Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/57/20241308

Dual attention-enhanced SSD: A novel deep learning model for object detection

Haotian Liu

Department of Aeronautics and Astronautics, Fudan University, Shanghai 200433, China

lee15634356527@gmail.com

Abstract. Object detection is a fundamental task in computer vision with significant implications across various applications, including autonomous driving, surveillance, and image understanding. The accurate and efficient detection of objects within images is crucial for enabling machines to interpret visual information and make informed decisions. In this paper, we present an enhanced version of the Single Shot MultiBox Detector (SSD) for object detection, leveraging the concept of dual attention mechanisms. Our proposed approach, named SSD-Dual Attention, integrates dual attention layers into the SSD framework. These dual attention layers are strategically positioned between feature maps and prediction convolutions, enhancing the model's ability to capture informative feature representations across a wide range of object scales and backgrounds. Experimental results on the PASCAL VOC 2007 and 2012 datasets validate the effectiveness of our approach. Notably, SSD-Dual Attention achieves an impressive mean Average Precision (mAP) of 78.1%, surpassing the performance of SSD models enhanced with attention mechanisms such as SSD-ECA, SSD-CBAM, SSD-Non-local attention and SSD-SE attention, as well as the original SSD. These results underscore the enhanced accuracy and precision of our object detection system, marking a substantial advancement in the field of computer vision. Code is available at https://github.com/AlexHunterLeo/Dual-attention-Enhanced-SSD-A-Novel-Deep-Learning-Model-for-Object-Detection

Keywords: Object Detection, SSD, Dual Attention, Position Attention, Channel Attention.

1. Introduction

Object detection [1], the task of identifying and precisely locating objects within images, stands as a foundational challenge in the realm of computer vision [2]. Thanks to the rapid advancements in deep convolutional neural networks (CNNs) [3], object detection has made significant strides, opening the doors to a multitude of practical applications. The prowess of object detection plays an indispensable role in a variety of automated systems [4], encompassing advanced surveillance systems [5], autonomous vehicles [6], and smart industrial automation [7]. In the realm of automated surveillance [8], object detection takes centre stage, facilitating the tracking of individuals and vehicles through video feeds to discern anomalies or suspicious activities [9]. For the domain of autonomous vehicles [6], the ability to detect objects such as cars, pedestrians, and traffic signs is paramount for safe and efficient navigation. Likewise, in the context of smart factories [10], object detection assumes a pivotal role, empowering automated visual inspection by reliably identifying defects and anomalies in real-time.

DOI: 10.54254/2755-2721/57/20241308

Historically, detection frameworks such as R-CNN [11] relied on regional proposal algorithms, which proved to be computationally inefficient. In contrast, one-stage detectors like SSD [12] have revolutionized the field by circumventing the need for proposal generation, directly predicting object classes and bounding boxes from feature maps. This unified approach has paved the way for real-time object detection [13], a critical requirement for applications like self-driving cars [6]. However, SSD encounters challenges when confronted with small objects and complex environments.

To bolster the performance of SSD, we introduce the integration of dual attention mechanisms [14], which are designed to focus on informative regions and channels while filtering out extraneous data. By selectively accentuating crucial features, attention mechanisms hold the potential to significantly enhance both object localization and classification. Our research endeavours to fortify SSD with a lightweight dual attention mechanism, thereby fostering robust object detection capabilities that transcend varying scales and environmental conditions. This enhanced SSD promises to elevate the accuracy of automated visual inspection in manufacturing settings, enabling the more precise and reliable detection of anomalies.

2. Previous works

Earlier CNN-based object detection frameworks, such as R-CNN [11], relied on external region proposal algorithms to identify potential object regions, which were then subjected to subsequent classification. While these methods delivered high accuracy, their reliance on intricate multi-stage pipelines led to inefficiencies during inference. To address this, approaches like Fast R-CNN and Faster R-CNN [15] were introduced, allowing for shared feature extraction across proposals using a unified network architecture. The introduction of the Region Proposal Network in Faster R-CNN further eliminated the need for external region proposal algorithms. Nevertheless, the two-stage pipeline [16] still imposed limitations on inference speed.

In contrast, one-stage detectors like SSD revolutionized object detection by obviating the need for proposal generation. Instead, they directly predict object classes and bounding boxes from feature maps of various scales using a single feed forward network. This streamlined approach has proven indispensable for real-time detection, particularly in applications like self-driving vehicles [17]. However, SSD has exhibited constraints in terms of generalization, particularly when confronted with small objects or challenging scenarios characterized by occlusion, clutter, and scale variations.

The burgeoning field of attention mechanisms has demonstrated significant potential in enhancing model generalization by focusing on pertinent information. Notably, squeeze-and-excitation networks [18] have showcased the effectiveness of channel attention, while models like CBAM [19] have integrated both channel and spatial attention. Additionally, efficient channel attention [20] also has showcased the effectiveness of channel attention, while models like non-local neural networks [21] have integrated non-local operations. In alignment with these advancements, our research endeavours to elevate the performance of SSD by integrating dual attention mechanisms, which aim to enhance feature representations. By strategically emphasizing crucial regions and channels, attention mechanisms have the potential to confer substantial benefits to object detection across a wide spectrum of scales and environmental conditions.

