# CBAM: Convolutional Block Attention Module

# CBAM：卷积块注意模块

Sanghyun Woo , Jongchan Park\* , Joon-Young Lee , and In So Kweon

Sanghyun Woo ，Jongchan Park\* ，Joon-Young Lee ，和 In So Kweon

Korea Advanced Institute of Science and Technology, Daejeon, Korea

韩国科学技术院，大田，韩国

{shwoo93, iskweon77 }@kaist.ac.kr

{shwoo93, iskweon77 }@kaist.ac.kr

2 Lunit Inc., Seoul, Korea

2 Lunit Inc.，首尔，韩国

jcpark@lunit.io

3 Adobe Research, San Jose, CA, USA

3 Adobe Research，圣何塞，加利福尼亚州，美国

jolee@adobe.com

Abstract. We propose Convolutional Block Attention Module (CBAM), a simple yet effective attention module for feed-forward convolutional neural networks. Given an intermediate feature map, our module sequentially infers attention maps along two separate dimensions, channel and spatial, then the attention maps are multiplied to the input feature map for adaptive feature refinement. Because CBAM is a lightweight and general module, it can be integrated into any CNN architectures seamlessly with negligible overheads and is end-to-end trainable along with base CNNs. We validate our CBAM through extensive experiments on ImageNet-1K, MS COCO detection, and VOC 2007 detection datasets. Our experiments show consistent improvements in classification and detection performances with various models, demonstrating the wide applicability of CBAM. The code and models will be publicly available.

摘要。我们提出了卷积块注意模块（CBAM），这是一种简单而有效的用于前馈卷积神经网络注意模块。给定一个中间特征图，我们的模块沿两个独立的维度（通道和空间）顺序推导出注意图，然后将注意图乘以输入特征图以进行自适应特征细化。由于CBAM是一个轻量级且通用的模块，它可以无缝地集成到任何CNN架构中，且几乎不会增加额外开销，并且可以与基础CNN一起端到端训练。我们通过在ImageNet-1K、MS COCO检测和VOC 2007检测数据集上进行大量实验来验证我们的CBAM。我们的实验显示了在各种模型中分类和检测性能的一致提高，证明了CBAM的广泛应用性。代码和模型将公开可用。

Keywords: Object recognition, attention mechanism, gated convolution

关键词：对象识别，注意机制，门控卷积

# 1 Introduction

# 1 引言

Convolutional neural networks (CNNs) have significantly pushed the performance of vision tasks based on their rich representation power. To enhance performance of CNNs, recent researches have mainly investigated three important factors of networks: depth, width, and cardinality.

卷积神经网络（CNNs）基于其丰富的表示能力，显著推动了基于视觉任务的性能 。为了提高CNN的性能，最近的研究主要调查了网络的三个重要因素：深度、宽度和基数。

From the LeNet architecture [4] to Residual-style Networks [5-8] so far, the network has become deeper for rich representation. VGGNet [9] shows that stacking blocks with the same shape gives fair results. Following the same spirit, ResNet [5] stacks the same topology of residual blocks along with skip connection to build an extremely deep architecture. GoogLeNet [10] shows that width is another important factor to improve the performance of a model. Zagoruyko and Komodakis [6] propose to increase the width of a network based on the ResNet architecture. They have shown that a 28-layer ResNet with increased

从LeNet架构[4]到至今的残差样式网络[5-8]，网络已经变得更深，以获得丰富的表示。VGGNet[9]表明，堆叠相同形状的块可以得到公平的结果。遵循同样的精神，ResNet[5]堆叠了相同的残差块拓扑，并通过跳跃连接构建了一个极其深的架构。GoogLeNet[10]表明，宽度是提高模型性能的另一个重要因素。Zagoruyko和Komodakis[6]提出基于ResNet架构增加网络的宽度。他们已经证明了一个28层的ResNet通过增加

\*Both authors have equally contributed.

\*两位作者贡献相等。

†The work was done while the author was at KAIST.

†该工作是在作者在KAIST期间完成的。

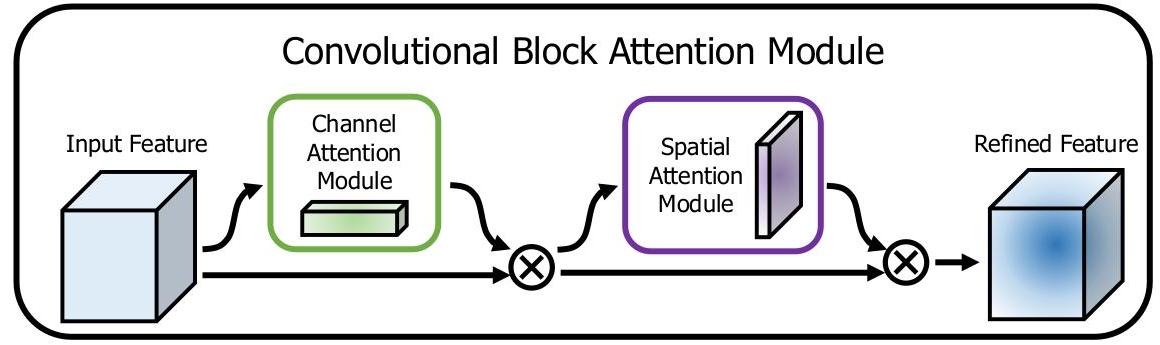


Fig. 1: The overview of CBAM. The module has two sequential sub-modules: channel and spatial. The intermediate feature map is adaptively refined through our module (CBAM) at every convolutional block of deep networks.

图1：CBAM的概述。该模块有两个顺序子模块：通道和空间。中间特征图通过我们的模块（CBAM）在深度网络的每个卷积块中自适应细化。

width can outperform an extremely deep ResNet with 1001 layers on the CI-FAR benchmarks. Xception [11] and ResNeXt [7] come up with to increase the cardinality of a network. They empirically show that cardinality not only saves the total number of parameters but also results in stronger representation power than the other two factors: depth and width.

宽度可以超越在CI-FAR基准上具有1001层的极深ResNet。Xception[11]和ResNeXt[7]提出增加网络的基数。他们通过实验表明，基数不仅节省了参数总数，而且比其他两个因素：深度和宽度，具有更强的表示能力。

Apart from these factors, we investigate a different aspect of the architecture design, attention. The significance of attention has been studied extensively in the previous literature . Attention not only tells where to focus, it also improves the representation of interests. Our goal is to increase representation power by using attention mechanism: focusing on important features and suppressing unnecessary ones. In this paper, we propose a new network module, named "Convolutional Block Attention Module". Since convolution operations extract informative features by blending cross-channel and spatial information together, we adopt our module to emphasize meaningful features along those two principal dimensions: channel and spatial axes. To achieve this, we sequentially apply channel and spatial attention modules (as shown in Fig. 1), so that each of the branches can learn ’what’ and ’where’ to attend in the channel and spatial axes respectively. As a result, our module efficiently helps the information flow within the network by learning which information to emphasize or suppress.

除了这些因素之外，我们还研究了架构设计的另一个方面，即注意力。注意力的重要性在之前的文献中已经被广泛研究 。注意力不仅告诉我们关注哪里，还能提高兴趣点的表示。我们的目标是利用注意力机制来增强表示能力：关注重要特征并抑制不必要的特征。在本文中，我们提出了一个新的网络模块，名为“卷积块注意力模块”。由于卷积操作通过融合跨通道和空间信息来提取信息特征，我们采用我们的模块来强调沿着这两个主要维度——通道和空间轴上的有意义特征。为了实现这一点，我们依次应用通道和空间注意力模块（如图1所示），以便每个分支可以分别在通道和空间轴上学习关注 ’什么’ 和 ’哪里’。因此，我们的模块通过学习强调或抑制哪些信息，有效地帮助网络内部的信息流动。

In the ImageNet-1K dataset, we obtain accuracy improvement from various baseline networks by plugging our tiny module, revealing the efficacy of CBAM. We visualize trained models using the grad-CAM [18] and observe that CBAM-enhanced networks focus on target objects more properly than their baseline networks. Taking this into account, we conjecture that the performance boost comes from accurate attention and noise reduction of irrelevant clutters. Finally, we validate performance improvement of object detection on the MS COCO and the VOC 2007 datasets, demonstrating a wide applicability of CBAM. Since we have carefully designed our module to be light-weight, the overhead of parameters and computation is negligible in most cases.

