ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2021.2.059

MENDEL Soft Computing Journal

A Streamlined Attention Mechanism for Image Classification and Fine-Grained Visual Recognition

D. Dakshayani Himabindu[1,2,✉], **S. Praveen Kumar**[3]

[1]Research Scholar, Department of CSE, GIT, GITAM University
[2]Assistant professor, Department of IT, VNRVJIET
[3]Assistant Professor, Department of CSE, GIT, GITAM University
dakshayanihimabindu_d@vnrvjiet.in✉

Abstract

In the recent advancements attention mechanism in deep learning had played a vital role in proving better results in tasks under computer vision. There exists multiple kinds of works under attention mechanism which includes under image classification, fine-grained visual recognition, image captioning, video captioning, object detection and recognition tasks. Global and local attention are the two attention based mechanisms which helps in interpreting the attentive partial. Considering this criteria, there exists channel and spatial attention where in channel attention considers the most attentive channel among the produced block of channels and spatial attention considers which region among the space needs to be focused on. We have proposed a streamlined attention block module *which helps in enhancing the feature based learning with less number of additional layers i.e., a GAP layer followed by a linear layer with an incorporation of second order pooling (GSoP) after every layer in the utilized encoder. This mechanism has produced better range dependencies by the conducted experimentation. We have experimented our model on CIFAR-10, CIFAR-100 and FGVC-Aircrafts datasets considering fine-grained visual recognition. We were successful in achieving state-of-the-result for FGVC-Aircrafts with an accuracy of 97%.*

Keywords: *Visual Attention, Spatial Attention, Channel Attention, Fine-Grained Visual Recognition, Image Classification, Deep Learning.*

Received: 12 December 2021
Accepted: 20 December 2021
Published: 21 December 2021

1 Introduction

The evolution of deep learning [13] in the world has immensely emerged providing a deeper understanding of convolutional neural networks [12, 14]. The proposal of deeply stacked convolutional neural networks [8, 17] has proved to be providing better interpretable features and patterns which have proved in gaining state-of-the-art results in computer vision tasks such as image classification, segmentation, recognition and, detection. There are various techniques employed on neural networks under the tasks of computer vision, one major technique that is employed to provide a better comprehension includes Attention-based modeling.

An attention mechanism was introduced [1, 19] to enhance the feature-based performance by extracting the focused and valuable information from a neural network under the motive of machine translation tasks. As we are aware that the attention mechanism plays a vital role in human visual system by focusing on partial parts of a scene which are salient enough in order to capture the structure in an efficient manner rather than capturing the whole, hence it later proved to perform equally well on tasks under computer vision as well. There exist several efficient approaches which are classified and presented in a structured form of attention extraction. While neural machine translation has attained its state-of-results under transduction tasks [6, 20, 23], in parallel attention-based mechanisms were considered favorable under vision tasks; mainly in image captioning. This captured the attention between visual features and their respective text generation. These effective approaches include global and local attention techniques which pursue soft and hard attention [18]. Global attention is performed through the "soft attention approach" which learns the aligned weights and the features that are extracted from the whole data i.e due to the application of attention over all the patches of the image which eventually results in computational expense. Soft attention is considered to be low performing due to its nature of deriving the attentive weighted sum average from the whole image patch. On the other hand, local attention performs both soft and hard attention, where hard attention focuses mainly on sub-patches of the image. This combination helps in finding corresponding aligned weights as well as proves better results compared to global attention. Due to its tendency of collecting useful information from sub-patches of the image, it proves to be less computationally expensive.

