SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

ConTEXTual Net 3D: Visual Grounding in PET/CT for Enhanced Interactive Reporting

Zachary Huemann, MS, MA, University of Wisconsin – Madison; Samuel Church, MS; Joshua D. Warner, MD; Daniel Tran; Xin Tie, MS; Junjie Hu, PhD; Steve Y. Cho, MD; Meghan G. Lubner, MD; Tyler J. Bradshaw, PhD

Introduction/Background

Visual grounding algorithms, which link text descriptions to specific image regions, have many potential applications in radiology. However, these algorithms require large training datasets of annotated image-text pairs, which currently do not exist for most imaging modalities. We developed a pipeline to extract reported descriptions of salient PET/CT findings and to automatically segment the corresponding image findings. We then applied this pipeline to generate a large annotated dataset for training a 3D vision-language visual grounding model, enabling interactive PET/CT reports.

Methods/Intervention

Our multi-step pipeline operates on PET/CT images and corresponding radiology reports, uses a series of large language models (LLMs) to extract text descriptions of PET findings, and then automatically segments the findings in the image based on the reported slice number and maximum standardized uptake value (SUVmax). Starting with 25,000 PET/CT exams retrospectively collected from 2010 to 2023, the final training/validation/test set consisted of 11,356 sentence-label pairs from 5,094 PET/CT exams. This dataset was used to train a novel 3D vision-language model adapted from ConTEXTual Net, which uses the sentence description, encodes it through an LLM, and fuses the text encodings with a 3D segmentation nn-UNet via cross-attention. The model was then evaluated on a holdout test set of 256 cases reviewed by a board-certified radiologist.

Results/Outcome

The automatic labeling pipeline's accuracy was 98% (251/256). ConTEXTual Net 3D achieved an F1-score of 0.78 on the holdout test set, with a sensitivity of 0.75 and a recall of 0.81. The model performed similarly on 18F-fluorodeoxyglucose (FDG) PET/CT exams (F1=0.78) and on non-FDG PET/CT exams (F1=0.76).

Conclusion

The proposed labeling pipeline demonstrated high accuracy in creating large annotated datasets of image-text pairs for PET/CT, allowing for the development of 3D visual grounding models.

Statement of Impact

Our method can be used to generate the necessary image-text training data to train a visual grounding model to segment key lesions in PET/CT. It opens the door to interactive reports that improve patient and provider comprehension and may allow for retrospective quantitative PET studies.

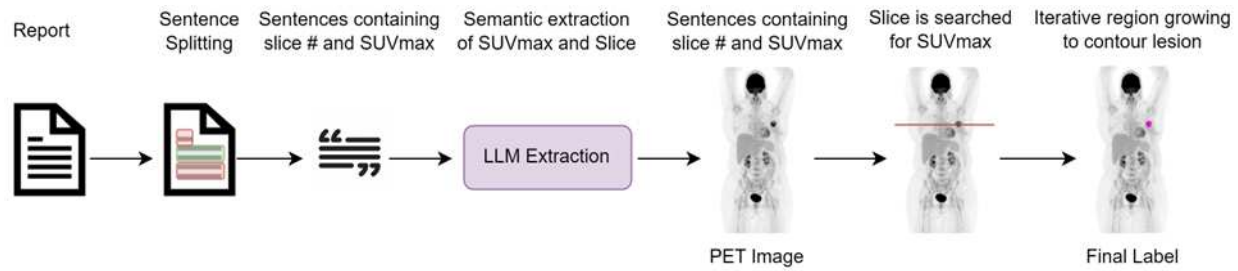

Figure 1. The data preprocessing pipeline is shown. First, we split the PET reports into individual sentences. We extract all sentences that contain a slice number and an SUVmax. Of those, we check which ones contain anatomical descriptor terms using RadGraph. We then used LLMs and in-context learning to filter out sentences describing prior imaging and sentences containing multiple findings. For imaging annotations, the reported slice number is searched for the specified SUVmax. If the SUVmax is found, we use an iterative region-growing method to create the label. This results in a training dataset of PET/CT images, descriptive sentences, and referring segmentations for developing a visual grounding model.

