엣지 디바이스에서 객체 탐지를 위한 그룹별 어탠션 기반 경량 디코더 연구

티엔투고[1], 엠디 딜로와르 호씬, 허의남[1]

[1]경희대학교 대학원 - 컴퓨터 과학과 공학과

thu.ngo@khu.ac.kr, delowar.cit@gmail.com, johnhuh@thu.ac.kr

A group-wise attention based decoder for lightweight salient object detection on edge-devices

Thien-Thu Ngo[1], Md Delowar Hossain[1], Eui-Nam Huh[1]
[1]Dept. of Computer Science and Engineering, Kyung Hee University

요 약

The recent scholarly focus has been directed towards the expeditious and accurate detection of salient objects, a task that poses considerable challenges for resource-limited edge devices due to the high computational demands of existing models. To mitigate this issue, some contemporary research has favored inference speed at the expense of accuracy. In an effort to reconcile the intrinsic trade-off between accuracy and computational efficiency, we present novel model for salient object detection. Our model incorporate group-wise attentive module within the decoder of the encoder-decoder framework, with the aim of minimizing computational overhead while preserving detection accuracy. Additionally, the proposed architectural design employs attention mechanisms to generate boundary information and semantic features pertinent to the salient objects. Through various experimentation across five distinct datasets, we have empirically substantiated that our proposed models achieve performance metrics comparable to those of computationally intensive state-of-the-art models, yet with a marked reduction in computational complexity.

1. Introduction

Salient object detection (SOD), advanced by the developments in deep learning, and has become an indispensable component in a multitude of vision-based applications. The methodology aims to emulate human visual attention mechanisms by distinguishing salient objects from their backgrounds and providing pixel-level saliency evaluations. Consequently, SOD has found applications in diverse fields, including but not limited to, autonomous vehicles, content-centric retrieval, image editing, and compression [1]. Given the pervasive adoption of SOD in consumer technologies, there is a pressing need to integrate it into resource-limited edge devices characteristic of the Internet of Things (IoT) landscape, such as smartphones and robotic systems. Achieving both real-time performance and high accuracy in these contexts presents considerable challenges. Therefore, there exists an unequivocal requirement for the development of adaptable and efficient SOD architectures suitable for streamlined deployment.

Contemporary state-of-the-art SOD models, while delivering exceptional performance, are often characterized by substantial computational demands and elevated memory usage. These attributes pose considerable impediments to their efficient deployment on edge computing devices. One primary contributor to this computational burden is the complexity of the decoder, particularly the computational overhead incurred across both shallow and deep layers, which results in protracted training durations and elevated inference costs. In response to these computational challenges, numerous researchers have concentrated their efforts on the development of computationally efficient SOD models. Such efficiency is realized through either the employment of streamlined model architectures [2] or the application of knowledge distillation techniques [3]. Nevertheless, some of these optimized models prioritize rapid inference and reduced model dimensions at the expense of accuracy, a compromise that could compromise service quality, especially in applications with stringent safety requirements. To solve these problems, we introduce a new framework that introduce group-wise attentive features in the decoder to reduce the complexity of the whole framework,

