On the Automatic Generation of FDG-PET-CT Reports

E. Sabri Eyuboglu *
Department of Computer Science
Stanford University
Stanord, CA 94305
eyuboglu@stanford.edu

Abstract

Labeling is major bottleneck in radiology machine learning workflows. We present a novel approach for training radiology scan encoders without labels. Specifically, we train a dual scan/report encoder in two new tasks (1) radiology masked language modeling and (2) report/scan mismatch detection. FDG PET/CT is a powerful medical imaging modality with applications in the diagnosis and staging of cancer. We develop our model and use PET/CT as a case study for its utility.

1 Introduction

Recent work has demonstrated that convolutional neural networks (CNNs) can learn to interpret radiological scans from wide range of imaging modalities. In some cases, CNNs exhibit radiologist-level accuracy in diagnosing and segmenting scans. These models are usually trained using large datasets labeled with well-structured labels. For example, CheXNet, a CNN trained to interpret chest x-rays, was trained on a dataset of over 100k images each labeled with up to 14 thoracic pathologies.

Acquiring well-structured labels is a significant bottleneck in machine learning for radiology. Even if hospitals are able to pull a large numbers of exams, they rarely come with labels that a traditional computer vision classifier could train on. Usually, physicians manually label each exam, either by reading the accompanying report or by re-interpreting the scan. The process is painstaking and time-intensive, which limits the size and number of datasets that we can train models on.

A medical imaging report is the main line of communication between a radiologist interpreting an exam and other physicians working with the patient. They are often the only record of an exams findings, besides the images themselves. This begs the question: could we bypass the labeling process and train computer vision directly on the report itself? Here, we propose a novel approach that leverages recent advances in NLP modeling and pre-training.

Imaging reports are comprised of several paragraphs of natural language describing the exam. Each report typically includes background information on the patient, the motivation for the exam, additional descriptions of the study procedure, and most importantly a discussion of abnormal findings in the image. Furthermore, reports make use of highly specialized vocabulary to communicate findings and conditions. Their length, heterogeneity and jargon make them difficult to model.

FDG PET/CT is a powerful medical imaging modality with applications in the diagnosis and staging of cancer. The scan is actually a combination of two different imaging modalities: positron emission tomography and computerized tomography. Together they provide a detailed view of metabolic activity in the body and can capture biochemical and physiologic phenomena that CT alone cannot [1].

*External Collaborators: Geoffrey Angus (MS Student in Computer Science), Dr. Matt Lungren (Radiologist Stanford Medical Center), Dr. Bhavik Patel (Radiologist Stanford Medical Center), Dr. Guido Davidzon (Radiologist Stanford Medical Center)

Tumors in the lymphatic system, lungs, head, neck and abdomen can be identified with high sensitivity in PET/CT, making it a critical imagining tool in clinical oncology. The PET/CT scan, which is often administered over the full-length of the body, is a detailed 3D-volume mapping physiologic phenomena in many organs. Identifying tumors in this large and heterogeneous imaging modality requires a deep understanding of human anatomy and the nuances of the metabolic expression of tumors. The pathologies captured by PET/CT are diverse, making it an ideal case study for our approach.

2 Existing Work

2.1 Dataset

We will be using a dataset of $8,251$ PET/CT exams from Stanford Hospital administered between 2004 and 2011. One exam includes a *normal/abnormal* label, the report text written by a radiologist at the time of the study, and sequence of 200 slices each consisting of one PET and one CT image each. The sequence of images spans from the top of the head to the mid-thighs. A *normal* label indicates that the radiologist found no abnormalities anywhere in the scan. An *abnormal* label indicates that there was at at least one abnormal finding. This implies that many slices in an *abnormal* exam will show no abnormality at all. Moreover, because PET/CT scans are usually ordered to stage or track known cancers, the distribution of labels in our dataset is highly unbalanced: only 1 in 9 exams is normal.

A report contains 574 words on average with a standard deviation of 169. Medical imaging reports are usually divided into sections. The "impression" is one important section included in most reports that synthesizes and summarizes the report, much like a scientific abstract does. In our dataset, 99.96% of reports include an impression section. On average, each impression is 53 words long with a standard deviation of 44.95. The large standard deviation can be explained by the many normal exams, whose impressions are usually short sentences to the effect of "no evidence of metabolically active disease." Impression sometimes contain multiple sentences. In our dataset, impressions are 3.34 sentences on average, though over 3,000 contain just one. Furthermore, the impression is often broken down into enumerated subsections.

