ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746895

# GANET: UNARY ATTENTION REACHES PAIRWISE ATTENTION VIA IMPLICIT GROUP CLUSTERING IN LIGHT-WEIGHT CNNS

*Cheng Zhuang and Yunlian Sun**

School of Computer Science and Engineering, Nanjing University of Science and Technology

## ABSTRACT

The attention mechanism has been widely explored to construct a long-range connection which is beyond the realm of convolutions. The two groups of attention, unary and pairwise attention, seem like being incompatible as fire and water due to the completely different operations. In this paper, we propose a Group Attention (GA) block to bridge the gap between these two attentions and merely leverage unary attention to lightweightly reach the effect of pairwise attention, based on the implicit group clustering of light-weight CNNs. Compared with the conventional pairwise attention, i.e, Non-Local networks, our method artfully bypasses the burdensome pixel-pair calculation to save a huge computational cost, that is a big advantage of our work. Experiments on the task of image classification demontrate the effectiveness and efficiency of our GA block to enhance the light-weight models. Code will be released at https://github.com/ChiSuWq/GANet.

***Index Terms*—** Non-Local networks, attention mechanism, light-weight CNNs, image classification

## 1. INTRODUCTION

The attention mechanism manages to capture a long-range connection between distant features which is beyond the realm of conventional convolutions. Based on the attention operation, this mechanism can fall into two groups: unary attention and pairwise attention.

The unary attention[1, 2, 3, 4, 5] leverages the pooling method and convolutions over channel or spatial dimension to enlarge the receptive field thus enable the interdependency of features' global information. Typical examples include SE[1] and CBAM[2]. In general, an attention mask is required to lightweightly recalibrate the original feature map for the suppression of trivial features and highlight of informative ones.

The pairwise attention, whose representative work is Non-Local networks, instead measures the pixel-to-pixel relation and adaptively aggregates key pixels' features into the query one[6, 7, 8]. However, the calculation of the pixel-pair relation introduces a huge occupation of computational resources and GPU memory, resulting in the low efficiency of Non-Local Networks[8, 9].

However, due to the completely different processing of these two attentions, they seem like being incompatible. Motivated by this observation, here comes to our idea: ***Is there a way to bridge the gap between unary attention and pairwise attention, i.e., represent pairwise attention via unary attention?***

Thanks to the nature that the semantic entity is learned in a group-wise form in image classification[10], which we call "implicit group clustering" and meanwhile the form of clustering is unique for light-weight CNNs, we can bridge the gap between unary and pairwise attentions specially. As illustrated in Figure 1, although we query pixels about different apples which belong to the same semantic entity, the returned attention map of Non-Local method is nearly identical, which indicates all the local apples are trying to do the same thing: "how to represent myself more like the concept 'apple'?". Meanwhile, the concept is well described by the global information of this semantic entity. The similar observation can be likewise found in the lamp from Figure 1. Accordingly, the pairwise attention for all these pixels is shared in the group thus can be reached using simple unary attention. Note that we argue that the clustering in Figure 1 is dedicated for light-weight CNNs, for that the channels of these models are limited. Therefore, the learned group-wise entity is coarse and comprises of more subtle sub-entities which could only learned by the heavy networks, like ResNet[11].

Based on the above motivation, we propose our attention block, named as Group Attention (GA) block, for our method is implemented in a group-wise form. The GA block is comprised of two unary attention modules to enhance implicit group clustering and reach pairwise attention, respectively. Our GA block can be easily plugged into the existing light-weight CNNs[12, 13, 14, 15], named GANet, to equip CNNs with long-range connection.

The main contributions of our paper are as follows:

- We propose a Group Attention (GA) block for light-weight networks, which bridges the gap between unary and pairwise attention via implicit group clustering. To the best of our knowledge, this is the first work to realize the compatibility of these two distinct attentions.

* Yunlian Sun is the corresponding author (yunlian.sun@njust.edu.cn).

Thanks to the National Natural Science Foundation of China under Grant 62076131 for funding.

- We exploit the unary attention to reach pairwise attention. Consequently, the computation cost is significantly reduced from $\mathcal{O}((H \times W) \times (H \times W) \times C)$ to $\mathcal{O}((H \times W) \times (\frac{C}{G} \times G))$ due to the bypassing of pixel-pair calculation.
- Experiments on image classification demonstrate that GANet achieves a leading performance compared with other state-of-the art attention modules.

