Improving Image Representations via MoCo Pre-Training for Multimodal CXR Classification

Francesco Dalla Serra[1,2(✉)], Grzegorz Jacenków[3], Fani Deligianni[2], Jeff Dalton[2], and Alison Q. O'Neil[1,3]

[1] Canon Medical Research Europe, Edinburgh, UK
[2] University of Glasgow, Glasgow, UK
[3] University of Edinburgh, Edinburgh, UK
francesco.dallaserra@mre.medical.canon

Abstract. Multimodal learning, here defined as learning from multiple input data types, has exciting potential for healthcare. However, current techniques rely on large multimodal datasets being available, which is rarely the case in the medical domain. In this work, we focus on improving the extracted image features which are fed into multimodal image-text Transformer architectures, evaluating on a medical multimodal classification task with dual inputs of chest X-ray images (CXRs) and the indication text passages in the corresponding radiology reports. We demonstrate that self-supervised Momentum Contrast (MoCo) pre-training of the image representation model on a large set of unlabelled CXR images improves multimodal performance compared to supervised ImageNet pre-training. In particular, MoCo shows a 0.6% absolute improvement in AUROC-macro, when considering the full MIMIC-CXR training set, and 5.1% improvement when limiting to 10% of the training data.
To the best of our knowledge, this is the first demonstration of MoCo image pre-training for multimodal learning in medical imaging.

Keywords: multimodal learning, multimodal BERT, image representation, self-supervised image pre-training, CXR classification

1 Introduction

Multimodal learning has recently gained attention for healthcare applications [1], due to the rich patient representation enabled by combination of different data sources *e.g.* images, reports, and clinical data. Recent works in multimodal learning have mainly focused on Transformer [2] architectures, with similar approaches adopted in the medical domain [3]. Whilst the role of the joint pre-training process has been widely explored [4], fewer works have focused on the single modality components of the models. In particular, the role of the image representation is frequently neglected. However, the task of multimodal representation learning is complex and one of the main challenges in the medical domain is the lack of large-scale, labeled datasets, compared to the millions of images

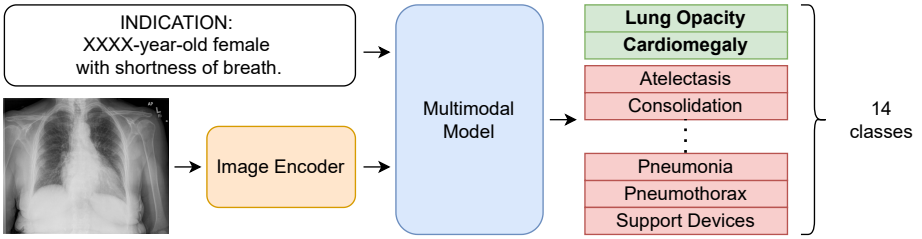

Fig. 1. Illustration of the multimodal CXR multi-label classification pipeline. The indication field and CXR image are dual input modalities, and the output is a set of positive (green) or negative (red) predictions for 14 radiographic findings labels, as annotated in the MIMIC-CXR dataset [13]–[15]. In this data, taken from the IU-Xray dataset [16], ages (and other patient-identifiable information) is replaced by a placeholder, here indicated by XXXX. The **image encoder** is the component that we investigate in this paper, to discover a strategy for learning a good image representation.

available in computer vision tasks in the general domain. Therefore, we seek to mitigate the complexity of multimodal learning by providing robust image representation as input.

In the general multimodal domain, the "bottom-up top-down" [5] approach is a popular image representation paradigm for multimodal Transformer architectures such as VisualBERT [6] and ViLBERT [7]. These models use Region of Interest (RoI) feature maps extracted from Faster R-CNN [8], which is pre-trained on large object detection datasets (*e.g.* VisualGenome [9]). Faster R-CNN requires large scale datasets with bounding box annotations. The image encoder is then frozen during fine-tuning of the Transformer model, based on the strong assumption that pre-trained detectors extract representative features for the downstream task. Other image representation strategies have been proposed. In Pixel-BERT [10], the image representation is defined as the feature map of the last convolutional layer of a convolutional neural network (CNN). Similarly, the discrete latent space of a variational autoencoder (VAE) has been adopted in DALL-E [11]. Alternatively, the Vision Transformer (ViT) [12] consists of directly feeding raw pixel patches as the input for Transformer architectures.

