applied sciences

MDPI

Article

Calibrated Convolution with Gaussian of Difference

Huoxiang Yang [1], Chao Li [1], Yongsheng Liang [1], Wei Liu [2] and Fanyang Meng [2,*]

[1] College of Electronic and Information Engineering, Shenzhen University, Shenzhen 518060, China; yhuoxiang@gmail.com (H.Y.); lichao2018@email.szu.edu.cn (C.L.); liangys@hit.edu.cn (Y.L.)
[2] Network Communication Center, Pengcheng Laboratory, Shenzhen 518055, China; liuwei@sziit.edu.cn
* Correspondence: mengfy@pcl.ac.cn; Tel.: +86-15814696029

Abstract: Attention mechanisms are widely used for Convolutional Neural Networks (CNNs) when performing various visual tasks. Many methods introduce multi-scale information into attention mechanisms to improve their feature transformation performance; however, these methods do not take into account the potential importance of scale invariance. This paper proposes a novel type of convolution, called Calibrated Convolution with Gaussian of Difference (CCGD), that takes into account both the attention mechanisms and scale invariance. A simple yet effective scale-invariant attention module that operates within a single convolution is able to adaptively build powerful scale-invariant features to recalibrate the feature representation. Along with this, a CNN with a heterogeneously grouped structure is used, which enhances the multi-scale representation capability. CCGD can be flexibly deployed in modern CNN architectures without introducing extra parameters. During experimental tests on various datasets, the method increased the ResNet50-based classification accuracy from 76.40% to 77.87% on the ImageNet dataset, and the tests generally confirmed that CCGD can outperform other state-of-the-art attention methods.

Keywords: convolutional neural network; scale-invariance; attention mechanism

Citation: Yang, H.; Li, C.; Liang, Y.; Liu, W.; Meng, F. Calibrated Convolution with Gaussian of Difference. *Appl. Sci.* **2022**, *12*, 6570. https://doi.org/10.3390/app12136570

Academic Editor: Federico Divina

Received: 21 May 2022
Accepted: 24 June 2022
Published: 29 June 2022

1. Introduction

Convolutional Neural Networks (CNNs) have proven to be effective at tackling a wide range of visual tasks [1,2]. To make features more robust and improve their performance, they often use attention mechanisms to suppress semantic noise, highlight regions with the correct semantic features and bring the network's operational character closer to human vision. Excellent attention-based networks often have strong feature transformation capabilities, thus providing more accurate representations that can support specific tasks. This has made research into how to enhance the feature extraction capability of attention-based networks an important topic in visual computing in recent years.

A number of different attention methods have been proposed with the goal of achieving more effective feature transformation [2–10]. These methods can be broadly categorized into two types: single-scale and multi-scale. SENet [3] provides a powerful, lightweight screening mechanism that can self-recalibrate the feature distribution via channel-wise interdependencies. Moving beyond a focus on channels, BAM [4] and CBAM [5] use spatial importance in a similar way for the calibration index. Later works [6–8] extend this even further by designing advanced attention blocks. However, single-scale attention methods lack the ability to make full use of self-similarity and structural information at different levels. This results in the attention regions being less than ideal (note, for instance, that in the third column of the first row in Figure 1, the activated region of nails represents only a small proportion of the overall group). One of the simplest ways to tackle this problem is to introduce multi-scale structures in the attention design. Typical examples, such as SKNet [9] and its variants [10,11], use multi-scale information to generate more responsive attention maps, given the same structural constraints. This produces compelling results when used for image classification tasks (it can be seen in the fourth column in Figure 1 that

SK has a stronger ability to represent features than the single-scale method, SE). However, while such methods are effective in carrying out a multi-scale feature fusion, they seldom touch the distinctive importance of invariance in the scale space. Still, scale-invariant features can be of crucial importance in a wide variety of visual tasks.