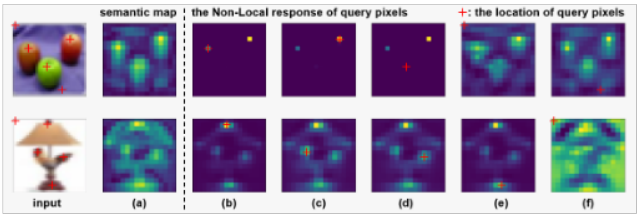


**Fig. 1**. The visualization of the conventional Non-Local module in a given group, i.e., in the same semantic entity. (a) refers to the current feature maps obtained by lightweight CNNs, e.g., Shufflenetv2[12] while the rests are pixel-pair attention maps for different query pixels. Surprisingly, we find that the pixels belong to the same semantic entity (the bright ones in (a)) have nearly the same response. Note that we also exhibit attention maps of the background pixels, which is obviously unlike those in foreground. Zoom in for a better view.

## 2. METHODS

### 2.1. Unary Attention for Unary Relation

Suppose $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ is the feature tensor in neural networks, $C$, $H$ and $W$ are the number of channels, the height and width of the feature, respectively. Then we divide $\mathbf{X}$ into $G$ groups along the channel dimension, named as $\mathbf{X}_i \in \mathbb{R}^{\frac{C}{G} \times H \times W}$, where $i$ refers to the $i$-th group.

In this section, our intention is to leverage unary attention for the enhancement of unary relation. Concretely, unary relation is regarded as the representation of each semantic entity corresponding to each group, i.e., implict group clustering. We adopt a simple yet effective $1 \times 1$ convolution to construct the unary attention mask $att_i^u \in \mathbb{R}^{H \times W}$ for each semantic group, followed by a normalization which is further introduced in Sec. 2.3:

$$att_i^u = \sigma(normalize(conv1 \times 1(\mathbf{X}))) \tag{1}$$

where $\sigma()$ is indicated as a sigmoid function. Note that the above $1 \times 1$ convolution exploits all channels, which means it considers all groups of semantic entities.

Then the original $\mathbf{X}_i$ is scaled by $att_i^u$ to obtain the enhanced feature $\mathbf{X}_i^u$:

$$\mathbf{X}_i^u = att_i^u \mathbf{X}_i \tag{2}$$

and all groups $\mathbf{X}_i^u$ form the entire feature map $\mathbf{X}^u$. To be specific, the above unary attention is served as a kind of spatial attention which suppresses irrespective pixels but highlights informative ones. Thus, the unary relation, i.e., the group-wise semantic entity is being more distinct.

### 2.2. Unary Attention for Pairwise Relation

Despite the fact that $\mathbf{X}^u$ has the benefit of unary relation, we argue that the learning of pairwise relation is still missing. To address the above issue, Non-Local networks come up with an inspiring idea that the relationship of each pixel-pair can be measured using pixel-to-pixel similarity. The global information of key pixels thus can be adaptively aggregated into each query pixel. However, this method introduces too much computational cost[8]. Thus, naturally here comes the question: ***How to save the computational cost as much as possible while keeping the pairwise relation?***.

Standing on our observation that each pixel in the identical entity of light-weight CNNs has the same response from the entire feature map, we argue that unary attention is able to reach pairwise attention via implicit group clustering. First, we need to generate the semantic entity. Considering that the semantic entity is the global information in the given group, the entity representation $\mathbf{y}_i^u \in \mathbb{R}^{\frac{C}{G} \times 1 \times 1}$ can be easily produced through a global weighted pooling with an attention mask $att_i^{pp} \in \mathbb{R}^{H \times W}$ generated by also a $1 \times 1$ convolution to decide how much the feature of each pixel contributes to the global information:

$$att_i^{pp} = \sigma_s(normalize(conv1 \times 1(\mathbf{X}_i^u))) \tag{3}$$

$$\mathbf{y}_i^u = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} (att_i^{pp})_{h,w} (\mathbf{X}_i^u)_{h,w} \tag{4}$$

where the superscript "$pp$" indicates that the attention is for the pooling process in the pairwise relation and $\sigma_s$ is a softmax function over the spatial dimension. Note that the generation of $att_i^{pp}$ have a convolution only considering features of the current group. Except that, the whole processing is completely the same as that in $att_i^u$.

