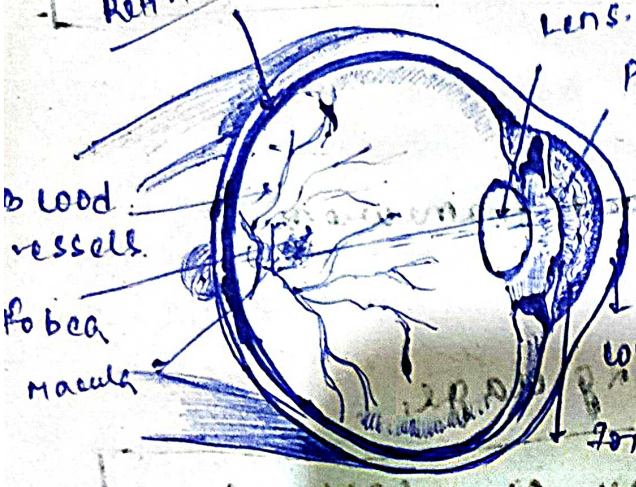
Understanding attention modules.

CBAM and BAM

Brief on human vision system.



Whenever light traverses from a single object of interest to the Macula, which is the primary functional region of Retina inside the eye.

But, whenever a diverse set of objects are present in the scene and lights from all the objects of interest traverses to Macula of Retina, which is primary functional region of Retina inside eye, our visual perception system kind of use an advanced attention mechanism to focus on single object of interest.

So, we want to have similar kind of research for our computer vision system, so that we need to incorporate attention mechanism to our vision based deep learning models. And at the same time, the attention mechanism was 1st popularized by the paper Neural Machine translation, (Attention are you need) for in visual system if we use such attention mechanism the object of interest becomes in focus and surrounding objects/ backgrounds are blurred/ faded.

What are spatial and channel attention modules

The words spatial and channel are most-referred terms in computer vision. these refers to convolution layers.

A convolution layer can be represented by a 3D volume of shape typically $C \times H \times W$.

where each slice of the volume refers to a feature map (or map or extracted features and mapped) over a depth of C channels.

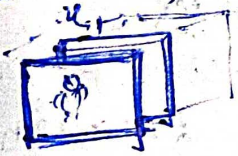
C = total channels of convolution volume,

(often can be considered total feature maps in the layer / total kernels used in the convolution layer)

What is mean by spatial attention?

① Spatial means the feature space / domain space encapsulated within feature map or in layman word is, the space / domain space which contains the feature map / or the feature space or domain space where the convolution features are mapped called feature map.

② Spatial attention mechanism / attention mechanism masks on the feature space which defines a single-cross-sectional area, which is the domain of interest in that feature space useful for classification.



For example in the given feature map 1 ⇒ the shape is (dragon) object is important to attention, so, spatial attention gives attention to features in that feature space (), so, spatial (which represent the object) attention enhances these features, by generating a spatial mask/ spatial attention mask, by refining spatial features. it will enhance the subsequent convolution layers, which improves performance of the model.

Why is channel attention

Usually in convolution layers, the weights inside the kernels which make the feature maps worked on due to the convolution volume are very small close to zero, which why various feature maps containing the convolution volume looks like learn similar information, seems like appearing copies of other.

Although, they look similar, but these filters are extremely useful in learning. different types of features in images like horizontal, vertical edges, diff. pattern, textures in image processing. But, channel attention attention provides weightage or influence parameter to the channel which is essential or relevant for learning the classification contributes to learn the task given to the network and boosts the overall model performance (channel attention enhance the weightage of channel, towards learning) only both, isn't one sufficient.

