International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Image Segmentation via Attention Module

Siddhartha Gupta , Shashank Shekhar Garg

Under guidance of Dr. Ankita Gupta

Computer Science Department, Maharaja Agrasen Institute of Technology

Abtract - Image segmentation is a crucial part of many systems for visual comprehension. Partitioning pictures (or video frames) into several parts or objects is necessary. Numerous applications, such as medical image analysis (such as tumour border extraction and measurement of tissue volumes), autonomous cars (such as navigable surface and pedestrian identification), video surveillance, and augmented reality, to name a few, all heavily rely on segmentation. Several approached have been used in the past, including Deep Convolutional Neural Networks. In this paper, we aim to achieve image segmentation of general everyday images via recently introduced attention modules, which have otherwise yielded optimal results in specific subset of biomedical images. We'll be employing an encoder-decoder architecture called U-Net. The main components of the encoder are a contracting path, also known as an encoder, which records the context of the image, and a symmetric expanding path, also known as a decoder, which enables precise location.

Keywords- Image Segmentation, Neural Network, Attention Module, U-Net

1. *INTRODUCTION*

The classification of pixels with semantic labels (semantic segmentation) or the division of distinct objects (instance segmentation) are two ways to formulate the problem of segmenting an image. Semantic segmentation performs pixel-level labelling for every image pixel using a set of object categories (such as human, car, tree, and sky), making it a more difficult task than image classification, which predicts a single label for the entire image. Instance segmentation expands the scope of semantic segmentation by identifying and separating each object of interest in the image (such as the division of distinct people).

Several techniques, including region growing, thresholding, watersheds, Otsu, k-means clustering, histogram-based clustering, graph cuts, and Markov random fields have been used in the literature for image segmentation. However, the majority of these earlier techniques segment objects using low-level features and cues.

Deep learning-based models have seen notable improvements in performance accuracy and time efficiency in recent years, leading to notable success. Numerous neural network-based techniques for object detection and classification have been developed by researchers.

A significant improvement in the performance of vision tasks has been made by convolutional neural networks (CNNs) due to their strong representational capabilities. Three critical network components—depth, width, and cardinality—have been the focus of recent research studies to improve the performance of CNNs.
The network has grown deeper to support rich representation from the LeNet architecture to Residual-style Networks thus far. VGGNet demonstrates that stacking blocks of the same shape yields reasonable results. ResNet builds an incredibly deep architecture in the same manner by stacking the same residual block topology and skipping connections.

But there are some issues that need to be resolved. Some of these flaws stem from the model architecture itself, while others are caused by the nature of the issue or field. A pedestrian may appear very small in the input image of the scene segmentation problem for self-driving cars, for instance, while the rest of the image is dominated by cars, trees, and the street. Another instance is when cars from the same class are close together and appear larger than the farther away cars in the image that was captured.

By using attention modules to force the architecture to concentrate on informative locations (objects of interest) in the input and ignore non-informative ones, researchers have attempted to solve these issues. In order to make the architecture of machine translation concentrate on a particular input word while producing a particular output word, attention networks first appeared in machine translation. In order to make the architecture concentrate on

a particular area of the input image while creating a specific output word, attention networks are also introduced for images in the image captioning problem.
In the field of medical imaging, attention modules have shown very promising results because they do away with the need for manual feature engineering and manual feature extraction.

Therefore, we will be aiming to apply an attention module , using a U-Net architecture, following an encoder-decoder approach , to a general , everyday life image dataset and thus answer the following research questions:

a) Can Attention Modules be used for image segmentation, other than medical images?

b) Will the applied attention module give the same (or better) results than the previously used techniques on a general image dataset?