在ImageNet-1K数据集中，通过插入我们的小型模块，我们从各种基线网络中获得了准确性的提升，揭示了CBAM的有效性。我们使用grad-CAM [18] 观察训练后的模型，并注意到CBAM增强的网络比基线网络更恰当地关注目标对象。考虑到这一点，我们推测性能提升来自于准确的注意力和对无关杂质的降噪。最后，我们在MS COCO和VOC 2007数据集上验证了对象检测的性能提升，证明了CBAM的广泛应用性。由于我们精心设计了轻量级的模块，因此在大多数情况下，参数和计算的开销是可以忽略不计的。

Contribution. Our main contribution is three-fold.

贡献。我们的主要贡献有三方面。

1. We propose a simple yet effective attention module (CBAM) that can be widely applied to boost representation power of CNNs.

1. 我们提出了一个简单而有效的注意力模块（CBAM），可以广泛用于增强CNNs的表示能力。

2. We validate the effectiveness of our attention module through extensive ablation studies.

2. 我们通过广泛的消融研究验证了我们的注意力模块的有效性。

3. We verify that performance of various networks is greatly improved on the multiple benchmarks (ImageNet-1K, MS COCO, and VOC 2007) by plugging our light-weight module.

我们验证了，通过接入我们轻量级模块，各种网络在多个基准测试（ImageNet-1K、MS COCO和VOC 2007）上的性能得到了显著提升。

# 2 Related Work

# 2 相关工作

Network engineering. "Network engineering" has been one of the most important vision research, because well-designed networks ensure remarkable performance improvement in various applications. A wide range of architectures has been proposed since the successful implementation of a large-scale CNN [19]. An intuitive and simple way of extension is to increase the depth of neural networks [9]. Szegedy et al. [10] introduce a deep Inception network using a multi-branch architecture where each branch is customized carefully. While a naive increase in depth comes to saturation due to the difficulty of gradient propagation, ResNet [5] proposes a simple identity skip-connection to ease the optimization issues of deep networks. Based on the ResNet architecture, various models such as WideResNet [6], Inception-ResNet [8], and ResNeXt [7] have been developed. WideResNet [6] proposes a residual network with a larger number of convolutional filters and reduced depth. PyramidNet [20] is a strict generalization of WideResNet where the width of the network gradually increases. ResNeXt [7] suggests to use grouped convolutions and shows that increasing the cardinality leads to better classification accuracy. More recently, Huang et al. [21] propose a new architecture, DenseNet. It iteratively concatenates the input features with the output features, enabling each convolution block to receive raw information from all the previous blocks. While most of recent network engineering methods mainly target on three factors depth , width , and cardinality , we focus on the other aspect,’attention’, one of the curious facets of a human visual system.

网络工程。"网络工程"一直是视觉研究中最重要的领域之一，因为设计良好的网络确保了在各种应用中性能的显著提升。自从大规模CNN [19] 的成功实施以来，已经提出了广泛的架构。扩展的一种直观而简单的方法是增加神经网络的深度[9]。Szegedy等人[10]引入了一种深度Inception网络，该网络使用多分支架构，其中每个分支都经过精心定制。然而，由于梯度传播的困难，简单地增加深度会导致饱和，因此ResNet [5] 提出了一个简单的恒等跳接连接，以缓解深度网络的优化问题。基于ResNet架构，已经开发出各种模型，如WideResNet [6]、Inception-ResNet [8] 和ResNeXt [7]。WideResNet [6] 提出了一个具有更多卷积滤波器和减少深度的残差网络。PyramidNet [20] 是WideResNet的严格泛化，其中网络的宽度逐渐增加。ResNeXt [7] 建议使用分组卷积，并表明增加基数可以提高分类精度。最近，Huang等人[21]提出了一个新的架构，DenseNet。它迭代地将输入特征与输出特征连接起来，使得每个卷积块都能接收到来自所有先前块的原信息。虽然大多数最近的网络工程方法主要针对三个因素：深度 、宽度 和基数 ，但我们关注另一个方面，即“注意力”，这是人类视觉系统的一个好奇特征。

Attention mechanism. It is well known that attention plays an important role in human perception [23-25]. One important property of a human visual system is that one does not attempt to process a whole scene at once. Instead, humans exploit a sequence of partial glimpses and selectively focus on salient parts in order to capture visual structure better [26].

注意力机制。众所周知，注意力在人类感知中扮演着重要角色 [23-25]。人类视觉系统的一个重要特性是，人们不会试图一次性处理整个场景。相反，人类利用一系列部分瞥见，并选择性地关注显著部分，以更好地捕捉视觉结构 [26]。

Recently, there have been several attempts [27, 28] to incorporate attention processing to improve the performance of CNNs in large-scale classification tasks. Wang et al. [27] propose Residual Attention Network which uses an encoder-decoder style attention module. By refining the feature maps, the network not only performs well but is also robust to noisy inputs. Instead of directly computing the attention map, we decompose the process that learns channel attention and spatial attention separately. The separate attention generation process for 3D feature map has much less computational and parameter overhead, and therefore can be used as a plug-and-play module for pre-existing base CNN architectures.

最近，已有一些尝试 [27, 28] 将注意力处理融入卷积神经网络（CNNs），以提高在大规模分类任务中的性能。Wang等人 [27] 提出了残差注意力网络，该网络使用编码器-解码器风格的注意力模块。通过优化特征图，网络不仅表现良好，而且对噪声输入具有鲁棒性。我们不是直接计算 注意力图，而是将学习通道注意力和空间注意力的过程分别分解。对于3D特征图的独立注意力生成过程，计算和参数的开销要小得多，因此可以用作现有基础CNN架构的即插即用模块。

More close to our work, Hu et al. [28] introduce a compact module to exploit the inter-channel relationship. In their Squeeze-and-Excitation module, they use global average-pooled features to compute channel-wise attention. However, we show that those are suboptimal features in order to infer fine channel attention, and we suggest to use max-pooled features as well. They also miss the spatial attention, which plays an important role in deciding ’where’ to focus as shown in [29]. In our CBAM, we exploit both spatial and channel-wise attention based on an efficient architecture and empirically verify that exploiting both is superior to using only the channel-wise attention as [28]. Moreover, we empirically show that our module is effective in detection tasks (MS-COCO and VOC). Especially, we achieve state-of-the-art performance just by placing our module on top of the existing one-shot detector [30] in the VOC2007 test set.

更接近我们工作的是，Hu等人[28]引入了一个紧凑的模块来利用通道间关系。在他们提出的Squeeze-and-Excitation模块中，他们使用全局平均池化特征来计算通道注意力。然而，我们证明了这些特征在推断精细通道注意力方面是次优的，并建议也使用最大池化特征。他们还忽略了空间注意力，正如[29]所示，这在决定“关注哪里”方面起着重要作用。在我们的CBAM中，我们基于一个高效的架构利用了空间和通道注意力，并通过实验验证了同时利用这两种注意力优于仅使用通道注意力[28]。此外，我们通过实验表明，我们的模块在检测任务（MS-COCO和VOC）中是有效的。特别是，我们仅通过将我们的模块置于现有的单次检测器[30]之上，就在VOC2007测试集上实现了最先进的性能。

# 3 Convolutional Block Attention Module

# 3 卷积块注意力模块

Given an intermediate feature map as input, CBAM sequentially infers a 1D channel attention map and a 2D spatial attention map as illustrated in Fig. 1. The overall attention process can be summarized as:

给定一个中间特征图 作为输入，CBAM依次推断出一个一维通道注意力图 和一个二维空间注意力图 ，如图1所示。整体注意力过程可以概括为：

where denotes element-wise multiplication. During multiplication, the attention values are broadcasted (copied) accordingly: channel attention values are broadcasted along the spatial dimension, and vice versa. is the final refined output. Fig. 2 depicts the computation process of each attention map. The following describes the details of each attention module.

