Chella Thiyagarajan N - ME17B179

Tarun - ME17B114

Aman Arora - ED17B031

CS6910 – Project

Fundamentals of Deep Learning

Contents

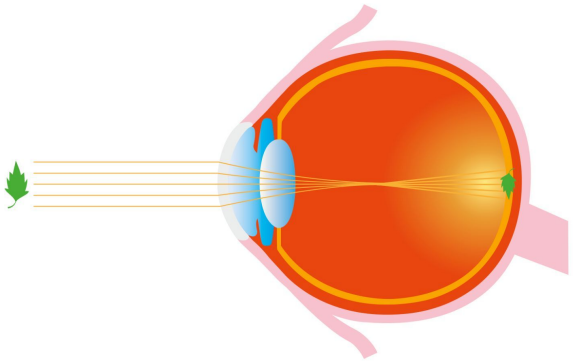
CBAM: Convolutional Block Attention Module

1. Abstract

Computer Vision has aimed to imitate human visual perception in terms of code-based algorithms. Human eye in coordination with the brain has an ability to aid concentrating our vision on a certain object and blur the surroundings. Light is transferred from the object of interest to the primary functional region of retina. We aim to apply a similar approach to our computer vision based models and give importance to essential features, by using Convolutional Block Attention Module (CBAM). CBAM contains two sequential sub-modules called the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), which can be applied in series or in parallel, and obtain a refined feature map.

2. Introduction

Modern-day techniques used in Neural Networks in the field of Computer Vision include "Attention Mechanisms." Before understanding the Attention Mechanisms used in Computer Vision, let's look at how attention mechanisms are inspired by human visual capabilities.
In the diagram below, the light traverses from the object of interest, the leaf, to the Macula, which is the predominant functional region of the Retina inside the human eye.



When our eye has to focus on a single object among a line-up of distinct objects in our scope of vision, the attention mechanism in our visual perception system uses a complicated group of filters to cause a blurring effect, similar to that in digital photographs. So, that the object of interest is in focus, while the surrounding is blurred or faded.

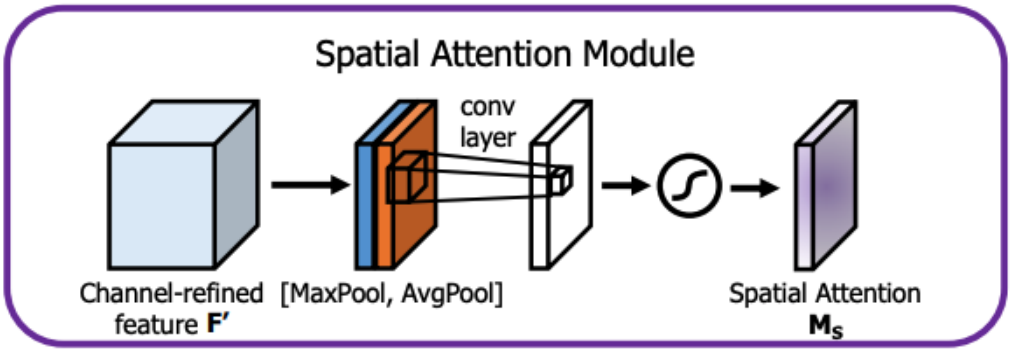
The Attention mechanism in Deep Learning is based on the concept of directing your focus and giving greater attention to certain factors when processing the data. While training computer vision based models, it is beneficial to give greater attention to useful features. This will aid us to improve accuracy of our task and also reduce the computation cost that is being used on training non-essential features. Attention modules makes CNN learn and

focus on the important information rather than learning non-essential background information. Considering the case of object detection, here useful information is the objects and target class crop that we are interested in classifying and localizing in an image.

We can have two kinds of attention module for image data:

1. Spatial Attention
2. Channel Attention

Spatial attention represents the attention mechanism on the feature map. In an image Spatial Attention will generate a mask which will enhance the features of the object of interest. Hence, refining the feature maps with the help of Spatial Attention, we enhance the input to the succeeding convolutional layers which thus improves the performance of the model. Spatial attention module consists of a simple 2D-convolutional layer.



Spatial Attention Map

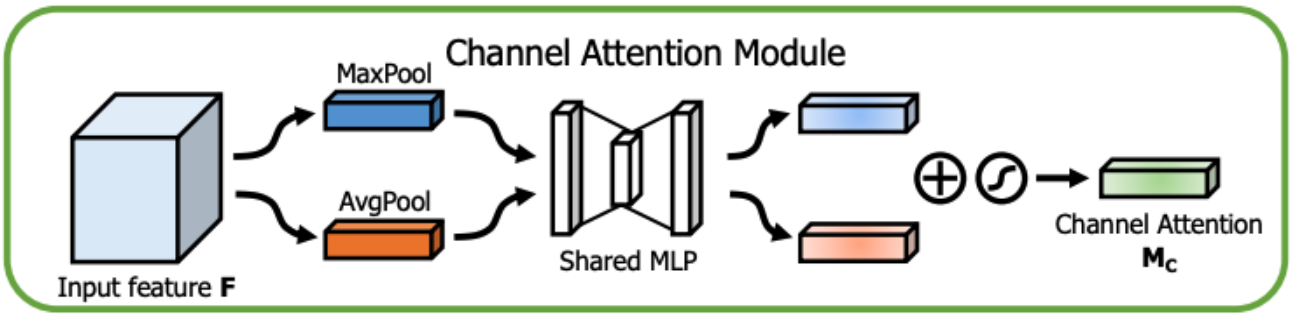
➔ Take the input feature map F and generate two intermediate feature maps. Fs(avg), and Fs(max).