After obtaining $\mathbf{y}_i^u$, we should decide how much $\mathbf{y}_i^u$ each pixel needs so that we can assign the semantic information to each pixel adaptively. Otherwise, if all pixels get access to the global information, irrelevant pixels (the value assumed to be 0) in this semantic group will be filled also with the vector of the corresponding entity, that is going to contaminate the spatial distribution of the semantic object. Thus, an attention mask $att_i^{pa} \in \mathbb{R}^{H \times W}$ is required to achieve the above target:

$$att_i^{pa} = \sigma(normalize(conv1 \times 1(\mathbf{X}_i^u))) \tag{5}$$

where $att_i^{pa}$ is obtained in the same way as $att_i^{pp}$ and "$pa$" means this attention is for the assigning process in the pairwise relation.

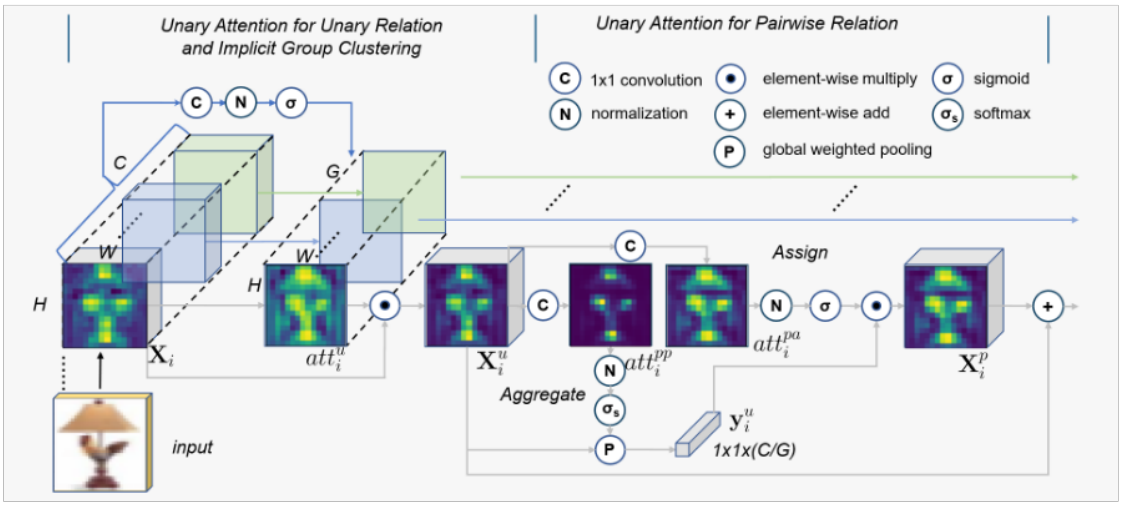


**Fig. 2**. The pipeline of our proposed GA block.

Then according to the attention mask, the semantic information $\mathbf{y}_i^u$ is adaptively allocated to every pixel for the final construction of pairwise relation $\mathbf{X}_i^p$:

$$\mathbf{X}_i^p = \mathbf{y}_i^u (att_i^{pa})_{h,w} \tag{6}$$

$$(\hat{\mathbf{X}}_i)_{h,w} = (\mathbf{X}_i^u)_{h,w} + \alpha_i \mathbf{X}_i^p \tag{7}$$

where $\alpha_i \in \mathbb{R}^{1\times1\times1}$ is a learnable parameter for each respective group to determine how much the pairwise relation should complement the unary relation. It is initialized as zero to support the learning of unary relation in the early training.

### 2.3. Normalization

For the reason that networks are trained on various images, we argue that diverse images might have biased magnitudes when generating the above three attention masks, therefore prohibiting the model from learning the positive mechanism. This is similarly mentioned in some previous works[10, 16]. Hence, we introduce an effective normalization scheme to stabilize the learning of our attention masks. To be concrete, given the attention feature $a_i \in \mathbb{R}^{H\times W}$, we normalize $a_i$ over the whole space:

$$\begin{aligned}(\overline{a}_i)_{h,w} &= \frac{(a_i)_{h,w} - \mu_i}{\sigma_i + \epsilon} \\ \mu_i &= \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}(a_i)_{h,w} \\ \sigma_i^2 &= \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}((a_i)_{h,w} - \mu_i)^2\end{aligned} \tag{8}$$

where $\epsilon$ (default=1e-5) is adopted for numerical stability. Then, an affine transformation with learnable parameters $\gamma$ and $\beta$ is applied to scale and shift the normalized value for the identity transform[17]:

$$\hat{a}_i = \gamma \overline{a}_i + \beta \tag{9}$$

The above processing is the whole content of $normalize()$, the formula used in Sec. 2.1 and 2.2. An ablation study about this normalization is later exhibited in the experiment section to demonstrate its validity.