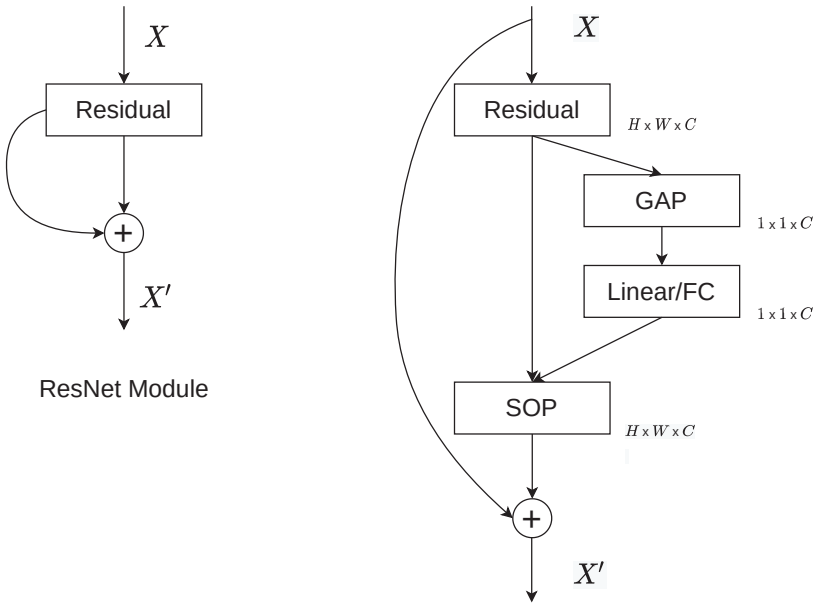


Figure 1: (a) Traditional ResNet module, (b) Streamlined ResNet attention module; where GAP denotes global average pooling and SOP denotes second order pooling.

Similarly, spatial and channel-wise attention mechanisms are also some of the recent advances in attention under the tasks of computer vision which are used in the architecture of grouped convnets. Spatial attention resembles the attention mechanism on the space encapsulated within the feature map helping in refining the resulting feature maps. This provides an enhancement of input proving an improvement in the performance. This mechanism was applied into architectures of several tasks such as image caption generation, visual questioning and etc. Bi-linear convolutional neural networks [16] were one of the advances in which spatial attention was involved in their architecture. Channel attention provides weights for those channels which prove in contributing towards learning and thus boosts the overall model performance. Channel attention came into existence after a while of well-performed models under the mechanisms of spatial attention, which were later proved to be partially efficient.

The combination of spatial and channel attention proved to perform well through some of the recent time advancements. The Squeeze-and-Excitation Network [9], SCANet [4], Convolutional Block Attention Module [27] were one of the evolution's which combined the spatial and channel attention-based mechanisms in order to bring out the features which are much effective and relevant. The recent advancements prove efficient channel attention and spatial attention by performing some reasonable experiments which tend to perform better than the previous models.

In this paper, considering fine-grained visual recognition we propose a streamlined attention block module which integrates both efficient channel attention and spatial attention through a mechanism called as second order pooling a.k.a element wise multiplication which involves a mathematical approach of deriving an outer product from two spatial matrices. We have implemented our proposed model on CIFAR-10, CIFAR-100, and Aircrafts datasets. This proves to provide better attention and gains better features with an accuracy proved in the Table 2, 3, 4 respectively.

2 Previous Work

Tsung-Yu Lin *et al.* [16] proposed Bilinear convolutional neural networks that captures the localized feature interactions when an input image is pooled and an outer product of features is obtained which is derived from the two convolutional neural networks that are accessed as Encoders, considering second order pooling at the end of the respective network. These two convolutional neural networks whose features are involved in generating an outer product proves to provide better fine-grained features such as texture-based and part-based feature representations. Jie Hu *et al.* [9] proposed the Squeeze-and-Excitation networks which illustrates a channel attention mechanism in which blocks of squeeze-and-excitation layers proves cardinality. The 'squeeze' operation by the means of its spatial dimensions produces a channel descriptor by aggregating the feature maps obtained. This embedding is later followed by another operation which takes the aggregation as input and produce per-channel modulation weights. This operation is hence called as the 'excitation' operation. Sanghyun Woo *et al.* [27] have intro-

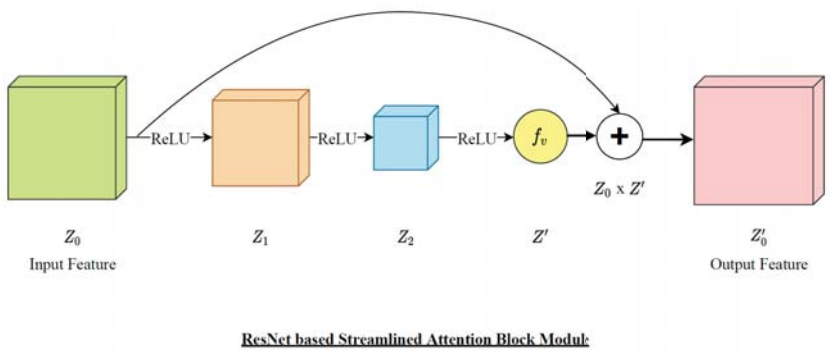


Figure 2: The proposed Streamlined attention block module. Input feature is initially passed through a GAP layer and FC layer. The obtained feature vector f_v undergoes SOP operation with the initiated input feature providing the succeeding block of feature maps.