3 Problem Formulation

Our objective is to to train a computer vision model that is can learn to extract salient features from a radiological scans, without well-structured labels. We will refer to the portion of our model that extracts features from the scan as the *scan encoder*.

In order to do so, we need to train our model on well-defined tasks that we believe will require learning an effective scan encoder.

3.1 Tasks

Radiology Masked Language Modeling In *radiology masked language modeling*, the model is given a scan and portion of the reports impression with some *key* words masked out. The model is tasked with recovering the masked word, using both the scan and the word's context to inform its decision. We take inspiration from the masked language modeling task used to pre-train BERT [2]. However, unlike Devlin *et al.*, we don't mask words at random, but rather mask words that are indicative of the pathology underlying the exam. Say we extract the sentence "FDG uptake in the inguinal lymph nodes." We might, for example, mask the word "inguinal", since it, more than any other word in the sentence, indicates the underlying pathology.

This approach introduces a challenge not present in traditional masked language modeling: for each exam we must identify words, which words are most indicative of pathology. This is especially challenging because indicative words are largely modality specific, (e.g. the words used to highlight an abnormality in a chest x-ray are likely very different from those used to specify the pathology in a PET/CT scan).

To tackle this problem in PET/CT, we first performed a statistical analysis of the report texts in our training dataset. This gave a us an empirical sense of which words and n-grams were most salient in

PET/CT REPORTS. Next, together with radiologists from Stanford's department of nuclear medicine, we crafted a hierarchy of pathologies common in PET/CT scans. For each term, we wrote a series of regular expressions that match terms related to the pathology. To generate training examples, we recursively traverse the hierarchy and search for regular expression matches in the report. When a match is found, with probability $p = k * 0.80$ we mask the word, with probability $p = k * 0.2$ we don't mask the word, but still include it as a label, and with probability $p = 1 - k$, we do nothing at all. The hyperparameter k is free to vary across the pathology hierarchy. This is important, since some pathologies are far more common than others.

We hypothesize that models trained in radiology masked language modeling will have strong scan encoders. Radiology masked language modeling is in some ways more challenging than traditional masked language modeling, because words indicative of pathology are, by design, often quite interchangeable in their lexical contexts. In our example above, "FDG uptake in the inguinal lymph nodes" could just as easily have been "FDG uptake in the supraclavicular lymph nodes". The only way the model could discriminate between these two cases is by effectively interpreting the scan and incorporating the interpretation in lexical prediction.

Report/Exam Mismatch Detection Like Devlin *et al.*, we also include a classification task. Next sentence prediction is of little appeal when it comes to learning effective scan encoders. After all, whether or not one sentence likely comes after another is mostly independent of the scan itself. Instead, we introduce a new task, which we call *report/exam mismatch detection*. Here, with some probability $p = q$, we match an exam with a random report sampled from our dataset. The model must then predict whether or not the scan and the report it received as input are mismatched. Note that if we do sample a random report, we should not want to perform radiology masked language modeling, since the report is mismatched with the scan.

3.2 Input

The model accepts as input both the PET/CT scan and a sentence randomly sampled from the report's impression.

1. A PET/CT scan, which combination of two imaging modalities: positron emission tomography and computerized tomography. Together we have two sequences of l images. We can thus represent the scan as a tensor $X \in \mathbb{R}^{2,h,w,l}$, where h and w are the height and width of the images, respectively.

2. A randomly selected sentence from the scan's report. The report is tokenized using word-piece tokenization [2]. In radiology masked language modeling, some of the tokens are replaced with the special token $[MASK]$. In report/mismatch prediction, the report and scan may not correspond.

4 Model

4.1 Tokenization and Vocabulary

Tokenizers trained on general natural language are ineffective on radiology datasets. Words that are very common in PET/CT reports like "hypermetabolic", are often split into many word pieces. We trained a PET/CT specific word-piece tokenizer on our training using Google's SentencePiece library [3].

Furthermore, we extended the vocabulary used by Devlin *et al.* with 3,000 word pieces common in PET-CT reports.