➔ Concatenate these two outputs, Global Average Pool and Max Pooling, and pass it through a small convolutional block of 7x7 kernel size. We use large kernel sizes.

$$\mathbf{M_s}(\mathbf{F}) = \sigma(f^{7\times7}([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})]))$$

Channels are basically the feature maps stacked in a tensor, where each cross-sectional slice is basically a feature map. Generally, in convolutional layers, the trainable weights making up the filters generally learn small values (close to zero). Thus, we see similar feature maps, with many appearing to be copies of one another. Even though they look alike, these filters are essential for learning different kinds of features. While some are specific to learn horizontal and vertical edges, while others are more generic and learn a certain texture in the image. The channel attention essentially assigns weightage to each of the channels and thus enhances those particular channels which have a greater contribution towards learning and thus boosts the overall model performance. Spatial attention module consists of multilayer perceptron, and at the end, we add a sigmoid

function to get a mask of the input feature map.



Channel Attention Map

Channel Attention Map follows a similar generation process, but here along with Average Pooling, Max Pooling is also added to get a better distinctive channel features.

$$\mathbf{M_c}(\mathbf{F}) = \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F})))$$

3. Paper explanation

Convolutional Block Attention Module (CBAM) is applied at every convolutional block in deep networks to get subsequent "Refined Feature Maps" from the "Input Intermediate Feature Maps." CBAM contains two sequential sub-modules called the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), which can be implemented in Parallel or Series:

Parallel

Once we finish computing the channel attenuation map, Mc(F) and spatial attention map Ms(F) for our input feature space F.
To design an efficient module, we compute the channel attention Mc(F) and the spatial attention Ms(F), at two separate branches, then compute the attention map M(F). Then, apply a sigmoid function. The outputs from both the branches are resized to C×H×W before addition.

$$\mathbf{M}(\mathbf{F}) = \sigma(\mathbf{M_c}(\mathbf{F}) + \mathbf{M_s}(\mathbf{F}))$$

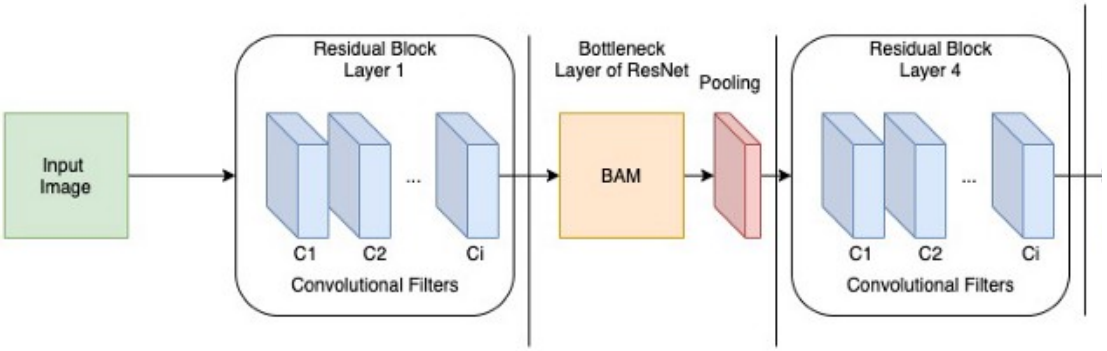
Mc and Ms are added element-wise to get the final attention map. We chose addition over multiplication due to the smooth gradient flow at the time of backpropagation.
For the given input feature map F. For both channel and spatial attention, the refined feature map F' is computed as

$$\mathbf{F'} = \mathbf{F} + \mathbf{F} \otimes \mathbf{M}(\mathbf{F})$$

Here ⊗ denotes element-wise multiplication.

We adopt a residual learning scheme along with the attention mechanism to facilitate the gradient flow. So after multiplying the attention mask, we again add the output with the input tensor F.

The parallel attention module is placed at every bottleneck of ResNet architecture. The attention module denoises low-level features such as background texture features at the initial stages and then eventually focuses on the exact target, which is a high-level semanticity.



Series

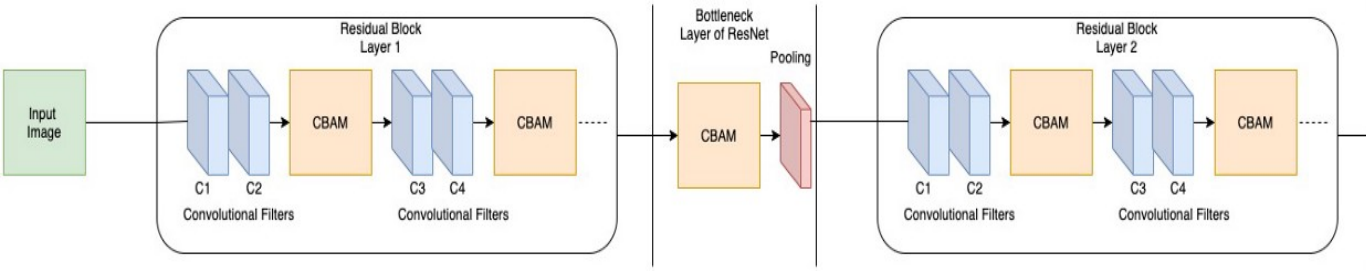
Given an intermediate feature map F as input, CBAM sequentially infers a 1D channel attention map Mc and a 2D spatial attention map Ms. The overall attention process can be summarized as,

$$\mathbf{F}' = \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F},$$
$$\mathbf{F}'' = \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}'$$

Here ⊗ denotes element-wise multiplication.

During multiplication, the attention values are broadcasted accordingly: channel attention values are broadcasted along the spatial dimension, and vice versa. F" is the final refined output.

Unlike parallel attention modules, series attention modules are placed inside residual blocks as well as at the bottlenecks.



First, the Channel attention map is implemented on the input feature map, followed by the application of the spatial map. Finally, F" is added to the previous input convolutional layer.

Difference between series and parallel attention modules

In the case of parallel attention modules, only Global Average Pooling was applied to obtain the statistics of the feature map in both spatial and channel dimensions. The series attention module also considered using MaxPooling with Average Pool. They proved that using Max Pooling accounts to generate the most salient features from the feature map and compensate for the global average pooling output, which encodes the global statistics softly.
In parallel attention, the convolutional operation was done using dilation value to increase the receptive field as we go deep in the network. Whereas, the series attention module used the larger kernel size of 7x7 and normal convolutional layer to incorporate the same.

