Attention

Attention

Channel Attention

- SE Net
- GSop-Net
- SRM
- GCT

Spatial Attention

- RAM
- STN
- DCN
- Self-attention and variants
- ViT

Others

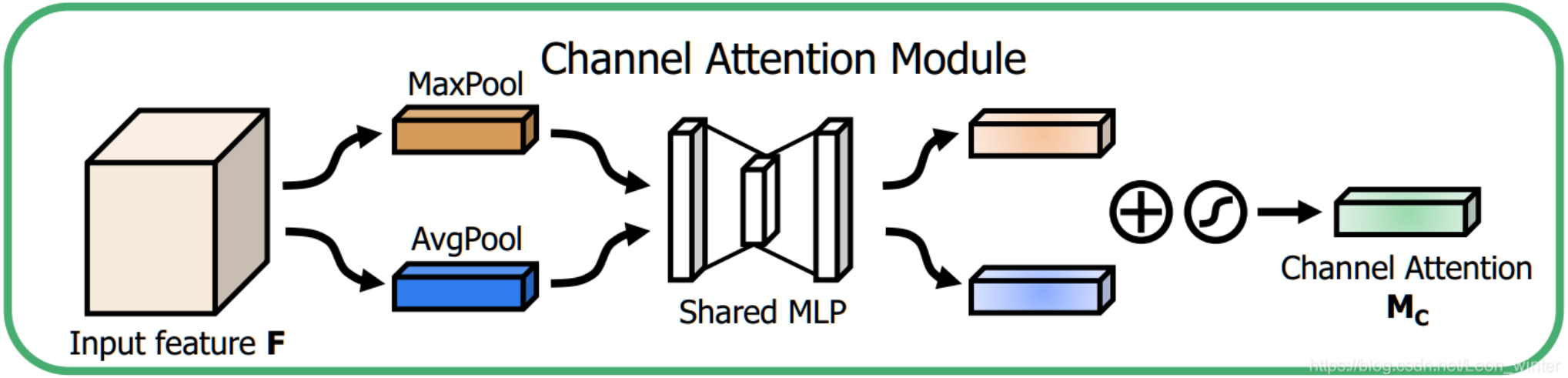
- Temporal attention
- Branch Attention
- Channel&Spatial Attention
- Spatial&Temporal Attention

Attention

- Channel Attention

$$Attention = f(g(x), x)$$

对象选择过程

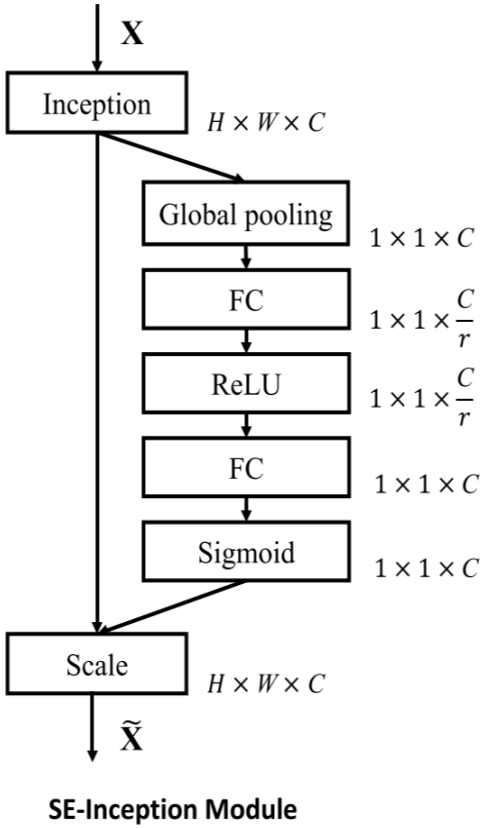


Attention

- SE Net（Squeeze-and-Excitation）

$$s = F_{se}(X, \theta) = \sigma(W_2\delta(W_1 \mathrm{GAP}(X)))$$

$$Y = sX$$



SE-Inception Module

Squeeze: 全局平均池化 ⇨ 收集全局信息

Excitation： 全连接层和非线性层 ⇨ 捕获通道关系并输出注意力向量

优势：

SE块起到强调重要通道同时抑制噪声的作用。
由于计算资源要求低，可以在每个残差单元之后添加一个 SE 块。

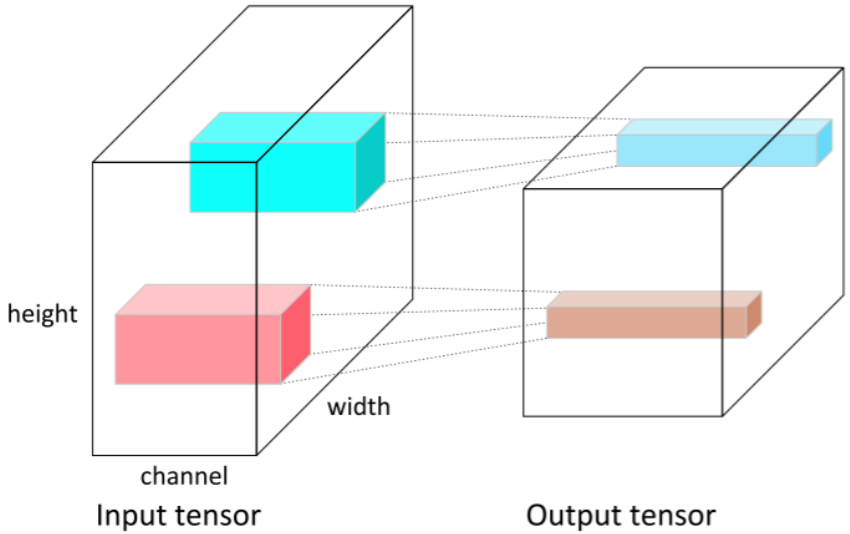
不足：

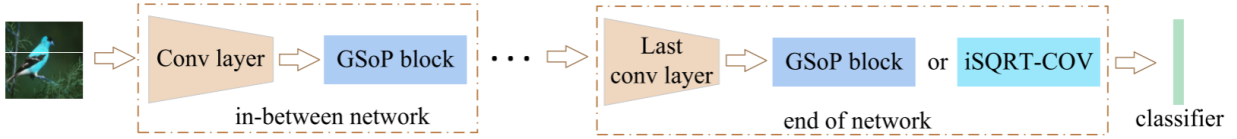
在挤压模块中，全局平均池化过于简单，无法捕获复杂的全局信息。
在激励模块中，全连接层增加了模型的复杂度。

Attention

- GSoP-Net（Global Second-order Pooling）

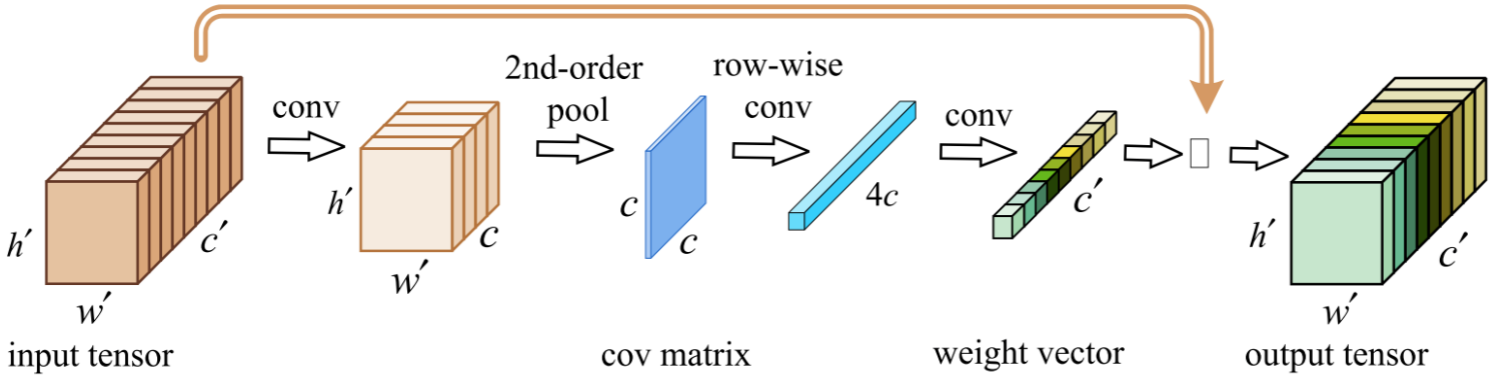
$$s = F_{\text{gsop}}(X, \theta) = \sigma(W\text{RC}(\text{Cov}(\text{Conv}(X))))$$
$$Y = sX$$