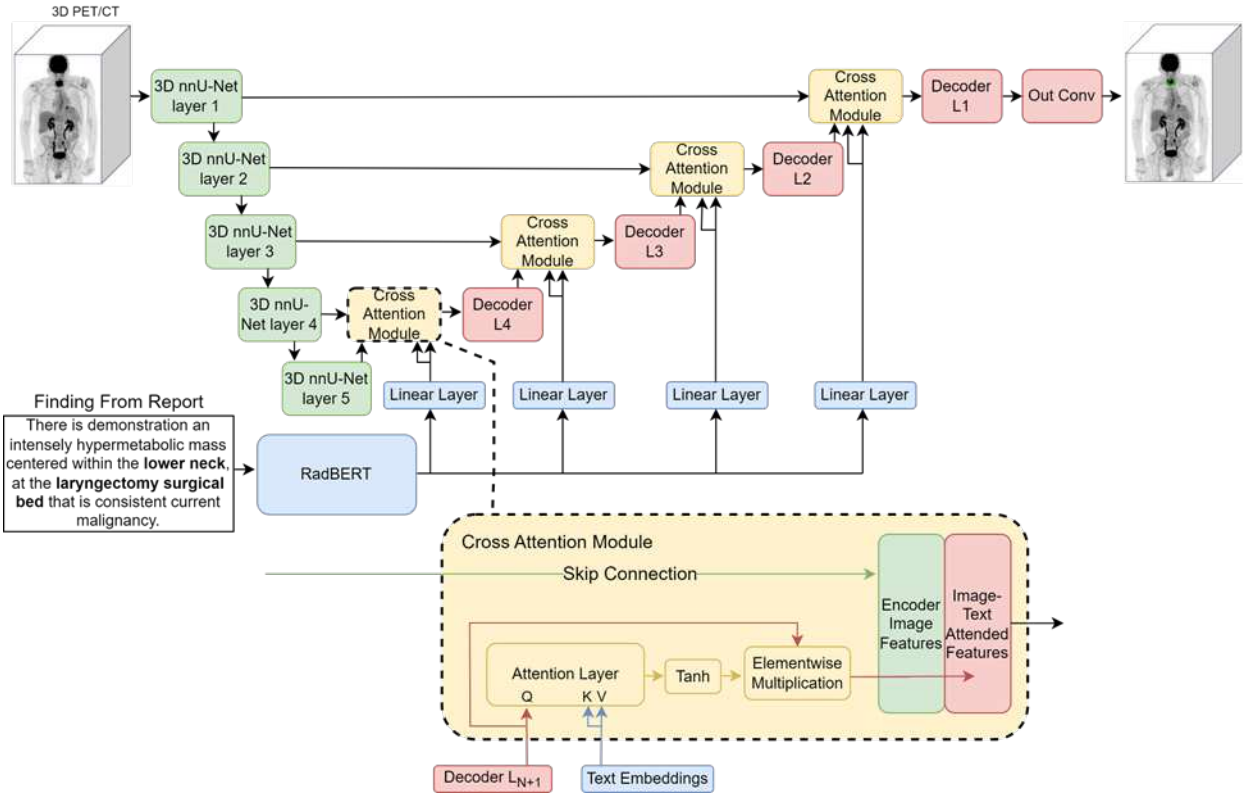

Figure 2. The 3D multimodal vision-language model used for visual grounding of PET findings. The 3D PET/CT is encoded via a 3D nn-Unet and the sentence is encoded via RadBERT. For cross-attention, the language embeddings are used as the key and the value with the vision features as the query, which produces pixel-wise attention maps which are then applied to the pixel space. This mechanism allows for text-guided segmentation.