In the parallel attention module, parallel generation of spatial and channel attention maps was considered, which was later added to obtain the final attention map. Whereas in the case of the series attention module, a sequential approach was used. Firstly, the channel attention map was computed, and then the Spatial map was finally derived from the generated intermediate feature map. The order of the sequential arrangement in the case of CBAM is,
Channel Attention → Spatial Attention

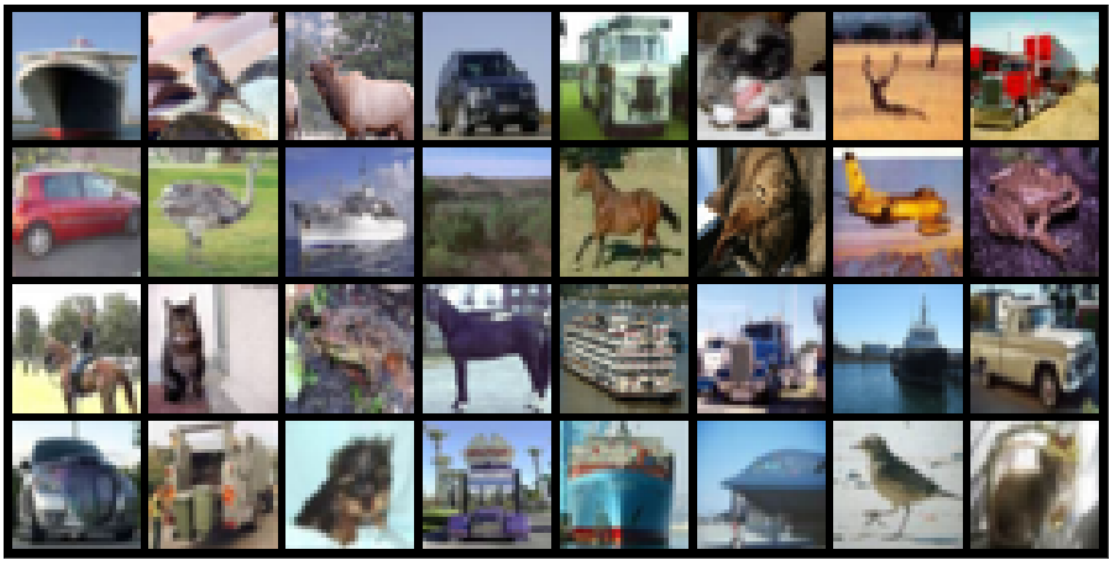
4. Experiments

Ablation Study

Ablation studies the performance of our model by removing certain components to understand the contribution of the component to the overall system. We apply Occlusion sensitivity test on CNN models without attention modules, models with spatial attention, models with channel attention, their series combination, their parallel combination. Looking at the performance, we aim to interpret which of the models gives better results and is robust.

We have used the CIFAR10 dataset for our purpose. The CIFAR-10 dataset contains 60000 32x32 color images in 10 classes, with 6000 images belonging to each class. It is split into 50000 training images and 10000 test images.

The dataset is split into one test batch and five training batches, each of which consists of 10000 images. The test batch contains exactly 1000 randomly-chosen images belonging to each class. The training batches have the remaining images in a random order, but some of the training batches may have more images from one class than the other. Between them, the training batches have 5000 images per class.



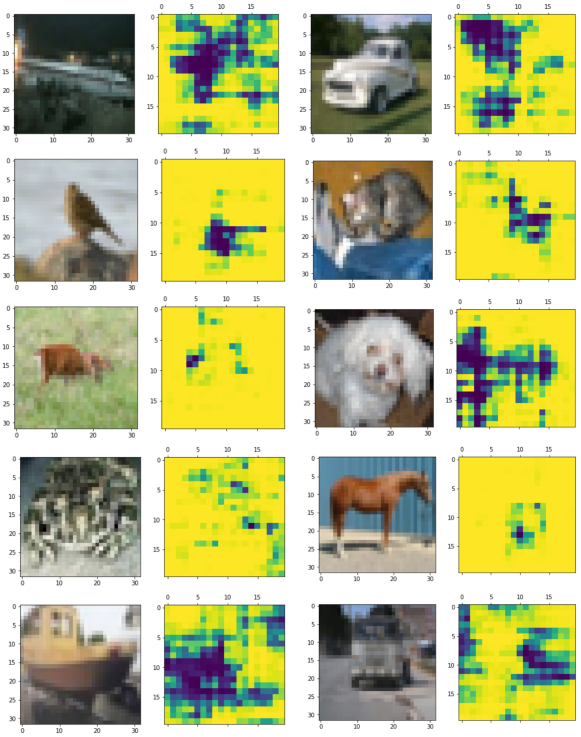
1.(a) Visualization of a random training set batch

We have trained our own CNN model with three convolution layers and without attention. The architecture of the model is as shown below.

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 32, 32, 32]	896
BatchNorm2d-2	[-1, 32, 32, 32]	64
ReLU-3	[-1, 32, 32, 32]	0
MaxPool2d-4	[[-1, 32, 16, 16], [-1, 32, 16, 16]]	
Conv2d-5	[-1, 64, 16, 16]	18,496
BatchNorm2d-6	[-1, 64, 16, 16]	128
ReLU-7	[-1, 64, 16, 16]	0
MaxPool2d-8	[[-1, 64, 8, 8], [-1, 64, 8, 8]]	
Conv2d-9	[-1, 128, 8, 8]	73,856
BatchNorm2d-10	[-1, 128, 8, 8]	256
ReLU-11	[-1, 128, 8, 8]	0
MaxPool2d-12	[[-1, 128, 4, 4], [-1, 128, 4, 4]]	
Linear-13	[-1, 1024]	2,098,176
ReLU-14	[-1, 1024]	0
Linear-15	[-1, 10]	10,250
Softmax-16	[-1, 10]	0

