Contrastive Cross-Modal Pre-Training: A General Strategy for Small Sample Medical Imaging

Gongbo Liang, *Member, IEEE*, Connor Greenwell, *Student Member, IEEE*, Yu Zhang, *Student Member, IEEE*, Xin Xing, *Student Member, IEEE*, Xiaoqin Wang, Ramakanth Kavuluru, and Nathan Jacobs, *Senior Member, IEEE*

***Abstract*—A key challenge in training neural networks for a given medical imaging task is the difficulty of obtaining a sufficient number of manually labeled examples. In contrast, textual imaging reports are often readily available in medical records and contain rich but unstructured interpretations written by experts as part of standard clinical practice. We propose using these textual reports as a form of weak supervision to improve the image interpretation performance of a neural network without requiring additional manually labeled examples. We use an image-text matching task to train a feature extractor and then fine-tune it in a transfer learning setting for a supervised task using a small labeled dataset. The end result is a neural network that automatically interprets imagery without requiring textual reports during inference. We evaluate our method on three classification tasks and find consistent performance improvements, reducing the need for labeled data by 67%–98%.**

***Index Terms*—Annotation-efficient modeling, pre-training, convolutional neural network, text-image matching.**

I. Introduction

CONVOLUTIONAL neural networks (CNNs) have been promising tools for imaging analysis on various tasks [1]–[5] and have been rapidly adopted in the medical imaging domain [6]–[12]. However, robust CNNs are typically trained using large quantities of manually annotated data, such as images with discrete labels or pixel-level labels [13]–[19]. Though large number of images are acquired each year in the medical domain, the number of images with manually annotated labels remains quite small due to the high annotation cost [20]–[22]. For instance, the cancer imaging archive (TCIA), one of the largest collections of cancer images that is available for public download, hosted 127 datasets as of December 1^{st}, 2020; only six of them contain more than 1000 cases, and a majority (80 out of 127) have fewer than a hundred cases [23]. The limited number of manually annotated images presents a barrier to adopting modern deep learning techniques in the medical imaging domain [20]–[22], [24].

In contrast to manually annotated labels, textual imaging reports, which are often readily available in medical records, contain interpretations written by experts as part of standard clinical practice (Fig. 1). These unstructured textual data provide rich information about the images, but it is difficult to directly integrate the data into CNN training as labels.

We propose using textual imaging reports as a weak supervision signal. Our model gleans the associations between text and images through a text and image matching network that we call TIMNet (Text-Image Matching Network). The network learns the image features from a large number of text-image pairs. These training pairs include both positive examples of images and associated reports and those that randomly pair an image with a report not associated with it. In this sense, our model learns to contrast cross-modal pairs that are valid matches with those that do not correspond to the same diagnostic imaging event. Various models for downstream application can then be built using the pre-trained image feature extractor. Intuitively speaking, TIMNet transfers the strong expert generated language signal to the image feature part of the architecture, priming it more powerfully for downstream training on supervised tasks. Since the feature extractor is pre-trained on a large and relevant dataset, transfer learning can be applied when building a downstream application that allows only a small labeled dataset to be used [25], [26].

Our method is widely applicable in the medical imaging domain, in which textual reports are often readily available

Manuscript received December 23, 2020; revised June 3, 2021 and August 24, 2021; accepted August 30, 2021. Date of publication September 8, 2021; date of current version April 13, 2022. This work was supported in part the National Science Foundation under Grant IIS-1553116, in part by the American Cancer Society under Grant IRG-19-140-31, and in part by the National Cancer Institute under Grant P30CA177558. *(Corresponding author: Gongbo Liang.)*

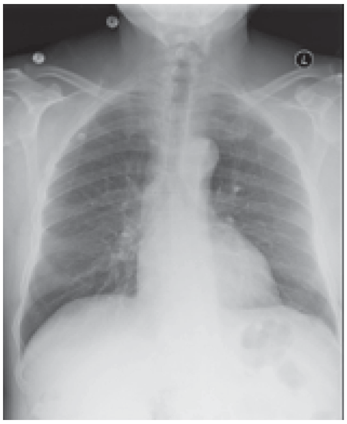
Gongbo Liang is with the Department of Computer Science, University of Kentucky, Lexington, KY 40506 USA, and also with the Department of Computer Science and Information Technology, Eastern Kentucky University, Richmond, KY 40475 USA (e-mail: tliang130@gmail.com).

Connor Greenwell, Yu Zhang, Xin Xing, and Nathan Jacobs are with the Department of Computer Science, University of Kentucky, Lexington, KY 40506 USA (e-mail: connor.greenwell@uky.edu; y.zhang@uky.edu; xxi242@g.uky.edu; jacobs@cs.uky.edu).

Xiaoqin Wang is with the Department of Radiology, University of Kentucky, Lexington, KY 40506 USA (e-mail: Xiaoqin.Wang@uky.edu).

Ramakanth Kavuluru is with the Division of Biomedical Informatic, University of Kentucky, Lexington, KY 40506 USA (e-mail: ramakanth.kavuluru@uky.edu).

Digital Object Identifier 10.1109/JBHI.2021.3110805



FINAL REPORT

HISTORY: ___-year-old female with chest pain.

COMPARISON: Comparison is made with chest radiographs from ___.

FINDINGS: The lungs are well expanded. A retrocardiac opacity is seen which is likely due to atelectasis although infection is hard to exclude. Given the linear shape of the opacity, atelectasis is perhaps more likely. The heart is top-normal in size. The cardiomediastinal silhouette is otherwise unremarkable. There is no pneumothorax or pleural effusion. Visualized osseous structures are unremarkable.

IMPRESSION: Retrocardiac opacity, likely due to atelectasis but possibly due to pneumonia in the appropriate setting.

Fig. 1. An example of a chest X-ray (left) with the radiology report (right) from the MIMIC-CXR dataset.

but obtaining manual labels is expensive. We demonstrate our method on three classification tasks, but our method uses textual data in a general way that is not limited to specific image analysis tasks. Our experimental results show that the proposed method reduces the need for manually annotated data by up to 98%. Our contributions in this work are as follows:

- We propose a novel weakly supervised method of pre-training medical imaging analysis networks, which learns image features from a large but unlabeled dataset with associated textual reports;
- We demonstrate the proposed method via three classification tasks and also investigate the transferability of the learned features between datasets;
- We discuss the challenges of this study and provide a clear research direction for future researchers.

II. Background

A. Textual Data in Image Analysis

Researchers are actively seeking solutions for using textual data effectively in the field of medical imaging analysis. However, most works focus on deriving manual labels from textual data in different ways. For instance, CheXpert [27] and NegBio [28] propose two natural language processing (NLP) algorithms that are used for clinical text parsing. Both of them extract keywords (e.g., `pneumonia` and `pulmonary edema`) from the textual reports, then assign attributes (e.g., `positive` or `negative`) to the extraction keywords. The output of such algorithms are pairs of keywords and attributes, for instance, `positive pneumonia` or `negative pulmonary edema`. One may further infer discrete image-level labels from such attribute-keyword pairs and use the derived labels in CNN training [29]–[32].