与现有的方法不同，GSoP-Net最早将该模型引入到中间层，以便在深度卷积网早期利用整体图像信息

Attention



Method:

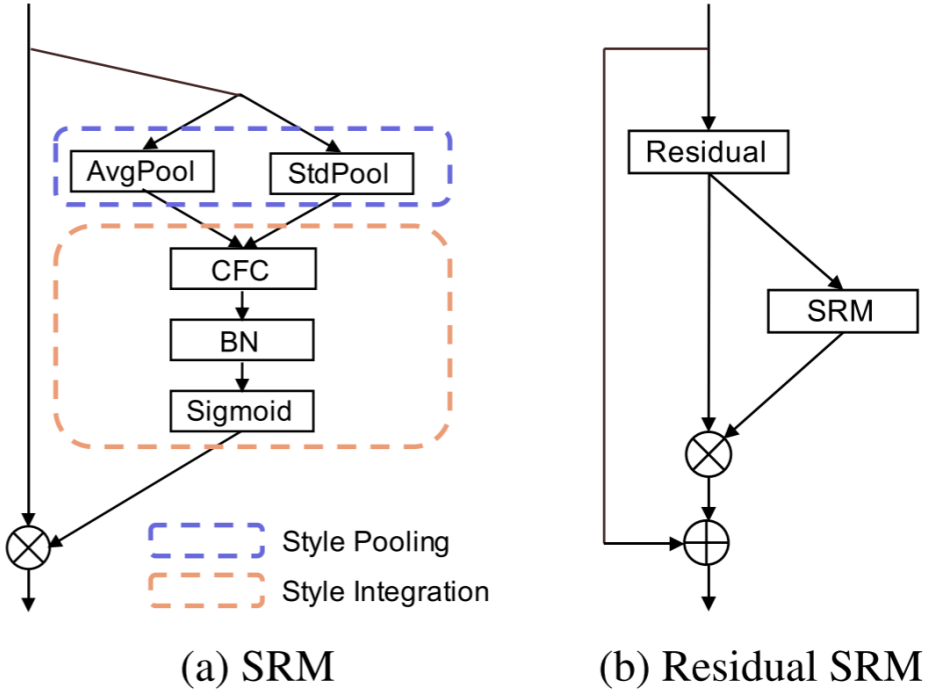
首先使用 1x1 卷积减少通道数
然后计算不同通道的协方差矩阵以获得它们的相关性
接下来，对协方差矩阵执行逐行归一化
执行逐行卷积以保持结构信息并输出向量
然后应用一个全连接层和一个 sigmoid 函数来得到一个 c 维的注意力向量

Attention

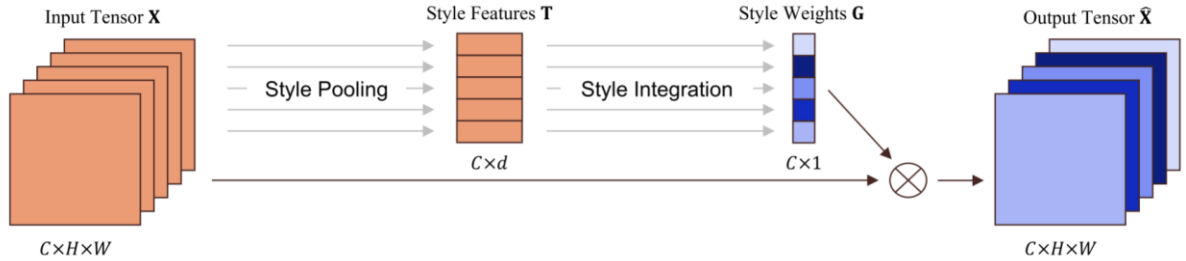
- SRM（Style-based Recalibration Module）

$$s = F_{\text{srm}}(X, \theta) = \sigma(\text{BN}(\text{CFC}(\text{SP}(X))))$$

$$Y = sX$$



(a) SRM　　(b) Residual SRM

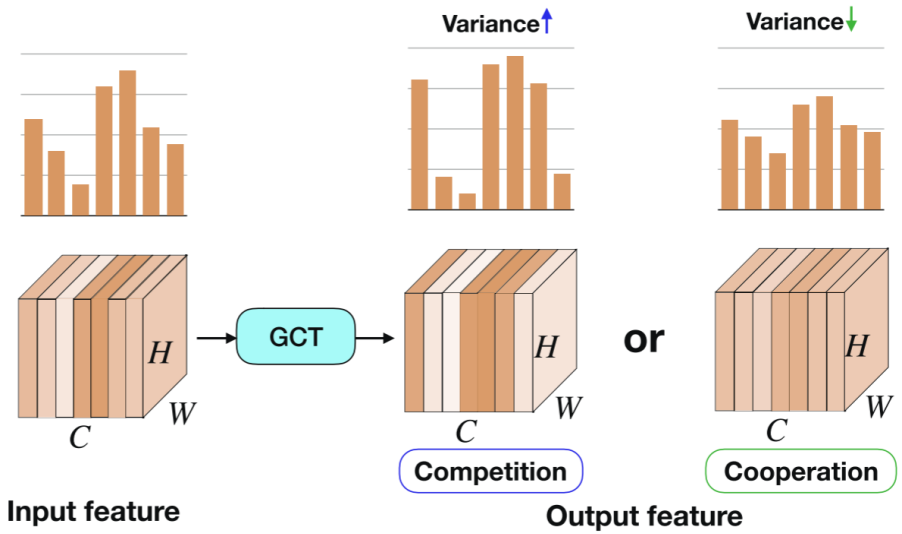


Style Pooling：
采用每个特征图的通道统计数据（平均值和标准差）作为样式特征
不同通道之间的相关性也可以包含在样式向量

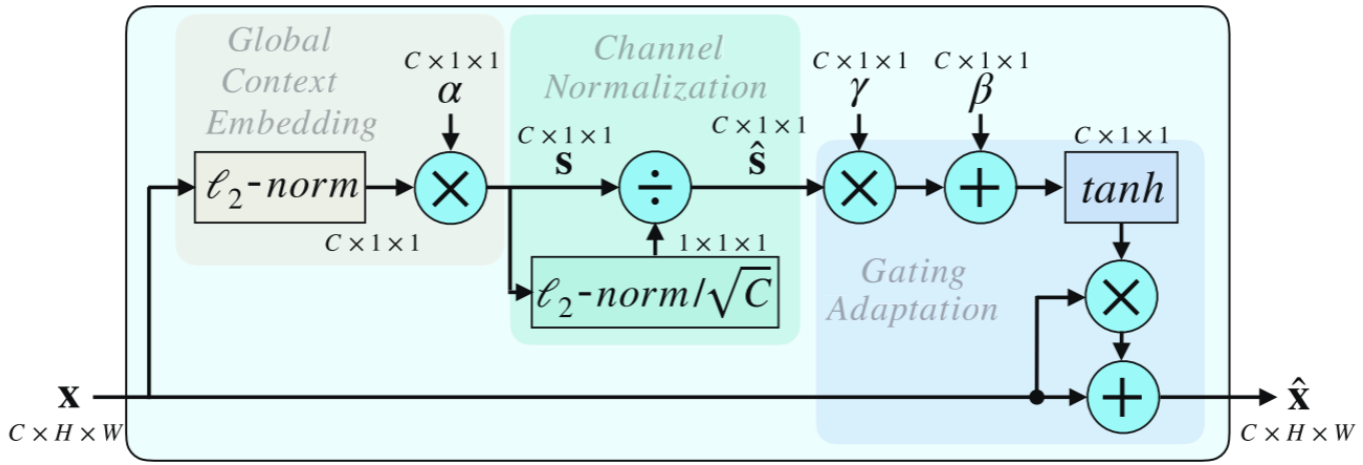
Style Integration：
Channel-wise Fully Connected
Batch Normalization
sigmoid activation function