1.(b) Model Architecture

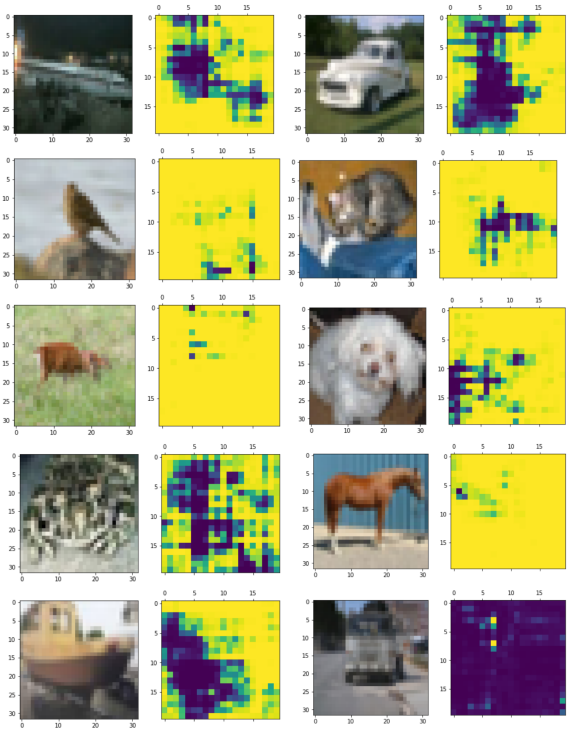
Occlusion sensitivity test was used to evaluate performance on this model in each of the classes, without attention module. Occlusion sensitivity is a simple technique to analyze which parts of an image are most important for the neural network. You can measure a network's sensitivity to occlusion in different regions of the data using small perturbations of the data.



2.(a) Occlusion sensitivity test on model without Attention modules

Model with only Channel attention

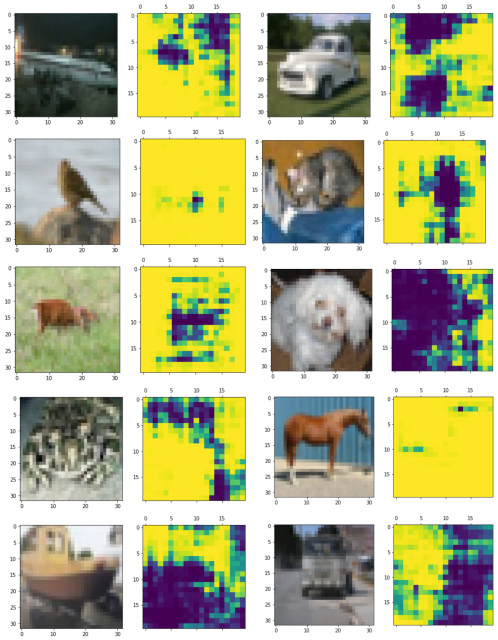
We have incorporated the channel attention module into our CNN model and trained our model on the same CIFAR10 dataset. To analyze and understand the effects of channel attention on our model performance. We have performed occlusion sensitivity test on this new model.



2.(b) Occlusion sensitivity test on model with only Channel Attention module

Model with only Spatial attention

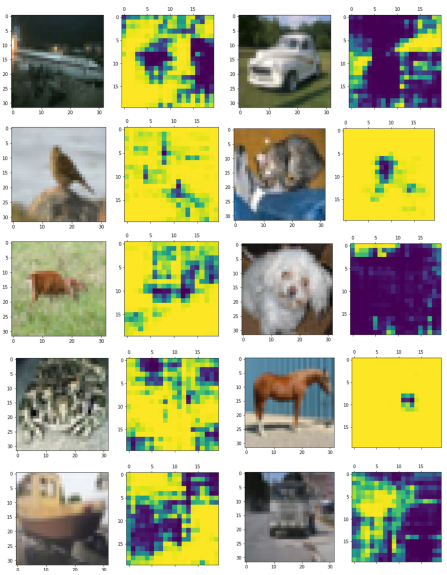
We have incorporated the spatial attention module into our CNN model and trained our model on the same CIFAR10 dataset. To analyze and understand the effects of spatial attention on our model performance. We have performed occlusion sensitivity test on this new model.



2.(c) Occlusion sensitivity test on model with only Spatial Attention module

Channel and Spatial attention in Parallel

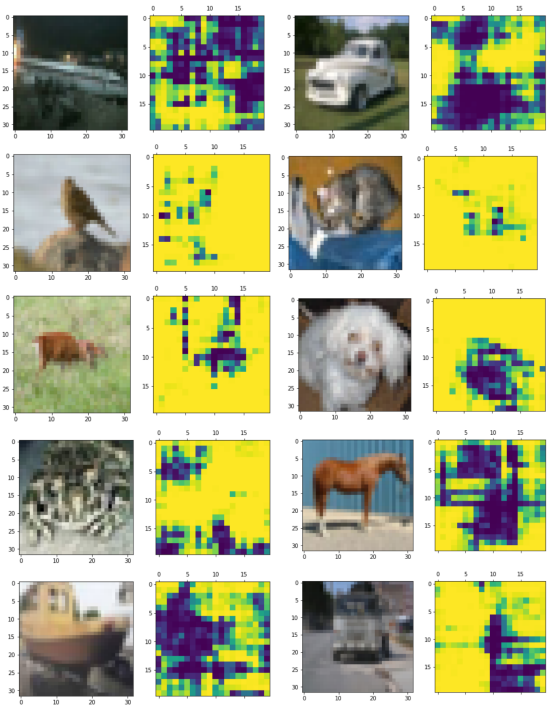
We have incorporated channel and spatial attention modules in a parallel architecture into our CNN model and trained our model on the same CIFAR10 dataset. To analyze and understand the effects of the two attention modules in parallel on our model performance. We have performed occlusion sensitivity test on this new mod



2.(d) Occlusion sensitivity test with Channel and Spatial Attention modules in Parallel

Channel and Spatial attention in Series

We have incorporated channel and spatial attention modules in a series architecture into our CNN model. With sequential arrangement the modules, first channel attention and then spatial attention. Then, we trained our model on the same CIFAR10 dataset. To analyze and understand the effects of the two attention modules in parallel on our model performance. We have performed occlusion sensitivity test on this new model.