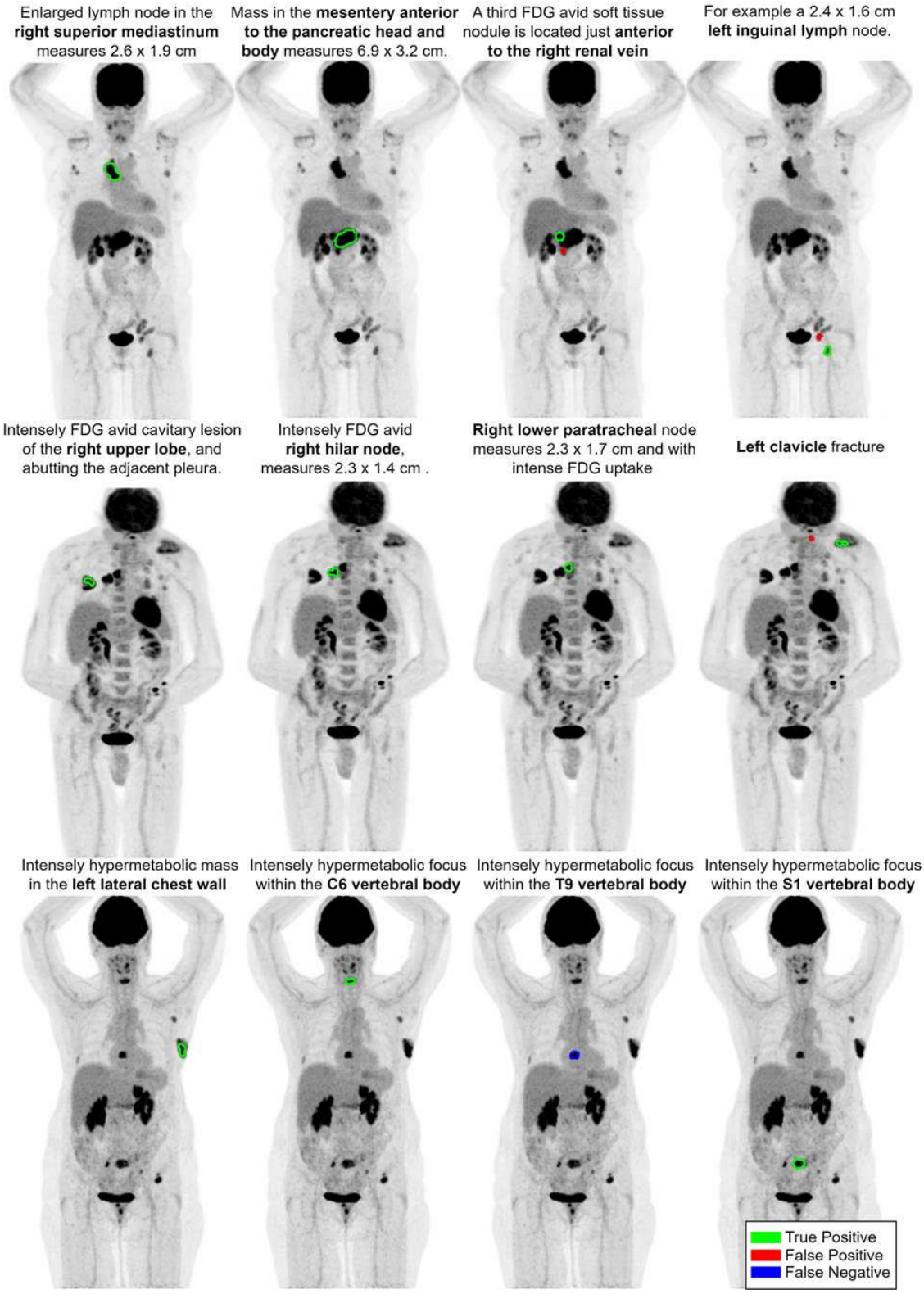

Figure 3. Example images and findings overlaid with the model outputs. True positives are shown in green, false positives are shown in red, and false negatives are shown in blue.

Keywords

Multimodal; Vision-Language Models; PET/CT; Large Language Models; Visual Grounding; Segmentation