Visual Grounding of Whole Radiology Reports for 3D CT Images

Akimichi Ichinose[1(✉)], Taro Hatsutani[1], Keigo Nakamura[1], Yoshiro Kitamura[1], Satoshi Iizuka[2], Edgar Simo-Serra[3], Shoji Kido[4], and Noriyuki Tomiyama[4]

[1] Medical Systems Research and Development Center, FUJIFILM Corporation, Tokyo, Japan
akimichi.ichinose@fujifilm.com

[2] Center for Artificial Intelligence Research, University of Tsukuba, Ibaraki, Japan

[3] Department of Computer Science and Engineering, Waseda University, Tokyo, Japan

[4] Graduate School of Medicine, Osaka University, Osaka, Japan

Abstract. Building a large-scale training dataset is an essential problem in the development of medical image recognition systems. Visual grounding techniques, which automatically associate objects in images with corresponding descriptions, can facilitate labeling of large number of images. However, visual grounding of radiology reports for CT images remains challenging, because so many kinds of anomalies are detectable via CT imaging, and resulting report descriptions are long and complex. In this paper, we present the first visual grounding framework designed for CT image and report pairs covering various body parts and diverse anomaly types. Our framework combines two components of 1) anatomical segmentation of images, and 2) report structuring. The anatomical segmentation provides multiple organ masks of given CT images, and helps the grounding model recognize detailed anatomies. The report structuring helps to accurately extract information regarding the presence, location, and type of each anomaly described in corresponding reports. Given the two additional image/report features, the grounding model can achieve better localization. In the verification process, we constructed a large-scale dataset with region-description correspondence annotations for 10,410 studies of 7,321 unique patients. We evaluated our framework using grounding accuracy, the percentage of correctly localized anomalies, as a metric and demonstrated that the combination of the anatomical segmentation and the report structuring improves the performance with a large margin over the baseline model (66.0% vs 77.8%). Comparison with the prior techniques also showed higher performance of our method.

Keywords: Deep Learning · Vision Language · Visual Grounding · Computed Tomography

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_59.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 611–621, 2023.
https://doi.org/10.1007/978-3-031-43904-9_59

1 Introduction

In recent years, a number of medical image recognition systems have been developed [6] to alleviate the increasing burden on radiologists [2,21,22]. In the development of such systems, the task of manually labeling images is a significant bottleneck. Auto-labeling, the process of automatically assigning labels to images using machine learning algorithms, has emerged as a promising solution to this problem. In cases where there are plenty of image and caption pairs, one potential approach to auto-labeling is visual grounding [12], which utilizes natural language descriptions to identify and localize objects in images.

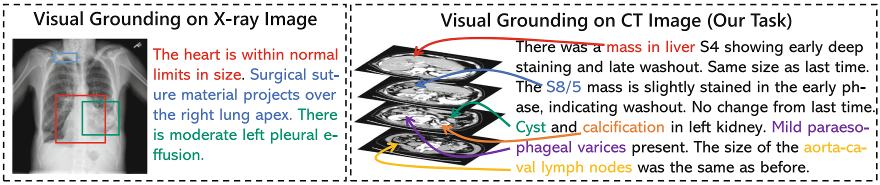

Visual Grounding on X-ray Image
The heart is within normal limits in size. Surgical suture material projects over the right lung apex. There is moderate left pleural effusion.
Visual Grounding on CT Image (Our Task)
There was a mass in liver S4 showing early deep staining and late washout. Same size as last time. The S8/5 mass is slightly stained in the early phase, indicating washout. No change from last time. Cyst and calcification in left kidney. Mild paraesophageal varices present. The size of the aorta-caval lymph nodes was the same as before.

Fig. 1. Comparison of the visual grounding task on X-ray image and on CT image.