2.(d) Occlusion sensitivity test with Channel and Spatial Attention modules in Series

Memory usage

The attention modules are memory and computationally efficient. The difference in the memory usage and computational cost compared to a model without attention modules is substantially low.

Without Attention

```
----------------------------------------------------------------
Input size (MB): 0.01
Forward/backward pass size (MB): 670.67
Params size (MB): 8.40
Estimated Total Size (MB): 679.08
----------------------------------------------------------------
```

Series Attention Modules

```
----------------------------------------------------------------
Input size (MB): 0.01
Forward/backward pass size (MB): 669.32
Params size (MB): 8.42
Estimated Total Size (MB): 677.76
----------------------------------------------------------------
```

Series Attention on ResNet

From the Ablation study performed above, we had concluded that the model with Channel and Spatial Attention modules in series performed better than the rest. To get a better understanding of this model, we have implemented series attention module on ResNet architecture.

We have trained the ResNet model on the CIFAR10 and CIFAR100 datasets, with and without using Convolutional Block Attention Module.

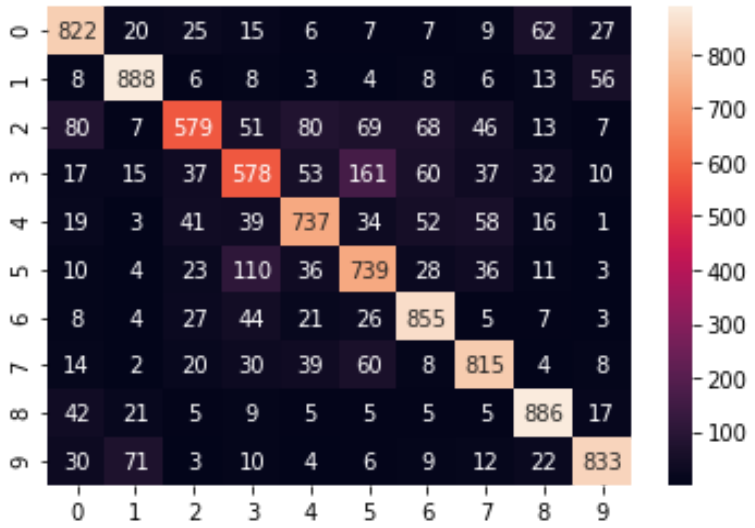
The CIFAR100 dataset is similar to CIFAR-10, except that it has 100 classes containing 600 images each. It is split into 500 training images and 100 testing images for each of the classes. The 100 classes in the CIFAR-100 dataset are grouped into 20 super-classes. Each image has a "fine" label, the class to which it belongs, and a "coarse" label, the super-class to which it belongs.

This will aid us in studying and understand the effect caused by the attention module on a better and well-proven architecture. Residual Neural Network (ResNet) is a neural network of a kind that builds on constructs known from pyramidal cells in the cerebral cortex. Residual neural networks do this by utilizing skip connections or shortcuts to jump over some layers.

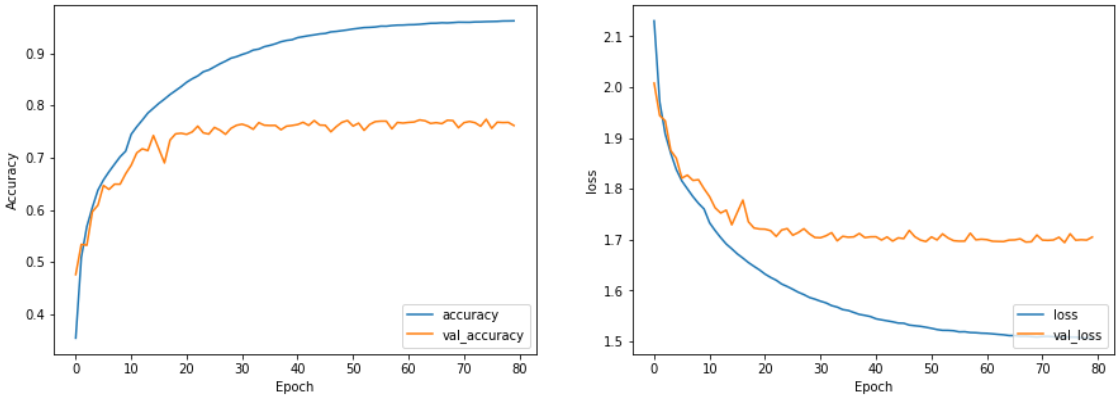
5. Results

Ablation Study

Model without Attention Modules

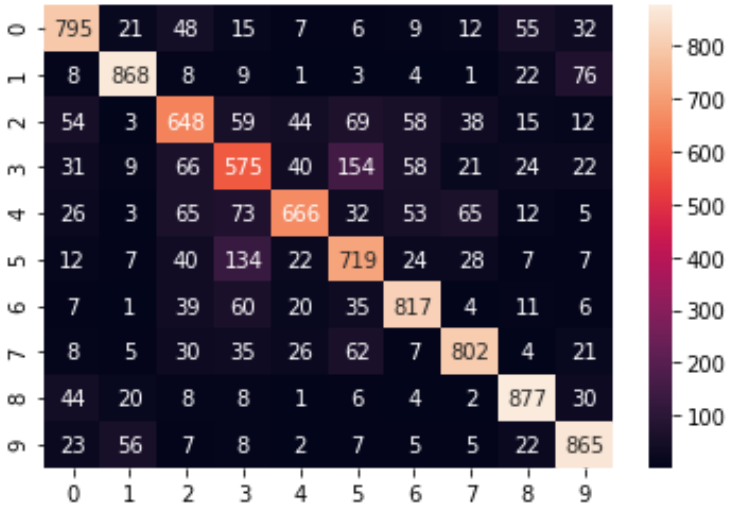


3.(a) Confusion Matrix of model performance on CIFAR10 dataset, without attention modules