In this paper, we are interested in multimodal CXR multi-label classification of medical images supported by the medical history of the patient which is available in free-text radiology reports (indication field), as shown in Figure 1. We use MIMIC-CXR [13]–[15], which is the largest open access multimodal medical dataset, to evaluate our proposed methodology, for the task of Chest X-Ray classification of 14 radiographic findings classes. In MIMIC-CXR, bounding boxes are not available, making the "bottom-up top-down" [5] approach unsuitable for this task. To the best of our knowledge, there are two works performing multimodal classification of CXR using the text indication section as an additional inference-time input: ChestBERT [3] and what we denote as "Attentive" [17]. Following ChestBERT [3], the state-of-the-art for this task, we adopt the multimodal bitransformer model (MMBT), which has a similar image represen-

Table 1. Summary of the considered image pre-training strategies suited to CXR image classification.

Method	Ease of training	Medical image suitability
Supervised ImageNet *Supervised training on 1000 ImageNet classes.*	No training required. Pre-trained weights available for standard CNN architectures.	Weak – trained on natural images.
Autoencoder *Encoder-decoder architecture trained on reconstruction loss.*	Easy – does not require large batches.	Flexible – can train on relevant medical image data (no labels required)
SimCLR *Contrastive learning approach*	Hard – requires high compute power to handle the large batches ($> 10^3$ images).	
MoCo *Contrastive learning approach*	Moderate – designed to work with a small batch size ($\sim 10^2$ images) & uses efficient updating of the large dynamic dictionary.	

tation to Pixel-BERT. Differently to the previously described multimodal BERT models, MMBT does not include a joint pre-training step. More recently, Liao et al [18] have shown a method of joint modality pre-training to be effective by maximising the mutual information between the encoded representations of images and their corresponding reports. At inference time, the image only is used for classification. However, we consider the situation where we may have limited task-specific labelled multimodal (paired image and text) training data, but ample unlabelled unimodal (imaging) data available for pre-training and therefore we investigate image-only pre-training techniques.

For learning good visual representations, many self-supervised contrastive learning strategies have shown promising results in the medical domain, for instance Momentum Contrast (MoCo) contrastive training [19] [20] and Multi-Instance Contrastive Learning (MICLe) [21] – an application of SimCLR [22] to medical imaging. In particular, MoCo pre-training has shown superior results in a similar chest X-Ray imaging classification task, outperforming other methods using standard supervised pre-training on ImageNet [20]. Similarly, MedAug [23] has extended the work of Sowrirajan et al [20], by considering different criteria to select positive pairs for MoCo. However, the best approach in [23] (which targets mixed-view classification) is to create pairs from lateral and frontal views of CXR, while we focus our work on frontal views only, making this method unsuitable for our task. MoCo works by minimising the embedding distance between positive pairs – generated by applying different data augmentations to an image – and maximising the distance to all other augmented images in the dataset [19]. MoCo maintains a large dynamic dictionary of negative samples as

a queue with fixed length (set as a hyperparameter) which is updated every step by adding the newest batch of samples and removing the oldest. This allows the model to have a large number of negative samples without the need for very large batches, unlike other contrastive learning approaches (*e.g.* SimCLR [22]), making MoCo a sensible choice when training on fewer GPUs[1]. In this work, for the imaging component of MMBT we experiment with two strategies for training a CNN image encoding: a) MoCo and b) a classic autoencoder (AE) strategy (as in DALL-E [11], although the VAE is not required for our task).