With the recent advances in cross-modal technology based on deep learning, many frameworks for visual grounding has been proposed [7,11]. Within the medical domain, several large scale datasets with radiology reports are available (e.g. OpenI [3], MIMIC-CXR [9]), and these produced researches on medical image visual grounding [1,25]. However, to the best of our knowledge, prior studies have focused on 2D X-ray images [28] or videos [15], and there has been no research applying visual grounding to 3D computed tomography (CT) images so far. Visual grounding on CT images has the following difficulties: 1) **Large number of anomaly types to detect**: Existing researches on visual grounding using X-ray images handled only chest X-ray images. The number of anomaly types to detect is at most dozen or so (e.g. 13 findings [8]). In contrast, our research handles CT images including various parts of the human body. Consequently, the number of anomaly types to be detected is larger than one hundred. 2) **Long and complex sentences**: Radiology reports on X-ray images are often simple, noting only the presence or absence of anomalies. On the other hand, in CT examinations, the qualitative diagnosis of each anomaly is often performed. In cases, multiple anomalies are simultaneously described in a sentence. Therefore, the description tend to be long and complicated with multiple sentences (Fig. 1). Visual grounding for CT images requires the extraction of information about the location and type of each anomaly from these complex sentences.

In this work, we propose a novel visual grounding framework for 3D CT images and radiology reports. The main idea is to separate the task into three parts: 1) anatomical segmentation on images, 2) report structuring, and 3) localization of described anomalies. In the anatomical segmentation, multiple organs and tissues are extracted using the deep learning based segmentation model and provided as landmarks. The report structuring model, which is based on BERT [5], is also introduced to extract information of each anomaly from a complex report. Both of these features are fed into the grounding model (3) to extrapolate medical domain knowledge, thereby enabling accurate visual grounding.

Our contributions are as follows:

- We show the first visual grounding results for 3D CT images that covers various body parts and anomalies.
- We introduce a novel grounding architecture that can leverage report structuring results of presence/type/location of described anomalies.
- We validate the efficacy of the proposed framework using a large-scale dataset with region-description correspondence annotations.

2 Related Work

Visual Grounding. Visual grounding task involves learning the correspondences between descriptions in the text and image regions from a given training set of region-description pairs [12]. There are mainly two approaches: one-stage approach and two-stage approach. Most studies follow a two-stage approach [14,17]. However, this approach usually employs a pre-trained object detector, and it leads to restrict the capability of categories and attributes in grounding. Accordingly, recent studies is shifting to employ the one-stage approach, in which visual grounding is performed by end-to-end training [4,10,27].

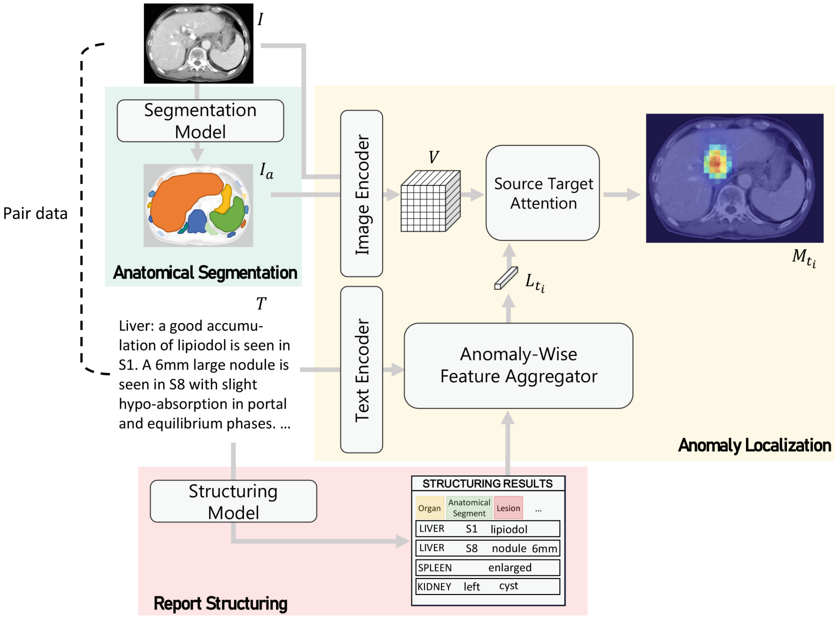
Vision-Language Tasks on Medical Image. The existence of public datasets with paired images and reports [3,9,26] has accelerated research on cross-modal tasks in the medical field [16,25]. Inspired by the success of visual grounding, several studies of visual grounding for medical images and radiology reports have also been reported [1,23,28]. These studies utilized a large scale dataset and an attention-based language interpretation model such as BERT [5] to ground the descriptions in the report. However, these studies have focused on X-ray images, and to the best of our knowledge, there have been no studies on CT images, which cover the entire body and have a complex report.