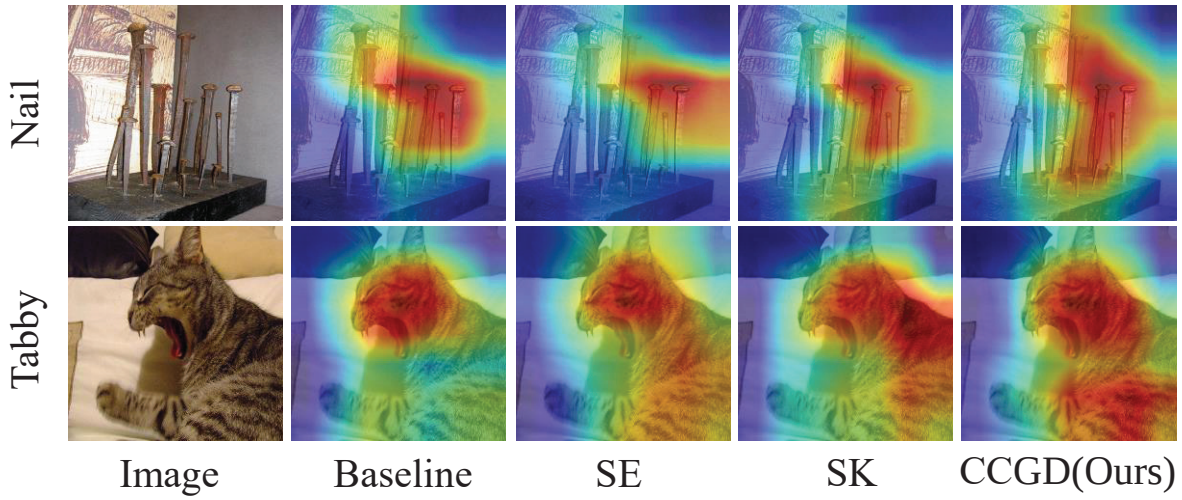


Figure 1. Heatmaps for different learning models using Grad-CAM [12]. All the selected models were based on ResNet50 [13] and trained on ImageNet [14]. It can be seen that CCGD offers more comprehensive region identification.

In view of the above issues, we have developed a novel convolution operator called Calibrated Convolution with Gaussian of Difference that unifies scale-invariance and attention mechanisms. The focus of the method is upon collecting scale-invariant information for the execution of self-calibration. This generates more discriminative feature representations by explicitly incorporating more accurate information. As is shown in Figure 1, CCGD is better to extract comprehensive salient regions than traditional convolutions or other attention mechanisms, no matter what shape they are. Figure 2 shows a structural diagram of CCGD. It is composed of two paths. One is a scale-invariant attention module that can transform a feature map into a pixel-level attention map. This attention map is used as a reference to guide the feature transformation process by means of its own invariant feature information. The other path is an original scale path. Here, feature maps are only executed as convolution operations to retain the original scale features and reduce the number of parameters. These two paths form a heterogeneously grouped structure that can aggregate contextual information at different levels. The work presented here offers the following contributions:

- A novel convolution operator, called Calibrated Convolution with Gaussian of Difference (CCGD), which introduces scale-invariant information into the attention mechanism and has a heterogeneously grouped structure. CCGD is especially well-suited to feature transformation because of its remarkable capacity for self-calibration and multi-scale representation.
- The capacity to replace traditional convolution and plug-and-play methods in modern CNNs to form CCGDNet without introducing any extra parameters. This enables them to have a more robust representation learning capability.

Comprehensive experiments were undertaken that demonstrate that our approach outperforms most methods on both the Cifar100 [15] and ImageNet [14] datasets when adopting the same standard training strategies.

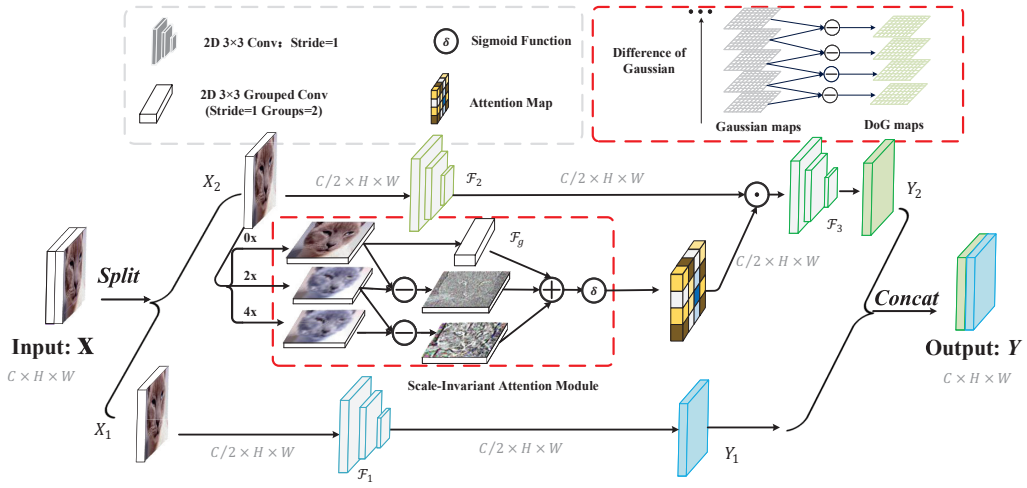


Figure 2. Structural diagram of the proposed Calibrated Convolution with Gaussian of Difference.

2. Related Works

As our work is focused on improving the feature representation learning capacity of CNNs by using a scale-invariant attention mechanism, we first review the related work on deep learning methods for network architecture design. We then look at existing attention mechanisms for CNNs.