In summary, our contributions are to:

1. Compare how different pre-training strategies of the image encoder perform in the multimodal setup: autoencoder (AE), MoCo, and standard (supervised) ImageNet pre-trained weights; finding MoCo to perform best.
2. Explore how these strategies degrade when the multimodal transformer is fine-tuned on a smaller subset of the training set, finding MoCo pre-trained weights to perform better in a limited data scenario.
3. Extend the work of [20] – which demonstrates the effectiveness of MoCo pre-training for image-only pleural-effusion classification – to a multimodal multi-label classification problem.
4. Apply Gradient-weighted Class Activation Mapping (Grad-CAM) [24] to evaluate the impact of the pre-training strategy on the image features that activate the model, and report quantitative results on a small subset of the ChestX-ray8 test set with annotated bounding boxes [25].

2 Method

Our proposed three step pipeline is shown in Figure 2. In particular, in this work, we explore the effectiveness of different image representations for the model by considering different pre-training strategies.

2.1 Model

The overall architecture for this work is based on the multimodal bitransformer model (MMBT) [26] as shown in Figure 2c. This builds on the BERT architecture [27], adapting it for multimodal data by introducing an additional visual input. Both the textual and the visual input are projected into the input embedding space and summed with the related positional and segment embedding; the segment embedding is then available to the model to discriminate between textual and visual inputs. The image embedding corresponds to the feature map outputted from the last convolutional layer of ResNet-50. This is flattened to obtain $N = 49$ embedding and projected by a single fully connected layer, indicated as $I = \{I_1 \dots I_N\}$. The textual input is tokenised into M BERT subword

[1] Due to the limited computing power, we decided to neglect the contrastive learning approach proposed by [21], trained on 16–64 Cloud TPU cores.

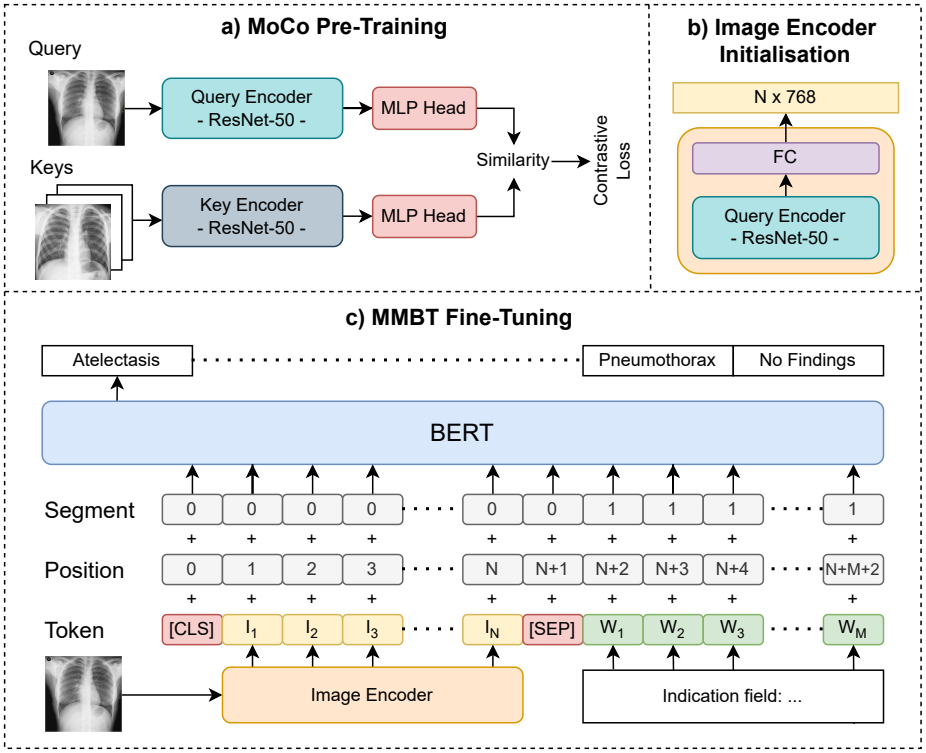

Fig. 2. Illustration of the proposed three step pipeline: a) MoCo pre-training of ResNet-50; b) initialisation of ResNet-50 using the query-encoder weights; and c) fine-tuning of MMBT on the 14 classes in MIMIC-CXR.

