Dual Wavelet Attention Networks for Brain Tumor Classification and Detection (MRI)

EE 678 - Wavelets

Group Members:

Mrudul Jambhulkar - 21d070044
Bhavik Yadav - 21d070090
Akhilesh Chauhan - 21d070010
Om Unhale - 210070058 (BTP student)

Contents

Abstract

There have been advancements in image processing tasks such as object detection, image classification due to Convolutional Neural Network (CNN) models. Current models employ the use of spatial domain relying on pixel values. Spectral analysis of images is an aspect where the model can be further improved upon fulfilling the objectives of noise frequency removal and identifying the unclear edges among others. This type of CNN architecture, combines the normal model with the multi-resolution analysis using the 2-D Wavelet Transform (WT) for feature extraction. 2D WT helps in capturing the approximate and detailed information of the image. Attention mechanisms are also used in CNN to learn critical information and reduce redundancy. Attention based models are know to provide better interpretability/ explainability and accuracy. In this work, a Dual Wavelet Attention network (DWAN) is proposed for brain tumor detection and classification. It consists of a wavelet channel attention and a wavelet spatial attention. [1] served as the primary source for most of the work in this study. .We used an augmented dataset from Kaggle for our experiments : Dataset .
The code used by us for the experiment is available here : Code

1 Introduction

The classification of brain tumors through MRI scan images becomes imperative since it aids in determining the right treatment mode for each tumor type. In this sense, all tumors are either classified 'benign' or 'malignant', the latter which occurs in greater than or equal to three different grades based on their aggressiveness, grade I being the least (or Pilocytic Astrocytoma) and grade IV being the most (or Glioblastoma). Although MRI has a high superiority in terms of tissue contrast and sensitivity, it is still relatively complex and poses trouble when analyzing manually. Several complications arise due to the great variety of tumors that exist with respect to their shape, size, position and intensity, thereby losing precision and necessitating automated and aggressive methodologies. It has been noticed that Convulutional Neural Networks (CNNs) help in addressing these issues by learning complex features from MRI images. Whereas in regular CNNs, they find it difficult to pay attention to specific or critical regions (especially small or low contrast tumors). Attention mechanism can be incorporated with CNNs, and

is particularly beneficial for medical imaging tasks, as it allows the network to emphasize significant spatial and channel features. Additionally, the 2D WT allows for inclusion of multi-resolution uses which facilitates in feature extraction of both approximate and detailed components of the image. Such combinations of Wavelet-CNNs with attention mechanism assist in achieving better accuracy of the model and its interpretability, as they bias on the key features of the model. In this work, we propose a Dual Wavelet Attention Network (DWAN) to detect and classify a brain tumor where the use of wavelet channel and spatial attention parallel will enhance performance to ensure faster and more accurate classification.

2 Attention Is All You Need!

RNNs , LSTMs and gated neural networks are widely used in language modeling and machine translation . Attention mechanisms are used to model the dependencies in the system despite their distances in input and output sequences . [2] highlights the transformer model which is based entirely on attention mechanism .

The Transformer model proposes a novel architecture specifically for sequence-to-sequence transduction tasks. It has largely abandoned recurrence and convolution in favor of attention mechanisms. Its architecture is basically split into two parts: an encoder and a decoder. Both components rely heavily on self-attention mechanisms, combined with feed-forward networks in stacked layers.This model has an encoder-decoder architecture. Again, both architectures consist of $N = 6$ identical layers. Such a structure is good at processing sequential data, again leveraging the power of attention mechanisms and feed-forward networks.