2.1. Architecture Design

Network architecture design is a long-standing concern in CNNs. The particular interest is the need for an efficient backbone module that can be used to enhance the network performance across different tasks, such as accurate classification of the image features in datasets. Pioneering network architecture designs include AlexNet [1], VGGNet [16], GoogleNet [17], ResNet [13], and DenseNet [18]. In these works, ResNet is the most widely used, with many researchers adopting it as a prototype network for the design of applications, e.g., WideResNet [19] and ResNext [20], which can improve network performance and convergence speed by increasing its width or by introducing lightweight 3×3 grouped convolutions. Res2Net [21] and HS-ResNet [22] build upon ResNet by having a fine-grained hierarchical split and concatenated connections within a single residual block. This makes it possible to build very deep and robust networks. Apart from ResNet-based improvements, there are augmented versions of DenseNet, such as CondenseNet [23], which can take advantage of dense connection mechanisms. Meanwhile, the NSA [24] series allows for the accumulation of predefined search spaces that can be incorporated into search strategies by means of having a network architecture that can vary according to performance estimates.

2.2. Attention Mechanisms

Attention mechanisms aim to suppress semantic noise and highlight effective information by building various long-range dependencies. Numerous attention-based methods have contributed to the success of deep learning in the field of computer vision. These methods can be broadly categorized into two categories according to the scale structure they use as the basis of the attention mechanism: single-scale attention and multi-scale attention.

Single-scale attention: Here, a single original scale feature map is used to generate the attention map. In an early version of this, SENet [3] adaptively recalibrated channel-wise feature responses by explicitly modeling single scale feature interdependencies between channels. Working more broadly from a spatial dimension, BAM [4] and CBAM [5] similarly used both channel-wise interdependencies and spatial locations for the purposes of reweighting. This can yield a better performance while using the same parameters. Other approaches, such as SRM [6], have sought to model style information dependencies and incorporate them into the attention block. However, single-scale attention methods

are generally limited by their inability to make full use of self-similarity information and structural information at different levels, leading to sub-optimal results.

Multi-scale attention: Here, a multi-scale feature set is used to generate the attention map. Thus, SKNet [9] has a built-in dynamic kernel selection mechanism that is guided by a combination of features from different multi-scale convolutions (3×3 and 5×5). This enhances network performance by having a smaller number of additional parameters and a lower computational complexity. SPA [10] introduces structural regularization in the form of spatial pyramid blocks that can also incorporate structural information. This is then used to build a more comprehensive multi-scale attention map. The recently developed SCNet [11] adaptively builds multi-scale spatial and inter-channel dependencies around each spatial location by means of a novel self-calibration operation. HMSA [25] uses a multi-scale attention module that learns to predict a dense mask that can combine multi-scale predictions. This not only improves performance in comparison to average-pooling but also allows the network to diagnostically visualize the importance of different scales for classes and scenes.

Most existing multi-scale attention methods have evolved from multi-scale inference for computer vision. As a result, they can effectively collect self-similarity information and structural information contained in features at different scales. However, most methods combine the multi-scale information by using simple addition operations, which introduces redundancy. Furthermore, the scale-invariant features contained in the multi-scale information are overlooked despite their importance to the feature calibration process.

3. Methods

3.1. Motivation

Before presenting details of our proposed method, we will give a brief introduction to the Difference of Gaussians (DoG) method present in the SIFT [26] algorithm to clearly establish the motivation underlying this paper.For any given general expression of an image $f(x, y)$, the goal of the DoG is to establish a large number of stable distinctive matching points, so as to improve geometric estimation. The DoG assumes that subtractive expansion can take the following form:

$$\begin{aligned} DOG &\triangleq G_{\sigma_1}(x, y) - G_{\sigma_2}(x, y) \\ &= \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sigma_1} e^{\frac{-\left(x^2+y^2\right)}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{\frac{-\left(x^2+y^2\right)}{2\sigma_2^2}} \right) \end{aligned} \tag{1}$$

where, $G(\cdot)$ is the Gaussian filtering function, and σ represents the difference smoothing parameter. As a result of this subtractive mechanism, the DoG method can extract the distinctive stable and invariant features of complex objects. This plays a key role in classical matching algorithms.

Inspired by the DoG, we incorporate invariant distinctive information into a pixel-level attention module. We argue that it is the DoG-based feature descriptors in our innovative attention module that enable it to take full advantage of distinctive invariant information. In the following section, we describe our CCGD in detail.

3.2. Scale-Invariant Calibrated Convolution

A structural diagram of the proposed method is shown in Figure 2. As with grouped convolution, CCGD divides the standard convolutional filters into multiple parts $[\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_g]$. Here, however, each filter is not positioned in parallel, but rather made responsible for executing a particular function.

