# Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks

# Gather-Excite：在卷积神经网络中利用特征上下文

Jie Hu\*

胡杰\*

Momenta

Momenta

hujie@momenta.ai

Li Shen\*

沈丽\*

Visual Geometry Group

视觉几何组

University of Oxford

牛津大学

lishen@robots.ox.ac.uk

Samuel Albanie\*

塞缪尔·阿尔巴尼亚\*

Visual Geometry Group

视觉几何组

University of Oxford

牛津大学

albanie@robots.ox.ac.uk

Gang Sun

孙刚

Momenta

Momenta

sungang@momenta.ai

Andrea Vedaldi

安德烈亚·维达利

Visual Geometry Group

视觉几何组

University of Oxford

牛津大学

vedaldi@robots.ox.ac.uk

Abstract

摘要

While the use of bottom-up local operators in convolutional neural networks (CNNs) matches well some of the statistics of natural images, it may also prevent such models from capturing contextual long-range feature interactions. In this work, we propose a simple, lightweight approach for better context exploitation in CNNs. We do so by introducing a pair of operators: gather, which efficiently aggregates feature responses from a large spatial extent, and excite, which redistributes the pooled information to local features. The operators are cheap, both in terms of number of added parameters and computational complexity, and can be integrated directly in existing architectures to improve their performance. Experiments on several datasets show that gather-excite can bring benefits comparable to increasing the depth of a CNN at a fraction of the cost. For example, we find ResNet-50 with gather-excite operators is able to outperform its 101-layer counterpart on ImageNet with no additional learnable parameters. We also propose a parametric gather-excite operator pair which yields further performance gains, relate it to the recently-introduced Squeeze-and-Excitation Networks, and analyse the effects of these changes to the CNN feature activation statistics.

虽然卷积神经网络（CNNs）中使用的自下而上的局部操作符与自然图像的一些统计特性相匹配，但它们也可能阻止模型捕捉上下文的长距离特征交互。在这项工作中，我们提出了一个简单、轻量级的方法，用于在CNN中更好地利用上下文。我们通过引入一对操作符来实现这一点：gather操作符，它有效地从大范围的空间中聚合特征响应；excite操作符，它将池化信息重新分配到局部特征。这些操作符成本低廉，无论是添加的参数数量还是计算复杂度，都可以直接集成到现有架构中以提高其性能。在多个数据集上的实验表明，gather-excite带来的好处可以与增加CNN深度的效果相媲美，但成本只有一小部分。例如，我们发现带有gather-excite操作符的ResNet-50能够在不增加可学习参数的情况下，超过其在ImageNet上的101层对比模型。我们还提出了一种参数化的gather-excite操作符对，它带来了进一步的性能提升，并将其与最近引入的Squeeze-and-Excitation网络联系起来，分析了这些改变对CNN特征激活统计的影响。

# 1 Introduction

# 1 引言

Convolutional neural networks (CNN) [21] are the gold-standard approach to problems such as image classification , object detection [32] and image segmentation [3]. Thus, there is a significant interest in improved CNN architectures. In computer vision, an idea that has often improved visual representations is to augment functions that perform local decisions with functions that operate on a larger context, providing a cue for resolving local ambiguities [39]. While the term "context" is overloaded [6], in this work we focus specifically on feature context, namely the information captured by the feature extractor responses (i.e. the CNN feature maps) as a whole, spread over the full spatial extent of the input image.

卷积神经网络（CNN）[21]是处理如图像分类 、目标检测[32]和图像分割[3]等问题的黄金标准方法。因此，人们对改进CNN架构有着浓厚的兴趣。在计算机视觉领域，一个经常改善视觉表示的想法是将执行局部决策的函数与在更大范围内操作的函数相结合，为解决局部模糊提供线索[39]。虽然“上下文”一词含义繁多[6]，但在这项工作中，我们特别关注特征上下文，即特征提取器响应（即CNN特征图）整体捕获的信息，遍布输入图像的整个空间范围。

In many standard CNN architectures the receptive fields of many feature extractors are theoretically already large enough to cover the input image in full. However, the effective size of such fields is in practice considerably smaller [27]. This may be one factor explaining why improving the use of context in deep networks can lead to better performance, as has been repeatedly demonstrated in object detection and other applications .

在许多标准的CNN架构中，许多特征提取器的感受野在理论上已经足够大，可以覆盖整个输入图像。然而，实际上这些字段的有效大小要小得多[27]。这可能是解释为什么在深度网络中改进上下文的使用可以导致性能提升的一个因素，这一点在目标检测和其他应用中已经反复得到验证 。

\*Equal contribution

\*同等贡献

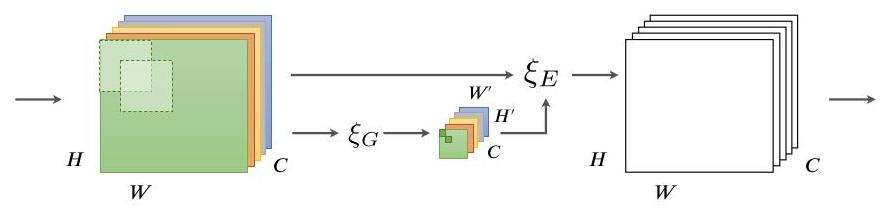


Figure 1: The interaction of a gather-excite operator pair, . The gather operator first aggregates feature responses across spatial neighbourhoods. The resulting aggregates are then passed, together with the original input tensor, to an excite operator that produces an output that matches the dimensions of the input.

图1：一对收集-激化操作符的交互 。收集操作符 首先在空间邻域内聚合特征响应。然后将这些聚合结果与原始输入张量一起传递给激化操作符 ，该操作符生成的输出与输入的尺寸相匹配。

Prior work has illustrated that using simple aggregations of low level features can be effective at encoding contextual information for visual tasks, and may prove a useful alternative to iterative methods based on higher level semantic features [44]. Demonstrating the effectiveness of such an approach, the recently proposed Squeeze-and-Excitation (SE) networks [15] showed that reweighting feature channels as a function of features from the full extent of input can improve classification performance. In these models, the squeeze operator acts as a lightweight context aggregator and the resulting embeddings are then passed to the reweighting function to ensure that it can exploit information beyond the local receptive fields of each filter.

先前的研究表明，使用低级特征的简单聚合可以在视觉任务中对上下文信息进行有效编码，并且可能成为基于高级语义特征的迭代方法的可行替代 [44]。最近提出的挤压与激发（Squeeze-and-Excitation，SE）网络 [15] 证明了这种方法的 effectiveness，它通过将特征通道的权重作为输入全范围特征的功能进行重新调整，可以提高分类性能。在这些模型中，挤压操作符充当轻量级的上下文聚合器，然后将得到的嵌入传递给重新加权函数，以确保它可以利用每个滤波器的局部感受野之外的信息。

In this paper, we build on this approach and further explore mechanisms to incorporate context throughout the architecture of a deep network. Our goal is to explore more efficient algorithms as well as the essential properties that make them work well. We formulate these "context" modules as the composition of two operators: a gather operator, which aggregates contextual information across large neighbourhoods of each feature map, and an excite operator, which modulates the feature maps by conditioning on the aggregates.

在本文中，我们基于这种方法，进一步探索在整个深度网络架构中整合上下文的机制。我们的目标是探索更高效的算法以及使它们有效工作的本质属性。我们将这些“上下文”模块构建为两个操作符的组合：一个聚集操作符，它跨每个特征图的大邻域聚合上下文信息；一个激发操作符，它通过条件聚合来调整特征图。

Using this decomposition, we chart the space of designs that can exploit feature context in deep networks and explore the effect of different operators independently. Our study leads us to propose a new, lightweight gather-excite pair of operators which yields significant improvements across different architectures, datasets and tasks, with minimal tuning of hyperparameters. We also investigate the effect of the operators on distributed representation learned by existing deep architectures: we find the mechanism produces intermediate representations that exhibit lower class selectivity, suggesting that providing access to additional context may enable greater feature re-use. The code for all models used in this work is publicly available at https://github.com/hujie-frank/GENet.

使用这种分解，我们绘制了可以在深度网络中利用特征上下文的设计空间，并独立地探索不同操作符的效果。我们的研究使我们提出了一对新的轻量级聚集-激发操作符，它们在不同的架构、数据集和任务中都能带来显著的改进，并且对超参数的调整最少。我们还研究了操作符对现有深度架构学到的分布式表示的影响：我们发现该机制产生的中间表示表现出较低的类别选择性，这表明提供额外的上下文信息可能使特征重用变得更加灵活。本文中使用的所有模型的代码都在 https://github.com/hujie-frank/GENet 公开可用。

# 2 The Gather-Excite Framework

# 2 聚集-激发框架

In this section, we introduce the Gather-Excite (GE) framework and describe its operation.

