Improving Small Objects Detection using Transformer

Shikha Dubey¹, Farrukh Olimov²,
Muhammad Aasim Rafique¹,
and Moongu Jeon ^{†1},

Abstract—General artificial intelligence is a trade-off between the inductive bias of an algorithm and its out-of-distribution generalization performance. The conspicuous impact of inductive bias is an unceasing trend of improved predictions in various problems in computer vision like object detection. Although a recently introduced object detection technique, based on transformers (DETR), shows results competitive to the conventional and modern object detection models, its accuracy deteriorates for detecting small-sized objects (in perspective). This study examines the inductive bias of DETR and proposes a normalized inductive bias for object detection using a transformer (SOF-DETR). It uses a lazy-fusion of features to sustain deep contextual information of objects present in the image. The features from multiple subsequent deep layers are fused with element-wise-summation and input to a transformer network for object queries that learn the long and short-distance spatial association in the image by the attention mechanism. SOF-DETR uses a global set-based prediction for object detection, which directly produces a set of bounding boxes. The experimental results on the MS COCO dataset show the effectiveness of the added normalized inductive bias and feature fusion techniques by detecting more small-sized objects than DETR.

Index Terms—Normalized Inductive Bias, Features Fusion, Small-Objects Detection, Transformer, Self-Attention.

I. INTRODUCTION

Object detection is one of the long-standing research topics in computer vision (CV), and more challenges surface with technology updates and artificial intelligence (AI) ventures. New applications demand rigorous, accurate, and efficient detection of all interesting objects in an image. AI enthralled by the advancement in deep learning (DL) generally benefits from the hard inductive bias of convolutional neural networks (CNN) with region proposal network (RPN) for two-stages object detection [1]–[4]. Later, single-shot object detection with anchors efficiently produced competitive results [5]–[7]. Despite outstanding results, RPN generates highly overlapping region proposals for the same set of objects demanding post-processing using handcrafted processing like Non-Maximum Suppression (NMS) based on Intersection-over-Union (IoU) [4]–[6]. Recently introduced paradigm, transformer architecture has shown outstanding performance in the natural language processing field [8] and led the researchers to utilize such paradigms in several fields like image captioning [9], speech processing [10], and even biomedical imaging [11].

Moreover, it inspired researchers from the computer vision field as well [12]–[14], and recently, a new architecture, detection using the transformer (DETR), was proposed [13]. This architecture efficiently outputs bounding boxes per image and classes in parallel using a self-attention mechanism. DETR eliminates the necessity of post-processing detection by using bipartite matching, which helps remove redundancy in proposals. DETR shows competitive results among conventional and modern object detection models. Nevertheless, detecting smaller-sized objects compared with medium and larger-sized objects is still challenging due to some factors as inadequate resolutions of objects and fewer training images containing small-sized objects than other-sized objects present in the training dataset. That reflects in the overall performance degradation of these detectors. This present work attempts to address such problems in the object detection task using transformers and exploiting inductive biases.

Generally, artificial neural networks are famous for their inductive bias exposit through their composition, either in convolutional [12], [15] or sequential networks [16]. Convolution neural networks explore hard inductive bias using diverse compositions of networks layers and the hyper-parameters for connected layers and provide performance improvements. Therefore, in this work, we propose a model, SOF-DETR, which leverages such inductive biases in the case of transformer detection to improve small-sized object detection. To provide the extensive features of small-sized objects without losing other objects' information, a normalized inductive bias is included in our proposed model, SOF-DETR. Normalized inductive bias includes a fusion technique. Fusion techniques are commonly used to combine multi-sensor or multi-source data [17]. In our study, we use it to join different feature layers. Multiple layers of the normalized convolutional layers are fused together to give some hard-encoded inductive bias to favor the small-sized objects. Then, these advanced deep features are progressed through the transformer network as object queries and learn associations among present objects in the image by the attention mechanism of a transformer. Furthermore, similar to studies, SOF-DETR predicts unique bounding boxes associated with a single object by utilizing a global set-based cost function for the bi-partite matching technique.

Contributions: This study contributes to the object detection field as follows. First, we introduce a normalized inductive bias for detection using a transformer to get extensive features from small-sized objects. Second, this normalized

¹ School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, 61005, South Korea;

² Threat Intelligence Team, Monitorapp, Seoul, South Korea;

[†] represents the corresponding author

Fig. 1. SOF-DETR Architecture.

inductive bias utilizing a fusion of multiple layers also supports generating more focused self-attention maps. Our extensive experiments demonstrate the effectiveness of all the modules. The detailed explanation, along with experimentation of SOF-DETR, is exhibited in the following sections.