## 3. EXPERIMENTS

### 3.1. Experiments on CIFAR100

The CIFAR100[18] dataset consists of 50k training and 10k testing images with a size of 32×32 pixels. There are 100 classes in total. We choose Shufflenetv2[12] as the baseline and several SOTA attentions for comparison. Unary attentions include SE[1], SGE[10] and CBAM[2] while pairwise attentions contain the conventional Non-Local (NL) module[6] and our specially designed group Non-Local (GNL) module considering the NL operation in a group-wise form. The reason why we choose Shufflenetv2 is that it has a suitable channel number for implicit group clustering and a Resnet-like structure which we can easily decide where to insert attention modules.

All the models expect for the Non-Local one are trained from scratch with a mini-batch size 64 for 200 epochs using SGD and a cosine shape learning decay with a learning rate from 0.1 to 0. The Non-Local module is carefully finetuned from the baseline with 100 epochs and 0.05 initial learning rate, due to the negative performance of training from scratch. Note that we do not perform downsampling in stage2 for clearer pixel-pair relation. All the attention modules are appropriately inserted into bottlenecks like SENet. The group number in SGE, GNL and GA is fixed to 15.

**Table 1**. Comparisons with state-of-the-art attention modules on CIFAR100 and ImageNet.

| | CIFAR100 | | | ImageNet | | |
|---|---|---|---|---|---|---|
| Backbone | Param. | MFLOPs | Top-1 Acc | Param. | MFLOPs | Top-1 Acc |
| Shufflenetv2 | 1.43M | 188.51 | 76.88% | 2.28M | 147.70 | 66.80% |
| SE-Shufflenetv2[1] | 1.65M | 189.39 | 77.41% | 2.47M | 156.45 | 68.08% |
| CBAM-Shufflenetv2[2] | 1.58M | 190.88 | 77.40% | 2.47M | 157.47 | 68.38% |
| SGE-Shufflenetv2[10] | 1.43M | 189.17 | 78.02% | 2.35M | 156.33 | 67.33% |
| NL-Shufflenetv2[6] | 1.43M | 259.28 | 77.56% | 2.35M | 192.78 | 67.26% |
| GNL-Shufflenetv2 | 1.43M | 259.28 | 78.22% | 2.35M | 192.78 | 68.11% |
| GA-Shufflenetv2 | 1.48M | 194.14 | 78.40% | 2.39M | 163.26 | 68.16% |

As illustrated in Table 1, the pairwise attention GNL surpasses all the unary attention modules. It indicates that a stronger long-range connection is modeled by pairwise attention than unary attention. Note that the NL module does not come up to expectations. We conjecture that it's because the NL module measures the pairwise relation along the whole channels which may entangle different semantic entities to prevent the learning of each other's pairwise attention, as similarly mentioned in the recent work[7]. However, the performance of pairwise attention is in a trade-off manner with a non-negligible computational complexity. To be concrete, the FLOPs of GNL-Shufflenetv2 is 259.28M while those of unary attention networks are about 190M.

On the contrary, our GA-Shufflenetv2 gives consideration to both performance and complexity. While reaching similar Top-1 accuracy to GNL-Shufflenetv2, our model's parameters and FLOPs are on par with unary attentions. Based on the above results, we demonstrate that the proposed GA block lightweightly bridges the gap between unary attention and pairwise attention.

Furthermore, we insert GA block into another lightweight model, Shufflenetv1[14]. The Top-1 accuracy of Shufflenetv1 is 75.76%. Compared with SGE(76.74%), GA presents a stronger performance(78.06%). It proves that our module is not specially designed for Shufflenetv2 but all the lightweight models with a suitable channel number.