tokens, indicated as $W = \{W_1 \dots W_M\}$. A $[CLS]$ token is used at the beginning of the input sequence, and its final hidden vector is used as the multimodal input sequence representation for classification. The $[SEP]$ token is used to separate the two input modalities.

2.2 Self-supervised Image Pre-training

We experiment with two self-supervised strategies: an Autoencoder (AE) and Momentum Contrast (MoCo).

The **AE** consists of a ResNet-50 encoder and decoder. The model is trained by minimising the reconstruction loss, defined as the mean squared error between the input and the reconstructed image. Following pre-training, the decoder is discarded and the ResNet-50 encoder weights are used as initialization for the MMBT image encoder.

As shown in Figure 2a, **MoCo** employs two ResNet-50 models – a query encoder and a key encoder – each followed by a MultiLayer Perceptron (MLP) projection head composed by two fully connected layers. The model is then

trained by optimising the Info Noise Contrastive Estimation (InfoNCE) loss function [28]. Due to the large dictionary, training of the key encoder through backpropagation is computationally intractable; instead, network parameters are updated using momentum updates in tandem with the query encoder. Following pre-training, the weights of the query encoder (without the MLP head) are used to initialise the ResNet-50 image encoder, shown in Figure 2b.

3 Experiments & Results

3.1 Experimental Setup

Dataset: We evaluated our method on MIMIC-CXR [13]–[15], which contains 377,110 CXR images with the associated radiology reports from 65,379 patients.

Using the CheXpert labeler [29], 14 different labels have been automatically extracted from the radiology report: *Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices*. The CheXpert labeler assigns a value of whether the label has a *positive*, *negative*, or *uncertain* mention in the report, or is not discussed (*missing*). For each label, we re-formulate the task as a multi-label binary classification task: *positive* vs. *others* (*negative*, *uncertain*, *missing*).

In this study, we select only images from a frontal view, either anteroposterior (AP) and posteroanterior (PA). Following the official MIMIC-CXR split, this yields 208,794 training pairs, 1,695 validation pairs and 2,920 test report/image pairs. As presented in [3], the text modality corresponds to the indication field (*i.e.* scan request text) extracted from the radiology reports. This is the part that would be available at imaging time and describes relevant medical history.

The self-supervised pre-training of the image encoder is performed on the CheXpert dataset [29] which consists of 224,316 CXR images from 65,240 patients; we ignore the available annotations and treat this dataset as a large unlabelled dataset. Input images are resized by matching the smaller edge to 224 pixels and maintaining the original aspect ratio.

Model Implementation & Training: For the self-supervised pre-training, we adopt the AE and MoCo implementations available from the PyTorch Lightning library[2]. During pre-training, the input images are resized by matching the smaller edge to 224 pixels and maintaining the original aspect ratio. Similar to Sowrirajan et al [20], we employ the following data augmentation techniques: random rotation ($-10° \leq \theta \leq 10°$), random horizontal flipping; and random crop of 224×224 pixels. The same data augmentations are also applied during the fine-tuning step.

At the fine-tuning stage, we adopt the MMBT implementation made available by the authors of ChestBERT [3][3], which uses the MultiModal Framework

[2] `https://pytorch-lightning-bolts.readthedocs.io/en/latest/self_supervised_models.html`

[3] `https://github.com/jacenkow/mmbt`

(MMF) [30]. We use the same training parameters as [3]: models are trained using a batch size of 128 and Adam optimiser with weight decay, with the learning rate set to 5×10^{-5}, and a linear warm-up schedule for the first 2000 steps, and micro F1 score computed on the validation set was used as the early stopping criterion and a patience of 4000 steps, up to a maximum of 14 epochs. Each experiment was repeated 5 times using different random seeds to initialise the model weights and randomise batch shuffling.