- GCT(Gated Channel Transformation)

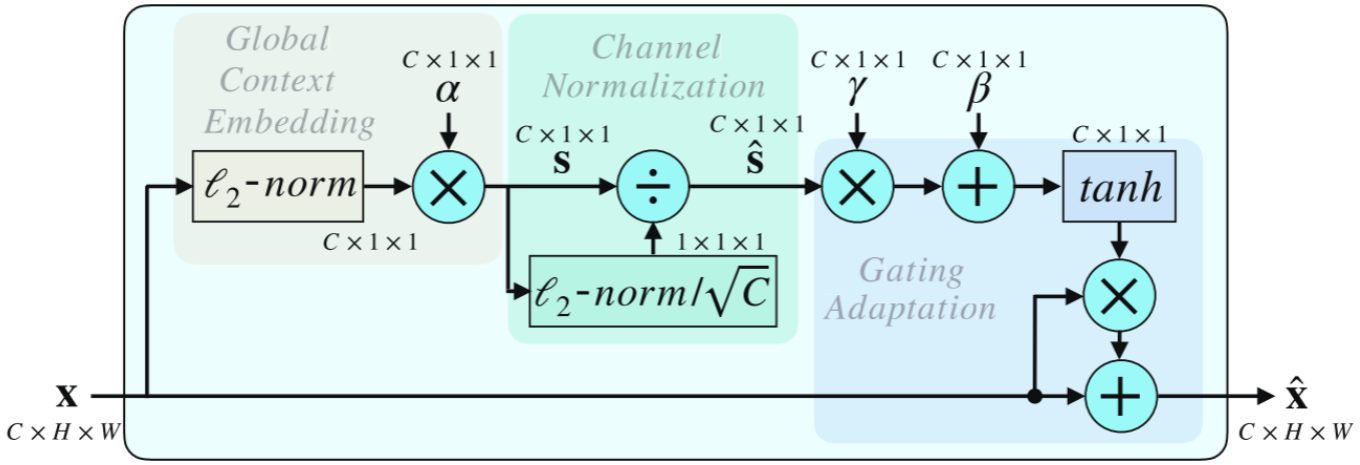
$$s = F_{\text{gct}}(X, \theta) = \tanh(\gamma CN(\alpha \text{Norm}(X)) + \beta)$$
$$Y = sX + X,$$



由于激励模块中全连接层的计算需求和参数数量，在每个卷积层之后使用 SE 块是不切实际的。此外，使用全连接层来模拟通道关系是一个隐含的过程。



Attention



$$s_c = \alpha_c ||x_c||_2 = \alpha_c \{ [\sum_{i=1}^{H} \sum_{j=1}^{W} (x_c^{i,j})^2] + \epsilon \}^{\frac{1}{2}},$$

$$\hat{s}_c = \frac{\sqrt{C} s_c}{||\mathbf{s}||_2} = \frac{\sqrt{C} s_c}{[(\sum_{c=1}^{C} s_c^2) + \epsilon]^{\frac{1}{2}}},$$

$$\hat{x}_c = x_c [1 + \tanh(\gamma_c \hat{s}_c + \beta_c)].$$

Method:

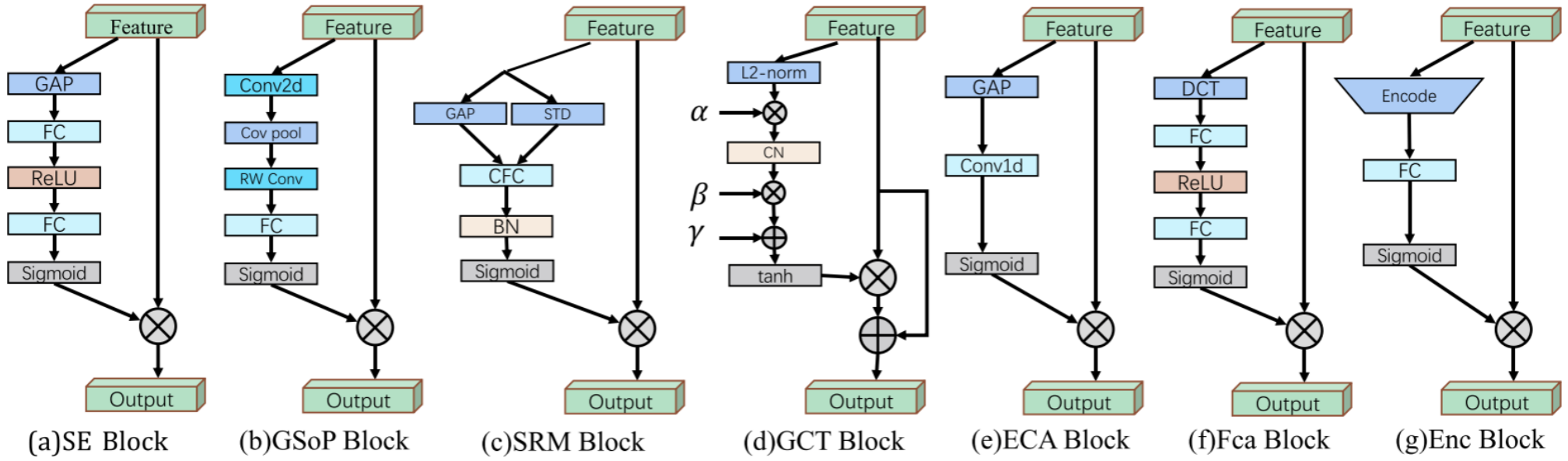
首先通过计算每个通道的 l2 范数来收集全局信息
接下来，应用一个可学习的向量 α 来缩放特征

然后通过通道归一化采用竞争机制在通道之间进行交互。
应用可学习的尺度参数 γ 和偏差 β 来重新缩放归一化。

GCT 采用 tanh 激活来控制注意力向量
最后，它不仅将输入乘以注意力向量，还添加了一个原型连接。

Attention

- Channel Attention



(a)SE Block (b)GSoP Block (c)SRM Block (d)GCT Block (e)ECA Block (f)Fca Block (g)Enc Block

Attention

Channel Attention

- SE Net
- GSop-Net
- SRM
- GCT

Spatial Attention

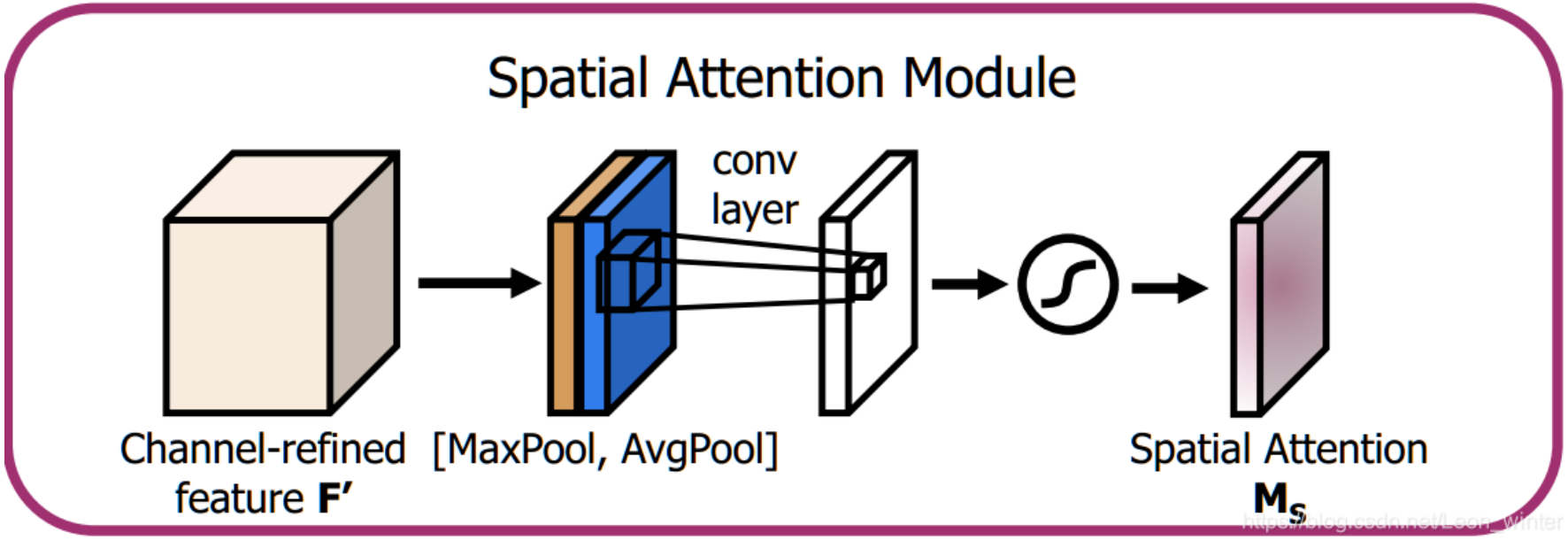
- RAM
- STN
- DCN
- Self-attention and variants
- ViT