Author's shown that both together worked better as compared to individual attention mechanism,

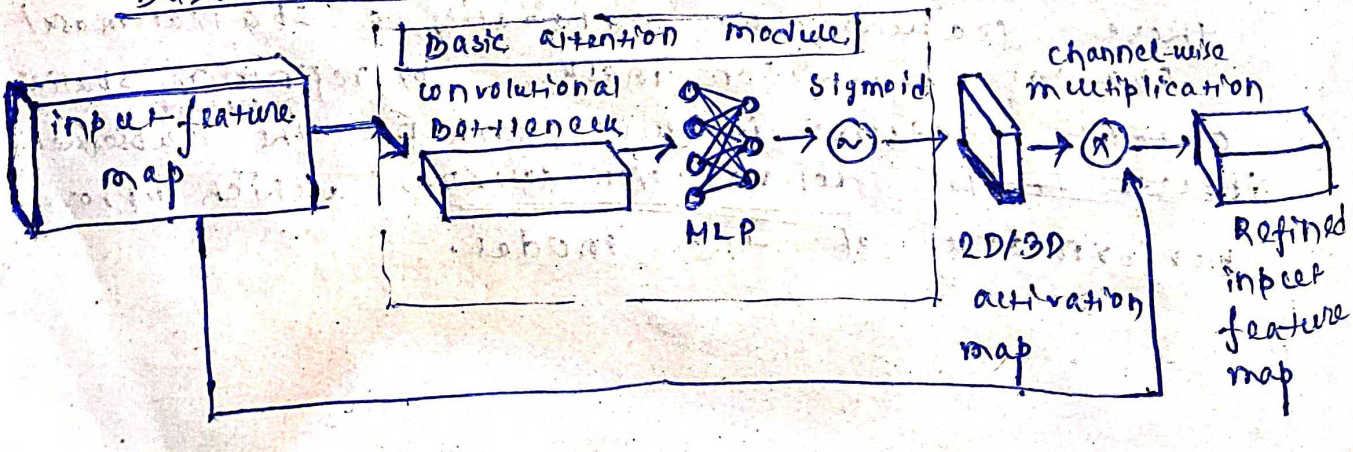
Understanding Attention Modules: CBAM and BAM

Recently different SOTA networks have leveraged these attention mechanisms that have significantly improved and refined real-time results.

Attention Module:

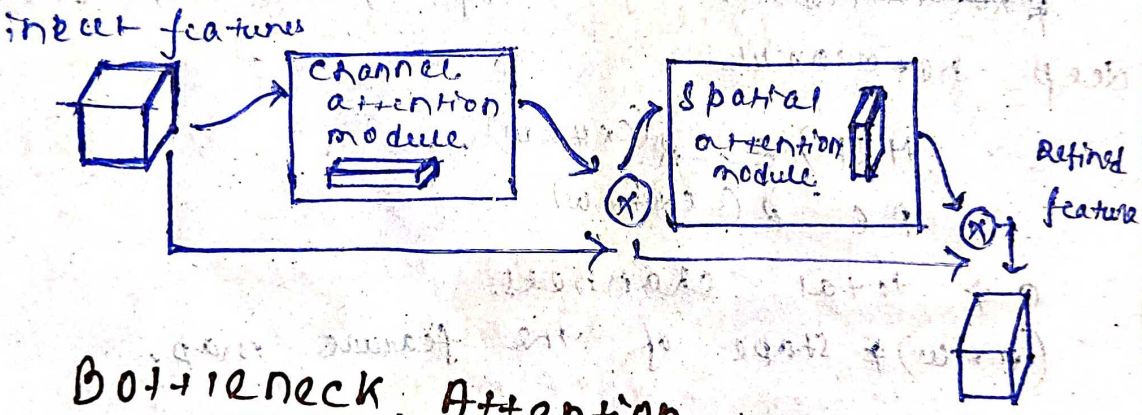
Attention modules are used to provide attention on relevant info in convolution volume, instead of unimportant background info not useful for the task given to the CNN.