Baselines: The chosen method is compared with two unimodal baselines, to verify the improvement brought by inputting both visual and textual modalities at once. Moreover, we compare MMBT with another multimodal approach which we denote "Attentive" [17], to justify the architecture design chosen for our multimodal experiments.

- **BERT** [27] - using a BERT model only (similar to the backbone of MMBT) a unimodal text classifier is trained, without the CXR image.
- **ResNet-50** [31] - using ResNet-50 only (similar to the network used for the image representation in MMBT) a unimodal image classifier is trained, without text information.
- **Attentive** [17] - this model follows a two stream approach where a) the CXR image is processed by a ResNet-50 model and b) the indication field is encoded by BioWordVec embeddings [32] followed by two sequential bi-directional Gated Recurrent Units (GRUs) [33]. The visual and textual feature representations are then fused using two multimodal attention layers.

Metrics & Experiments: We report the F1 score and the Area Under the Receiver Operating Characteristic (AUROC), multiplying all metrics by 100 for ease of reading. To assess whether a pre-training strategy helps in a limited training data scenario, the same experiments are conducted using only a 10% random sample of the original training set.

3.2 Comparison of Self-Supervised Pre-training Strategies

Here we compare MMBT with the baselines, adopting different pre-training strategies for the image encoder, as described in Section 2.2. The AE and MoCo pre-trained ResNet-50 are compared against: (1) random initialization – to verify the benefit of starting from pre-trained weights; (2) ImageNet initialization – widely adopted in computer vision.

Results: As shown in Table 2 (top), both unimodal baselines (text-only BERT and image-only ResNet-50) obtain lower classification scores compared to the multimodal approaches (Attentive and MMBT); with MMBT achieving the best results, as previously reported in [3]. However, in the limited data scenario (Table 2 (bottom)), the gap between unimodal and multimodal approaches is reduced when considering the standard ImageNet initialization. This suggests that the image modality is not processed effectively by the multimodal architectures,

Table 2. Results on the MIMIC-CXR test set, comparing different ResNet-50 pre-training strategies. The models are fine-tuned on the full training set (top) and on 10% of the training set (bottom).

100% Training Set

Model	Image Pre-Training		F1		AUROC	
	Method	Dataset	Macro	Micro	Macro	Micro
BERT	-		$24.4_{\pm 1.0}$	$40.2_{\pm 0.5}$	$71.5_{\pm 0.3}$	$82.1_{\pm 0.4}$
ResNet-50	Supervised	ImageNet	$27.2_{\pm 0.6}$	$48.4_{\pm 0.9}$	$75.8_{\pm 1.2}$	$85.3_{\pm 0.8}$
ResNet-50	MoCo	CheXpert	$\mathbf{28.5_{\pm 0.7}}$	$\mathbf{49.5_{\pm 0.6}}$	$\mathbf{76.3_{\pm 0.3}}$	$\mathbf{85.5_{\pm 0.1}}$
Attentive	Supervised	ImageNet	$29.3_{\pm 0.5}$	$51.1_{\pm 0.6}$	$76.3_{\pm 0.5}$	$85.9_{\pm 0.5}$
Attentive	MoCo	CheXpert	$\mathbf{31.9_{\pm 0.3}}$	$\mathbf{53.2_{\pm 0.5}}$	$\mathbf{77.8_{\pm 0.6}}$	$\mathbf{86.4_{\pm 0.4}}$
MMBT	*Random Initialization*		$32.0_{\pm 1.2}$	$49.7_{\pm 0.7}$	$76.1_{\pm 0.3}$	$85.2_{\pm 0.3}$
MMBT	Supervised	ImageNet	$34.3_{\pm 2.1}$	$54.7_{\pm 0.7}$	$79.8_{\pm 1.1}$	$87.4_{\pm 0.8}$
MMBT	AE	CheXpert	$34.5_{\pm 1.2}$	$52.4_{\pm 0.3}$	$77.9_{\pm 0.4}$	$86.3_{\pm 0.4}$
MMBT	MoCo	CheXpert	$\mathbf{36.7_{\pm 1.4}}$	$\mathbf{55.3_{\pm 0.6}}$	$\mathbf{80.4_{\pm 0.3}}$	$\mathbf{87.6_{\pm 0.4}}$

