Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing

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‡

Microsoft Health Futures

Abstract. Multi-modal data abounds in biomedicine, such as radiology images and reports. Interpreting this data at scale is essential for improving clinical care and accelerating clinical research. Biomedical text with its complex semantics poses additional challenges in vision–language modelling compared to the general domain, and previous work has used insufficiently adapted models that lack domain-specific language understanding. In this paper, we show that principled textual semantic modelling can substantially improve contrastive learning in self-supervised vision–language processing. We release a language model that achieves state-of-the-art results in radiology natural language inference through its improved vocabulary and novel language pretraining objective leveraging semantics and discourse characteristics in radiology reports. Further, we propose a self-supervised joint vision–language approach with a focus on better text modelling. It establishes new state of the art results on a wide range of publicly available benchmarks, in part by leveraging our new domain-specific language model. We release a new dataset with locally-aligned phrase grounding annotations by radiologists to facilitate the study of complex semantic modelling in biomedical vision–language processing. A broad evaluation, including on this new dataset, shows that our contrastive learning approach, aided by textual-semantic modelling, outperforms prior methods in segmentation tasks, despite only using a global-alignment objective.

Keywords: self-supervision, multi-modal, weak supervision, radiology

1 Introduction

Advances in deep learning have enabled automated diagnosis systems that operate near or above expert-level performance, paving the way for the use of machine learning systems to improve healthcare workflows, for example by supporting fast triaging and assisting medical professionals to reduce errors and omis-

* These authors contributed equally.
† Work conducted during Benedikt Boecking's internship at Microsoft Research.
‡ Corresponding author: ozan.oktay@microsoft.com

sions [9,20,54,72]. A major hurdle to the widespread development of these systems is a requirement for large amounts of detailed ground-truth clinical annotations for supervised training, which are expensive and time-consuming to obtain. Motivated by this challenge, there has been a rising interest in multi-modal self-supervised learning [31,45] and cross-modal weak supervision [19,21,33,72,76] (using partial and imperfect labels derived from the auxiliary modality), in particular for paired image–text data. Such data is collected routinely during clinical practice, and common examples are X-ray images [19,33,76] or computed tomography (CT) scans [9,19,21,72] paired with reports written by medical experts. Importantly, while many remain private, some paired clinical datasets [3,15,34] have been released to the research community such as MIMIC-CXR [34].

This article focuses on self-supervised vision–language processing (VLP) for paired image and text data in the biomedical domain. The goal is to jointly learn good image and text representations that can be leveraged by downstream applications such as zero-/few-shot image classification, report generation and error detection, and disease localisation. Self-supervised VLP has several advantages over supervised learning, not just because it does not require laborious manual annotations, but also because it does not operate on a fixed number of predetermined conditions or object categories, since the joint latent space is learned from raw text. However, in contrast to the general domain setting, self-supervised VLP with biomedical data poses additional challenges. Take radiology as an example, publicly available datasets [3,15,34] are usually smaller, on the order of a few hundred thousand pairs rather than millions in general-domain vision–language processing (e.g. [61] collected 400M text–image pairs on the Internet for self-supervision). Furthermore, linguistic challenges are different in biomedical settings, including common usage of negations, expressions of uncertainty, long-range dependencies, more frequent spatial relations, the use of domain-specific modifiers, as well as scientific terminology rarely found in the general domain. Taking negation as an example, "there is no dog in this picture" would be a highly unusual caption on social media, but "there is no evidence of pneumonia in the left lung" or "there are no new areas of consolidation to suggest the presence of pneumonia" are descriptions commonly found in radiology reports. Moreover, pretrained models including object detectors often used in general domain visual grounding are typically unavailable or under-perform in domain-specific applications (see also Supp. in [31]). Additionally, imbalance in underlying latent entities of interest (e.g., pulmonary findings) can cause larger numbers of false negatives in contrastive learning objectives that sample at random, which can lead models to degrade and memorise irrelevant text and image aspects. For example, radiology images and text reports with normal findings occur much more frequently compared to exams that reveal abnormal conditions such as pneumonia or pneumothorax (also see [11]). Supp. B.1 provides further discussion of these challenges.

Related self-supervised VLP work [30,31,45,56,85] has achieved impressive downstream classification and zero-shot classification performance. However, our study reveals that suboptimal text modelling due to insufficient vocabulary ad-

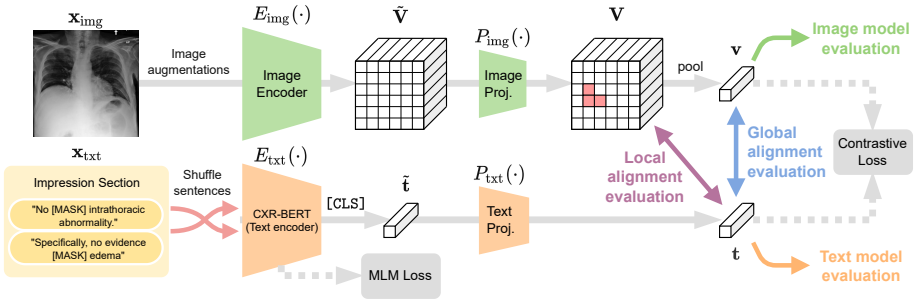


Fig. 1: BioViL leverages our radiology-specific text encoder (CXR-BERT), text augmentation, regularisation, and maintains language model quality via a masked language modelling (MLM) loss. We conduct a broad evaluation of models and representations that includes zero-shot classification, phrase grounding, and natural language inference.

justment, fine-tuning, and language grounding appears to have gone unnoticed, all of which are shown to degrade the quality of joint latent representations. In particular, a more thorough benchmarking of the text, image, and shared embeddings, across a multitude of downstream benchmarks, reveals that large improvements in performance are possible by taking care to build highly specialised text models and by maintaining their performance during joint training. Free-text image descriptions provide a semantically dense learning signal compared to image-only contrastive methods and supervised classification [16]. Further, extracting shared semantics of images and text pairs is easier for text, as the modality is already discretised. Thus, making the most of text modelling before and during joint training can lead to large improvements in not just the text model, but also of the image model and joint representations. We present the following contributions in this work:

1. We introduce and release a new chest X-ray (CXR) domain-specific language model, CXR-BERT[1] (Fig. 2). Through an improved vocabulary, a novel pretraining procedure, regularisation, and text augmentation, the model considerably improves radiology natural language inference [54], radiology masked token prediction [17,48], and downstream VLP task performance.
2. We propose and release a simple but effective self-supervised VLP approach for paired biomedical data, which we name BioViL[1,2] (Fig. 1), and evaluate in the radiology setting. Through improvements in text modelling, text model grounding, augmentation, and regularisation, the approach yields new state-of-the-art performance on a wide range of public downstream benchmarks. Our large-scale evaluation (see Table 2) includes phrase grounding, natural language inference [54], as well as zero-/few-shot classification and zero-shot segmentation via the RSNA Pneumonia dataset [66,76]. Notably, our approach achieves improved segmentation performance despite only using a global alignment objective during training.