An input, X, is evenly divided into two subsets, X_1 and X_2, each of which then enters a designated pathway for learning feature representations at a different level. The scale-invariant attention pathway is further subdivided into two sub-paths. The first component of the top path is a 3×3 convolution. Meanwhile, the X_2 features are sent along a bottom

path to execute a Scale-Invariant Calibration (SIC) operation. Motivated by the DoG, this module contains continuous down-sampling and up-sampling operations, $S_r(\cdot)$, with different rates, to mimic the Gaussian filtering in Equation (1). This minimizes the extra computational load and parameters. A Gaussian blur approximation, G, is then obtained, which has the same resolution as the input, X_2 . This can be written as follows:

$$G = \{G_i\}_{i=1}^{n} = S_r(X_2) \tag{2}$$

where n denotes the number of iterations for $S_r(\cdot)$. The difference operation on the bottom path is performed upon G as follows:

$$D = \sum_{i=1}^{n-1} (G_{i+1} - G_i)^2 \tag{3}$$

where D is a set of difference maps. To avoid any possible differentiation problems in the back-propagation and to enhance the effect of the difference maps, we use a square function to replace the absolute value operation. The calibrated operation can therefore be defined as follows:

$$X_2' = \mathcal{F}_2(X_2) \odot \delta\big(D + \mathcal{F}_g(X_2)\big) \tag{4}$$

where δ represents the sigmoid activation, and $\odot$ denotes element-wise multiplication. A 3×3 group convolution, $\mathcal{F}_g(\cdot)$, is used to introduce a learnable feature region as an extra replenishment.

For the attention mechanism in CNNs, the mathematical description can be expressed as follows:

$$X' = \mathcal{F}(X) \odot AM(X) \tag{5}$$

where $AM(\cdot)$ represents the attention map, which is used to assist in the calibration of the output features of a convolution. Correspondingly, $\delta\big(D + \mathcal{F}_g(X_2)\big)$ is the attention map that calibrates the output features of $\mathcal{F}_2(\cdot)$. Where, the DoG is a key technology for generating robust attention maps. Then, the final output after CCGD can be formulated as follows:

$$Y = Y_2 \cup Y_1 = \mathcal{F}_3\big(X_2'\big) \cup \mathcal{F}_1(X_1) \tag{6}$$

The advantages of CCGD are three-fold. First, by utilizing the calibrated steps defined in Equation (4), each feature is allowed to adaptively build powerful pixel-level dependencies on the basis of its scale-invariant information, enabling recalibration of the intermediate feature representation. Second, as $\mathcal{F}_2(\cdot)$ and $\mathcal{F}_3(\cdot)$ are combined into series form, the receptive field of the scale-invariant attention pathway can be expanded explicitly, thus providing a feature map that is different from the original pathway, $\mathcal{F}_1(\cdot)$, in the scale space. This gives CCGD a strong multi-scale representation ability. Third, instead of introducing added parameters, the scale-invariant calibration operation only employs lightweight continuous down-sampling, up-sampling, and 3×3 grouped convolutions. The deeper the network, the more significant this becomes.

3.3. Instantiations

As indicated above, CCGD can enhance the ability of CNNs to accurately represent features. To further investigate the viability of the proposed method, we have developed what we term CCGDNet by incorporating CCGD into ResNet. All 18-layer, 50-layer, and 101-layer bottleneck structures were used as the trunk for the network. Here, we simply replaced the 3×3 convolutional layer in each building block with the proposed CCGD (see Figure 3), while leaving other relevant hyperparameters intact. By default, the number of different maps in our CCGD is set to 2. This will be further explained in Section 4.3.

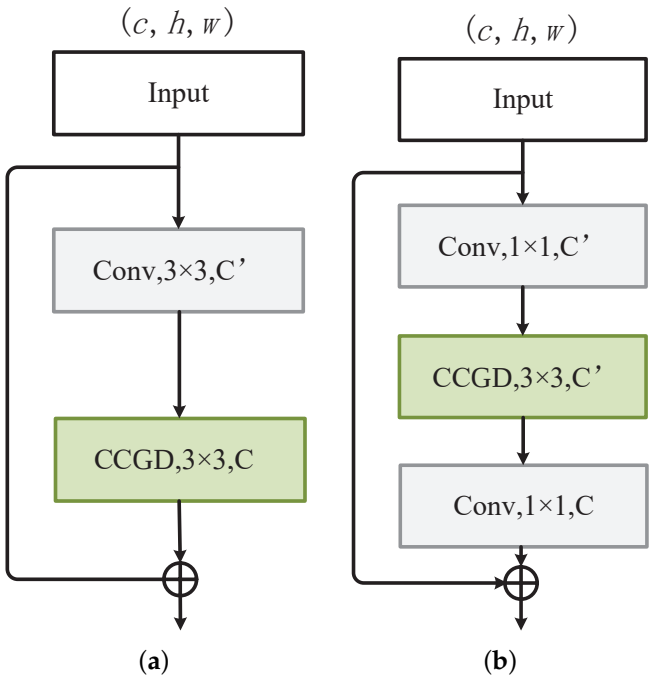


Figure 3. CCGD integrated with different residual block in ResNet: (**a**) basicblock; (**b**) bottlenck block.