在这一节中，我们介绍聚集-激发（GE）框架并描述其操作。

The design is motivated by examining the flow of information that is typical of CNNs. These models compute a hierarchy of representations that transition gradually from spatial to channel coding. Deeper layers achieve greater abstraction by combining features from previous layers while reducing resolution, increasing the receptive field size of the units, and increasing the number of feature channels.

设计灵感来源于研究卷积神经网络中典型的信息流。这些模型计算一个表示层次，从空间编码逐渐过渡到通道编码。更深的层通过组合前一层的特征来实现更高层次的抽象，同时降低分辨率，增加单元的感知野大小和特征通道的数量。

The family of bag-of-visual-words models demonstrated the effectiveness of pooling the information contained in local descriptors to form a global image representation out of a local one. Inspired by this observation, we aim to help convolutional networks exploit the contextual information contained in the field of feature responses computed by the network itself.

包含视觉词模型族 证明了将局部描述符中的信息池化以形成全局图像表示的有效性。受到这一观察的启发，我们旨在帮助卷积网络利用网络本身计算的特征响应场中包含的上下文信息。

To this end, we construct a lightweight function to gather feature responses over large neighbourhoods and use the resulting contextual information to modulate original responses of the neighbourhood elements. Specifically, we define a gather operator which aggregates neuron responses over a given spatial extent, and an excite operator which takes in both the aggregates and the original input to produce a new tensor with the same dimensions of the original input. The GE operator pair is illustrated in Fig. 1.

为此，我们构建了一个轻量级函数，用于在大邻域内收集特征响应，并使用由此产生的上下文信息来调节邻域元素的原响应。具体来说，我们定义了一个聚集操作符 ，它在一个给定的空间范围内聚合神经元响应，以及一个激发操作符 ，它接收聚合结果和原始输入，以产生一个与原始输入具有相同维度的新的张量。GE操作符对在图1中有所说明。

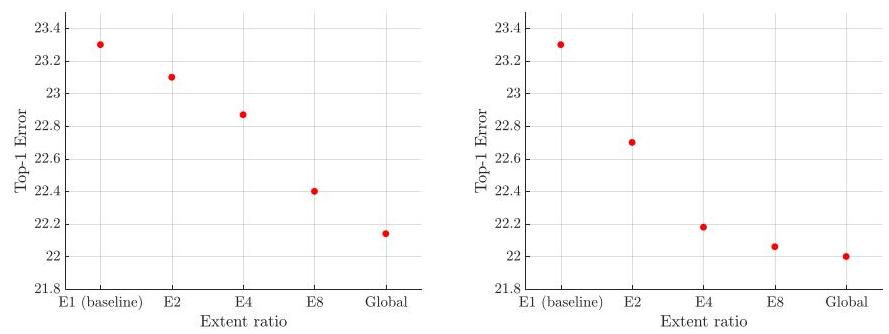


Figure 2: Top-1 ImageNet validation error (%) for the proposed (left) GE- and (right) GE- designs based on a ResNet-50 architecture (the baseline label indicates the performance of the original ResNet-50 model in both plots). For reference, ResNet-101 achieves a top-1 error of 22.20%. See Sec. 3 for further details.

图2：基于ResNet-50架构的提议的GE- （左）和GE- （右）设计的ImageNet验证错误百分比（基线标签表示两个图中原ResNet-50模型的性能）。作为参考，ResNet-101实现了22.20%的top-1错误。有关更多详细信息，请参见第3节。

More formally, let denote a collection of feature maps produced by the network. To assess the effect of varying the size of the spatial region over which the gathering occurs, we define the selection operator where represents the extent ratio of the selection. We then define a gather operator with extent ratio to be a function that satisfies for any input the constraint , where , denotes the indicator tensor and is the Hadamard product. This notation simply states that at each output location of the channel , the gather operator has a receptive field of the input that lies within a single channel and has an area bounded by . If the field envelops the full input feature map, we say that the gather operator has global extent. The objective of the excite operator is to make use of the gathered output as a contextual feature and takes the form , where is the map responsible for rescaling and distributing the signal from the aggregates.

更正式地说，令 表示由网络产生的特征图集合。为了评估变化聚集发生的空间区域大小对效果的影响，我们定义了选择算子 ，其中 表示选择的范围比例。然后我们定义了一个具有范围比例 的聚集算子，它是一个满足对于任何输入 约束 的函数 ，其中 ， 表示指示张量， 是 Hadamard 积。这个记号简单地说明，在通道 的每个输出位置 上，聚集算子有一个输入的感受野，该感受野位于单个通道内，并且面积由 确定。如果该字段覆盖了整个输入特征图，我们说聚集算子具有全局范围。激励算子的目标是将聚集的输出作为上下文特征来使用，其形式为 ，其中 是负责重新缩放和分配聚集信号的映射。

# 3 Models and Experiments

# 3 模型和实验

In this section, we explore and evaluate a number of possible instantiations of the gather-excite framework. To compare the utility of each design, we conduct a series of experiments on the task of image classification using the ImageNet 1K dataset [33]. The dataset contains 1.2 million training images and validation images. In the experiments that follow, all models are trained on the training set and evaluated on the validation set. We base our investigation on the popular ResNet- 50 architecture which attains good performance on this dataset and has been shown to generalise effectively to a range of other domains [9]. New models are formed by inserting gather-excite operators into the residual branch immediately before summation with the identity branch of each building block of ResNet-50. These models are trained from random initialisation [10] using SGD with momentum 0.9 with minibatches of 256 images, each cropped to pixels. The initial learning rate is set to 0.1 and is reduced by a factor of 10 each time the loss plateaus (three times). Models typically train for approximately 300 epochs in total (note that this produces stronger models than the fixed 100-epoch optimisation schedule used in [15]). In all experiments, we report single-centre-crop results on the ImageNet validation set.

在本节中，我们探索和评估了 gather-excite 框架的多种可能实例化。为了比较每种设计的实用性，我们在使用 ImageNet 1K 数据集 [33] 的图像分类任务上进行了一系列实验。该数据集包含120万张训练图像和 验证图像。在接下来的实验中，所有模型都在训练集上进行训练，在验证集上进行评估。我们的研究基于流行的 ResNet-50 架构，该架构在该数据集上表现出良好的性能，并且已经证明可以有效地推广到其他多种领域 [9]。新模型是通过在每个 ResNet-50 构建块的残差分支与身份分支相加之前立即插入 gather-excite 操作符而形成的。这些模型从随机初始化 [10] 开始，使用动量为0.9的 SGD，以及每个包含256张图像的小批量，每个图像裁剪为 像素。初始学习率设置为0.1，每次损失停滞时（共三次）就将其降低10倍。模型通常总共训练大约300个周期（请注意，这比 [15] 中使用的固定100个周期的优化计划产生了更强的模型）。在所有实验中，我们报告了在 ImageNet 验证集上的单中心裁剪结果。

# 3.1 Parameter-free pairings

# 3.1 无参数配对

We first consider a collection of GE pairings which require no additional learnable parameters. We take the gather operator to be average pooling with varying extent ratios (the effect of changing the pooling operator is analysed in the suppl. material). The excite operator then resizes the aggregates, applies a sigmoid and multiplies the result with the input. Thus, each output feature map is computed as , where interp denotes resizing to the original input size via nearest neighbour interpolation. We refer to this model as GE- , where the notation is used to denote that the operator is parameter-free . A diagram illustrating how these operators are integrated into a residual unit can be found in Fig. 4 of the supplementary material.

我们首先考虑一组不需要额外可学习参数的GE配对。我们将聚集操作符 视为具有不同扩展比例的平均池化（改变池化操作符的效果在补充材料中进行分析）。激励操作符随后调整聚集的大小，应用sigmoid函数，并将结果与输入相乘。因此，每个输出特征图计算为 ，其中interp 表示通过最近邻插值将大小调整到原始输入尺寸。我们将这个模型称为GE- ，其中记法 用于表示该操作符是无参数的 。一个说明这些操作符如何集成到残差单元中的图示可以在补充材料的图4中找到。

|  | top-1 err. | top-5 err. | GFLOPs | #Params |
| --- | --- | --- | --- | --- |
| ResNet-50 (Baseline) | 23.30 | 6.55 | 3.86 | 25.6 M |
| GE- (stage2) | 23.29 | 6.50 | 3.86 | 28.0 M |
| GE- (stage3) | 22.70 | 6.24 | 3.86 | 27.2 M |
| GE- (stage4) | 22.50 | 6.20 | 3.86 | 26.8 M |
| GE- (all) | 22.00 | 5.87 | 3.87 | 31.2 M |

Table 1: Effect (error %) of inserting GE operators at different stages of the baseline architecture ResNet-50.