duced a module involving sequentially inferring attention maps where channel attention module is followed by spatial attention module that is incorporated in a block based architecture. These separated attentions are multiplied with the input feature maps obtained through the passage of the respective module products in order to obtain adaptive feature refinement. This block is considered to feasible enough to integrate the features obtained from the combination of channel and spatial attention in any CNN architecture. Qilong Wang *et al.* [25] proposed an efficient channel attention network that moreover concentrates on avoiding the dimensionality reduction by using the appropriate cross-channel interaction which can help in reduction of model complexity and preserve performance. The provided model can be efficiently implemented via a 1-D convolution. Considering the ongoing trend of involving second order pooling layer at the end of the model proposed earlier, Zilin Gao *et al.* [7] proposed a model named GSoP, where the model involves a global second order networks i.e second order pooling layer is linked at the end of each layer of the respective encoder. This has proved to attain efficient higher order representations by taking non-linear modelling under consideration.

3 Methodology

We have proposed a streamlined attention block module which includes the combination of efficient channel attention and spatial attention in a divergent manner. This attention block succeed in proving highly relevant feature based information. It includes less number of additional layers proving to provide greater impact in capturing long range dependencies. The architecture with respect to the proposed model has been illustrated in Section 3.1.1, in detail.

3.1 Data Description and Architecture

In order to prove fine-grained visual recognition task we have used the following datasets for this model i.e., CIFAR-10, CIFAR-100, FGVC-Aircrafts. Each dataset contains certain set of training and testing images which are utilized for the respective processing. The details with respect to each dataset has been illustrated in Table 1. The specific reason for choosing these set of datasets is to understand the varying reason between low scale classification, middle scale classification and fine-grained visual recognition. In order to capture, higher range dependencies in fine-grained visual recognition we would be requiring similar kind of images in order to train the model appropriately and efficiently. These datasets were observed to be appropriate, the reason is further discussed in section 4.

By taking performance under consideration, ResNet [8] module has been utilized as the encoder for the proposed architecture. This module has proved to perform better on the preceding works as well. The model architecture involves an additional block of Streamlined attention module after each layer of the traditional ResNet module. The proposed module includes a global average pooling layer succeeding with a fully connected layer. Further, an operation of second order pooling takes place which will be discussed in the later sections. The detailed structure of the proposed module have illustrated in Fig. 1.

3.2 Channel Attention

The proposed streamlined attention block module has been improved by considering the mechanisms i.e., [27], [25], and [9], as these proposed works highly focus on how efficient a channel attention can be proved to get better. In [27] Sanghyun Woo *et al.* proposed that the channel wise dependencies are captured through the obtained block/combination of the feature maps which are considered to be efficient enough to provide long range feature dependencies. This mainly helps in providing high attention with lesser number of additional layers. This exploits the features obtained via cross-channel interaction considering the fact of "what" needs to be focused in a channel attention block. Following that work, [25] had proved similar

Table 1: Description of the utilized datasets

Datasets	No. of classes	Train Samples	Test Samples
Cifar-10	10	50k	10k
Cifar-100	100	50k	10k
FGVC-Aircrafts	100	6.6k	3.3k

results by following a norm of "dimensionality reduction" through an introduction of a kernel which eventually considers the neurons as initialized. In [9], they have concentrated on the channel attention by considering the respective mechanism of squeezing the feature maps and exciting them after a point. The proposed attention module depicts the mechanism of following global average pooling and second order pooling as mentioned earlier in the Section 3.1. This mechanism is considered to be meticulous due to its tendency of grasping attention with lesser number of effective layers.

When a block of feature maps are passed through the provided attention module, the feature maps attain channel attention as they focus on the long range dependencies obtained from the feature maps of the previous layers and utilized via global average pooling layer as illustrated in the Fig. 3. Global average pooling plays a vital role in applying attention in a model. This helps in bringing down the feature dimensions of H x W x C to 1 x 1 convolution based feature map with C number of channels having highly useful information. Further, passage through the linear/ fully connected layer accomplishes the point of training an additional layer for attaining the second order pooling mechanism. Later, gaining is proved through the spatial attention mechanism which is discussed in the Section 3.3.