Although such kinds of approaches are useful in automatically labeling a large amount of data, NLP models for label inference need to be trained separately, which increases the difficulty of using such methods. In addition, the accuracy of the derived labels could be an issue. As shown in [33], the automatically generated labels using CheXpert and NegBio do not always agree with each other in a large number of cases. Empirically speaking, the label inference errors often manifest in a non-uniform manner, making the method even more problematic when training downstream application using the labels with nonuniform noise levels.

Instead of deriving labels from the clinical textual data, we propose using the textual data as weak supervision for image feature learning through text-image matching tasks. Our method uses textual data in a general way that is not limited to specific image analysis tasks. More importantly, there is no need to train a separate NLP model for label inferring that is independent of the image analysis tasks.

B. Text and Image Matching

Text and image matching has become a popular research topic in recent years. One commonly used strategy is global representation matching, that usually involves three procedures: 1) image feature embedding, 2) text feature embedding, and 3) measures of the distance between the two embeddings. For instance, Kiros *et al.* [34] used a CNN to encode images and a long short-term memory network (LSTM) [35] to encode the full text. The triplet loss is used to pull the embeddings of the matched text and image closer to each other and push the unmatched ones further apart. Wehrmann *et al.* [36] encoded the text data using an efficient character-level inception module that convolves over characters in the text. Sarafianos *et al.* [37] trained a text-image matching network by using a ResNet101 [13] network for imaging processing and a transformer-based model [38] with an LSTM based encoder for text processing.

Most of the prior efforts focused on the matching problem from the perspective of text-based querying systems for image retrieval. In contrast, we aim to imbue image analysis models with the knowledge obtained from learning to match a medical image with the associated clinician-generated textual report. We use the global matching approach with a two-branch setup one each for image processing and text processing. The absolute difference between the two output feature vectors is fed to a classification network, which predicts if the input image and text are a valid pair — the text snippet is in fact the correct findings report for the image.

C. Contrastive Learning

Contrastive learning is a machine learning strategy that learns the general features of a dataset by comparing the similarity and dissimilarity between data samples from the same class and different classes. In the imaging domain, contrastive learning can be done through Siamese networks that typically contain two identical subnetworks, which share the same weights [39]–[42]. The subnetworks take a pair of images either from the same class or different classes as input and output two feature vectors. A distance-based loss function—such as triplet loss [42]–[44] or contrastive loss [45]–[47]—is then used to compare the outputs that minimize the loss of samples from the same class and maximize the loss of samples from different classes.

Essentially, the Siamese network performs a binary classification task because each pair of input data is either from the "same" class or "not," which makes the binary cross-entropy loss a natural choice for Siamese network training. As triplet loss or contrastive loss, binary cross-entropy-based strategy has also been widely used [48]–[52]. For instance, Koch *et al.* firstly used such a method on image recognition and used binary cross-entropy loss to estimate the similarity between two feature vectors [48]; Wu *et al.* applied such a method on video-based person re-identification and got a promising result [49]; Yang *et al.* used binary cross-entropy to train a Siamese network that aims to overcome the extreme imbalance issue on text classification [50]. In addition, binary cross-entropy-based training is also used by major scientific computation software companies; for instance MATLAB uses it in their contrastive learning tutorial [52].

Contrastive learning is a natural choice to build our image and text matching network. We propose a Siamese-style network with the slightly variation where the subnetworks have different architectures given the inherent variation in processing and representing text and image data. We use the binary cross-entropy method to form the distance-based learning model.

D. Contextualized Natural Language Encoding

Contextualized natural language embedding represents words as dense embeddings in a real vector space [53]–[55]. Unlike the conventional word2vec embedding method [56], [57], the embeddings of contextualized methods are dependent on the surrounding context in which a word appears. The embeddings of the same word may be different according to the sentences in which the word occurs. Thus, contextualized embedding methods are more capable of addressing polysemy and homonymy issues in natural language processing. Popular contextualized embedding methods may include EMLo [54], transformers [58], and BERT [38].

The transformer architecture [58] is a well-known context-aware neural network architecture for natural language processing. It uses an encoder-decoder structure and attention mechanism to translate one sequence to another sequence. It encodes the contextual information of a word from distant parts of a sentence. In addition, unlike a conventional recurrent neural network (RNN), a transformer does not require sequential data to be processed in order, leading to new practical gains in efficiency.

The *bidirectional encoder representations from transformers* (BERT) [38] architecture trains a transformer by jointly conditioning on both left and right contexts in all layers. A pre-trained BERT model can be fine-tuned to create state-of-the-art models for various tasks without substantial task-specific architectural modifications [59], [60]. The BERT model contains twelve layers of transformer encoders. The outputs of these layers can be used as contextualized embeddings. In this study, we use a pre-trained BERT model as the contextualized natural language embedding model that encodes a given text as a feature vector. Specific details of this are presented in Section III-A.

III. Method

We assume two datasets (X_P and X_L) exist, where X_P has paired 2-tuples of images and associated textual reports and X_L consists of labeled images. Specifically, $X_P = \{(x_i^{(j)}, x_t^{(j)}) : x_i^{(j)} \in x_i, x_t^{(j)} \in x_t\}$ is an image and text paired dataset where x_i is a set of images and x_t is a set of textual notes such that $x_t^{(j)}$, the the j-report, corresponds to $x_i^{(j)}$, the j-th image. Also, $X_L = \{x_l, y\}$ is a labeled imaging dataset, where x_l is a set of images and y represents the set of corresponding labels, which could be image-level labels, bounding boxes, or pixel-level labels. Typically, we have $|X_L| \ll |X_P|$.

We first learn an image encoding function f using X_P that encodes x_i to a feature space. Then, f can be used to build a downstream application using X_L that maps x_l to y. We learn f through TIMNet, which predicts whether a given pair of (x_i, x_t) is naturally related.

The proposed idea is illustrated in Fig. 2. The network consists of two modules: 1) a weakly supervised image feature learning module and 2) a downstream application module for a specific task. The image encoding function f is a CNN image feature extractor that is shared between the modules.