Base structure

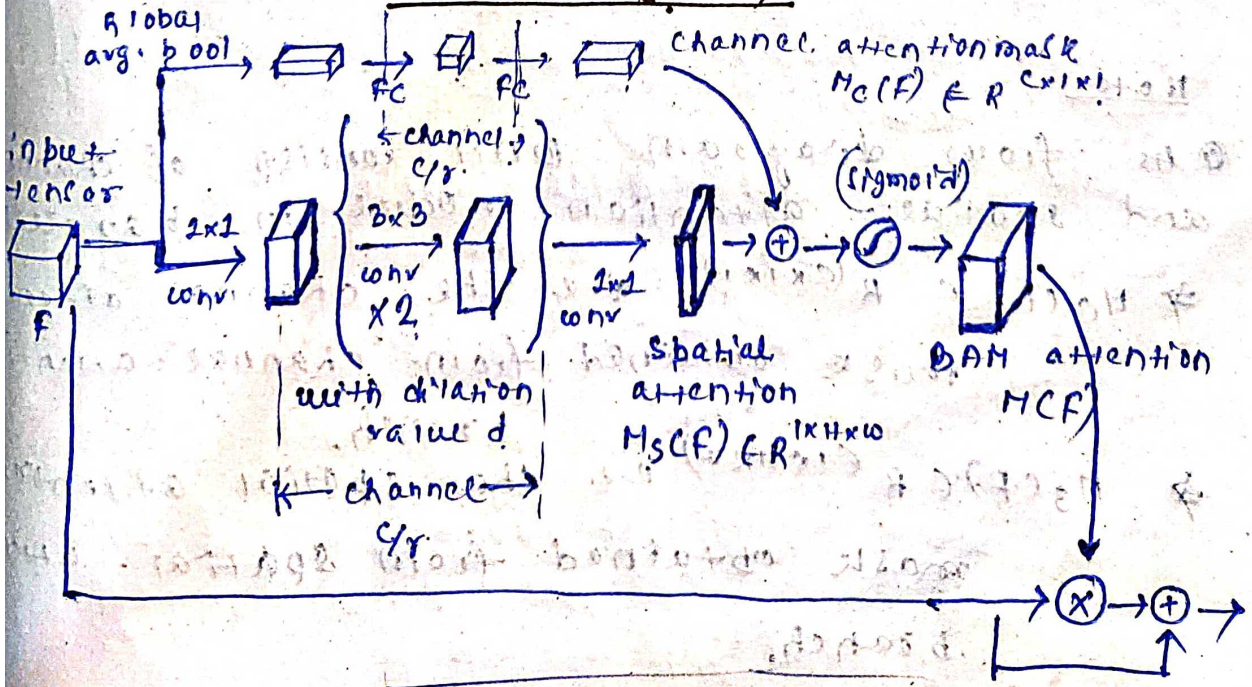


- It takes $(C \times H \times W)$ input feature map as input. $(1 \times H \times W)$ output feature map or $(C \times H \times W)$ for 3D attention map. This attention map / attention mask is multiplied with input feature map to get refined output feature map. Which relevant information are enhanced as compared to unimportant info.
- Generally attention mechanisms are applied to spatial and channel dimensions. These two attention mechanisms can be applied sequentially or parallelly.
- Attention mechanism can be experimented in residual architectures.

Over all structure (BAM):



Bottleneck Attention Module (BAM)



[structure of BAM]

$f \Rightarrow$ input feature map

$f' \Rightarrow$ output refined/enhanced input feature map.

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

$\otimes \Rightarrow$ defined elementwise multiplication

$M(F) \Rightarrow$ attention mask generated as a function of F from (BAM)

The step $\{F + \boxed{F \otimes M(F)}\}$ acts as the residual step where we add the output of previous layer to the input to next layer so as to reduce the error due to vanishing gradient descent and ensure the flow of information over the deep networks.

$M(F) \in R^{(C \times H \times W)}$

$F \in R^{(C \times H \times W)}$

$C \Rightarrow$ total channels

$(H \times W) \Rightarrow$ shape of the feature map.

Note:

① As from diagram BAM consists of channel and spatial attention module in parallel.

$\Rightarrow M_C(F) \in R^{(C \times 1 \times 1)}$ i.e. the channel attention mask obtained from channel attention branch.

$\Rightarrow M_S(F) \in R^{(1 \times H \times W)}$ i.e. the spatial attention mask obtained from spatial attention branch.