4. Experiments

4.1. Implementations

We conducted detailed experiments using the generic PyTorch framework. For fairness of comparison, we implemented all of the classification experiments by adopting the same standard practice [13]. For the input, we used the Cifar100 [15] and ImageNet [14] datasets. In the case of ImageNet, we used SGD to optimize all the models. The momentum and weight decay were initialized at 0.9 and 5×10^{-4}, respectively. We executed all of the methods tested for 100 epochs with a 0.1 initial learning rate. This was attenuated by a factor of 10 after every 30 epochs. For the Cifar100 data, the batch size was set to 128, and the training was conducted over 200 epochs. The optimization strategy was similar to the one used for ImageNet. By default, all tasks were performed on a server with an 8 GTX1080Ti card and 256 GB of RAM.

4.2. Results for the Cifar100 Datasets

We first integrated the proposed CCGD into ResNet [13] and its variant, ResNetXt [20], to form CCGDNet, and then explored the performance gain for the Cifar100 dataset. After that, we conducted extra experiments to compare CCGDNet against other state-of-the-art attention-based methods. As with our prior instantiations, we only replaced the original 3×3 convolution with CCGD. The results in Figure 4 show that our CCGD+ResNet outperforms ResNet by a significant margin. For instance, CCGD+ResNet18 outperforms ResNet50 by a margin of 1.63%, and CCGD+ResNetXt further obtains the considerable performance improvements. The same trend can be observed for deeper networks. Table 1 shows the top-1 accuracy for the various attention model when applied to the ResNet. When compared with other state-of-the-art attention-based methods, we can see that CCGDNet achieved more advanced performance improvements. For instance, SENet achieved an accuracy of only 80.18%,whereas the baseline with CCGD achieved an accuracy of 81.36%, an increase of 1.18%. It is interesting to note that the advantage of this model not only improves network performance but also reduces the number of parameters, e.g., SENet50 (26.24 M) vs. CCGDNet50 (22.30 M), an obvious decrease of 15%.

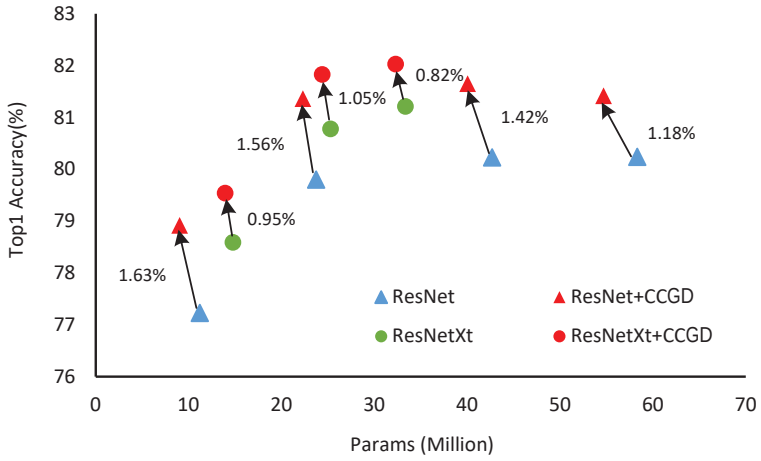


Figure 4. The results of using CCGD on Cifar100 dataset with ResNet and ResNetXt backbone.

Table 1. The results of using different attention-based methods on Cifar100 dataset with ResNet backbone. Results in bold are the best in each column with the same backbone. ↑ means the increase from baseline.

Methods	Backbone	Params	Flops	Top-1 (%)
ResNet (Baseline)	ResNet18	11.22 M	0.56 G	77.56
SE [3]		11.31 M	0.57 G	78.11
BAM [4]		11.25 M	0.57 G	77.77
CBAM [5]		11.31 M	0.57 G	77.94
SKNet [9]		11.55 M	0.57 G	78.76
GC [8]		11.32 M	0.57 G	77.92
SRM [6]		11.23 M	**0.57 G**	78.05
SC [11]		9.83 M	0.60 G	78.50
CCGD(Ours)		**9.04 M**	0.66 G	**79.19** (↑ 1.63)
ResNet (Baseline)	ResNet50	23.76 M	1.31 G	79.80
SE [3]		26.24 M	1.31 G	80.18
BAM [4]		26.28 M	1.31 G	80.24
CBAM [5]		26.28 M	1.31 G	80.45
SK [9]		24.03 M	1.33 G	80.84
GC [8]		26.28 M	1.32 G	79.97
SRM [6]		23.76 M	1.31 G	80.34
SC [11]		23.71 M	**1.25 G**	80.47
CCGD(Ours)		**22.30M**	1.36 G	**81.36** (↑ 1.56)