3.3 Spatial Attention

As mentioned the feature refinement is obtained through the encapsulated spatial area of a specific feature map. This is proved via inter-spatial relationship as in [27]. The resulting spatial map has proved to be effective by providing highly interpretable information. Spatial attention mainly focuses on 'where' in the image should the attention be gained for finer feature extraction. The feature maps computed in this block are considered to be 2-dimensional due to them being spatial $\mathbb{R}^{1\times H\times W}$. Hence, the obtained features from the Linear/ FC layer are spatially distributed among the map, to which the input is further combined to successfully establish second order pooling operation, which was illustrated earlier. The attained spatial attention is due to the preceding operation, as second order pooling combines the two projected feature maps spatially. Further discussion regarding second order pooling has been illustrated in section 3.4 in detail.

3.4 Importance of Second Order Pooling

The main intention behind the convolutional neural networks is to capture multiple classes and objects which are defined in high dimensional space. Whenever the higher order representations are captured effectively and efficiently by having the capability of enriching the non-linear modelling, then the model can be concluded to be highly efficient. One of such methodology to prove capturing of higher order representations is global second order pooling (GSoP).GSoP layers are observed to be successful in proving meaningful image representations in form of covaraince matrices providing state-of-the-art results in [15], [5], [26] and [24] under object recognition, detection, video classification and fine-grained visual recognition tasks. GSoP proves to perform well and produce state-of-the-art results in some of the previous like B-CNN [16], DeepO$_2$P [10] by training the convolutional neural networks in an end-to-end manner by further attaching the GSoP layer at the last layers i.e at the end of the whole model. Considering GSoP layer after each layer in the part of encoder, [7] were successful enough to prove the state-of-the-art results. We have proved results for model by utilizing this mechanism in the architecture of the proposed streamlined attention block module. The structure and interpretation of the proposed architecture is proved in detail in section 3.5

3.5 The Proposed Streamlined Attention Block Module

This methodology includes an extended function connected to each of the 3rd convolutional block in the bottleneck architecture of the respective encoder- ResNet. The traditional architecture of the residual networks include residual connections, where we attach an additional streamlined attention block module which provides meticulous neural attention for obtaining long-range dependencies with lesser complexity compared to existing literature. In streamlined attention block module, we include the mentioned global average pooling in the prior level of the function considering the dimensions of the provided input to depict the averaged weights for the succeeding layers of the respective attention block module. The succeeding weights would be passing through a linear/ fully connected layer of the channel based dimensions. The following step would include an element-wise multiplication i.e. second order pooling with respect to the proved weights of the preceding layers. The importance of second order pooling have been addressed in section 3.1.4, which ultimately helps in considering that the attained attention

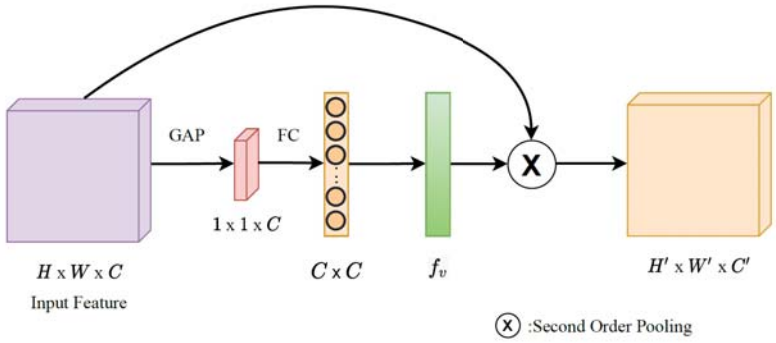


Figure 3: The proposed Streamlined attention block module. Input feature is initially passed through a GAP layer and FC layer. The obtained feature vector f_v undergoes SOP operation with the initiated input feature providing the succeeding block of feature maps.

Table 2: Performance comparison with existing literature over CIFAR-10 dataset

CIFAR10				
Attention	Model	No. of parameters	Training Time	Test Accuracy
No	ResNet18	11M+	9 mins + 30 Epochs	85.95 ± 1.00
	ResNet34	21M+	11 min + 30 Epochs	88.63 ± 1.70
	ResNet50	25M+	18 min + 30 Epochs	88.67 ± 1.45
Yes	CBAM-ResNet50		13 min + 30 Epochs	83.75 ± 1.30
	ECANet-ResNet50		8 min + 19 Epochs	88.89 ± 1.59
	Steamlined CBAM-ResNet50	43.6M	19 min + 25 Epochs	**89.45 ± 8.7**
	Steamlined CBAM-ResNet101	80.4M	25 min + 31 Epochs	**89.42 ± 7.3**
	Steamlined CBAM-ResNet152	110.7M	32 min + 30 Epochs	**89.20 ± 6.5**