表1：在不同阶段插入GE操作符对基准架构ResNet-50的效果（错误%）

Spatial extent: This basic model allows us to test the central hypothesis of this paper, namely that providing the network with access to simple summaries of additional feature context improves the representational power of the network. To this end, our first experiment varies the spatial extent ratio of the GE- design: we consider values of , as well a global extent ratio using global average pooling. The results of this experiment are shown in Fig. 2 (left). Each increase in the extent ratio yields consistent improvements over the performance of the ResNet-50 baseline (23.30% top-1 error), with the global extent ratio achieving the strongest performance (22.14% top-1 error). This experiment suggests that even with a simple parameter-free approach, context-based modulation can strengthen the discriminative power of the network. Remarkably, this model is competitive with the much heavier ResNet-101 model (22.20% top-1 error). In all following experiments, except where noted otherwise, a global extent ratio is used.

空间范围：这个基本模型允许我们测试本文的核心假设，即提供网络访问附加特征上下文的简单摘要可以提升网络的表征能力。为此，我们的第一个实验改变了GE- 设计的空间范围比例：我们考虑了 的值，以及使用全局平均池化的全局范围比例。这个实验的结果显示在图2（左）中。每次增加范围比例都会在ResNet-50基线（23.30% top-1错误率）的性能上带来一致的改进，全局范围比例实现了最强的性能（22.14% top-1错误率）。这个实验表明，即使是简单的无参数方法，基于上下文的调制也可以增强网络的判别力。值得注意的是，这个模型与更重的ResNet-101模型（22.20% top-1错误率）具有竞争力。在所有后续实验中，除非另有说明，都使用了全局范围比例。

# 3.2 Parameterised pairings

# 3.2 参数化配对

We have seen that simple gather-excite operators without learned parameters can offer an effective mechanism for exploiting context. To further explore the design space for these pairings, we next consider the introduction of parameters into the gather function, . In this work, we propose to use strided depth-wise convolution as the gather operator, which applies spatial filters to independent channels of the input. We combine with the excite operator described in Sec. 3.1 and refer to this pairing as GE- .

我们已经看到，没有学习参数的简单收集-激发操作可以提供一个有效的机制来利用上下文。为了进一步探索这些配对的设计空间，我们在收集函数 中引入参数。在这项工作中，我们提议使用步进深度卷积作为收集操作符，它对输入的独立通道应用空间滤波器。我们将 与第3.1节中描述的激发操作符结合起来，并将这种配对称为GE- 。

Spatial extent: We begin by repeating the experiment to assess the effect of an increased extent ratio for the parameterised model. For parameter efficiency, varying extent ratios is achieved by chaining stride 2 depth-wise convolutions ( such convolutions are performed in total). For the global extent ratio, a single global depth-wise convolution is used. Fig. 2 (right) shows the results of this experiment. We observe a similar overall trend to the GE- study and note that the introduction of additional parameters brings expected improvements over the parameter-free design.

空间范围：我们首先通过重复实验来评估增加范围比例对参数化模型的影响。为了参数效率，通过链接 步长为 2 的深度可分卷积（总共执行了 这样的卷积）来实现不同的范围比例。对于全局范围比例，使用单个全局深度可分卷积。图 2（右）展示了这个实验的结果。我们观察到与 GE- 研究类似的总体趋势，并注意到引入额外的参数带来了预期的改进，超过了无参数设计。

Effect on different stages: We next investigate the influence of GE- on different stages (here we use the term "stage" as it is defined in [9]) of the network by training model variants in which the operators are inserted into each stage separately. The accuracy, computational cost and model complexity of the resulting models are shown in Tab. 1. While there is some improvement from insertion at every stage, the greatest improvement comes from the mid and late stages (where there are also more channels). The effects of insertion at different stages are not mutually redundant, in the sense that they can be combined effectively to further bolster performance. For simplicity, we include GE operators throughout the network in all remaining experiments, but we note that if parameter storage is an important concern, GE can be removed from Stage 2 at a marginal cost in performance.

对不同阶段的影响：接下来，我们通过训练在每个阶段单独插入操作符的模型变种，来研究 GE- 对网络不同阶段（在这里我们使用 [9] 中定义的“阶段”一词）的影响。结果模型的准确性、计算成本和模型复杂度在表 1 中显示。虽然在每个阶段插入都有些许改进，但最大的改进来自于中期和后期阶段（那里也有更多的通道）。不同阶段的插入效果并不是相互冗余的，在意义上它们可以有效地组合起来进一步强化性能。为了简化，我们在所有剩余实验中都在整个网络中包含 GE 操作符，但我们注意到如果参数存储是一个重要考虑因素，可以从第二阶段移除 GE，而性能损失很小。

Relationship to Squeeze-and-Excitation Networks: The recently proposed Squeeze-and-Excitation Networks [15] can be viewed as a particular GE pairing, in which the gather operator is a parameter-free operation (global average pooling) and the excite operator is a fully connected subnetwork. Given the strong performance of these networks (see [15] for details), a natural question arises: are the benefits of parameterising the gather operator complementary to increasing the capacity of the excite operator? To answer this question, we experiment with a further variant, GE- , which combines the GE- design with a convolutional channel subnetwork excite operator (supporting the use of variable spatial extent ratios). The parameterised excite operator thus takes the form interp , where matches the definition given in [15], with reduction ratio 16). The performance of the resulting model is given in Tab. 2. We observe that the model not only outperforms the and models, but approaches the performance of the considerably larger 152 layer ResNet at approximately one third of the computational complexity.

与 Squeeze-and-Excitation 网络的关系：最近提出的 Squeeze-and-Excitation 网络 [15] 可以看作是一种特殊的 GE 配对，其中聚集操作是无参数的操作（全局平均池化），激发操作是一个全连接子网络。鉴于这些网络的表现强劲（详情见 [15]），一个自然的问题是：参数化聚集操作的好处是否与增加激发操作容量的好处互补？为了回答这个问题，我们实验了一种变种，GE- ，它将 GE- 设计与 卷积通道子网络激发操作相结合（支持使用可变空间范围比）。因此，参数化的激发操作形式为 interp ，其中 符合 [15] 中给出的定义，具有 16 的缩减比。表 2 给出了所得模型的性能。我们观察到 模型不仅超过了 和 模型，而且接近于计算复杂度大约为其三分之一的 152 层 ResNet 的性能。

Throughout this work, we use the term "parameter-free" to denote a model that requires no additional learnable parameters. Under this definition, average pooling and nearest neighbour interpolation are parameter-free operations.

在本工作中，我们使用 "无参数" 这个术语来表示不需要额外可学习参数的模型。根据这个定义，平均池化和最近邻插值是无参数操作。

|  | top-1 err. | top-5 err. | GFLOPs | #Params |
| --- | --- | --- | --- | --- |
| ResNet-101 | 22.20 | 6.14 | 7.57 |  |
| ResNet-50 (Baseline) | 23.30 | 6.55 | 3.86 |  |
| SE | 22.12 | 5.99 | 3.87 | 28.1 M |
| GE- | 22.14 | 6.24 | 3.86 | 25.6 M |
| GE- | 22.00 | 5.87 | 3.87 | 31.2 M |
| GE- | 21.88 | 5.80 | 3.87 | 33.7 M |

Table 2: Comparison of differing GE configurations with a ResNet-50 baseline on the ImageNet validation set (error %) and their respective complexities. The ResNet-101 model is included for reference.

表 2：不同 GE 配置与 ResNet-50 基线在 ImageNet 验证集（错误%）及其各自复杂度上的比较。ResNet-101 模型仅供参考。

|  | top-1 err. | top-5 err. | GFLOPs | #Params |
| --- | --- | --- | --- | --- |
| ResNet-152 | 21.87 | 5.78 | 11.28 | 60.3 M |
| ResNet-101 (Baseline) | 22.20 | 6.14 | 7.57 | 44.6 M |
| SE | 20.94 | 5.50 | 7.58 | 49.4 M |
| GE- | 21.47 | 5.69 | 7.58 | 44.6 M |
| GE- | 21.46 | 5.45 | 7.59 | 53.7 M |
| GE- | 20.74 | 5.29 | 7.59 | 58.4 M |

Table 3: Comparison of differing GE configurations with a ResNet-101 baseline on the ImageNet validation set (error %) and their respective complexities. The GE- (101) model outperforms a deeper ResNet-152 (included above for reference).