3 Methods

We first formulate the problem. Next, we explain three key components of anatomical segmentation, report structuring, and anomaly localization in our framework. In our framework, multiple organ labels obtained as the output of anatomical segmentation encourage the grounding model to learn detailed anatomy, and report structuring allows the grounding model to accurately extract the features of the target anomaly from complex sentences.

3.1 Problem Formulation

Our research assumes that a dataset of image-report pairs with region-description correspondence annotations is provided for training. We show the overall framework in Fig. 2. We denote an image and a paired report as I and T respectively. Let I_a be a label image in which multiple organs are extracted from I. Each report T contains descriptions of multiple (image) anomalies. We denote each anomalies as $t_i \in \{t_1, t_2, ..., t_N\}$. Given an image I and corresponding organ label images I_a encoded as $V \in \mathbb{R}^{d_z \times d_y \times d_x \times d}$ and a description about an anomaly t_i encoded as $L_{t_i} \in \mathbb{R}^d$, the goal of our framework is to generate a segmentation map M_{t_i} that represents the location of the anomaly t_i.



Liver: a good accumulation of lipiodol is seen in S1. A 6mm large nodule is seen in S8 with slight hypo-absorption in portal and equilibrium phases. ...

Organ	Anatomical Segment	Lesion
LIVER	S1	lipiodol
LIVER	S8	nodule 6mm
SPLEEN		enlarged
KIDNEY	left	cyst

Fig. 2. The proposed framework for 3D-CT visual grounding.

3.2 Anatomical Segmentation

The task of the anatomical segmentation is to extract relevant anatomies that can be clues for visual grounding. We use the commercial version of the 3D image analysis software (Synapse 3D V6.8, FUJIFILM corporation, Japan) to extract 32 organs and tissues (See Appendix Table. A1). In this software, anatomies are extracted using U-Net based architectures [13,18]. The extracted anatomical label images are I_a.

3.3 Report Structuring

The tasks of the report structuring are as follows: 1) anatomical prediction, 2) phrase recognition, and 3) relationship estimation between phrases (See Appendix Fig. A1). The anatomical prediction is a sentence-wise prediction to determine which organ or body part is mentioned in each sentence. The organs and body parts to be recognized are shown in Appendix Table. A2. The sentences belonging to the same class are concatenated, then the phrase recognition and the relationship estimation are performed for each class.

The phrase recognition module extracts phrases and classifies each of them into 9 classes (See Appendix Table. A2). Subsequently, the relationship estimation module determines whether there is a relationship between anomaly phrases (e.g. 'nodule', 'fracture') and other phrases (e.g. '6mm', 'Liver S6'), resulting in the grouping of phrases related to the same anomaly. If multiple anatomical phrases are grouped in the same group, they are split into separate groups on a rule basis (e.g. [‘right S1’, ‘left S6’, ‘nodule’] -> [‘right S1’, ‘nodule’], [‘left S6’, ‘nodule’]). More details of implementation and training methods are reported in Nakano et al. [20] and Tagawa et al [24].

3.4 Anomaly Localization

The task of the anomaly localization is to output a localization map of the anomaly mentioned in the input report T. The CT image I and the organ label image I_a are concatenated along the channel dimension and encoded by a convolutional backbone to generate a visual embedding V. The sentences in the report T are encoded by BERT [5] to generate embeddings for each character. Let $r = \{r_1, r_2, ..., r_{N_C}\}$ be the set of character embeddings where N_C is the number of characters. Our framework next adopt the Anomaly-Wise Feature Aggregator (AFA). For each anomaly t_i, AFA generates a representative embedding L_{t_i} by aggregating the embeddings of related phrases based on report structuring results. The final grounding result M_{t_i} is obtained by the following Source-Target Attention.

$$M_{t_i} = \text{sigmoid}(L_{t_i} W_Q (V W_K)^T) \tag{1}$$

where W_Q, $W_K \in \mathbb{R}^{d \times d_n}$ are trainable variables.