3. Dataset and Preprocessing

The PASCAL VOC dataset was adopted in our experiment, included VOC2007 and VOC2012. PASCAL VOC dataset was created for the purpose of object detection, which Labels objects in images with bounding boxes and their corresponding category labels. The dataset contains 20 object categories, which include various animals (person, bird, cat, cow, dog, horse, sheep), vehicles (airplane, bicycle, boat, bus, car, motorbike, train), and indoor objects (bottle, chair, dining table, potted plant, sofa, TV/monitor), seeing in Table 1, and there are some image examples with its ground truth labels and boxes in Fig.1. The images in the dataset have various resolutions. Most images in the PASCAL VOC dataset have resolutions of width x height in the range of 500 x 375 pixels or 375 x 500 pixels. Some images might be larger or smaller than this typical size. The reasons why we used PASCAL VOC dataset

DOI: 10.54254/2755-2721/57/20241308

are PASCAL VOC dataset contains both large objects and small objects for evaluating a detection model's ability to generalize across different size objects, PASCAL VOC dataset includes various challenges such as occlusions, truncations, and objects in different poses and also the original SSD paper used PASCAL VOC dataset, which makes us easier to compare our model's performance and mean average precision to the original SSD.

Table 1. Number of objects in the train set and test set.

Train set		Test set	
animals	number of objects	animals	number of objects
person	15576	person	5227
bird	1820	bird	576
cat	1616	cat	370
cow	1058	cow	329
dog	2079	dog	530
horse	1156	horse	395
sheep	1347	sheep	311
vehicles	number of objects	vehicles	number of objects
airplane	1285	airplane	311
bicycle	1208	bicycle	389
boat	1397	boat	393
bus	909	bus	254
car	4008	car	1541
motorbike	1141	motorbike	369
train	984	train	302
indoor objects	number of objects	indoor objects	number of objects
bottle	2116	bottle	657
chair	4338	chair	1374
dining table	1057	dining table	299
potted plant	1724	potted plant	592
sofa	1211	sofa	396
TV/monitor	1193	TV/monitor	361
total images	16551	total images	4952
total objects	49653	total objects	14856

(a) (b)

DOI: 10.54254/2755-2721/57/20241308

Figure 1. Some image examples with its ground truth labels and boxes from PASCAL VOC dataset.

For the data pre-processing, in the first step we applied photometric distortions to the image. The distortions mainly focused on brightness, contrast, saturation, and hue. The factors used to adjust these four values were picked from a uniform distribution. Each distortion had a 50% chance of being applied, and the order they were applied is random. In the second step, we applied expand to our image with a 50% probability. We created an empty image which was filled with the mean of ImageNet data that our base VGG was trained on. The length and width of the empty image were 1 to 4 times the length and width of the original image. The original image was then placed at random coordinates within this larger empty image. In the third step, we applied crop operation to our image. The length and width of the cropped region were 0.3 to 1 times the length and width of the original image. Then computed intersection over union (IoU) of the cropped region and all the bounding boxes of the original image, which were used to decide whether or not this cropped region was picked. After cropping, recalculated the four corners' coordinates of the bounding boxes whose centres located in the cropped region. In the fourth step we apply flipping to our image with a 50% probability. Horizontally flip an image and all the bounding boxes in the image. In the fifth step we resize our image. We resize our image into 300 x 300. Because the bounding boxes coordinates data we read is not fractional coordinates, the bounding boxes coordinates need to be divide by the former size of the image, then multiply 300. For the pre-train, we use VGG-16 weights for our VGG layer [22]. VGG-16 was one of the top-performing architectures on the ImageNet classification challenge. Initializing our model with weights from a pre-trained VGG-16 model leads to better convergence during training and improved detection accuracy.

4. Model

4.1. Dual attention-Enhanced SSD structure

From our Fig. 2, there are four parts in our model VGG convolution part, auxiliary convolution part, prediction convolution part, attention part. For the VGG convolution part, the Conv1 has two convolution layers, one pooling layer. Conv2 has two convolution layers, one pooling layer. Conv3 has three convolution layers, one pooling layer. Conv4 has three convolution layers, one pooling layer. Conv5 has three convolution layers, one pooling layer. Conv6 has one convolution layer. Conv7 has one convolution layer. For the auxiliary convolution part, the Conv8 has two convolution layers. Conv9 has two convolution layers. Conv10 has two convolution layers. Conv11 has two convolution layers. For the feature map, we got them from the last convolution layer of Conv4, the last convolution layer of Conv7, the last convolution layer of Conv8, the last convolution layer of Conv9, the last convolution layer of Conv10, the last convolution layer of Conv11. We sent these feature map to attention part then to the prediction convolution part. For prediction convolution part, in the Fig. 1 we draw one convolution

DOI: 10.54254/2755-2721/57/20241308

block for each feature map. But in fact, we used two convolution blocks for each feature map, one for the prediction of the bounding boxes, one for the prediction scores for each class.

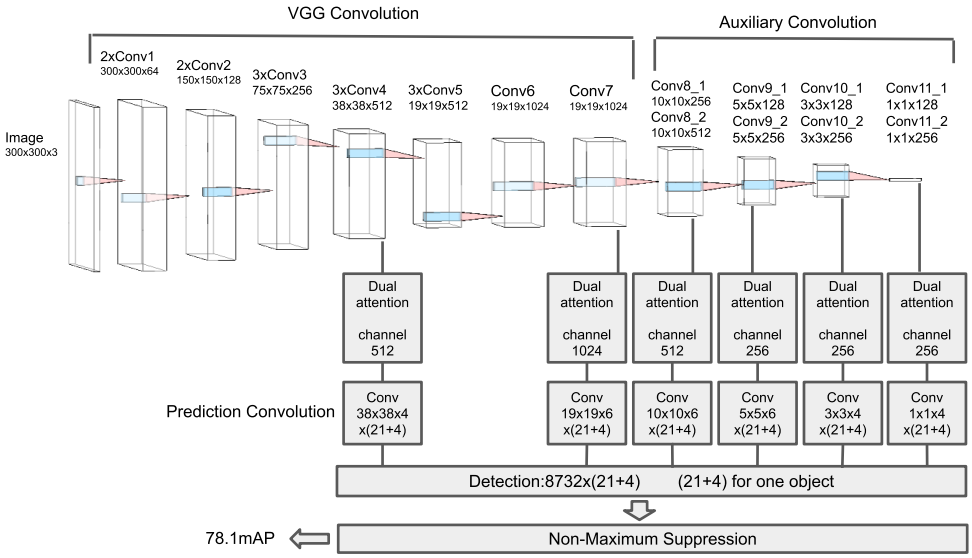

Figure 2. Dual attention-Enhanced SSD.