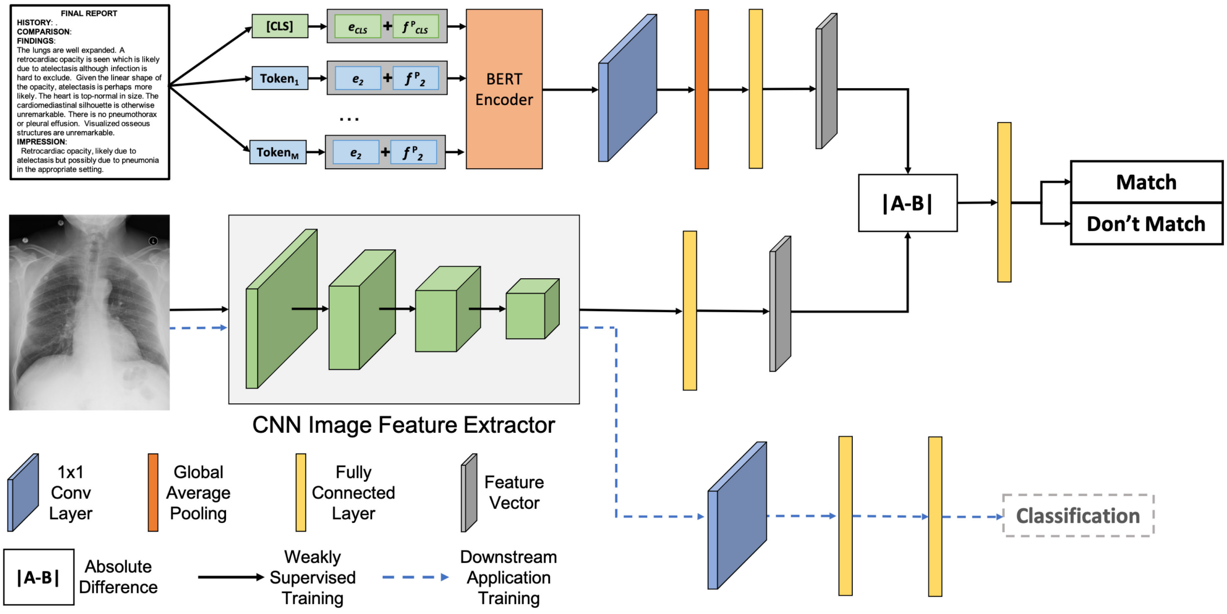
A. Weakly Supervised Image Feature Learning

Weakly supervised image feature learning, i.e., image encoding is carried out through a global cross-modal matching approach with a two-branch network, one for text processing and the other for imaging processing. The input of the network is a text-image pair with a label indicating whether the text and image correspond to the same imaging event. The pair of the text and image is first projected into a vector space. Then, the absolute difference between the two output feature vectors is fed to a classification network, which predicts whether the input image and text are a true pair.

Mathematically, the text processing branch is defined as:

$$v_t^{(j)} = h_t(x_t^{(j)}), \tag{1}$$

where $h_t(\cdot)$ is the text processing network, $x_t^{(j)}$ is the j^{th} input text, and $v_t^{(j)}$ is the output vector. The network is composed of a BERT encoder, a 1×1 convolutional (Conv) layer, a global average pooling (GAP) layer, and a fully-connected (FC) layer. The BERT encoder is used as a text feature extractor, which encodes the input text as contextualized embedding vectors. The vectors are then passed through the 1×1 Conv layer following



FINAL REPORT
HISTORY: .
COMPARISON:
FINDINGS:
The lungs are well expanded. A retrocardiac opacity is seen which is likely due to atelectasis although infection is hard to exclude. Given the linear shape of the opacity, atelectasis is perhaps more likely. The heart is top-normal in size. The cardiomediastinal silhouette is otherwise unremarkable. There is no pneumothorax or pleural effusion. Visualized osseous structures are unremarkable.
IMPRESSION:
Retrocardiac opacity, likely due to atelectasis but possibly due to pneumonia in the appropriate setting.

Fig. 2. The TIMNet cross-modal matching architecture. 1) weakly supervised image feature learning through a text-image matching network (solid black line). 2) Downstream application training using a small dataset (dashed blue line).

GAP. The FC layer projects the vectors into a feature space as a feature vector. The image processing branch is defined as:

$$v_i^{(j)} = h_i(x_i^{(j)}), \tag{2}$$

where $h_i(\cdot)$ is the image processing network, $x_i^{(j)}$ is the j^{th} input image, and $v_i^{(j)}$ is the output feature vector. The Conv layers of a ResNet model [13] are used as a feature extractor in $h_i(\cdot)$. An FC layer is added after the feature extractor that projects the image features into a feature space as a feature vector. The absolute difference of $v_t^{(j)}$ and $v_i^{(j)}$ is then passed to a shallow classification network, $h_{cls}(\cdot)$, for the matching prediction. Cross-entropy loss is used during the training of $h_{cls}(\cdot)$ to contrast valid matches from mismatched instances.

In general, the global matching network for weakly supervised image feature learning is expressed as:

$$h(x^{(j)}) = h_{cls}(|h_t(x_t^{(j)}) - h_i(x_i^{(j)})|), \tag{3}$$

where $x^{(j)}$ is the j^{th} input, $x^{(j)} = \{x_t^{(j)}, x_i^{(j)}\}$, $x^{(j)} \in X_P$, $|\cdot|$ denotes the absolute difference, and $h(\cdot)$ is the whole matching network. The absolute difference combining with cross-entropy loss forms a contrastive-style network that learns image features under a weakly supervised fasion.

B. Fine-Tuning for Downstream Tasks

At the end of the weakly supervised feature learning phase described earlier, the hypothesis is that the matching process has pre-trained the parameterized image feature extractor f to the extent that it needs fewer supervised instances for a downstream image-related task. This, in turn, relies on our high-level intuition that there is a nontrivial transferable signal available in the textual annotations to improve imaging tasks down the line.

The fine-tuning of the image processing branch in TIMNet for downstream tasks is fairly straightforward. Although it can be done for a variety of applications, in this study, we demonstrate the proposed method for classification tasks.

To build a downstream application model, we can either add additional Conv layers and FC layers to the image processing branch $h_i(\cdot)$, or use it as-is by retraining the FC layers. Since we need to optimize only the few additional layers from scratch, while the rest of the network is already pre-trained on a larger set of images from the same domain, the total number of required training instances for fine-tuning can be much smaller than training the entire network from scratch.

C. Implementation

The HuggingFace BERT with `bert-base-uncased` weights [61] is used as the backbone of the text processing branch. A 1×1 Conv layer, a GAP layer, and an FC layer with 512 neurons are added to the BERT encoder. The ResNet-18 model with ImageNet pre-trained weights is used as the backbone of the imaging processing branch. A 1×1 Conv layer and an FC layer with 512 neurons are added before and after the GAP layer, respectively. Both of the 1×1 Conv layers in the text processing and image processing branches are followed by batch normalization [62] and a rectified linear unit (ReLU) [63].

All the Conv layers of the image processing branch are used as a feature extractor in the downstream networks. A shallow CNN classification network is added on top of the feature extractor. The CNN classification network contains a 1×1 Conv layer, an FC layer with 512 neurons, and an output layer with various numbers of neurons for different tasks. Cross-entropy loss and an Adam optimizer [64] with a learning rate of 10^{-4} are used for both the weakly supervised feature learning stage and the downstream application training stage.