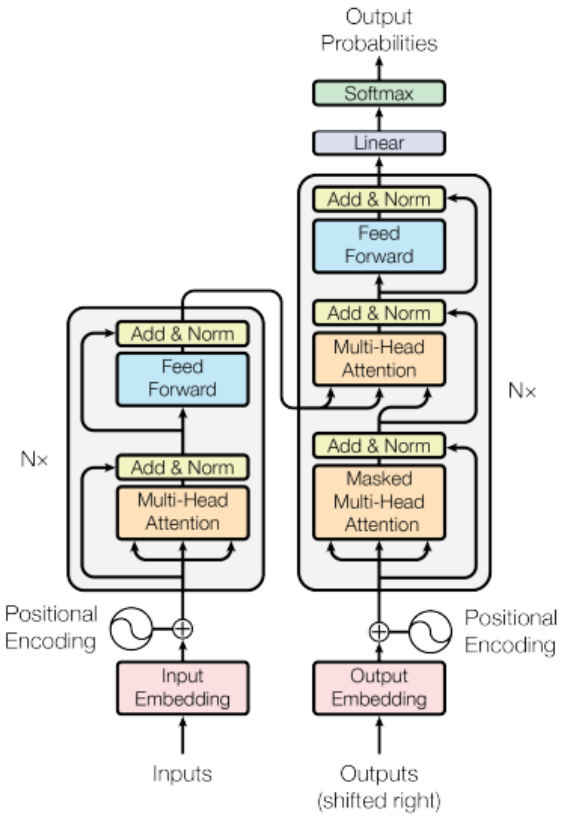

Figure 1: The Transformer - model architecture [2]

Left half is the encoder part and the right half is the decoder part in the above figure.

Encoder Stack: An encoder stack transforms an input sequence into a continuous representation. Each sub-layer within the layer consists of two components: a multi-head self-attention mechanism and a position-wise feed forward network (FFN). Self-attention enables a model to attend to information from different positions in the input sequence at once, therefore capturing both long and short range dependencies. The FFN adds more non-linear transformations to the model. Residual connections and layer normalization are applied around both sub-layers in order to stabilize training and ensure

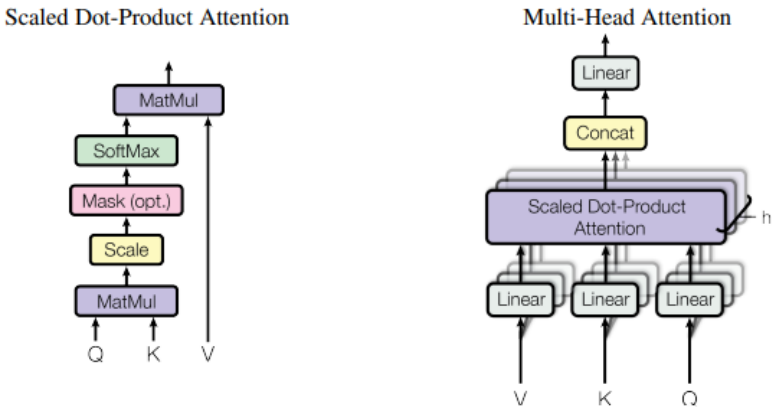

Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel [2]

that the gradients flow perfectly.

Decoder Stack: The decoder stack produces the output sequence from an encoded representation. In addition to the two sub-layers contained in the encoder, each of the decoder layers comprises a third sub-layer for encoder-decoder attention, which allows the decoder to attend to relevant parts of the output of the encoder. Self-attention in the decoder is also masked so that the predictions for any position depend only on the positions that have been seen so far, preserving the auto-regressive property. Just like the encoder, we apply residual connections and layer normalization to benefit training stability and convergence.

This architecture allows the Transformer to better model complex dependencies in sequences, making it quite pertinent for applications such as machine translation.

Attention: The attention function could be defined as a mapping, the input to which would be query and set of key-value pairs to get the output. All query, keys, values, and output are vectors in nature. The output is computed to be the weighted sum of the values, where the weight for each value is provided by a compatibility function, which will establish the level of relationship of the query with the respective key.

Scaled Dot-Product Attention: A scaled dot-product attention mechanism is one of the key constituents of Transformer architecture. It processes

three kinds of inputs: queries Q, keys K, and values V. The dot product of a query with each key gives an attention score that is scaled by the square root of the key dimension. Subsequently, a softmax operation is applied to result in attention weights:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Scaling the dot products by square root of the key dimension prevents them from getting too large, which may cause softmax to run into regions with terribly small gradients. It pushes the softmax into regions where learning becomes difficult. The softmax ensures attention weights sum up to 1; thus more focus could be given to relevant parts of the input sequence.

Multi-Head Attention: While scaled dot-product attention computes a single attention function, multi-head attention enhances that by projecting the queries, keys, and values into multiple subspaces using learned linear projections. Each head computes attention independently in its subspace, focusing on different parts of the input, and their outputs are concatenated and linearly transformed:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{ head }_1, \ldots, \text{ head }_\text{h}\right) W^O \\ \text{where head } &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{2}$$

Where the projections are parameter matrices $W_i^Q \in R^{d_{\text{model}} \times d_k}, W_i^K \in R^{d_{\text{model}} \times d_k}, W_i^V \in R^{d_{\text{model}} \times d_v}$ and $W^O \in R^{hd_v \times d_{\text{model}}}$.