Table 2 shows the top-1 accuracy for the various models when applied to the Cifar100 dataset. When compared to the original backbone networks, our method improved the performance of all baseline models. In the case of ResNetXt, the performance increased from 80.78% to 81.83%, a 1.05% gain that confirmed the superiority and generalization ability of our method for general backbone networks. When compared with other state-of-the-art attention-based methods, we can see that CCGDNet achieved the best performance in several scenarios, but not in all. SE, for instance, outperformed CCGD in the case of DenseNet. This is contrary to what we would have expected. As already mentioned, a heterogeneously grouped structure is used for CCGD composition. This boosts its multi-scale representation ability. However, DenseNet and CCGD both use heterogeneously grouped structures and feature reuse mechanisms, both of which could be reformulated as a homogeneous strategy. This may, therefore, lead to over-fitting on small-scale datasets.

Table 2. Performance comparisons of different models on the Ciafr100 dataset. Results in bold are the best in each column.

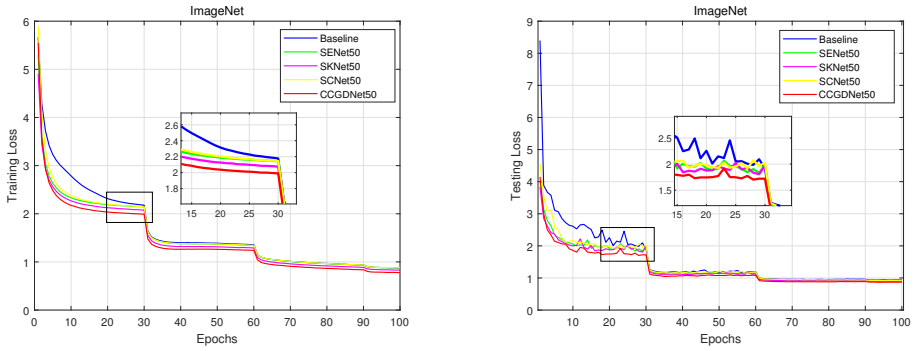
Methods	Backbone		
	ResNetXt	**DenseNet**	**DLANet**
Baseline	80.78	80.24	80.43
SE	81.40	**80.66**	-
BAM	81.15	80.48	-
SK	-	80.24	80.79
SGE	81.35	80.08	80.74
CCGD(Ours)	**81.83**	80.53	**81.10**

4.3. Results for the ImageNet Datasets

Next, we explored whether the superior performance of our method could be generalized to other datasets apart from Cifar100. Comparative experiments were therefore carried out using the ImageNet dataset. Due to space limitations, we only report here the detailed metrics for ResNet [13]. Table 3 shows that, when compared to other attention methods, CCGDNet achieved a better performance while using fewer parameters, i.e., for SENet50, BAMNet50, and SRMNet50, the top-1 accuracies were 0.76%, 0.97%, and 0.74% better, respectively. Clear advantages were also visible in the loss curves (see Figure 5). We also explored the effect of using different network depths on the ImageNet dataset. As can be seen in Table 3, as the depth increased, the performance of CCGDNet steadily surpassed the other competitors, given the same configuration. These results demonstrate the effectiveness of CCGD for complex datasets.

Table 3. The results of using different attention-based methods on ImageNet dataset with ResNet backbone. Results in bold are the best in each column with the same backbone. ↑ means the increase from baseline.