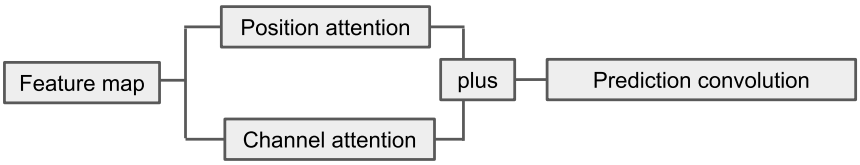

Figure 3. Dual attention.

The attention we used was dual attention. Dual attention combined position attention and channel attention, seeing in Fig. 3. Each feature map would be passed in position attention block and channel attention block, then we added up the result of position attention block and channel attention block. Then we sent the sum of these two to the prediction convolution block. By incorporating dual attention after each feature map but before each prediction convolution block to enhance the SSD, prediction block could focus on relevant regions and channels, the model might reduce the impact of background noise or irrelevant regions in the image, leading to better generalization.

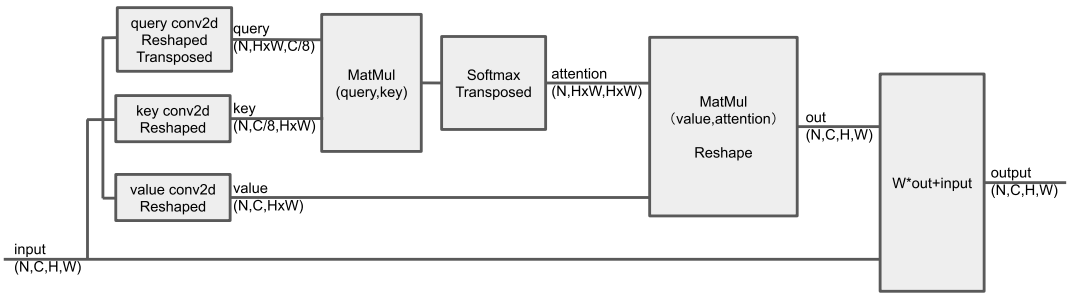

Figure 4. Position attention.

DOI: 10.54254/2755-2721/57/20241308

We used the input from the feature map to create query, key, value matrices by convolution layer, reshaping, transposing. The attention score between different positions in the input was computed by taking the dot product of the query and key matrices. Softmax was applied to the scores along the last dimension to produce the attention weights. We multiplied the value matrix with the attention weights which is transposed to get the attention output, and then reshaped it. We multiplied the attention output with a learnable weight and added the original input to get the final output. Position attention focused on the location of the aspect. It emphasized certain spatial locations in the feature map that are more informative or relevant to the task. The Fig. 4 is the calculation process of position attention.

We used the input from the feature map to create query, key, value matrices by reshaping, transposing without convolution layer. The dot product of the query and key was taken to calculate the energy. Then we took the maximum value along each row, expanded it to the shape of the original energy tensor, subtracted the original energy tensor to get the scores. Softmax was applied to the scores along the last dimension to produce the attention weights. We multiplied the attention weights with the value matrix to get the attention output, and then reshaped it. We multiplied the attention output with a learnable weight and added the original input to get the final output. Channel attention focused on the class of aspect. By adjusting the focus on the channels of the feature maps, the model can emphasize certain channels that are more informative for the detection task. The Fig.5 is the calculation process of channel attention.

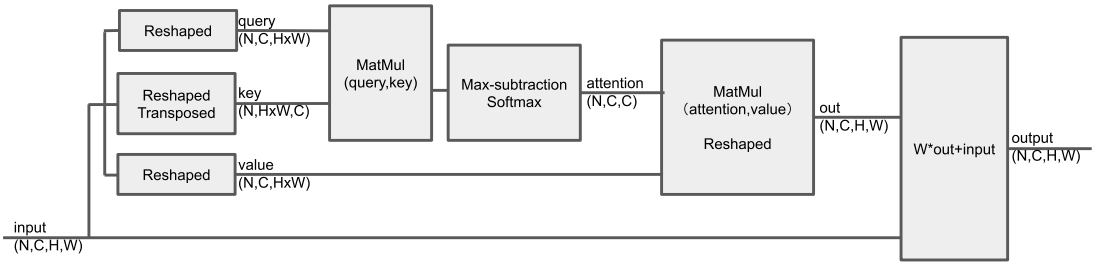

Figure 5. Channel attention.

The following table is the comparison of parameters between our dual attention-enhanced SSD and the original SSD, seeing table 2.