其中 表示逐元素乘法。在乘法过程中，注意力值相应地广播（复制）：通道注意力值沿空间维度广播，反之亦然。 是最终的细化输出。图2描述了每个注意力图的计算过程。以下描述了每个注意力模块的细节。

Channel attention module. We produce a channel attention map by exploiting the inter-channel relationship of features. As each channel of a feature map is considered as a feature detector [31], channel attention focuses on ’what’ is meaningful given an input image. To compute the channel attention efficiently, we squeeze the spatial dimension of the input feature map. For aggregating spatial information, average-pooling has been commonly adopted so far. Zhou et al. [32] suggest to use it to learn the extent of the target object effectively and Hu et al. [28] adopt it in their attention module to compute spatial statistics. Beyond the previous works, we argue that max-pooling gathers another important clue about distinctive object features to infer finer channel-wise attention. Thus, we use both average-pooled and max-pooled features simultaneously. We empirically confirmed that exploiting both features greatly improves representation power of networks rather than using each independently (see Sec. 4.1), showing the effectiveness of our design choice. We describe the detailed operation below.

通道注意力模块。我们通过利用特征之间的通道关系来生成通道注意力图。由于特征图中的每个通道被视为一个特征检测器 [31]，通道注意力关注的是给定输入图像的 ’什么’ 是有意义的。为了有效地计算通道注意力，我们压缩了输入特征图的空间维度。为了聚集空间信息，到目前为止通常采用平均池化。Zhou等人[32]建议使用它来有效地学习目标对象的范围，而Hu等人[28]在其注意力模块中采用它来计算空间统计。在之前的工作之外，我们认为最大池化收集了关于独特对象特征的另一个重要线索，以推断更细的通道注意力。因此，我们同时使用平均池化和最大池化的特征。我们从经验上确认，利用这两种特征大大提高了网络的表示能力，而不是单独使用每一种（见第4.1节），这表明了我们的设计选择的有效性。我们将在下面详细描述操作。

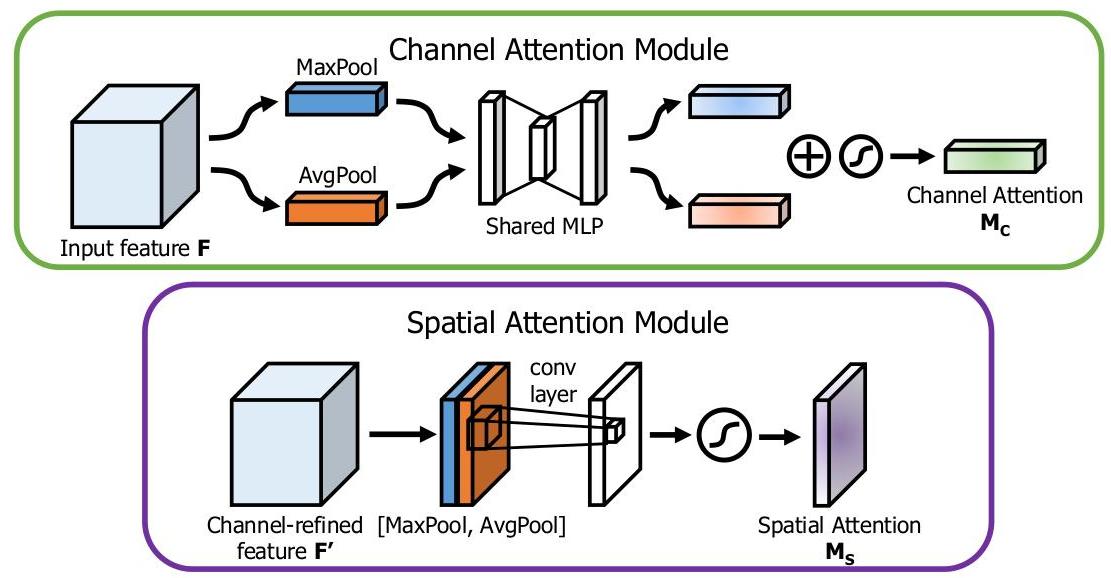


Fig. 2: Diagram of each attention sub-module. As illustrated, the channel sub-module utilizes both max-pooling outputs and average-pooling outputs with a shared network; the spatial sub-module utilizes similar two outputs that are pooled along the channel axis and forward them to a convolution layer.

图2：每个注意力子模块的图解。如图所示，通道子模块使用共享网络同时利用最大池化输出和平均池化输出；空间子模块利用沿着通道轴池化的类似两个输出，并将它们传递到卷积层。

We first aggregate spatial information of a feature map by using both average-pooling and max-pooling operations, generating two different spatial context descriptors: and , which denote average-pooled features and max-pooled features respectively. Both descriptors are then forwarded to a shared network to produce our channel attention map . The shared network is composed of multi-layer perceptron (MLP) with one hidden layer. To reduce parameter overhead, the hidden activation size is set to , where is the reduction ratio. After the shared network is applied to each descriptor, we merge the output feature vectors using element-wise summation. In short, the channel attention is computed as:

我们首先通过使用平均池化和最大池化操作聚合特征图的空间信息，生成两种不同的空间上下文描述符： 和 ，分别表示平均池化特征和最大池化特征。这两个描述符随后被送入共享网络以生成我们的通道注意力图 。共享网络由具有一个隐藏层的多层感知器（MLP）组成。为了减少参数开销，隐藏层的激活大小设置为 ，其中 是缩减比例。在共享网络应用于每个描述符之后，我们通过逐元素相加合并输出特征向量。简而言之，通道注意力的计算如下：

where denotes the sigmoid function, , and . Note that the MLP weights, and , are shared for both inputs and the ReLU activation function is followed by .

其中 表示sigmoid函数， 和 。请注意，MLP权重 和 对于两个输入是共享的，并且在 后面跟有ReLU激活函数。

Spatial attention module. We generate a spatial attention map by utilizing the inter-spatial relationship of features. Different from the channel attention, the spatial attention focuses on ’where’ is an informative part, which is complementary to the channel attention. To compute the spatial attention, we first apply average-pooling and max-pooling operations along the channel axis and concatenate them to generate an efficient feature descriptor. Applying pooling operations along the channel axis is shown to be effective in highlighting informative regions [33]. On the concatenated feature descriptor, we apply a convolution

空间注意力模块。我们通过利用特征之间的空间关系生成空间注意力图。与通道注意力不同，空间注意力关注于 ’哪里’ 是一个信息丰富的部分，这与通道注意力是互补的。为了计算空间注意力，我们首先沿通道轴应用平均池化和最大池化操作，并将它们连接起来生成一个有效的特征描述符。沿通道轴应用池化操作被证明在突出显示信息区域方面是有效的[33]。在连接的特征描述符上，我们应用一个卷积层来生成空间注意力图 ，该图编码了需要强调或抑制的位置。

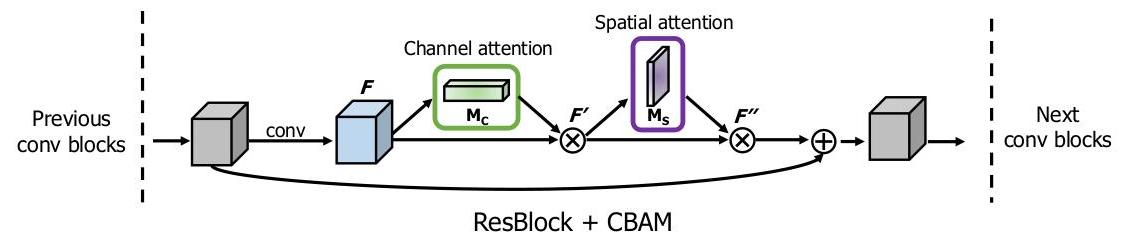


Fig. 3: CBAM integrated with a ResBlock in ResNet[5]. This figure shows the exact position of our module when integrated within a ResBlock. We apply CBAM on the convolution outputs in each block.

图3：ResNet[5]中与ResBlock集成的CBAM。该图显示了我们的模块在ResBlock内部集成的确切位置。我们在每个块的卷积输出上应用CBAM。

layer to generate a spatial attention map which encodes where to emphasize or suppress. We describe the detailed operation below.