10% Training Set

Model	Image Pre-Training		F1		AUROC	
	Method	Dataset	Macro	Micro	Macro	Micro
BERT	-		$21.3_{\pm 2.7}$	$36.6_{\pm 1.7}$	$67.4_{\pm 0.4}$	$79.7_{\pm 1.3}$
ResNet-50	Supervised	ImageNet	$22.1_{\pm 0.9}$	$42.1_{\pm 0.7}$	$68.0_{\pm 1.9}$	$79.7_{\pm 3.4}$
ResNet-50	MoCo	CheXpert	$\mathbf{23.6_{\pm 1.1}}$	$\mathbf{43.8_{\pm 1.8}}$	$\mathbf{70.8_{\pm 0.9}}$	$\mathbf{81.3_{\pm 0.9}}$
Attentive	Supervised	ImageNet	$21.7_{\pm 0.9}$	$42.1_{\pm 1.4}$	$65.1_{\pm 1.1}$	$78.9_{\pm 0.6}$
Attentive	MoCo	CheXpert	$\mathbf{22.8_{\pm 1.0}}$	$\mathbf{44.3_{\pm 1.9}}$	$\mathbf{70.2_{\pm 0.5}}$	$\mathbf{82.7_{\pm 0.4}}$
MMBT	*Random Initialization*		$25.1_{\pm 2.1}$	$40.7_{\pm 3.0}$	$69.6_{\pm 0.7}$	$81.6_{\pm 0.6}$
MMBT	Supervised	ImageNet	$26.4_{\pm 2.1}$	$44.3_{\pm 1.5}$	$69.0_{\pm 0.4}$	$79.3_{\pm 1.8}$
MMBT	AE	CheXpert	$27.6_{\pm 1.2}$	$44.2_{\pm 1.1}$	$70.5_{\pm 0.4}$	$82.1_{\pm 0.3}$
MMBT	MoCo	CheXpert	$\mathbf{28.5_{\pm 2.4}}$	$\mathbf{48.8_{\pm 1.1}}$	$\mathbf{74.1_{\pm 0.7}}$	$\mathbf{84.5_{\pm 0.9}}$

which motivates us to investigate how to improve the image representations to maintain the benefit of using both modalities with limited data.

Table 2 shows a consistent improvement from adopting MoCo initialization of the image encoder (ResNet-50), which demonstrates that MMBT benefits from such domain-specific image pre-training strategy. The margin of improvement from ImageNet increases with a limited training set, aligned with the results in [20]. Compared to Sowrirajan et al [20] — who showed the benefit of MoCo pre-training only on pleural effusion classification, using an image-only CNN – we broaden the paradigm to multimodal classification of 14 different classes. Furthermore, we report the AUROC scores for each class in Table 4. This shows that MoCo pre-trained MMBT yields the highest scores for most classes, when fine-tuned on the full MIMIC-CXR training set, and more obviously when fine-tuned on a 10% random subset of the training set.

On the contrary, AE seems to be a less effective pre-training strategy. This might be attributed to the reconstruction loss, which encourages the model to

Table 3. IoU results computed on the ChestX-ray8 test set, containing bounding box annotations. We evaluate only on the five classes that overlap with MIMIC-CXR.