Table 2. Comparison of parameters between SSD and Dual attention-Enhanced SSD

SSD			Dual attention-Enhanced SSD		
Layer	Output Shape	Number of parameters	Layer	Output Shape	Number of parameters
VGG convolution part1	[512,38,38]	7,635,264	VGG convolution part1	[512,38,38]	7,635,264
			Dual attention-4_3	[512,38,38]	328,320
VGG convolution part2	[1024,19,19]	12,848,640	VGG convolution part2	[1024,19,19]	12,848,640
			Dual attention-7	[1024,19,19]	1,312,000
Auxiliary convolution part1	[512,10,10]	1,442,560	Auxiliary convolution part1	[512,10,10]	1,442,560
			Dual attention-8_2	[512,10,10]	328,320
Auxiliary convolution part2	[256,5,5]	360,832	Auxiliary convolution part2	[256,5,5]	360,832
			Dual attention-9_2	[256,5,5]	82,240

DOI: 10.54254/2755-2721/57/20241308

Table 2: (continued).

Auxiliary convolution part3	[256,3,3]	328,064	Auxiliary convolution part3	[256,3,3]	328,064
			Dual attention-10_2	[256,3,3]	82,240
Auxiliary convolution part4	[256,1,1]	328,064	Auxiliary convolution part4	[256,1,1]	328,064
			Dual attention-11_2	[256,1,1]	82,240
Prediction convolution	[8732,4] [8732,21]	3,341,550	Prediction convolution	[8732,4] [8732,21]	3,341,550
Total parameters		26,284,974	Total parameters		28,500,334
Trainable parameters		26,284,974	Trainable parameters		28,500,334

4.2. MultiBox loss, Non-Maximum Suppression and Mean Average Precision

We generated anchor boxes for every feature map, resulting in a total of 8732 anchor boxes generated. Then, we computed the Intersection over Union (IoU) between the ground truth bounding boxes of an image and the 8732 anchor boxes, resulting in an IoU matrix with dimensions (number of objects in the image, 8732). For each anchor box, we determined the most likely corresponding ground truth bounding box and indexed it in a tensor. Similarly, for each ground truth bounding box, we identified the most suitable anchor box. These mappings ensured that each ground truth object was associated with an anchor box. Anchor boxes with an IoU below a specified threshold were labeled as 0. Using these labels, we derived encoded ground truth boxes from the anchor boxes. A boolean tensor was used to identify non-zero labels, allowing us to select the relevant ground truth and prediction boxes. We computed the location loss using the L1Loss and the confidence loss using the CrossEntropyLoss. The final loss value combined both of these losses, taking into account positive and hard negative confidence losses.

The predicted boxes were decoded using anchor boxes to obtain decoded predicted boxes for object detection. For each class, 8732 boxes were extracted and filtered based on a threshold score. These filtered boxes were subsequently ranked by their scores, and an Intersection over Union (IoU) matrix was computed among them. Non-maximum suppression was applied, starting from the box with the highest score, and any boxes with a high IoU value relative to it were suppressed. This process continued, ensuring that already suppressed boxes remained suppressed. Ultimately, only a few unsuppressed boxes remained for each class. These unsuppressed boxes for all classes constituted the final predicted ground truth boxes.

For every class, we calculated the Average Precision (AP). AP represents precision averaged over a set of recall values. Precision and recall are two primary metrics in object detection. Precision measures the accuracy of the detections, while recall measures how many of the true objects were detected. We calculated the number of true positives and false positives for each detection for the given class. Using these true positives and false positives, we computed the cumulative precision and recall at each detection. Precision values were extracted for various recall thresholds, and their mean provided the AP for that class. After obtaining the AP for each class, the mean of these values yielded the Mean Average Precision (mAP).

5. Results

5.1. Experiment Result including mAP and Detection Visualization

In our experiments, we utilized the VGG16 architecture, which had been pre-trained on the ImageNet dataset. We made specific architectural modifications by converting the fully connected layers fc6 and fc7 into convolutional layers. Additionally, we removed all dropout layers and the fc8 layer. During the training process, we employed a batch size of 8, initialized the learning rate at 10^-3, decayed it to 10^-

DOI: 10.54254/2755-2721/57/20241308

4 at iteration 80000, and further decreased it to 10^-5 at iteration 100000. We conducted a total of 120000 iterations.

On the PASCAL VOC dataset, which encompasses both VOC2007 and VOC2012, we conducted a comparative evaluation against the original SSD, SSD-ECA (Efficient Channel Attention), SSD-CBAM (Convolutional Block Attention Module), SSD-Non-local attention and SSD-SE (Squeeze-and-Excitation) attention models. All methods underwent fine-tuning using the same pre-trained VGG16 network. In the context of training on the PASCAL VOC dataset, Table 3 demonstrates that our SSD-Dual attention model outperforms the original SSD, SSD-ECA, SSD-CBAM, SSD-Non-local attention and SSD-SE attention models in terms of accuracy. The mean Average Precision (mAP) of the original SSD was reported in the original SSD paper. Our SSD-Dual attention model achieves a 3.8% higher mAP compared to the original SSD.

Table 3. PASCAL VOC dataset test detection results

Model	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
SSD	74.3	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3
SSD-ECA	76.9	77.9	84.2	77.5	70.4	46.0	86.8	85.8	88.0	59.3	81.1
SSD-CBAM	77.3	79.3	**85.4**	76.9	**71.8**	46.7	86.8	**86.4**	88.0	58.9	82.0
SSD-Non-local attention	77.4	**82.4**	84.3	77.4	65.2	48.8	**87.7**	86.1	88.5	58.9	83.5
SSD-SE attention	77.5	81.4	85.2	77.8	71. 7	47.3	87.4	86.4	87.7	57.4	79.7
SSD-Dual attention	**78.1**	81.1	84.5	**78.2**	71.5	**49.2**	86.0	85.9	**88.5**	**60.7**	**86.7**
Model	mAP	table	dog	horse	mbike	person	plant	sheep	sofa	train	TV
SSD	74.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD-ECA	76.9	75.6	85.3	87.8	83.9	78.8	52.7	77.5	77.9	**87.2**	75.2
SSD-CBAM	77.3	75.1	85.3	87.3	**85.0**	78.4	52.8	76.5	**80.7**	86.8	75.1
SSD-Non-local attention	77.4	**76.6**	85.1	88.0	84.7	78.8	51.7	**78.8**	79.9	85.9	**76.6**
SSD-SE attention	77.5	75.7	**85.4**	**88.4**	84.0	79.0	54.1	78.6	78.5	86.6	77.1
SSD-Dual attention	**78.1**	75.7	85.1	88.3	84.4	**79.8**	**54.4**	78.5	80.2	87.0	76.0

The following pictures are the detection results of original SSD and SSD-Dual attention.