Others

- Temporal attention
- Branch Attention
- Channel&Spatial Attention
- Spatial&Temporal Attention

- Spatial Attention

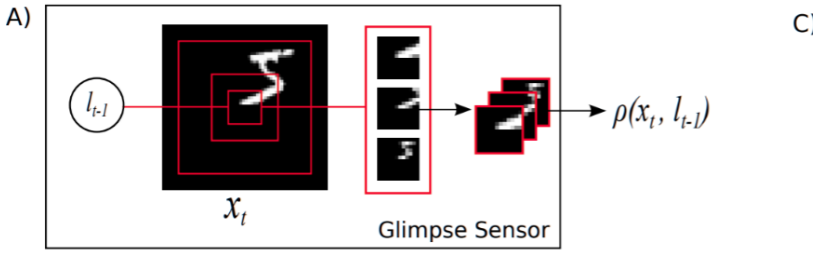
看作是自适应的空间区域选择，关注哪里

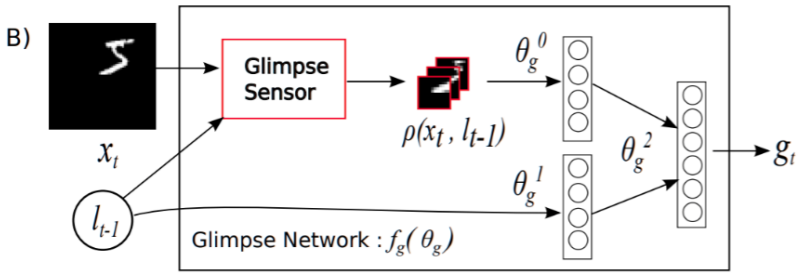


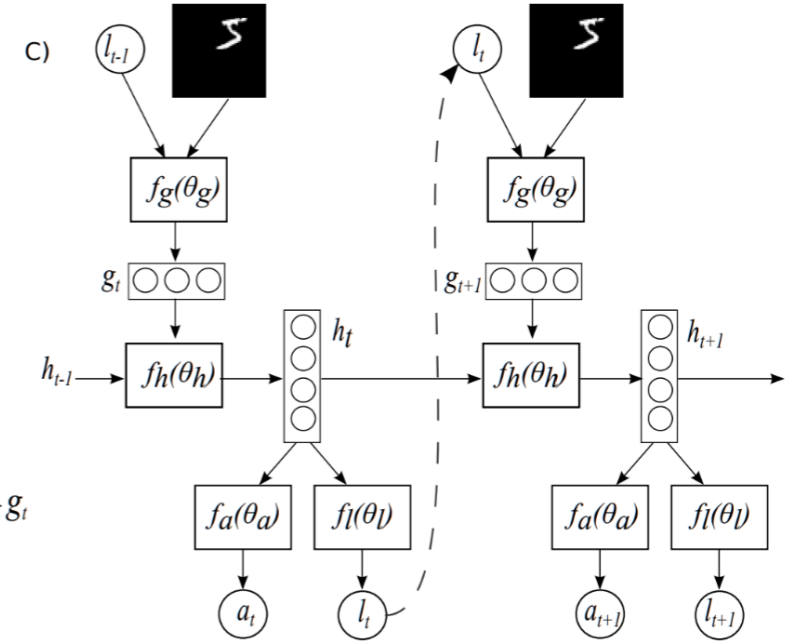
Attention

- RAM(Recurrent Attention Model)
- Glimpse Network

$$g_t = f_{\text{image}}(X) \cdot f_{\text{loc}}(l_t)$$





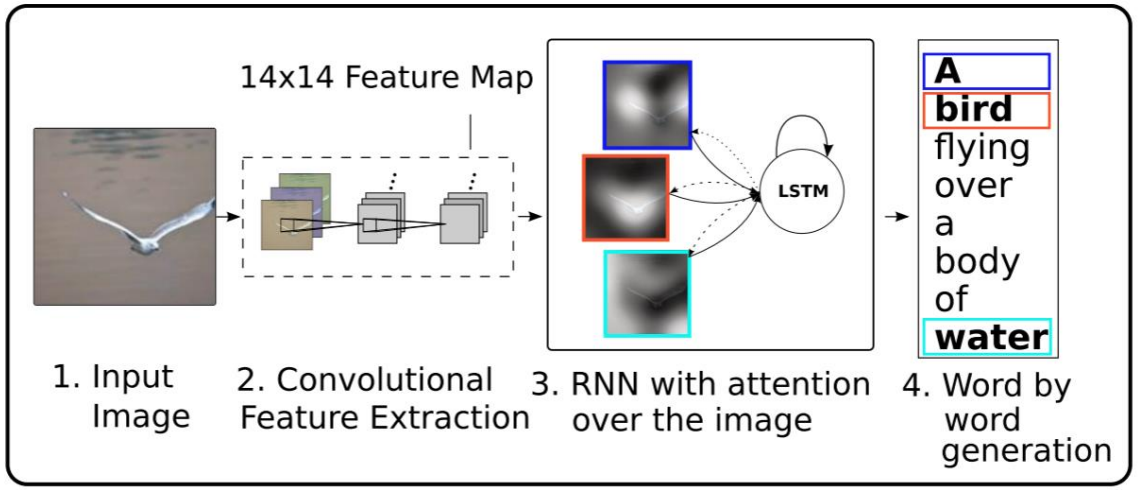


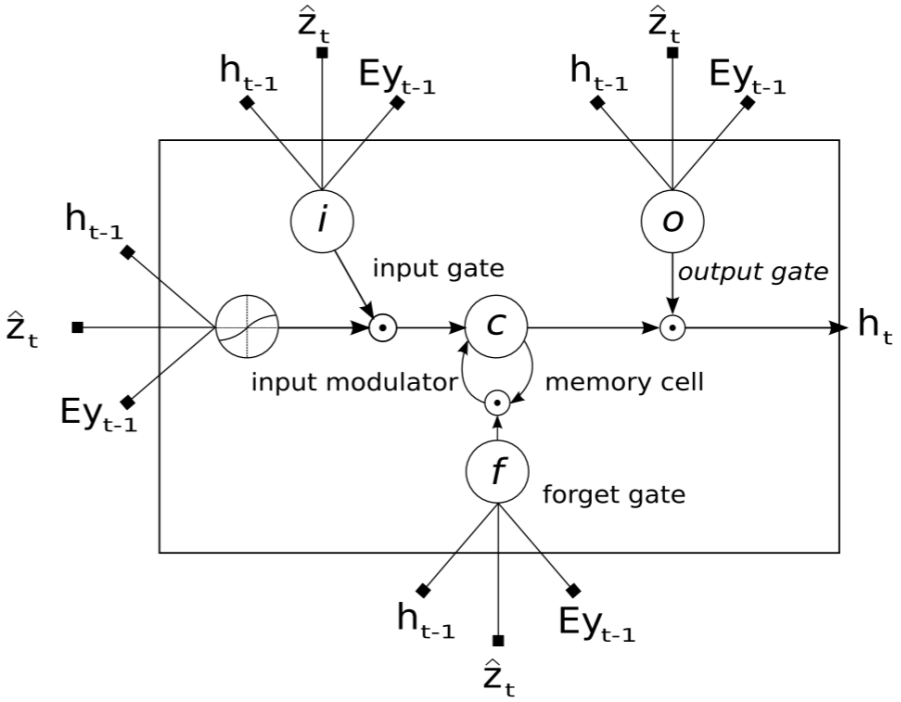
$$r_t^{(1)} = f_{\text{rec}}^{(1)}(g_t, r_{t-1}^{(1)})$$
$$r_t^{(2)} = f_{\text{rec}}^{(2)}(r_t^{(1)}, r_{t-1}^{(2)})$$