Multi-head attention allows the model to jointly attend to information from different representation subspaces at each position, which helps the model learn diverse dependencies across the sequence. The Transformer strikes a balance between model complexity and computational cost by using h=8 heads.

Position-wise Feed-Forward Networks: Each block in both the encoder and decoder has a position-wise feed-forward network. That is essentially two linear transformations with a ReLU activation in between, and it does this entirely independently on each position:

$$\text{FFN}(x) = \max\left(0, xW_1 + b_1\right)W_2 + b_2 \tag{3}$$

Another way of describing this is as two convolutions with kernel size 1.These layers enable the model to transform its representations produced by attention in such a way as to improve its expressibility power.

Embeddings and SoftMax: It does use learned embeddings as input, since it translates input tokens into possibly very dense vector representations. These are shared between the encoder and decoder's input layers and the pre-softmax linear transformation in the output layer, scaled by the square root of model dimension.

At the output, then, a linear transformation followed by a softmax maps the decoder's outputs to probabilities over the target vocabulary.

Positional Encoding: Because the Transformer does not apply recurrence or convolution, it adds positional encodings to the input embeddings, which signify the order of an element in a sequence. The authors utilize fixed sinusoidal functions to encode such positions.

$$\begin{aligned} PE_{(pos,2i)} &= \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \\ PE_{(pos,2i+1)} &= \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \end{aligned} \tag{4}$$

$$\text{where pos is the position and } i \text{ is the dimension.} \tag{5}$$

These encodings enable the model to identify both absolute and relative positional dependencies.

WHY SELF ATTENTION: The central mechanism in the Transformer architecture is self-attention, with very strong advantages over recurrent and convolutional networks. Motivating our use of self-attention we consider three desiderata.

Computational Efficiency: Self-attention is parallelizable, so its time complexity during the training process is significantly reduced. RNNs typically read tokens sequentially, but in self-attention one can consider all tokens at the same time with $O(n^2d)$ complexity where n is the sequence length and d is the model dimension. It pays off for big sequences; recurrent models apply there terribly slow because of the sequential nature.

Parallelization: The self-attention mechanism also outperforms in parallelization. As you can see in the table,it has the constant number of sequential operations with respect to any sequence length, while RNN relies on the O (n) operations. It makes it possible for the Transformer to employ modern hardware such as GPUs much more powerfully, which enables faster training and inference.

Path Length for Dependencies: Learning long-range dependencies is challenging for sequence models. The self-attention mechanism minimizes the worst path length between any two words to O(1), which in turn makes it easier for the model to learn dependencies across distant positions. That is

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Figure 3: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention [2]

quite different from RNNs and CNNs for which the path length grows linearly or logarithmically with the length of the input sequence.

Interpretability: The attention weights in self-attention provide clear transparency on exactly how the model is processing information and which parts of the input the model pays attention to at the time of making a specific prediction. This interpretability can then be leveraged to understand the decision-making processes of the model much better.

3 Dual Wavelet Attention Networks

3.1 I. Related Works

3.1.1 Wavelet Transforms in Image Processing

Wavelet transforms have become increasingly popular in deep learning-based image processing because they can capture both spatial and frequency details. However, earlier methods often did not use attention mechanisms for tasks like image super-resolution, classification, inpainting, denoising, and restoration. Recent research has addressed this by combining wavelet transforms with attention techniques. For example, AWNet [4] combines non-local attention with Discrete Wavelet Transform (DWT) to improve image signal processing (ISP) for smartphone images. Similarly, WAEN [3] uses an attention embedding network and a wavelet embedding network to improve video super-resolution. A method in [5] integrates DWT and its inverse into an attention module to remove rain streaks from images. For image classification,

[6] combines features extracted by 2D DWT with those from convolutional neural networks to improve accuracy. Wavelet pooling, introduced in [7], uses wavelet decomposition to break an image into subbands. It keeps only the low-frequency subband, reducing feature size and avoiding overfitting better than max pooling. These developments show how wavelet transforms and attention mechanisms together can significantly enhance image processing techniques.