(a) (b)

DOI: 10.54254/2755-2721/57/20241308

Figure 6. Comparison of detection results between original SSD and SSD-Dual attention. (a)(c) the result of original SSD. (b)(d) the result of SSD-Dual attention.

Figure 6 provides clear visual evidence of the enhanced performance of SSD-Dual attention compared to the original SSD. Particularly, in Fig.6 (b), SSD-Dual attention successfully detects the chair within the image, whereas the original SSD fails to detect any object. This observation strongly suggests that our model has significantly improved detection capabilities in comparison to the original SSD. Subsequently, we computed the confusion matrix for both the original SSD and SSD-Dual attention. We established a threshold of 0.5 for the Intersection over Union (IoU) between the final predicted boxes and the ground truth boxes, and the results are presented in Figure 7.

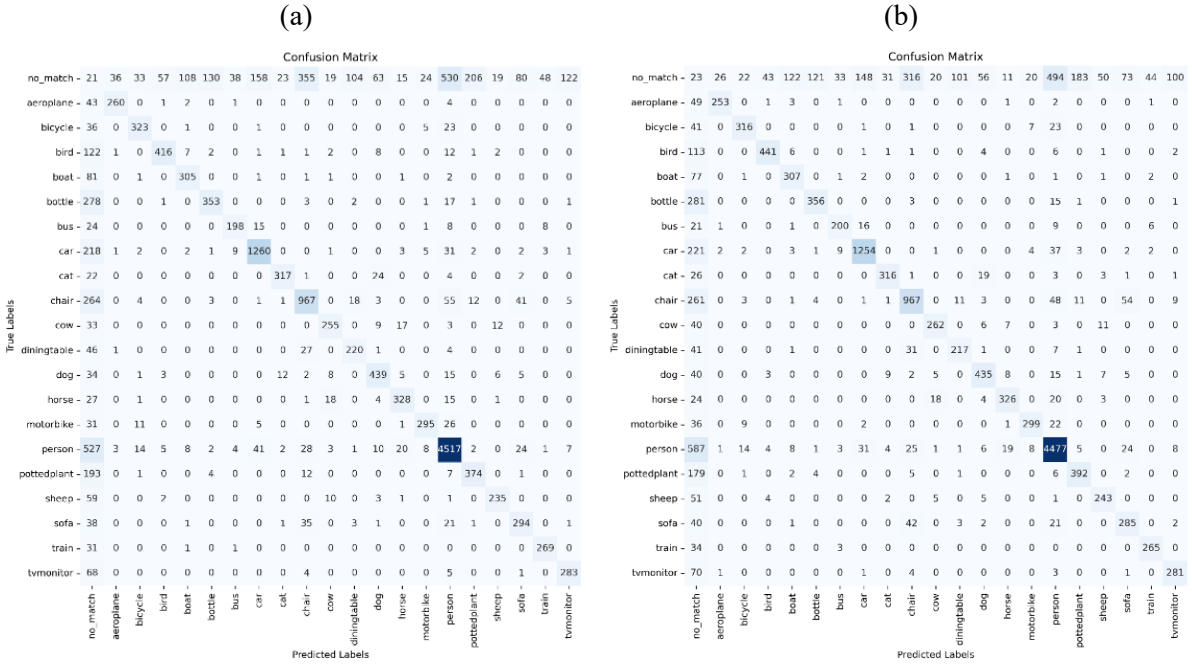

Figure 7. Confusion matrix when IoU threshold is 0.5. (a) confusion matrix of original SSD when IoU threshold is 0.5. (b) confusion matrix of SSD-Dual attention when IoU threshold is 0.5.

Figure 7 provides valuable insights into the performance of SSD-Dual attention compared to the original SSD. Notably, SSD-Dual attention predicts a significantly higher number of matching boxes

DOI: 10.54254/2755-2721/57/20241308

for the classes bird, chair, cow, and plant. Additionally, it predicts slightly more matching boxes for classes such as boat, bottle, bus, motorbike, and sheep. This analysis indicates that the primary challenge in object detection, both for the original SSD and SSD-Dual attention, lies in accurately predicting the positions of bounding boxes. Once this aspect is successfully addressed, the task of class classification becomes relatively easier. To assess the accuracy of the final predicted boxes, we further increased the threshold for IoU between the final predicted boxes and the ground truth boxes to 0.8. The resulting outcomes are presented in Figure 8.