(1) And sigmoid is used to obtain the $M_c(F)$ attention mask from combination of $M_c(F)$ and $M_s(F)$, thus

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

as, $\left\{ \begin{array}{l} M_c(F) \in R^{(C \times 1 \times 1)} \\ M_s(F) \in R^{(1 \times H \times W)} \end{array} \right\}$ but, $M(F) \in R^{(C \times H \times W)}$

(2) before addition to ~~obtain~~ $M_c(F)$, $M_s(F)$ are resized to $(C \times H \times W)$ (by method of broadcasting will be discussed later)

(3) Note here, to obtain $M(F)$, to combine $M_c(F)$ and $M_s(F)$ for sigmoid we choose elementwise addition over multiplication due to smooth gradient flow at the time of backprop. what does that mean,

if its addition,

$$\frac{\partial e}{\partial M_c(F)} = \frac{\partial e}{\partial M(F)} \cdot \frac{\partial M(F)}{\partial M_c(F)}$$

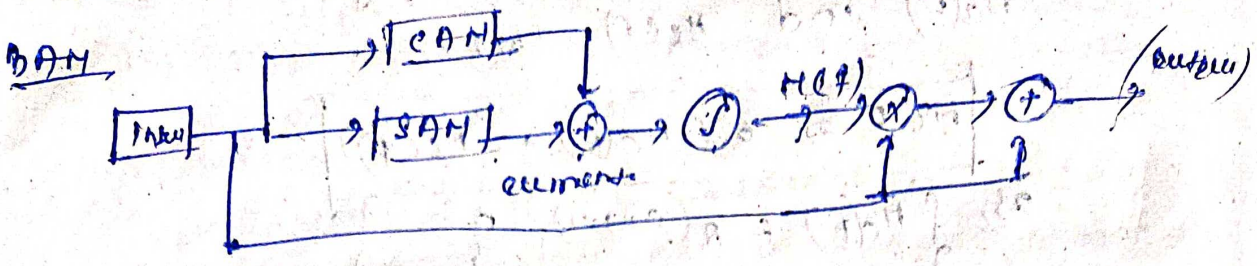
$$= \frac{\partial e}{\partial (M(F))} \cdot 1 \cdot$$

but, if multiplication,

$$\frac{\partial e}{\partial M_c(F)} = \frac{\partial e}{\partial M(F)} \cdot M_s(F) \cdot 1$$

so, this gradient term also dependent on $M_s(F)$, thus that is not robust to $M_c(F)$ term only, i.e. may not smooth flow and gradient on $M_s(F)$ branch (may be kinda recurrent or turbulent flow in backprop)

Channel attention Module



{ Channel Attention Module }

channel attention modules takes the intermediate feature map $F \in R^{(C \times H \times W)}$ and generates channel attention mask $M_c(F) \in R^{(C \times 1 \times 1)}$

Steps to generate channel attention map are.

1. DO global average pooling of feature map F and get a channel vector $F_c \in R^{(C \times 1 \times 1)}$
2. pass F_c to small MLP of one hidden layer followed by one output layer and hidden layer has dimension (C/r). r is the reduction ratio for that channel attention branch.

   so, if F has, 1024 channels and reduction ratio is 16, then

   to tal neurons in hidden layer is.

   $(1024/16) = 64$

3. Add batch norm layer at the from of this MLP.

$$\therefore M_c(F) = BN(MLP(AvgPool(F)))$$
$$= BN(W_2(W_0 AvgPool(F) + b_0) + b_1)$$

where, $W_0 \in R^{(C/r \times C)}$, $b_0 \in R^{(C/r)}$

$W_1 \in R^{(C \times C/r)}$, $b_1 \in R^{C}$

we used reduction ratio, as MLP acts like an auto encoder and the manifold learning is done and hidden layer acts as latent space if Hidden space $\simeq$ C then it means

acts as identity learning and will not learn low dimensional manifold, so, it avoided cause overfitting from the perpective of the ~~CBAM~~ channel activation branch.

Spatial Attention Map (SAM)

Spatial attention map is used to provide attention in spatial domain,

So, takes input $F \in R^{(C \times H \times W)}$

output

$M_S(F) \in R^{(1 \times H \times W)}$

1. Input feature map F is passed through chains of (1×1), (3×3) convolution layers. the (3×3) convolution layer consists of dilated convolution having a dilation rate $(d = 4)$. Dilation is used to increase effective receptive field (refer dilated convolution note), these (3×3) convolution layer contains (C/r) ~~modules~~ kernels.