表 3：不同 GE 配置与 ResNet-101 基线在 ImageNet 验证集（错误%）及其各自复杂度上的比较。GE- (101) 模型超过了更深的 ResNet-152（如上所述，供参考）。

# 3.3 Generalisation

# 3.3 泛化

Deeper networks: We next ask whether the improvements brought by incorporating GE operators are complementary to the benefits of increased network depth. To address this question, we train deeper ResNet-101 variants of the GE- , GE- and GE- designs. The results are reported in Tab. 3. It is important to note here that the GE operators themselves add layers to the architecture (thus this experiment does not control precisely for network depth). However, they do so in an extremely lightweight manner in comparison to the standard computational blocks that form the network and we observe that the improvements achieved by GE transfer to the deeper ResNet-101 baseline, suggesting that to a reasonable degree, these gains are complementary to increasing the depth of the underlying backbone network.

更深的网络：我们接下来询问，加入GE操作符带来的改进是否与增加网络深度的好处相互补充。为了回答这个问题，我们训练了更深层的ResNet-101变体，包括GE- ，GE- 和GE- 设计。结果报告在表3中。在这里需要注意的是，GE操作符本身向架构中添加了层（因此这个实验并没有精确控制网络深度）。然而，与构成网络的标准化计算块相比，它们以极其轻量级的方式这样做，并且我们观察到GE实现的改进转移到了更深的ResNet-101基线，这表明在合理的程度上，这些收益与增加底层骨干网络的深度是相互补充的。

Resource constrained architectures: We have seen that GE operators can strengthen deep residual network architectures. However, these models are largely composed of dense convolutional computational units. Driven by demand for mobile applications, a number of more sparsely connected architectures have recently been proposed with a view to achieving good performance under strict resource constraints . We would therefore like to assess how well GE generalises to such scenarios. To answer this question, we conduct a series of experiments on the ShuffleNet architecture [50], an efficient model that achieves a good tradeoff between accuracy and latency. Results are reported in Tab. 4. In practice, we found these models challenging to optimise and required longer training schedules to reproduce the performance of the baseline model reported in [50] (training curves under a fixed schedule are provided in the suppl. material). We also found it difficult to achieve improvements without the use of additional parameters. The GE- variants yield improvements in performance at a fairly modest theoretical computational complexity. In scenarios for which parameter storage represents the primary system constraint, a naive application of GE may

资源受限架构：我们已经看到GE操作符可以加强深度残差网络架构。然而，这些模型主要由密集的卷积计算单元组成。由于移动应用的需求，最近提出了一些连接更为稀疏的架构，目的是在严格的资源约束下实现良好的性能 。因此，我们希望评估GE在如此场景下泛化的效果如何。为了回答这个问题，我们在ShuffleNet架构 [50] 上进行了一系列实验，这是一个高效的模型，它在准确性和延迟之间取得了良好的平衡。结果报告在表4中。实际上，我们发现这些模型很难优化，并且需要更长的训练计划 来复现 [50] 中报告的基线模型的性能（在固定计划下的训练曲线提供在补充材料中）。我们还发现，在不使用额外参数的情况下很难实现改进。GE- 变体在相当温和的理论计算复杂度下实现了性能提升。在参数存储代表系统主要约束的场景中，GE的简单应用可能

| ShuffleNet variant | top-1 err. | top-5 err. | MFLOPs | #Params |
| --- | --- | --- | --- | --- |
| ShuffleNet (Baseline) | 32.60 | 12.40 | 137.5 | 1.9 M |
| SE | 31.24 | 11.38 | 139.9 | 2.5 M |
| GE- (E2) | 32.40 | 12.31 | 138.9 | 2.0 M |
| GE- (E4) | 32.32 | 12.24 | 139.1 | 2.1 M |
| GE- (E8) | 32.12 | 12.11 | 139.2 | 2.2 M |
| GE- | 31.80 | 11.98 | 140.8 | 3.6 M |
| GE- | 30.12 | 10.70 | 141.6 | 4.4 M |

Table 4: Comparison of differing GE configurations with a ShuffleNet baseline on the ImageNet validation set (error %) and their respective complexities. Here, ShuffleNet refers to "ShuffleNet " in [50].

表4：不同GE配置与ShuffleNet基线在ImageNet验证集（错误%）上的比较以及它们的复杂性。在此，ShuffleNet指的是文献[50]中的“ShuffleNet ”。

|  | ResNet-110 [10] | ResNet-164 [10] | WRN-16-8 [49] |
| --- | --- | --- | --- |
| Baseline |  |  |  |
| SE |  |  |  |
| GE- |  |  |  |
| GE- |  |  |  |
| GE- |  | 4.07 / 20.85 |  |

Table 5: Classification error (%) on the CIFAR-10/100 test set with standard data augmentation (padding 4 pixels on each side, random crop and flip).

表5：在CIFAR-10/100测试集上使用标准数据增强（每边填充4个像素，随机裁剪和翻转）的分类错误（%）。

be less appropriate and more care is needed to achieve a good tradeoff between accuracy and storage (this may be achieved, for example, by using GE at a subset of the layers).

可能不太合适，需要更多的关注以达到准确性和存储之间的良好权衡（例如，通过在层的子集上使用GE来实现）。

Beyond ImageNet: We next assess the ability of GE operators generalise to other datasets beyond ImageNet. To this end, we conduct additional experiments on the CIFAR-10 and CIFAR-100 image classification benchmarks [19]. These datasets consist of color images drawn from 10 classes and 100 classes respectively. Each contains 50k train images and 10k test images. We adopt a standard data augmentation scheme (as used in ) to facilitate a useful comparative analysis between models. During training, images are first zero-padded on each side with four pixels, then a random patch is produced from the padded image or its horizontal flip before applying mean/std normalization. We combine GE operators with several popular backbones for CIFAR: ResNet-110 [10], ResNet-164 [10] and the Wide Residual Network-16-8 [49]. The results are reported in Tab. 5. We observe that even on datasets with considerably different characteristics (e.g. pixels), GE still yields good performance gains.

超出ImageNet：接下来，我们评估GE操作符在其他数据集（如ImageNet之外）的泛化能力。为此，我们在CIFAR-10和CIFAR-100图像分类基准数据集[19]上进行了额外的实验。这些数据集分别包含 个来自10个类别和100个类别的彩色图像。每个数据集包含50k个训练图像和10k个测试图像。我们采用标准数据增强方案（如文献 中使用的那样）来进行有用的模型比较分析。在训练过程中，图像首先在每侧进行零填充，填充四个像素，然后从填充的图像或其水平翻转图像中产生一个随机的 补丁，在应用均值/标准差归一化之前。我们将GE操作符与CIFAR的几种流行主干网络相结合：ResNet-110 [10]、ResNet-164 [10]和Wide Residual Network-16-8 [49]。结果报告在表5中。我们观察到，即使在具有相当不同特征（例如 像素）的数据集上，GE仍然能够带来良好的性能提升。

Beyond image classification: We would like to evaluate whether GE operators can generalise to other tasks beyond image classification. For this purpose, we train an object detector on MS COCO [25], a dataset which has approximately 80k training images and 40k validation images (we use the train-val splits provided in the 2014 release). Our experiment uses the Faster R-CNN framework [32] (replacing the RoIPool operation with RoIAlign proposed in [8]) and otherwise follows the training settings in [9]. We train two variants: one with a ResNet-50 backbone and one with a GE- (E8) backbone, keeping all other settings fixed. The ResNet-50 baseline performance is mAP. Incorporating the GE- backbone improves the baseline performance to mAP.