2. *Attention Module Used*

Convolutional Block Attention Module (CBAM) is a novel strategy for enhancing CNN networks' capacity for representation. The authors of the original paper [4] claim that the CBAM mechanism combines two distinct attention blocks: the channel attention module, which is a feature detector and aims to produce a channel attention map by utilising the inter-channel relationship of features; and the spatial attention module, which is a complement to the channel attention module and aims to produce a spatial attention map by utilising the inter-spatial relationship of features (see Figure 2)

With an input of intermediate feature map $F \in R^{C\times H\times W}$, CBAM sequentially infers a 1D channel attention map $Mc \in R^{C\times 1\times 1}$ and a 2D spatial attention map $Ms \in R^{1\times H\times W}$. The complete attention process can be formulated as:

$$F' = M_c(F) \otimes F,$$
$$F'' = M_s(F') \otimes F', \quad (1)$$

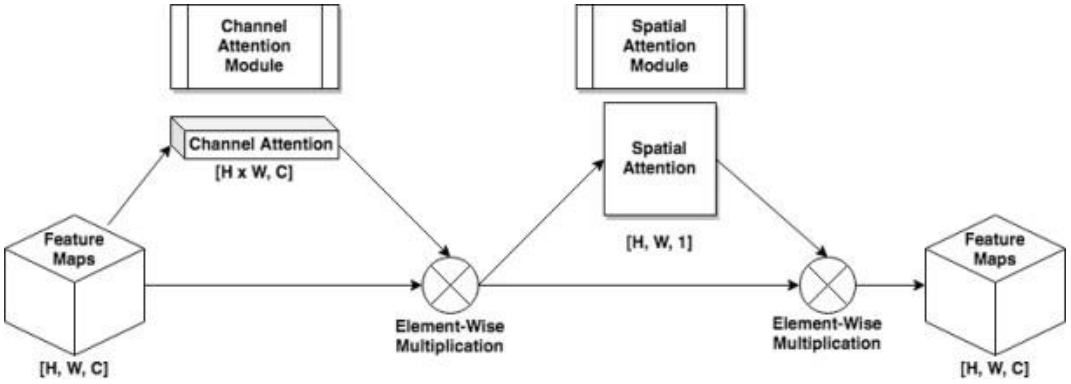


Figure 1: Block diagram of CBAM, according to Woo et al. [2018].

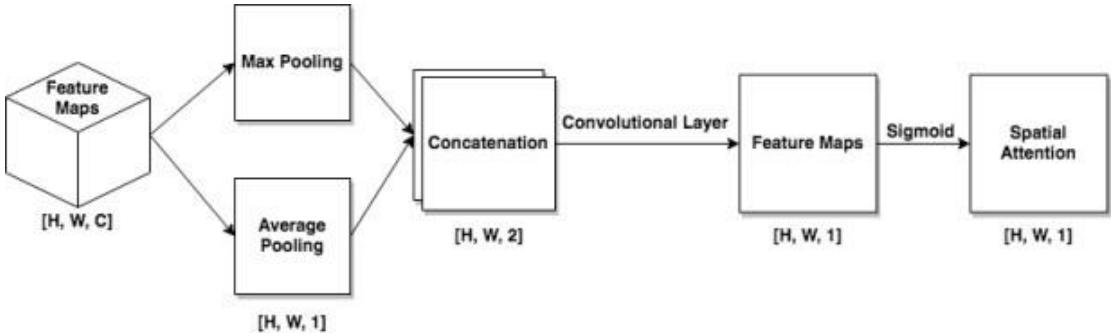


Figure 2: Block diagram of Spatial Attention CBAM, according to Woo et al. [2018].

3. *Related Works*

Attention mechanism

It is widely acknowledged that human perception depends heavily on attention. Human vision does not attempt to process an entire scene at once, which is an important characteristic of the human visual system. Instead, in order to better understand visual structure, humans make use of a series of partial glimpses and selectively focus on salient portions.

There have been a number of recent attempts to incorporate attention processing to enhance CNN performance in complex classification tasks. Residual Attention Network, which employs an encoder-decoder-style attention module, is the idea put forth by Wang et al. [2]. The network not only performs well but also is resistant to noisy inputs thanks to the feature map refinement.

We demonstrate that those features are inadequate for determining fine channel attention, and we recommend using max-pooled features as well. Additionally, they neglect spatial attention, which is crucial in choosing "where" to focus.

