Using BERT Models to Label Radiology Reports

John R. Zech, MD, MA

John R. Zech, MD, MA, is a radiology resident at New York Presbyterian–Columbia and incoming musculoskeletal radiology fellow at NYU. He previously served on the *Radiology: Artificial Intelligence* trainee editorial board. He is currently pursuing a Radiological Society of North America Resident Research Grant–funded project to develop a deep learning–based tool to assist radiologists in detecting pediatric upper extremity fractures.

Generating accurate labels for radiology datasets remains a primary barrier to artificial intelligence (AI) model development. Without labels, the large amount of historical imaging data stored in radiology information systems sits idle and cannot be harnessed for AI research. Many creative attempts have been made by researchers to obtain labels for their data, including the U.S. National Cancer Institute's "Crowds Cure Cancer" exhibit at prior Radiological Society of North America annual meetings, which invited radiologists to manually annotate studies to help further oncologic imaging research. Nevertheless, having experts manually label studies is time intensive and limits the amount of imaging data that can be used in a model. Automated methods to infer labels from radiology reports are a promising avenue toward generating labeled data relatively inexpensively.

Popular, freely available chest radiograph datasets such as ChestXray14, CheXpert, and MIMIC-CXR have been labeled using complicated "rule-based" labeling tools (1,2). Broadly, these rule-based tools attempt to parse sentences grammatically and use specific researcher-defined language patterns to infer when a term is positively or negatively mentioned. Such methods have been used for decades on medical text but are time intensive to design and may not perform as intended. For example, researchers have criticized the original ChestXray14 dataset labels generated by a rule-based process and have demonstrated that their validated accuracy is substantially lower than originally reported (3).

Researchers have investigated how machine learning methods could be applied to this problem to improve labeling performance (4). Much of this work predates exciting recent developments in AI for natural language processing (NLP), which has been catalyzed by a transformer network architecture (5). Previously, researchers used recurrent neural network architectures to analyze text data using deep learning. Transformers improved on these models in multiple ways, notably by changing the way that word order is encoded (4). These changes allowed larger datasets to be used in such models and have driven a wave of effort to develop new model architectures based on the transformer concept. One of these is the bidirectional encoder representations from transformers (BERT) model used in the article by Tejani et al (6) in this issue of *Radiology: Artificial Intelligence*. Pretrained BERT models are well-suited to being retrained ("fine-tuned") on small labeled natural language datasets and offer a promising method for labeling radiology reports.

Tejani et al (6) have evaluated how five variants of the BERT model performed in labeling chest radiograph reports for the presence of four support devices: endotracheal tube, enteric tube, central venous catheter, and Swan-Ganz catheter. The authors labeled 1004 reports manually and performed several experiments. First, they trained each model on 60% of their labeled data (approximately 600 reports), validated on 20% (approximately 200 reports), and reported performance on 20% held-out test data (approximately 200 reports). They repeated this process five times with different data splits to make test predictions for all 1004 labeled reports. The models demonstrated strong classification performance, with area under the receiver operating characteristic curve (AUC) values between 0.936 and 0.996 on the various line detection tasks. Second, the authors focused on one of these data splits and evaluated performance as they gradually reduced the number of training cases available from 603 cases to 406, 201, 149, 101, and finally only 48 reports. Expectedly, performance degraded as fewer examples were available, but the most modern model architectures (DeBERTa, DistilBERT, RoBERTa) consistently outperformed older models (BERT, PubMedBERT) on smaller datasets. The best-performing model on small datasets, DeBERTa, achieved AUCs between 0.845 and 0.925 on the smallest (n = 48) training dataset and AUCs of 0.899–0.964 on the training dataset of 101 reports. Finally, the authors reported the average time to train each model and infer labels on a larger collection of 69 095 reports.

Strengths of this study include, first, the strong classification performance, particularly when training data were limited. Given the cost of expert labeling, data efficiency is of primary importance. Additionally, most real-world datasets are imbalanced, with many negative cases and relatively few positive cases, and it is important to maximize yield from available positive examples. Second, models were trained and offered predictions quickly, despite using modest hardware (Nvidia 1080ti GPUs). Third, a variety of transformer-based models were compared. Results demonstrated strong overall performance for all models when trained with sufficient training data, as well as a performance advantage for the most modern model architectures when training data were limited.

From the Department of Radiology, New York Presbyterian–Columbia University Irving Medical Center, 622 W 168th St, New York, NY 10032. Received June 26, 2022; revision requested June 28; revision received July 2; accepted July 7. **Address correspondence to** the author (email: *jrz2111@columbia.edu*).

Author declared no funding for this work. See also the article by Tejani et al in this issue. Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2022; 4(4):e220124 • https://doi.org/10.1148/ryai.220124 • Content codes: AI CH • ©RSNA, 2022

Several limitations of the study should be noted. First, this study focused on labeling chest radiograph reports only for the presence or absence of specific support devices. How might these different BERT models have comparatively performed in labeling these reports for a wider variety of clinically important findings, such as the National Institutes of Health ChestXray14 labels, which include diagnoses such as pneumonia, atelectasis, and cardiomegaly (1)?

Second, no simpler benchmark model is offered as a comparison to these BERT models. Although transformer-based natural language models clearly represent state-of-the-art technology, computationally simpler methods—such as logistic regression applied to a term frequency–inverse document frequency matrix—have demonstrated strong performance in chest radiograph report classification (7).

Third, a challenge in using pretrained deep learning NLP models is that certain words may not be in the pretrained dictionary, such as, in the case of the models used by Tejani et al, "Swan-Ganz." There are various practical ways to deal with this limitation. The authors deconstructed the word into fragments that existed in the pretrained vocabulary, and the algorithm had to learn to associate these various fragments with the finding. The challenge of learning the relationship between the finding and multiple semantically unrelated fragments ("sw," "an," "gan," "z") may partially explain the comparatively weaker performance in detecting Swan-Ganz catheters compared with other devices.

Tejani et al make several contributions compared with existing work; the most relevant comparison work to this study is CheXbert, which trained modified versions of the original BERT model on data from CheXpert and MIMIC-CXR datasets (8) . First, they evaluate how these models perform with small training datasets (CheXbert had 1687 labeled reports and used 75% for training). Second, they demonstrate how more modern refinements of the BERT architecture can outperform the original model when data are scarce. Third, they identify specific support devices, in comparison to the catchall "support device" category used in these other datasets that grouped all these devices into a single category.

A fundamental challenge in automatically extracting labels from reports is the wide variety of language used by different radiologists and different institutional templates to communicate the same findings. Prior work has demonstrated degraded external performance of deep learning models trained to extract findings (7) and to correct typographical errors (9) in chest radiograph reports. Information extracted from images themselves using deep learning can be combined with information from radiology reports to minimize labeling error. Ultimately, some degree of randomized manual review of automatically generated deep learning labels is critical to assure quality in deployment.

With a sufficiently large dataset, the superior contextual awareness of BERT models should prove to be an advantage in identifying when a concept is part of a negative expression ("negation detection"), a constant challenge in this work. For example, the phrase "there has been interval removal of an endotracheal tube" indicates absence of the device, and a model needs to use context to infer that this device is absent despite being mentioned. While the authors note in their review of model errors that models did not always recognize the word "removal," given sufficient training data, this class of models should have a greater ability to detect negation because of their strong contextual awareness (4). The superior contextual awareness of these models could help avoid the complex and brittle rule-based efforts often currently made to detect negation.

Binary labels generated through this type of process can be used directly only to train weakly supervised models that classify images as positive or negative for pathologic findings overall, rather than strongly supervised ones that specifically localize pathologic features (10). These weakly supervised models are challenging for radiologists to interpret (10). If researchers' ultimate goal is the localization of important findings for radiologist consideration, they may wish to include object localization in associated imaging directly into their labeling process.

In conclusion, Tejani et al demonstrate that pretrained BERT deep learning models achieved strong performance in labeling chest radiograph reports for the presence of specific lines, with the most recent model variants performing impressively even on very small training datasets. These models can be trained quickly on modest hardware, and they can rapidly label tens of thousands of reports. The use of these automated methods could avoid the use of complex and brittle rule-based systems commonly used to label reports currently. A particularly exciting feature of these models is their superior ability to detect linguistic context. This ability could facilitate reliable automatic negation detection, an ongoing challenge in this work. Direct comparison of these state-of-the-art methods with simpler machine learning methods would be informative. Given the varied language used by different radiologists and institutions to describe the same findings, it may prove challenging to achieve strong automated labeling performance on reports from varied external institutions. Binary labels generated through such models can be used directly only to train weakly supervised imaging models, which are challenging for radiologists to interpret. Careful expert review is ultimately needed to ensure the accuracy of these pipelines in deployment.

Disclosures of conflicts of interest: J.R.Z. Former member of *Radiology: Artificial Intelligence* trainee editorial board.

References

1. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv:1705.02315 [preprint] http://arxiv.org/abs/1705.02315. Posted May 5, 2017. Accessed June 23, 2022.
2. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031 [preprint] http://arxiv.org/abs/1901.07031. Posted January 21, 2019. Accessed June 23, 2022.
3. Oakden-Rayner L. Exploring large-scale public medical image datasets. Acad Radiol 2020;27(1):106–112.
4. Mozayan A, Fabbri AR, Maneevese M, Tocino I, Chheang S. Practical guide to natural language processing for radiology. RadioGraphics 2021;41(5):1446–1453.
5. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv:1706.03762 [preprint] http://arxiv.org/abs/1706.03762. Posted June 12, 2017. Accessed June 23, 2022.
6. Tejani AS, Ng YS, Xi Y, Fielding JR, Browning TG, Rayan JC. Performance of multiple pre-trained BERT models to automate and accelerate data annotation for large datasets. Radiol Artif Intell 2022;4(4):e220007.

7. Drozdov I, Forbes D, Szubert B, Hall M, Carlin C, Lowe DJ. Supervised and unsupervised language modelling in Chest X-Ray radiological reports. PLoS One 2020;15(3):e0229963.
8. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arXiv:2004.09167 [preprint] http://arxiv.org/abs/2004.09167. Posted April 20, 2020. Accessed June 23, 2022.
9. Zech J, Forde J, Titano JJ, Kaji D, Costa A, Oermann EK. Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models. Ann Transl Med 2019;7(11):233.
10. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol Artif Intell 2021;3(6):e200267.