(a) (b)

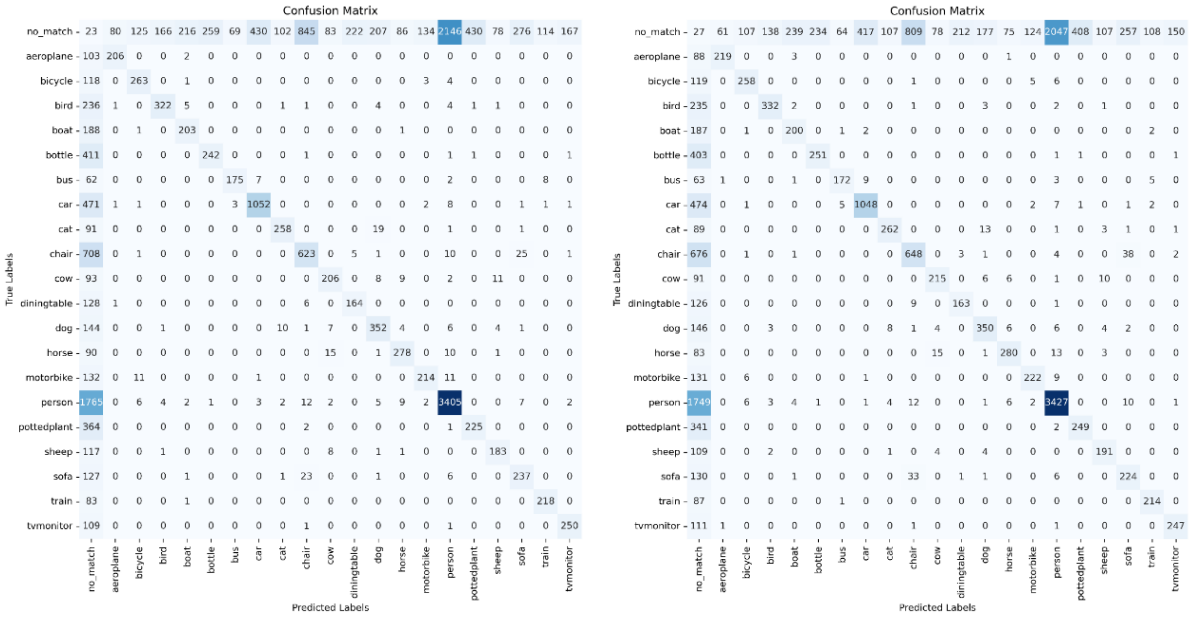

Figure 8. Confusion matrix when threshold IoU is 0.8. (a) confusion matrix of original SSD when IoU threshold is 0.8. (b) confusion matrix of SSD-Dual attention when IoU threshold is 0.8.

From Fig.8, we can observe that SSD-Dual attention predicts more matching boxes for most of the classes at a higher IoU threshold. This indicates that the bounding boxes predicted by SSD-Dual attention are more accurate. From Fig.7 and Fig.8, we can conclude that whether or not the model could detect the objects in the image has not improved significantly from original SSD to SSD-Dual attention, but for the accuracy of the detections of the model has improved.

5.2. Improvement Effect of Dual Attention on SSD Structure Performance

The position attention in the dual attention mechanism typically generates a map of the same height and width as the feature map, with each value indicating the importance of the corresponding spatial location. This enables the model to concentrate on specific spatial regions within a feature map. On the other hand, the channel attention component assigns weights to each channel in a feature map, enabling some channels to be amplified while others are suppressed. Channel attention allows the model to prioritize certain channels over others.

One of the key challenges in SSD is effectively handling small objects and objects with varying aspect ratios. By introducing attention mechanisms between the feature maps and the prediction convolution layers, the model can dynamically select the most crucial features for predicting bounding boxes and object classes. This can be especially advantageous when detecting small objects or objects with non-standard aspect ratios. Regions in the image that are irrelevant or unimportant feature channels can introduce noise into the model's predictions. Attention mechanisms help mitigate this issue by suppressing these irrelevant regions or channels, enabling the model to make predictions based on more pertinent information. By focusing on specific regions or channels, attention can also lead to a reduction

DOI: 10.54254/2755-2721/57/20241308

in the computational resources required for prediction convolution layers to make accurate predictions of bounding boxes and classes. This resource efficiency is particularly valuable in resource-constrained environments or real-time applications.

5.3. Performance Comparison between SSD-ECA, SSD-CBAM, SSD-Non-local attention, SSD-SE attention and SSD-dual attention

SSD-ECA deploys efficient channel attention that captures channel-wise interdependencies without the need for the expensive squeeze operation. Specifically, the feature maps undergo an adaptive average pooling layer, a 1D convolutional layer, a sigmoid activation function to produce channel-wise attention weights. Multiply the weights with original feature maps. ECA focuses only on channel attention and is designed to be lightweight, it might not compete with the richer feature capturing capability of dual attention, which leads to enhanced performance.

SSD-CBAM incorporates both spatial and channel attention but uses a slightly different architecture than dual attention. CBAM introduces attention maps sequentially along two separate dimensions rather than in parallel or through shared operations. While this sequential approach can be beneficial, it may lead to a slight performance decrease if the channel attention suppresses critical channels required by subsequent position attention. Additionally, the channel attention component of CBAM may not be as effective.

SSD-Non-local attention uses non-local operations which is to compute the response at a position as a weighted sum of the features at all positions in the input feature map. Specifically, use the pairwise function which is implemented as a dot product of transformed feature maps to determine the weights. Multiply the weights which are normalized by softmax function with transformed feature map. Reshape, project the result and add to the original feature maps. Non-local attention can capture long-range dependencies in the data, but it may not be as effective as dual attention in capturing relevant features.

SSD-SE attention employs channel attention to adjust the importance of different channels within an image. It calculates channel weights based on global information and utilizes these weights to assign importance to each channel. This mechanism introduces only channel attention, resulting in a relatively low increase in computational cost and model parameters when compared to position attention mechanisms. By focusing on channel weight adjustment, SSD-SE attention can be particularly effective in cases where certain channels play a crucial role, although it may not capture spatial dependencies as effectively.

