with progressive dilated convolutional networks," *Medical physics*, vol. 45, no. 12, pp. 5482–5493, 2018.

[28] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[29] M. P. Heinrich, O. Oktay, and N. Bouteldja, "Obelisk-net: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions," *Medical image analysis*, vol. 54, pp. 1–9, 2019.

[30] A. Jesson and T. Arbel, "Brain tumor segmentation using a 3d fcn with multi-scale loss," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 392–402.

[31] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[32] X. Chen, J. H. Liew, W. Xiong, C.-K. Chui, and S.-H. Ong, "Focus, segment and erase: an efficient network for multi-label brain tumor segmentation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 654–669.

[33] J. Liu, F. Chen, C. Pan, M. Zhu, X. Zhang, L. Zhang, and H. Liao, "A cascaded deep convolutional neural network for joint segmentation and genotype prediction of brainstem gliomas," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1943–1952, 2018.

[34] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[35] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] S. Andermatt, S. Pezold, and P. C. Cattin, "Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 31–42.

[39] T. H. N. Le, R. Gummadi, and M. Savvides, "Deep recurrent level set for segmenting brain tumors," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 646–653.

[40] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical image analysis*, vol. 43, pp. 98–111, 2018.

[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[42] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[43] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 421–429.

[44] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[48] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.