$$y = f_{\text{cls}}(r_t^{(1)})$$

Attention

- Hard and Soft Attention





$$\mathbf{i}_t = \sigma(W_i E\mathbf{y}_{t-1} + U_i \mathbf{h}_{t-1} + Z_i \hat{\mathbf{z}}_t + \mathbf{b}_i),$$
$$\mathbf{f}_t = \sigma(W_f E\mathbf{y}_{t-1} + U_f \mathbf{h}_{t-1} + Z_f \hat{\mathbf{z}}_t + \mathbf{b}_f),$$
$$\mathbf{c}_t = \mathbf{f}_t \mathbf{c}_{t-1} + \mathbf{i}_t \tanh(W_c E\mathbf{y}_{t-1} + U_c \mathbf{h}_{t-1} + Z_c \hat{\mathbf{z}}_t + \mathbf{b}_c),$$
$$\mathbf{o}_t = \sigma(W_o E\mathbf{y}_{t-1} + U_o \mathbf{h}_{t-1} + Z_o \hat{\mathbf{z}}_t + \mathbf{b}_o),$$
$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t).$$

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right),$$

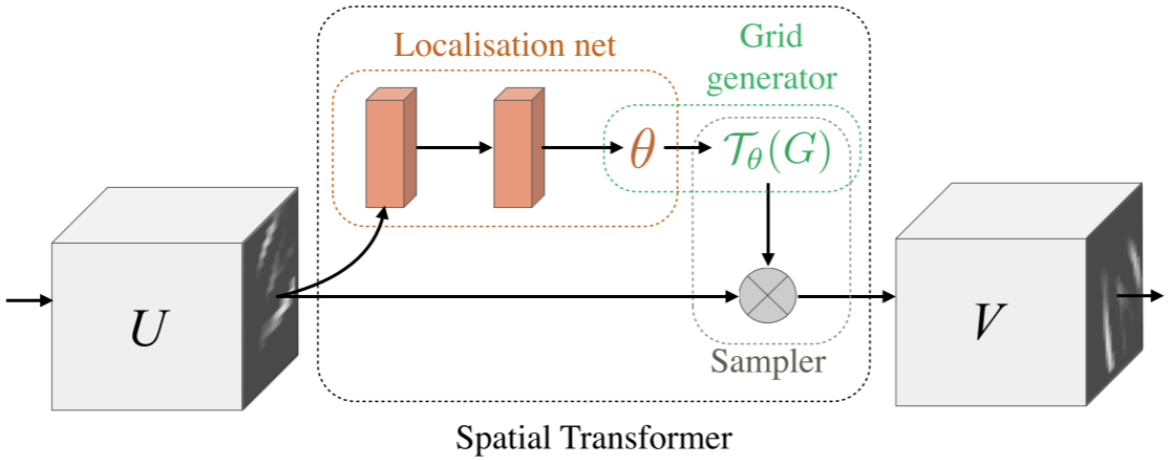
$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \ \mathbf{a}_i \in \mathbb{R}^D$$

$$e_{t,i} = f_{\text{att}}(a_i, h_{t-1})$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{L} \exp(e_{t,k})}$$

Attention

- STN(Spatial Transformer Networks)



通过为每个输入样本生成适当的变换来主动地对图像（或特征图）进行空间变换。然后在整个特征图（非局部）上执行转换，可以包括缩放、裁剪、旋转以及非刚性变形

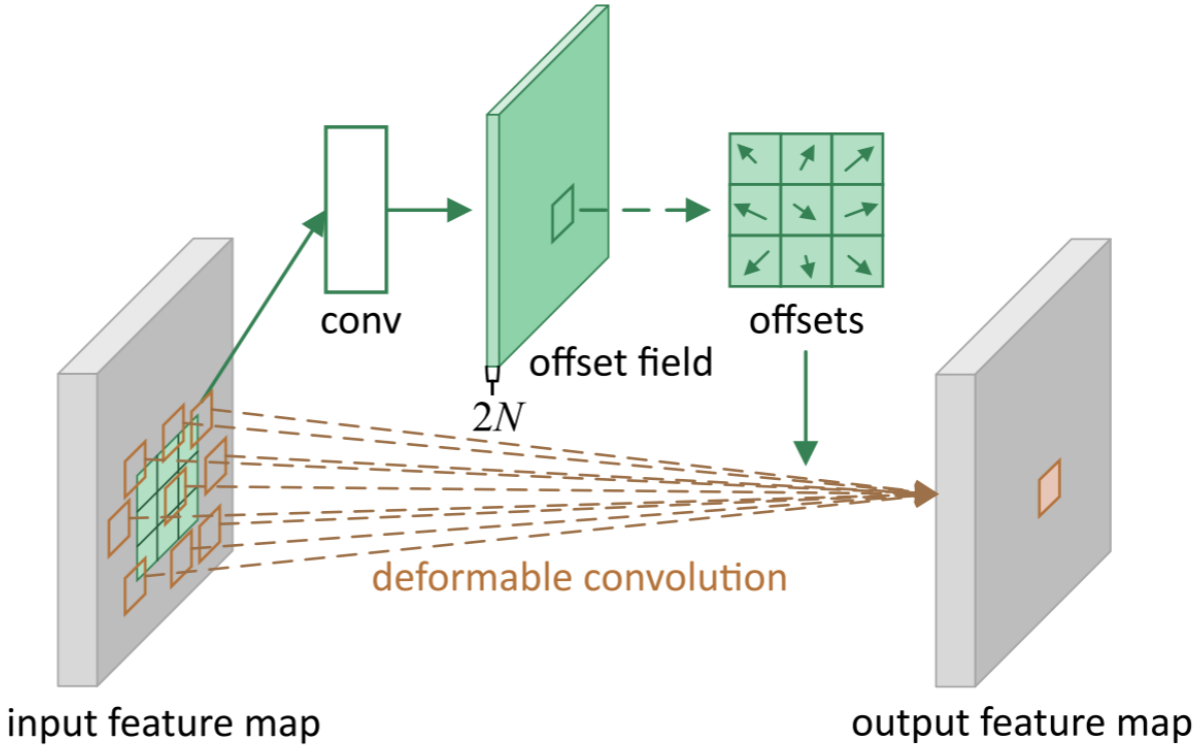
$$\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} = f_{\text{loc}}(U)$$

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}.$$

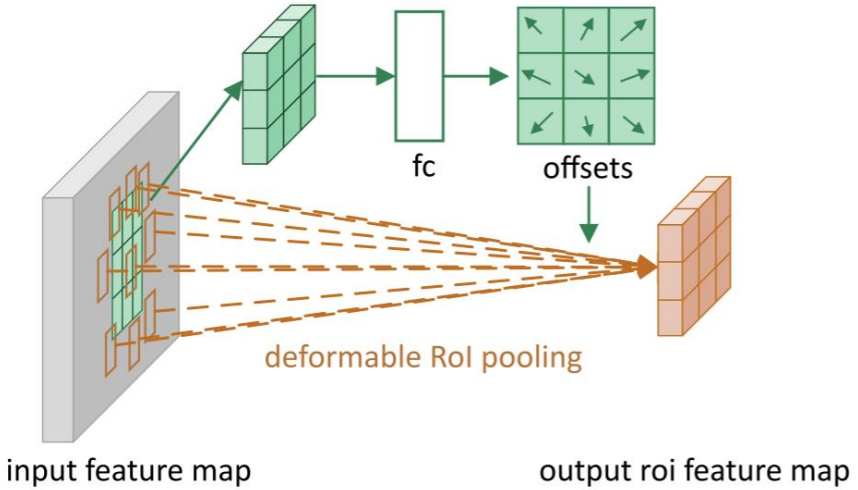
使用显式程序来学习平移、缩放、旋转和其他更一般的扭曲的不变性，使网络关注最相关的区域

Attention

- Deformable Convolutional Networks



基于使用额外偏移量增加模块中的空间采样位置并从目标任务中学习偏移量的想法，而无需额外的监督



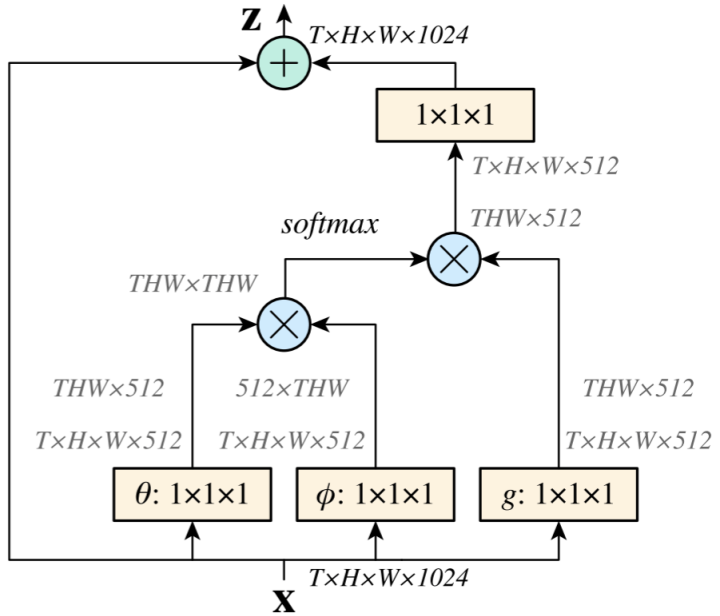
二维偏移添加到标准卷积中的常规网格采样位置。它使采样网格能够自由变形,偏移量是通过额外的卷积层从前面的特征图中学习的

Attention

- Self-attention and variants
- Non-Local Neural Network

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j).$$

$$f(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j).$$



通过计算特征图中每个空间点之间的相关矩阵来生成巨大的注意力图，然后注意力引导密集的上下文信息聚合

非局部操作将某个位置的响应计算为所有位置特征的加权和

优点:

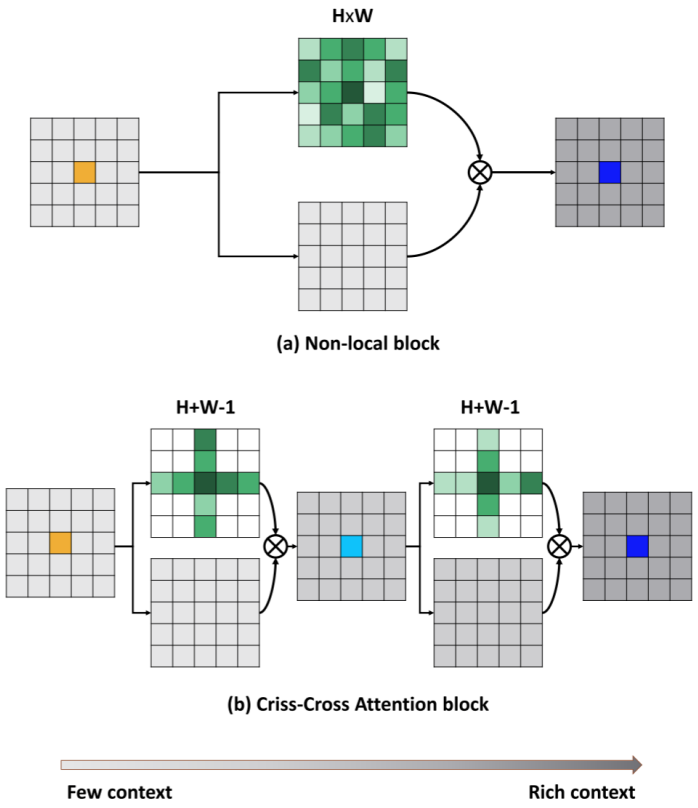
非局部操作通过计算任意两个位置之间的交互直接捕获远程依赖关系

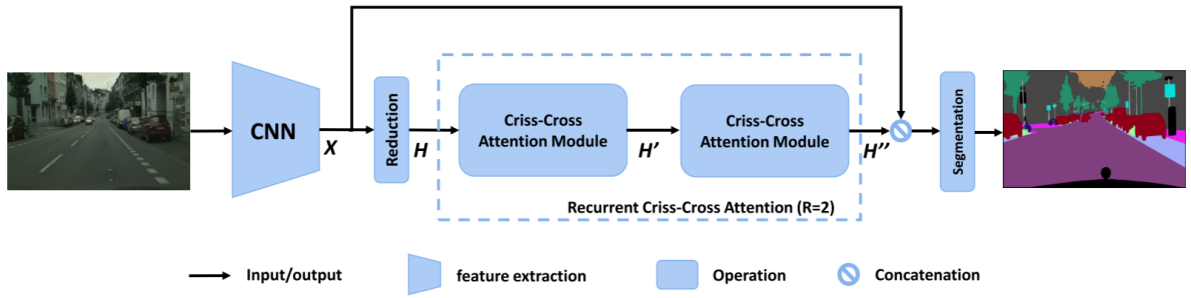
非局部操作是有效的，即使只有几层（例如 5 层）也能达到最佳效果

非局部操作保持可变的输入大小，并且可以很容易地与其他操作（例如，我们将使用的卷积）相结合

Attention

- CCNet



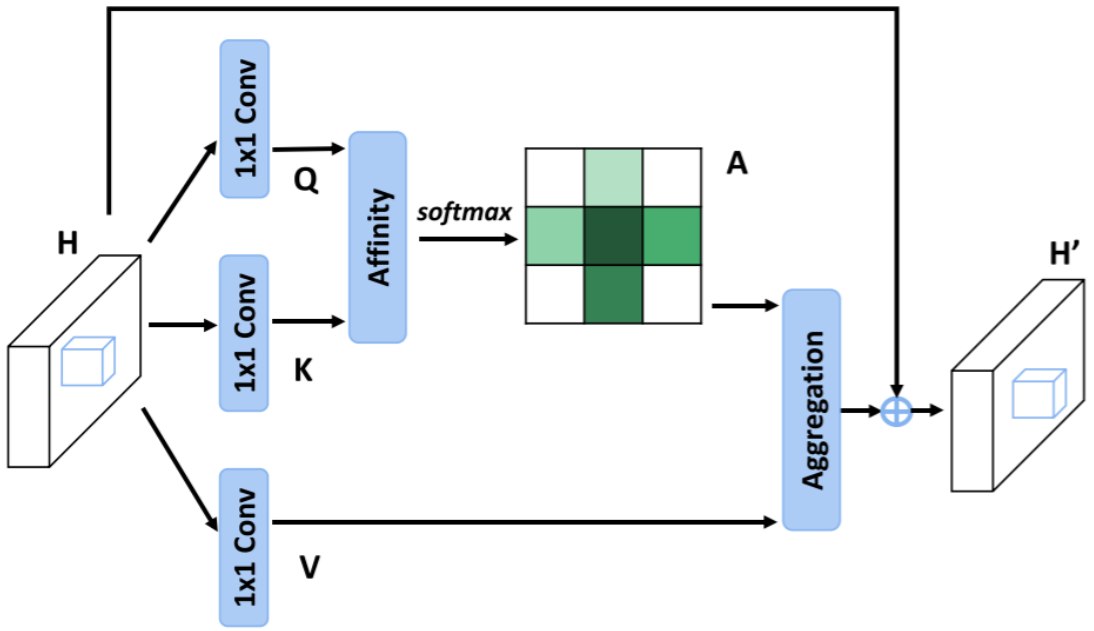


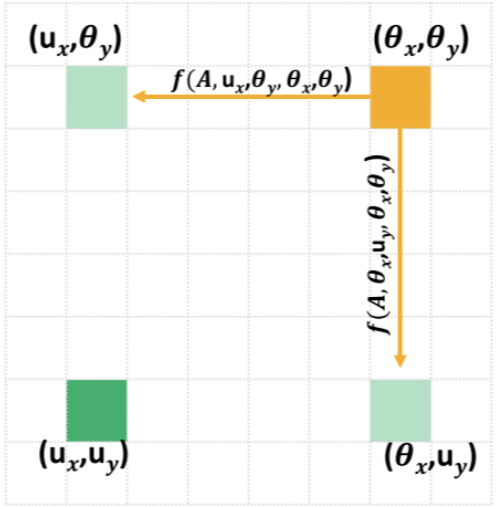
CCNet 可以通过一个新的交叉注意力模块在交叉路径上收集其周围像素的上下文信息。通过进一步的循环操作，每个像素最终可以捕获所有像素的远程依赖关系

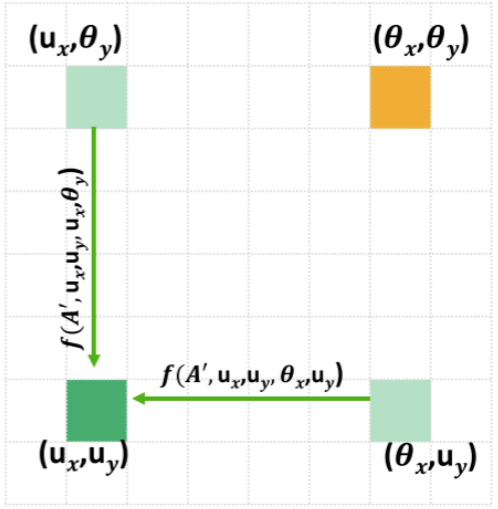
优点:

GPU 内存友好
计算效率高
最先进的性能

Attention



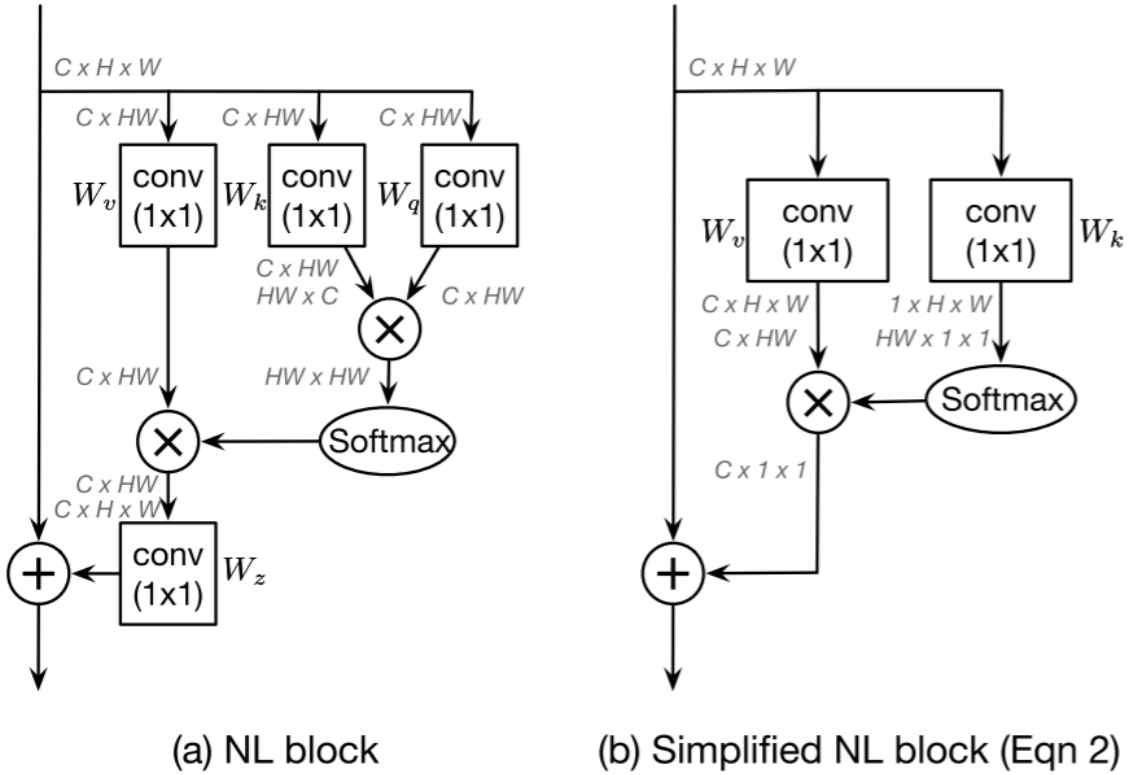




特征图中的每个位置通过自适应预测的注意力图与所有其他位置连接，从而收获各种范围上下文信息

Attention

- GCNet



(a) NL block (b) Simplified NL block (Eqn 2)

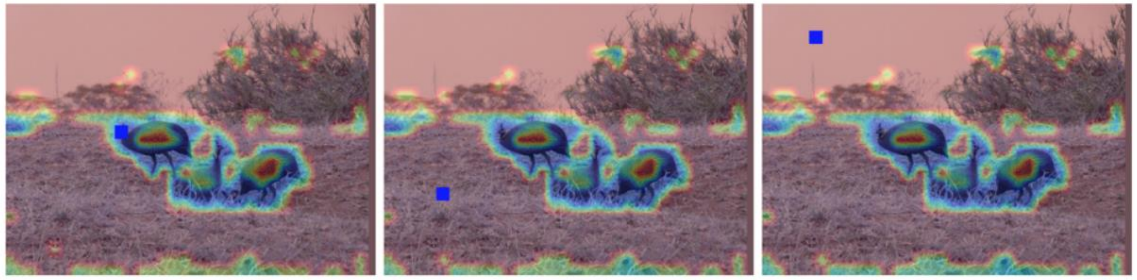
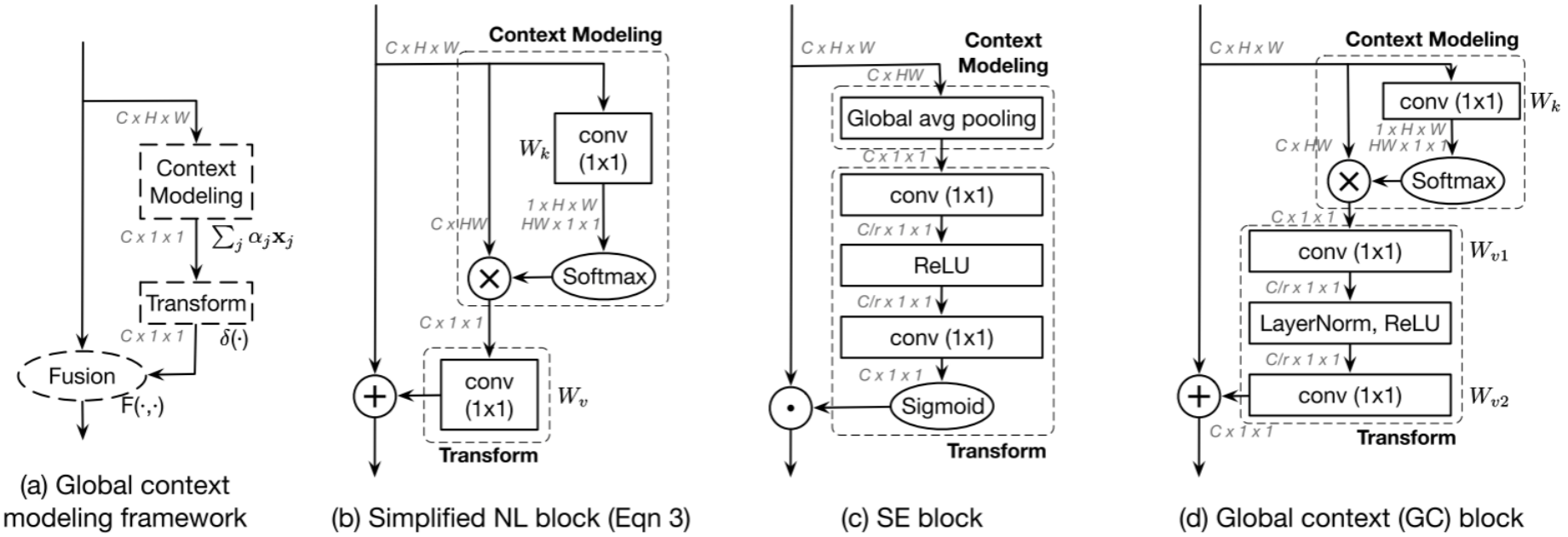


Figure 1: Visualization of attention maps (heatmaps) for different query positions (red points) in a non-local block on COCO object detection. The three attention maps are all almost the same. More examples are in Figure 2.

$$\mathbf{z}_i = \mathbf{x}_i + W_z \sum_{j=1}^{N_p} \frac{f(\mathbf{x}_i, \mathbf{x}_j)}{\mathcal{C}(\mathbf{x})} (W_v \cdot \mathbf{x}_j),$$

$$\mathbf{z}_i = \mathbf{x}_i + \sum_{j=1}^{N_p} \frac{\exp(W_k \mathbf{x}_j)}{\sum_{m=1}^{N_p} \exp(W_k \mathbf{x}_m)} (W_v \cdot \mathbf{x}_j),$$

Attention



(a) Global context modeling framework

(b) Simplified NL block (Eqn 3)

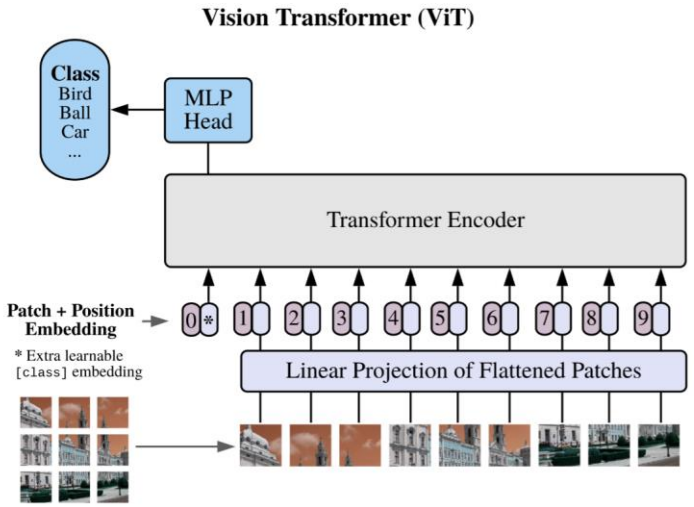
(c) SE block

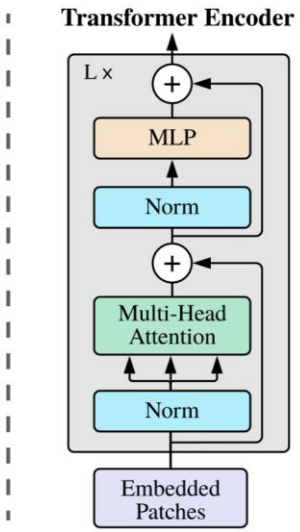
(d) Global context (GC) block

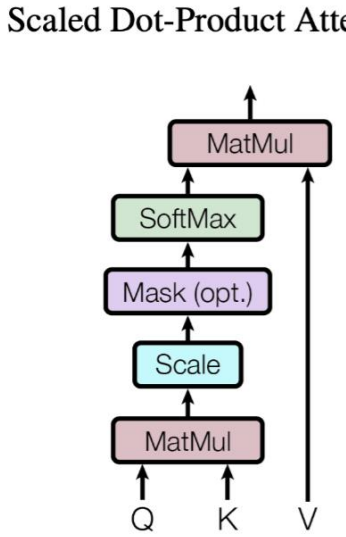
对所有查询位置显式使用与查询无关的注意力图来简化非局部块。然后我们使用这个注意力图将相同的聚合特征添加到所有查询位置的特征中，以形成输出.
结合SE和NL块优点，构建了GC块

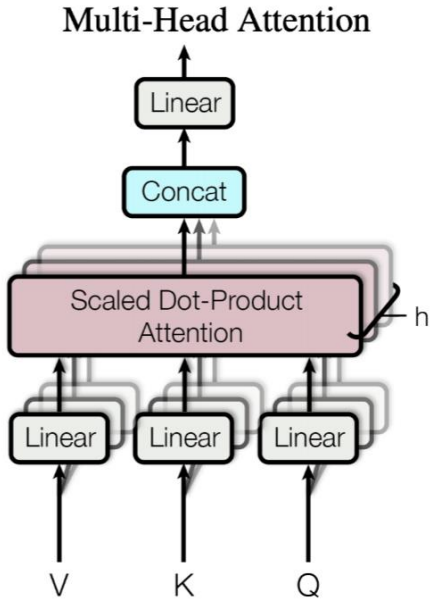
Attention

- ViT









Attention

Channel Attention

- SE Net
- GSop-Net
- SRM
- GCT

Spatial Attention

- RAM
- STN
- DCN
- Self-attention and variants
- ViT

Others

- Temporal attention
- Branch Attention
- Channel&Spatial Attention
- Spatial&Temporal Attention

Attention

- Temp Attention

  时间注意可以看作是一种动态的时间选择机制，决定何时注意，因此通常用于视频处理

- Temp Attention

  动态的分支选择机制

- Channel&Spatial Attention
  - DANet

$$Q,\quad K,\quad V = W_q X,\quad W_k X,\quad W_v X$$
$$Y^{\text{pos}} = X + V\text{Softmax}(Q^T K)$$
$$Y^{\text{chn}} = X + \text{Softmax}(X X^T) X$$
$$Y = Y^{\text{pos}} + Y^{\text{chn}}$$

- Spatial&Temporal Attention

Attention

Reference:

[1]L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "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, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 6298–6306. [Online]. Available: https://doi.org/10.1109/CVPR.2017.667

[2] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in Proceedings of the 27th International Conference on Machine Learning. Omnipress, 2010, pp.807–814.

[3] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[4] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi,and A. Agrawal, "Context encoding for semantic segmentation," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018.IEEE Computer Society, 2018, pp. 7151–7160. [Online].Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Context_Encoding_for_CVPR_2018_paper.html

[5] G. Zilin, X. Jiangtao, W. Qilong, and L. Peihua, "Global second-order pooling convolutional networks," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Attention

Reference：

[6] H. Lee, H.-E. Kim, and H. Nam, "Srm : A style-based recalibration module for convolutional neural networks," 2019.
[7] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition, 2020, pp.11 794–11 803.
[8] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," 2021.
[9] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh,J. Gall, and L. V. Gool, "Spatio-temporal channel correlation networks for action classification," 2019.
[10] Z. Chen, Y. Li, S. Bengio, and S. Si, "You look twice: Gaternet for dynamic filter selection in cnns," 2019.
[11] H. Shi, G. Lin, H. Wang, T.-Y. Hung, and Z. Wang, "Spsequencenet: Semantic segmentation network on 4d point clouds," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4574–4583.
[12] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," 2019.X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empiricalstudy of spatial attention mechanisms in deep networks," 2019.
[13] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
[14] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in International Conference on Computer Vision, 2019. [Online]. Available: http://arxiv.org/abs/1908.07678