The overall architecture of this module is illustrated in Appendix Fig. A2.

Anomaly-Wise Feature Aggregator. The results of the report structuring $m_{t_i} \in \mathbb{R}^{N_C}$ are defined as follows:

$$m_{t_{ij}} = \begin{cases} c_j & \text{if a } j\text{-th character is related to an anomaly } t_i, \\ 0 & \text{else.} \end{cases} \tag{2}$$

$$m_{t_i} = \{m_{t_{i1}}, m_{t_{i2}}, ...m_{t_{iN_C}}\} \tag{3}$$

where c_j is the class index labeled by the phrase recognition module (Let C be the number of classes). In this module, aggregate character-wise embeddings based on the following formula.

$$e_k = \{r_j | m_{t_{ij}} = k\} \tag{4}$$

$$L_{t_i} = \mathrm{LSTM}([v_{organ}; p_1; e_1; p_2; e_2; ..., p_C; e_C]) \tag{5}$$

where v_{organ} and p_k are trainable embeddings for each organ and each class label respectively. $[\cdot;\cdot]$ stands for concatenation operation. In this way, embeddings of characters related to the anomaly t_i are aggregated and concatenated. Subsequently, representative embeddings of the anomaly are generated by an LSTM layer. In the task of visual grounding focused on 3D CT images, the size of the dataset that can be created is relatively small. Considering this limitation, we use an LSTM layer with strong inductive bias to achieve high generalization performance.

4 Dataset and Implementation Details

4.1 Clinical Data

We retrospectively collected 10,410 CT studies (11,163 volumes/7,321 unique patients) and 671,691 radiology reports from one university hospital in Japan. We assigned a bounding box to each anomaly described in the reports as shown in Appendix Fig. A3. The total category number is about 130 in combination of anatomical regions and anomaly types (The details are in Fig. 4) For each anomaly, a correspondence annotation was made with anomaly phrases in the report. The total number of annotated regions is 17,536 (head: 713 regions, neck: 285 regions, chest: 8,598 regions, and abdomen: 7,940 regions). We divide the data into 9,163/1,000/1,000 volumes as a training/validation/test split.

4.2 Implementation Details

We use a VGG-like network as Image Encoder, with 15 3D-convolutional layers and 3 max pooling layers. For training, the voxel spacings in all three dimensions are normalized to 1.0 mm. CT values are linearly normalized to obtain a value of [0–1]. The anatomy label image, in which only one label is assigned to each voxel, is also normalized to the value [0–1], and the CT image and the label image are concatenated along the channel dimension. As our Text Encoder, we use a BERT with 12 transformer encoder layers, each with hidden dimension of 768 and 12 heads in the multi-head attention. At first, we pre-train the BERT using 6.7M sentences extracted from the reports in a Masked Language Model task. Then we train the whole architecture jointly using dice loss [19] with the first 8 transformer encoder layers of the BERT frozen. Further information about implementation are shown in Appendix Table. A3.

5 Experiments

We did two kinds of experiments for comparison and ablation studies. The comparison study was made against TransVG [4] and MDETR [10] that are one-stage visual grounding approaches and established state-of-the-art performances on photos and captions. To adapt TransVG and MDETR for the 3D modality, the backbone was changed to a VGG-like network with 3D convolution layers, the same as the proposed method. We refer one of the proposed method without anatomical segmentation and report structuring as the baseline model.

5.1 Evaluation Metrics

We report segmentation performance using Dice score, mean intersection over union (mIoU), and the grounding accuracy. The output masks are thresholded to compute mIoU and grounding accuracy score. The mIoU is defined as an average IoU over the thresholds [0.1, 0.2, 0.3, 0.4, 0.5]. The grounding accuracy is defined as the percentage of anomalies for which the IoU exceeds 0.1 under the threshold 0.1.

5.2 Results

The experimental results of the two studies are shown in Table. 1. Both of MDETR and TransVG failed to achieve stable grounding in this task. A main difference between these models and our baseline model is using a source-target attention layer instead of the transformer. It is known that a transformer-based algorithm with many parameters and no strong inductive bias is difficult to generalize with such a relatively limited number of training data. For this reason, the baseline model achieved a much higher accuracy than the comparison methods.