[1] Pretrained models available on HuggingFace: https://aka.ms/biovil-models
[2] Code can be found at: https://aka.ms/biovil-code

3. We also release a ***L**ocal **A**lignment **Che**st **X**-ray dataset*, `MS-CXR`[3], to encourage reproducible evaluation of shared latent semantics learned by biomedical image-text models. This large, well-balanced phrase grounding benchmark dataset contains carefully curated image regions annotated with descriptions of eight radiology findings, as verified by board-certified radiologists. Unlike existing chest X-ray benchmarks, this challenging phrase grounding task evaluates joint, local image-text reasoning while requiring real-world language understanding, e.g. to parse domain-specific location references, complex negations, and bias in reporting style.

2 Making the Most of Free-Text Supervision

We assume that we are given a set $\mathcal{D}$ of pairs of radiology images and reports $(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}})$. Let $\mathbf{w} = (w_1, \ldots, w_T)$ denote a vector of T (sub-)word tokens of a text document $\mathbf{x}_{\text{txt}}$ (after tokenisation). Recall that a BERT [73] encoder E_{txt} outputs a feature vector for each input token w_t as well as a special global `[CLS]` token used for downstream classification. Let $\tilde{\mathbf{t}} = [E_{\text{txt}}(\mathbf{w})]_{\texttt{[CLS]}}$ denote the `[CLS]` token prediction by E_{txt} based on input $\mathbf{w}$, and $\mathbf{t} = P_{\text{txt}}(\tilde{\mathbf{t}})$ its lower-dimensional projection by a model P_{txt}.

2.1 CXR-BERT: Domain-Specific Language Model Pretraining

We introduce CXR-BERT (Fig. 2), a specialised chest X-ray (CXR) language model with an adjusted vocabulary, pretrained in three phases to capture dense semantics in radiology reports [4]. To achieve this specialisation to the CXR report domain despite limited data availability, our approach includes pretraining on larger data from closely related domains. The phases proceed as follows: **(I)** First, we construct a custom WordPiece [80] vocabulary of 30k tokens from PubMed abstracts[4] (15 GB), MIMIC-III [35] clinical notes (3.5 GB), and MIMIC-CXR radiology reports (0.1 GB). With this custom vocabulary, our model produces fewer sub-word breakdowns (Table 1). **(II)** Second, we pretrain a randomly initialised BERT model via Masked Language Modelling (MLM) on the PubMed + MIMIC-III + MIMIC-CXR corpora. We largely follow RoBERTa [48] pretraining configurations, i.e. dynamic whole-word masking for MLM and packing of multiple sentences into one input sequence. This phase aims to build an initial domain-specific BERT model in the biomedical and clinical domains. **(III)** Third, we continue pretraining on MIMIC-CXR only to

Table 1: Vocabulary comparison of common radiology terms with ClinicalBERT (Wiki/Book, cased), PubMedBERT (PubMed, uncased), and CXR-BERT (PubMed+MIMIC-III/CXR, uncased). ✓ marks that a word appears in the vocabulary, otherwise its sub-tokens are shown.

Full word	ClinicalBERT	PubMedBERT	CXR-BERT
pneumonia	✓	✓	✓
opacity	op-acity	✓	✓
effusion	e-ff-usion	✓	✓
pneumothorax	p-ne-um-oth-orax	✓	✓
atelectasis	ate-lect-asis	ate-le-ct-asis	✓
cardiomegaly	card-io-me-gal-y	cardio-me-gal-y	✓
bibasilar	bi-bas-ila-r	bib-asi-la-r	✓

[3] The `MS-CXR` dataset can be found on PhysioNet https://aka.ms/ms-cxr.
[4] Obtained via https://pubmed.ncbi.nlm.nih.gov/download/

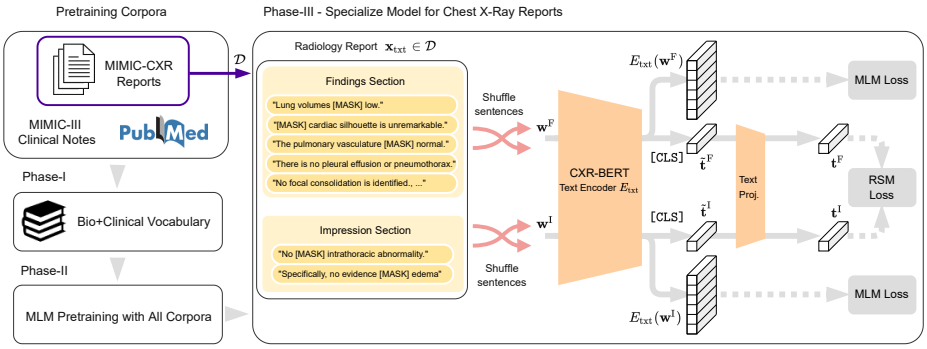


Fig. 2: The proposed CXR-BERT text encoder has three phases of pretraining and uses a domain-specific vocabulary, masked language modelling (MLM) and radiology section matching (RSM) losses, regularisation, and text augmentations.

further specialise our CXR-BERT to the CXR domain. Here, we also add a novel sequence prediction task to the objective to obtain better sequence representations, as explained below.

Note that a raw radiology report $\mathbf{x}_{\text{txt}}$ typically consists of several sections, including a 'FINDINGS' section that details clinical observations, and an 'IMPRESSION' section summarising the clinical assessment [74,77]. Our sequence prediction objective of phase (III) aims to take advantage of this structure. Specifically, we continually run MLM pretraining on MIMIC-CXR radiology reports and propose to add a radiology section matching (RSM) pretraining task, formulated to match IMPRESSION to FINDINGS sections of the same study.