Chaudhari et al. [2021] [3] have proposed an insightful and thorough taxonomy for classifying attention mechanisms for language, text, and speech. The encoder-decoder model is how the authors explain the majority of the concepts. Therefore, in order to clarify, let's define the following: input patterns, the mode's input vectors

In light of this justification, Chaudhari et al. The various attention mechanisms are categorised by Chaudhari et al. [3] into several (non-exclusive) groups:

The number of representation/feature levels where the model will learn the attention weights is considered in the category of "number of abstraction levels." At this level, we can think about two different kinds of attention: single-level attention, where the attention weights are only calculated for the initial input sequence, and multi-level attention, where the attention mechanism is applied to several levels of abstraction of the initial input sequence, typically sequentially.

The number of input sequence positions where the attention weights are learned is taken into account in this category. At this level, the following attentional categories may be taken into account: To create the context vector, soft attention uses a weighted average of all the hidden states in the input sequence (i.e., the same as candidate states).

The number of distinct feature representations of the input sequence used in the learning task is taken into account in this category. At this level, the following attentional categories can be taken into account:
the most typical type of attention involves learning the attention weights using only a single feature representation, or single-representational attention; multi-representational attention, which entails learning the attention weights using several representations (of the same input sequence), and producing a context vector that is the result of a weighted combination of these various representations and their attention weights;

Sequence count: This category accounts for the quantity of input and output sequences. At this level, the following attentional categories may be taken into account: when the candidate and query states are associated with two distinct input and output sequences, respectively; Co-attention occurs when several input sequences are taken into account at once with the primary objective being to jointly determine the attention weights of each; When the candidate states and the query are from the same input sequence, self-attention occurs.

4. *Dataset*

A sizable object detection, segmentation, key-point detection, and captioning dataset is the MS COCO (Microsoft Common Objects in Context) dataset. There are 328K images in the dataset. In 2014, the MS COCO dataset's initial version was made public. There are 164K images total, divided into 83K training images, 41K validation images, and 41K test images. All of the previous test images as well as 40K new images were included in the 81K image additional test set that was released in 2015.

A portion of the COCO dataset that is one-fifth the size of the original dataset was used to train the model. The dataset is balanced and accurate and represents the entire dataset.

5. *Methodology*

Our architecture is based on U-Net and is supported by ResNet50 weights. A CBAM attention block has been added to each skip attention to help the decoder comprehend features better

As input to the model, a single RGB image with dimensions of 224*224*3 is provided. Encoder and decoder are the two components that make up the modified U-Net model.

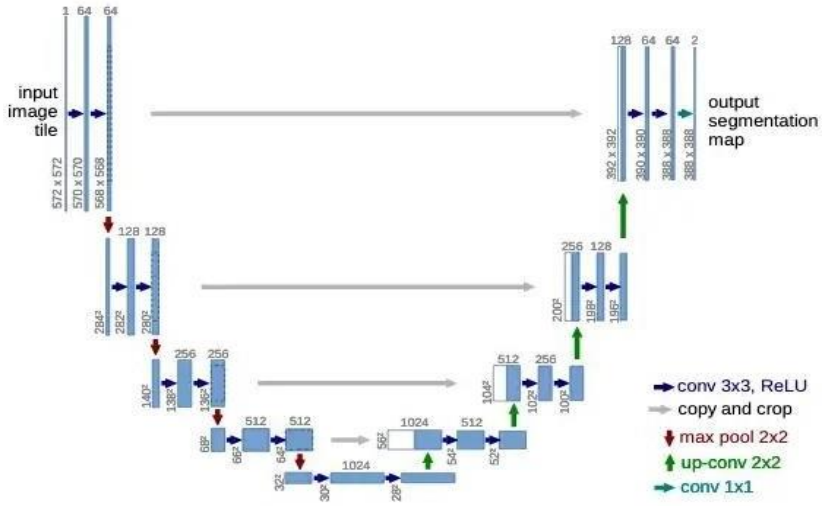


Figure 3: U-Net Architecture. (Ronneberger et al., 2015) [5].