Table 1. Results of the comparison/ablation studies. '-' represents 'not converged'.

Method	Anatomical Seg.	Report Struct.	Dice [%]	mIoU [%]	Accuracy [%]
MDETR [10]	–	–	N/A	–	–
TransVG [4]	–	–	N/A	8.5	21.8
Baseline	✗	✗	27.4	15.6	66.0
Proposed	✓	✗	28.1	16.6	67.9
	✗	✓	33.0	20.3	75.9
	✓	✓	**34.5**	**21.5**	**77.8**

The ablation study showed that the anatomical segmentation and the report structuring can improve the performance. In Fig. 3 (upper row), we demonstrate several cases that facilitate an intuitive understanding of each effect. Longer reports often mention more than one anomaly, making it difficult to recognize the grounding target and cause localization errors. The proposed method can explicitly indicate phrases such as the location and size of the target anomaly, reducing the risk of failure. Figure 3 (lower row) shows examples of grounding results when a query that is not related to the image is inputted. In this case, the grounding results were less consistent with the anatomical phrases. The results suggest that the model performs grounding with an emphasis on anatomical information against the backdrop of abundant anatomical knowledge.

The grounding performance for each combination of organ and anomaly type is shown in Fig. 4. The performance is relatively high for organ shape abnormalities (e.g. swelling, duct dilation) and high-frequency anomalies in small organs (e.g. thyroid/prostate mass). For these anomaly types, our model is considered to be available for automatic training data generation. On the other hand, the performance tends to be low for rare anomalies (e.g. mass in small intestine) and anomalies in large body part (e.g. limb). Improving grounding performance for these targets will be an important future work.

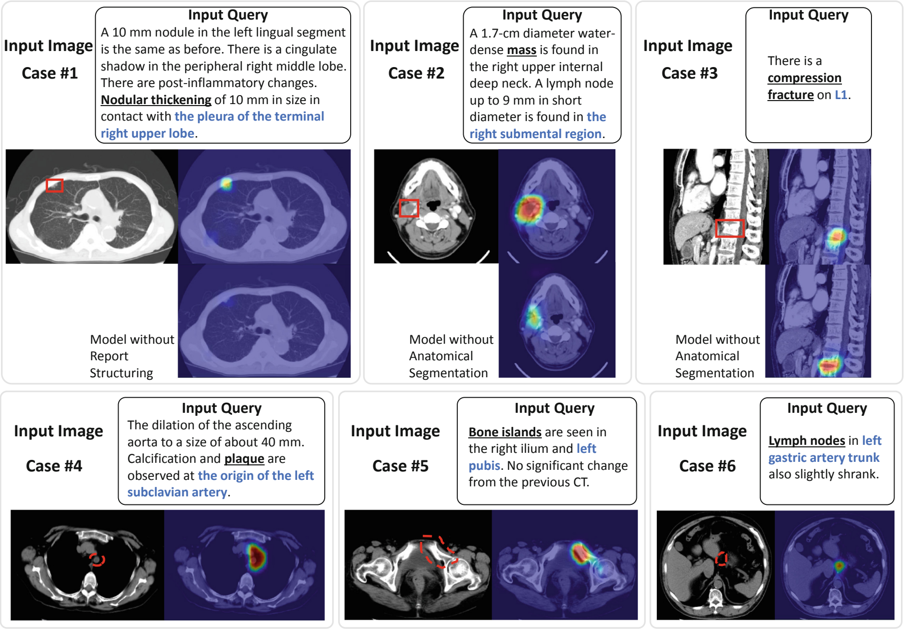
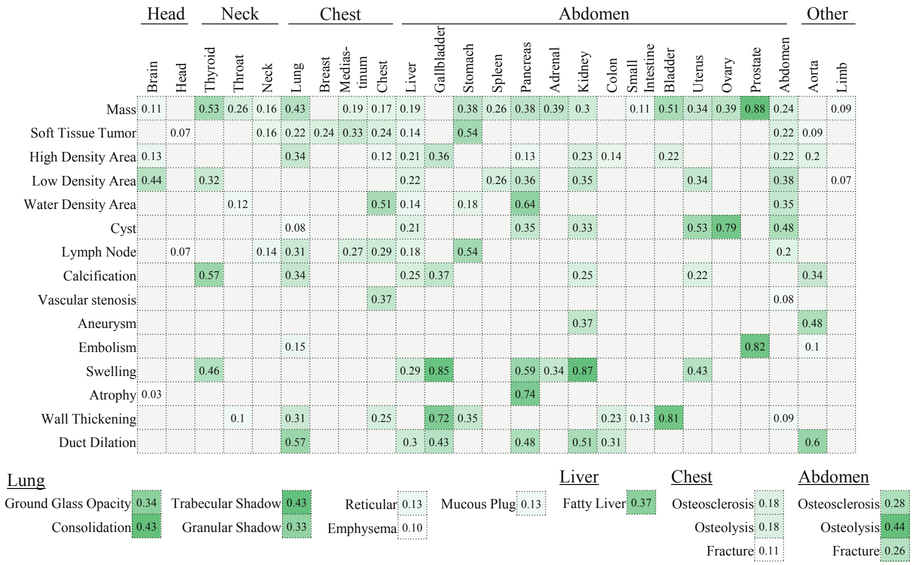