超越图像分类：我们希望评估 GE 操作符是否能够推广到图像分类以外的其他任务。为此，我们在 MS COCO [25] 数据集上训练一个对象检测器，该数据集大约有 80k 张训练图像和 40k 张验证图像（我们使用 2014 版本中提供的训练-验证划分）。我们的实验使用 Faster R-CNN 框架 [32]（将 RoIPool 操作替换为 [8] 中提出的 RoIAlign），其余训练设置遵循 [9]。我们训练了两个变体：一个使用 ResNet-50 主干网络，另一个使用 GE- (E8) 主干网络，保持所有其他设置不变。ResNet-50 基线的性能是 mAP。加入 GE- 主干网络后，基线性能提升至 mAP。

# 4 Analysis and Discussion

# 4 分析与讨论

Effect on learned representations: We have seen that GE operators can improve the performance of a deep network for visual tasks and would like to gain some insight into how the learned features may differ from those found in the baseline ResNet-50 model. For this purpose, we use the class selectivity index metric introduced by [28] to analyse the features of these models. This metric computes, for each feature map, the difference between the highest class-conditional mean activity and the mean of all remaining class-conditional activities over a given data distribution. The resulting measurement is normalised such that it varies between zero and one, where one indicates that a filter only fires for a single class and zero indicates that the filter produced the same value for every class. The metric is of interest to our work because it provides some measure of the degree to which features are being

对学习表征的影响：我们已经看到 GE 操作符可以提升深度网络在视觉任务上的性能，并希望了解学习到的特征与基线 ResNet-50 模型中的特征有何不同。为此，我们使用了 [28] 引入的类别选择性指数指标来分析这些模型的特征。该指标计算每个特征图上最高类别条件均值活动与给定数据分布下所有其他类别条件活动的均值之间的差异。得到的测量结果归一化，使其在零和一之间变化，其中一表示过滤器仅对一个类别起作用，零表示过滤器对所有类别产生相同的值。这个指标对我们工作很有意义，因为它提供了一种衡量特征被多少程度区分开了度量。

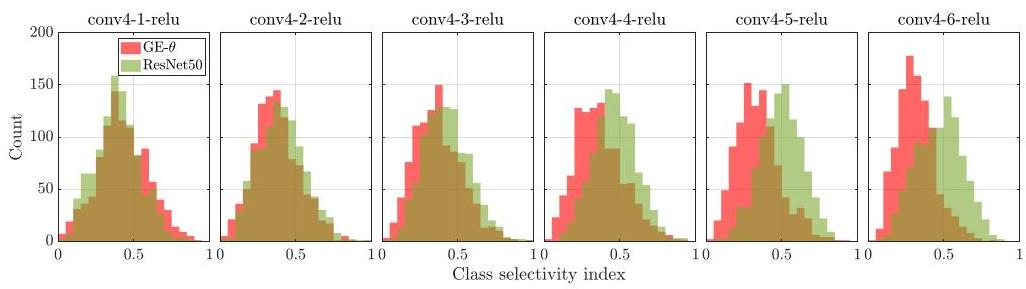


Figure 3: Each figure depicts the class selectivity index distribution for features in both the baseline ResNet-50 and corresponding GE- network at various blocks in the fourth stage of their architectures. As depth increases, we observe that the GE- model exhibits less class selectivity than the ResNet-50 baseline.

图3：每个图形展示了基线ResNet-50和相应GE- 网络在架构第四阶段的各个块中特征类别选择性指数分布。随着深度的增加，我们观察到GE- 模型比ResNet-50基线展现出更少的类别选择性。

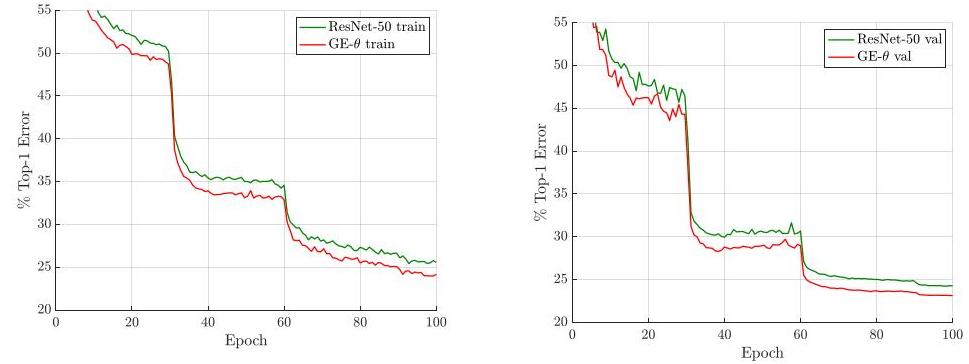


Figure 4: Top-1 error (%) on the ImageNet training set (left) and validation set (right) of the ResNet-50 baseline and proposed GE- (global extent) model under a fixed-length training schedule.

图4：在固定长度训练计划下，ResNet-50基线和提出的GE- （全局范围）模型在ImageNet训练集（左）和验证集（右）上的Top-1错误率（%）。

# shared across classes, a central property of distributed representations that can describe concepts efficiently [12].

# 在类别之间共享，这是分布式表示的一个中心属性，可以高效地描述概念 [12]。

We compute the class selectivity index for intermediate representations generated in the fourth stage (here we use the term "stage" as it is defined in [9]). The features of this stage have been shown to generalise well to other semantic tasks [31]. We compute class selectivity histograms for the last layer in each block in this stage of both models, and present the results of GE- and ResNet-50 in Fig. 3. An interesting trend emerges: in the early blocks of the stage, the distribution of class selectivity for both models appears to be closely matched. However, with increasing depth, the distributions begin to separate, and by conv4-6-relu the distributions appear more distinct with GE- exhibiting less class selectivity than ResNet-50. Assuming that additional context may allow the network to better recognise patterns that would be locally ambiguous, we hypothesise that networks without access to such context are required to allocate a greater number of highly specialised units that are devoted to the resolution of these ambiguities, reducing feature re-use. Additional analyses of the SE and GE- models can be found in the suppl. material.

我们计算了第四阶段生成的中间表示的类别选择性指数（在这里我们使用“阶段”一词，如[9]中定义）。这一阶段特征已被证明可以很好地推广到其他语义任务 [31]。我们计算了此阶段每个块的最后层的类别选择性直方图，并在图3中展示了GE- 和ResNet-50的结果。一个有趣的趋势出现了：在阶段的早期块中，两个模型的类别选择性分布看起来非常接近。然而，随着深度的增加，分布开始分离，到了conv4-6-relu，分布看起来更加不同，GE- 展现出的类别选择性比ResNet-50少。假设额外的上下文可能允许网络更好地识别本地模糊的模式，我们假设没有访问此类上下文的网络需要分配更多的专门化单元来解析这些模糊性，减少特征重用。SE和GE- 模型的额外分析可以在补充材料中找到。

Effect on convergence: We explore how the usage of GE operators play a role in the optimisation of deep networks. For this experiment, we train both a baseline ResNet-50 and a GE- model (with global extent ratio) from scratch on ImageNet using a fixed 100 epoch schedule. The learning rate is initialised to 0.1 and decreased by a factor of 10 every 30 epochs. The results of this experiment are shown in Fig. 4. We observe that the GE- model achieves lower training and validation error throughout the course of the optimisation schedule. A similar trend was reported when training with SE blocks [15], which as noted in Sec. 3.2, can be interpreted as a parameter-free gather operator and a parameterised excite operator. By contrast, we found empirically that the GE- model does not exhibit the same ease of optimisation and takes longer to learn effective representations.

对收敛性的影响：我们探讨 GE 操作符的使用在深度网络优化中的作用。为此实验，我们从零开始训练了一个基线 ResNet-50 模型和一个带有全局扩展比例的 GE- 模型，在 ImageNet 上使用固定的 100 个训练周期。初始学习率设为 0.1，每 30 个周期减少 10 倍。本实验的结果显示在图 4 中。我们观察到在整个优化过程中，GE- 模型在训练和验证误差上都取得了更低的水平。在训练时使用 SE 块 [15] 也报告了类似趋势，如第 3.2 节所述，可以将 SE 块解释为无参数的聚集操作符和参数化的激发操作符。相比之下，我们从经验上发现 GE- 模型并不表现出同样的优化容易度，并且需要更长时间来学习有效的表示。

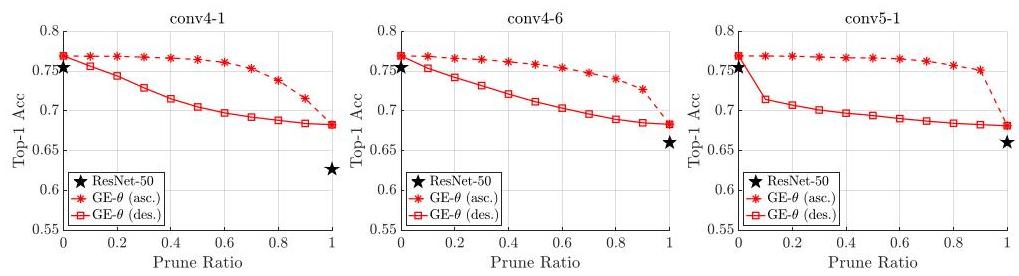


Figure 5: Top-1 ImageNet validation accuracy for the GE- model after dropping a given ratio of feature maps out the residual branch for each test image. Dashed line denotes the effect of dropping features with the least assigned importance scores first. Solid line denotes the effect of dropping features with the highest assigned importance scores first. For reference, the black stars indicate the importance of these feature blocks to the ResNet-50 model (see Sec. 4 for further details).