Table 3: Performance comparison with existing literature over CIFAR-100 dataset

CIFAR100				
Attention	Model	No. of parameters	Training Time	Test Accuracy
No	ResNet18	11M+	18 min + 30 Epochs	65.35 ± 1.91
	ResNet34	21M+	23 min + 37 Epochs	67.23 ± 1.55
	ResNet50	25M+	36 min + 33 Epochs	69.68 ± 1.47
Yes	CBAM-ResNet50		27 min + 27 Epochs	69.27 ± 4.78
	ECANet-ResNet50		13 min + 30 Epochs	69.79 ± 3.60
	Streamlined CBAM-ResNet50	43.6M	14 min + 18 Epochs	**70.19 ± 10.87**
	Streamlined CBAM-ResNet101	80.4M	25 min + 23 Epochs	69.87 ± 12.9
	Streamlined CBAM-ResNet152	110.7M	31 min + 28 Epocha	69.23 ± 8.75

from the provided attention is considered to be highly efficient.

Let us suppose the subsequent feature block from the fully connected layer to be denoted as Z_a and the initial input as Z, such that we can represent the forthcoming operation of second order pooling as,

$$\boldsymbol{Z_a} \otimes \boldsymbol{Z} = \boldsymbol{Z'}$$

where $\otimes$ is the respective denomination of second order pooling. Hence the resulting feature output is considered as Z'.

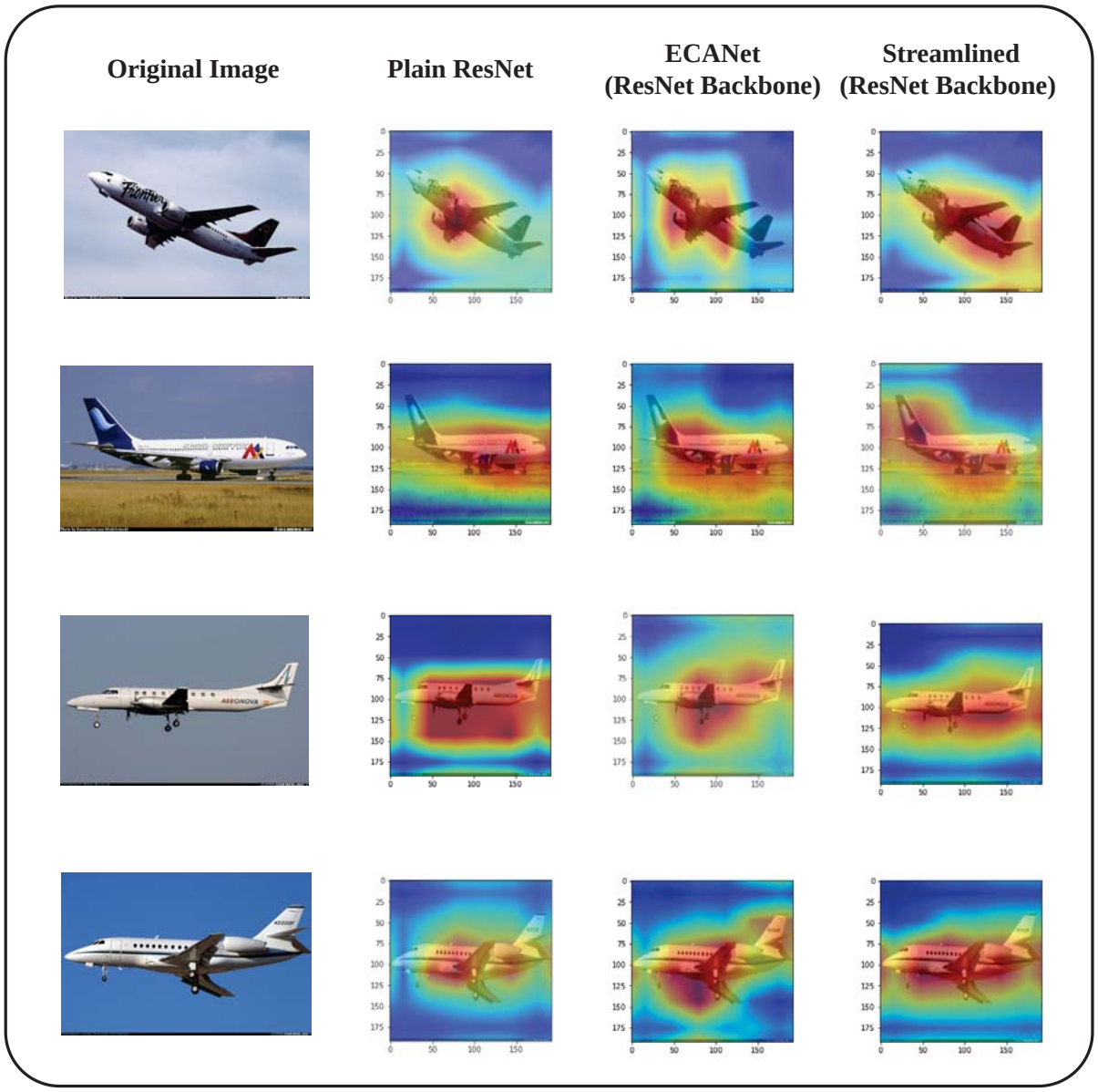
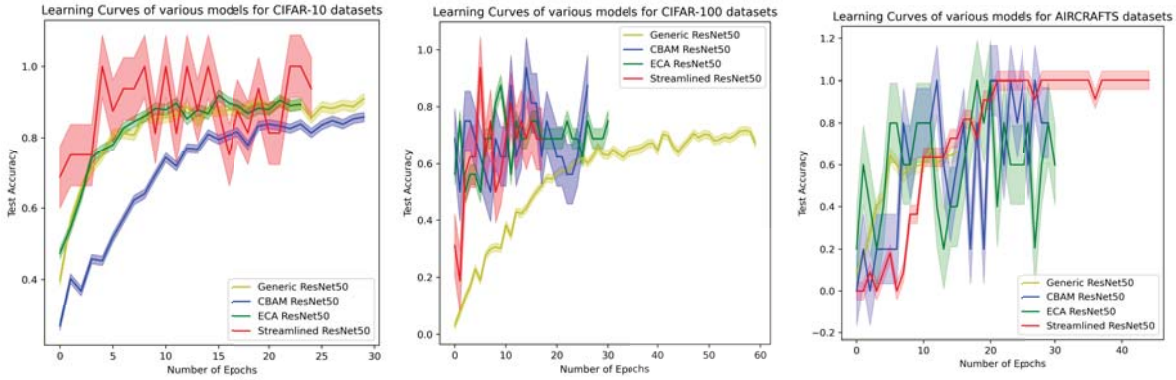