层来生成空间注意力图 ，该图编码了需要强调或抑制的位置。我们将在下面详细描述操作。

We aggregate channel information of a feature map by using two pooling operations, generating two 2D maps: and . Each denotes average-pooled features and max-pooled features across the channel. Those are then concatenated and convolved by a standard convolution layer, producing our 2D spatial attention map. In short, the spatial attention is computed as:

我们通过使用两种池化操作来聚合特征图上的通道信息，生成两个2D地图： 和 。每一个都表示跨通道的平均池化特征和最大池化特征。然后将它们连接并通过标准卷积层卷积，产生我们的2D空间注意力图。简而言之，空间注意力的计算方式如下：

where denotes the sigmoid function and represents a convolution operation with the filter size of .

其中 表示sigmoid函数， 代表滤波器大小为 的卷积操作。

Arrangement of attention modules. Given an input image, two attention modules, channel and spatial, compute complementary attention, focusing on ’what’ and ’where’ respectively. Considering this, two modules can be placed in a parallel or sequential manner. We found that the sequential arrangement gives a better result than a parallel arrangement. For the arrangement of the sequential process, our experimental result shows that the channel-first order is slightly better than the spatial-first. We will discuss experimental results on network engineering in Sec. 4.1.

注意力模块的排列。给定一个输入图像，两个注意力模块，通道和空间，分别计算补充注意力，分别关注’什么’和’哪里’。考虑到这一点，两个模块可以以并行或顺序方式放置。我们发现顺序排列比并行排列得到更好的结果。对于顺序过程的排列，我们的实验结果表明，先通道后空间的顺序略优于先空间后通道。我们将在第4.1节讨论网络工程的实验结果。

# 4 Experiments

# 4 实验部分

We evaluate CBAM on the standard benchmarks: ImageNet-1K for image classification; MS COCO and VOC 2007 for object detection. In order to perform better apple-to-apple comparisons, we reproduced all the evaluated networks [5- in the PyTorch framework [35] and report our reproduced results in the whole experiments.

我们在标准基准上评估CBAM：ImageNet-1K用于图像分类；MS COCO和VOC 2007用于对象检测。为了进行更好的逐项比较，我们在PyTorch框架[35]中复现了所有评估的网络[5- ，并在整个实验中报告了我们的复现结果。

To thoroughly evaluate the effectiveness of our final module, we first perform extensive ablation experiments. Then, we verify that CBAM outperforms all the baselines without bells and whistles, demonstrating the general applicability of CBAM across different architectures as well as different tasks. One can seamlessly integrate CBAM in any CNN architectures and jointly train the combined CBAM-enhanced networks. Fig. 3 shows a diagram of CBAM integrated with a ResBlock in ResNet [5] as an example.

为了彻底评估我们最终模块的有效性，我们首先进行了广泛的消融实验。然后，我们验证了CBAM在没有任何花哨功能的情况下优于所有基线，证明了CBAM在不同架构以及不同任务上的一般适用性。用户可以无缝地将CBAM集成到任何CNN架构中，并共同训练结合了CBAM增强的网络。图3显示了CBAM与ResNet中的ResBlock集成的示例图。

# 4.1 Ablation studies

# 4.1 消融研究

In this subsection, we empirically show the effectiveness of our design choice. For this ablation study, we use the ImageNet-1K dataset and adopt ResNet-50 [5] as the base architecture. The ImageNet-1K classification dataset [1] consists of 1.2 million images for training and 50,000 for validation with 1,000 object classes. We adopt the same data augmentation scheme with for training and apply a single-crop evaluation with the size of at test time. The learning rate starts from 0.1 and drops every 30 epochs. We train the networks for 90 epochs. Following , we report classification errors on the validation set.

在本小节中，我们实证地展示了我们的设计选择的有效性。为了这项消融研究，我们使用了ImageNet-1K数据集并采用ResNet-50 [5] 作为基础架构。ImageNet-1K分类数据集 [1] 包含了120万张用于训练的图像和5万张用于验证的图像，共有1000个对象类别。我们采用了与 相同的数据增强方案进行训练，并在测试时应用了大小为 的单裁剪评估。学习率从0.1开始，每30个周期下降一次。我们训练网络90个周期。遵循 ，我们在验证集上报告分类错误。

Our module design process is split into three parts. We first search for the effective approach to computing the channel attention, then the spatial attention. Finally, we consider how to combine both channel and spatial attention modules. We explain the details of each experiment below.

我们的模块设计过程分为三个部分。我们首先寻找计算通道注意力的有效方法，然后是空间注意力。最后，我们考虑如何结合通道和空间注意力模块。我们将在下面解释每个实验的细节。

Channel attention. We experimentally verify that using both average-pooled and max-pooled features enables finer attention inference. We compare 3 variants of channel attention: average pooling, max pooling, and joint use of both poolings. Note that the channel attention module with an average pooling is the same as the SE [28] module. Also, when using both poolings, we use a shared MLP for attention inference to save parameters, as both of aggregated channel features lie in the same semantic embedding space. We only use channel attention modules in this experiment and we fix the reduction ratio to 16 .

通道注意力。我们实验性地验证了同时使用平均池化和最大池化特征能够使注意力推断更细致。我们比较了通道注意力的3个变体：平均池化、最大池化以及两种池化的联合使用。注意，使用平均池化的通道注意力模块与SE [28] 模块相同。此外，当使用两种池化时，我们为了节省参数，对注意力推断使用共享的MLP，因为聚合的通道特征位于相同的语义嵌入空间。我们在这个实验中只使用通道注意力模块，并将缩减比例固定为16。

Experimental results with various pooling methods are shown in Table 1. We observe that max-pooled features are as meaningful as average-pooled features, comparing the accuracy improvement from the baseline. In the work of SE [28], however, they only exploit the average-pooled features, missing the importance

表1展示了使用不同池化方法的实验结果。我们观察到，最大池化特征与平均池化特征一样有意义，通过比较基线的准确性提升可以看出。然而，在SE [28] 的工作中，他们只利用了平均池化特征，忽略了其重要性。

| Description | Parameters | GFLOPs | Top-1 Error(%) | Top-5 |
| --- | --- | --- | --- | --- |
| ResNet50 (baseline) |  | 3.86 | 24.56 | 7.50 |
| ResNet50 + AvgPool (SE [28]) |  | 3.94 | 23.14 | 6.70 |
| ResNet50 + MaxPool |  | 3.94 | 23.20 | 6.83 |
| ResNet50 + AvgPool & MaxPool |  | 4.02 | 22.80 | 6.52 |

Table 1: Comparison of different channel attention methods. We observe that using our proposed method outperforms recently suggested Squeeze and Excitation method [28].

表1：不同通道注意力方法的比较。我们观察到使用我们提出的方法优于最近提出的挤压和激发方法 [28]。

| Description | Param. | GFLOPs | Top-1 Error(%) | Top-5 |
| --- | --- | --- | --- | --- |
| ResNet channel (SE [28]) |  | 3.860 | 23.14 | 6.70 |
| ResNet channel |  | 3.860 | 22.80 | 6.52 |
| ResNet50 + channel + spatial (1x1 conv, k=3) | 28.10M | 3.862 | 22.96 | 6.64 |
| ResNet channel + spatial (1x1 conv, ) | 28.10M | 3.869 | 22.90 | 6.47 |
| ResNet channel + spatial (avg&max, |  | 3.863 | 22.68 | 6.41 |
| ResNet channel + spatial (avg&max, ) | 28.09M | 3.864 | 22.66 | 6.31 |

Table 2: Comparison of different spatial attention methods. Using the proposed channel-pooling (i.e. average- and max-pooling along the channel axis) along with the large kernel size of 7 for the following convolution operation performs best.

表2：不同空间注意力方法的比较。使用提出的通道池化（即沿通道轴的平均池化和最大池化）以及随后的卷积操作中使用7的大核尺寸表现最佳。

| Description | Top-1 Error(%) | Top-5 |
| --- | --- | --- |
| ResNet channel (SE [28]) | 23.14 | 6.70 |
| ResNet50 + channel + spatial | 22.66 | 6.31 |
| ResNet50 + spatial + channel | 22.78 | 6.42 |
| ResNet50 + channel & spatial in parallel | 22.95 | 6.59 |

Table 3: Combining methods of channel and spatial attention. Using both attention is critical while the best-combining strategy (i.e. sequential, channel-first) further improves the accuracy.