图 5：在移除给定比例的特征图后，GE- 模型在 ImageNet 上的 Top-1 验证准确率。虚线表示首先移除分配的重要性分数最低的特征的效果。实线表示首先移除分配的重要性分数最高的特征的效果。作为参考，黑色星号表示这些特征块对 ResNet-50 模型的重要性（详情见第 4 节）。

Feature importance and performance. The gating mechanism of the excite operator allows the network to perform feature selection throughout the learning process, using the feature importance scores that are assigned to the outputs of the gather operator. Features that are assigned a larger importance will be preserved, and those with lower importance will be squashed towards zero. While intuitively we might expect that feature importance is a good predictor of the contribution of a feature to the overall network performance, we would like to verify this relationship. We conduct experiments on a GE- network, based on the ResNet-50 architecture. We first examine the effect of pruning the least important features: given a building block of the models, for each test image we sort the channel importances induced by the gating mechanism in ascending order (labelled as "asc." in Fig. 5), and set a portion (the prune ratio) of the values to zero in a first-to-last manner. As the prune ratio increases, information flow flows through an increasingly small subset of features. Thus, no feature maps are dropped out when the prune ratio is equal to zero, and the whole residual branch is dropped out when the ratio is equal to one (i.e., the information of the identity branch passes through directly). We repeat this experiment in reverse order, dropping out the most important features first (this process is labelled "des." in Fig. 5). This experiment is repeated for three building blocks in GE- (experiments for SE are included in the suppl. material). As a reference for the relative importance of features contained in these residual branches, we additionally report the performance of the baseline ResNet-50 model with the prune ratio set to 0 and 1 respectively. We observe that preserving the features estimated as most important by the excite operator retains the much of the overall accuracy during the early part of the pruning process before an increasingly strong decay in performance occurs. When reversing the pruning order, the shape of this performance curve is inverted, suggesting a consistent positive correlation between the estimated feature importance and overall performance. This trend is clearest for the deeper conv5-1 block, indicating a stronger dependence between primary features and concepts, which is consistent with findings in previous work . While these feature importance estimates are instance-specific, they can also be used to probe the relationships between classes and different features [15], and may potentially be useful as a tool for interpreting the activations of networks.

特征重要性与性能。激励操作符的 gating 机制允许网络在整个学习过程中执行特征选择，使用分配给 gather 操作符输出的特征重要性得分。被分配较大重要性的特征将被保留，而那些重要性较低的特征将被压缩趋近于零。直观上，我们可能会期望特征重要性是特征对整体网络性能贡献的良好预测指标，我们希望验证这种关系。我们在基于 ResNet-50 架构的 GE- 网络上进行实验。首先，我们检查修剪最不重要特征的影响：对于模型的每个构建块，我们对由 gating 机制引起的每个测试图像的通道重要性进行升序排序（在图 5 中标记为 "asc."），并以从先到后的方式将一部分（修剪比例）的值设为零。随着修剪比例的增加，信息流通过的特征子集越来越小。因此，当修剪比例等于零时，没有特征图被丢弃，而当比例等于一时（即，恒等分支的信息直接通过），整个残差分支被丢弃。我们以相反的顺序重复这个实验，首先丢弃最重要的特征（这个过程在图 5 中标记为 "des."）。这个实验在 GE- 的三个构建块中重复进行（SE 的实验包括在补充材料中）。作为这些残差分支中特征相对重要性的参考，我们还报告了基线 ResNet-50 模型在修剪比例设置为 0 和 1 时的性能。我们观察到，在修剪过程的早期阶段，保留由激励操作符估计为最重要的特征可以保持大部分整体准确性，然后性能出现逐渐增强的衰减。当反向修剪顺序时，这个性能曲线的形状是颠倒的，表明估计的特征重要性与整体性能之间存在一致的正相关关系。这种趋势在更深的 conv5-1 块中最明显，表明主要特征与概念之间的依赖性更强，这与以前工作 的发现一致。虽然这些特征重要性估计是特定于实例的，但它们也可以用来探测类与不同特征之间的关系 [15]，并且可能作为解释网络激活的工具具有潜在用途。

# 5 Related Work

# 5 相关工作

Context-based features have a rich history of use in computer vision, motivated by studies in perception that have shown that contextual information influences the accuracy and efficiency of object recognition and detection by humans . Several pioneering automated vision systems incorporated context as a component of sophisticated rule-based approaches to image understanding [36, 7]; for tasks such as object recognition and detection, low-dimensional, global descriptors have often proven effective as contextual clues . Alternative approaches based on graphical models represent another viable mechanism for exploiting context and many other forms of contextual features have been proposed [6]. A number of works have incorporated context for improving semantic segmentation (e.g. [48, 23]), and in particular, ParseNet [26] showed that encoding context through global feature averaging can be highly effective for this task.

基于上下文的特征在计算机视觉领域有着丰富的应用历史，这一应用受到了感知研究中关于上下文信息影响人类对象识别和检测的准确性和效率的启发 。一些开创性的自动视觉系统将上下文作为复杂规则基础图像理解方法的一个组成部分 [36, 7]；在诸如对象识别和检测等任务中，低维的全局描述符常常被证明是有效的上下文线索 。基于图形模型的替代方法代表了另一种利用上下文的可行机制 ，并且已经提出了许多其他形式的上下文特征 [6]。许多研究已经将上下文整合到提高语义分割中（例如 [48, 23]），特别是ParseNet [26] 显示了通过全局特征平均编码上下文对于这项任务可以非常有效。

The Inception family of architectures popularised the use of multi-scale convolutional modules, which help ensure the efficient aggregation of context throughout the hierarchy of learned representations [17]. Variants of these modules have emerged in recent work on automated architecture search [51], suggesting that they are components of (at least) a local optimum in the current design space of network blocks. Recent work has developed both powerful and generic parame-terised attention modules to allow the system to extract informative signals dynamically . Top-down attention modules [42] and self-attention [41] can be used to exploit global relationships between features. By reweighting features as a generic function of all pairwise interactions, non-local networks [43] showed that self-attention can be generalised to a broad family of global operator blocks useful for visual tasks.

Inception架构家族 推广了多尺度卷积模块的使用，这些模块有助于确保在整个学习到的表征层次中有效地聚合上下文 [17]。这些模块的变体已经在最近的自动架构搜索工作中出现 [51]，这表明它们至少是当前网络块设计空间中局部最优的组成部分。最近的工作已经开发出了强大且通用的参数化注意力模块，以允许系统动态地提取信息信号 。自顶向下的注意力模块 [42] 和自注意力 [41] 可以用来利用特征之间的全局关系。通过将特征作为所有成对交互的通用函数进行重新加权，非局部网络 [43] 显示了自注意力可以推广到一系列适用于视觉任务的全局操作块。

There has also been considerable recent interest in developing more specialised, lightweight modules that can be cheaply integrated into existing designs. Our work builds on the ideas developed in Squeeze-and-Excitation networks [15], which used global embeddings as part of the SE block design to provide context to the recalibration function. We draw particular inspiration from the studies conducted in [44], which showed that useful contextual information for localising objects can be inferred in a feed-forward manner from simple summaries of basic image descriptors (our aim is to incorporate such summaries of low, mid and high level features throughout the model). In particular, we take the SE emphasis on lightweight contextual mechanisms to its logical extreme, showing that strong performance gains can be achieved by the GE- variant with no additional learnable parameters. We note that similar parameterised computational mechanisms have also been explored in the image restoration community [18], providing an interesting alternative interpretation of this family of module designs as learnable activation functions.

近年来，开发更专业、轻量级的模块以低成本集成到现有设计中受到了相当大的关注。我们的工作建立在 Squeeze-and-Excitation 网络 [15] 的思想之上，该网络在 SE 块设计中使用全局嵌入作为重新校准函数的上下文信息。我们特别从 [44] 的研究中汲取灵感，该研究显示，可以通过简单的基本图像描述子的摘要以前馈方式推断出用于定位物体的有用上下文信息（我们的目标是在整个模型中整合这些低级、中级和高级特征的摘要）。特别是，我们将 SE 对轻量级上下文机制的强调推向逻辑极致，证明了通过 GE- 变体（无需额外可学习参数）可以实现强烈的性能提升。我们注意到，图像修复社区 [18] 也探索了类似的参数化计算机制，为这类模块设计的可学习激活函数提供了有趣的替代解释。

# 6 Conclusion and Future Work

# 6 结论与未来工作

In this work we considered the question of how to efficiently exploit feature context in CNNs. We proposed the gather-excite (GE) framework to address this issue and provided experimental evidence that demonstrates the effectiveness of this approach across multiple datasets and model architectures. In future work we plan to investigate whether gather-excite operators may prove useful in other computer vision tasks such as semantic segmentation, which we anticipate may also benefit from efficient use of feature context.

