Improving Zero-Shot Detection of Low Prevalence Chest Pathologies using Domain Pre-trained Language Models

Aakash Mishra[*1] AMISHRA@COLLEGE.HARVARD.EDU
Rajat Mittal[*1] RAJATMITTAL@COLLEGE.HARVARD.EDU
Christy Jestin[1] CHRISTYJESTIN@COLLEGE.HARVARD.EDU
Kostas Tingos[1] KTINGOS@COLLEGE.HARVARD.EDU
Pranav Rajpurkar[1] PRANAV_RAJPURKAR@HMS.HARVARD.EDU
[1] *Cambridge, MA*

Editors: Under Review for MIDL 2023

Abstract

Recent advancements in zero-shot learning have enabled the use of paired image-text data to replace structured labels, replacing the need for expert annotated datasets. Domain pre-trained models, such as CXR-BERT, BlueBERT, and ClinicalBERT, offer the potential to improve the performance of CLIP-like models with specific domain knowledge by replacing BERT weights at the cost of breaking the original model's alignment. We evaluate the performance of zero-shot classification models with domain-specific pre-training for detecting low-prevalence pathologies. Even though replacing the weights of the original CLIP-BERT degrades model performance on commonly found pathologies, we show that pre-trained text towers perform exceptionally better on low-prevalence diseases. This motivates future ensemble models with a combination of differently trained language models for maximal performance.

Keywords: Zero-shot, fine-tuning, Pre-training, Multi-modal, self-supervised learning

1. Introduction

Contrastive learning methods have helped create deep learning models with robust zero-shot performance in the field of medical imaging. New innovations in deep learning have enabled the automation of tasks such as medical image interpretation at an expert level (Litjens et al., 2017; Rajpurkar et al., 2017; Tiu et al., 2022). Many of these methods, however, rely on the existence of a large dataset of annotated examples. These annotations often take significant expert time to assign (Smit et al., 2020), and the resulting approaches are limited to only predicting diseases that are explicitly annotated. Recently, the emergence of contrastive learning approaches for multimodal zero-shot learning (Radford et al., 2021) have allowed for paired image-text data to replace structured labels and still achieve competitive performance on many downstream tasks. Many works like Tiu et al. (2022), (Jang et al., 2022), Kumar et al. (2022) and van der Sluijs et al. (2023) achieved sometimes expert-level, zero-shot performance for evaluating and diagnosing chest x-ray images.

While many models such as CheXzero are initialized with weights derived from natural image-caption training, there have been a number of open-source language models that specialize in specific medical domains such as CXR-BERT (Boecking et al., 2022), BlueBERT

[*] Contributed equally

© 2023 CC-BY 4.0, A. Mishra, R. Mittal, C. Jestin, K. Tingos & P. Rajpurkar.

(Peng et al., 2019), and ClinicalBERT (Alsentzer et al., 2019). These present the opportunity to enhance the performance of CLIP-like models with domain knowledge. In particular, domain pre-trained models see a larger medical corpora than is available in chest x-ray reports, and as a result, form richer embeddings for low-prevalence pathologies (infrequently mentioned pathologies in the image-report dataset). The tradeoff is that using a pre-trained text tower breaks the alignment between the vision and language modalities. In this work, we evaluate the performance of zero-shot classification models with domain-specific pre-training and find that they perform well for detecting low-prevalence pathologies.

2. Methods

We utilize contrastive learning with image–text pairs to learn a representation that enables zero-shot multi-label classification. The key idea behind this approach is to have or learn two embedding models: a vision embedder and a text embedder that produce same-sized embeddings that can be assessed for similarity after training via a contrastive loss. In our approach, we do not start our vision and language towers from scratch. Instead, we replace one tower with an existing generalized pretrained model (CLIP) and the other with a domain-specific pretrained model, unlike our baseline, CheXzero (Tiu et al., 2022), which begins training with both towers initialized with aligned CLIP weights. The first of three language domain pretraining methods we leverage is Microsoft's CXR-BERT, a chest x-ray (CXR) domain-specific language model (Boecking et al., 2022). CXR-BERT is first pre-trained from a randomly initialized BERT model via Masked Language Modeling (MLM) on PubMed and MIMIC-III. Then it is further pre-trained on MIMIC-CXR to specialize towards the chest x-ray domain. BlueBERT was pre-trained on PubMed abstracts and MIMIC-III clinical notes and achieves good performance on the Biomedical Language Understanding Evaluation (BLUE) benchmark (Peng et al., 2019). The final domain pre-trained text stack we use is ClinicalBERT "Bio+Clinical BERT" model, initialized from BioBERT and trained on MIMIC-III data (Alsentzer et al., 2019). These domain pre-trained text models do not inherently create a text embedding with the same dimensionality as that of our image stack. As a result, we include a text projection on top of the pre-trained models to complete our text stack, which produce a final embedding of size 128. After training the model on the alignment task for 5 epochs, we evaluate on the downstream chest x-ray pathology classification task. The model is presented with a chest X-ray as well as a negative prompt (e.g. no pneumonia) and a positive prompt (e.g. pneumonia). The similarities between the image and the prompts are calculated and used as logits to produce a probability for having or not having a disease.

Additionally, we retrained the baseline CheXzero model from CLIP's weights since the published model is an ensemble. This new baseline offers a better comparison to our experiments which use single-models. We trained each model for 5 epochs, batch size of 32, learning rate of 5e-6, and SGD optimizer with a momentum of 0.9.

3. Results & Conclusion

Domain pretrained text towers have extremely strong performance on rarely mentioned diseases. We find that pre-trained text towers perform extremely well across categories with little mention in the reports in the MIMIC-CXR alignment training dataset.