IV. Evaluation and Discussion

A. Datasets

We use the MIMIC-CXR [33] and Mendeley-V2 [65] datasets in this study. The MIMIC-CXR dataset is used in both the TIMNet pre-training and downstream application training. Mendeley-V2 is used only for downstream application training.

1) MIMIC-CXR: This dataset contains 227,835 radiographic studies of 64,588 patients with 368,948 chest X-rays and the associated radiology reports. The dataset also provides 14 labels (with 13 labels for different abnormalities and one label for the normal case), which are derived from the radiology reports using NLP tools, such as NegBio and CheXpert [27], [28]. In the official training/validation/testing split, the validation set and test set have 2,991 and 5,159 images, respectively. We combine the official validation and test sets to form our validation set. The chest X-ray images are resized to 500×500. We use the official training set and our validation set in both TIMNet pre-training and downstream application training.

2) Mendeley-V2: This dataset is a pediatric chest X-ray dataset that includes 4,273 pneumonia images and 1,583 normal images. Although the imaging modality is the same as that of MIMIC-CXR, the patient demographics are different. In addition, the images in Mendeley-V2 may have been acquired with devices from different vendors. We used the original training/validation split of this dataset to train a downstream application model that evaluates the transferability of pre-trained weights between different datasets while retaining the same imaging modality.

B. Evaluation Setup

We evaluate the proposed method through the downstream application performance and the degree of need for labeled instances for supervision in downstream application training

1) TIMNet Pre-Training: The TIMNet is pre-trained for 50 epochs using both the true text-image pairs and negative pairs from the training set of MIMIC-CXR. A true pair contains an image and the associated report. A negative pair contains an image and a report randomly selected from the training set. The "findings" portion of the radiology reports is used as the text input. The length of each piece of text is preprocessed to 256 words at the word embedding stage. We add 0 s for texts shorter than 256 words, and we snip much longer texts to 256 words. The output of this first weak-supervision phase is a probability estimate of the input text-image pair being a true match. All the Conv layers of the image processing branch are used as a feature extractor in the downstream networks. Shallow CNN classification networks are added on top of the feature extractor with the same architecture discussed in Sec III-C for various downstream applications.

2) Downstream Applications: The downstream networks are trained using varying amounts of the training data and image-level labels from the MIMIC-CXR and Mendeley-V2 datasets, ranging from 0.5% to 100%. Three downstream applications are trained, namely, 1) a binary classification model for abnormality diagnosis of MIMIC-CXR; 2) a multi-label classification model for lung disease diagnosis of MIMIC-CXR; and 3) a binary classification model for pediatric pneumonia diagnosis of Mendeley-V2. The downstream applications aim to test the TIMNet pre-trained feature extractor on the pre-training dataset (i.e., MIMIC-CXR) and an external dataset (i.e., Mendeley-V2). Each downstream network is trained for 100 epochs with a batch size of 16. No text is needed for training the downstream networks. We use the accuracy (ACC), the area under the receiver operating characteristic curve (auROC), precision (Prec), recall (Recall), F1 score (F1), and average precision (AP) as the evaluation metrics for the binary classification tasks and the auROC and AP for the multi-label classification tasks.

C. Classification on Pre-Training Dataset

We first present the evaluation results of downstream applications that are trained using the MIMIC-CXR dataset by comparing the proposed pre-training method (*Ours*) with two other methods, *Base* and *ImageNet*. The *Ours* model is pre-trained on MIMIC-CXR using the proposed text-image matching method. The *Base* model is randomly initialized using the default settings. The *ImageNet* model is pre-trained on the ImageNet dataset [66]. We use the PyTorch provided ImageNet weights in this study.

Two downstream applications are trained and tested using the MIMIC-CXR dataset: a binary classification model and a multi-label classification model. The binary classification model predicts whether an abnormality exists in an image, and the multi-label classification model predicts what kind of abnormality exists in an image. The output of the binary classification model is Boolean, and the output for the multi-label task is a multi-hot vector, with 1 indicating the presence of a particular class.

1) Binary Classification: Fig. 3 shows the result of the binary classification task on the MIMIC-CXR dataset. The results reveal that the proposed method has the best performance in general, with better gains when only a few labeled images are available. For instance, when using 0.5% of labeled data ($\approx$ 1,800 instances), the *Base* model has an accuracy of 66.41%, *ImageNet* has a 70.05% accuracy, while *Ours* has a 71.81% accuracy. The highest accuracy of *Base* is 76.38%, when it is trained with 90% of the labeled data ($\approx$ 325,000 instances). Both *Ours* and *ImageNet* surpass the best performance of *Base* with only 30% of the training data ($\approx$ 108,000 instances). Thus the need for manual labels is reduced by 67% when using the proposed method or ImageNet pre-trained method. However, the highest accuracy of *ImageNet* is 77.52%. *Ours* is able to further improve the result to 78.30%.

Similar trends are seen in auROC, F1, and AP scores, where the pre-trained models have better performance than *Base* across all settings; the performance of the proposed method is better than the ImageNet pre-trained models in almost all of the settings. One thing worth noting is the proposed method has a consistent performance across the four metrics (accuracy,

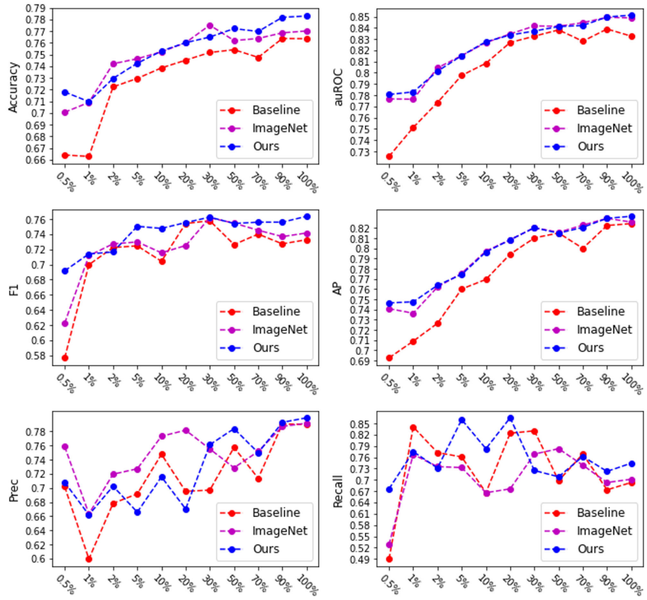


Fig. 3. Binary classification results on MIMIC-CXR dataset.

auROC, F1, and AP). However, *ImageNet* has a relatively poor performance on the F1 score. For instance, when only using 0.5% of training data (≈ 1,800 instances), the *Base* model has an F1 score of 0.5768. *ImageNet* has an F1 score of 0.6227, which is relatively low when compared with the *ImageNet* performance with 0.5% training on the other three metrics. However, *Ours* has an F1 score of 0.6917 that is within the performance range of the other three metrics when using 0.5% of training data. With 10% of training data (≈ 36,000 instances), *Ours* improves F1 score to 0.7477, but *ImageNet*'s score is 0.7155. The F1 score of *ImageNet* does reach the same range as *Ours* (≈ 0.76) with 30% of training data. However, *ImageNet* performance starts declining when increasing the size of the training dataset, but *Ours* maintains the same level after that.