表3：通道和空间注意力方法的结合。同时使用两种注意力至关重要，而最佳结合策略（即顺序的，通道优先）进一步提高了准确性。

of max-pooled features. We argue that max-pooled features which encode the degree of the most salient part can compensate the average-pooled features which encode global statistics softly. Thus, we suggest to use both features simultaneously and apply a shared network to those features. The outputs of a shared network are then merged by element-wise summation. We empirically show that our channel attention method is an effective way to push performance further from SE [28] without additional learnable parameters. As a brief conclusion, we use both average- and max-pooled features in our channel attention module with the reduction ratio of 16 in the following experiments.

的最大池化特征。我们认为，编码最显著部分程度的最 大池化特征可以补偿编码全局统计的平均池化特征。因此，我们建议同时使用这两种特征，并对这些特征应用共享网络。共享网络的输出然后通过逐元素相加合并。我们从经验上证明了我们的通道注意力方法是一种在不增加额外可学习参数的情况下，从SE [28]进一步提高性能的有效方法。简而言之，我们在随后的实验中在通道注意力模块中同时使用平均池化和最大池化特征，并使用16的降低比例。

Spatial attention. Given the channel-wise refined features, we explore an effective method to compute the spatial attention. The design philosophy is symmetric with the channel attention branch. To generate a 2D spatial attention map, we first compute a 2D descriptor that encodes channel information at each pixel over all spatial locations. We then apply one convolution layer to the 2D descriptor, obtaining the raw attention map. The final attention map is normalized by the sigmoid function.

空间注意力。给定通道精炼的特征，我们探索了一种计算空间注意力的有效方法。设计理念与通道注意力分支对称。为了生成2D空间注意力图，我们首先计算一个2D描述符，该描述符在所有空间位置上的每个像素编码通道信息。然后我们对2D描述符应用一层卷积，得到原始注意力图。最终的注意力图通过sigmoid函数进行归一化。

We compare two methods of generating the 2D descriptor: channel pooling using average- and max-pooling across the channel axis and standard convolution reducing the channel dimension into 1 . In addition, we investigate the effect of a kernel size at the following convolution layer: kernel sizes of 3 and 7. In the experiment, we place the spatial attention module after the previously designed channel attention module, as the final goal is to use both modules together.

我们比较了两种生成2D描述符的方法：在通道轴上使用平均池化和最大池化的通道池化，以及标准 卷积将通道维度减少到1。此外，我们还研究了下一个卷积层中核大小的影响：核大小为3和7。在实验中，我们在之前设计的通道注意力模块后放置了空间注意力模块，因为最终目标是同时使用这两个模块。

Table 2 shows the experimental results. We can observe that the channel pooling produces better accuracy, indicating that explicitly modeled pooling leads to finer attention inference rather than learnable weighted channel pooling (implemented as convolution). In the comparison of different convolution kernel sizes, we find that adopting a larger kernel size generates better accuracy in both cases. It implies that a broad view (i.e. large receptive field) is needed for deciding spatially important regions. Considering this, we adopt the channel-pooling method and the convolution layer with a large kernel size to compute spatial attention. In a brief conclusion, we use the average- and max-pooled features across the channel axis with a convolution kernel size of 7 as our spatial attention module.

表2显示了实验结果。我们可以观察到通道池化产生了更好的准确度，表明显式建模的池化导致了比可学习的加权通道池化（实现为 卷积）更精细的注意力推理。在比较不同卷积核大小时，我们发现采用较大的核大小在两种情况下都产生了更好的准确度。这表明需要一个广泛的视野（即大的感受野）来决定空间中的重要区域。考虑到这一点，我们采用了通道池化方法和带有大核大小的卷积层来计算空间注意力。简而言之，我们使用通道轴上平均池化和最大池化的特征，以及核大小为7的卷积层作为我们的空间注意力模块。

Arrangement of the channel and spatial attention. In this experiment, we compare three different ways of arranging the channel and spatial attention submodules: sequential channel-spatial, sequential spatial-channel, and parallel use of both attention modules. As each module has different functions, the order may affect the overall performance. For example, from a spatial viewpoint, the channel attention is globally applied, while the spatial attention works locally. Also, it is natural to think that we may combine two attention outputs to build a 3D attention map. In the case, both attentions can be applied in parallel, then the outputs of the two attention modules are added and normalized with the sigmoid function.

通道和空间注意力的排列。在这个实验中，我们比较了三种不同的通道和空间注意力子模块的排列方式：顺序通道-空间，顺序空间-通道，以及并行使用两个注意力模块。由于每个模块具有不同的功能，顺序可能会影响整体性能。例如，从空间角度来看，通道注意力是全局应用的，而空间注意力是局部工作的。此外，自然地认为我们可以结合两个注意力输出构建一个3D注意力图。在这种情况下，两个注意力可以并行应用，然后两个注意力模块的输出相加并通过sigmoid函数进行归一化。

Table 3 summarizes the experimental results on different attention arranging methods. From the results, we can find that generating an attention map sequentially infers a finer attention map than doing in parallel. In addition, the channel-first order performs slightly better than the spatial-first order. Note that all the arranging methods outperform using only the channel attention independently, showing that utilizing both attentions is crucial while the best-arranging strategy further pushes performance.

表3总结了不同注意力排列方法的实验结果。从结果中我们可以发现，顺序生成注意力图比并行生成注意力图推断出更精细的注意力图。此外，通道优先顺序的表现略优于空间优先顺序。注意，所有排列方法的表现都优于仅独立使用通道注意力的方法，这表明同时利用两种注意力是至关重要的，而最佳排列策略进一步提升了性能。

Final module design. Throughout the ablation studies, we have designed the channel attention module, the spatial attention module, and the arrangement of the two modules. Our final module is as shown in Fig. 1 and Fig. 2: we choose average- and max-pooling for both channel and spatial attention module; we use convolution with a kernel size of 7 in the spatial attention module; we arrange the channel and spatial submodules sequentially. Our final module(i.e. ResNet50 + CBAM) achieves top-1 error of 22.66%, which is much lower than SE [28](1.e. ResNet50 + SE), as shown in Table 4.

最终模块设计。通过消融研究，我们设计了通道注意力模块、空间注意力模块以及这两个模块的排列。我们的最终模块如图1和图2所示：我们为通道和空间注意力模块都选择了平均池化和最大池化；在空间注意力模块中使用了核大小为7的卷积；我们依次排列通道和空间子模块。我们的最终模块（即ResNet50 + CBAM）实现了22.66%的top-1错误率，远低于SE [28]（即ResNet50 + SE），如表4所示。

# 4.2 Image Classification on ImageNet-1K

# 4.2 ImageNet-1K上的图像分类

We perform ImageNet-1K classification experiments to rigorously evaluate our module. We follow the same protocol as specified in Sec. 4.1 and evaluate our

我们进行ImageNet-1K分类实验以严格评估我们的模块。我们遵循第4.1节中指定的相同协议并评估我们的

Woo, Park, Lee, Kweon module in various network architectures including ResNet [5], WideResNet [6], and ResNext [7].

Woo、Park、Lee、Kweon模块在各种网络架构中的表现，包括ResNet [5]、WideResNet [6]和ResNext [7]。

| Architecture | Param. | GFLOPs | Top-1 Error (%) | Top-5 Error (%) |
| --- | --- | --- | --- | --- |
| ResNet18 [5] | 11.69M | 1.814 | 29.60 | 10.55 |
| ResNet18 [5] SE | 11.78M | 1.814 | 29.41 | 10.22 |
| ResNet18 [5]+ CBAM | 11.78M | 1.815 | 29.27 | 10.09 |
| ResNet34 | 21.80M | 3.664 | 26.69 | 8.60 |
| ResNet34 [5] SE [28] | 21.96M | 3.664 | 26.13 | 8.35 |
| ResNet34 [5]+ CBAM | 21.96M | 3.665 | 25.99 | 8.24 |
| ResNet50 [5] |  | 3.858 | 24.56 | 7.50 |
| ResNet50 5 SE |  | 3.860 | 23.14 | 6.70 |
| ResNet50 [5]+ CBAM |  | 3.864 | 22.66 | 6.31 |
| ResNet101 [5 | 44.55M | 7.570 | 23.38 | 6.88 |
| ResNet101 | 49.33M | 7.575 | 22.35 | 6.19 |
| ResNet101 CBAM | 49.33M | 7.581 | 21.51 | 5.69 |
| WideResNet18 [6] (widen=1.5) |  | 3.866 | 26.85 | 8.88 |
| WideResNet18 [6] (widen=1.5) + SE [28 |  | 3.867 | 26.21 | 8.47 |
| WideResNet18 [6] (widen=1.5) + CBAM |  | 3.868 | 26.10 | 8.43 |
| WideResNet18 |  | 6.696 | 25.63 | 8.20 |
| WideResNet18 [6] (widen=2.0) + SE [28] |  | 6.696 | 24.93 | 7.65 |
| WideResNet18 [6] (widen=2.0) + CBAN |  | 6.697 | 24.84 | 7.63 |
| ResNeXt50 [7] (32x4d) |  | 3.768 | 22.85 | 6.48 |
| ResNeXt50 SE |  | 3.771 | 21.91 | 6.04 |
| ResNeXt50 CBAM |  | 3.774 | 21.92 | 5.91 |
| ResNeXt101 [7] (32x4d) | 44.18M | 7.508 | 21.54 | 5.75 |
| ResNeXt101 [7] (32x4d) + SE [28] | 48.96M | 7.512 | 21.17 | 5.66 |
| ResNeXt101 [7] (32x4d) + CBAM | 48.96M | 7.519 | 21.07 | 5.59 |

\* all results are reproduced in the PyTorch framework.

\* 所有结果均在PyTorch框架中复现。

Table 4: Classification results on ImageNet-1K. Single-crop validation errors are reported.

表4：ImageNet-1K上的分类结果。报告了单次裁剪验证错误率。

Table 4 summarizes the experimental results. The networks with CBAM outperform all the baselines significantly, demonstrating that the CBAM can generalize well on various models in the large-scale dataset. Moreover, the models with CBAM improve the accuracy upon the one of the strongest method – SE [28] which is the winning approach of the ILSVRC 2017 classification task. It implies that our proposed approach is powerful, showing the efficacy of new pooling method that generates richer descriptor and spatial attention that complements the channel attention effectively.

表4总结了实验结果。带有CBAM的网络在所有基线中表现显著优于其他网络，表明CBAM在大规模数据集上的各种模型中具有很好的泛化能力。此外，带有CBAM的模型在准确性上超过了最强的方法——SE [28]，SE是ILSVRC 2017分类任务的获胜方法。这表明我们提出的方法具有强大的效力，显示了新池化方法生成更丰富描述子和空间注意力有效地补充了通道注意力的有效性。

Fig. 4 depicts the error curves of various networks during ImageNet-1K training. We can clearly see that our method exhibits lowest training and validation error in both error plots. It shows that CBAM has greater ability to improve generalization power of baseline models compared to SE [28].

图4描绘了在ImageNet-1K训练期间各种网络的误差曲线。我们可以清楚地看到，我们的方法在两个误差图中的训练和验证误差都是最低的。这表明CBAM相比SE [28]具有更大的能力来提高基线模型的泛化能力。

We also find that the overall overhead of CBAM is quite small in terms of both parameters and computation. This motivates us to apply our proposed module CBAM to the light-weight network, MobileNet [34]. Table 5 summarizes the experimental results that we conducted based on the MobileNet architecture. We have placed CBAM to two models, basic and capacity-reduced model(i.e. adjusting width multiplier to 0.7). We observe similar phenomenon as shown

我们还发现CBAM在参数和计算方面的整体开销非常小。这激励我们将我们提出的模块CBAM应用到轻量级网络MobileNet [34]中。表5总结了基于MobileNet架构进行的实验结果。我们在两个模型中应用了CBAM，分别是基本模型和容量降低模型（即调整宽度乘数 至0.7）。我们观察到与表4所示相似的现象。

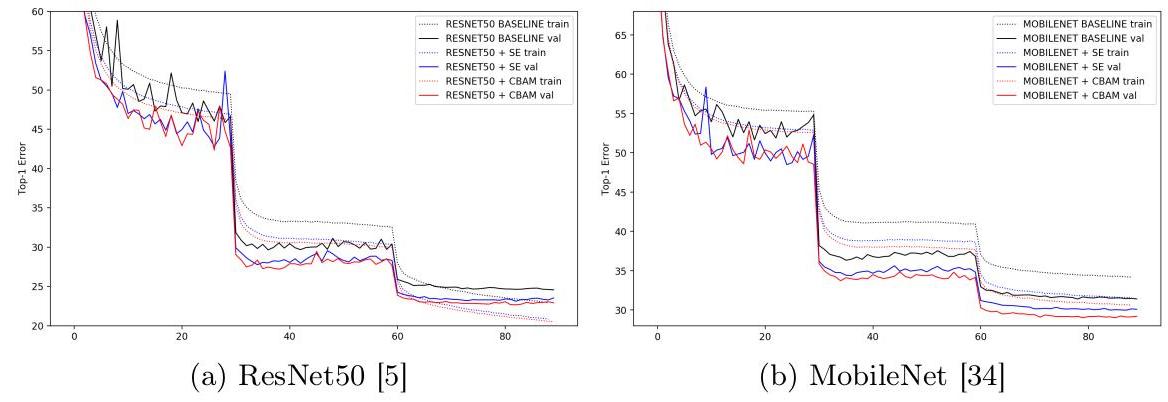


Fig. 4: Error curves during ImageNet-1K training. Best viewed in color.

图4：ImageNet-1K训练期间的误差曲线。彩色查看效果最佳。

| Architecture | Parameters | GFLOPs | Top-1 Error (%) | Top-5 Error (%) |
| --- | --- | --- | --- | --- |
| MobileNet [34] |  | 0.283 | 34.86 | 13.69 |
| MobileNet [34] SE [28] | 2.71M | 0.283 | 32.50 | 12.49 |
| MobileNet [34] |  | 0.289 | 31.51 | 11.48 |
| MobileNet [34] | 4.23M | 0.569 | 31.39 | 11.51 |
| MobileNet [34] SE [28] |  | 0.570 | 29.97 | 10.63 |
| MobileNet [34] CBAM |  | 0.576 | 29.01 | 9.99 |

\* all results are reproduced in the PyTorch framework.

\* 所有结果均在PyTorch框架中复现。

Table 5: Classification results on ImageNet-1K using the light-weight network, MobileNet [34]. Single-crop validation errors are reported.

表5：使用轻量级网络MobileNet [34]在ImageNet-1K上的分类结果。报告了单裁剪验证误差。

in Table 4. CBAM not only boosts the accuracy of baselines significantly but also favorably improves the performance of SE [28]. This shows the great potential of CBAM for applications on low-end devices.

如表4所示。CBAM不仅显著提高了基线的准确性，而且有利于改善SE [28]的性能。这显示了CBAM在低端设备应用上的巨大潜力。

# 4.3 Network Visualization with Grad-CAM [18]

# 4.3 使用Grad-CAM [18]进行网络可视化

For the qualitative analysis, we apply the Grad-CAM [18] to different networks using images from the ImageNet validation set. Grad-CAM is a recently proposed visualization method which uses gradients in order to calculate the importance of the spatial locations in convolutional layers. As the gradients are calculated with respect to a unique class, Grad-CAM result shows attended regions clearly. By observing the regions that network has considered as important for predicting a class, we attempt to look at how this network is making good use of features. We compare the visualization results of CBAM-integrated network (ResNet50 + CBAM) with baseline (ResNet50) and SE-integrated network (ResNet50 + SE). Fig. 5 illustrate the visualization results. The softmax scores for a target class are also shown in the figure.

对于定性分析，我们应用了Grad-CAM [18] 对ImageNet验证集中的图像使用不同的网络。Grad-CAM是一种最近提出的可视化方法，它使用梯度来计算卷积层中空间位置的重要性。由于梯度是针对一个独特的类别计算的，Grad-CAM结果清晰地显示了关注的区域。通过观察网络认为对预测一个类别重要的区域，我们尝试了解这个网络是如何很好地利用特征的。我们比较了集成CBAM的网络（ResNet50 + CBAM）的可视化结果与基线（ResNet50）和集成SE的网络（ResNet50 + SE）。图5展示了可视化结果。图中还显示了目标类的softmax得分。

In Fig. 5, we can clearly see that the Grad-CAM masks of the CBAM-integrated network cover the target object regions better than other methods. That is, the CBAM-integrated network learns well to exploit information in target object regions and aggregate features from them. Note that target class

在图5中，我们可以清楚地看到，集成CBAM的网络的Grad-CAM遮罩比其他方法更好地覆盖了目标物体区域。也就是说，集成CBAM的网络很好地学会了利用目标物体区域的信息并从中汇聚特征。注意目标类别

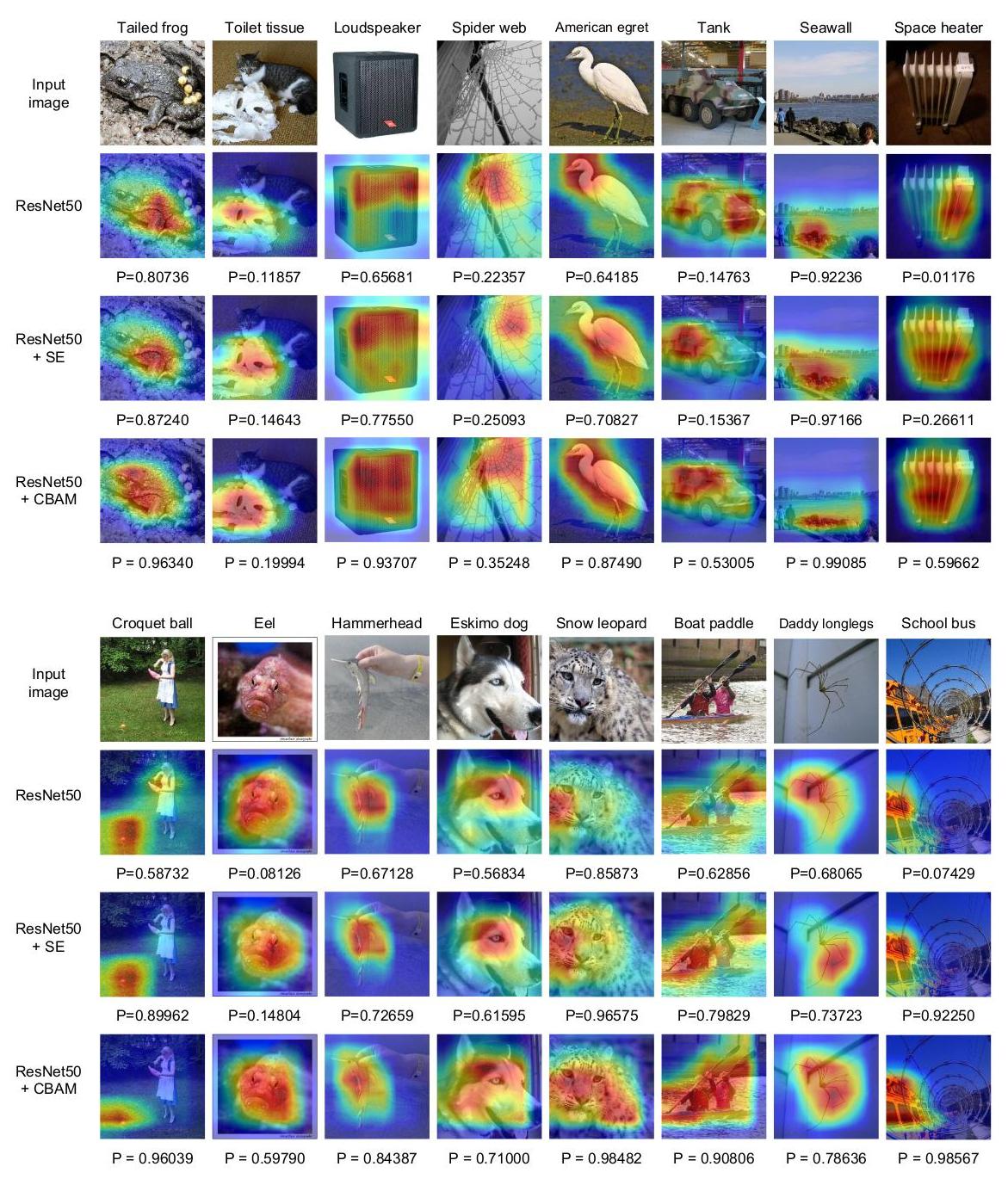


Fig. 5: Grad-CAM [18] visualization results. We compare the visualization results of CBAM-integrated network (ResNet50 + CBAM) with baseline (ResNet50) and SE-integrated network (ResNet50 + SE). The grad-CAM visualization is calculated for the last convolutional outputs. The ground-truth label is shown on the top of each input image and denotes the softmax score of each network for the ground-truth class.

图5：Grad-CAM [18] 可视化结果。我们比较了集成CBAM的网络（ResNet50 + CBAM）、基线（ResNet50）和集成SE的网络（ResNet50 + SE）的可视化结果。Grad-CAM可视化是基于最后一个卷积输出的计算结果。每个输入图像顶部显示了真实标签， 表示每个网络对真实类别的softmax得分。

| Backbone | Detector | mAP@.5 | mAP@.75 |  |
| --- | --- | --- | --- | --- |
| ResNet50 [5] | Faster-RCNN | 46.2 | 28.1 | 27.0 |
| ResNet50 [5]+ CBAM | Faster-RCNN | 48.2 | 29.2 | 28.1 |
| ResNet101 [5] | Faster-RCNN [4] | 48.4 | 30.7 | 29.1 |
| ResNet101 | Faster-RCNN[41] | 50.5 | 32.6 | 30.8 |

\* all results are reproduced in the PyTorch framework.

\* 所有结果都是在PyTorch框架中复现的。

Table 6: Object detection mAP(%) on the MS COCO validation set. We

表6：在MS COCO验证集上的对象检测mAP(%)。我们

adopt the Faster R-CNN [41] detection framework and apply our module to the base networks. CBAM boosts mAP@[.5, .95] by 0.9 for both baseline networks.

采用了Faster R-CNN [41] 检测框架并将我们的模块应用到基础网络中。CBAM使基线网络的mAP@[.5, .95] 提高了0.9。

| Backbone | Detector | mAP@.5 | Parameters (M) |
| --- | --- | --- | --- |
| VGG16 [9] | SSD [39] | 77.8 | 26.5 |
| VGG16 [9] | StairNet [30] | 78.9 | 32.0 |
| VGG16 [9] | StairNet [30] SE [28] | 79.1 | 32.1 |
| VGG16 [9] | StairNet [30] + CBAM | 79.3 | 32.1 |
| MobileNet [34] | SSD [39] | 68.1 | 5.81 |
| MobileNet [34] | StairNet [30] | 70.1 | 5.98 |
| MobileNet [34] | StairNet [30] SE [28] | 70.0 | 5.99 |
| MobileNet [34] | StairNet [30] + CBAM | 70.5 | 6.00 |

\* all results are reproduced in the PyTorch framework.

所有结果都在 PyTorch 框架中复现。

Table 7: Object detection mAP(%) on the VOC 2007 test set. We adopt the StairNet [30] detection framework and apply SE and CBAM to the detectors. CBAM favorably improves all the strong baselines with negligible additional parameters.

表7：在 VOC 2007 测试集上的目标检测 mAP(%)。我们采用 StairNet [30] 检测框架，并在检测器中应用 SE 和 CBAM。CBAM 以可忽略的额外参数有利于改进所有强大的基线。

scores also increase accordingly. From the observations, we conjecture that the feature refinement process of CBAM eventually leads the networks to utilize given features well.

分数也相应增加。从观察中，我们推测 CBAM 的特征精炼过程最终使网络能够很好地利用给定的特征。

# 4.4 MS COCO Object Detection

# 4.4 MS COCO 目标检测

We conduct object detection on the Microsoft COCO dataset [3]. This dataset involves 80k training images ("2014 train") and 40k validation images ("2014 val"). The average mAP over different IoU thresholds from 0.5 to 0.95 is used for evaluation. According to , we trained our model using all the training images as well as a subset of validation images, holding out 5,000 examples for validation. Our training code is based on [40] and we train the network for iterations for fast performance validation. We adopt Faster-RCNN [41] as our detection method and ImageNet pre-trained ResNet50 and ResNet101 [5] as our baseline networks. Here we are interested in performance improvement by plugging CBAM to the baseline networks. Since we use the same detection method in all the models, the gains can only be attributed to the enhanced representation power, given by our module CBAM. As shown in the Table 6, we observe significant improvements from the baseline, demonstrating generalization performance of CBAM on other recognition tasks.

