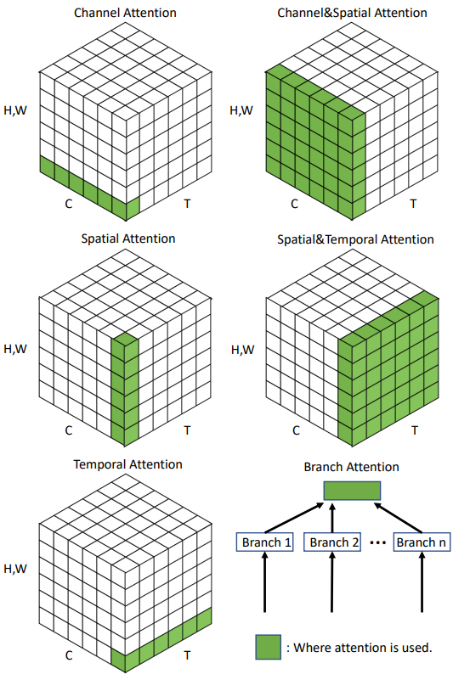
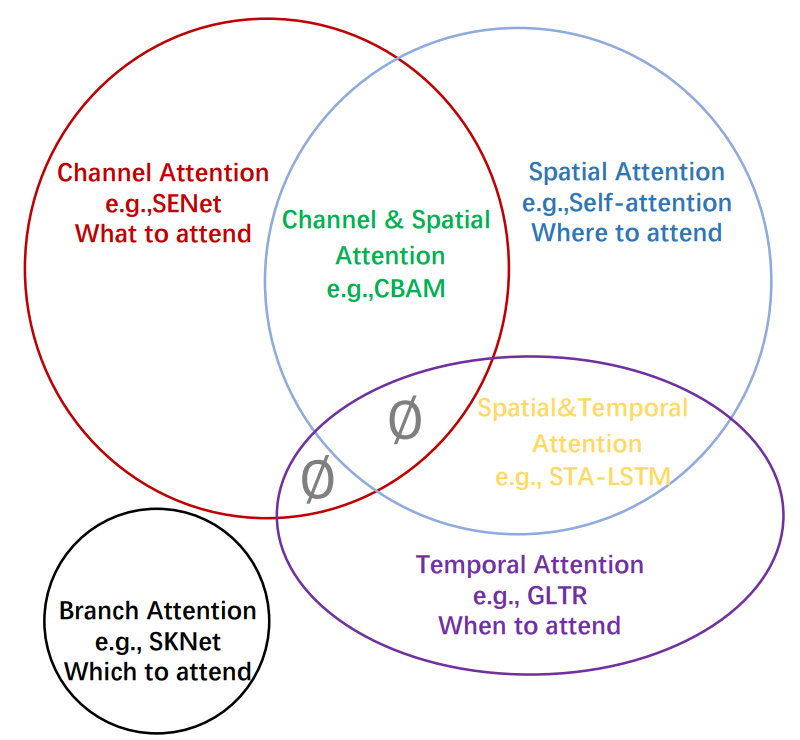
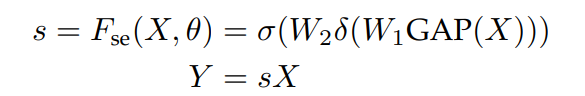
**Attention Mechanisms in Computer Vision:**