II. METHOD

The architecture of the proposed model, SOF-DETR, is depicted in Fig. 1. SOF-DETR introduces a new composition of the neural networks in the transformer-based detection model (DETR). It channelizes the hard inductive bias of a CNN to the soft inductive bias of a transformer by interleaving a normalized inductive bias on fused feature map from different convolutional blocks. The model consists of three modules: convolutional backbone, normalized inductive bias, and transformer detection. For unique detection, this study follows set-based object detection [18] to associate a unique bounding box with one object. The composition of each module is given in the following sections¹.

A. Convolutional Backbone

In this study, we have utilized the standard convolutional model, ResNet [19], as the backbone of SOF-DETR for extracting deep structural and contextual features of all objects in the image. First, we provide the given input image I of $3 * H * W$ dimensions, where H, W are the height and width of the image, respectively. We have experimented our model on the ResNet-50.

B. Normalized Inductive Bias

DETR practices the CNN hard inductive bias for features and soft inductive bias to learn the relationship. There are still challenges in the case of small-sized object detection due to

their inadequate resolutions resulting in missing indispensable features of these objects for further processing, which leads to poor attention maps for such objects. Consequently, to overcome this challenge in the case of transformer-based detection, our proposed model, SOF-DETR, channelize hard inductive bias and normalized it before passing it to the soft inductive bias of the transformer encoder. The normalized hard inductive bias fuses the features extracted at different convolution blocks so that the essential features of the small-sized are intact for further processing. It accepts the output of the last two layers (layers 3 and 4) of the backbone model and down-project the features from high-level (layer 4) using a de-convolutional layer of 1024-size and transfers it through a convolutional layer of 1024-size (concatenating more layers decreases the overall performance). These features are fused after group normalization using element-wise-summation as given in equation (1).

$$F_c = \text{Conv}(\text{Deconv}(F(I, L))) \oplus \text{Conv}(F(I, L - 1)) \quad (1)$$

where F stands for extracted features from an image, I and layer L or $L - 1$, and F_c stands for combined features. These combined features are then passed through the relu layer, followed by a convolutional layer of size 1024 for extracting deep features from these combined high-level and low-level features. Concatenating these features has not improved the current model; therefore, we have experimented with the element-wise summation of the low-level and high-level features. Now, we transfer these contextual features, including inductive bias, to the transformer module for further processing.

C. Transformer Detection

The transformer architecture [13] for object detection consists of a simple encoder-decoder architecture and a feed-forward network (FFN) for ultimate predictions. **The encoder layer** consists of standard multi-head self-attention and FFN modules [8] for correlating the queries with keys and values.

¹The code will be publicly available on <https://github.com/shikha-gist/SOF-DETR/>

Positional encoding is added for image data [20]. These encoded features of N objects of n -dimension are passed through the decoder layers. N is greater than the maximum number of objects present in the image. **The decoder layer** also follows the standard transformer-based architecture. The embedded N features generate attention maps using a multi-head self-encoder-decoder attention mechanism. SOF-DETR follows the DETR decoder layer where all the embeddings of N objects are passed together as queries and generate the output for each object simultaneously as decoder embedded form. At last, these embedded decoder outputs are individually decoded into bounding box coordinates, and class labels by passing it through FFN consists of three-layer fully-connected layers with ReLU activation function and provide us N number of object predictions with normalized center coordinates, height, and width. The linear layer provides the class of the object.

Following the studies [13], [18], we have utilized a set-based loss function for training the network and giving a unique bounding box for each object. A set of fixed-size of N embedded output from the decoder is generated, where N is significantly larger than maximum number of objects present in the image. If the ground truth for image I has o number of objects in set, we pad ground truth with ϕ for no object case to make it equal to N number of detected objects $\hat{d} = \{\hat{d}_i\}_{i=1}^N$. Now, we find the best permutation, $\hat{\vartheta} : \vartheta \in \alpha_N$, of N elements present in these two sets, o , and $\hat{d}$, which gives the lowest cost for finding bipartite matching as below:

$$\hat{\vartheta} = \arg \min_{\vartheta \in \alpha_N} \sum_i^N C_{bm}(o_i, \hat{d}_{\vartheta(i)}) \quad (2)$$