Methods	Backbone	Years	Parameters	FLOPS	Top-1 (%)	Top-5 (%)
ResNet (baseline)	ResNet-18	CVPR-16	11.69 M	1.82 G	69.83	89.10
SE [3]		CVPR-18	11.78 M	1.82 G	70.86	89.78
BAM [4]		BMVC-18	11.71 M	1.83 G	71.12	89.99
CBAM [5]		ECCV-18	11.78 M	**1.82 G**	70.73	89.91
TA [27]		WACV-21	11.69 M	1.83 G	71.19	89.99
CCGD (Ours)			**9.16** M	2.05 G	**71.22** (↑ 1.39)	**90.03** (↑ 0.93)
ResNet (baseline)	ResNet-50	CVPR-16	25.56 M	4.12 G	76.40	92.94
SENet [3]		CVPR-18	28.08 M	4.13 G	77.11	93.40
BAM [4]		BMCV-18	25.92 M	4.21 G	76.90	93.66
CBAM [5]		ECCV-18	28.09 M	4.13 G	77.34	93.69
GALA [28]		ICLR-19	29.40 M	-	77.27	93.66
GC [8]		CVPR-19	28.10 M	4.13 G	77.70	93.76
SK [9]		CVPR-19	26.15 M	4.19 G	77.54	93.62
SRM [6]		CVPR-19	25.62 M	4.12 G	77.13	93.70
SC [11]		CVPR-20	25.60 M	**3.95 G**	77.52	93.78
TA [27]		WACV-21	25.56 M	4.17 G	77.48	93.68
CCGD (Ours)			**24.15** M	4.27 G	**77.87**(↑ 1.47)	**93.95** (↑ 1.01)
ResNet (baseline)	ResNet-101	CVPR-16	44.46 M	7.84 G	78.20	93.91
SE [3]		CVPR-18	49.32 M	7.86 G	78.46	94.10
BAM [4]		BMVC-18	44.91 M	7.93 G	78.46	94.02
CBAM [5]		ECCV-18	49.33 M	7.86 G	78.49	94.31
SK [9]		CVPR-19	45.68 M	7.98 G	78.79	94.26
SRM [6]		CVPR-19	44.68 M	7.85 G	78.47	93.75
SC [11]		CVPR-20	44.56 M	**7.20 G**	78.60	93.98
TA [27]		WACV-21	44.56 M	7.95 G	78.03	93.77
CCGD (Ours)			**41.89 M**	-	**78.84**(↑ 0.64)	**94.32**(↑ 0.41)



(**a**) (**b**)

Figure 5. The loss curves of using different attention-based methods on ImageNet dataset with ResNet50: (**a**) training loss; (**b**) testing loss.

4.4. Ablation Analysis

Experiments were also conducted to assess the contribution of different aspects of the proposed CCGD. As described in Section 3.2, we developed an SIC operation and grouped structure to enhance the representation learning capacity. It can be seen in Table 4 that, when no SIC operation was used, the result was already much better than it was for the original ResNet-50. This demonstrates the validity of the grouped structure. When the SIC operation was added, there was a further performance gain of 0.8%. We also examined the effectiveness of the specific components of the SIC operation. Here, the results show that, when the other configurations were kept constant, adopting a square function or extra replenishment yielded a performance gain of about 0.19% or 0.34%, respectively. This confirmed the validity of our original reasoning. In the SIC module, the number of differential maps controls the intensity of the calibration mechanism. It can be seen in Table 5 that, when the differential number was set to two, the performance gain had already reached the peak value, which slowly declined during the next stage. We believe that an excessive zoom factor will destroy an image or result in a loss of feature information during the continuous down-sampling and up-sampling operations. This, in turn, will affect the quality of the calibration weight.

Table 4. Ablation experiments on ImageNet dataset. Result in bold is the best.

Methods	Design Choice			Top1-Acc
	SIC Operation	FR	Square Function	
ResNet	-	-	-	76.40
CCGD	-	-	-	77.07
	✓	-	-	77.56
	✓	✓	-	77.68
	✓	-	✓	77.53
	✓	✓	✓	**77.87**

Table 5. The results of using different numbers of difference maps on ImageNet dataset with ResNet-50 backbone. Results in bold are the best in each column.

Methods	Number	Top1-Acc (%)	Top5-Acc (%)
CCGD	0	77.10	93.22
	1	77.43	93.25
	2	**77.87**	**93.95**
	3	77.84	93.84
	4	77.63	93.80

4.5. Visualization with Grad-CAM

To further examine the feature representation ability of the CCGD, we used Grad-CAM to observe the effectiveness of the class activation mapping. All of the models were based on ResNet50 and trained on ImageNet. In Figure 6a, we can see that, when compared with SE, the CCGD activation mapping covered more regions of the relevant objects, such as the 'Hare' and the 'Ballpoint'. SE, by contrast, only managed to cover part of the object. CCGD is also able to reduce the influence of background noise, so that the activation map is better centered on the target, as can be seen in the case of the 'Airliner' image. Similar phenomena can be observed at different stages (see Figure 6b). In conclusion, the proposed CCGD is able to improve the accuracy of the network by strengthening the representational power of the feature map and its robustness against noise.