2. output $\in R^{(C \times H \times W)}$ after that, (1×1) convolution layer, ~~def~~ outputs a spatial attention mask.

pass, $R^{(C \times H \times W)}$ passes through (1×1) convolution layer to generate spatial attention mask,

- (C/r) reduced channel helps as ~~encoder~~ latent space to learn similar to manifold. hard feature extraction and reduce total number of parameters

3. Add batchnormalization at the front of convolution layers stack.

$$M_S(F) = BN(f_3^{1\times1}(f_2^{3\times3}|_{dilate=4}(f_1^{1\times1}(F))))$$

then we combine.

$$M(F) = \phi(M_c + M_s)$$

$$F' = F + F \otimes M(F)$$

as, ~~it~~ CNN's structure followed by the DAH and resembles the Resnet family, as it was implemented on Resnet.

Now $M_c(F) \in R^{(C \times 1 \times 1)}$

① $M_c(F) \in R^{(C \times 1 \times 1)} \Rightarrow$ broadcasted to $(C \times H \times W)$

$\{c_1, c_2 \ldots c_n\}$

$$\begin{bmatrix} c_1 & c_1 & \cdots \\ 1 & & \\ \vdots & & \\ c_1 & \cdots & c_1 \end{bmatrix}_{(1 \times W)} \cdots \begin{bmatrix} c_N & \cdots & c_N \\ & & \\ c_N & \cdots & c_N \end{bmatrix}_{(1 \times W)}$$

② Similarly, $M_s(F) \in R^{(1 \times H \times W)}$ broadcasted to $(C \times H \times W)$

$$\begin{bmatrix} p_1 & p & \cdots \\ & & \\ \cdots & p_{HW} & \end{bmatrix}_{(H \times W)} \quad \begin{bmatrix} p & \cdots \\ & \\ p_{HW} & \end{bmatrix}_{(1 \times H \times W)}$$

1 $\cdots$

to ~~the~~ where they are added.

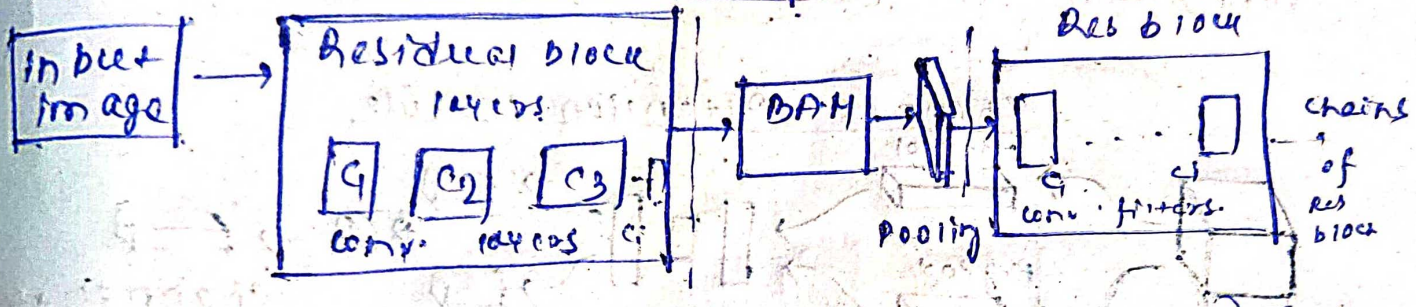
for $M_s|_{(C \times H \times W)}$ $\Rightarrow$ all the channels have spatial attention having the focus or enhance, on the relevant features on feature map.

and on other hand

$M_c|_{(C \times H \times W)}$ $\Rightarrow$ focus on the channel which has highest priority for the goal to achieve by NN.

So, when they add as a combined we obtain a convolution volume (ex HxW) and which has feature map, where spatially attention or enhanced on relevant features along with feature maps having the weightage enchanced based on their relevancy.

Where to keep BAM?