3.1.2 Attention Mechanisms in Deep Learning

The attention mechanism does not have a strict mathematical definition. Traditional methods like local image feature extraction, saliency detection, and sliding window-based techniques can also be considered forms of attention. Recently, the attention mechanism has become a key area of research in deep learning, often added to neural networks to enhance both theirperformance and interpretability. The concept of visual attention gained traction with the highway network [8], which introduced a gating mechanism to improve the flow of feature information in deep networks. Later, [9] proposed triplet attention, which models spatial and channel attention efficiently without reducing dimensionality. FcaNet [10] extended channel attention by including a multi-spectral component using frequency analysis, linking global average pooling (GAP) with the discrete cosine transform. These advancements highlight the growing importance of attention mechanisms in deep learning models for improved efficiency.

3.2 II. Dual Wavelet Attention Networks for Image Classification

These attention networks are designed based upon the dual wavelet attention block. Also, utilising the property of wavelet attentions being able to capture spatial features, we aim to incorporate spatial attention in FcaNet as well.

3.2.1 A. 2D DWT

We define the coefficients of the discrete cosine transform (DCT) as:

$$B_{mn} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos\left(\frac{\pi(2x+1)m}{2M}\right) \cos\left(\frac{\pi(2y+1)n}{2N}\right), \tag{6}$$

where $f(x,y)$ is an image of size $M \times N$ and $0 \leq x \leq M-1$, $0 \leq y \leq N-1$.

Next, the discrete wavelet transform (DWT) coefficients after the scale function can be generated by

$$W_\phi(j_0,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot \phi_{j_0,m,n}(x,y), \tag{7}$$

where $\phi_{j_0,m,n}(x,y)$ represents the scale functions.

The scale function of the orthogonal wavelet basis satisfies $\phi(x,y) = \phi(x)\phi(y)$. Thus,

$$W_\phi(j_0,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot \phi_{j_0,m}(x) \cdot \phi_{j_0,n}(y), \tag{8}$$

which is equivalent to the convolutional operations with the scale kernels.

Using these equations, we can represent the low frequency component of the 2D DWT using the Haar Wavelet basis.

$$\sum_{m,n \in \{0,1,2,\ldots,2^j-1\}} W_\phi(j_0,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) = \mathrm{GAP}(f(x,y))\sqrt{MN} \tag{9}$$

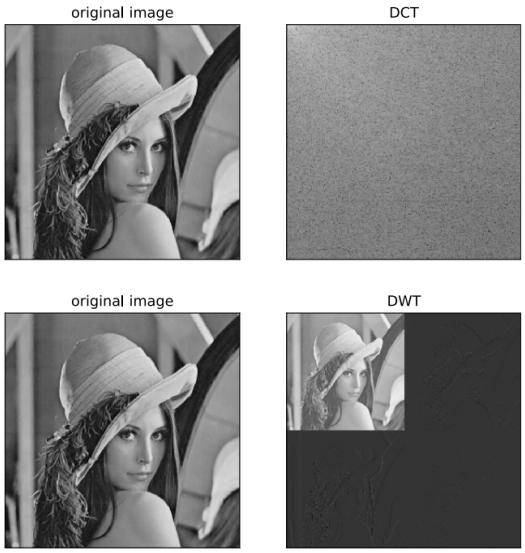

Figure 4: DCT vs DWT for an example image [1]

We can see the results of performing the DCT and DWT operations on Lena in Fig. 4. It's easy to see that DWT is better suited to maintain the spatial features from this figure due it's property of frequency localisation in both time and frequency.

3.2.2 B. Dual Wavelet Attention

The dual wavelet attention mechanism works as shown in Fig. 5. It aims to use both the high and low frequency components to construct the feature model. The final output feature map given by this mechanism is an element-wise multiplication of the input feature map with the wavelet channel attention map followed by another multiplication with the wavelet spatial attention map.

