2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669743

CT-CAD: Context-Aware Transformers for End-to-End Chest Abnormality Detection on X-Rays

Qiran Kong[1,2], Yirui Wu[1,2,*], Chi Yuan[1,2], Yongli Wang[3]
[1]Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing, China
Email: kds809545917@gmail.com, wuyirui@hhu.edu.cn, steve.yuan1990@gmail.com
[2]College of Computer and Information, Hohai University, Nanjing, China
[3]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
Email: yongliwang@njust.edu.cn

***Abstract*—Supervised based deep learning methods have achieved great success in medical image analysis domain. Essentially, most of them could be further improved by exploring and embedding context knowledge for accuracy boosting. Moreover, they generally suffer from slow convergency and high computing cost, which prevents their usage in a practical scenario. To tackle these problems, we present CT-CAD, context-aware transformers for end-to-end chest abnormality detection on X-Ray images. The proposed method firstly constructs a context-aware feature extractor, which enlarges receptive fields to encode multi-scale context information via an iterative feature fusion scheme and dilated context encoding blocks. Afterwards, deformable transformer detector are built for category classification and location regression, where their deformable attention block attend to a small set of key sampling points, thus allowing the transformer to focus on feature subspace and accelerate convergence speed. Through comparative experiments on Vinbig Chest and Chest Det 10 Datasets, the proposed CT-CAD demonstrates its effectiveness and outperforms the existing methods in mAP and training epoches.**

***Index Terms*—Chest X-Ray Images, Abnormality Detection, Context-Aware Feature Extractor, Deformable Transformer Detector**

I. Introduction

Chest X-Ray (CXR) Image is one of the most preferred diagnostic tools in medical practice, which has an important role in the diagnosis of thoracic diseases.

Applying deep learning methods to build automatical CXR diagnose tools is thus becoming a hot research topic, due to their scalability to process either big data or small size data and significant power to analyze complex CXR data with highly nonlinear modeling capability. Following such idea, researchers have made tremendous progress on chest abnormality detection, which is inspired by the great success of object detection methods in computer vision. For example, Baltruschat et al. [1] firstly pre-train a neural network on the ImageNet dataset for classification of natural images, and then utilize transfer learning for chest radiography analysis, which proves the efficiency of proper knowledge transfer on medical image analysis domain. Furthermore, Annarumma et al. [2] develop and test an artificial intelligence (AI) system for automated real-time triaging of adult chest radiographs, which uses an ensemble of two deep CNNs to predict the clinical priority from radiologic appearances only.

However, superimposition and overlapping of different anatomical structures locate along the projection direction, leading to the diversity of chest abnormalities. Hence, it's very difficult to detect abnormalities in some cases. Without special focus on these difficulties, most of the existing CXR abnormality detection methods derive their ideas or structure designs from object detection methods, which leads them to suffer from domain shift, requiring additional and specific knowledge embedding. Moreover, the structure of deep neural networks brings several inherent disadvantages, i.e., slow convergence and high computation cost, which would be worse facing various patterns of chest abnormalities.

Facing these challenges, we propose CT-CAD, context-aware transformers for end-to-end chest abnormality detection task. The proposed CT-CAD consists of two modules, i.e., context-aware feature extractor and deformable transformer detector. To address the issue of extracting multi-scale context information for small abnormalities locating, we not only design dilated context encoding blocks to enlarge receptive fields, but also propose an iterative feature fusion scheme to fuse multi-scale features. Regarded as a powerful network architecture based on attention mechanisms for machine translation, deformable transformer detector adaptively aggregates the key and distinguish features without any hand-designed components, thus enhancing feature representation capability to solve difficulties of multiple and complex pattern discovery. The core of deformable transformer detector, i.e., deformable attention block, attends to a small set of sampling locations as a pre-filter for prominent key elements out of the whole feature space, which could be considered as context information on feature subspace and greatly decrease computation and memory cost at both training and testing.

Authorized licensed use limited to: Danmarks Tekniske Informationscenter. Downloaded on February 06,2023 at 14:50:42 UTC from IEEE Xplore. Restrictions apply.

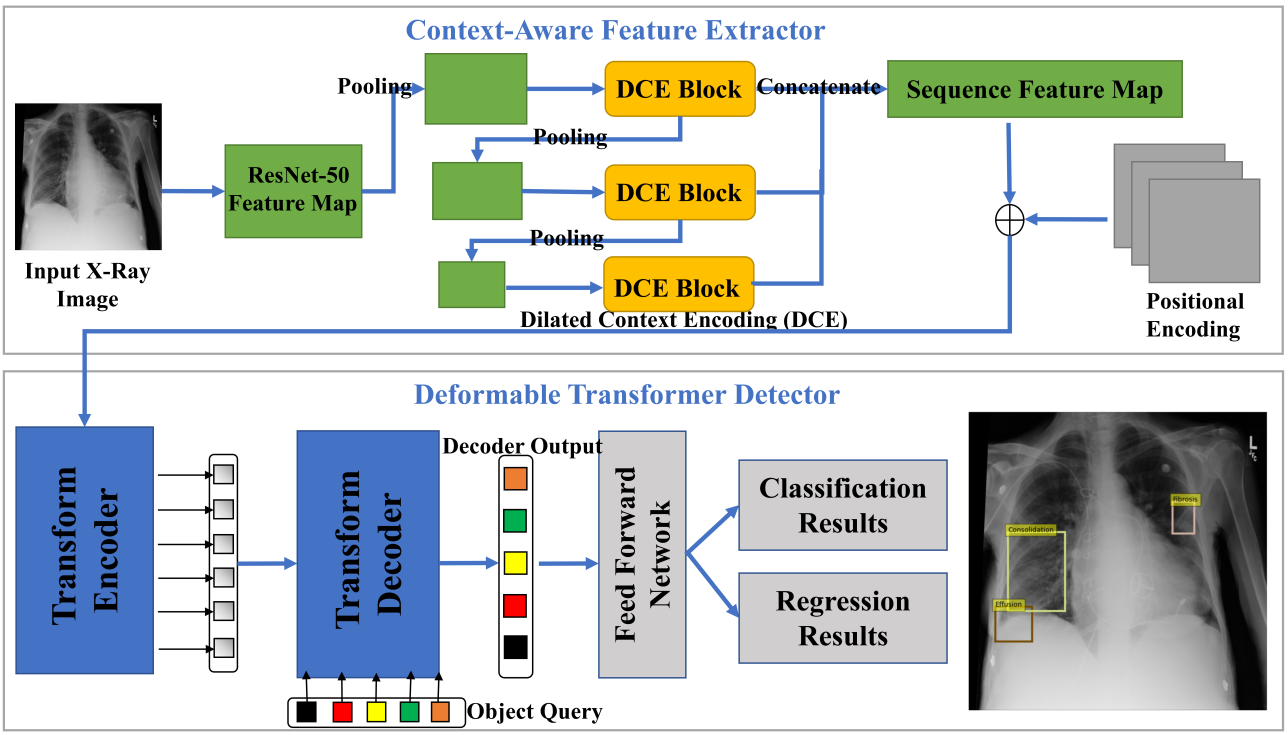

Fig. 1. Network architecture of the proposed CT-CAD method. It's noted that CT-CAD could output a set of predictions without pre- and post-processing steps in an end-to-end manner.

II. The Proposed Method

A. Network Structure Overview

As shown in Fig. 1, we design two main modules, i.e., context-aware feature extractor and deformable transformer detector. The former module contains ResNet-50 backbone, dilated context encoding (DCE) block, and positional encoding structure, while the latter one contains transformer encoder, transformer decoder and a feed-forward network for classification and regression tasks.

It's noted that the proposed CT-CAD method require a fixed number N_{obj} for possible predictions, each with a coordinate regression results and an abnormality classification result. Let y the ground truth and $\hat{y} = \{\hat{y}_i\}_{i=1}^{N_{obj}}$ denotes the set of N_{obj} predictions. The total loss for both regression and classification tasks is achieved by searching for a permutation $\omega \in \Omega_{N_{obj}}$ of the N_{obj} predictions with Hungarian algorithm, which could be described as:

$$\hat{\omega} = \underset{\omega \in \Omega_N}{\arg\min} \sum_{i=1}^{N} L_{\text{match}} \left(y_i, \hat{y}_{\omega(i)}\right) \tag{1}$$

where y is padded to the size of N_{obj}, $\hat{y_{\omega(i)}}$ is the ith element of the predictions. Each element of the prediction refers to $\hat{y}_{\omega(i)} = \left(\hat{p}_{\omega(i)}\left(c_i\right), \hat{b}_{\omega(i)}\right)$, where $\hat{b}_{\omega(i)}$ represents the bounding box and $\hat{p}_{\omega(i)}\left(c_i\right)$ represents the probability of the class with the maximum probability.

The loss function for training is a combination of the box loss and classification loss, which is defined as:

$$L(\hat{y}, y) = \sum_{i=1}^{N} \left[\alpha_1 L_{cls}\left(c_i, \hat{p}_{\omega(i)}\left(c_i\right)\right) + \alpha_2 L_{loc}\left(b_i, \hat{b}_{\omega_{(i)}}\right)\right] \tag{2}$$

where the classification loss is the cross entropy, represented as

$$L_{cls}\left(c_i, \hat{p}_{\omega(i)}\left(c_i\right)\right) = \sum_{i=1}^{N} -\log \hat{p}_{\omega(i)}\left(c_i\right) \tag{3}$$

and the bounding box loss is

$$L_{loc}\left(b_i, \hat{b}_{\omega_{(i)}}\right) = \sum_{i=1}^{N} \left[\beta_1 L_{iou}\left(b_i, \hat{b}_{\omega_{(i)}}\right) + \beta_2 L_{reg}\left(b_i, \hat{b}_{\omega_{(i)}}\right)\right] \tag{4}$$

which is essentially the summation of IoU loss and L_1 loss. It's noted that α_1, α_2, β_1 and β_2 are all hyper-parameters. Specifically, we adopt GIoU [3] to balance the loss between large and small objects. Parameters of the proposed CT-CAD method are updated based on the loss obtained by the best search of permutation, which enables the proposed network to be trained in an end-to-end manner without many hand designed components.

B. Design of Context-Aware Feature Extractor

Inspired by [4], we design the proposed iterative feature fusion scheme for multi-scale feature fusion as shown in Fig. 1. Essentially, the proposed feature fusion scheme builds on

Authorized licensed use limited to: Danmarks Tekniske Informationscenter. Downloaded on February 06,2023 at 14:50:42 UTC from IEEE Xplore. Restrictions apply.

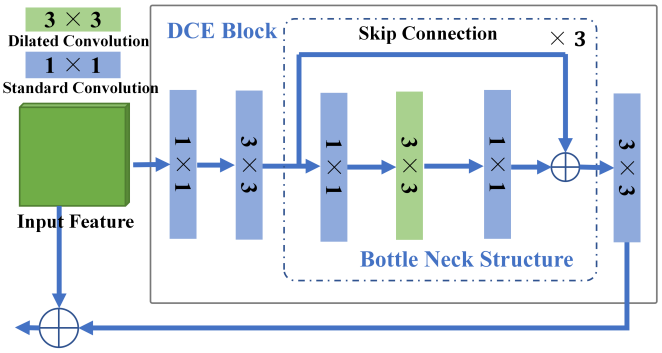

Fig. 2. Architecture design of the proposed DCE block, where the dilated convolution filter is adopted to enlarge the receptive field, thus acquiring quantity of multi-scale context information for further processing.

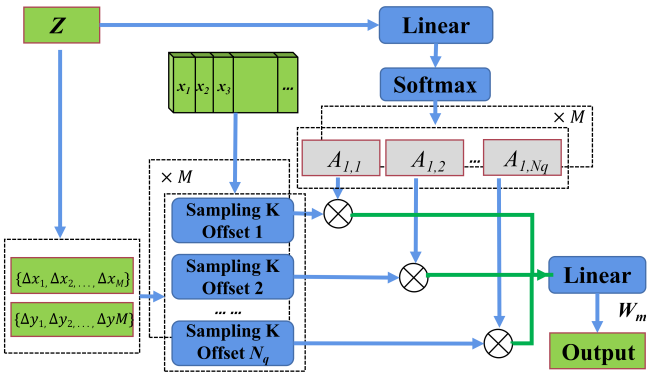

Fig. 3. Architecture design of the proposed deformable attention block, which is the core component of deformable transformer detector. It's noted that K points are sampled from the input multi-scale feature map.

top of the Feature Pyramid Networks (FPN) [5] by iteratively and progressively refining scaled feature map from the top layers to the bottom-up ones. Unrolling the iterative structure to a sequential implementation, we obtain feature map for abnormality detector that looks at the images twice or more with structures of multiple stages, and much more carefully with DCE blocks to enhance feature representation. Similar to the cascaded detector in Cascade structure, the proposed feature fusion scheme iteratively enhances original feature map of FPN to generate increasingly powerful representations. In other words, the proposed feature fusion scheme acts as a multi-scale feature encoding scheme in a global sense by directly resizing feature map, meanwhile DCE blocks encodes multi-scale information in a local sense by enlarging receptive via fields convolutional filters with different sizes. Such iterative feature fusion operations could be represented as

$$\begin{cases} F_l = F_{l-1} + f_{DCE}(F_{l-1}) \\ F_{l+1} = f_{down}(F_l) \end{cases} \tag{5}$$

where F_l refers to the lth feature map after $l-1$ times pooling operations, functions $f_{DCE}()$ and $f_{down}()$ represent single-in-single-out operator of DCE block and down-sampling operator, l varies from 2 to 4 in the proposed method.

Inspired by YOLOF [6], we design structure of DCE block as shown in Fig. 2, where dilated convolution and skip connections are used to enlarge the receptive field and capture more local context information. Essentially, this powerful one-level feature successfully finds a way to generate an output feature with various receptive fields, compensating for the lack of multiple-level features. Therefore, it exceeds the range of scales matching to the scaled feature's receptive field, which benefits the detection performance for abnormalities across various scales.

Specifically, we first design a 1×1 and a 3×3 standard convolution layer as a projector, which is used for feature refinement. The main component in DCE block is the residual block, which consists of two 1×1 convolution layer with a 3×3 dilated convolution layer. Then, we stack several residual blocks with residual connection to build a short-way for gradient flow. Each residual block has a different dilated rates with different receptive field, covering all scales and extracting extensive contextual information. Finally, we sum the resulting feature map with the original feature map for output.

C. Design of Deformable Transformer Detector

The proposed deformable attention block is illustrated in Fig. 3 with single-scale and multi-head attention property. Given a sequence input feature, we first obtain query feature $\boldsymbol{z}$ and feature map $\boldsymbol{x}$ via several linear layers. By applying linear layers on $\boldsymbol{z}$, we can compute multi-head offsets $\{\Delta \boldsymbol{x}_m, \Delta \boldsymbol{y}_m\}_{m=1}^{M}$ and the corresponding attention weights A. It's noted that each pair of offsets is used to sample k points from the feature map $\boldsymbol{x}$. Afterwards, single-scale and multi-head deformable attention block can be defined as:

$$At(A, x) = \sum_{m=1}^{M} W_m [\sum_{k=1}^{K} A_{m,k} \cdot f_{off}([x, \Delta x_m, \Delta y_m]_k)] \tag{6}$$

where m indexes the attention head, k indexes the sampled keys, M and K are total number of attention heads and sampling points, respectively.

Furthermore, the efficiency property of single-scale and multi-head deformable attention block leads multi-scale deformable attention block to be easily built as:

$$\begin{aligned} &MsAt(\{A^l\}_{l=1}^{L}, \{x^l\}_{l=1}^{L}) = \\ &\textstyle\sum_{m=1}^{M} \boldsymbol{W}_m [\sum_{l=1}^{L} \sum_{k=1}^{K} A_{m,k}^{L} \cdot f_{off}([x^l, \Delta x_m^l, \Delta y_m^l]_k)] \end{aligned} \tag{7}$$

where l refers to the index of layers and L is the total number of layers. We stack 6 deformable encoder and decoder layers with deformable attention blocks to achieve decoder output, whose size is (N_{obj}, c_{out}). It's noted that N_{obj} is the number of the abnormalities detected and c_{out} is the output dimension of decoder layers.

Authorized licensed use limited to: Danmarks Tekniske Informationscenter. Downloaded on February 06,2023 at 14:50:42 UTC from IEEE Xplore. Restrictions apply.

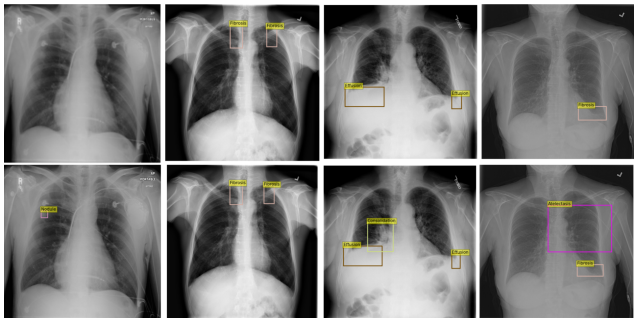

Fig. 4. Comparisons of detection results between Faster R-CNN the proposed CT-CAD. Faster R-CNN is above and misses some abnormalities.

TABLE I
PERFORMANCE COMPARISON AMONG CT-CAD AND THE EXISTING METHODS, WHERE BOLD TEXTS REFER TO THE BEST PERFORMANCE.

Dataset	Method	AP50
VinBig	Faster R-CNN with FPN	29.1
	Yolov3 [7]	26.2
	DETR [8]	33.5
	Cascade R-CNN [9]	33.5
	Yolo Modified	29.5
	Faster R-CNN Modified	30.3
	Ensemble Model 1	35.7
	Ensemble Model 2	34.3
	Ensemble Model 3	33.9
	CT-CAD	**36.3**
Chest Det-10	Faster R-CNN with FPN	39.3
	Yolov3 [7]	37.7
	DETR [8]	41.5
	Cascade R-CNN [9]	41.1
	DenseNet [10]	42.7
	CT-CAD	**43.6**

III. EXPERIMENTS AND ANALYSIS

We adopt two datasets to conduct chest X-Ray abnormality detection, i.e., Vinbig Chest X-Ray Dataset and ChestX Det-10 Dataset. For former dataset, we select a subset for experiments, which contains 5000 training images and 1063 testing images in total. With annotations of bounding boxes and the corresponding class labels, all images are labeled by a panel of experienced radiologists for the presence of 14 critical radiographic findings. The latter one is a subset Dataset with box annotations of a public dataset NIH Chest-14, which contains 3001 and 541 images in the training set and testing set, respectively. It's noted each image is annotated with 10 common categories of diseases.

Experimental results of performance comparison on VinBig Dataset and ChestX Det-10 Dataset are shown in Table. I. From Table. I, we could observe that accuracy in VinBig Dataset is generally lower than ChestX Det-10 Dataset, since CXR images in VinBig Dataset not only correspond to more categories of abnormalities, but also vary in appearance with more complex patterns. It's observed that the proposed CT-CAD has achieved the highest AP50 on both datasets, which outperforms Faster R-CNN, YoLo and their modified versions by a large margin. All these facts prove structures of deformable transformer detector and dilated context encoder are helpful to improve detection accuracy.

On the challenging VinBig dataset, CT-CAD achieves competitive performance comparing with the ensemble baseline 1, which is a complicated structure that ensembles the results of five different detectors. So are the other two ensemble baseline methods. All these facts point out that complexity in structure design not always brings advantages on performance boosting. When comparing with DETR, the better performance obtained by CT-CAD shows that the proposed deformable attention block can help focus on informative feature subspace without having to look over the entire space, which might bring noise information to decrease accuracy of detection results.

In Fig. 3, we compare the abnormality detection accuracy between the proposed CT-CAD and Faster R-CNN, where we can view that CT-CAD is capable to detect hard cases, such as nodules that are ignored by Faster R-CNN.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Central Universities under Grant B200202177.

REFERENCES

[1] I. M. Baltruschat, L. Steinmeister, H. Ittrich, G. Adam, H. Nickisch, A. Saalbach, J. von Berg, M. Grass, and T. Knopp, "When does bone suppression and lung field segmentation improve chest x-ray disease classification?" in *Proceedings of 16th IEEE International Symposium on Biomedical Imaging*, 2019, pp. 1362–1366.

[2] M. Annarumma, S. J. Withey, R. J. Bakewell, E. Pesce, V. Goh, and G. Montana, "Automated triaging of adult chest radiographs with deep artificial neural networks," *Radiology*, vol. 291, no. 1, pp. 196–202, 2019.

[3] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[4] S. Qiao, L. Chen, and A. L. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," *CoRR*, vol. abs/2006.02334, 2020.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[6] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," *arXiv preprint arXiv:2103.09460*, 2021.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[8] 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*. Springer, 2020, pp. 213–229.

[9] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," in *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

Authorized licensed use limited to: Danmarks Tekniske Informationscenter. Downloaded on February 06,2023 at 14:50:42 UTC from IEEE Xplore. Restrictions apply.