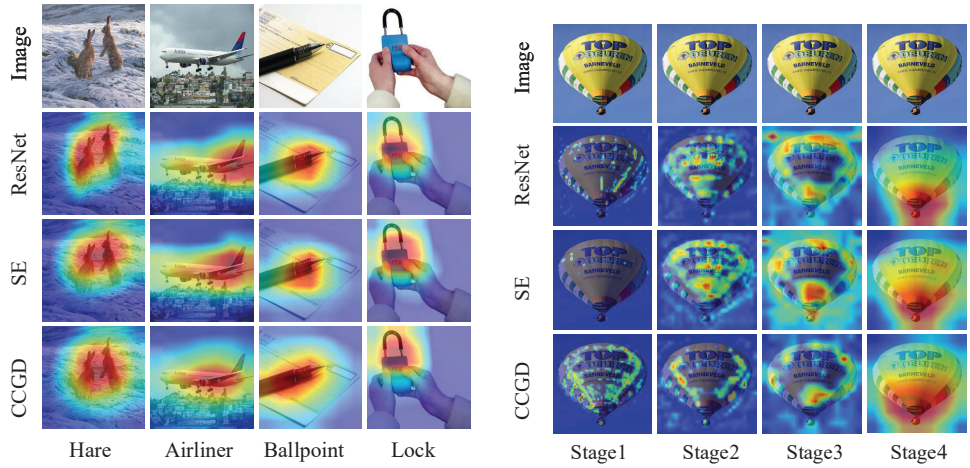


Figure 6. Visualization of class activation mapping using ResNet50, SENet50, and CCGDNet50 as backbone networks: (**a**) CAM of different objects; (**b**) CAM at different stages.

5. Conclusions

In this paper, we present a CCGD method that can improve the basic convolutional feature transformation process of CNNs, thus helping them to generate more discriminative representations. The image classification results obtained during our experiments agree well with our initial intuition and theory. In our future research, we aim to conduct further tests to assess whether CCGD is effective for other kinds of visual tasks, such as object detection and semantic segmentation.

Author Contributions: Conceptualization, H.Y.; methodology, H.Y. and F.M.; software, H.Y. and C.L.; formal analysis, Y.L.; investigation, H.Y. and W.L.; data curation, H.Y. and C.L.; writing—original draft preparation, H.Y.; writing—review and editing, Y.L. and W.L.; visualization, F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No.61871154, No. 62031013), by the Youth Program of National Natural Science Foundation of China (61906103,61906124), by the Basic and Applied Basic Research Fund of Guangdong Province (2019A1515011307).

Informed Consent Statement: Informed consent was implied from all subjects involved in the study by completing the survey.

Data Availability Statement: Data are available upon reasonable request to the submitting author.

Conflicts of Interest: There are no conflict of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
2. Zhang, Y.; Wallaceand, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv* **2015**, arXiv:1510.03820.
3. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
4. Woo, S.; Park, J.; Lee, J.Y. Bam: Bottleneck attention module. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018; pp. 67–86.
5. Park, J.; Woo, S.; Lee, J.Y. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
6. Lee, H.; Kim, H.E.; Nam, H. Srm: A style-based recalibration module for convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1854–1862.
7. Bello, I.; Zoph, B.; Vaswani, A.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3286–3295.
8. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE International Conference on Computer Vision Workshops (CVPR), Long Beach, CA, USA, 16–20 June 2019.
9. Li, X.; Wang, W.; Hu, X.L.; Yang, J. Selective kernel networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
10. Guo, J.; Ma, X.; Sansom, A.; McGuire, M. Spanet: Spatial Pyramid Attention Network for Enhanced Image Recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Beijing, China, 13–16 October 2020; pp. 1–6.
11. Liu, J.J.; Hou, Q.; Cheng, M.M.; Wang, C.; Feng, J. Improving Convolutional Networks with Self-Calibrated Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10096–10105.
12. Russakovsky, O.; Deng, J.; Satheesh, S. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 21–26 July 2017; pp. 618–626.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp.770–778.
14. Russakovsky, O.; Deng, J.; Satheesh, S. Learning multiple layers of features from tiny images. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
15. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images (Technical Report). University of Toronto, Canada, 2009. Available online: https://www.cs.toronto.edu/~kriz/cifar.html (accessed on 23 June 2022).
16. Karen, S.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
17. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2828.
18. Huang, G.; Liu, Z.; Maaten, V.D.; Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
19. Zagoruyko, S.; Nikos, K. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
20. Xie, S.; Girshick, R.; Dollár, A.P. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
21. Gao, S.; Cheng, M.M.; Zhao, K.Z. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
22. Yuan, P.; Lin, S.; Cui, C. Hierarchical-Split Block on Convolutional Neural Network. *arXiv* **2020**, arXiv:2010.07621.
23. Huang, G.; Liu, S.; Maaten, L.; Weinberger, K.Q. An efficient densenet using learned group convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2752–2761.
24. Barret, Z.; Vijay, V.; Jonathon, S.; Quoc, V.L.E. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8697–8710.
25. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical multi-scale attention for semantic segmentation. *arXiv* **2020**, arXiv:2005.10821.
26. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
27. Misra, D.; Nalamada, T.; Ajay, U.; Hou, Q.B. Rotate to Attend: Convolutional Triplet Attention Module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3139–3148.
28. Linsley, D.; Shiebler, D.; Eberhardt, S.; Serre, T. Learning what and where to attend. *arXiv* **2018**, arXiv:1805.08819.