我们在 Microsoft COCO 数据集 [3] 上进行目标检测。这个数据集包括 80k 训练图像（"2014 训练"）和 40k 验证图像（"2014 验证"）。使用从 0.5 到 0.95 的不同 IoU 阈值计算的平均 mAP 用于评估。根据 ，我们使用所有训练图像以及验证图像的一个子集来训练我们的模型，保留 5,000 个样本用于验证。我们的训练代码基于 [40]，并且我们训练网络 次迭代以进行快速性能验证。我们采用 Faster-RCNN [41] 作为我们的检测方法，以及 ImageNet 预训练的 ResNet50 和 ResNet101 [5] 作为我们的基线网络。这里我们关注的是通过将 CBAM 插入基线网络来提升性能。由于我们在所有模型中使用了相同的检测方法，因此收益只能归因于我们模块 CBAM 增强的表示能力。如表6所示，我们从基线观察到显著的改进，证明了 CBAM 在其他识别任务上的泛化性能。

# 4.5 VOC 2007 Object Detection

# 4.5 VOC 2007 目标检测

We further perform experiments on the PASCAL VOC 2007 test set. In this experiment, we apply CBAM to the detectors, while the previous experiments (Table 6) apply our module to the base networks. We adopt the StairNet [30] framework, which is one of the strongest multi-scale method based on the SSD [39]. For the experiment, we reproduce SSD and StairNet in our PyTorch platform in order to estimate performance improvement of CBAM accurately and achieve and mAP@. 5 respectively, which are higher than the original accuracy reported in the original papers. We then place SE [28] and CBAM right before every classifier, refining the final features which are composed of up-sampled global features and corresponding local features before the prediction, enforcing model to adaptively select only the meaningful features. We train all the models on the union set of VOC 2007 trainval and VOC 2012 trainval (“07+12”), and evaluate on the VOC 2007 test set. The total number of training epochs is 250 . We use a weight decay of 0.0005 and a momentum of 0.9 . In all the experiments, the size of the input image is fixed to 300 for the simplicity.

我们进一步在PASCAL VOC 2007测试集上进行了实验。在这个实验中，我们将CBAM应用于检测器，而之前的实验（表6）则是将我们的模块应用于基础网络。我们采用了StairNet [30]框架，这是基于SSD [39]的较强多尺度方法之一。为了进行实验，我们在PyTorch平台上复现了SSD和StairNet，以准确评估CBAM的性能提升，并分别达到了 和 mAP@. 5，这高于原论文中报告的原始精度。然后，我们在每个分类器之前放置SE [28]和CBAM，优化由上采样全局特征和相应的局部特征组成的最终特征，在预测前强制模型自适应选择仅对有意义特征。我们在VOC 2007 trainval和VOC 2012 trainval的并集（“07+12”）上训练所有模型，并在VOC 2007测试集上进行评估。总的训练周期数为250。我们使用权重衰减为0.0005和动量为0.9。在所有实验中，输入图像的大小固定为300，以简化处理。

The experimental results are summarized in Table 7. We can clearly see that CBAM improves the accuracy of all strong baselines with two backbone networks. Note that accuracy improvement of CBAM comes with a negligible parameter overhead, indicating that enhancement is not due to a naive capacity-increment but because of our effective feature refinement. In addition, the result using the light-weight backbone network [34] again shows that CBAM can be an interesting method to low-end devices.

实验结果总结在表7中。我们可以清楚地看到，CBAM提高了所有强基线在两种骨干网络上的精度。请注意，CBAM的精度提升带来的参数开销是可以忽略不计的，这表明增强并非是由于简单的容量增加，而是因为我们有效的特征优化。此外，使用轻量级骨干网络 [34] 的结果再次表明，CBAM可能成为低端设备的一个有趣方法。

# 5 Conclusion

# 5 结论

We have presented the convolutional bottleneck attention module (CBAM), a new approach to improve representation power of CNN networks. We apply attention-based feature refinement with two distinctive modules, channel and spatial, and achieve considerable performance improvement while keeping the overhead small. For the channel attention, we suggest to use the max-pooled features along with the average-pooled features, leading to produce finer attention than SE [28]. We further push the performance by exploiting the spatial attention. Our final module (CBAM) learns what and where to emphasize or suppress and refines intermediate features effectively. To verify its efficacy, we conducted extensive experiments with various state-of-the-art models and confirmed that CBAM outperforms all the baselines on three different benchmark datasets: ImageNet-1K, MS COCO, and VOC 2007. In addition, we visualize how the module exactly infers given an input image. Interestingly, we observed that our module induces the network to focus on target object properly. We hope CBAM become an important component of various network architectures.

我们提出了卷积瓶颈注意力模块（CBAM），这是一种提高卷积神经网络（CNN）表征能力的新方法。我们应用了基于注意力的特征精炼，包括两个独特的模块：通道和空间，在保持较低计算开销的同时实现了性能的显著提升。对于通道注意力，我们建议使用最大池化特征和平均池化特征相结合，从而产生比SE [28]更精细的注意力。我们进一步通过利用空间注意力来提升性能。我们的最终模块（CBAM）学习在什么位置以及如何强调或抑制，并有效地精炼中间特征。为了验证其有效性，我们使用各种最先进的模型进行了广泛的实验，并确认CBAM在三个不同的基准数据集上均优于所有基线：ImageNet-1K、MS COCO和VOC 2007。此外，我们还可视化了这个模块在给定输入图像时是如何进行推断的。有趣的是，我们观察到我们的模块能够引导网络正确地关注目标对象。我们希望CBAM能成为各种网络架构中的重要组成部分。

# References

# 参考文献

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2009)

2. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images

3. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proc. of European Conf. on Computer Vision (ECCV). (2014)

4. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11) (1998) 2278-2324

5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016)

6. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)

7. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)

8. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proc. of Association for the Advancement of Artificial Intelligence (AAAI). (2017)

9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

10. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2015)

11. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. arXiv preprint arXiv:1610.02357 (2016)

12. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention." advances in neural information processing systems. In: Proc. of Neural Information Processing Systems (NIPS). (2014)

13. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. (2014)

14. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. (2014)

15. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. (2015)

16. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. (2015)

17. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Proc. of Neural Information Processing Systems (NIPS). (2015)

18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 618-626

19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proc. of Neural Information Processing Systems (NIPS). (2012)

20. Han, D., Kim, J., Kim, J.: Deep pyramidal residual networks. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2017)

21. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)

22. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016)

23. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. In: IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). (1998)

24. Rensink, R.A.: The dynamic representation of scenes. In: Visual cognition 7.1-3. (2000)

25. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. In: Nature reviews neuroscience 3.3. (2002)

26. Larochelle, H., Hinton, G.E.: Learning to combine foveal glimpses with a third-order boltzmann machine. In: Proc. of Neural Information Processing Systems (NIPS). (2010)

27. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017)

28. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)

29. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2017)

30. Sanghyun, W., Soonmin, H., So, K.I.: Stairnet: Top-down semantic aggregation for accurate one shot detection. In: Proc. of Winter Conference on Applications of Computer Vision (WACV). (2018)

31. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proc. of European Conf. on Computer Vision (ECCV). (2014)

32. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE (2016) 2921-2929

33. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR. (2017)

34. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., An-dreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

35. : Pytorch. http://pytorch.org/ Accessed: 2017-11-08.

36. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Proc. of European Conf. on Computer Vision (ECCV). (2016)

37. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Proc. of European Conf. on Computer Vision (ECCV). (2016)

38. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016)

39. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proc. of European Conf. on Computer Vision (ECCV). (2016)

40. Chen, X., Gupta, A.: An implementation of faster rcnn with study for region sampling. arXiv preprint arXiv:1702.02138 (2017)

41. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proc. of Neural Information Processing Systems (NIPS). (2015)