4.2 Pre-training

We use **BERT** base from Devlin *et al.* pre-trained on BookCorpus and English Wikipedia. We perform domain-specific fine-tuning with masked language modeling and abnormality detection on our training dataset of PET/CT reports.

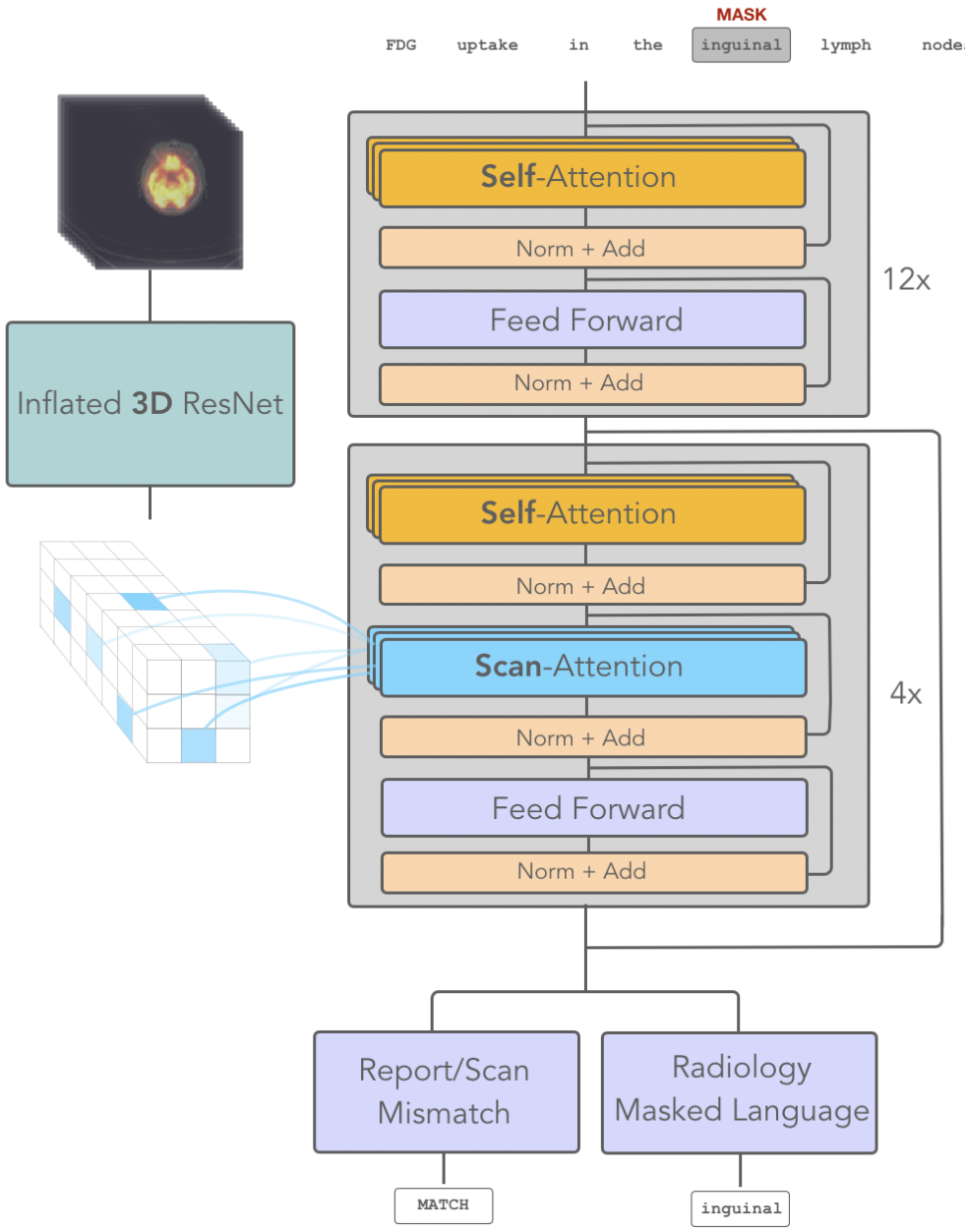

Figure 1: A high-level overview of model architecture.