There is no clear winner for precision and recall, but the proposed method still scores better across most settings. For both of the metrics, *Ours* wins 11 out of 22 training fractions, *ImageNet* wins 6 out of 22 training fractions, and *Base* wins 5 out of 22 training fractions. In addition, the figure reveals that the worst performances are generated by the *Base* model with 0.5% or 1% of training data. The best performances are achieved by *Ours*.

One particular observation is that for very small amounts of training data, such as 0.5% to 1% of the full dataset (≈ 1,800 to 3,600 instances), TIMNet performs consistently well across all evaluation metrics. ImageNet pre-trained model performs well on most of the metrics but scores poorly on F1 and Recall. The *Base* model has the worst performance across all the metrics when the training data is limited.

2) Multi-Label Classification: Fig. 4, the multi-label classification results on MIMIC-CXR, also shows the superior performance of the pre-training models compared to the *Base* model. Both TIMNet and ImageNet pre-training are able to significantly reduce the need for manual labels while achieving a comparable performance. The *Base* model reaches its best performance (0.9152 auROC) with 100% of the training data, while *Ours* and *ImageNet* can achieve a similar performance (0.9148 and 0.9137 auROCs, respectively) with only 30% of the training data. With 50% of the training data, *Ours* reaches 0.9169 auROC, while *ImageNet* reaches 0.9159 auROC.

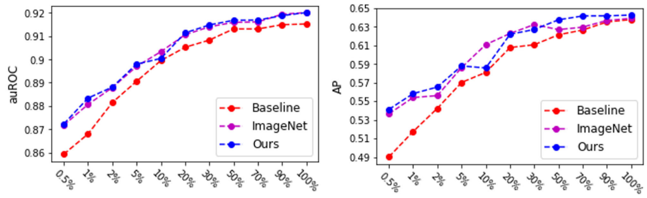


Fig. 4. Multi-label learning results on MIMIC-CXR dataset.

It is hard to pick a winner when comparing *Ours* against *ImageNet*. Both of the methods have similar performance on this multi-label classification task, especially for auROC. Though *Ours* scores better in 9 out of 11 training fractions, most of them are only marginally better than *ImageNet*. Slightly larger gains are observed in AP, which may show some advantage of TIMNet pre-training.

D. Classification on an External Dataset

This section aims to evaluate the TIMNet pre-trained weights on the dataset that is not used for the pre-training. We choose the Mendeley-V2 dataset for this evaluation because it is a pediatric pneumonia diagnosis dataset with the same medical imaging modality as MIMIC-CXR and different patient demographics.

Four models are trained and compared for pneumonia diagnosis using Mendeley-V2, namely, *Ours*, *Base*, *ImageNet*, and *C2L* [67]. The *Ours*, *Base*, and *ImageNet* follow the same setup with the previous experiment, which has a TIMNet pre-trained feature extractor, a randomly initialized feature extractor, or an ImageNet pre-trained feature extractor, respectively. The feature extractor in *C2L* is pre-trained using the *comparing to learn* method, the recent top method for self-supervised pre-training in medical imaging analysis tasks [67]. We use the author-released PyTorch code and pre-trained weights for ResNet-18 in this study[1]. These weights are also trained on the MIMIC-CXR dataset.

Fig. 5 shows the performance of the compared models. The results reveal that TIMNet's pre-trained features work surprisingly well and achieve the best performance, in general, on this dataset among the four methods on all evaluation metrics, except Recall. It is able to reduce the need for labeled data by 98.33% compared with the Base. The *Base* model achieves its highest accuracy of 87.52% using 30% of the training data, and the highest auROC of 0.9333 also uses 30% of the training data, while *Ours* outperforms *Base* with using only 0.5% of training data with an 88.14% accuracy and a 0.9352 auROC. Thus the reduction in training instances is $(30 - 0.5)/30 = 98.33\%$. The highest performances of *Ours* are 91.87% accuracy and 0.9613

[1] [Online]. Available: https://github.com/funnyzhou/C2L_MICCAI2020

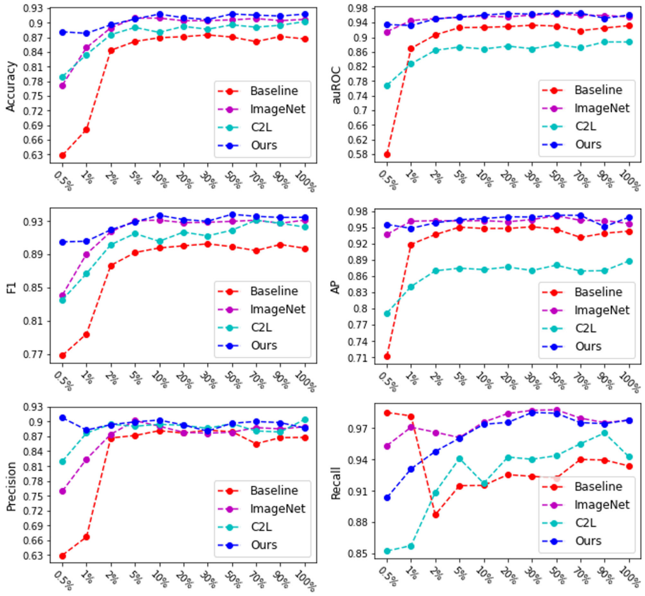


Fig. 5. Binary classification results on Mendeley-V2 dataset.

