Review on Deep Learning based Medical Image Processing

Aman Agarwal
Research Scholar
Computer Science and Engineering
Chandigarh University,
Punjab, India
aman.agar85@gmail.com

Rakesh Kumar
Professor
Computer Science and Engineering
Chandigarh University,
Punjab, India
rakesh77kumar@gmail.com

Meenu Gupta
Associate Professor
Computer Science and Engineering
Chandigarh University,
Punjab, India
gupta.meenu5@gmail.com

2022 IEEE International Conference on Current Development in Engineering and Technology (CCET) | 978-1-6654-5415-5/22/$31.00 ©2022 IEEE

***Abstract*—Deep learning (DL) has made extensive progress in many exploration regions. Computer vision is one of the most trending fields advancing due to extensive research in developing DL models, mainly focusing on image pattern recognition. These days, numerous specialists attempt to track down answers for issues in different fields under the illumination of DL techniques. In this review, important information is added about DL models and testing points specialists can use in DL. This study explored DL and concentrates on what advancements are made in the most famous fields, for example, Medical Image segmentation, automated disease classification, 4D Motion Estimation, etc. Moreover, this study discusses the leftover difficulties of these examination regions that DL can tackle and talks about future subjects to help scientists, especially in handling Medical Images.**

Keywords— Convolutional Neural Network (CNN), DL, medical image segmentation, U-Net, residual learning

I. INTRODUCTION

Medical Image Analysis is the backbone of Medical Research. Computer vision, pattern recognition, image mining, and machine learning have become increasingly common in medical image processing. It provides the cardinal evidence needed in medical research done in laboratories and the diagnosis made by specialized individuals. As a result, highly specialized medical imaging techniques were developed over the years, providing the base for Medical Image Analysis. Some of them include Computer Tomography (CT), Ultrasound (US), Magnetic Resonance Imaging (MRI), and X-Rays. Among these, CT has a higher resolution on high-density tissue, but it relies on clinicians' skills and harms human bodies like X-Rays. But X-Rays, cheaper and more convenient, are preferred for first medical reports. This shows that different imaging techniques have their own set of characteristics, but DL algorithms could match them effectively. CNN is one of the most classical methods used in image analysis. CNN helps reduce the number of parameters without compromising the quality of models, which is of utmost importance when dealing with images with high dimensionality and is preferred over traditional Neural Networks (NN) [1].

CNN delivers the best precision compared to state-of-the-art methods, making DL reach an incomparable level of precision in image classification and analysis. The typical use of CNN was in classification. However, in biomedical image processing tasks, the desired output needs localization, or each pixel is supposed to be assigned a class or a label.CNN techniquehas been used for a long time in medical image-processing tasks. This work can be seen as a compilation of some of the most popular CNN models used for medical image processing.

This work has been further classified into the following sections- Section II discusses the background study of the development of CNN. Section III reviews attention modules that can be used with CNN to get better results. Further, Section IV discusses the different techniques used to segment medical images. Finally, this work is concluded in Section V.

II. BACKGROUND STUDY

CNN, an excellent feature extractor, is widely used for medical image processing tasks. Yann LeCunn introduced the founding stone of the modern-day CNN in 1988 who developed LeNet-5 [2], which is also regarded as the standard architecture for modern-day CNNs. LeNet-5 consists of around 60000 parameters with 2 convolutional and 3 fully connected layers. Another architecture similar to LeNet was AlexNet [3], introduced in 2012. However, it is similar to LeNet but has some constraints, like adding more filters for categorizing more objects. The development of different DL models over the years is shown in fig. 1.

Next, VGG-16 was developed as the authors in [4] believed that stacking more layers in the CNN model was the best way to improve the accuracy. It was generated using 13 convolutional layers and 3 deep layers. It provided good results, but as deeper networks started getting developed, the degradation problem was exposed. With the increase in depth of the network, accuracy started getting saturated. To carry out semantic segmentation, thousands of training images are required, which is usually beyond reach in biomedical tasks.

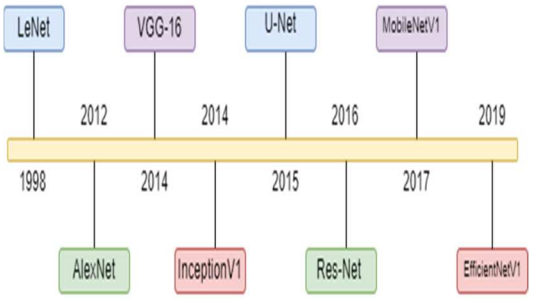

Fig. 1. Timeline showing the development of CNN over the years

Keeping this in mind, the authors of [5] built upon the "Fully Convolutional Network" (FCN) [6], where the architecture of CNN was modified in such a way that it worked with very few training images and yielded a more precise segmentation. FCN provided great results compared to the state-of-the-art methods as it trained pixel-to-pixel on semantic segmentation without much change in

the machinery. The authors in [5] introduced a new architecture by introducing a new operation to the existing FCN, which was the concatenation operation. The output of the first convolutional layer was concatenated with the output of upsampling layer. This architecture was named as U-Net as it has a U shape, as shown in fig. 2.

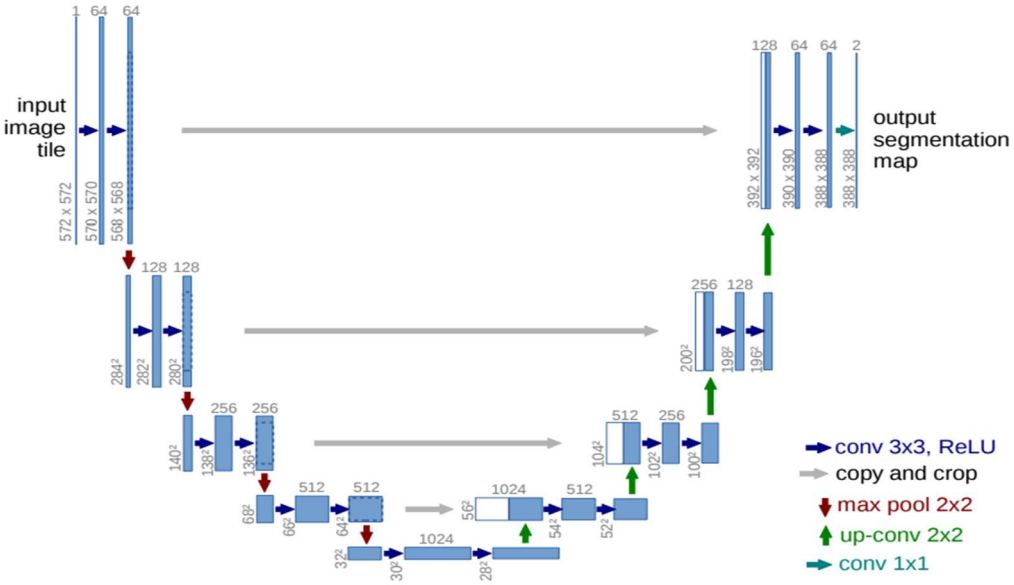

Fig. 2. U-Net architecture diagram

The architecture looks similar to Autoencoders [7]. Autoencoders have a downsampling step followed by upsampling. It was used to remove noise from the image, but it can be tricked to perform segmentation tasks by providing the segmented mask as output to the model. The down-sampling layer (left side) is similar to traditional NN with repeated 3x3 convolutions followed by a 2x2 max-pooling operation with a stride of 2. The activation function used was Rectified Linear Unit (ReLU).At each downsampling step, the number of feature channels was doubled. After this, an up-sampling operation (right side) was performed, followed by a 2x2 convolution. Concatenation with the corresponding feature map from the downsampling path was also done. Following this, 3x3 convolution was applied, with the activation function being ReLU. At the final layer, 1x1 convolution is used to map each 64-component feature vector to the desired number of classes.

Data augmentation with elastic deformation played a significant role in helping the U-Net model achieve very precise results in various biomedical image segmentation tasks, as shown in Fig. 3. The model required very few training images because of data augmentation. This gave confidence to the authors that U-Net could be used in image segmentation tasks which proved to be very accurate as most of the models that are developed to date, somewhere, use U-Net. Another problem faced by researchers while training Deep Neural Networks (DNN) was the problem of vanishing gradients [8]. The sigmoid, as well as the ReLU activation functions, faced the problem of vanishing gradients. The authors [9] solved this problem by introducing the residual network or the ResNet. Here, the authors introduced a new concept of skip connections. The skip connections skip training from a few layers and connect it directly to the output. It reduces the overall loss of features that are supposed to be lost in deeper networks, which are carried from one block and added to a block later.

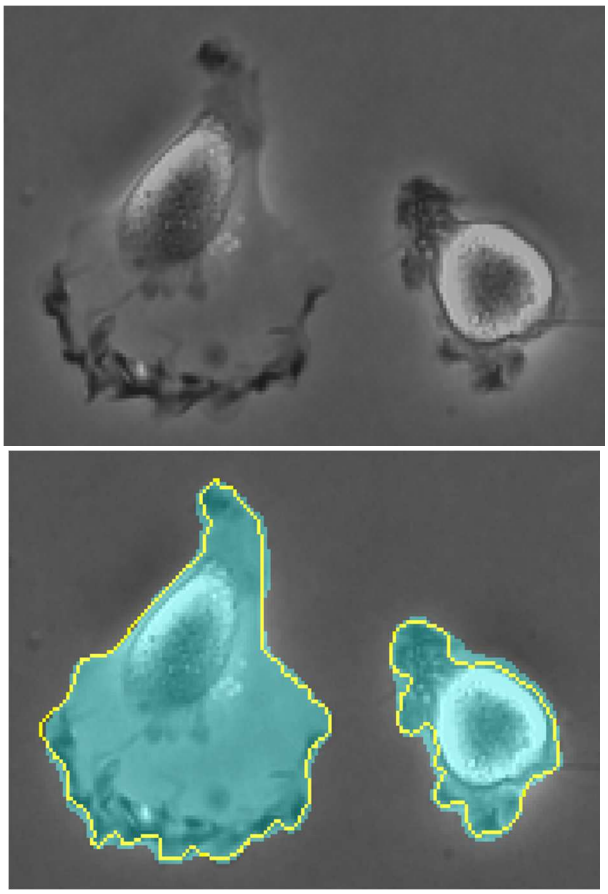

Fig. 3. Results of ISBI cell tracking challenge

The authors hence explicitly let the layers fit a residual mapping. Let us consider an input x to a convolution layer followed by activation function ReLU whose output is again passed through a similar layer with activation function ReLU (Let this operation be $F(x)$). Now to introduce residual learning, the output of this entire operation is stacked with the original input. The final function then becomes $H(x) = F(x) + x$. This was how the authors reduced the error and developed a less time-consuming model. Figure 4 shows the block diagram for the same.

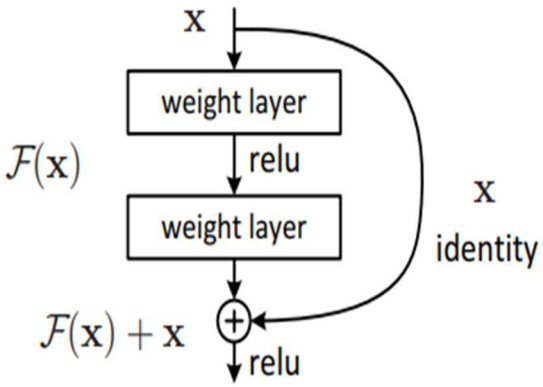

Fig. 4. Building block of Residual Learning

This model enabled the construction of deeper models without producing higher training errors. ResNet was a heavy model and could not be used for faster computations required for mobile applications. In [10], the authors introduced a lightweight network called MobileNetV1. It used depth-wise separable convolutions andwas mainly used for object detection and classification tasks. MobileNetV3 was introduced in 2019 and developed to be even lighter than MobileNetV1 and could be used for segmentation tasks. In [11], the authors introduced EfficientNet, a scaling method. Using a compound coefficient, it can uniformly scale all width, resolution, and depth dimensions. Its application is mainly in classification tasks.

III. ATTENTION MODULES

The attention mechanism has recently been an essential element in DL research. Initially, this mechanism was introduced in Natural Language Processing (NLP) tasks by the authors of [12] to tackle the forgetting issue in encoder and decoder architecture and was further extended to other fields, such as Computer Vision. Attention can be described with its literal meaning, i.e., to focus on something. Similarly, the model is made to learn how to focus on the essential features in the input. Cognitive science asserts that the human visual nerve gets vast volumes of information that it cannot handle. As a result, the human brain weighs the input and concentrates solely on the pertinent data. Researchers have adopted a similar idea in many fields and developed numerous attention methods to enhance the performance of DNN models in machine translation, reinforcement learning, generative models, visual recognition, etc. Recent advances in machine learning, especially DL, and the growing ability to analyzevast and multiple input data streams have led to these innovations.

Attention modules can be broadly classified into two major categories- (I) Soft Attention and (II) Hard Attention. (I) can be further classified into Spatial, Channel, and Self-Attention, whereas (II) can be classified into Bayesian, Gaussian, and reinforced attention modules. In total, there are 50+ attention modules developed so far, which are used in various tasks. Channel attention focuses on extracting channel-wise features as each feature map provides information about specific input parts. It can be seen as an extension to Squeeze and Excitation (SE) Attention which takes an input and converts each channel into a single value. For this purpose, it uses global-average pooling. The output is then passed to fully connected layers [13]. The complexity of SE attention was reduced with the introduction of Efficient Channel Attention. Spatial attention's primary function is to find a critical area in an image. Instead of channel attention, spatial attention focuses on a spatial map's essential areas. Hence, this attention is helpful in medical image segmentation and object detection tasks. On the other hand, it is suggested that self-attention can extract the links between input sequence tokens to encode higher-order interactions and contextual information. Soft functions like softmax and sigmoid are frequently used in soft attention approaches, which calculate the attention scores as the weighted sum of all the input entities. These approaches can be trained using back-propagation strategies because they are differentiable.

Hard attention chooses one of the states as the attention score rather than utilizing the weighted average of the hidden states. Using latent random variables as attention scores is the primary concept behind both Bayesian and variational attention. Gaussian attention utilizes a 2D Gaussian kernel, while reinforced attention substitutes a Bernoulli-sigmoid unit for softmax. In [14], the authors introduce the "Residual Attention Network," a CNN which uses an attention mechanism and can blend very well with the state-of-art forward propagating network architecture. Attention modules are introduced in Residual Neural Networks, which help generate attention-aware features. The features change as the network goes deeper. Importantly, attention residual learning was proposed to train deep attention residual networks.

IV. MEDICAL IMAGE SEGMENTATION

Medical image segmentation is one of the most crucial jobs in medical image analysis. Traditional medical image segmentation often entails intuitively detecting image features like lines and edges and mathematically tracing image gradients along object boundaries using techniques like graph cuts, active contours, level-set, etc. The DL-based segmentation approach can be used in routine radiation treatment procedures since it may speed up the contouring process, enhance contour accuracy and consistency, and encourage adherence to demarcation criteria. When applied to highly specialized medical imaging techniques like CT and X-Ray, Segmentation takes the information from the image's background and generates a mask. This can be done both with 2D and 3D data. There exist lots of tools to perform segmentation tasks, including manual and semi-automatic techniques.

The fully manual techniques include manually annotating the data. However, the manual way of segmenting medical images is a tedious and highly time-consuming task that becomes even more complicated with large databases [15].

Cardiologists primarily use MRI,which is challenging to segment due to its anisotropic nature with distant 2D slices. Hence, an automated approach is needed to perform this segmentation task. Recently, most state-of-art segmentation methods are based on DL approaches, which substantially improves the performance of previous methods. DL methods are broadly divided into two categories- the 2D methods segment each slice independently, whereas 3D methods segment multiple slices together as a volume. When working with 2D methods, as no 3D context is considered, it might be challenging to maintain 3D consistency between the segmentation of different slices. Still, 2D methods are popular as they are lightweight, and dividing the 3D stacks into multiple 2D images increases the number of training images.

In [16], the authors found that segmenting each slice separately was beneficial and presented a fully automatic framework for segmenting Cardiac MRI. The results from the proposed model were compared with different 2D and 3D methods, where it was found that 2D methods outperformed 3D methods in terms of both training time and accuracy. The mean dice score for Left Ventricle, Right Ventricle, and Myocardium was 0.914.In [17], the authors used U-Net-inspired architecture to segment cardiac structures. To combine the strength of 2D and 3D methods, Recurrent Neural Networks (RNN) was used to process all the slices in the same stack from the base to the apex. However, the correlation between the slices was low except for adjacent slices. Also, the prediction made on each slice did not depend on existing predictions. In [18], the authors proposed V-Net for volumetric medical image segmentation, a fully CNN able to segment 3D images directly. Data augmentation was also performed, and a special objective function was introduced which could deal with a strong imbalance between foreground and background voxels.

In [19], the authors have designed a novel DL method to perform cardiac image segmentation on a short-axis MRI image stack. A novel variant of U-Net is applied to segment the stacks iteratively [20]. The segmentation of a slice largely depends on the existing segmented mask of the previous slice, and therefore, 3D consistency is maintained. Dice Similarity Coefficient (DSC), Intersection Over Union (IoU) or Jaccard Index, and Average Hausdorff Distance (AHD) are some widely used evaluation metrics for segmentation tasks. The equations for the metrics are mentioned in Eq. (1) to Eq. (3), respectively. DSC is the measure of overlap between two areas, namely the predicted area and ground truth, as shown in Eq. (1).

$$D(a,b) = 2 * \frac{a \cap b}{a \cup b} \tag{1}$$

Here, a and b refer to two areas. The value varies from 0 (complete mismatch) to 1(perfect match). IoU is quite similar to DSC, defined as a union of the area between predicted segmentation and ground truth divided by the overlap between the two, as given in Eq. (2).

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

The surface distances between the expected and ground truth segmentations are measured by AHD given in Eq. (3).

$$d_{AHD}(A,B) = \frac{\frac{1}{A}\sum_{a\in A, b\in B} \min \mathrm{d}(a,b) + \frac{1}{B}\sum_{b\in B, a\in A} \min d(a,b)}{2} \tag{3}$$

The sum of all minimum distances between all points in point sets$A$ and B divided by the total number of points in A yields the directed average Hausdorff distance from point set A to point set B. The directed average Hausdorff distance from point A to point B and the directed average Hausdorff distance from point B to point A can be used to compute the average Hausdorff distance.

V. CONCLUSION

This work discusses the role of DL in Medical Image Processing. Medical Image Segmentation, Image classification for diseases, and detection are automated using DL. This work reviewed and discussed the development of CNNs used for different image-processing tasks over the years. U-Net was one of the best models for image segmentation, and many models have been developed that have used U-Net as a backbone. Medical Image data is mostly volumetric in nature, and processing 3D data is found to be computationally expensive. Hence, 3D data is converted to 2D to ease model training, and it is also seen that 2D methods deliver better results compared to 3D methods. The role of attention models in enhancing DL models is also discussed in this article. Using the attention modules on U-Net and ResNet improves the models, which can better predict and classify diseases. The use of attention modules with other models makes the model computationally expensive. In the future, techniques combined with attention can be developed that is lightweight without compromising much on accuracy.

REFERENCES

[1] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," J. big Data, vol. 8, no. 1, pp. 1–74, 2021.

[2] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review", Neural Comput., vol. 29, no. 9, pp. 2352–2449, 2017.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv Prepr. arXiv1409.1556, 2014.

[5] 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, pp. 234–241, 2015.

[6] 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, pp. 3431–3440, 2015.

[7] D. Bank, N. Koenigstein, and R. Giryes, “Autoencoders,” arXiv Prepr. arXiv2003.05991, 2020.
[8] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 6, no. 02, pp. 107–116, 1998.
[9] 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, pp. 770–778, 2016.
[10] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv Prepr. arXiv1704.04861, 2017.
[11] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in International conference on machine learning, pp. 6105–6114, 2019.
[12] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv Prepr. arXiv1409.0473, 2014.
[13] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, “Visual Attention Methods in Deep Learning: An In-Depth Survey,” arXiv Prepr. arXiv2204.07756, 2022.
[14] F. Wang et al., “Residual attention network for image classification,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164, 2017.
[15] H. McGrath et al., “Manual segmentation versus semi-automated segmentation for quantifying vestibular schwannoma volume on MRI,” Int. J. Comput. Assist. Radiol. Surg., vol. 15, no. 9, pp. 1445–1455, 2020.
[16] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, “An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation,” in International Workshop on Statistical Atlases and Computational Models of the Heart, pp. 111–119, 2017.
[17] R. P. K. Poudel, P. Lamata, and G. Montana, “Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation,” in Reconstruction, segmentation, and analysis of medical images, Springer, pp. 83–94, 2016.
[18] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in 2016 fourth international conference on 3D vision (3DV), pp. 565–571, 2016.
[19] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache, “3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation,” IEEE Trans. Med. Imaging, vol. 37, no. 9, pp. 2137–2148, 2018.
[20] Bose, K., Shubham, K., Tiwari, V. and Patel, K.S., 2022. “Insect Image Semantic Segmentation and Identification Using UNET and DeepLab V3+”. In ICT Infrastructure and Computing: Proceedings of ICT4SD, pp. 703-711, 2022.