4.3 Architecture

Scan Encoder We use an inflated ResNet to encode a set of C feature vectors $\mathbf{A} = \{\mathbf{a_1}, \mathbf{a_2}, ...\mathbf{a_C}\}$ from the input scan [4]. Each feature vector $\mathbf{a}_i \in \mathbb{R}^d$ is a d-dimensional embedding corresponding to a 3-dimensional section of the input scan. If one were to divide the input scan into C equal sized cubes (like in figure (1) below), each feature vector would correspond roughly to one of those sections. We use an encoding of size $d = 1024$.

Report Encoder To encode a representation of the masked impression sentence we use **BERT** base from Devlin *et al.* pre-trained on BookCorpus and English Wikipedia. We also perform domain-specific fine-tuning with masked language modeling and abnormality detection on our training

dataset of PET/CT reports. Specifically, we use 12 transformer blocks, each with self-attention, a feed-forward network, batch normalization and residual connections.

Report/Scan Decoder Our report/scan decoder integrates the scan and report encodings into one representation which is used by task specific classification heads. Our report scan decoder is a stack of four modified transformer encoder blocks. First self-attention is applied to the hidden state representation of the sentence from the preceding transformer block. Next, we apply scan-attention on the 3D representation of the scan output by the inflated ResNet. Scan-attention allows the transformer block to attend to specific regions of the scan. Specifically, in each head we generate (1) a query vector $\mathbf{q}$ for each of the report hidden states, (2) a key vector $\mathbf{k}$ for each feature vector $\mathbf{a}_i$ in the scan encoding, (3) and also a value vector $\mathbf{v}$ for each feature vector in the scan. We compute attention scores and aggregate the value vectors as is suggested by the original authors of the transformer paper [5]. The output of the report/scan decoder is added to a residual connection from the scan encoder before being fed to the task heads.

Task Heads We use Devlin *et al.*'s MLM task head for radiology masked language modeling and Devlin *et al.*'s NSP task head for report/scan mismatch detection.

Loss and Training We train the model to simultaneously perform radiology masked language modeling and report/scan mismatch detection. Our loss function is a sum average of the cross entropy loss on the two tasks.

We use Adam and a pre-training learning rate of 1e-4 and a fine-tuning learning rate of 5e-3.

Implementation Our implementation is available in the private repository found at https://github.com/seyuboglu/fdg-pet-ct. I implemented all of the NLP functionality, and most of the general training infrastructure and image model. To compare my contributions to those of my collaborators, please refer to the commit history.

5 Results

5.1 Pre-training

We performed pre-training on traditional masked language modeling. We achieve a peak masked language modeling accuracy of **0.456**.

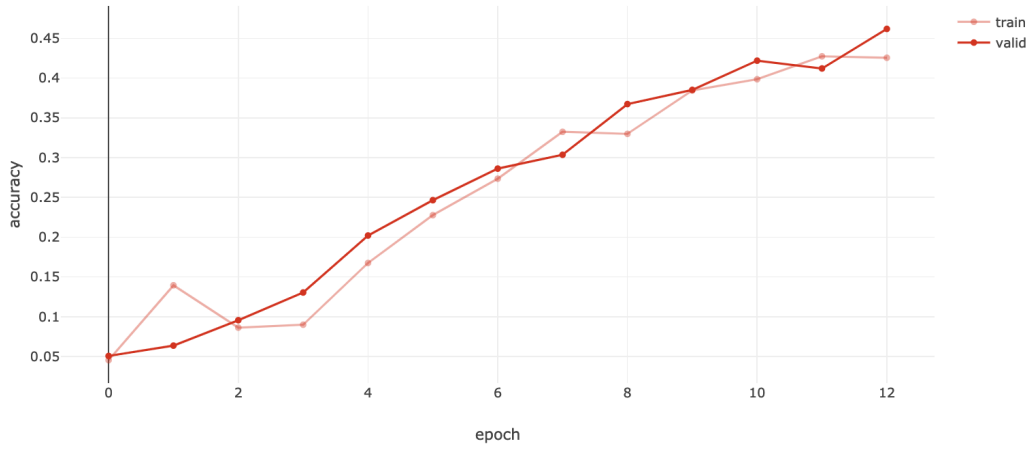

Figure 2: Learning curve for masked language modeling.