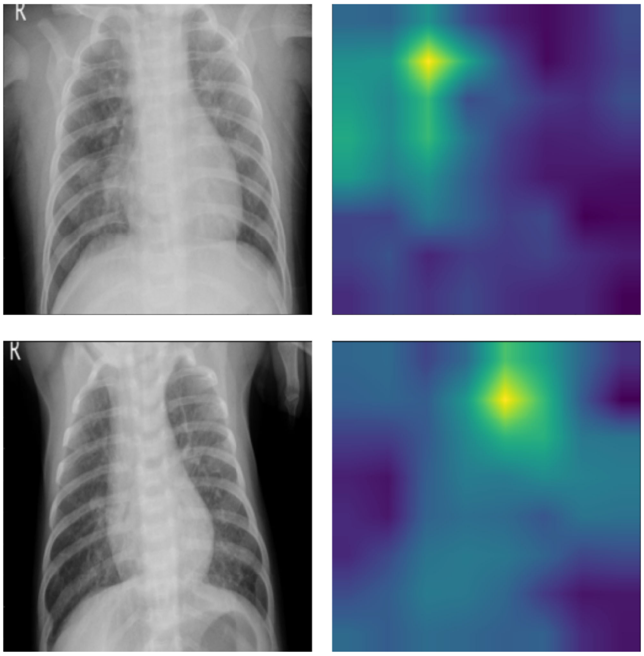


Fig. 6. Two CAM visualizations for pediatric chest X-ray pneumonia classification on the Mendeley-V2 dataset with chest X-ray on the left and CAM on the right. The CAMs reveal that the model correctly focuses on the corresponding areas show some concerns about pneumonia.

TABLE I
TEXT AND IMAGE MATCHING RESULTS

Dataset	Accuracy	auROC	F1 Score	Prec	Recall	AP
MIMIC-CXR	0.74	0.83	0.74	0.67	0.82	0.77

auROC. These are also nearly 5% (ACC) and 3% (auROC) higher than those of the *Base* model.

ImageNet achieves the second overall best performance, which is close but worse than the proposed method. One particular advantage of the proposed method is when the training set is small. For instance, when using only 0.5% of the training data, *Ours* gets an approximately 88% accuracy and an F1 score of 0.9. *ImageNet* only gets about 77% and 0.84, respectively. The difference between *Ours* and *ImageNet* reduces when increasing the amount of training data. On average, *Ours* performed about 1 to 2% better than *ImageNet* on all evaluation metrics, except recall, on which *Ours* is about 1% worse than *ImageNet*.

C2L performs significantly better than the *Base* when the training data size is extremely small. For instance, the method has about 21% higher accuracy and 33% higher auROC than the *Base* when trained using only 0.5% of the training data. However, a mixed result shows up when increasing the size of training data. On accuracy, F1 score, precision, and recall, *C2L* generates the third-best result among the four methods. It surpasses the best performance of *Base* on accuracy using only 2% of the training data and gets a comparable performance on F1 score using also only 2% of the training data, resulting in a 93.33% data reduction. However, *C2L* performs worse than *Base* on auROC and AP. C2L's highest auROC and AP are 0.8873 and 0.8877, respectively, which are 4.93% or 6.66% less than the *Base* scores, respectively.

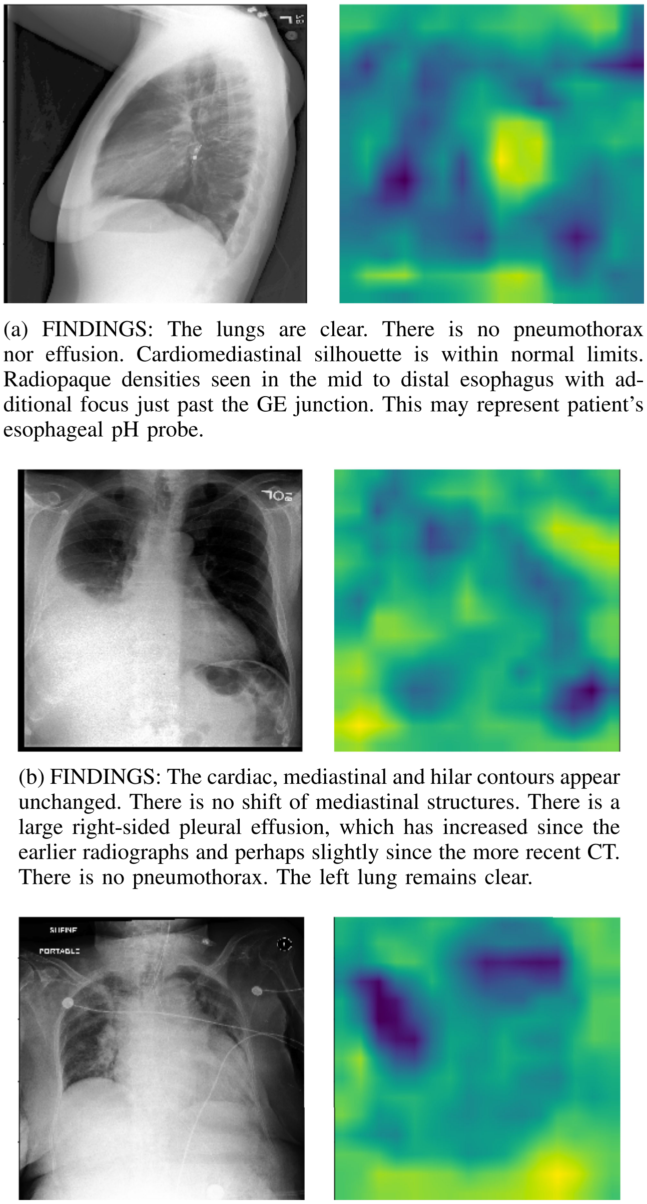
The high accuracy but low auROC and AP may imply the *C2L* model favors one class when making the prediction. The high F1 score may indicate the method works well when using 0.5 as the threshold for the binary classification task, but low auROC and AP may reveal that when changing the threshold to other values, the model performance may decrease. One feasible explanation for such a phenomenon is that as a self-supervise learning approach, *C2L* may heavily rely on the assumption that features from the same image should be similar. Without providing any supervision from domain experts, the feature learned under such assumption is open-ended, which may not robust enough in some cases.

Fig. 6 shows two class activation mapping (CAM) [68] visualizations for the pediatric pneumonia diagnosis that performed by *Ours*. The pixel values in the CAMs are associated with the contribution to the classification decision. A higher value (brighter color) indicates a higher contribution to the class decision. Both cases in the figure are ground-truth positive cases. The CAMs reveal that the model correctly focuses on the corresponding areas that show some concerns about pneumonia.

E. Discussion

We proposed learning meaningful feature representations of medical images using TIMNet via a text-image matching task without acquiring additional manually labeled annotations. Our method can help build effective models for downstream predictions that rely only on image input. During our experiments, we discovered that better text-image matching performances usually lead to improved downstream application performances in terms of higher accuracy and reduced need for labeled images.

The pre-trained TIMNet used in this study has a text-image matching performance of 74% accuracy and 0.83 auROC. Table I shows the relevant results of the evaluation. Fig. 7 shows CAM



(a) FINDINGS: The lungs are clear. There is no pneumothorax nor effusion. Cardiomediastinal silhouette is within normal limits. Radiopaque densities seen in the mid to distal esophagus with additional focus just past the GE junction. This may represent patient's esophageal pH probe.