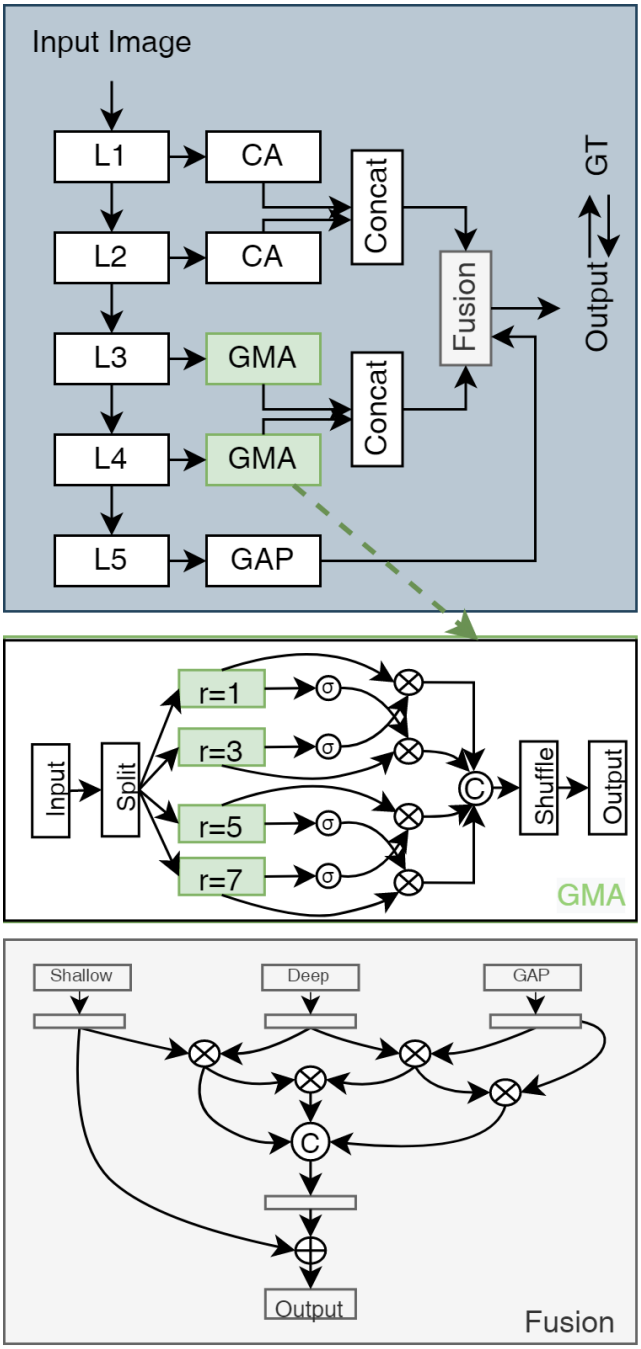

Figure 1: The architecture of the proposed framework

and employ attention mechanism to generate the boundary information and semantic features from shallow to deep layers. The experiments on five benchmark datasets shows that our proposal can significantly reduce the complexity in terms of FLOPS, Number of parameters while maintain the accuracy.

2. Related Works

a) Deep learning based salient object detection

Recent advancements in CNNs have yielded models that surpass traditional handcrafted methods, chiefly attributable to their enhanced capability for robust semantic feature extraction [4]. However, these CNN-based architectures are not without substantial limitations, such as increased computational complexity, extended inference latency, and expansive model size. These constraints present significant impediments to the effective deployment of such large-scale algorithms in edge computing environments.

b) Attention mechanism

The incorporation of attention mechanisms serves as a cornerstone in the realm of vision-related tasks, most notably in the specialized field of SOD [5]. Within the scope of SOD, attention mechanisms are generally categorized into spatial and channel-based categories. Spatial attention is geared towards the identification of image regions with a high likelihood of containing salient objects. This is commonly realized through the assignment of varying weights to disparate areas of the feature map, thereby directing the model's focus towards regions of probable saliency. Conversely, channel-based attention aims to discern the most informative feature channels pertinent to salient object detection, often by allocating greater weights to channels that exhibit higher discriminative power for the task. The integration of these attention mechanisms enables SOD models to adeptly capture complex interrelationships between diverse spatial locations and feature channels.

3. The proposed framework

Upon a comprehensive analysis of the constraints inherent in existing computationally intensive models, we introduce a novel framework designed to seamlessly incorporate group-wise features into an attention-based decoder, thereby enhancing computational efficiency and reducing overhead. As delineated in Figure 1, the architecture of our proposed network is predicated upon a ResNet-50 backbone. Notably, we have opted to omit the fully connected layers and the terminal pooling operation to tailor the backbone specifically for the requirements of SOD.

In the initial shallow layers of the network, we employ the Channel Attention mechanism from the Convolutional Block Attention Module (CBAM) [6] to augment the inter-channel feature correlations, thereby preserving essential boundary information. This process is mathematically represented as:

$$CA = BN(FC(AvgPool\ (L)))$$

Here, AvgPool denotes global average pooling, FC signifies the dimensional reduction applied to the output of AvgPool, and BN is Batch Normalization..