在这项工作中，我们考虑了如何在卷积神经网络（CNNs）中高效利用特征上下文的问题。我们提出了聚集-激励（GE）框架来解决这个问题，并通过实验证明了这种方法在多个数据集和模型架构中的有效性。在未来的工作中，我们计划研究聚集-激励操作符在其他计算机视觉任务中（如语义分割）是否也证明是有用的，我们预计这些任务也可能从特征上下文的有效利用中受益。

Acknowledgments. The authors would like to thank Andrew Zisserman and Aravindh Mahendran for many helpful discussions. Samuel Albanie is supported by ESPRC AIMS CDT. Andrea Vedaldi is supported by ERC 638009-IDIU.

致谢。作者们想要感谢 Andrew Zisserman 和 Aravindh Mahendran 的许多有益讨论。Samuel Albanie 由 ESPRC AIMS CDT 支持。Andrea Vedaldi 由 ERC 638009-IDIU 支持。

# References

# 参考文献

[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In CVPR, 2016.

[2] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive psychology, 1982.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI, 2018.

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, 2017.

[5] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In ECCV Workshop, 2004.

[6] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In CVPR, 2009.

[7] A Hanson. Visions: A computer system for interpreting scenes. Computer vision systems, 1978.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In ICCV, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In ECCV, 2016.

[11] Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In ECCV, 2008.

[12] Geoffrey E Hinton, James L McClelland, David E Rumelhart, et al. Distributed representations. Parallel distributed processing: Explorations in the microstructure of cognition, 1986.

[13] Howard S Hock, Gregory P Gordon, and Robert Whitehurst. Contextual relations: the influence of familiarity, physical plausibility, and belongingness. Perception & Psychophysics, 1974.

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, 2018.

[16] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In ECCV, 2016.

[17] Tsung-Wei Ke, Michael Maire, and X Yu Stella. Multigrid neural architectures. In CVPR, 2017.

[18] Idan Kligvasser, Tamar Rott Shaham, and Tomer Michaeli. xUnit: Learning a spatial activation function for efficient image restoration. In CVPR, 2018.

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. Tech Report, 2009.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998.

[22] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In ICML, 2009.

[23] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In CVPR, 2016.

[24] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In ICLR, 2014.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[26] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. In ICLR workshop, 2016.

[27] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In NIPS, 2016.

[28] Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In ICLR, 2018.

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In CVPR, 2014.

[30] Kevin P Murphy, Antonio Torralba, and William T Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In NIPS, 2004.

[31] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In CVPR, 2018.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. ImageNet large scale visual recognition challenge. IJCV, 2015.

[34] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. IJCV, 2013.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[36] Thomas M Strat and Martin A Fischler. Context-based vision: recognizing objects using information from both 2D and 3D imagery. IEEE TPMI, 1991.

[37] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In ICLR Workshop, 2016.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In CVPR, 2015.

[39] Antonio Torralba. Contextual priming for object detection. IJCV, 2003.

[40] Antonio Torralba, Kevin P Murphy, William T Freeman, Mark A Rubin, et al. Context-based vision system for place and object recognition. In ICCV, 2003.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.

[42] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In CVPR, 2017.

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In CVPR, 2018.

[44] Lior Wolf and Stanley Bileschi. A critical view of context. IJCV, 2006.

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In ECCV, 2018.

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.

[47] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In MIR, 2007.

[48] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016.

[49] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In BMVC, 2016.

[50] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In CVPR, 2018.

[51] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In CVPR, 2018.

# A Appendix

# A 附录

Optimization curves. In Fig. 6 we show the effect of training ShuffleNet and the GE- variant under a fixed, 100 epoch schedule to allow a direct comparison. As noted in main paper, for the results reported in Sec. 3 we used a longer training schedule to reproduce the baseline ShuffleNet performance.

优化曲线。在图6中，我们展示了在固定的100个训练周期计划下训练ShuffleNet及其GE- 变体的影响，以便进行直接比较。如主论文中所述，为了重现基线ShuffleNet的性能，我们在第3节报告的结果中使用了更长的训练计划 。

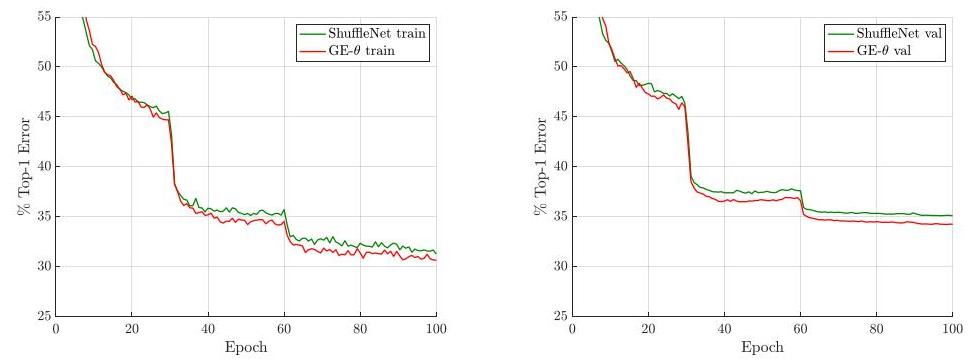


Figure 6: Top-1 error (%) on (left) the ImageNet training set and (right) the ImageNet validation set of the ShuffleNet Baseline and GE- variant of this architecture (with global extent) trained with a fixed 100 epoch schedule.

图6：在固定的100个训练周期计划下训练的ShuffleNet基线和GE- 变体（具有全局范围）在ImageNet训练集（左）和ImageNet验证集（右）上的Top-1错误率（%）。

| GE- variant | top-1 err. | top-5 err. |
| --- | --- | --- |
| ResNet-50 | 23.30 | 6.55 |
| (E4, max) | 23.10 | 6.39 |
| (E4, avg) | 22.87 | 6.40 |
| (global, max) | 23.65 | 6.62 |
| (global, avg) | 22.14 | 6.24 |

Table 6: Influence (error %) of different pooling operators for parameter-free GE designs. Each GE- variant is described by its (extent-ratio, pooling-type) pair.

表6：不同池化操作符对参数免费GE设计的影响（错误%）。每个GE- 变体都由其（范围-比例，池化类型）对描述。

Pooling method. We conduct additional experiments to assess the influence of the pooling method for GE- networks (shown in Tab. 6). Average pooling aggregates the neighbouring elements with equal contribution, while max pooling picks a single element to represent its neighbours. We observe that average pooling consistently outperforms max pooling. While adding contextual information by max pooling over a moderate extent can help the baseline model, it hurts the performance when the full global extent is used. This suggests that in contrast to averaging, a naive, parameter-free application of max pooling makes poor use of contextual information and inhibits, rather than facilitates, its efficient propagation. However, we note that when additional learnable parameters are introduced, this may no longer be the case: an interesting study presented in [45] has shown that there can be benefits to combining the output of both max and average pooling.

池化方法。我们进行了额外的实验来评估池化方法对GE- 网络的影响（如表6所示）。平均池化以相等的贡献聚合邻近元素，而最大池化选择一个元素来代表其邻近元素。我们观察到平均池化始终优于最大池化。虽然在适度范围内使用最大池化添加上下文信息可以帮助基线模型，但在使用完整全局范围时则会损害性能。这表明，与平均池化相比，简单、参数免费的 最大池化应用对上下文信息的利用较差，并且会抑制而不是促进其有效传播。然而，我们注意到，当引入额外的可学习参数时，情况可能不再如此：一篇有趣的论文[45]表明，结合最大池化和平均池化的输出可能是有益的。

Class selectivity indices. Following the approach described in Sec. 4 of the paper, we compute histograms of the class selectivity indices for and models and compare them with the baseline ResNet-50 in Fig. 7. We observe a weaker, but similar trend to the histograms of GE- reported in the main paper, characterised by a gradually emerging gap between the histograms of features at greater depth in stage four.

类选择指数。遵循论文第4节中描述的方法，我们计算了 和 模型的类选择指数直方图，并与图7中的基线 ResNet-50 进行比较。我们观察到一种较弱但与主论文中报告的 GE- 直方图相似的趋势，特征在于第四阶段深度较大的特征直方图之间逐渐出现的差距。

Feature importance. We repeat the pruning experiments described in Sec. 4 of the main paper (under the paragraph entitled Feature importance and performance) for an SE network based on a ResNet-50 backbone. The results of this experiment are reported in Fig. 8. We observe that the curves broadly match the trends seen in the GE- curves depicted in the main paper.