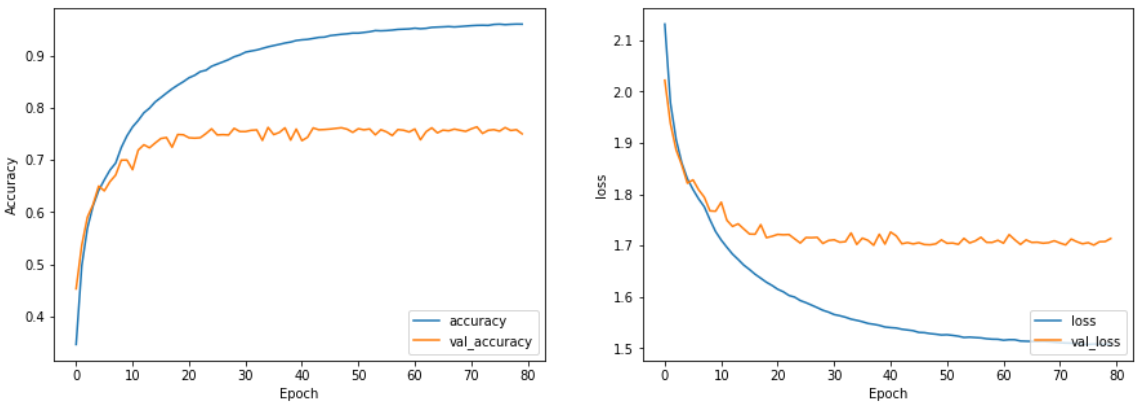


4.(a) Loss and Accuracy plots, CIFAR10 dataset, without attention modules

Model with CBAM



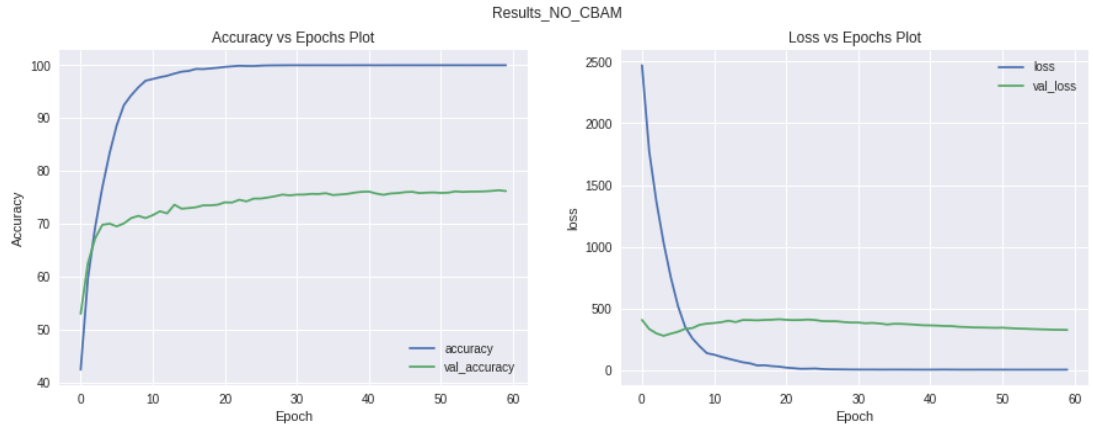
3.(b) Confusion Matrix of model performance on CIFAR10 dataset, with CBAM



4.(b) Loss and Accuracy plots, CIFAR10 dataset, with CBAM

Series Attention on ResNet

Without Attention Modules

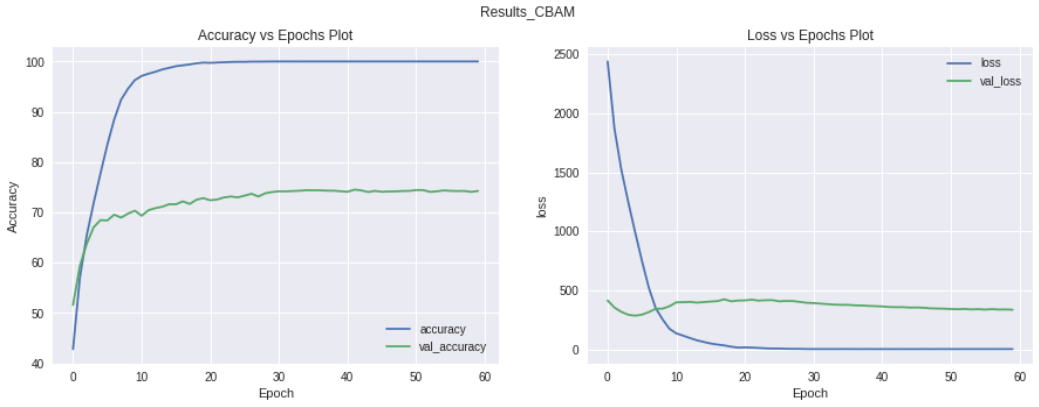


5.(a) Accuracy and Loss plots of ResNet model without CBAM

```
Test Set Classwise Accuracy after Training
{'bird': 66.4,
 'car': 88.4,
 'cat': 54.2,
 'deer': 72.6,
 'dog': 64.2,
 'frog': 84.0,
 'horse': 78.8,
 'plane': 81.89999999999999,
 'ship': 86.8,
 'truck': 84.6}
```