(b) FINDINGS: The cardiac, mediastinal and hilar contours appear unchanged. There is no shift of mediastinal structures. There is a large right-sided pleural effusion, which has increased since the earlier radiographs and perhaps slightly since the more recent CT. There is no pneumothorax. The left lung remains clear.

(c) FINDINGS: ET tube is seen with tip approximately 1.8 cm from the carina. Enteric tube seen passing below the inferior field of view. Lower lung volumes are noted on the current exam with bilateral parenchymal opacities which could be due to edema or infection. Prominence of the right hilum is again noted. Moderate cardiomegaly and appears to have progressed since prior could potentially being part due to changes in positioning. No acute osseous abnormalities. Surgical clips project over the left chest wall/axilla.

Fig. 7. CAM visualizations of text and image matching on MIMIC-CXR with chest X-ray on the left and CAM on the right. The CAMs reveal that the model focuses on the corresponding areas that show some concerns in the textual findings.

visualizations of TIMNet on the text-image matching task. The findings of the radiology reports are displayed below the images. The CAMs suggest that the decisions made by TIMNet are reasonable. For instance, in Fig. 7(a), the radiology report mentions radiopaque densities in the mid to distal esophagus, and the CAM appears to show that in the middle part of the image. For Fig. 7(b), the radiology report indicates increased right-sided pleural effusion, and CAM shows more significant contributions near the effusion areas on the right-hand side of the figure. Fig. 7(c) shows a similar correspondence between the CAM-based image segment contributions to the model decision and the textual report.

We also observed that text-image matching tasks in the medical domain are typically **more challenging** than in the natural imaging domain, specifically due to the following situations:

- Textual data in the medical domain usually contain **more words** per instance than the natural imaging domain. The vast majority of natural image captions contain 7 to 15 words [69]. However, this can be ten times or higher in the medical imaging domain [33].
- Textual descriptions in the medical domain usually convey **more uncertainty and ambiguity** than those in the natural imaging domain. For example, the findings of Fig. 7(c) are, *"moderate cardiomegaly and appears to have progressed since prior* **could potentially***being part due to changes in positioning."* Such hedge statements are usually not found in natural imaging captions.
- Textual data in the medical domain usually includes **information not explicitly present** in the image, such as the findings of Fig. 7(b), which mention that *"the cardiac, mediastinal and hilar* **contours appear unchanged**." Without comparison with previous studies, the network may not be able to link the words "**contours appear unchanged**" to any regions in the image.
- Textual data in the medical domain usually contain **redundant and uninformative text**. The majority of words of a report could be describing situations that convey normal health, which may not be very helpful when making the decision in text-image matching. Similar normal health condition statements may appear in many examples.

These nuances make text-image matching a difficult task in the medical imaging domain.

V. Conclusion

The main objective of our work is to demonstrate the potential of the clinician-authored textual reports that accompany most medical images in improving image-related supervised ML applications by pretraining a feature extractor without requiring additional manually labeled annotations. Such textual narratives are readily available and routinely curated as part of healthcare operations and are therefore a natural resource to leverage. The central premise of our effort is the insight that in a latent neural dense vector representation space, it may be possible to transfer linguistic signals that characterize expert summaries of images to downstream image-based tasks through weak supervision. Based on the experiments and evaluations in this paper, we believe we have successfully verified this insight for classification tasks.

At the core of our methodology is a two-branch architecture, TIMNet, that identifies whether a textual finding corresponds to the supplied image. The main purpose of text-image matching tasks is to use the textual finding as weak supervision to learn the image feature representations. In this way, the feature extractor is trained on a large and relevant dataset without requiring additional manually labeled annotations. Subsequently, the image branch is further fine-tuned for downstream supervised tasks using a small labeled dataset. Our experiments show that with the proposed method, small fractions (2%–30%) of the available full training data are needed to achieve the same performance as baseline models that do not exploit textual reports. Additionally, the benefits persist across datasets, which is an excellent benefit when transferring signals from models learned on deidentified textual reports (e.g., MIMIC-CXR) to other classification settings that have fewer or no textual annotations (due to HIPAA and other privacy restrictions).

During our experiments, we discovered that better text-image matching performances usually lead to improved downstream application performances. Thus, one of our future directions will be to further innovate the matching framework to improve the associated performance. Another important future direction is to see whether our text-image matching setup can actually transfer the image signal to downstream tasks in the NLP domain for clinical text. We believe this bidirectional feedback may help in extracting named entities (e.g., drugs, comorbidities, and anatomical sites) and relations connecting such entities (e.g., adverse drug reactions) from a variety of notes. These information extraction tasks are also usually plagued by a lack of large training datasets (esp. public ones) due to strict regulatory constraints governing textual data in medicine. Overall, we hope our work spurs further interest in exploring the synergy between text and images.

References

[1] T. Salem, S. Workman, M. Zhai, and N. Jacobs, “Analyzing human appearance as a cue for dating images,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–8.

[2] M. Zhai, T. Salem, C. Greenwell, S. Workman, R. Pless, and N. Jacobs, “Learning geo-temporal image features,” in *Proc. British Mach. Vision Conf.*, 2018, *arXiv:1909.07499*.

[3] Y. Zhang, G. Liang, T. Salem, and N. Jacobs, “Defense-PointNet: Protecting pointnet against adversarial attacks,” in *Proc. Int. Conf. Big Data*, 2019, pp. 5654–5660.

[4] Y. Su *et al.*, “A deep learning view of the census of galaxy clusters in illustristng,” *Monthly Notices Roy. Astronomical Soc.*, vol. 498, no. 4, pp. 5620–5628, 2020.

[5] G. Liang *et al.*, “Alzheimer’s disease classification using 2D convolutional neural networks,” in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021.

[6] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[7] L. Liu, S. Yang, L. Meng, M. Li, and J. Wang, “Multi-scale deep convolutional neural network for stroke lesions segmentation on CT images,” in *Int. MICCAI Brainlesion Workshop*, 2018, pp. 283–291.

[8] R. P. Mihail, G. Liang, and N. Jacobs, “Automatic hand skeletal shape estimation from radiographs,” *IEEE Trans. Nanobiosci.*, vol. 18, no. 3, pp. 296–305, Jul. 2019.

[9] Y. Zhang *et al.* “2D convolutional neural networks for 3D digital breast tomosynthesis classification,” in *Proc. Int. Conf. Bioinf. Biomed*, 2019, pp. 1013–1017.

[10] G. Liang, Y. Zhang, X. Wang, and N. Jacobs, “Improved trainable calibration method for neural networks on medical imaging classification,” in *Proc. British Mach. Vis. Conf.*, 2020.

[11] X. Xing *et al.*, “Dynamic image for 3D MRI image Alzheimer’s disease classification,” in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2020, pp. 355–364.