Figure 4: Class Activation Maps (CAMs) obtained for Generic ResNet,ECANet and Proposed Streamlined for AIRCRAFTS dataset. The images are randomly chosen from the testing set of the AIRCRAFT'S data set.

This block of streamlined attention module is attached to the encoder. Considering the feature vector resulted from the proposed module to be Z' and the input of the mainstream as Z_0, the operation regarding summation of the respective skip connections in residual neural networks can be justified as

$$Z' \oplus Z_o$$

Additionally, during back-propagation, the non trained neuron with respect to the linear layer are also freshly trained which are considered to be cause for explicit attention. This non-linear mapping with the use of FC layer eventually trains at the training phase when data is fed and able to understand the complex relationships and again map these features to successive convolution layered features. The process of using FC layer aids the learning process to acquire definite features and impart attention where ever it is required. As a note, it should be understood that, this attention mechanism need not require any new hyper-parameter to obtain attention. It is a tedious process to choose a hyper-parameter while training which is very similar to that of ECANet. Hence, we provide end-to-end attention without the requirement of additional hyper-parameter. The weights which are obtained and updated through the process of back-propagation would eventually understand where to provide *attention* and the second-order pooling would provide consistent features without loss of information. These insights led the model perform superior to other and the results obtained are discussed in the subsequent section.



(a) Accuracy w.r.t each epoch under performance criteria for CIFAR-10 dataset

(b) Accuracy w.r.t each epoch under performance criteria for CIFAR-100 dataset

(c) Accuracy w.r.t each epoch under performance criteria for AIRCRAFTS dataset

Figure 5: Learning curves obtained for distinct models for the three datasets i.e., CIFAR-10, 100 and AIRCRAFT's.