{res. block ⇒ conv ⇒ BN ⇒ activation ⇒ pooling but skip connection}

After each conv (residual block) BAM is used to ~~attention~~ denoise low level features. as early stages and ~~as~~ then at deeper layers it helps to enhance or focuses on exact targets which need a high level of semanticity

{semanticity ≡ meaningfulness}

{semantic segmentation ⇒ pixel wise classification which is more meaningful for that class}

semanticity ⇒ the quality that a linguistic ~~has~~ has ~~of be~~ so that it can be able to convey meanings, in particularly by means reference to the world of physical reality

Inform to HULENR (i) {implement WBAD from AD3 to AP3}

~~PAP has [illegible]~~ we can't acces PAP f/PI transform

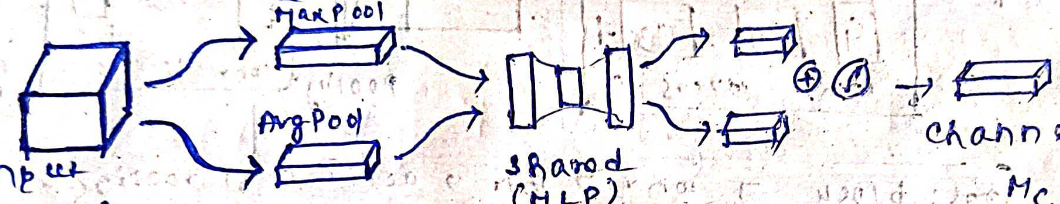
(ii)

(iii) Shift chane

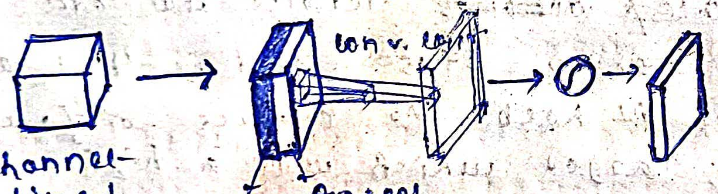
Convolution Block Attention Module (CBAM)

CBAM consists of a sequence of channel attention module and spatial attention module connected in cascaded structure.

Channel Attention Module



Spatial Attention Module



An intermediate feature map $F \in R^{(C \times H \times W)}$ passed to channel attention module.

$$F' = M_c(F) \otimes F$$

and, F' is then sequentially passed to a spatial attention module.

$$F'' = M_s(F') \otimes F'$$

$M_c(F) \in R^{(C \times 1 \times 1)}$
$M_s(F) \in R^{(1 \times H \times W)}$ } with the multiplication we need.

$M_c(F) | \in R^{(C \times H \times W)}$

$\equiv M_s(F') | \in R^{(C \times H \times W)}$

So, $M_s(F)$ and $M_c(F)$ are then broadcasted to respective dimensions.

(2) ⊗ element wise multiplication.

Channel attention map

follows same generation process as DAH. Here we pass two version

↳ (1) Avg. pool of F. } instead the direct
(11) Max pool of F } Global avg. pool of DAH then.

AvgPool(F) passed to MLP()
also MaxPool(F) passed to MLP()

so, here AvgPool and MaxPool both are added to obtain more distinctive ~~feature~~ channel feature.

$$M_c(F) = \delta(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$
$$= \delta(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max})))$$

$\delta \Rightarrow$ sigmoid function.

$W_0 \in (C/r \times C)$, $W_1 \in (C \times C/r)$

ReLU activation is followed by W_0,

Note) Here similar reduction ratio concept is used to obtain a encoder like structure to obtain a relevant information at the latent space or, $W_0 \Rightarrow$ layer (C/r) nodes.

and these transformation is non-linear in nature. It can understand non-linear features or non-linear information which is instilled inside the data, thus helps to denoise the array of $(C \times 1 \times 1)$ vector to obtain influence or weight of different channels.