In contrast, SSD-Dual Attention combines position and channel attention mechanisms in a straightforward manner, applying them in parallel or simultaneously to feature maps. Position and channel attentions are applied independently and then combined, leveraging both spatial and channel-wise information equally to achieve superior performance. Among the three attention mechanisms, SSD-Dual Attention performs the best in terms of mean Average Precision (mAP). The computational overhead and added complexity associated with SSD-Dual Attention are deemed acceptable for our specific application.

6. Discussion

From Table 3, it is evident that all attention mechanisms have exhibited improvements in Mean Average Precision (mAP) when compared to the baseline SSD model. Among them, SSD-Dual attention stands out with the highest overall mAP performance. Several object classes have shown significant enhancements with the integration of attention mechanisms. For instance, in the "Bird" category, all attention mechanisms surpassed the baseline SSD, with SSD-Dual attention leading at 78.2 mAP. The "Cow" category also saw improvements across the board, with all attention mechanisms surpassing the baseline SSD, with SSD-Dual attention leading at 86.7 mAP. On the other hand, there were object categories that exhibited marginal or no improvements with the incorporation of attention mechanisms. Notably, the "Bottle" category showed minimal improvement across all models, suggesting that attention mechanisms may not be as effective for certain object categories. Likewise, in the "Dog" category, improvements were minimal across all attention mechanisms compared to the baseline.

DOI: 10.54254/2755-2721/57/20241308

Throughout the evaluation, SSD-Dual attention consistently ranked at the top or near the top across all object categories. By incorporating dual attention into the SSD architecture, this study offers a means to enhance object detection performance in intricate settings. Scenarios characterized by overlapping objects, varying scales, or occlusions stand to benefit from the integration of dual attention. This model concurrently directs its focus towards pivotal regions while accentuating pertinent features, ultimately resulting in improved object detection in complex environments. Remarkably, our approach implements dual attention in a lightweight manner, augmenting the SSD structure without imposing substantial computational overhead. Consequently, an SSD equipped with dual attention stands to achieve superior performance, which holds particular relevance for engineering applications demanding lightweight models.

The combination of position and channel attention proved to be a highly effective and adaptive approach. Positional attention proves invaluable in assessing the significance of distinct spatial regions within each feature map. This proves especially advantageous given that SSD extracts features from multiple layers, each representing different scales. The application of positional attention can potentially amplify regions within each feature map that hold greater relevance for the detection of objects of corresponding scales. Furthermore, channel attention aids in evaluating the importance of various feature channels. In the context of SSD, where features from diverse layers are utilized, channel attention helps prioritize feature channels with greater relevance to the final prediction. In essence, this serves as an adaptive fusion mechanism, accentuating the most informative channels from different scales prior to the prediction phase.

Introducing dual attention mechanisms does indeed increase computational complexity, potentially leading to longer training times. While dual attention has demonstrated significant improvements across numerous categories, there are instances, such as in the case of "dog," where the improvement is marginal. It is conceivable that challenges in detection for certain classes may arise from factors beyond feature distinction, which attention mechanisms may not fully address. Moreover, the inclusion of additional parameters through attention mechanisms carries a potential risk of overfitting, particularly when training on smaller datasets.

In our experiments, we employ VGG16 as our foundational convolutional structure due to its simplicity and strong performance. Nevertheless, as deep learning research has advanced, newer architectures have emerged that offer superior accuracy and efficiency compared to VGG16. One such architecture is ResNet[23], which introduces the innovative concept of residual blocks. Each residual block incorporates skip connections that bypass one or more layers, effectively mitigating the vanishing gradient problem. This characteristic allows ResNet to achieve much greater depth than VGG. The increased depth of the ResNet architecture enables it to capture more abstract and intricate patterns within images, which can be highly advantageous for object detection.

To further enhance object detection performance, additional deconvolutional layers can be introduced after the auxiliary convolutional part of SSD. These layers facilitate the upsampling of feature maps, effectively incorporating context into higher-resolution layers. This approach significantly improves detection accuracy for smaller objects, which are typically challenging for SSD. By introducing these extra layers, which generate more feature maps, we can experiment with various anchor box scales and aspect ratios tailored to specific datasets or object types, thereby achieving further performance enhancements.

The attention weights of dual attention mechanisms are computed using a fixed formula based on the dot product. While there are learnable parameters involved, the overall structure of the attention computation remains the same. We can try to use adaptive attention mechanism to have an attention mechanism that can alter its behaviour more drastically based on the content of the input. Using a small neural network to process the feature map for producing attention weights is one of the solutions. The network can learn more complex ways to generate attention weights based on the specifics of the input feature map. We can also apply hierarchical attention to have multiple layers or levels of attention mechanisms, where coarser levels guide the attention of finer levels. At the coarser levels, the attention mechanism provides a broad overview to help in focusing on larger, more general regions or aspects of

DOI: 10.54254/2755-2721/57/20241308

the input data. Guided by the coarse-level attention, the finer levels then focus on details on specific regions or features indicated by the higher-level context.

We could conduct attention visualization for our SSD-Dual attention model in the future. Our model incorporates a lot of convolutional layers, we can use attention visualization to understand how models focus on specific parts of the input data, such as gradient-weighted class activation mapping[24] (Grad-CAM) which used to visualize where convolutional neural networks look in an image to make decisions. Grad-CAM doesn't directly visualize attention weights but rather shows which regions in the image were most influential in producing a particular class output. For our dual attention component, visualizing attention is more straightforward. For position attention, it produces a spatial attention map, we can directly visualize this map as a heatmap. For channel attention, we can project the channel attention weights onto the spatial dimensions using the feature maps, essentially showing which features are highlighted by the channel attention.