Let θ denote the weights of our language model and $m \subset \{1, \dots, T\}$ denote mask indices for M masked tokens, randomly sampled for each token vector $\mathbf{w}$ at every iteration. Given a batch $\mathcal{B}$ of token vectors $\mathbf{w} = (w_1, \dots, w_T)$, we write the MLM loss as the cross-entropy for predicting the dynamically masked tokens: $\mathcal{L}_{\text{MLM}} = -\frac{1}{|\mathcal{B}|} \sum_{\mathbf{w} \in \mathcal{B}} \log p_\theta(\mathbf{w}_m \,|\, \mathbf{w}_{\setminus m})$. Further, let $(\tilde{\mathbf{t}}_i^{\text{F}}, \tilde{\mathbf{t}}_i^{\text{I}})$ denote a pair of `[CLS]` tokens corresponding to the FINDINGS and IMPRESSION sections of the same i^{th} report, and let $(\mathbf{t}_i^{\text{F}}, \mathbf{t}_i^{\text{I}})$ denote the pair projected to a lower dimension via a two-layer perceptron P_{txt}. We introduce a contrastive loss on the text modality that favours IMPRESSION and FINDINGS text pair from the same report over unmatched ones. Specifically, for a batch of N such pairs, the RSM loss is defined as:

$$\mathcal{L}_{\text{RSM}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\mathbf{t}_i^{\text{F}} \cdot \mathbf{t}_i^{\text{I}} / \tau_1)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^{\text{F}} \cdot \mathbf{t}_j^{\text{I}} / \tau_1)} + \log \frac{\exp(\mathbf{t}_i^{\text{I}} \cdot \mathbf{t}_i^{\text{F}} / \tau_1)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^{\text{I}} \cdot \mathbf{t}_j^{\text{F}} / \tau_1)} \right), \quad (1)$$

where $\tau_1 > 0$ is a scaling parameter to control the margin. The resulting total loss of the specialisation phase (III) is $\mathcal{L}_{\text{III}} = \mathcal{L}_{\text{RSM}} + \lambda_{\text{MLM}} \mathcal{L}_{\text{MLM}}$. An additional important component for regularising the RSM loss is the use of increased dropout (25%), including on attention. We set $\tau_1 = 0.5$ and $\lambda_{\text{MLM}} = 0.1$, determined by a limited grid-search measuring $\mathcal{L}_{\text{GA}}$ (Eq. (2)) of the joint model on a

validation set. We also note that similar losses to the RSM loss, over the same or separate text segments, have been explored successfully for sentence representation learning [23,50] in other settings. As such, we empirically observed that an objective as in [23] using masked FINDINGS to FINDINGS matching can achieve similar performance and may be an appropriate replacement in other biomedical settings with differing text structure.

Text Augmentation. As domain-specific datasets are often quite small, effective text augmentation can induce large benefits. In the radiology domain, the sentences of the FINDINGS and IMPRESSION sections, which contain the detailed description and summary of the radiological findings, are usually permutation-invariant on the sentence level (cf. [60]). We thus find that randomly shuffling sentences within each section is an effective text-augmentation strategy for both pretraining of CXR-BERT as well as during joint model training.

2.2 BioViL: Vision-Language Representation Learning

We now introduce BioViL, a simple but effective self-supervised VLP setup for the biomedical domain (Fig. 1), which we study in a chest X-ray (CXR) application setting. BioViLuses a convolutional neural network (CNN) [38] image encoder E_{img}, our CXR-BERT text encoder E_{txt}, and projection models P_{img} and P_{txt} to learn representations in a joint space. The CNN model allows us to obtain a grid of local image embeddings $\tilde{\mathbf{V}} = E_{\text{img}}(\mathbf{x}_{\text{img}})$, which is fine-grained enough to be useful for segmentation (e.g. 16×16). Each encoder is followed by a modality-specific two-layer perceptron projection model P, which projects the encoded modality to a joint space of 128 dimensions–e.g., $\mathbf{V} = P_{\text{img}}(\tilde{\mathbf{V}})$–where the representation is ℓ_2-normalised. Note that projection should be applied to local embeddings before mean-pooling $\mathbf{v} = \text{pool}(P_{\text{img}}(\tilde{\mathbf{V}}))$, which gives us the global image embedding $\mathbf{v}$. The text branch uses the IMPRESSION section's projected `[CLS]` token $\mathbf{t}^{\text{I}}$ as the text representation in the joint space, as it contains a succinct summary of radiological findings. To align the representations and learn a joint embedding, we propose to use two loss terms. For a batch of size N, a symmetric contrastive loss [58] for *global alignment* of the image and text projections helps us learn the shared latent semantics:

$$\mathcal{L}_{\text{GA}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(\mathbf{v}_i\cdot\mathbf{t}_i^{\text{I}}/\tau_2)}{\sum_{j=1}^{N}\exp(\mathbf{v}_i\cdot\mathbf{t}_j^{\text{I}}/\tau_2)} + \log\frac{\exp(\mathbf{t}_i^{\text{I}}\cdot\mathbf{v}_i/\tau_2)}{\sum_{j=1}^{N}\exp(\mathbf{t}_i^{\text{I}}\cdot\mathbf{v}_j/\tau_2)}\right). \quad (2)$$

where $\tau_2 > 0$ is a scaling parameter. Further, we maintain the $\mathcal{L}_{\text{MLM}}$ loss during joint training, resulting in the final joint loss $\mathcal{L}_{\text{joint}} = \lambda_{\text{GA}}\mathcal{L}_{\text{GA}} + \mathcal{L}_{\text{MLM}}$. We set $\tau_2 = 0.5$ and $\lambda_{\text{GA}} = 0.5$, determined by a limited grid search measuring $\mathcal{L}_{\text{GA}}$ on a validation set.

Augmentations, Regularisation, and Image Encoder Pretraining. Due to the small dataset sizes expected in biomedical applications, we use image and text augmentations to help learn known invariances. We use a ResNet-50 [29] architecture as our image encoder and pretrain the model on MIMIC-CXR images

using SimCLR [6] with domain-specific augmentations as detailed in Section 4.1. For text, we use the same sentence-shuffling augmentation as in pretraining of CXR-BERT (see Section 4.1 for details). Furthermore, as in phase (III) of CXR-BERT training, we apply higher text encoder dropout (25%) than in standard BERT settings [17,73]. We find that the combination of all these components, including continuous MLM optimisation, is important to improve downstream performance across the board (see ablation in Table 4).

Zero-shot Classification. After joint training, we use text prompts to cast the zero-shot classification problem into an image–text similarity task as in [31,61,62]. For C classes, subject-matter experts design C text prompts representing the target labels $c \in \{1, \ldots, C\}$, e.g. for presence or absence of pneumonia (see Section 4.5). Each class prompt is represented as a vector of tokens $\mathbf{w}^c$ and passed to the text encoder and projector of BioViL to obtain ℓ_2-normalised text features $\mathbf{t}^c = P_{\text{txt}}(E_{\text{txt}}(\mathbf{w}^c)) \in \mathbb{R}^{128}$. For each input image $\mathbf{x}_{\text{img}} \in \mathbb{R}^{H \times W}$, we use the image encoder and projection module to obtain patch embeddings $\mathbf{V} = P_{\text{img}}(E_{\text{img}}(\mathbf{x}_{\text{img}})) \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 128}$ for segmentation tasks or the pooled embedding $\mathbf{v} = \text{pool}(\mathbf{V}) \in \mathbb{R}^{128}$ for instance-classification. We use dilated convolutions [82] to obtain higher-resolution feature maps. Probabilities for classes/regions can then be computed via a softmax over the cosine similarities between the image (or region) and prompt representations.