### 3.2. Experiments on ImageNet

The ImageNet 2012 dataset includes 1.28 million images for training and 50k images for validation from 1,000 classes[19]. Like CIFAR100, we follow the standard training scheme[14] while train 100 epochs using a cosine shape learning decay with a learning rate from 0.1 to 0 and a batch size of 512.

As presented in Table 1, GA-Shufflenetv2 also has a leading performance in ImageNet dataset. It indicates that our module is not only dedicated to CIFAR100 but being generalized in the task of image classification. Even though the improvement from our GA block is close to those of SE and CBAM, we argue that GA block is a spatial attention which has no relation with channel attention. It means GANet could be further promoted with orthogonal equipment of the other attentions, like SE.

**Table 2**. Performance on CIFAR100 test set for different components of GA block. U4U and U4P refer to the unary attention for unary relation and pairwise attention, respectively.

| GANet | -U4P | -U4U | -Normalize | -GA |
|---|---|---|---|---|
| 78.40% | 78.10% | 77.90% | 77.86% | 76.88% |

### 3.3. Ablation Study

In order to give more insights into each component of the proposed module, here we perform a series of ablation studies on CIFAR100 based on GA-Shufflenetv2.

As elaborated in Sec. 2, GA block is comprised of two unary attentions for unary relation (U4U) and pairwise relation (U4P), respectively. To investigate the effect of each component, we remove U4U and U4P from GANet, respectively. Table 2 indicates that both unary relation and pairwise relation are positive for the feature enhancement while the integration of them, i.e., GANet, outperforms the separated ones. We conjecture that in GA block, the unary relation first suppresses irrelevant noises in the given group. It reduces the risk of mistaking trivial pixels as informative ones in the following pairwise relation stage, thus making this relation more easily learned. What's more, the normalization is proved to be explicitly essential for GA block by a considerable margin (78.40% vs 77.86%).

## 4. CONCLUSION

In this paper, we observe that the implicit group clustering of light-weight CNNs is unique and pixels belong to the same group-wise entity share the Non-Local response. Therefore, we propose a brand new attention block, Group Attention(GA) block, which merely leverages the unary attention to reach pairwise attention with saving a huge amount of computational complexity. Experiments on image classification demonstrate the efficiency and effectiveness of our method.

## 5. REFERENCES


[1] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *The IEEE conference on computer vision and pattern recognition(CVPR)*, 2018, pp. 7132–7141.

[2] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *The European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[3] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li, "Fcanet: Frequency channel attention networks," *arXiv preprint arXiv:2012.11879*, 2020.

[4] Wanli Chen, Xinge Zhu, Ruoqi Sun, Junjun He, Ruiyu Li, Xiaoyong Shen, and Bei Yu, "Tensor low-rank reconstruction for semantic segmentation," in *The European Conference on Computer Vision(ECCV)*. Springer, 2020, pp. 52–69.

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu, "Dynamic convolution: Attention over convolution kernels," in *The IEEE conference on computer vision and pattern recognition(CVPR)*, 2020, pp. 11030–11039.

[6] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *The IEEE conference on computer vision and pattern recognition(CVPR)*, 2018, pp. 7794–7803.

[7] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu, "Disentangled non-local neural networks," in *The European Conference on Computer Vision(ECCV)*. Springer, 2020, pp. 191–207.

[8] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *The IEEE conference on computer vision and pattern recognition(CVPR)*, 2019, pp. 603–612.

[9] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *The IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[10] Xiang Li, Xiaolin Hu, and Jian Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," *arXiv preprint arXiv:1905.09646*, 2019.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *The IEEE conference on computer vision and pattern recognition(CVPR)*, 2016, pp. 770–778.

[12] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *The European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *The IEEE conference on computer vision and pattern recognition(CVPR)*, 2018, pp. 4510–4520.

[14] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *The IEEE conference on computer vision and pattern recognition(CVPR)*, 2018, pp. 6848–6856.

[15] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019, pp. 9190–9200.

[16] Xu Ma and Song Fu, "Position-aware recalibration module: Learning from feature semantics and feature position.," in *The International Joint Conference on Artificial Intelligence(IJCAI)*, 2020, pp. 797–803.

[17] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning(ICML)*. PMLR, 2015, pp. 448–456.

[18] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *The IEEE conference on computer vision and pattern recognition(CVPR)*. IEEE, 2009, pp. 248–255.