where C_{bm} is element-wise bipartite matching cost, which can be calculated using the Hungarian algorithm similar to [18], among elements of sets of ground-truth o and predictions $\hat{d}$ with index $\vartheta(i)$. Set of ground-truth o has elements as $o_i = (c_i, bbox_i)$ and c_i , $bbox_i$ are class and bounding box of the i^{th} object. Similarly, prediction for i^{th} object $\hat{p}_i = (\hat{c}_i, \hat{bbox}_i)$ with predicted class $\hat{c}_i$ of index $\vartheta(i)$ with probability $\hat{p}_{\vartheta(i)}(c_i)$ and bounding box $\hat{bbox}_i$. Similar to [13], [18], this bipartite matching takes place for both predictions class predictions and bounding box predictions. The loss function, Hungarian loss for all matched pairs, for predicted classes, and bounding boxes can be calculated similarly to work [13], [18]. Class imbalances are handled likewise to work [4]. The bounding box loss, C_{bbox} , can be calculated utilizing l_1 loss and generalized IoU loss, L_{IoU} similar to the work [21]. The loss function and shared layer-norm are added to each decoder layer.

Our model, SOF-DETR, is trained using this loss function and predicts a unique bounding box and class for each object present in the image.

III. EXPERIMENTATION AND DISCUSSION

A. Dataset

We have experimented with and evaluated SOF-DETR on the MS COCO object detection dataset, which categorizes the small-sized objects (2017). In this study, the denotation of objects' size is appraised in accordance with the MS COCO

Fig. 2. Qualitative analysis of SOF-DETR, where first-row and second-row illustrate detections from DETR and SOF-DETR, respectively. Additional detections through SOF-DETR are represented with a pink color class box. Detections only with more than 90% confidence value are displayed.

annotations: A labeled object with a bounding box having an area less than 32^2 (pixels) is a small-sized object (41%), the area between 32^2 and 96^2 is a medium-sized object (34%), and the objects with an area greater than 96^2 are large-sized (24%). There are at least 7 objects and at most 63 objects present (labeled) in a single image of the training set. In the case of occluded objects, the area is computed based on the size of the partially labeled bounded box. We have used data-augmentation techniques like scaling, resizing all the images by $800 * 800$ random cropping with a probability of 0.5, followed by resizing.

B. Evaluation Metrics

We have evaluated the performance of SOF-DETR using the standard evaluation metric: mAP (mean average precision) of bounding boxes averaged on thresholds $\in [0.5 : 0.05 : 0.95]$ for all detection. Moreover, we have also compared the proposed model's performance using conventional precision, recall, and F-1 scores @0.5 for small-sized, medium-sized, and large-sized objects. Furthermore, the PR curve (Precision-Recall curve) is also utilized for the performance analysis.

C. Implementation Details

The training of SOF-DETR is similar to DETR: It uses an AdamW optimizer with an initial learning rate of $10^{(-4)}$. The learning rate for the backbone (ResNet-50) is $10^{(-5)}$, and weight decay is $10^{(-4)}$. SOF-DETR uses 0.1 dropout and Xavier initialization [22] for starting weights. We have experimented SOF-DETR without any additional dilation layer [23], and future work can extend this work. We have reported the results after 500 epochs of training for an overall evaluation, with a learning rate dropping by 10 after 300 epochs. SOF-DETR is trained on 8-V100 Nvidia-GPUs on a single node with a batch size of 3 on each GPU (performs better than 2, 4 and 6 batch-sizes) and tested on a single GPU and single node with 2 batch sizes.

D. Performance Analysis of SOF-DETR

1) *Quantitative Analysis*: We have compared the SOF-DETR with DETR [13] and Faster-RCNN [4] algorithms. Table I shows the detailed qualitative results. We have also

Fig. 3. Effectiveness of normalized inductive bias on objects' self-attention maps of the small, medium, and large-sized objects.

Fig. 4. PR-Curves comparison on top-3 object classes for small-sized objects. The average precision of each object class is also given at the top-right corner.

compared the Faster-RCNN results reported in the paper DTER after fine-tuning [13] as well as pre-trained faster-RCNN without any fine-tuning [4]. SOF-DETR outperforms DETR with a 0.7% improvement in mAP score overall, and 1.2%, 0.1% and 0.4% improvement in detecting small-sized, medium-sized, and large-sized objects, respectively. The results indicate normalized inductive bias in SOF-DETR improves the detection of small-sized objects without affecting the medium-sized and large-sized objects detection. To emphasis the efficacy of our proposed model in small-sized object detection, we have also presented the quantitative analysis of SOF-DETR using $AP@0.5$ metric and PR curve as shown in Fig. 4 and Fig. 5, respectively. In the case of class imbalance, PR-Curve presents a more reliable comparison

TABLE I
PERFORMANCE ANALYSIS USING MAP METRICS, WHERE S,M, AND L STAND FOR SMALL, MEDIUM, AND LARGE-SIZED OBJECTS.