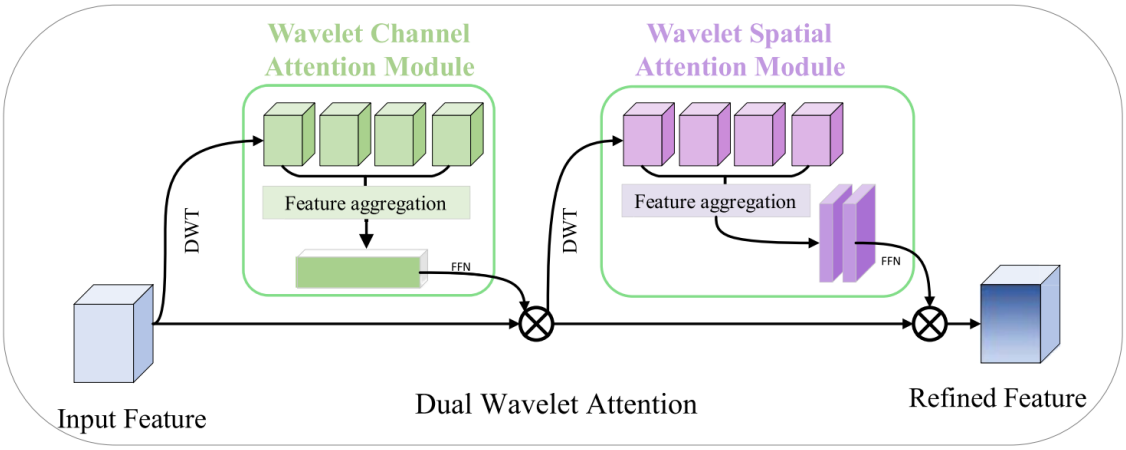

Figure 5: Dual Wavelet Attention Block Diagram [1]

The wavelet channel attention is an attention mechanism which aims to enhance the channel-wise feature representation in CNNs. It mainly uses the DWT to understand the high and low frequency components. Considering the 2D DWT can decompose the original input into the approximation (LL), horizontal detail (LH), vertical detail (HL), and diagonal detail (HH) sub-bands, we can aggregate their coefficients as scalars via element-wise summation and find their statistical feature as:

$$W_c = \sum_{i=0}^{M/2} \sum_{j=0}^{N/2} (LL + LH + HL + HH), \tag{10}$$

This is followed by implementing a simple feed-forward network (FFN) before the final output of wavelet channel attention block. This FFN consists of two fully connected layers in the form of a RELU activation function followed by a sigmoid activation function. The final representation of the wavelet channel attention map can be seen in Fig. 6.

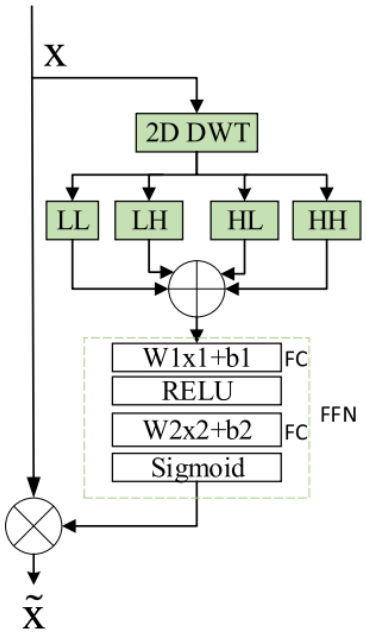

Figure 6: Wavelet Channel Attention Diagram [1]

To extract the spatial features of the image, the wavelet spatial attention mechanism is deployed. We can obtain the key image features and structure from wavelet decomposition. The mechanism works by concatenating the low-pass and the rest of the sub-bands aggregate of the 2D DWT. Thus, the aggregate feature map can be defined as:

$$W_s = CAT(LL, HL + LH + HH), \tag{11}$$

where CAT represents the concatenation operation. So, the final wavelet spatial attention map can be found by implementing the same FFN as earlier defined with the aggregate map being the input. The flow diagram is further representated in fig. 7.

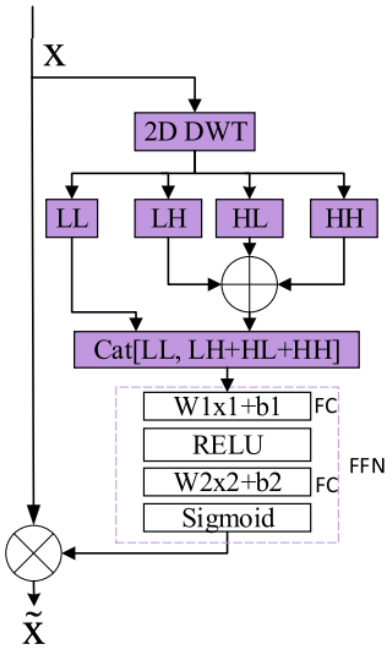

Figure 7: Wavelet Spatial Attention Diagram [1]

As compared with the earlier spatial attention mechanisms, by aggregating the low and high-frequency elements we can obtain more feature representations. It is also possible to generate the attention maps for different channels, which can better curve the spatial attention for each channel. Thus, by generating spatial representation for each channel of the input feature map, we obtain a better feature structure as shown in Fig. 8 further.