5.2 Radiology Masked Language Modeling

Radiology masked language modeling is a distinct task from traditional masked language modeling. Here we only mask words indicative of the pathology underlying the exam. We achieve a peak accuracy of **0.443**.

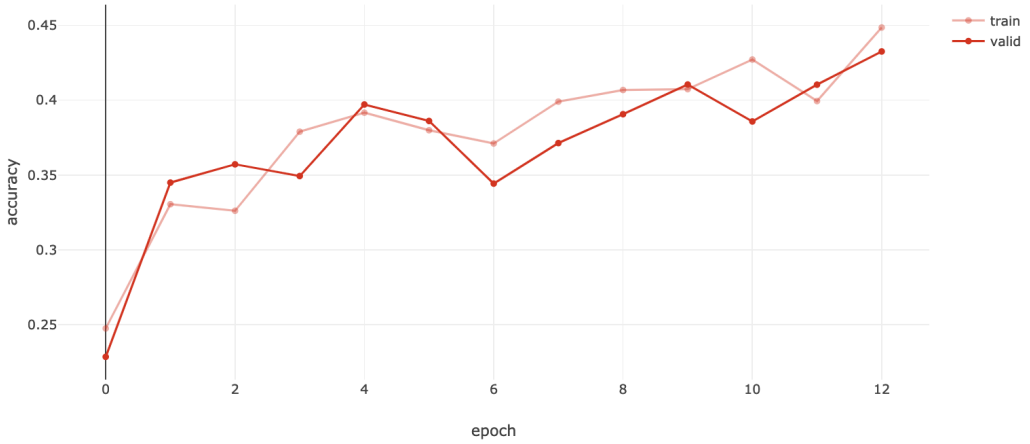

Figure 3: Learning curve for radiology masked language modeling

5.3 Scan/Report Mismatch Detection

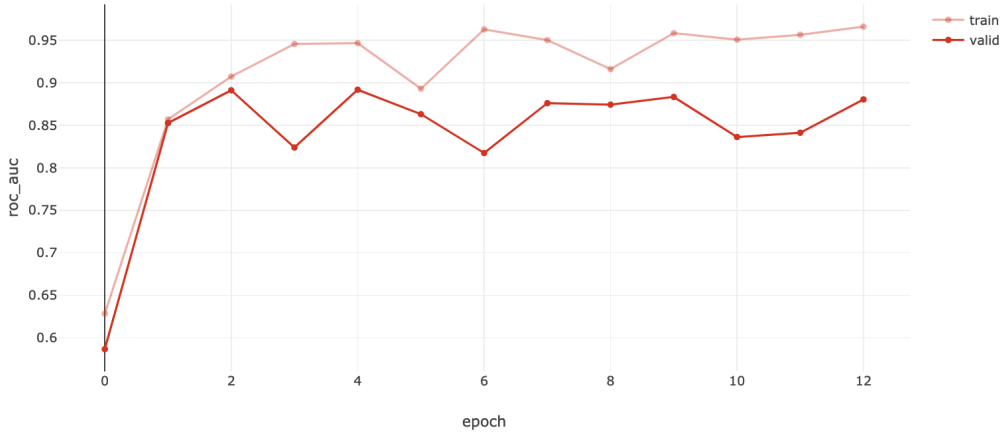

Figure 4: Learning curve for scan/report mismatch detection

We achieve a peak AUROC of **0.891** on this task.

Note we did not evaluate on our test set because this is an ongoing research project.

6 Future Work

Significant and exciting future work remains on this project. Most importantly, the approach must be better validated.

Furthermore, we will assess the performance of the model in zero-shot learning in classification tasks by feeding the model sentences like "FDG uptake in the [MASK]".

This approach presents a promising new way to learn effective scan encodings without labels.

References

[1] Landis K. Griffeth. Use of PET/CT scanning in cancer patients: technical and practical considerations. *Proceedings (Baylor University. Medical Center)*, 18(4):321–330, October 2005.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, October 2018. arXiv: 1810.04805.

[3] Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *arXiv:1804.10959 [cs]*, April 2018. arXiv: 1804.10959.

[4] João Carreira, Andrew Zisserman, Zisserman@google Com, and † Deepmind. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Technical report.

[5] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. Technical report.