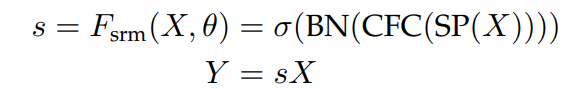
**A Survey**

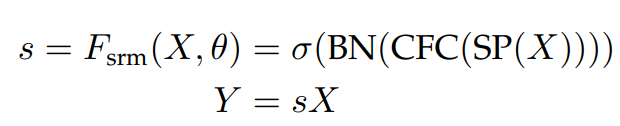
1. Six categories

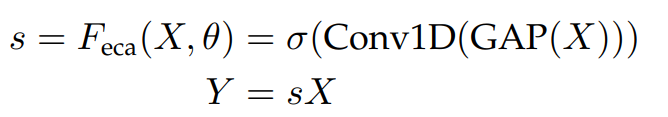


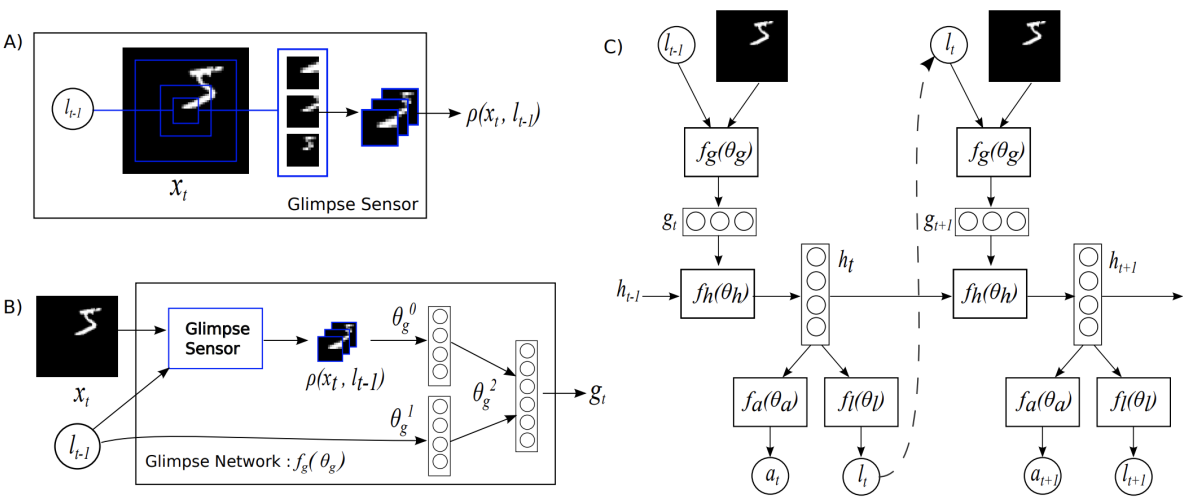
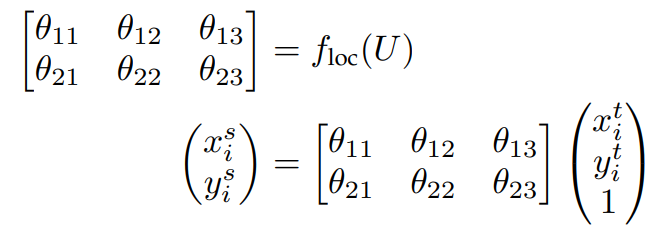
1. Attention Methods
   1. General Form (necessary but not sufficient)
      1. generate attention which corresponds to the process of attending to the discriminative regions
      2. process input x based on the generated attention
2. Channel Attention
   1. Motivation
      1. different channels in different feature maps usually represent different objects/features -- object selection process
   2. SENet
      1. SE block: squeeze and excitation block

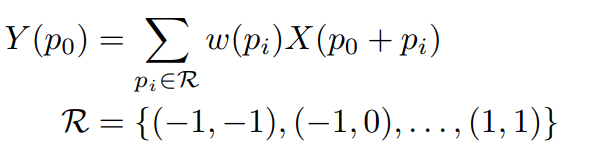


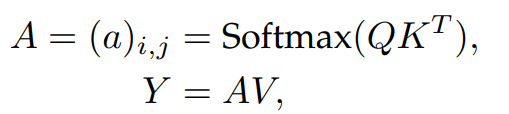
* + 1. Squeeze: capture global spatial information -- GAP, global average pooling -- global pooling outputs 1 response for every feature channel
    2. Excitation: (output attention vector) -- capture channel-wise relationships => output: attention vector (FC + Activation) -- input channels are scaled with the attention vector
    3. Very low computation cost – can be added after each residual unit
  1. GSoP-Net
     1. GSoP block: global second-order pooling – extract global second-order pooling
     2. Squeeze
        1. reduce # of channels with 1x1 convolution (from c to c’)
        2. computer a c’ x c’ covariance matrix — for different channels
        3. row-wise normalization on covariance matrix (it is implemented via BatchNorm2d)
        4. explicitly express the relation between the channels via covariance matrix GAP only considers 2-D information; covariance matrix considers 3-D information
     3. Excitation
        1. row-wise convolution + FC + sigmoid => 1 x c vector (attention vector)
        2. row-wise convolution is realized by group convolution — kernel size = (c’, 1), output = c’ x 4; FC output = c x 4
  2. SRM
     1. SRM module -- style-based recalibration module