7. Conclusion

In conclusion, our study focused on the integration of different attention mechanisms into the Single Shot MultiBox Detector (SSD) framework for object detection. The performance of these variations was evaluated based on the mean Average Precision (mAP) metric, with the following overall rankings, from highest to lowest: SSD-Dual Attention, SSD-SE Attention, SSD-Non-local attention, SSD-CBAM, SSD-ECA and the original SSD. This ranking underscores the positive impact of incorporating attention mechanisms into the SSD architecture, resulting in improved object detection capabilities.

Notably, SSD-Dual Attention emerged as the top-performing model, demonstrating its effectiveness across various object categories. Particularly noteworthy improvements were observed in categories such as birds, chairs, cows, and plants. This consistent enhancement across diverse object classes underscores the generalizability of the dual attention mechanism, allowing the model to adaptively capture both local and global context. This adaptability proves advantageous for detecting objects of varying sizes and characteristics.

An important contribution of our study is the empirical evidence that integrating dual attention mechanisms with SSD leads to enhanced detection performance across multiple categories. This highlights the benefits of combining both position and channel attention in object detection tasks, as the dual attention mechanism strikes a balance between these two types of attention. Furthermore, SSD-Dual Attention offers several innovations and improvements. Firstly, it enhances feature representations by enabling the model to focus on crucial spatial details and relevant feature channels. This refinement leads to improved object detection accuracy. Secondly, the dual attention mechanism enhances generalization, reduces noise in feature maps, and enhances robustness, making it suitable for object detection in challenging scenarios, including cluttered or low-quality images.

The applications of SSD-Dual Attention are diverse. It can be adapted to new tasks with other datasets, such as detecting anomalies or regions of interest in medical images or objects in the field of autonomous driving. Transfer learning from larger datasets like PASCAL VOC can also be employed to fine-tune the model on smaller datasets, overcoming challenges related to limited annotated examples. Moreover, the dual attention mechanism has practical applications in visualization, aiding in the interpretation of the model's decisions. This interpretability is crucial in domains where understanding the model's reasoning is essential.

Beyond object detection, dual attention mechanisms can be applied to various computer vision tasks, including image classification, image segmentation, image super-resolution, and image captioning. In each of these tasks, the dual attention mechanism enhances the model's ability to capture informative features, making it a valuable tool in the broader field of computer vision.

References

[1] 2011 Object Detection and Recognition (SpringerReference)
[2] 1983 Computer vision (Computer Vision, Graphics, and Image Processing) vol 22, pp 410–411

DOI: 10.54254/2755-2721/57/20241308

[3] Krizhevsky A, Sutskever I and Hinton G E 2017 ImageNet classification with deep convolutional neural networks (Commun. ACM) vol 60, pp 84–90
[4] Heimberger M, Horgan J, Hughes C, McDonald J and Yogamani S 2017 Computer vision in automated parking systems: Design, implementation and challenges (Image and Vision Computing) vol 68, pp 88–101
[5] Sagrebin M and Pauli J 2009 Real-Time Moving Object Detection for Video Surveillance (2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance) pp 31–36
[6] Massoud Y and Laganiere R 2022 Learnable Fusion Mechanisms for Object Detection in Autonomous Vehicles
[7] 2020 TT06 Computer Vision and Human-Machine Interaction in Industrial and Factory Automation (2020 25th IEEE International Conference on Emerging Technologies and Factory Automation) pp 502–504
[8] Saikrishnan V and Karthikeyan M 2023 Automated Object Detection and Classification using Metaheuristics with Deep Learning on Surveillance Videos (2023 International Conference on Sustainable Computing and Data Communication Systems) pp 29–34
[9] Kogut G T and Trivedi M M 2001 Maintaining the identity of multiple vehicles as they travel through a video network (Proceedings 2001 IEEE Workshop on Multi-Object Tracking) pp 29–34
[10] Ku B, Kim K and Jeong J 2022 Real-Time ISR-YOLOv4 Based Small Object Detection for Safe Shop Floor in Smart Factories (Electronics) vol 11, p 2348
[11] Girshick R 2015 Fast R-CNN (2015 IEEE International Conference on Computer Vision) pp 1440–1448
[12] Liu W, et al 2016 SSD: Single Shot MultiBox Detector. in Computer Vision (ECCV, 2016) ed Leibe B, Matas J, Sebe N and Welling M (Springer International Publishing, 2016) vol 9905, pp 21–37
[13] Redmon J, Divvala S, Girshick R and Farhadi A 2015 You Only Look Once: Unified, Real-Time Object Detection
[14] Fu J, et al 2018 Dual Attention Network for Scene Segmentation
[15] Ren S, He K, Girshick R and Sun J 2015 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
[16] Du L, Zhang R and Wang X 2020 Overview of two-stage object detection algorithms (J. Phys.: Conf. Ser.) 1544, 012033
[17] Ouyang Z, Niu J, Liu Y and Guizani M 2020 Deep CNN-Based Real-Time Traffic Light Detector for Self-Driving Vehicles (IEEE Trans. on Mobile Comput) vol 19, pp 300–313
[18] Hu J, Shen L, Albanie S, Sun G and Wu E 2017 Squeeze-and-Excitation Networks
[19] Woo S, Park J, Lee J-Y and Kweon I S 2018 CBAM: Convolutional Block Attention Module
[20] Wang Q, et al 2019 ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks
[21] Wang X, Girshick R, Gupta A and He K 2017 Non-local Neural Networks
[22] Simonyan K and Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition
[23] He K, Zhang X, Ren S and Sun J 2015 Deep Residual Learning for Image Recognition
[24] Selvaraju R R, et al 2016 Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization