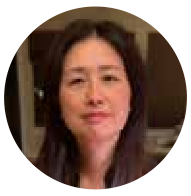

Ying Ding

Ajay Jaiswal

RadBERT-CL: Factually-Aware Contrastive Learning For Radiology Report Classification

Ajay Jaiswal* AJAYJAISWAL@UTEXAS.EDU
Liyan Tang* LYTANG@UTEXAS.EDU
Meheli Ghosh** MEHELIGHOSH69@GMAIL.COM
Justin F Rousseau* JUSTIN.ROUSSEAU@AUSTIN.UTEXAS.EDU
Yifan Peng*** YIP4002@MED.CORNELL.EDU
Ying Ding* YING.DING@ISCHOOL.UTEXAS.EDU
**The University of Texas at Austin, United States*
***Central University of Gujarat, India*
****Weill Cornell Medicine, United States*

Abstract

Radiology reports are unstructured and contain the imaging findings and corresponding diagnoses transcribed by radiologists which include clinical facts and negated and/or uncertain statements. Extracting pathologic findings and diagnoses from radiology reports is important for quality control, population health, and monitoring of disease progress. Existing works, primarily rely either on rule-based systems or transformer-based pre-trained model fine-tuning, but could not take the factual and uncertain information into consideration, and therefore generate false positive outputs. In this work, we introduce three sedulous augmentation techniques which retain factual and critical information while generating augmentations for contrastive learning. We introduce RadBERT-CL, which fuses these information into BlueBert via a self-supervised contrastive loss. Our experiments on MIMIC-CXR show superior performance of RadBERT-CL on fine-tuning for multi-class, multi-label report classification. We illustrate that when few labeled data are available, RadBERT-CL outperforms conventional SOTA transformers (BERT/BlueBert) by significantly larger margins (6-11%). We also show that the representations learned by RadBERT-CL can capture critical medical information in the latent space.

Keywords: Thoracic Disorder, Contrastive Learning, Radiology Reports, Chest-Xray, Classification

1. Introduction

Chest radiography is a critical medical imaging technique used for diagnosis, screening, and treatment of many perilous diseases. Radiology reports are documented by radiologists after examining a patient's medical history and diagnostic imaging, and represent complex anatomical and medical terms written for healthcare providers, along with indications of the presence or absence of any disease. Classifying radiology reports according to their description of abnormal findings is important for quality assurance and can mitigate the risks of diagnostic radiation exposure in children [24]. Additionally, the Precision Medicine Initiative (PMI) initiated by NIH and multiple research centers has highlighted the importance of text mining techniques to enable cohort phenotyping of patients for population health (Shin et al., 2017). Classifying radiology reports can help to identify patient cohorts and enable precision medicine on a large scale. Labeling radiology reports with disease types can also assist in the development of deep learning applications for automated-diagnosis (Rajpurkar et al., 2017; Han et al., 2020; Yao et al., 2018).

ChestX-ray14 (Wang et al., 2017), MIMIC-CXR (Johnson et al., 2019), and OpenI (Demner-Fushman et al., 2016) are some of the largest radiology datasets available, and many classification algorithms have been developed based on the training sets provided by these datasets to classify reports into diseases. CheXpert (Irvin et al., 2019) is an automated rule-based labeler consisting of three stages: mention extraction, mention classification, and mention aggregation, to extract observations from the free text radiology re-

Jaiswal, A., Tang, L., Ghosh, M., Rousseau, J.F., Peng, Y. &amp; Ding, Y.. (2021). RadBERT-CL: Factually-Aware Contrastive Learning For Radiology Report Classification. Proceedings of Machine Learning for Health, in Proceedings of Machine Learning Research: 158:196-208