[12] Q. Ying *et al.*, “Multi-modal data analysis for Alzheimer’s disease diagnosis: An ensemble model using imagery and genetic features,” in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[15] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[16] G. Liang, S. Fouladvand, J. Zhang, M. A. Brooks, N. Jacobs, and J. Chen, “GANai: Standardizing CT images using generative adversarial network with alternative improvement,” in *Proc. IEEE Int. Conf. Healthcare Informat.*, 2019, pp. 1–11.

[17] A. Y. Hannun *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature Med.*, vol. 25, no. 1, pp. 65–69, 2019.

[18] T. Falk *et al.*, “U-Net: Deep learning for cell counting, detection, and morphometry,” *Nature Meth.*, vol. 16, no. 1, pp. 67–70, 2019.

[19] G. Liang, X. Wang, Y. Zhang, and N. Jacobs, “Weakly-supervised self-training for breast cancer localization,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 1124–1127.

[20] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med.Image Anal.*, vol. 42, pp. 60–88, 2017.

[21] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, “Clinical Big Data and deep learning: Applications, challenges, and future outlooks,” *Big Data Mining Analytics*, vol. 2, no. 4, pp. 288–305, 2019.

[22] M. J. Willemink *et al.*, “Preparing medical imaging data for machine learning,” *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[23] “The cancer imaging archive,” *The Cancer Imaging Archive (TCIA)*, 2020. [Online]. Available: https://www.cancerimagingarchive.net/

[24] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger, and N. Jacobs, “Inconsistent performance of deep learning models on mammogram classification,” *J. Amer. College Radiol.*, vol. 17, no. 6, pp. 796–803, 2020.

[25] K. Mendel, H. Li, D. Sheth, and M. Giger, “Transfer learning from convolutional neural networks for computer-aided diagnosis: A comparison of digital breast tomosynthesis and full-field digital mammography,” *Academic Radiol.*, vol. 26, no. 6, pp. 735–743, 2019.

[26] G. Liang *et al.*, “Joint 2D-3D breast cancer classification,” in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 692–696.

[27] J. Irvin *et al.*, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 590–597.

[28] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, “NegBio: A high-performance tool for negation and uncertainty detection in radiology reports, ”*Proc. AMIA Summits Trans. Sci.*, 2018, Art. no. 188.

[29] J. Wu *et al.*, “Automatic bounding box annotation of chest X-ray data for localization of abnormalities,” in *Proc. 17th IEEE Int. Symp. Biomed. Imag.*, 2020, pp. 799–803.

[30] L. K. Tam, X. Wang, E. Turkbey, K. Lu, Y. Wen, and D. Xu, “Weakly supervised one-stage vision and language disease detection using large scale pneumonia and pneumothorax studies,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2020, pp. 45–55.

[31] M. Moradi, A. Madani, Y. Gur, Y. Guo, and T. Syeda-Mahmood, “Bimodal network architectures for automatic generation of image annotation from text,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2018, pp. 449–456.

[32] J. Yuan, H. Liao, R. Luo, and J. Luo, “Automatic radiology report generation based on multi-view image fusion and medical concept enrichment,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2019, pp. 721–729.

[33] A. E. Johnson *et al.*, “Mimic-cxr, a de-identified publicly availabled atabase of chest radiographs with free-text reports,” *Scientific Data*, vol. 6, no. 1, pp. 1–8, 2019.

[34] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *NIPS Deep Learn. Workshop*, 2014, *arXiv:1411.2539*.

[35] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] J. Wehrmann and R. C. Barros, “Bidirectional retrieval made simple,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7718–7726.

[37] N. Sarafianos, X. Xu, and I. A. Kakadiaris, “Adversarial representation learning for text-to-image matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2019, pp. 5814–5824.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguis.,: Human Lang. Tech.*, 2019, pp. 4171–4186.

[39] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a ‘siamese’ time delay neural network,” *Adv. Neural Inf. Process. Syst.*, vol. 6, pp. 737–744, 1993.

[40] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1735–1742.

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2014, pp. 1701–1708.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[43] M. Schultz and T. Joachims, “Learning a distance metric from relative comparisons,” *Adv. Neural Inf. Process. Syst.*, vol. 16, pp. 41–48, 2004.

[44] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, “Large scale online learning of image similarity through ranking,” *J. Mach. Learn. Res.*, vol. 11, no. 3, pp. 1109–1135, 2010.

[45] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[46] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[47] I. Misra and L. V. D. Maaten, “Self-supervised learning of pretext-invariant representations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6707–6717.

[48] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in *Proc. ICML Deep Learn. Workshop*, vol. 2, 2015.

[49] L. Wu, Y. Wang, J. Gao, and X. Li, “Where-and-when to look: Deep siamese attention networks for video-based person re-identification,” *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1412–1424, Jun. 2019.

[50] W. Yang, J. Li, F. Fukumoto, and Y. Ye, “MSCNN: A monomeric-siamese convolutional neural network for extremely imbalanced multi-label text classification,” in *Proc. Conf. Empirical Meth. Natural Lang. Process.*, 2020, pp. 6716–6722.

[51] M. Shorfuzzaman and M. S. Hossain, “MetaCOVID: A siamese neural network framework with contrastive loss for *n*-shot diagnosis of COVID-19 patients,” *Pattern Recognit.*, vol. 113, 2021, Art. no. 107700.

[52] “Train a siamese network to compare images,” *MATLAB & Simulink - MathWorks*, 2021. [Online]. Available: https://ww2.mathworks.cn/help/deeplearning/ug/train-a-siamese-network-to-compare-images.html

[53] B. McCann, J. Bradbury, C. Xiong, and R. Socher, “Learned in translation: Contextualized word vectors,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6297–6308.

[54] M. Peters *et al.*, “Deep contextualized word representations,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 2227–2237.

[55] N. Smith, “Contextual word representations: Putting words into computers,” *Commun. ACM*, vol. 17, no. 6, pp. 66–74, 2020.

[56] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. Adv. Neural Inf. Process. Syst*., 2013, pp. 3111–3119.

[57] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proc. 1st Int. Conf. Learn. Represent*., 2013.

[58] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst*., 2017, pp. 5998–6008.

[59] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune bert for text classification?,” in *Proc. China Nat. Conf. Chin. Comput. Linguistics*., Springer, 2019, pp. 194–206.

[60] J. Lee *et al.*, “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[61] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.

[62] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn.*, 2015, *arXiv:1502.03167*.

[63] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Represent*., 2015.

[65] D. Kermany and M. Goldbaum, “Labeled optical coherence tomography (OCT) and chest X-ray images for classification,” *Mendeley Data*, vol. 2, no. 2, 2018.

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[67] H.-Y. Zhou, S. Yu, C. Bian, Y. Hu, K. Ma, and Y. Zheng, “Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2020, pp. 398–407.

[68] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016, pp. 2921–2929.

[69] X. Hu *et al.*, “Vivo: Visual vocabulary pre-training for novel object captioning,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1575–1583.