Input Image
Case #1
Input Query
A 10 mm nodule in the left lingual segment is the same as before. There is a cingulate shadow in the peripheral right middle lobe. There are post-inflammatory changes. Nodular thickening of 10 mm in size in contact with the pleura of the terminal right upper lobe.
Model without Report Structuring
Input Image
Case #2
Input Query
A 1.7-cm diameter water-dense mass is found in the right upper internal deep neck. A lymph node up to 9 mm in short diameter is found in the right submental region.
Model without Anatomical Segmentation
Input Image
Case #3
Input Query
There is a compression fracture on L1.
Model without Anatomical Segmentation
Input Image
Case #4
Input Query
The dilation of the ascending aorta to a size of about 40 mm. Calcification and plaque are observed at the origin of the left subclavian artery.
Input Image
Case #5
Input Query
Bone islands are seen in the right ilium and left pubis. No significant change from the previous CT.
Input Image
Case #6
Input Query
Lymph nodes in left gastric artery trunk also slightly shrank.

Fig. 3. The grounding results for several input queries. Underlines in the input query indicate the target anomaly phrase to be grounded. The phrases highlighted in bold blue indicate the anatomical locations of the target anomaly. The red rectangles indicate the ground truth regions. Case #4-#6 are the grounding results when an unrelated input query is inputted. The region surrounded by the red dashed line indicates the anatomical location corresponding to the input query.



	Head		Neck			Chest				Abdomen													Other	
	Brain	Head	Thyroid	Throat	Neck	Lung	Breast	Medias-tinum	Chest	Liver	Gallbladder	Stomach	Spleen	Pancreas	Adrenal	Kidney	Colon	Small Intestine	Bladder	Uterus	Ovary	Prostate	Abdomen	Aorta	Limb
Mass	0.11		0.53	0.26	0.16	0.43		0.19	0.17	0.19		0.38	0.26	0.38	0.39	0.3		0.11	0.51	0.34	0.39	0.88	0.24		0.09
Soft Tissue Tumor		0.07			0.16	0.22	0.24	0.33	0.24	0.14		0.54											0.22	0.09	
High Density Area	0.13					0.34			0.12	0.21	0.36			0.13		0.23	0.14		0.22				0.22	0.2	
Low Density Area	0.44		0.32							0.22			0.26	0.36		0.35				0.34			0.38		0.07
Water Density Area				0.12					0.51	0.14		0.18		0.64									0.35		
Cyst						0.08				0.21				0.35		0.33				0.53	0.79		0.48		
Lymph Node		0.07			0.14	0.31		0.27	0.29	0.18		0.54											0.2		
Calcification			0.57			0.34				0.25	0.37					0.25				0.22				0.34	
Vascular stenosis									0.37														0.08		
Aneurysm																0.37								0.48	
Embolism						0.15																0.82		0.1	
Swelling			0.46							0.29	0.85			0.59	0.34	0.87				0.43					
Atrophy	0.03													0.74											
Wall Thickening				0.1		0.31			0.25		0.72	0.35					0.23	0.13	0.81				0.09		
Duct Dilation						0.57				0.3	0.43			0.48		0.51	0.31							0.6	