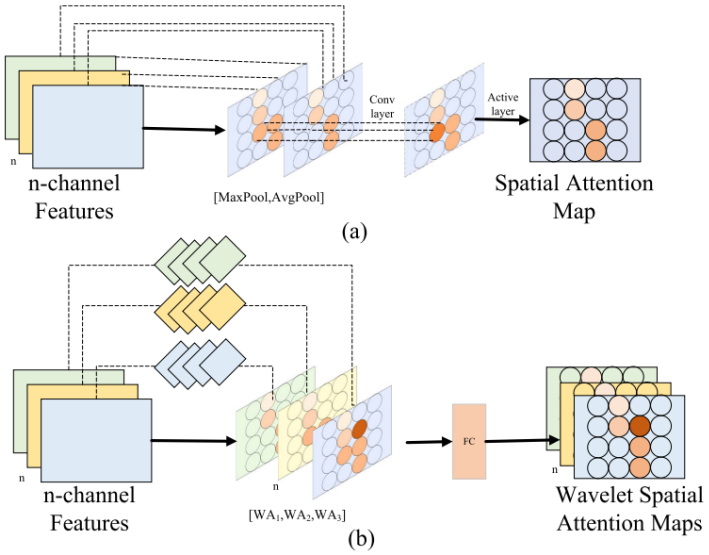

Figure 8: Comparison of Wavelet Spatial Attention mechanism with the earlier methods [1]

3.2.3 C. The Architecture of the Dual Wavelet Attention Networks

We can see the wavelet attention architecture combined with the CNNs (taking ResNet as an example) in Fig. 6. The figure shows how the proposed wavelet channel attention, wavelet spatial attention, and dual wavelet attention can all be embedded in different feature layers of the CNNs. The network construction of ResNet with the wavelet channel attention mechanism is shown in Fig. 9 (a) followed by implementations with the wavelet spatial attention and dual wavelet attention in Fig. 9 (b) and Fig. 9 (c) respectively.

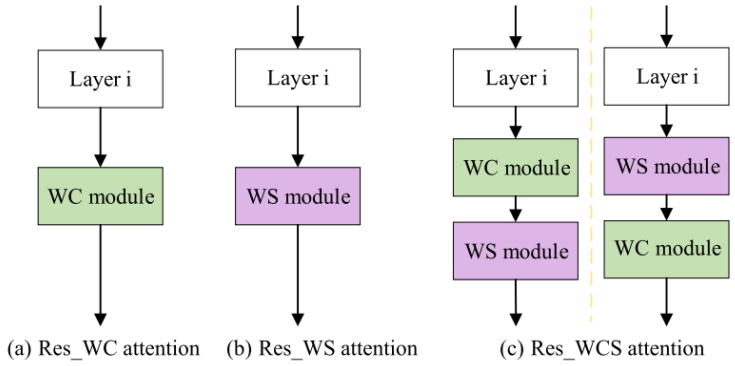

Figure 9: Example architecture of the wavelet attention mechanism incorporated with the CNN model [1]

3.3 III. Experiments

Dataset from https://github.com/sartajbhuvaji/brain-tumor-classification-dataset was used for the experiment . This dataset consists of MRI images of 3 different types of tumors and also images with no tumors . It comprises a total of - 926 images of glioma tumor , 937 images of meningioma tumor , 901 images of pituitary tumor and 500 images of MRI scans with no tumors including training and test data . The dataset was split into 80% for training and 20% for testing. Figure 10 shows sample images from these datasets .

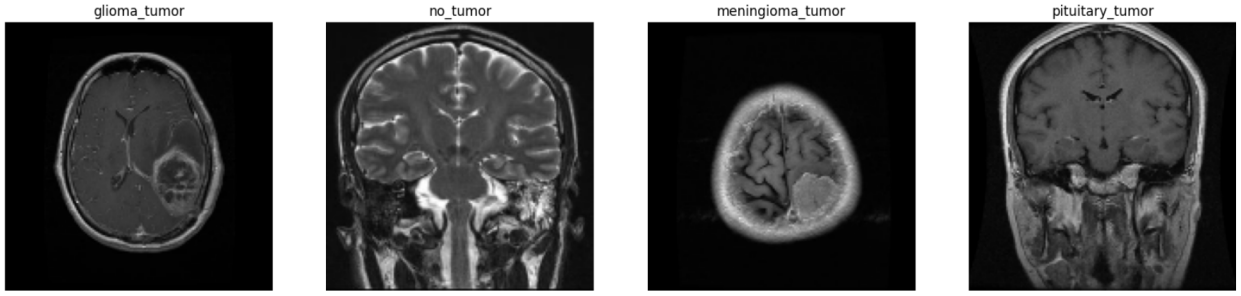

Figure 10: MRI images

ResNet18 was used as the base architecture . A wavelet channel attention module was followed by a wavelet spatial attention module after each ResNet

layers . This model is typically made for training on CIFAR datasets . So , the original MRI images (150x150) were resized to 32x32 resolution . The model was trained over 20 epochs with a learning rate of 0.001 . The batch size was set to 32 . Figure 11 shows the curves for training loss and accuracy vs the number of epochs . This model was then evaluated on testing data and a test accuracy of 90.05% was observed .