特征重要性。我们重复了主论文第4节中描述的剪枝实验（在标题为特征重要性与性能的段落下），针对基于 ResNet-50 骨架的 SE 网络。该实验的结果报告在图8中。我们观察到曲线大致符合主论文中描绘的 GE- 曲线的趋势。

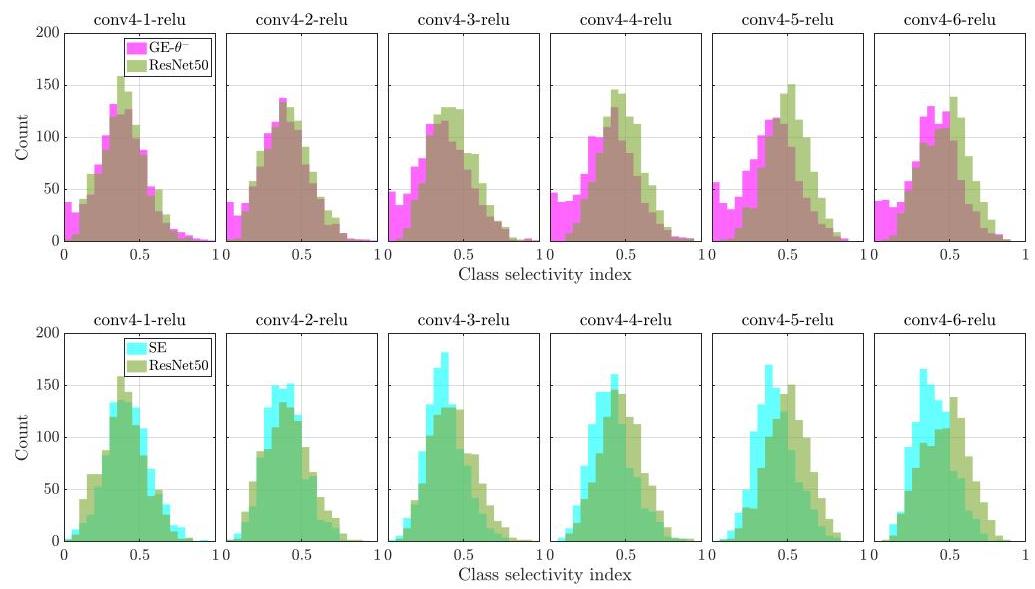


Figure 7: Each figure compares the class selectivity index distribution of the features of ResNet-50 against the GE- (top row) and SE (bottom row) networks at various blocks in the fourth stage of their architectures.

图7：每个图表比较了 ResNet-50 的特征与 GE- （顶部行）和 SE（底部行）网络在它们架构的第四阶段的各个块中的类选择指数分布。

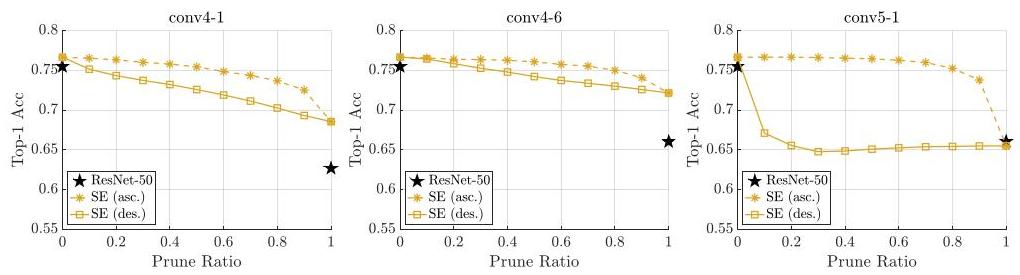


Figure 8: Top-1 ImageNet validation accuracy for the SE model after dropping a ratio of feature maps out for each test image. Dashed lines denote the effect of dropping features with the least assigned importance scores first. Solid lines denote the effect of dropping features with the highest assigned importance scores first. For reference, the black stars indicate the importance of these feature blocks to the ResNet-50 model.

图8：对于每个测试图像丢弃一定比例的特征图后，SE模型的 ImageNet 验证集上的 Top-1 准确率。虚线表示首先丢弃分配的重要性分数最低的特征的效果。实线表示首先丢弃分配的重要性分数最高的特征的效果。作为参考，黑色星号表示这些特征块对 ResNet-50 模型的重要性。

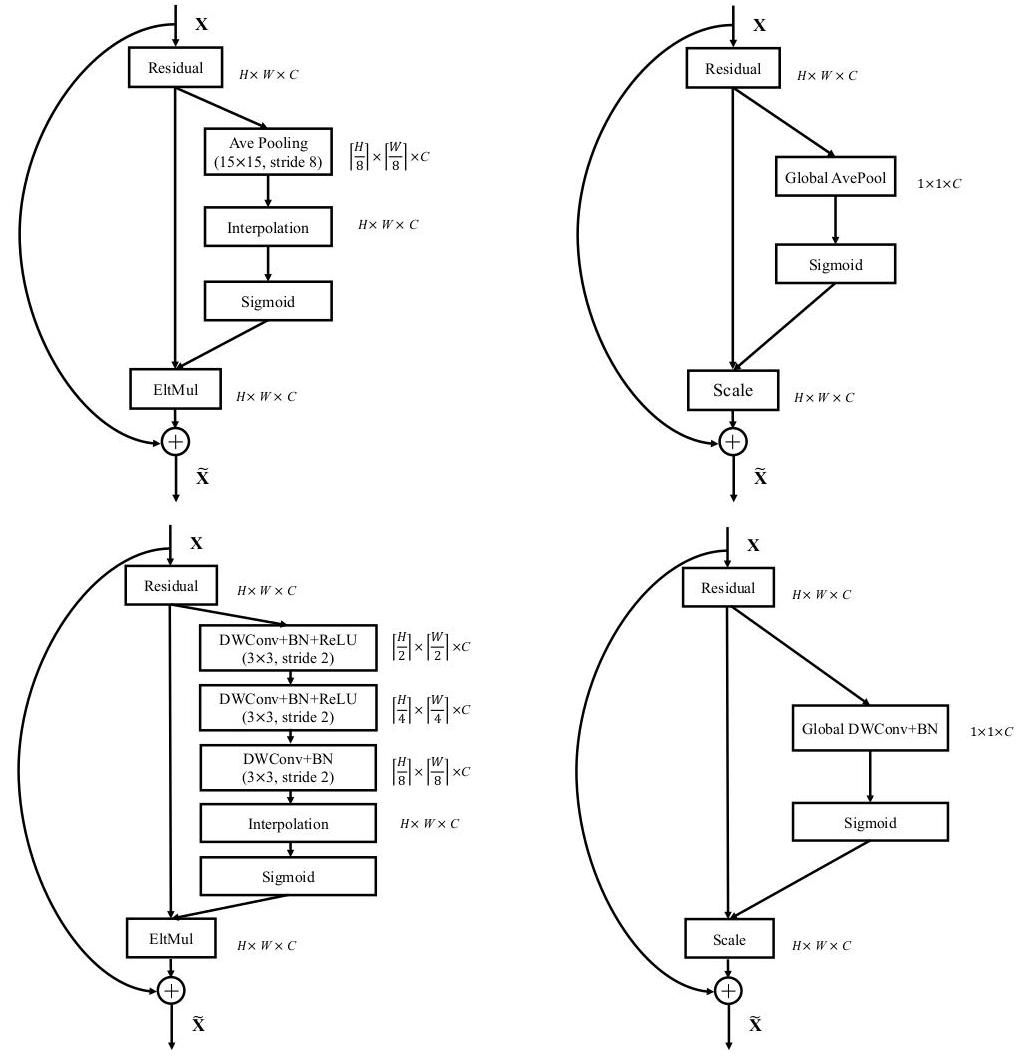


Figure 9: The schema of Gather-Excite modules. Top-left: GE- (E8). Top-right: GE- . Bottom-left: GE- (E8). Bottom-right: GE- .

图9： Gather-Excite 模块的架构图。左上角：GE- (E8)。右上角：GE- 。左下角：GE- (E8)。右下角：GE- 。

Operator diagrams. In Fig. 9, we illustrate diagrams of several of the GE variants described in the main paper, showing how they are integrated into residual units.

运算符图。在图9中，我们展示了主论文中描述的 GE 变体的几个运算符图，显示了它们如何集成到残差单元中。