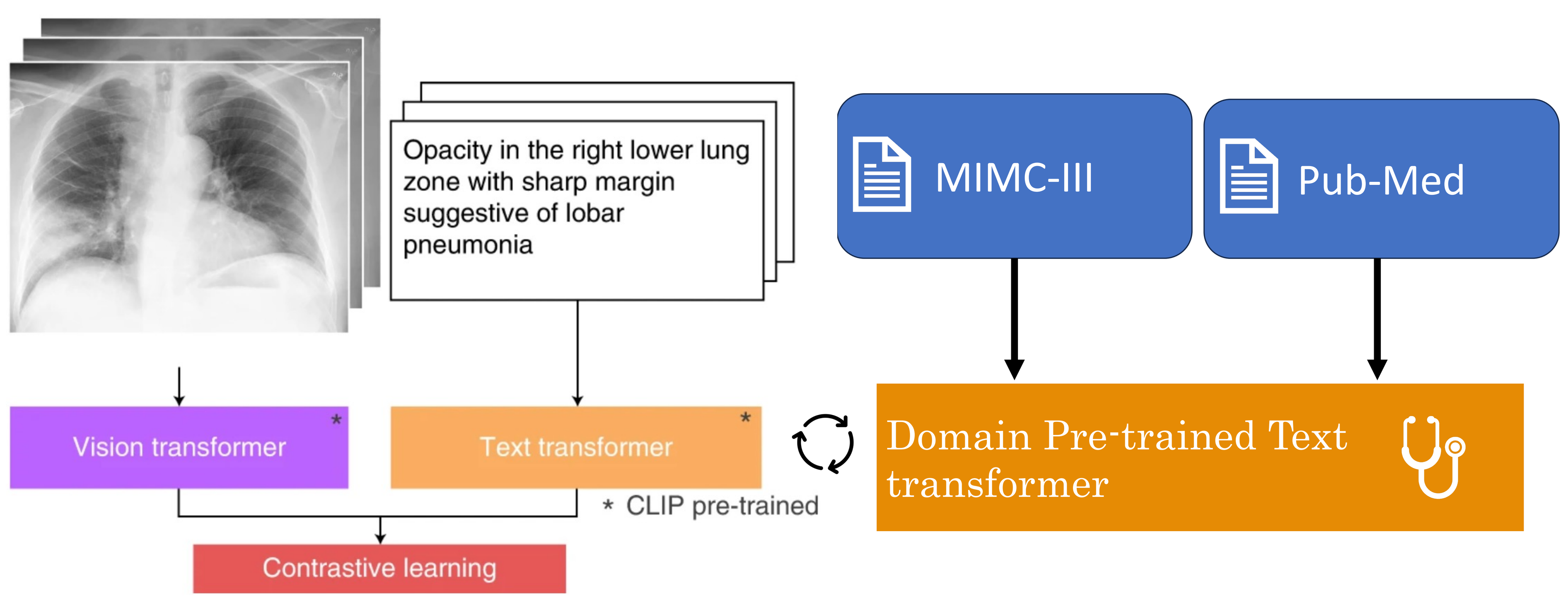
Improving Zero-Shot Detection of Low Prevalence Chest Pathologies using Domain Pre-trained Language Models

Aakash Mishra[1], Rajat Mittal[1], Christy Jestin[1], Kostas Tingos[1], Pranav Rajpurkar[1,2]
Harvard University[1], Harvard Medical School[2]

Abstract

Recent advances in zero-shot learning have enabled the use of paired image-text data to replace structured labels, replacing the need for expert annotated datasets. Models such as CLIP-based CheXzero utilize these advancements in the domain of chest X-ray interpretation. We hypothesize that 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.

Methods



We employ contrastive learning with image–text pairs to achieve zero-shot multi-label classification, utilizing two embedding models: a vision embedder and a text embedder. The embedding models are trained via a contrastive loss and the similarity is assessed. Instead of initializing both towers with CLIP weights like the baseline model, CheXzero (Tiu et al., 2022), we keep the existing vision tower from the generalized pretrained model and the replace the CLIP text tower with a domain-specific language model. We employ three language domain pretraining models: CXR-BERT, BlueBERT, and ClinicalBERT, all of which were pretrained on PubMed and MIMIC-III (Peng et al., 2019; Alsentzer et al., 2019; Boecking et al., 2022). CXR-BERT was further pretrained on MIMIC-CXR to specialize in the chest X-ray domain. Our text stack includes a text projection to complete our pretrained models and produce a final embedding size of 128 since originally the embedding size for each of the three language models was not standardized.

We then trained each model on the contrastive alignment task for 5 epochs using MIMIC-CXR data and then proceeded with a zero-shot evaluation on the chest X-ray pathology classification task. We follow the same zero shot inference strategy as Tiu et al. (2022). Furthermore, we trained a single CheXzero baseline model from CLIP weights to provide a better comparison for our single-model experiments since the published model is an ensemble. All models were trained for 5 epochs with a batch size of 32, a learning rate of 5e-6, and an SGD optimizer with a momentum of 0.9

Results

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 bootstrapped 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 cannot be said for common pathologies. Occurrences are defined as the number of MIMIC-CXR reports with impressions that contain the name of the pathology. Format: AUC values and 95% confidence intervals

Conclusions

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. This motivates the need for pre-trained text towers depending on the use-case of a zero-shot classifier. When diseases have little-to-no mention in the alignment training dataset, the baseline is automatically at a disadvantage. At zero-shot classification time, in the ideal scenario, the model will 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 the text tower has rarely or never been exposed to the new pathology as is the case with CheXzero, which was initialized with CLIP weights, the likelihood of the text embedding being meaningful in reference to the image embedding is small. Meanwhile, pre-trained text towers that have been trained on large datasets like MIMIC III and PubMed abstracts have encountered impressions of these low-prevalence pathologies in their masked-language modeling pre-training task. The surprising result is that even after alignment training where the pathology is rarely mentioned, the text embedding model is still able to perform well on these rarer categories. 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 pretraining 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 pretrained language model and a general BERT model.

References

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis Langlotz, Andrew Ng, and Pranav Ra-jpurkar. Expert-level detection of pathologies from unannotated chest x-ray imagesvia self-supervised learning. Nature Biomedical Engineering, pages 1–8, 09 2022. doi:10.1038/s41551-022-00936-9.4
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.
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
Ekin Tiu, Ellie Talius, Pujan Patel, Curtis Langlotz, Andrew Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray imagesvia self-supervised learning. Nature Biomedical Engineering, pages 1–8, 09 2022. doi:10.1038/s41551-022-00936-9.4