The component of the model that utilises Resnet50's weights during training is the encoder. Its weights have been frozen due to a lack of computational resources. The image is down sampled and given more channels by the encoder component of the model. By down sampling the image and adding more channels to it, the encoder tries to extract both the high-level and low-level features of the image, thereby expanding the receptive field. As a result, when the model reaches the encoder's final stage, it is aware of both the image's high-level and low-level features.

Decoder: This component is in charge of up sampling the image to its original size. Half as many features channels are available, and the size of image is doubled.

Attention block: Attention blocks have been added to each skip connection in order to extract useful features from the encoder and use them for the decoder. In order to extract useful features from intermediate channel features produced by the encoder, the CBAM attention module uses both spatial and channel attention.

The Technology stack for the same includes: Python , KERAS ,TENSORFLOW , GPU(nvidia and cloud) , PYTORCH , KAGGLE Notebook and Pandas

6. Metrics for Performance Review

A model should ideally be assessed in a variety of ways, including speed (inference time), storage needs, and quantitative accuracy (memory footprint). The majority of research projects to date, though, have concentrated on metrics for assessing model accuracy. The two most common metrics for evaluating the precision of segmentation algorithms are summarised below.

One of the most frequently employed metrics in semantic segmentation is intersection over union (IoU), also known as the Jaccard Index. The area of intersection between the predicted segmentation map and the ground truth is what is used to define it, and it is calculated as the area of intersection between the two divided by the area of union between the two:

$$IoU = J(A.B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Where A and B stand for the ground truth and the predicted segmentation maps, respectively. It has a 0 to 1 range.

Another widely used metric is mean-IoU, which is the average IoU across all classes. It is frequently employed in reporting the effectiveness of contemporary segmentation algorithms.

Popular metrics for reporting the accuracy of many traditional image segmentation models

include precision, recall, and F1 score. The following definitions of precision are applicable to both the aggregate level and each class:

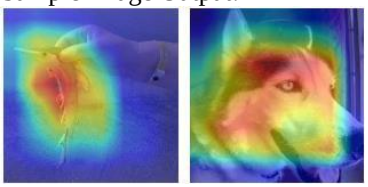
$$Precision = \frac{TP}{TP+FP} \quad (3)$$

7. *Result*

After running the above stated process over the MS-COCO Dataset, the resultant mIoU was **26.3** while the AP was **17.1**

Sample Image Input:

Sample Image Output:



8. *Conclusion*

We have successfully implemented an attention module using the U-Net architecture and ResNet-50 backbone on a huge image dataset. We have been able to apply CBAM over general, everyday life images (non-medical), achieving the first research motive successfully. While the results at not at par with the previously used models for image segmentation on the same MS COCO dataset, yet this opens up a lot of opportunity to dive further and not just perfect this model , but also explore other plausible applications and manifestations of the same. This shall be another step on the journey of making a human smart artificial intelligence system, perfectly able to achieve the goal of Computer vision.

9. Reference

[1] Huang, Y.; Xu, H. Fully convolutional network with attention modules for semantic segmentation. 2021
[2] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X.,Tang, X.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017)
[3] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An attentive survey of attention models. ACM Trans. Intell. Syst. Technol., 12(5), 10 2021. ISSN 2157-6904.
[4] Sanghyun Woo, Jongchan Park, Joon-Young Lee3 and In So Kweon. CBAM: Convolutional Block Attention Module, Adobe Research, San Jose, CA, USA
[5] Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation
[6] S. Minaee and Y. Wang, “An admm approach to masked signal decomposition using subspace representation,” IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3192–3204, 2019.
[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017
[8] Xiao Yang. An overview of the attention mechanisms in computer vision. In Journal of Physics: Conference Series, volume 1693, page 012173. IOP Publishing, 2020
[9] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, “Gather-excite: Exploiting feature context in convolutional neural networks,” in Conference on Neural Information Processing Systems, 2018
[10] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, “Exploring the limits of weakly supervised pretraining,” in ECCV, 2018