Table 4: Performance comparison with existing literature over AIRCRAFTS dataset

		Aicrafts		
Attention	Model	No. of parameters	Training Time	Test Accuracy
No	ResNet18	11M+	1hr 18 min+30 Epochs	62.33 ± 4.01
	ResNet34	21M+	1hr 30 min+30 Epochs	68.11 ± 5.92
	ResNet50	25M+	48 min+16 Epochs	61.72 ± 3.73
Yes	CBAM-ResNet50		1hr 24 min+30 Epochs	86 ± 6.46
	ECANet ResNet50		1hr 15 min+30 Epochs	63.63 ± 8.85
	Streamlined CBAM-ResNet50		2hr 33 min+45 Epochs	**97.27 ± 4.39**
	Streamlined CBAM-ResNet101		2hr 22 min+39 Epochs	**97.67 ± 1.89**
	Streamlined CBAM-ResNet152		2hr 50 min+45 Epochs	**97.77 ± 4.11**

4 Results and Discussions

We have experimented the proposed attention model on the system with a GPU (Graphical Processing Unit) NVIDIA RTX A4000 graphic card with configuration of 16GB VRAM and ≈ 6000 Cuda cores . For experimenting with our method, we have implied three highly reliable datasets Of which, two are used for standard classification they are CIFAR-10 and CIFAR-100. The other dataset is used to understand the attention mechanisms for the models imbibed with attention i.e., specifically for fine-grained visual recognition task. This eventually helps to acquire a set of features captivating the region of interest i.e. the *attentive portion*. As the Aircraft's dataset is specifically utilized for fine-grained visual recognition and if the developed model can compete with recent state-of-the-art methods then, eventually we can affirm that the developed model provides visual attention. Furthermore, with these data, we try to illustrate the performance of the proposed model with attention and depict its *attentive* ability with class activation maps (CAMs).

The tables were evaluated with detailed experimentation. Each individual value obtained was executed for 7-10 setting till it reach convergence. After executing the model for those numerous times, we chose to obtain the mean and standard deviation for the individual execution. The training for each model was varied till the convergence was reached and each model took discrete amount of time to train on each data set. The number of parameters consumed for the data sets CIFAR-10 and 100 were same for all the models but,, it varied for the Aircraft's data set. The data sets CIFAR-10 and 100 are designed for a smaller image size i.e., $32 \times 32 \times 3$. Where as, the Aircraft's images are varied in size and for that reason all the images were isotopically reshaped to $190 \times 190 \times 3$. So, the parameters consumed for the datasets and the no of epochs trained varied. We have used early stopping with a patience of 12 epochs. This obliged to halt the execution whenever the test loss was not decaying and found no learning for a successive 12 epochs. The batch size for each data set were remained constant. For CIFAR-10 and 100, 512 no of samples as batch were fed into the network. Whereas, Aircraft's data set size is less but,

the image resolution much higher compared to that of remaining two data set. So, a batch size of 128 was chosen to train those images and test them accordingly.

During experimentation, we have implied the generic/plain convolution networks i.e. the networks which were developed for ImageNet classification such as, ResNet-18, 34, 50 respectively. These methods do not specifically provide visual attention but, as previously mentioned, there are specific attention methods imposed to learn detailed features. So comparing these variant methods from generic neural architectures can produce a better perspective for learning features from basic classification tasks to fine-grained visual recognition tasks. All these methods with and without attention are illustrated in the Table 2, 3, and 4 respectively.

The process of experimentation included a 10x iterative procedure of execution. This process helped us in interpreting the range of the proposed model in keen manner which eventually helped in highlighting the attention based partials in an initiated image. The optimizer for the model has been chosen as Adam with a learning rate of 10^{-4} [11], the advantage of being computationally efficient and feasible application on sparse data would also help in the respective task of fine-grained visual recognition. It is observed that, adam tend to decrease the computational time and converged with higher rate.

We have chosen categircal crossentropy as our loss function for efficient backpropagation of the respective gradients in a multi-class classification task i.e,

$$\text{Loss} = -\sum_{i=1}^{\substack{\text{OP}\\ \text{size}}} y_i \cdot \log \hat{y}_i$$

where OP size is the output size which resembles the number of scalar values in the respective model output, $\hat{y}_i$ is the i-th is the scalar value in the model output, y_i is the corresponding target value. The OP size varies from dataset to dataset which means, for CIFAR-10 is is set to be 10 and for CIFAR-100 it is chosen as 100.

The visualisation with respect to the predicted classes has been proved via class activation maps (CAMs). These class activation maps not only visually represent the class that the specific network predicts whereas it mainly helps in concentrating on the localized object with respect to the predicted class which helps the user to detect the predicted portion and also without explicitly extracting through a specific bounding box. The CAMs depcit the ability of a model to provide visual attention by aggregating the final layer feature maps. These feature maps aggregation is also termed as heat maps [3, 22, 28].