Few-shot Tasks with BioViL. To further assess the representation quality, linear probing is applied to local ($\mathbf{V}$) and global ($\mathbf{v}$) image representations, by learning $\boldsymbol{\beta} \in \mathbb{R}^{128 \times C}$ weights and a bias term. Unlike [31,85], we leverage the pretrained projectors and class text embedding $\mathbf{t}^c$ from the zero-shot setting by using them for initialisation, which leads to improved performance and further reduces the need for manual label collection. Specifically, in few-shot classification settings, the weights and bias are initialised with $\boldsymbol{\beta} = [\mathbf{t}^1, \ldots, \mathbf{t}^C]$ and zeros, respectively.

3 Evaluating Self-Supervised Biomedical VLP

Accurate local alignment between modalities is an important characteristic of successful joint image-text training in healthcare, in particular since image and report samples often contain multiple clinical findings, each of which correspond to distinct image regions. Standard global-alignment approaches may attain high classification accuracy by overfitting to spurious image features for a given finding (e.g., chest tubes in images correlating with mentions of pneumothorax in reports). Image classification, the most frequently evaluated downstream task in related work [31,45,56,85], requires only scene-level labels, hence a less sophisticated understanding of natural-language image descriptions. Image classification tasks can largely be solved by simply detecting a small set of words and maintaining some understanding of negation, as exemplified by the development of automated, rule-based text-labellers such as CheXpert [33]. Instance-level image-text retrieval tasks address some evaluation limitations, but do not require the level of language reasoning needed to solve local correspondence between phrases

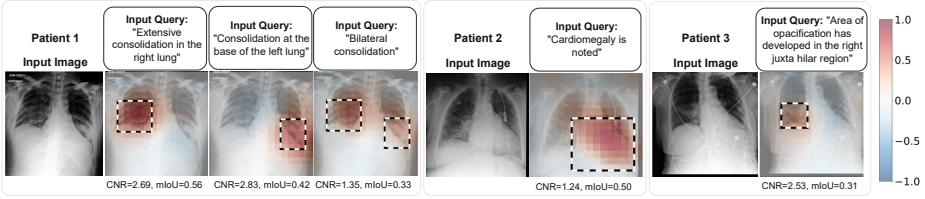


Fig. 3: Examples from the newly released MS-CXR phrase grounding dataset with BioViL latent vector similarity for different input text queries superimposed as heatmaps. Dashed boxes are ground-truth annotations by radiologists. X-ray images are mirrored horizontally.

and image regions. Existing public CXR benchmark datasets to evaluate local aspects of VLP have one or more of the following limitations (see Section 5 and Supp. C,D for more details): bounding boxes without corresponding free text descriptions, a limited number of samples, a limited number of abnormalities, and non-curated phrases impacting evaluation quality.

With this motivation in mind, we design MS-CXR, a radiology visual-grounding benchmark that has domain-specific language (e.g., paraphrasing and negations) and forms a more challenging real-world image-text reasoning task compared to existing evaluation datasets. To name just a few challenges, the phrase grounding task requires the ability to parse domain specific location modifiers, the ability to deal with reporting style biases, and understanding of complex negations, all while relating the correct findings to specific image regions.

3.1 MS-CXR – A Chest X-ray Phrase Grounding Benchmark

We publicly release MS-CXR, a new dataset containing image bounding box labels paired with radiology text descriptions, annotated and verified by two board-certified radiologists (see examples in Figs. 3 and C.1). MS-CXR provides 1153 image–sentence pairs of bounding boxes and corresponding phrases, collected across eight different cardiopulmonary radiological findings, with an approximately equal number of pairs for each finding (see Table C.1). It is curated to ensure gold-standard evaluation of phrase grounding. The phrases in MS-CXR are not simple short captions, but genuine descriptions of radiological findings from original radiology reports [34] and dictated transcripts [37]. Thus, compared to existing evaluation datasets, the proposed benchmark is a more challenging real-world image-text reasoning task.

All the benchmark samples are chosen from the public MIMIC-CXR dataset [24,34]. To collect a set of bounding-box labels, we first select samples from a set of studies with pre-existing image annotations (e.g., ellipses) [37,71] and verify their correctness. To link each image region with candidate phrases, we sampled sentences from the report of each study by extracting the highest matching sentences to the annotated labels using scores of the CheXbert classifier [69], and also used transcriptions of dictations when available [37]. Next, to better balance findings, we sampled additional studies at random as well as the ones used in the ImaGenome dataset [79], the latter being a dataset of annotations

Table 2: Comparing evaluations conducted in recent CXR image-text alignment studies.

Downstream task	Used in ref.*	Image encoder	Text encoder	Phrase reasoning	Findings localisation	Latent alignment	Annotation availability
Natural language inference	[B]	-	✓	✓	-	-	Scarce
Phrase grounding	[B]	✓	✓	✓	✓	✓	Scarce
Image classification	[B,C,G,L,M]	✓	-	-	-	-	High
Zero-shot image classif.	[B,G]	✓	✓	-	-	✓	Moderate
Dense image prediction (e.g. segmentation)	[B,G,L]	✓	-	-	✓	-	High
Global image–text retrieval	[C,G]	✓	✓	-	-	✓	High

*B, BioViL (Proposed); C, ConVIRT [85]; G, GLoRIA [31]; L, LoVT [56]; M, Local MI [45].

of anatomical regions. Note that these sampled studies do not have preexisting region proposals. Radiologists then manually reviewed separate sets of candidates. If a bounding box was not available, the radiologists manually annotated the corresponding region(s) in the image with new bounding boxes. Radiologists rejected studies where no correct phrase candidates were available and where existing bounding boxes were placed incorrectly (e.g., covering too large an area). To ensure a high quality, consistent benchmark, the phrase-image samples that do not adhere to our guidelines (see Supp. C.1) were filtered out, such as phrases containing multiple abnormalities in distinct lung regions.

4 Experiments

We conduct a comprehensive evaluation of our CXR-BERT language model as well as the proposed BioViL self-supervised VLP approach, and compare both to state-of-the art counterparts. Table 2 shows how our evaluation coverage compares to recent related studies. We begin by demonstrating CXR-BERT's superior performance and improved vocabulary, including on a radiology-specific NLI benchmark. Next, we assess joint image-and-text understanding of BioViL on our new `MS-CXR` benchmark, which evaluates grounding of phrases describing radiological findings to the corresponding image regions. We also investigate zero-shot classification and fine-tuning performance of BioViL on image- and pixel-level prediction tasks via the RSNA pneumonia dataset [66,76].