6.(a) Class-wise Test accuracy of ResNet model without CBAM, on CIFAR10 dataset

CIFAR10 dataset with CBAM

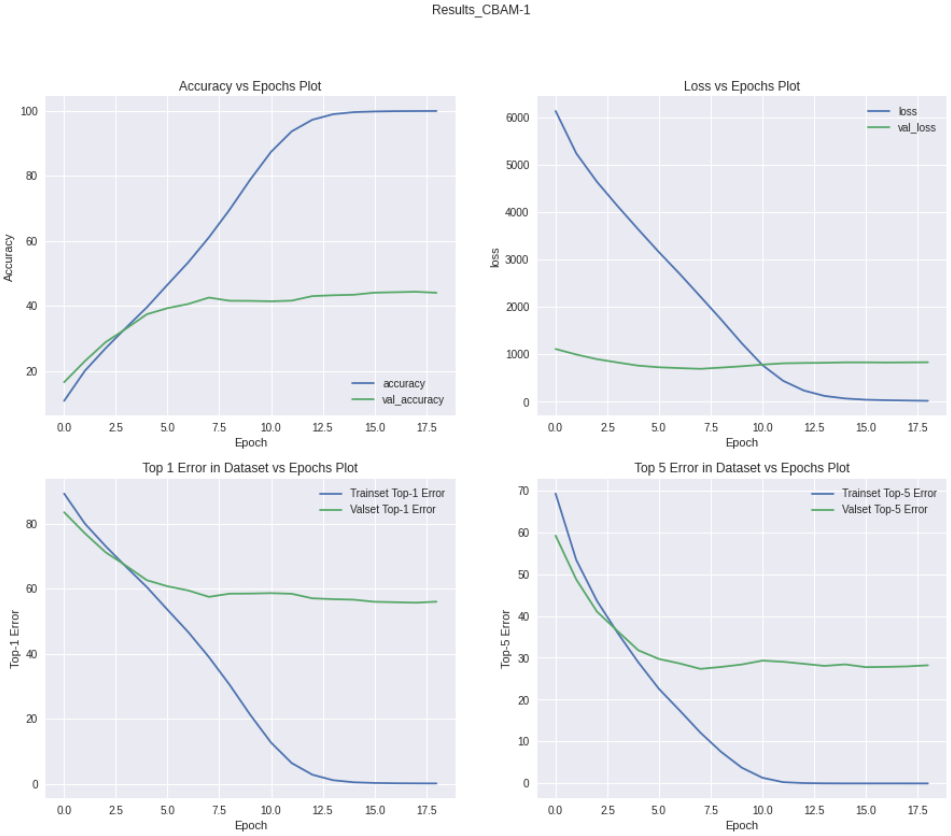


5.(a) Accuracy and Loss plots of ResNet model with CBAM, on CIFAR10 dataset

```
Test Set Classwise Accuracy after Training
{'bird': 61.199999999999996,
 'car': 88.1,
 'cat': 53.900000000000006,
 'deer': 68.0,
 'dog': 64.8,
 'frog': 82.69999999999999,
 'horse': 77.3,
 'plane': 77.60000000000001,
 'ship': 85.2,
 'truck': 83.8}
```

6.(b) Class-wise Test accuracy of ResNet model withCBAM, on CIFAR10 dataset

CIFAR100 dataset with CBAM



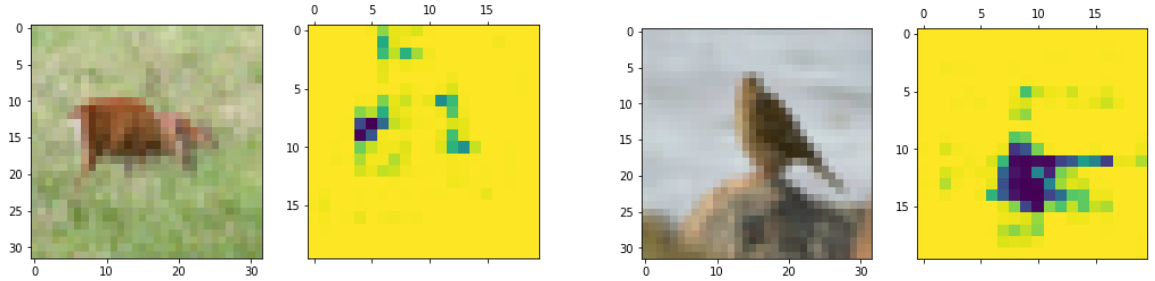
6. Observations

To get precise observations of the difference caused by Attention modules, we will consider images that will have a maximum impact due to spatial attention and channel attention, respectively.

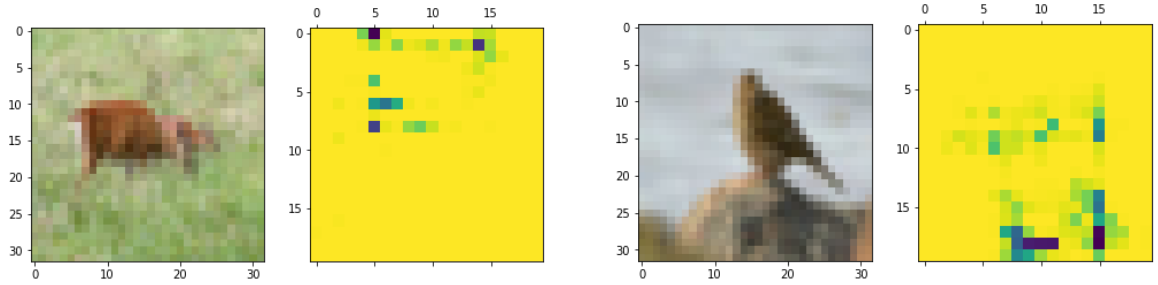
There will be a maximum impact due to spatial attention on images in which the object of interest is small. Hence, spatial attention will have a more significant effect to obtain a refined feature map.