Models	AP^{all}	AP^S	AP^M	AP^L	AP^s	AP^l
Pre-trained Faster-RCNN [4]	36.7	-	-	14.1	32.9	47.1
Fine-tuned Faster-RCNN-DC5 [13]	39.0	60.5	42.3	21.4	43.5	52.5
DETR (ResNet-50) [13]	42.0	62.4	44.2	20.5	45.8	61.1
<i>SOF-DETR (ResNet-50) (Ours)</i>	42.7	61.8	45.4	21.7	45.9	61.5

Fig. 5. Precision, Recall, and F1-Values comparison between DETR and SOF-DETR. All values are given on a scale of 0-100%.

of evaluation. These curves display the correlation between precision (positive predictive value) and recall (sensitivity) values for each attainable cut-off value. A curve higher than the other curve exhibits better performance and Fig.4 indicates that SOF-DETR outperforms other algorithms in most cases.

2) *Qualitative Analysis*: We have shown qualitative results of the proposed technique in Fig.2, and Fig. 3. The figures also depict DETR results for a comparative study. It is evident in Fig. 2, SOF-DETR detects small-sized present objects with high confidence scores, which DETR misses. Furthermore, SOF-DETR produces a higher confidence score for other small-sized objects than the confidence score of DETR. Fig. 2, also signifies higher confidence in predicting the small-sized objects without affecting the performance of medium-sized and large-sized objects. Another potential trace of the perceptible normalized inductive bias introduced in this study is depicted in Fig.3, which compares self-attention activation of SOF-DETR with self-attention maps of DETR. It is evident that the SOF-DETR attention maps are more vibrant for small-sized objects, including medium-sized and large-sized objects.

IV. CONCLUSION

This study improves the hard inductive bias of DETR for small-sized object detection without affecting the performance of medium-sized and large-sized objects. The normalized inductive bias is introduced using a lazy fusion of feature maps before passing it to the transformer layers of our proposed technique, SOF-DETR. The proposed technique shows higher confidence scores for detected small-sized objects and overall better performance than DETR. Future studies can explore another direction of improving inductive biases for small objects, like introducing a penalty in loss function for object sizes. This can be introduced in parallel to the normalized inductive biases. Future work can also include the study of such inductive biases in segmentation, panoptic segmentation, and instance segmentation.

ACKNOWLEDGMENT

This work was partially supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2014-3-00077, AI National Strategy Project) and by the Ministry of Culture, Sports, and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the culture Technology (CT) Research Development Program (R20200600020) 2021.

REFERENCES

- [1] R. Girshick, “Fast r-cnn,” in *IEEE International Conference on Computer Vision (ICCV)*, DOI 10.1109/ICCV.2015.169, pp. 1440–1448, 2015.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, pp. 1904–1916, 2015.
- [3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, DOI 10.1109/CVPR.2017.106, pp. 936–944, 2017.
- [4] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.
- [5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European Conference on Computer Vision (ECCV)*, pp. 21–37, 2016.
- [6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
- [7] B. Wu, F. Iandola, P. Jin, and K. Keutzer, “Squeezednet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pp. 446–454, 07 2017.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc.
- [9] S. Dubey, F. Olimov, M. A. Rafique, J. Kim, and M. Jeon, “Label-attention transformer with geometrically coherent objects for image captioning,” *ArXiv*, vol. abs/2109.07799, 2021.
- [10] N. Chen, S. Watanabe, J. Villalba, P. Åelasko, and N. Dehak, “Non-autoregressive transformer for speech recognition,” *IEEE Signal Processing Letters*, vol. 28, pp. 121–125, 2021.
- [11] A. Khan and B. Lee, “Gene transformer: Transformers for the gene expression-based classification of cancer subtypes,” *ArXiv*, vol. abs/2108.11833, 2021.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ICLR*, 2021.
- [13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.
- [14] G. Zhang, H.-C. Wong, and S.-L. Lo, “Multi-attention network for unsupervised video object segmentation,” *IEEE Signal Processing Letters*, vol. 28, pp. 71–75, 2021.
- [15] B. Neyshabur, “Towards learning convolutions from scratch,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, 2020.
- [16] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR*, vol. abs/1409.0473, 2015.
- [17] S. Fang, X. Pan, S. Xiang, and C. Pan, “Meta-msnet: Meta-learning based multi-source data fusion for traffic flow prediction,” *IEEE Signal Processing Letters*, vol. 28, pp. 6–10, 2021.
- [18] R. Stewart and M. Andriluka, “End-to-end people detection in crowded scenes,” 2015.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [20] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, and A. Ku, “Image transformer,” *ICML*, 2018.
- [21] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [22] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *AISTATS*, 2010.
- [23] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, “Fully convolutional instance-aware semantic segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.