Spatial Attention Map

Take input feature map F and generate two intermediate feature map F_{avg} and F_{max}

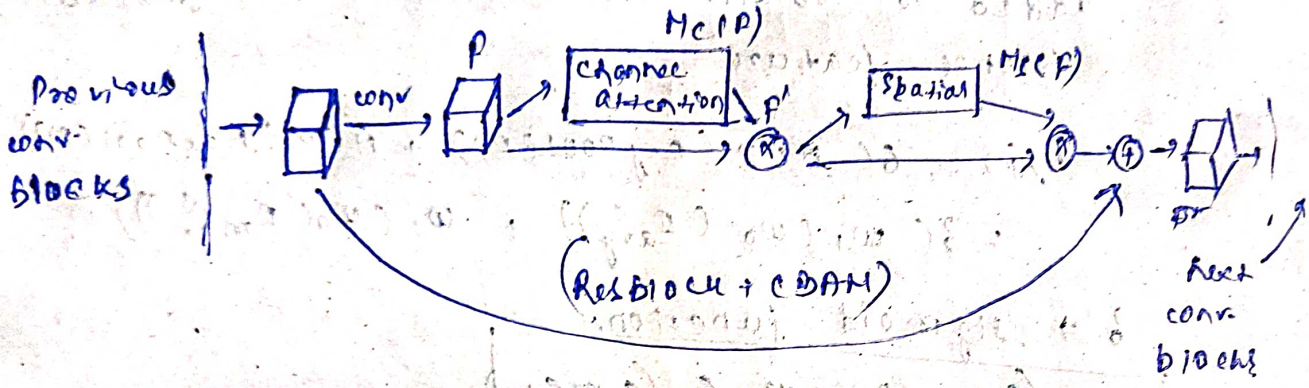
$F_{avg} \Rightarrow$ average pool
$F_{max} \Rightarrow$ max pool
} $\in \mathbb{R}^{C \times H \times W}$

1. Concatenate AvgPool and MaxPool, and pass through single convolution block of kernel size (7×7) with dilation = 1

Note: Unlike BAM, where to increase receptive field used d=4, whereas here, to do so (7×7) kernel used.

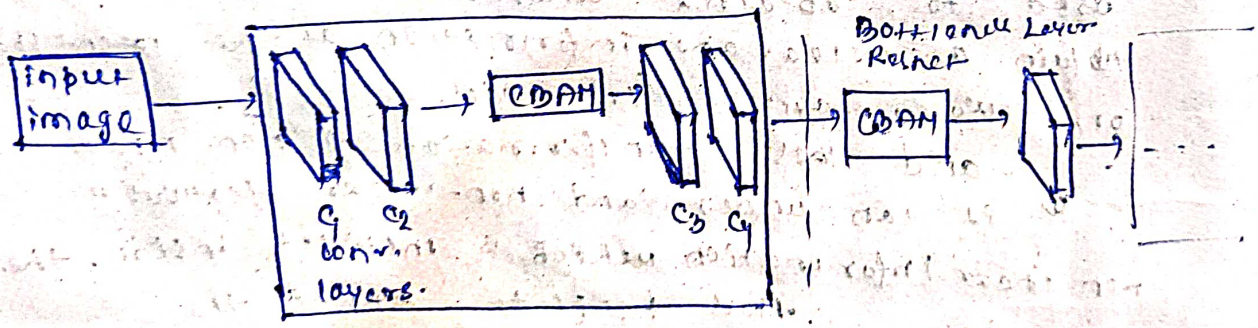
$$M_S(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$
$$= \sigma(f^{7\times7}([F^s_{avg}; F^s_{max}]))$$

How to use



Where to keep CBAM

Unlike BAM, CBAM are placed inside Resnet blo as well as the bottlenecks.



What is the difference between BAM and CBAM

(1) In case of BAM, GAP (Global average pooling) was used to obtain global statistics over the channels of feature map in ~~that~~ channel attention whereas in CBAM, maxpool and avgpool used and it was observed, max and avg pool combinedly helps to generate more refined features from feature map and compensate GAP outputs and encode global statistic softe

and, max and avg pool at combine retrieve
across layers features from feature maps and
compensate RAP o/p and obtain global safety

i) In DAN to increase receptive field in (SAM), dilated conv. used with d=2, same in CDAM is done by using kernel size (7×7), normal convolution layer with d=1

ii) in DAM, CAM and SAM are done in parallel and then combined by element-wise addition followed by sigmoid activation, whereas CDAM CAM and SAM were used sequentially. and at each step input is element-wise multiplied with CAM and SAM attention maps (same as DAM)