4.1 Setup

Datasets. We conduct experiments on the MIMIC-CXR v2 [34,24] chest radiograph dataset, which provides 227,835 imaging studies with associated radiology reports for 65,379 patients, all collected in routine clinical practice. We only use frontal view scans (AP and PA) and also discard studies without an Impression section. From this data, we establish a training set of 146.7k samples and a set of 22.2k validation samples, ensuring that all samples used for the different downstream evaluations are kept in a held-out test set. We emphasise that no labels are used during pretraining; for early stopping only a loss on validation data is tracked. For evaluation, we use RadNLI [54] to assess the proposed CXR-BERT text model in isolation, the new `MS-CXR` assesses joint image–text understanding

via phrase grounding, and the RSNA Pneumonia dataset [66,76] to test zero-shot segmentation, as well as zero-shot and fine-tuned classification performance.

Image and Text Pre-processing. We downsize and centre crop images to a resolution of 512×512 whilst preserving image aspect ratios. We perform image augmentations during training including: random affine transformations, random colour jitter, and horizontal flips (only for image fine-tuning tasks). For text model pre-training we utilise the 'FINDINGS' and 'IMPRESSION' sections of reports, while joint training is performed using only the latter. During training, we perform sentence shuffling within sections as text-augmentation. Additionally, we perform limited automatic typo correction as in [5].

Comparison Approaches. The proposed CXR-BERT text model is compared to the other specialised PubMedBERT [26] and ClinicalBERT [2] models. Note that ClinicalBERT was used in most related studies [31,45,85,56]. We compare BioViL to the closely related, state-of-the-art ConVIRT [85], LoVT [56] and GLoRIA [31] approaches (see Section 5). Lastly, we create BioViL-L by extending BioViL with the local loss term introduced in [31] to illustrate the complementary role of proposed pre-training strategy to recent advances in biomedical VLP.

Metrics. We report segmentation results via mean intersection over union (mIoU) and contrast-to-noise ratio (CNR), and report the Dice score [10] to compare to [56]. We first compute the cosine similarity between a projected phrase embedding $\mathbf{t}$ and local image representations $\mathbf{V}$, resulting in a grid of scores between $[-1, 1]$. The similarities are later thresholded to compute mIoU and Dice score. The mIoU is defined as an average over the thresholds [0.1, 0.2, 0.3, 0.4, 0.5]. The CNR measures the discrepancy between scores inside and out of the bounding box region, without requiring hard thresholds. This evaluation of local similarities is important as some clinical downstream applications may benefit from heatmap visualisations as opposed to discrete segmentations. For CNR, let A and $\overline{A}$ denote the interior and exterior of the bounding box, respectively. We then compute $\text{CNR} = |\mu_A - \mu_{\overline{A}}|/(\sigma_A^2 + \sigma_{\overline{A}}^2)^{\frac{1}{2}}$, where μ_X and σ_X^2 are the mean and variance of the similarity values in region X.

4.2 Text Model Evaluation

Natural Language Understanding. We use the RadNLI benchmark [54] to evaluate how well the proposed CXR-BERT text model captures domain-specific semantics. The dataset contains labelled hypothesis and premise pairs, sourced from MIMIC-CXR radiology reports, with the following label categories: (1) entailment, i.e. the hypothesis can be inferred from the premise; (2) contradiction, i.e. the hypothesis cannot be inferred from the premise; and (3) neutral, i.e. the inference relation is undetermined. RadNLI provides expert-annotated development and test sets (480 examples each), but no official training set. Thus, following [54], we use MedNLI [67] for training, which has 11k samples sourced from MIMIC-III discharge summaries, with equally distributed NLI labels. We fine-tune the language models up to 20 epochs and use early stopping by monitoring accuracy scores on the RadNLI development set. Table 3 summarises

Table 3: Evaluation of text encoder intrinsic properties and fine-tuning for radiology natural language inference: (1) RadNLI fine-tuning scores (average of 5 runs); (2) Mask prediction accuracy on MIMIC-CXR val. set; (3) Vocabulary comparison, number of tokens vs. original number of words in FINDINGS, increase shown as percentage.

	RadNLI accuracy (MedNLI transfer)	Mask prediction accuracy	Avg. # of tokens after tokenization	Vocabulary size
RadNLI baseline [54]	53.30	-	-	-
ClinicalBERT	47.67	39.84	78.98 (+38.15%)	28,996
PubMedBERT	57.71	35.24	63.55 (+11.16%)	28,895
CXR-BERT (after Phase-III)	60.46	77.72	58.07 (+1.59%)	30,522
CXR-BERT (after Phase-III + Joint Training)	65.21	81.58	58.07 (+1.59%)	30,522

the NLI evaluation, masked token prediction, and subword tokenisation results. Using only MedNLI training samples, our model achieves a good accuracy of 65.21%, and far outperforms fine-tuned ClinicalBERT, PubMedBERT, and the score reported in RadNLI [54]. Another important result is that RadNLI accuracy improves after joint training with images (last row of Table 3).

Mask Prediction Accuracy. While mask prediction accuracy does not always translate to downstream application performance, it is an auxiliary metric that captures important aspects of a language model's grasp of a target domain. We report Top-1 mask prediction accuracy on radiology reports in the MIMIC-CXR validation set (Table 3), and follow the standard masking configuration (15% masking probability). Despite being trained on closely related data, our CXR-BERT displays a much better mask prediction accuracy compared to ClinicalBERT (trained on MIMIC-III, which includes radiology reports) and PubMedBERT (trained on biomedical literature text). This suggests that radiology text significantly differs from other clinical text or biomedical literature text, highlighting the need for specialised text encoder models.

Ablation. We also conduct an ablation of the various aspects of CXR-BERT, measuring the impact after joint training. Table 4 shows that all components of CXR-BERT contribute to improved downstream and NLI performance, both in terms of alignment between related sentences (entailments) and of discrimination of contradictions. In particular, note the substantial improvement on these scores due to keeping the MLM objective during joint finetuning.

Table 4: CXR-BERT ablation. CNR and mIoU are macro averages of BioViL performance on all categories of `MS-CXR`. *Syn. sim.* denotes the average cosine similarity between RadNLI entailments. *Cont. gap* is the average similarity gap of RadNLI entailment and contradiction pairs. CXR-BERT is the combination of all components below the first row.

	RadNLI		Grounding	
Model or pretraining stage	Syn. sim.	Cont. gap	mIoU	CNR
ClinicalBERT	.657	.609	.182	0.791
Pretrain & Vocab (I–II)	.749	.646	.194	0.796
+ MLM loss added to joint training	.871	.745	.209	0.860
+ Use of attention drop-out (III)	.893	.802	.217	0.945
+ RSM Pretrain (III)	.877	.779	.220	1.012
+ Sentence shuffling (CXR-BERT)	.884	.798	.220	1.031

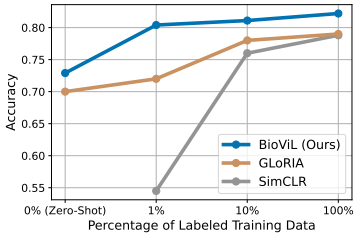
4.3 Local Alignment Evaluation – Phrase Grounding

We perform a phrase grounding evaluation of the pretrained BioViL model on the `MS-CXR` dataset. For each image–phrase pair, the image is passed to the CNN

Table 5: Contrast-to-noise ratio (CNR) obtained on the newly released `MS-CXR` dataset, averaged over four runs with different seeds. The results are collected using different text encoder and training objectives (e.g., G&L: Global and local loss).

Method	Objective	Text encoder	Atelectasis	Cardiomegaly	Consolidation	Lung opacity	Edema	Pneumonia	Pneumothorax	Pl. effusion	Avg.
Baseline	Global	ClinicalBERT	0.70±.03	0.53±.04	1.15±.07	0.75±.12	0.83±.04	0.85±.09	0.29±.01	1.05±.05	0.769±.02
Baseline	Global	PubMedBERT	0.72±.08	0.64±.05	1.22±.07	0.69±.07	0.80±.04	0.91±.09	0.21±.07	0.99±.03	0.773±.05
ConVIRT [85]	Global	ClinicalBERT	0.86±.04	0.64±.06	1.25±.06	0.78±.07	0.68±.07	1.03±.05	0.28±.08	1.02±.03	0.818±.01
GLoRIA [31]	G&L	ClinicalBERT	0.98±.04	0.53±.31	1.38±.03	1.05±.04	0.66±.03	1.18±.04	0.47±.02	1.20±.04	0.930±.03
BioViL	Global	CXR-BERT	1.02±.06	0.63±.08	1.42±.02	1.05±.06	0.93±.03	1.27±.04	0.48±.06	1.40±.06	1.027±.02
BioViL-L	G&L	CXR-BERT	1.17±.04	0.95±.21	1.45±.03	1.19±.05	0.96±.05	1.19±.01	0.74±.05	1.50±.03	1.142±.04

Table 6: RSNA Pneumonia zero-shot and fine-tuned classification. We compare to GLoRIA scores reported in [31] which outperforms ConVIRT [85] (see [31]). Training size: GLoRIA ($N = 186k$, private dataset), BioViL ($N = 146.7k$ of MIMIC-CXR).



Method	Type	Text model	Loss	% of labels	Acc.	F1	AUROC
SimCLR [6]	Image only	-	Global	1%	0.545	0.522	0.701
				10%	0.760	0.639	0.802
				100%	0.788	0.675	0.849
GLoRIA [31]	Joint	ClinicalBERT	Global & local	Zero-shot	0.70	0.58	-
				1%	0.72	0.63	0.861
				10%	0.78	0.63	0.880
				100%	0.79	0.65	0.886
Baseline	Joint	ClinicalBERT	Global	Zero-shot	0.719	0.614	0.812
BioViL	Joint	CXR-BERT	Global	Zero-shot	0.732	0.665	0.831
				1%	0.805	0.723	0.881
				10%	0.812	0.727	0.884
				100%	0.822	0.733	0.891

image encoder and projected to obtain a grid of image representations $\mathbf{V}$ in the joint space. Similarly, the phrase is embedded via the text encoder and projected to the joint space to obtain $\mathbf{t}$. Cosine similarity between $\mathbf{t}$ and elements of $\mathbf{V}$ produces a similarity grid, which is evaluated against the ground-truth bounding boxes. Table 5 shows the superior phrase grounding results achieved by BioViL across radiological findings and further shows that the addition of local losses as in our BioViL-L can improve phrase grounding performance for almost all findings. Moreover, the ablation in Table 4 demonstrates that there are clear gains to be had in visual grounding performance by improving the text model.

4.4 Global Alignment Evaluation – Zero-shot & Linear Probing

To measure global alignment quality, the joint models are also benchmarked on zero-/few-shot binary pneumonia classification problems (image-level) using the external RSNA dataset [66]. Fine-tuning is done via linear probing, i.e. only a last linear layer is trained. The evaluation is conducted on $\mathcal{D}_{\text{test}} = 9006$ images as in [31] (30% eval. / 70% train.) using the dataset's ground-truth labels. We define two simple text prompts for BioViL, representing presence/absence of pneumonia: "Findings suggesting pneumonia" and "No evidence of pneumonia". The image encoders are utilised and fine-tuned as described in Section 2.2.

The zero-shot and fine-tuned results in Table 6 show that our focus on better text modelling results in improved joint modelling of shared latent information between text-image pairs. Note that, to achieve its superior performance here and in Section 4.5, BioViL does not require extensive human expert text-prompt engineering (see Supp. A.1 for a sensitivity analysis) as for example conducted in GLoRIA [31], where variations over severity and/or location were created.

4.5 Local Alignment Evaluation – Semantic Segmentation

We evaluate models on an RSNA pneumonia segmentation task, using grid-level image representations in the joint latent space. We use the same text prompts as in the previous section for all models, and evaluate against ground-truth bounding boxes of the RSNA pneumonia dataset ($|\mathcal{D}_{\text{train}}| = 6634$ and $|\mathcal{D}_{\text{test}}| = 2907$). Table 7 shows that BioViL significantly reduces the need for dense annotations as compared to similar multi-modal and image-only pretraining approaches, outperforming them when using the same number of labelled data points. Note that our proposed modelling framework BioViL(Fig. 1), uses neither a local loss term [31,56], nor a separate object detection [63] or segmentation network [65]. Further, while Table 7 shows results using two simple queries, we find that BioViL continues to outperform related work even when more prompts are used for all models as in [31]. Dice and IoU are computed using the same threshold of 0.6 on predictions scaled between $[0, 1]$.

Table 7: RSNA pneumonia segmentation, showing *Zero-shot* and *linear probing* results. Related work is reproduced in the same experimental setup except for LoVT [56].

Method	% of Labels	Supervision	IoU	Dice	CNR
LoVT [56]	100%	Lin. prob.	-	0.518	-
ConVIRT [85]	-	Zero-shot	0.228	0.348	0.849
GLoRIA [31]	-	Zero-shot	0.245	0.366	1.052
BioViL	-	Zero-shot	0.355	0.496	1.477
SimCLR [6]	5%	Lin. prob.	0.382	0.525	1.722
SimCLR [6]	100%	Lin. prob.	0.427	0.570	1.922
BioViL	5%	Lin. prob.	0.446	0.592	2.077
BioViL	100%	Lin. prob.	0.469	0.614	2.178

5 Related Work

We refer the reader to Supp. D for a more detailed review of related work.

Biomedical Vision–Language Processing. Multiple studies explore joint representation learning for paired image and text radiology data [30,31,45,56,85]. [85] follow a contrastive learning formulation for instance-level representation learning, while [31,56] introduce approaches that combine instance-level radiology image–report learning with local terms. An alternative, local-only objective is explored by [45], approximating the mutual information between local image features and sentence-level text features. While most related approaches use no ground truth, [5] study a semi-supervised edema severity classification setting, and [28] assume sets of seen and unseen labels towards CXR zero-shot classification.

Related medical VLP work commonly uses publicly available contextual word embedding models including BioBERT [39], ClinicalBERT [2], BioClinicalBERT [2], or PubMedBERT [26]. The models are either trained from scratch or fine-tuned via continual pretraining using an MLM objective. Additional objectives such as adversarial losses [47] are added infrequently. The specialised corpora these models use include PubMed abstracts and PubMed Central full texts (see [2]), as well as MIMIC-III [35] clinical notes.

Local Alignment Datasets. Presently, no datasets exist that allow for phrase grounding of radiology findings, but some enable different forms of local image evaluations. VinDr [57], RSNA Pneumonia [66], and the NIH Chest X-ray Dataset [76] provide bounding-box annotations, but lack free-text descriptions.

REFLACX [37] provides gaze locations (ellipses) captured with an eye tracker, dictated reports, and some ground truth annotations for gaze locations, but no full phrase matches to image regions. Phrase annotations for MIMIC-CXR data released in [71] are of small size (350 studies), only contain two abnormalities, and for some samples have shortened phrases that were adapted to simplify the task. The ground-truth set of ImaGenome [79] only contains 500 studies, bounding-box regions annotate anatomical regions rather than radiological findings, and its sentence annotations are not curated for grounding evaluation.

6 Conclusion

We show that careful attention to text modelling can lead to large benefits for all learned models in self-supervised vision language processing (VLP) frameworks for medical applications. We introduce a novel pretraining procedure and publicly release a radiology domain-specific language model: CXR-BERT. It has an improved vocabulary and understanding of radiology sentences, contributing to improved downstream performance for all aspects of VLP approaches, e.g., the superior performance on a radiology natural language inference benchmark.

We also present BioViL, as a simple yet effective baseline for self-supervised multi-modal learning for paired image–text radiology data, with a focus on improved text modelling. The approach displays state-of-the-art performance on a large number of downstream tasks evaluating global and local aspects of the image model, text model, and joint latent space. On zero-shot tasks, the model does not require extensive text-prompt engineering compared to prior work. Notably, it outperforms related work on segmentation without requiring a local loss term or an additional vision model to produce region proposals. In that regard, it is complementary to local contrastive losses, and the combination of the two yields improved phrase grounding performance (Table 5).

To support the research community in evaluating fine-grained image–text understanding in the radiology domain, we also publicly release a chest X-ray phrase grounding dataset called `MS-CXR`. It presents a more challenging benchmark for joint image–text understanding compared to existing datasets, requiring reasoning over real-world radiology language and scans to ground findings in the correct image locations. Limitations of the proposed joint approach include that it does not explicitly deal with false negatives in the contrastive losses. Furthermore, co-occurrence of multiple abnormalities could enable contrastive methods to focus only on a subset to match pairs, e.g. pneumothorax and chest tubes commonly occur together [25]. Amongst its failure cases (see Supp. A.2 for more), we have seen that the approach struggles with very small structures, likely due to image resolution limits. Future work will expand the evaluated radiological findings, and explore using larger image resolution.

Acknowledgements: We would like to thank Dr Javier González and Fernando Pérez-García for their valuable feedback and contributions, Hannah Richardson for helping with the compliance review of the datasets, and Dr Matthew Lungren for their clinical input and data annotations provided to this study.

References

1. Akbari, H., Karaman, S., Bhargava, S., Chen, B., Vondrick, C., Chang, S.F.: Multi-level multimodal common semantic space for image-phrase grounding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 12476–12486. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01276
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). https://doi.org/10.18653/v1/W19-1909, https://aclanthology.org/W19-1909
3. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: PadChest: A large chest X-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
4. Casey, A., Davidson, E., Poon, M., Dong, H., Duma, D., Grivas, A., Grover, C., Suárez-Paniagua, V., Tobin, R., Whiteley, W., et al.: A systematic review of natural language processing applied to radiology reports. BMC medical informatics and decision making **21**(1), 1–18 (2021)
5. Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., Golland, P.: Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–539. Springer (2020)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020), http://proceedings.mlr.press/v119/chen20j.html
7. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
8. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.112, https://aclanthology.org/2020.emnlp-main.112
9. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. The Lancet **392**(10162), 2388–2396 (2018)
10. Crum, W.R., Camara, O., Hill, D.L.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE transactions on medical imaging **25**(11), 1451–1461 (2006)
11. Dai, S., Wang, Q., Lyu, Y., Zhu, Y.: BDKG at MEDIQA 2021: System report for the radiology report summarization task. In: Proceedings of the 20th Workshop on Biomedical Language Processing. pp. 103–111. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.bionlp-1.11, https://aclanthology.org/2021.bionlp-1.11

12. Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., Divakaran, A.: Align2Ground: Weakly supervised phrase grounding guided by image-caption alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 2601–2610. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00269
13. Datta, S., Roberts, K.: A hybrid deep learning approach for spatial trigger extraction from radiology reports. In: Proceedings of the Third International Workshop on Spatial Language Understanding. vol. 2020, pp. 50–55. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.splu-1.6, https://aclanthology.org/2020.splu-1.6
14. Datta, S., Si, Y., Rodriguez, L., Shooshan, S.E., Demner-Fushman, D., Roberts, K.: Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning. Journal of biomedical informatics **108**, 103473 (2020)
15. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
16. Desai, K., Johnson, J.: VirTex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11162–11173 (2021)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
18. Dligach, D., Bethard, S., Becker, L., Miller, T., Savova, G.K.: Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association **21**(3), 448–454 (2014)
19. Dunnmon, J.A., Ratner, A.J., Saab, K., Khandwala, N., Markert, M., Sagreiya, H., Goldman, R., Lee-Messer, C., Lungren, M.P., Rubin, D.L., Re, C.: Cross-modal data programming enables rapid medical machine learning. Patterns **1**(2), 100019 (2020)
20. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
21. Eyuboglu, S., Angus, G., Patel, B.N., Pareek, A., Davidzon, G., Long, J., Dunnmon, J., Lungren, M.P.: Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. Nature communications **12**(1), 1–15 (2021)
22. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 1473–1482. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298754
23. Gao, T., Yao, X., Chen, D.: SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910 (2021)

24. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
25. Graf, B., Sitek, A., Katouzian, A., Lu, Y.F., Krishnan, A., Rafael, J., Small, K., Xie, Y.: Pneumothorax and chest tube classification on chest X-rays for detection of missed pneumothorax. Machine Learning for Health (ML4H) NeurIPS Workshop: Extended Abstract (2020), https://arxiv.org/abs/2011.07353
26. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
27. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. In: 16th European Conference on Computer Vision, ECCV 2020. pp. 752–768. Springer (2020)
28. Hayat, N., Lashen, H., Shamout, F.E.: Multi-label generalized zero shot learning for the classiffcation of disease in chest radiographs. In: Machine Learning for Healthcare Conference. pp. 461–477. PMLR (2021)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90
30. Hsu, T.M.H., Weng, W.H., Boag, W., McDermott, M., Szolovits, P.: Unsupervised multimodal representation learning across medical images and reports. Machine Learning for Health (ML4H) NeurIPS Workshop (2018), https://arxiv.org/abs/1811.08615
31. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
32. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)
33. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Thirty-Third AAAI Conference on Artificial Intelligence. pp. 590–597. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.3301590
34. Johnson, A., Pollard, T., Berkowitz, S., Mark, R., Horng, S.: MIMIC-CXR database (version 2.0.0). PhysioNet (2019)
35. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
36. Joulin, A., Van Der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: European Conference on Computer Vision. pp. 67–84. Springer (2016)
37. Lanfredi, R.B., Zhang, M., Auffermann, W.F., Chan, J., Duong, P.A.T., Srikumar, V., Drew, T., Schroeder, J.D., Tasdizen, T.: REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. arXiv preprint arXiv:2109.14187 (2021). https://doi.org/10.13026/e0dj-8498, https://arxiv.org/abs/2109.14187

38. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
39. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
40. Li, A., Jabri, A., Joulin, A., Van Der Maaten, L.: Learning visual n-grams from web data. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 4183–4192. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.449, http://doi.ieeecomputersociety.org/10.1109/ICCV.2017.449
41. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. vol. 34(7), pp. 11336–11344. AAAI Press (2020), https://aaai.org/ojs/index.php/AAAI/article/view/6795
42. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
43. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021), https://arxiv.org/abs/2110.05208
44. Li, Y., Wang, H., Luo, Y.: A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1999–2004. IEEE (2020)
45. Liao, R., Moyer, D., Cha, M., Quigley, K., Berkowitz, S., Horng, S., Golland, P., Wells, W.M.: Multimodal representation learning via maximization of local mutual information. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2021)
46. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest X-ray report generation. In: Machine Learning for Healthcare Conference. pp. 249–269. PMLR (2019)
47. Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., Gao, J.: Adversarial training for large neural language models. arXiv preprint arXiv:2004.08994 (2020)
48. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019), https://arxiv.org/abs/1907.11692
49. Liu, Y., Wan, B., Ma, L., He, X.: Relation-aware instance refinement for weakly supervised visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5612–5621 (2021)
50. Logeswaran, L., Lee, H.: An efficient framework for learning sentence representations. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=rJvJXZb0W

51. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=Bkg6RiCqY7
52. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 13–23 (2019), https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html
53. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 11–20. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.9
54. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5288–5304. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.416, https://aclanthology.org/2021.naacl-main.416
55. Mu, Z., Tang, S., Tan, J., Yu, Q., Zhuang, Y.: Disentangled motif-aware graph learning for phrase grounding. AAAI (2021)
56. Müller, P., Kaissis, G., Zou, C., Rückert, D.: Joint learning of localized representations from medical images and reports. arXiv preprint arXiv:2112.02889 (2021), https://arxiv.org/abs/2112.02889
57. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. arXiv preprint arXiv:2012.15029 (2020). https://doi.org/10.13026/3akn-b287, https://arxiv.org/abs/2012.15029
58. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018), https://arxiv.org/abs/1807.03748
59. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
60. Preechakul, K., Piansaddhayanon, C., Naowarat, B., Khandhawit, T., Sriswasdi, S., Chuangsuwanich, E.: Set prediction in the latent space. Advances in Neural Information Processing Systems **34** (2021)
61. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
62. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: DenseCLIP: Language-guided dense prediction with context-aware prompting. arXiv preprint arXiv:2112.01518 (2021)
63. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
64. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neu-

ral Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada **28**, 91–99 (2015), https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html
65. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
66. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
67. Shivade, C.: MedNLI - A natural language inference dataset for the clinical domain. PhysioNet (Oct 2019)
68. Simard, P., Steinkraus, D., Platt, J.: Best practices for convolutional neural networks applied to visual document analysis. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. pp. 958–963. IEEE (2003)
69. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.117, https://aclanthology.org/2020.emnlp-main.117
70. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: pre-training of generic visual-linguistic representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2019), https://openreview.net/forum?id=SygXPaEYvH
71. Tam, L., Wang, X., Turkbey, E., Lu, K., Wen, Y., Xu, D.: Weakly supervised one-stage vision and language disease detection using large scale pneumonia and pneumothorax studies. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020 (March 2020)
72. Titano, J.J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., Swinburne, N., Zech, J., Kim, J., Bederson, J., et al.: Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nature medicine **24**(9), 1337–1341 (2018)
73. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
74. Wallis, A., McCoubrie, P.: The radiology report—are we getting the message across? Clinical radiology **66**(11), 1015–1022 (2011)
75. Wang, Q., Tan, H., Shen, S., Mahoney, M., Yao, Z.: MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2030–2038. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.159, https://aclanthology.org/2020.emnlp-main.159
76. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised clas-

sification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2097–2106. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.369
77. Wilcox, J.R.: The written radiology report. Applied Radiology **35**(7), 33 (2006)
78. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019), https://arxiv.org/abs/1910.03771
79. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W.G., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset for clinical reasoning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
80. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016), https://arxiv.org/abs/1609.08144
81. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 (2017), https://arxiv.org/pdf/1708.03888v3.pdf
82. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016), http://arxiv.org/abs/1511.07122
83. Yu, T., Hui, T., Yu, Z., Liao, Y., Yu, S., Zhang, F., Liu, S.: Cross-modal omni interaction modeling for phrase grounding. In: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. pp. 1725–1734 (2020). https://doi.org/10.1145/3394171.3413846
84. Zhang, Y., Ding, D.Y., Qian, T., Manning, C.D., Langlotz, C.P.: Learning to summarize radiology findings. In: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis. pp. 204–213. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-5623, https://aclanthology.org/W18-5623
85. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020), https://arxiv.org/abs/2010.00747
86. Zhang, Z., Zhao, Z., Lin, Z., He, X., et al.: Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems **33**, 18123–18134 (2020)