Subsequently, in the deeper layers, we introduce a Group-wise Multi-Scale Attention (GMA) module. In this module, dilated convolutions with varying kernel sizes are utilized to extract features at multiple scales. Each input within these deep layers is partitioned into four sub-features Fj, where j∈

{1,2,3,4}. Dilated convolutions with distinct dilation rates r ∈{1,3,5,7} are then applied to broaden the receptive field within each sub-branch. The attention weights for each branch are computed as follows:

$$\mathrm{Wj} = Sig(DW\mathrm{r}(\mathrm{Fj})$$

Here, DW represents the depth-wise convolutional layer, and Sig is the sigmoid activation function. Element-wise multiplication is subsequently performed to combine the attention weights with the sub-features from the four branches. Specifically, the features with dilation rates 1 and 3 are complemented, and a similar operation is conducted for features with dilation rates 5 and 7. The four sub-features are then concatenated and subjected to shuffle operations to produce the final output, represented as:

$$GMA = Shuffle(Concat(W_j * F_j)$$

Ultimately, the features from both shallow and deep layers are concatenated and forwarded to a Fusion module for final feature aggregation prior to output generation.

framework with a ResNet-50 backbone. All input images underwent resizing to a uniform dimension of 224x224 pixels and were subjected to a series of data augmentation techniques, including random cropping, flipping, and normalization. The Adam optimizer was employed for model training on the DUTS dataset, with an initial learning rate set at 5×10^{-5}, which was subsequently reduced by a factor of 10 after every 15 epochs. The weight decay parameter was configured at 5×10^{-4}.

5. Performance Analysis

In this section, we undertake a rigorous evaluation of the proposed framework's effectiveness relative to existing state-of-the-art methodologies. For the purpose of this performance assessment, we have selected a range of models and procured their corresponding saliency maps from their official project repositories. Specifically, the proposed framework is compared against ten state-of-the-art models, namely: RFCN [7], NLDF [8], DSS [9], PicaNet [10], DGRL [11], PAGR [12], C2S [13], RAS [14], R3Net [15], and EGNet [16].

In Table 1, we furnish a meticulous comparative analysis between the proposed methodology and other selected state-of-the-art approaches, employing evaluative metrics such as

Models	Flops (G)	Params (M)	FPS	ECSSD			PASCAL-S			DUTS-TE			DUT-OMRON			HKU-IS		
				$F\uparrow$	$S\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$M\downarrow$
Heavyweight Models																		
RFCN	102.8	134.69	0.4	.898	.859	.095	.837	.808	.118	.782	.792	.089	.738	.773	.094	.894	.858	.079
NLDF	263.9	35.49	18.5	.905	.875	.063	.831	.803	.099	.812	.815	.065	.753	.770	.079	.902	.879	.048
DSS	114.6	62.23	7.0	.921	.882	.052	.831	.797	.093	.830	.822	.052	.781	.788	.063	.916	.879	.040
PicaNet	37.1	32.85	5.6	.935	.917	.046	.857	.853	.076	.860	.868	.050	.803	.831	.065	.919	.904	.044
DGRL	24	126.35	3.6	.922	.903	.041	.854	.836	.072	.829	.841	.050	.774	.806	.062	.910	.895	.036
PAGR	100.4	23.63	-	.927	.889	.061	.856	.818	.093	.855	.837	.056	.771	.775	.071	.918	.887	.048
C2S	20.5	137.03	25	.896	.881	.059	.829	.826	.087	.790	.817	.066	.733	.779	.079	.883	.872	.051
RAS	35.6	21.23	20	.921	.892	.056	.829	.798	.101	.831	.838	.059	.799	.825	.058	.927	.906	.036
R3Net	56.16	60.24	-	.934	.910	.040	.835	.806	.092	-	-	-	.795	.816	.063	.915	.895	.035
EGNet	270.8	108.07	12.7	.947	.924	.037	.865	.852	.074	.889	.887	.039	.815	.840	.053	.935	.917	.031
Ours	**9.07**	**30.70**	**30**	**.935**	**.915**	**.039**	**.855**	**.848**	**.070**	**.853**	**.866**	**.046**	**.789**	**.821**	**.064**	**.922**	**.907**	**.036**

Table 1: We compared the performance of the proposed framework with state-of-the-art heavyweight models using three metrics: maximum F-measure (↑), S-measure (higher is better) and Mean Square Error (↓) (smaller is better).