Image Pre-Training		IoU				
Method	Dataset	Atelectasis	Cardiomegaly	Effusion	Pneumonia	Pneumothorax
Supervised	ImageNet	1.3	3.8	5.9	0.0	0.1
MoCo	CheXpert	**2.6**	**16.0**	**11.9**	**1.8**	**2.7**

focus on the intensity variation of CXRs rather than other meaningful features (*e.g.* shapes and textures) to discriminate between different classes.

Table 2 shows a consistent improvement achieved by adopting MoCo pre-trained weights also for the image encoder of the Attentive model and the image-only ResNet-50. This confirms that both unimodal and multimodal models benefit from the MoCo pre-training of the image encoder.

3.3 Model Explainability

To investigate the impact of pre-training on the learned features, we visually assess the quality of the activation maps obtained by two of the pre-training strategies: supervised ImageNet pre-training and MoCo pre-training on CheXpert. First, we fine-tune the fully connected layer of the ResNet-50 architecture on the full training set of MIMIC-CXR, while freezing the remaining pre-trained weights. Second, we apply Grad-CAM [24] to the final 7×7 activation map, computed before the fully connected layer. Finally, we assess if the generated maps highlight the correct anatomical location of the pathology, by computing the Intersection over Union (IoU) between the bounding boxes – annotated in the ChestX-ray8 dataset [25] – and the regions in the activation map that contribute positively to the classification of a target label. In this final step, we only consider the subset of ChestX-ray8 labels overlapping with those in MIMIC-CXR: *Atelectasis, Cardiomegaly, Pleural Effusion, Pneumonia, Pneumothorax.*

The mean IoU scores for each class are reported in Table 3. Although the overlap between the positive areas of the activation maps and the bounding boxes is low for both pre-training strategies, it can be observed that MoCo pre-training outperforms ImageNet for each class. This suggests that, when adopting MoCo pre-training, the CNN learns more meaningful features of CXRs that can be effectively exploited by the model for the downstream classification task. This is shown visually in Figure 3, where MoCo pre-trained ResNet-50 focuses more accurately to the areas matching with the bounding boxes. However, both pre-training strategies frequently focus on incorrect areas in the images.

4 Conclusion

In this work we have demonstrated the benefit of domain-specific self-supervised MoCo pre-training of the MMBT image encoder for multimodal multi-label CXR classification. To the best of our knowledge, this is the first study to compare