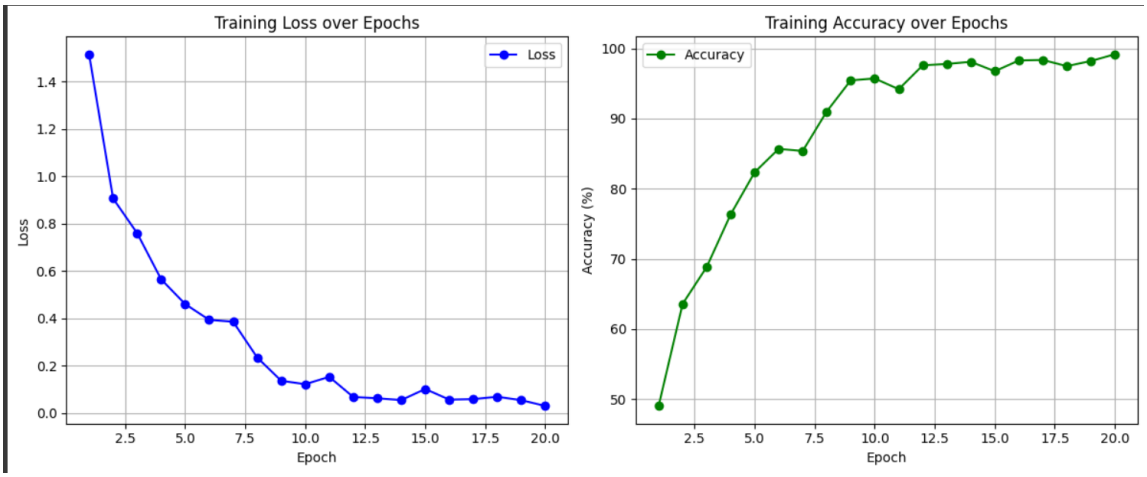

Figure 11: Loss and accuracy curves over different number of epochs

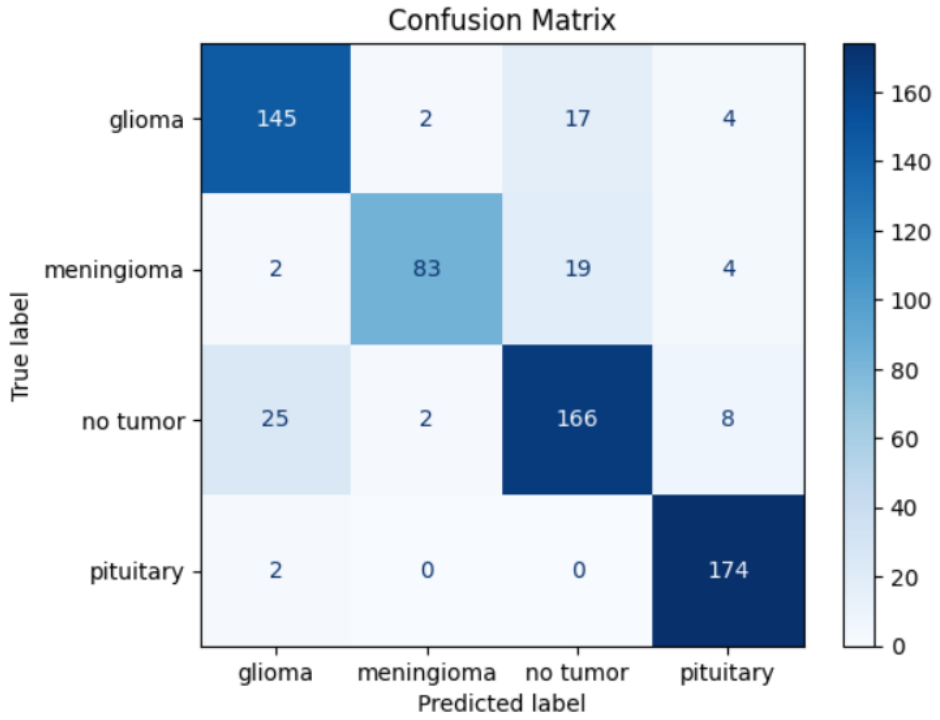

Figure 12: Confusion Matrix

Similarly , normal ResNet18 model (without any attention mechanism) was trained using this dataset with same hyperparameters : batch size of 32 and 0.001 learning rate , and an accuracy of 82.58% was obtained . Thus , the dual wavelet attention model provides superior performance .

4 Conclusion

Thus , experiments confirm that the DWAN offers better performance than standard models without attention in terms of accuracy and interpretability or explainability . Standard models are also prone to overfitting . Though attention mechanisms are used in some CNN architectures like SENet is a channel attention model which employs GAP but pooling often leads to loss of spatial information and can affect the performance of models in tasks like object detection , image classification and segmentation . Wavelet channel attention extracts features in wavelet domain and captures and aggregates

low and high frequency information from the image . Also , the spatial attention helps in preserving important spatial information from the image. Thus, DWAN offers better interpretability or explainability and is less prone to overfitting . The proposed model is very powerful for detection and classification of brain tumors . It can help in timely diagnosis and treatment of neurological disorders such as Alzheimer's disease (AD), schizophrenia, and dementia .

Similar , dual attention mechanisms can be used in other CNN architectures like VGGnet , AlexNet , ZFNet etc . DWAN can also be extended for the purpose of brain tumor segmentation , tumor localization etc .

Due to limited access to GPUs the model was trained over smaller number of epochs . Increasing the number of epochs and training on larger datasets can further improve the performance of the network . Also , the resizing of image may cause loss of some features , so the architecture can be modified such that the network can directly input the image of desired resolution to further improve the performance of model .

References and Supplementary Materials

[1] Y. Yang et al., "Dual Wavelet Attention Networks for Image Classification," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 4, pp. 1899-1910, April 2023, doi: 10.1109/TCSVT.2022.3218735.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[3] Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Wavelet attention embedding networks for video super-resolution," in 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7314–7320, 2021.

[4] L. Dai, X. Liu, C. Li, and J. Chen, "Awnet: Attentive wavelet network for image ISP," in ECCV Workshops, pp. 1–15, 2020.

[5] H.-H. Yang, C.-H. H. Yang, and Y.-C. F. Wang, "Wavelet channel attention module with a fusion network for single image deraining," in 2020 IEEE International Conference on Image Processing (ICIP), pp. 883–887, 2020.

[6] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," arXiv preprint arXiv:1805.08620, 2018.

[7] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in Proc. Int. Conf. Learn. Represent., pp. 1–12, 2018.

[8] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," arXiv preprint arXiv:1505.00387, 2015.

[9] H. Peng, X. Chen, and J. Zhao, "Residual pixel attention network for spectral reconstruction from RGB images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2020.

[10] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 783–792, October 2021.

[11] Brain Tumor Classification Dataset. Available: https://github.com/sartajbhuvaji/brain-tumor-classification-dataset

[12] Experimental Code on Google Colab. Available: https://colab.research.google.com/drive/1YYjcgjhpjic--m9lLjMXav6PI7AkX211?usp=sharing