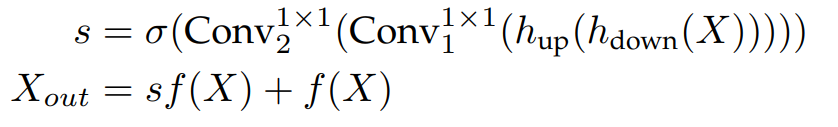
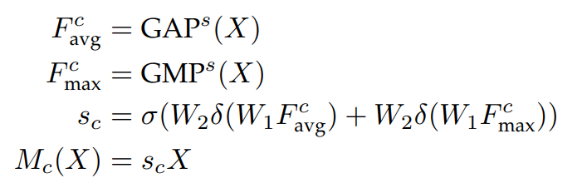
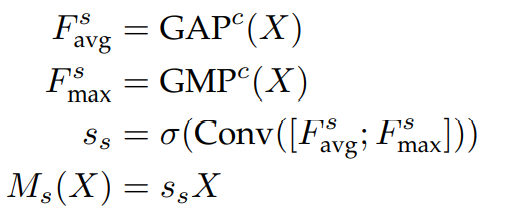


* + 1. SP, style pooling = global average pooling + global standard deviation pooling
    2. Squeeze: motivated by style transfer — mean & stdvar as style features + style integration (recalibration) => style features => attention vector
    3. Excitation: CFC, channel-wise fully connected layer -- FC per (output) channel
  1. ECANet -- efficient channel attention
     1. Motivation: reduce # of channel is not good — it is not a direct way to model the relationship between channels
        1. reduce number of channels is not mentioned in SENet part – it is an optional choice
        2. there is no use of reducing # of channel, proved via experiment in ECANet
     2. Excitation
        1. 1-D convolution (kernel is 2d, but only slide in 1-D) — replace reducing # of channels
        2. only consider the interaction between each channel and its k-nearest channels (it is achieved via channel-wise 1-D conv, whose kernel size is in shape of k, k is manual selected)

1. Spatial Attention
   1. RAM -- recurrent attention model
      1. glimpse sensor — output multiple resolution patches – similar to anchor box
      2. glimpse network — output feature representation
      3. RNN model — output next center coordinate (for glimpse sensor) and next action (target function)
   2. STN -- spatial transformer networks
      1. Motivation: CNN: translation equivariance (conv + pooling); lack other transformation invariance such as rotational invariance, scaling invariance and warping invariance.  
         STN -- explicit procedure to learn invariance  
         
      2. feature map U => conv/FC/Pool/Relu => theta, it is the transformation of the feature (eg. for affine transformation, theta in R^(2\*3))
      3. based on theta, map the input feature map to transformed feature maps
      4. note: STN is to find ROI (feature position),   
         but feature position->feature score is not continuous  
         change in feature position may not influence feature score, which means the net is not differentiable  
         bilinear sampling (just some kind of transformation that makes feature position->feature score differentiable)
   3. Deformable Convolutional Networks (MSRA)
      1. Motivation: standard convolution use standard grid sampling, which is vulnerable to feature translation => adding an offset to each sample point of convolution kernel



* + 1. Details
       1. offsets are learned via another lightweight CNN
       2. introduce deformable ROI pooling — for object detection
  1. Self-Attention
     1. Motivation: due to the localization of the convolution, CNNs have inherently narrow receptive fields, which limits the ability of CNNs to understand scenes globally
     2. self-attention first computes the queries, keys and values (linear + reshape)  
        A in R^(N\*N), N=H x W, is the attention matrix; a\_ij is the relationship between the i\_th and j\_th elements
  2. Vision Transform – pure attention -- not well defined in this paper

1. Branch Attention – multi-branch structure
   1. Gating mechanism / multi-kernel + fuse
2. Channel & Spatial Attention
   1. Residual Attention Network
      1. Motivation: very deep convolutional residual attention network (RAN) by combining an attention mechanism with residual connections
      2. Attention model: mask + trunk + (any structure with pre-activation residual unit and an inception block)  
          
      3. Mask:
         1. bottom-up (max-pooling after residual units) to increase reception field
         2. top-down (linear interpolation to keep the size) to learn a mask of same size
         3. skip-connections between the two parts to capture spatial & cross-channel information  
            3-D attention map
         4. fail to capture global spatial information & high cost to predict a 3-D attention map
   2. CBAM
      1. CBAM: convolutional block attention module  
         reduce computation complexity and capture global spatial information
      2. two sequential sub-modules
         1. global-max-pooling + global-average-pooling in parallel to capture global information
         2. channel:  
             
         3. spatial:  
             

spatial branch works as a complementary to channel attention