Lung	
Ground Glass Opacity	0.34
Consolidation	0.43
Trabecular Shadow	0.43
Granular Shadow	0.33
Reticular	0.13
Emphysema	0.10
Mucous Plug	0.13

Liver	
Fatty Liver	0.37

Chest	
Osteosclerosis	0.18
Osteolysis	0.18
Fracture	0.11

Abdomen	
Osteosclerosis	0.28
Osteolysis	0.44
Fracture	0.26

Fig. 4. Grounding performance for representative anomalies. The value in each cell is the average dice score of the proposed method.

6 Conclusion

In this paper, we proposed the first visual grounding framework for 3D CT images and reports. To deal with various type of anomalies throughout the body and complex reports, we introduced a new approach using anatomical recognition results and report structuring results. The experiments showed the effectiveness of our approach and achieved higher performance compared to prior techniques. However, in clinical practice, radiologists write reports from comparing multiple images such as time-series images, or multi-phase scans. Realizing such sophisticated diagnose process by a visual grounding model will be a future research.

References

1. Bhalodia, R., et al.: Improving pneumonia localization via cross-attention on medical images and reports. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 571–581. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_53
2. Dall, T.: The Complexities of Physician Supply and Demand: Projections from 2016 to 2030. IHS Markit Limited (2018)
3. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
4. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: TransVG: end-to-end Visual Grounding With Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1769–1779 (2021)

5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
6. Ebrahimian, S., et al.: FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. Acad. Radiol. **29**(4), 559–566 (2022)
7. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 108–124. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_7
8. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
9. Johnson, A.E., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 317 (2019)
10. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: MDETR-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1780–1790 (2021)
11. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
12. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of Advances in Neural Information Processing System, pp. 1889–1897 (2014)
13. Keshwani, D., Kitamura, Y., Li, Y.: Computation of total kidney volume from ct images in autosomal dominant polycystic kidney disease using multi-task 3D convolutional neural networks. In: Shi, Y., Suk, H.-I., Liu, M. (eds.) MLMI 2018. LNCS, vol. 11046, pp. 380–388. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00919-9_44
14. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision, pp. 201–216 (2018)
15. Li, B., Weng, Y., Sun, B., Li, S.: Towards visual-prompt temporal answering grounding in medical instructional video. arXiv preprint arXiv:2203.06667 (2022)
16. Li, Y., Wang, H., Luo, Y.: A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, pp. 1999–2004. IEEE (2020)
17. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv. Neural. Inf. Process. Syst. **32**, 13–23 (2019)
18. Masuzawa, N., Kitamura, Y., Nakamura, K., Iizuka, S., Simo-Serra, E.: Automatic segmentation, localization, and identification of vertebrae in 3d ct images using cascaded convolutional neural networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 681–690. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_66
19. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the International Conference on 3D Vision, pp. 565–571. IEEE (2016)

20. Nakano, N., et al.: Pre-training methods for creating a language model with embedded knowledge of radiology reports. In: Proceedings of the annual meeting of the Association for Natural Language Processing (2022)
21. Nishie, A., et al.: Current radiologist workload and the shortages in Japan: how many full-time radiologists are required? Jpn. J. Radiol. **33**, 266–272 (2015)
22. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Med. J. (Online) 359 (2017)
23. Seibold, C., et al.: Detailed Annotations of Chest X-Rays via CT Projection for Report Understanding. arXiv preprint arXiv:2210.03416 (2022)
24. Tagawa, Y., et al.: Performance improvement of named entity recognition on noisy data using teacher-student training. In: Proceedings of the annual meeting of the Association for Natural Language Processing (2022)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: TieNet: text-image embedding network for common thorax disease classification and reporting in Chest X-rays. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9049–9058 (2018)
26. Yan, K., Wang, X., Lu, L., Summers, R.M.: DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J. Med. Imaging **5**(3), 036501–036501 (2018)
27. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4683–4693 (2019)
28. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: AlignTransformer: hierarchical alignment of visual regions and disease tags for medical report generation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 72–82. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_7