4. Experiments Results

To rigorously assess the performance of the proposed model, evaluations were conducted across five benchmark datasets, namely ECSSD, HKUIS, DUTOMRON, DUTS-TE, and PASCAL-S, which contain 1000, 4447, 5168, 5019, and 850 images and their corresponding labels, respectively. For the purpose of accuracy quantification, a comprehensive set of metrics was employed, including F-measure (F), mean absolute error (MAE), and S-measure. To ascertain computational complexity, three specific metrics were utilized: the number of parameters, denominated in millions as Param (M); the number of floating-point operations, quantified in Giga FLOPS (G); and the inference speed, measured in frames per second (FPS).

The training and testing procedures were executed on an NVIDIA GTX 1080 GPU, leveraging the PyTorch 3.8

F-measure, MAE, and S-measure. The proposed model demonstrates performance metrics that are congruent with other robust models, while concurrently achieving substantial reductions in both the number of parameters and FLOPS, as well as attaining real-time inference speed (FPS). Notably, our model outperforms nine of the ten robust models in evaluative metrics. In terms of computational efficiency, the proposed model surpasses the others heavyweight models (EGNet, RFCN, NLDF, PAGR), realizing around 90%~96% reducing in FLOPS and around 50~70% in number of parameters. Particularly, it sustains performance metrics commensurate with EGNet, while effecting an approximate 80% reduction in FLOPS and manifesting an inference speed that is 14-fold faster (FPS). These comparative assessments corroborate the model's successful reconciliation of accuracy and computational efficiency.

6. Conclusion

In the present study, we have unveiled a novel framework that seamlessly integrates group-wise attentive features into an encoder-decoder architecture specifically designed for the task of salient object detection. By capitalizing on the attention mechanism to synthesize features at both shallow and deep layers, the proposed framework attains performance metrics that are commensurate with existing state-of-the-art models, while substantially mitigating computational overhead, as evidenced by reductions in the number of parameters, frames per second (FPS), and floating-point operations per second (FLOPS). The introduction of the proposed models substantiates the efficacy of employing group-wise attentive features as a promising strategy for reducing computational complexity in the development of lightweight salient object detection systems.

Acknowledgement

This work was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00258649) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2202-0-00047, Development of Microservices Development/Operation Platform Technology that Supports Application Service Operation Intelligence).

References

[1] Borji, A., Cheng, M.M., Hou, Q., Jiang, H. and Li, J., 2019. Salient object detection: A survey. Computational visual media, 5, pp.117-150.

[2] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection," IEEE Transactions on Image Processing, vol. 30, pp. 3804–3814, 2021.

[3] Zhang, J., Liang, Q. and Shi, Y., 2022. KD-SCFNet: Towards more accurate and efficient salient object detection via knowledge distillation. arXiv preprint arXiv:2208.02178

[4] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H. and Yang, R., 2021. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(6), pp.3239-3259.

[5] Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M. and Hu, S.M., 2022. Attention mechanisms in computer vision: A survey. *Computational visual media*, *8*(3), pp.331-368.

[6] Woo, S., Park, J., Lee, J.Y. and Kweon, I.S., 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).

[7] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in European conference on computer vision, pp. 825–841, Springer, 2016.

[8] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Nonlocal deep features for salient object detection," in Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 6609–6617, 2017

[9] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3203–3212, 2017

[10] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3089–3098, 2018

[11] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, Detect globally, refine locally: A novel approach to saliency detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3127–3135, 2018

[12] Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G., 2018a. Progressive attention guided recurrent network for salient object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp.714–722

[13] Li, X., Yang, F., Cheng, H., Liu, W., Shen, D., 2018. Contour knowledge transfer for salient object detection, in: Proceedings of the European conference on computer vision (ECCV), pp. 355–370

[14] Chen, S., Tan, X., Wang, B., Hu, X., 2018. Reverse attention for salient object detection, in: Proceedings of the European conference on computer vision (ECCV), pp. 234–250

[15] Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A., 2018. R3net: Recurrent residual refinement network for saliency detection, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press Menlo Park, CA, USA. pp. 684–690

[16] Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M., 2019. EGNet: Edge guidance network for salient object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 8779–8788