The object localization is visualized through heat maps denoting the concentrated region represented as red. In mundane terms, the color which varies eventually depicts what portion of image the model is focusing on i.e, the more the red in color the model is more *attentive* towards that region. As our model is based on attention mechanism, the level of concentration over a specific region is highly focused and hence the usage of the discriminative localization on the proposed model is highly relevant to prove the concentration of the obtained attention based output.Hence, we specifically shuffled the testing data and chosen four images at random and plotted them in the Fig. 4. Further, it is seen that attention tend to provide superior performance even in agricultural images for detailed visual recognition of infected leaves classification [2, 21].

The ability to learn for a model eventually varies from epoch to epoch. As we trained our model numerous times, we plotted the learning curves for individual model for all the three data sets. Thus, we take the devation produced for every epoch and produce the learning curves by combining all the possible cases produced. In the Fig. 5, the shaded proportion for each model defines the deviation and the solid line witch specific color retains the mean value obtained. The model accuracy have been displayed in the Fig. 5 and the prediction based class activation maps have been displayed through the Fig. 4.

From the results obtained, it was observed that, the model tend to provide keen visual attention for fine-grained visual recognition and can perform Superior for large-scale visual recognition. This is taken as future endeavour to build a model which is capable of providing state-of-the-art outcomes not only for Aircrafts dataset but also for other variant data sets.

5 Limitations

Even though we were successful enough in producing state-of-the-results for FGVC-Aircrafts dataset, there exists few complications in the proposed model with respect to the number of parameters and the layer based second order pooling. We were able to produce effective and efficient results from the proposed model whereas, the reason for the extensive number of parameters is due to the added layers i.e global average pooling and a linear layer for every block of the proposed module. The issue with respect to the second order pooling is due to its attachment with every layer of the same.

6 Conclusion

In this research, we produced a novel block for extracting fine-grained features and it eventually tend to improve performance for standard classification and fine-grained visual recognition tasks. We successfully depicted that, the model tend to produce visual attention with class activation maps. As a future endeavour, we try to understand the varying attention mechanisms and try to integrate with Long short term memory (LSTM's) and Gated Reccurent Units (GRU's) to produce a model which understands the temporal sequence of patterns. Also, we try to implement our Streamlined model for various bio-medical imaging and create social impact by depciting it's performance at large scale.

References

[1] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *CoRR abs/1409.0473* (2015).

[2] Ch, R., et al. *Deep Bi-linear Convolution Neural Network for Plant Disease Identification and Classification.* 06 2021, pp. 293–305.

[3] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)* (2018), IEEE, pp. 839–847.

[4] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., and Chua, T.-S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 6298–6306.

[5] Cui, Y., Zhou, F., Wang, J., Liu, X., Lin, Y., and Belongie, S. J. Kernel pooling for convolutional neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 3049–3058.

[6] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Gao, Z., Xie, J., Wang, Q., and Li, P. Global second-order pooling convolutional networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 3019–3028.

[8] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[9] Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7132–7141.

[10] Ionescu, C., Vantzos, O., and Sminchisescu, C. Matrix backpropagation for deep networks with structured layers. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 2965–2973.

[11] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2015).

[12] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25, Curran Associates, Inc.

[13] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521* (05 2015), 436–44.

[14] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[15] Li, P., Xie, J., Wang, Q., and Gao, Z. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 947–955.

[16] Lin, T.-Y., RoyChowdhury, A., and Maji, S. Bilinear cnn models for fine-grained visual recognition. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 1449–1457.

[17] Liu, S., and Deng, W. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (2015), pp. 730–734.

[18] Luong, T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In *EMNLP* (2015).

[19] Mnih, V., Heess, N. M. O., Graves, A., and Kavukcuoglu, K. Recurrent models of visual attention. In *NIPS* (2014).

[20] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training.

[21] Rao, D. S., et al. Plant disease classification using deep bilinear cnn. *INTELLIGENT AUTOMATION AND SOFT COMPUTING 31*, 1 (2022), 161–176.

[22] Selvaraju, R. R., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 618–626.

[23] Vaswani, A., et al. Attention is all you need. In *Advances in neural information processing systems* (2017), pp. 5998–6008.

[24] Wang, H., et al. Multi-scale location-aware kernel representation for object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 1248–1257.

[25] Wang, Q., et al. Eca-net: Efficient channel attention for deep convolutional neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 11531–11539.

[26] Wang, Y., Long, M., Wang, J., and Yu, P. S. Spatiotemporal pyramid network for video action recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 2097–2106.

[27] Woo, S., Park, J., Lee, J.-Y., and Kweon, I.-S. Cbam: Convolutional block attention module. In *ECCV* (2018).

[28] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2921–2929.