Configuration	Lung tumor	Aortic enlargement	Enlarged PA	Clavicle fracture
Occurrences	1 ($\leq$ 1%)	6 ($\leq$ 1%)	38 ($\leq$ 1%)	69 ($\leq$ 1%)
Baseline	0.687 (0.633, 0.725)	0.580 (0.547, 0.619)	0.713 (0.557, 0.838)	0.664 (0.401, 0.848)
BlueBert	0.654 (0.605, 0.699)	0.640 (0.609, 0.670)	0.667 (0.537, 0.805)	0.585 (0.264, 0.809)
ClinicalBert	0.713 (0.665, 0.753)	0.596 (0.567, 0.628)	0.757 (0.580, 0.891)	0.762 (0.445, 0.978)
CXRBert	**0.714 (0.666, 0.751)**	**0.760 (0.737, 0.782)**	**0.854 (0.777, 0.935)**	**0.853 (0.839, 0.872)**
Configuration	Pleural effusion	Pneumonia	Atelectasis	Pneumothorax
Occurrences	78644 (18.6%)	65204 (15.7%)	56377 (13.6%)	45994 (11.1%)
Baseline	0.864 (0.823, 0.897)	**0.849 (0.823, 0.874)**	**0.770 (0.727, 0.814)**	**0.811 (0.762, 0.865)**
BlueBERT	0.857 (0.821, 0.882)	0.540 (0.506, 0.580)	0.732 (0.680, 0.779)	0.689 (0.591, 0.783)
ClinicalBERT	0.773 (0.732, 0.805)	0.821 (0.799, 0.845)	0.639 (0.589, 0.699)	0.590 (0.490, 0.687)
CXRBERT	**0.895 (0.871, 0.918)**	0.819 (0.801, 0.841)	0.675 (0.624, 0.714)	0.683 (0.602, 0.782)

Table 1: Pretrained text tower zero-shot boostrapped AUC performance on low-prevalence pathologies in the MIMIC-CXR training set, evaluated on the VinDr-CXR dataset. Both ClinicalBERT and CXRBert outperform the baseline on the majority of low-prevalence pathologies. The same does not hold for common pathologies. Occurrences are defined as the number of reports in MIMIC-CXR with impressions that contain the name of the pathology. `Format:(Value, (Confidence Interval))`

In particular, VinDr-CXR contains labels for many rarely mentioned diseases, including lung tumor, aortic enlargement, enlarged pulmonary artery, and clavicle fracture which all occur as strings less than 100 times in the more than 300 thousand report impressions in MIMIC-CXR. Table 1 shows that the zero-shot AUC of both CLIP-vision + domain pre-trained (ClinicalBERT and CXRBERT) language models exceed that of the baseline in every category on these rare diseases. However, when we observe the diseases with the highest number of direct mentions in the training set, the opposite trend holds. The baseline outperforms the majority of mixed CLIP-vision + domain pre-trained text models.

This suggests a strong motivation for pre-trained text towers depending on the use-case of a zero-shot classifier. When diseases have little-to-no mention in the training dataset, the baseline is automatically at a disadvantage. At zero-shot classification time, the model will ideally produce a relevant image embedding containing information about the new pathology and the text tower will embed the new pathology to an aligned encoding. However, if text tower has rarely or never been exposed to the new pathology, the possibility of this text embedding being meaningful and aligning with the image embedding is small. Meanwhile, pre-trained text towers that have been trained on large datasets like MIMIC III and PubMed abstracts have likely encountered these low-prevalence pathologies in their masked-language modeling pre-training task. As such, the embeddings produced for these diseases by the pre-trained text tower are likely richer. The surprising result is that even after alignment training where the pathology is rarely mentioned, the embeddings are not lost to gradient information-loss. This suggests the image tower learned to align to these new emebeddings and as a result - the vision tower was able to use the text embedding space to characterize the image of the unseen pathology.

Our research suggests that a zero-shot classifier designed to classify a broad range of both common and low-prevalence chest x-ray pathologies will likely need some form of a domain pretrained text tower. One option is to include pre-training in the text encoder in a way that does not hurt performance on common diseases, a task of future research. The other option is a weighted ensemble that has both a domain pre-trained language model and general BERT model.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323, 2019. URL http://arxiv.org/abs/1904.03323.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing, 2022. URL https://arxiv.org/abs/2204.09817.

Jongseong Jang, Daeun Kyung, Seung Hwan Kim, Honglak Lee, Kyunghoon Bae, and Edward Choi. Significantly improving zero-shot x-ray pathology classification via fine-tuning pre-trained image-text encoders. *arXiv preprint arXiv:2212.07050*, 2022.

Bhawesh Kumar, Anil Palepu, Rudraksh Tuwani, and Andrew Beam. Towards reliable zero shot classification in self-supervised models with conformal prediction. *arXiv preprint arXiv:2210.15805*, 2022.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.07.005. URL https://www.sciencedirect.com/science/article/pii/S1361841517301135.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474, 2019. URL http://arxiv.org/abs/1906.05474.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.org/abs/1711.05225.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *CoRR*, abs/2004.09167, 2020. URL https://arxiv.org/abs/2004.09167.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis Langlotz, Andrew Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 09 2022. doi: 10.1038/s41551-022-00936-9.

Rogier van der Sluijs, Nandita Bhaskhar, Daniel Rubin, Curtis Langlotz, and Akshay Chaudhari. Exploring image augmentations for siamese representation learning with chest x-rays. *arXiv preprint arXiv:2301.12636*, 2023.