Channel attention will cause a maximum impact on images that have greater variation in channels. Each of the image channels has a vast difference, images that have fewer regions of white, black, and gray. Channel attention will have a more significant impact on colorful images.
As shown in two of the chosen example images:

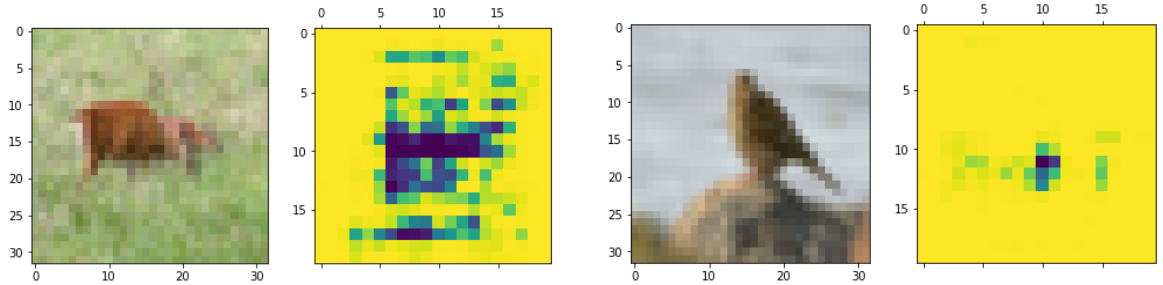
Without Attention Module



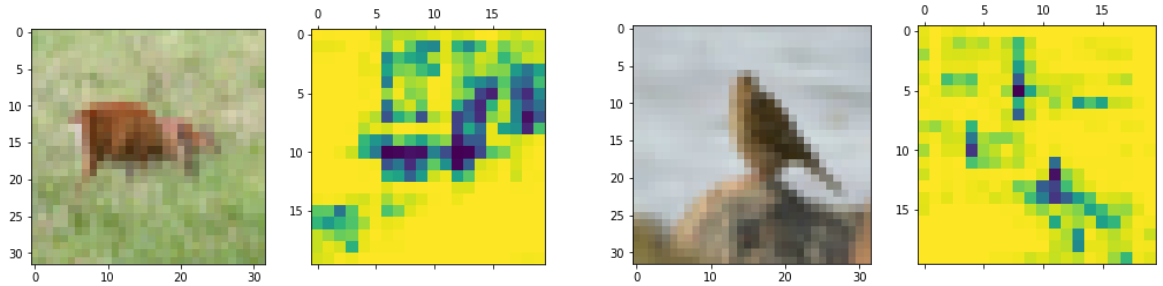
Only Channel Attention



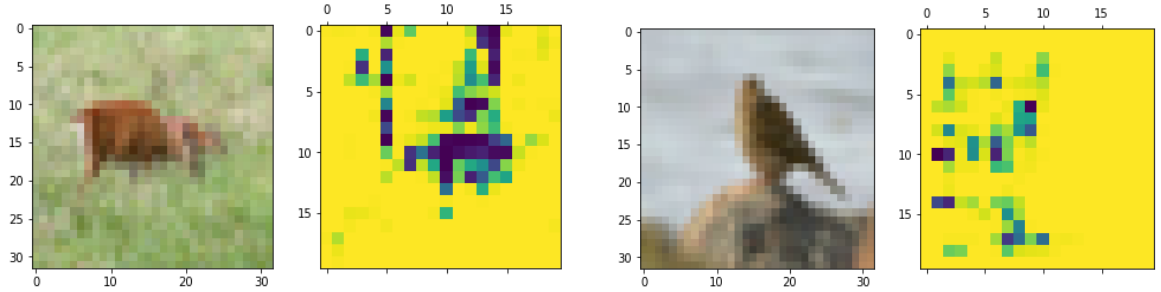
Only Spatial Attention



Channel and Spatial attention in Parallel



Channel and Spatial attention in Series



While validating the performance of CBAM over the conventional baseline models, we evaluate the performance of the Resnet 34 model with CBAM and without CBAM on CIFAR 10 and CIFAR 100 datasets. The accuracy and loss performance with epochs on the CIFAR10 dataset are similar, but we get a lower test accuracy without CBAM than the case with CBAM.

7. Conclusion

In the ablation tests performed on the CIFAR10 dataset, we can analyze the differences in the different models of the attention module. For each of the following cases, (I) channel only module, (II) spatial only module, (III) channel and spatial module in series, and (IV) channel and spatial module in parallel, loss versus epochs, accuracy versus epochs was plotted for the train, and test sets and the confusion matrices were plotted. Between the channel only and spatial only modules, it can be seen that the loss and accuracy graphs are very similar, and we can only assume that their performance is similar. The confusion matrix generated for both the cases indicates that their performances are very consistent across classes. In the cases where we are comparing the ones with both channel and spatial modules, respectively. It can be seen that there is no significant improvement against the case where we are using only one of the two modules. There is no significant difference between the cases where we are using the modules in series or parallel, respectively. The confusion matrices are also quite similar. Thus, in contrast to the results shown in the paper, we are not able to generate evidence to suggest that the use of CBAM in any combination improves the performance of the model.

To validate the performance of CBAM over the conventional baseline models, we evaluate the performance of the Resnet 34 model with CBAM and without CBAM on CIFAR 10 and CIFAR 100 datasets. The accuracy and loss performance with epochs on the CIFAR10 dataset are similar, but we get a lower test accuracy in the case without CBAM than the case with CBAM. Thus we are unable to show here that the model's performance improves with CBAM. On the CIFAR 100 dataset, besides the loss and accuracy, we are also able to study the top 1 and top 5 error % as there are 100 classes. The top 5% error doesn't really make sense for CIFAR 10 as there are only 10 classes in total. However, even in this dataset, we don't obtain evidence of CBAM improving the model's performance. One of the key differences with the paper is that it was performed

significantly on the Image net dataset. Since this dataset is no longer publicly available, the studies had to be performed on the CIFAR 10 and CIFAR 100 datasets.

But from our understanding, Channel attention identifies which feature map is important for learning and is refined. Meanwhile, the spatial attention conveys what within the feature map is essential to learn. Hence, robustly combining both attention modules enhances the Feature Maps and thus justifies the significant improvement in model performance.

--- THANK YOU ---