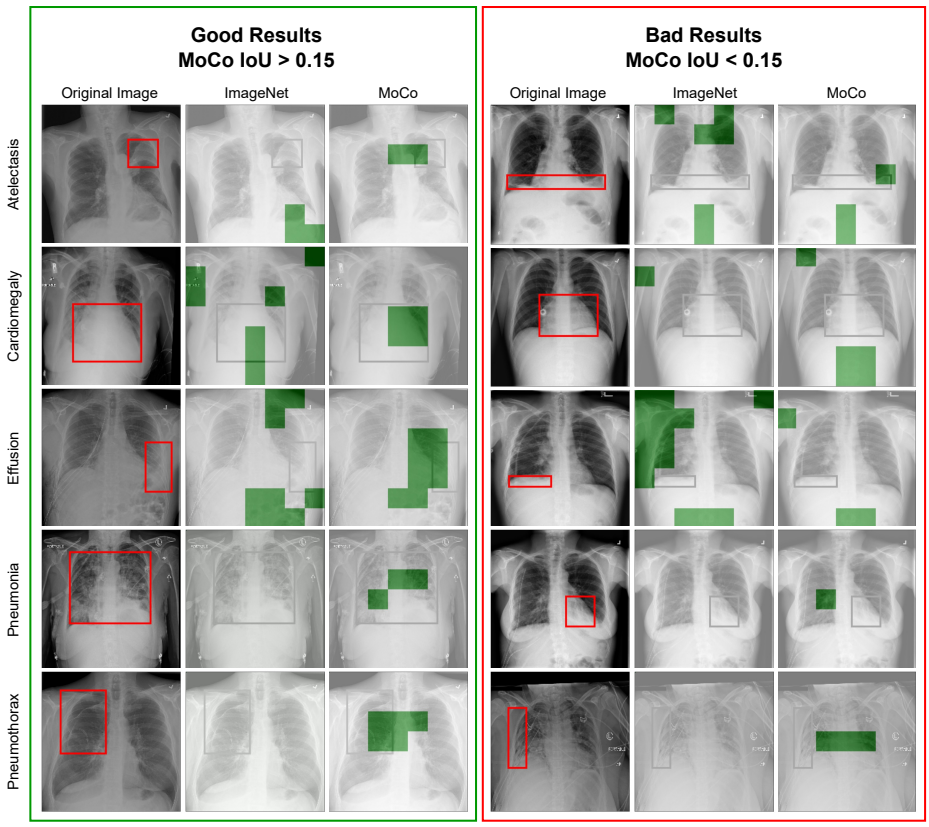

Fig. 3. Examples of CXRs taken from ChestX-ray8 dataset with the corresponding bounding box annotations highlighted in red. Grad-CAM is computed on the last 7×7 activation map, before the fully connected layer of ResNet-50, for both ImageNet and MoCo pre-training. The green regions show the activations thresholded at 0 i.e. all positive activations (activations can also be negative). The left side images are selected having an IoU score greater than 0.15 between the bounding box and the positive regions, using MoCo pre-trained weights; the right side images are selected with an IoU score lower than 0.15.

how different self-supervised pre-training strategies affect multimodal performance in the medical domain. Our results show that the choice of image encoder plays a substantial role, especially with limited annotated data, where ResNet pre-trained using MoCo achieves the best performances. In future research, it would be interesting to combine unsupervised unimodal pre-training, as demonstrated in this paper, followed by an unsupervised multimodal pre-training step, as demonstrated in [18], to see if a cumulative improvement could be obtained.

References

[1] S. C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *Digital Medicine*, no. 1, 2020.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[3] G. Jacenków, A. Q. O'Neil, and S. A. Tsaftaris, "Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers," *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*,

[4] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *Transactions of the Association for Computational Linguistics*, pp. 570–585, 2021.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[6] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[7] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, 2019.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, 2015.

[9] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, no. 1, pp. 32–73, 2017.

[10] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.

[11] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8821–8831.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-

jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[14] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, no. 1, 2019.

[15] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *circulation*, no. 23, e215–e220, 2000.

[16] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, no. 2, pp. 304–310, 2016.

[17] T. van Sonsbeek and M. Worring, "Towards automated diagnosis with attentive multi-modal learning using electronic health records and chest X-rays," in *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, T. Syeda-Mahmood, K. Drechsler, H. Greenspan, A. Madabhushi, A. Karargyris, M. G. Linguraru, C. Oyarzun Laura, R. Shekhar, S. Wesarg, M. Á. González Ballester, and M. Erdt, Eds., Cham: Springer International Publishing, 2020, pp. 106–114.

[18] R. Liao, D. Moyer, M. Cha, K. Quigley, S. Berkowitz, S. Horng, P. Golland, and W. M. Wells, "Multimodal representation learning via maximization of local mutual information," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 273–283.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[20] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "MoCo pretraining improves representation and transferability of chest X-ray models," in *Medical Imaging with Deep Learning*, PMLR, 2021, pp. 728–744.

[21] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, *et al.*, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3478–3488.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[23] Y. N. T. Vu, R. Wang, N. Balachandar, C. Liu, A. Y. Ng, and P. Rajpurkar, "Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation," in *Machine Learning for Healthcare Conference*, PMLR, 2021, pp. 755–769.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[25] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[26] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[27] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[28] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[29] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 590–597.

[30] A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh, *MMF: A multimodal framework for vision and language research*, `https://github.com/facebookresearch/mmf`, 2020.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] Z. Yijia, Q. Chen, Z. Yang, H. Lin, and Z. lu, "BioWordVec, improving biomedical word embeddings with subword information and MeSH," *Scientific Data*, 2019.

[33] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," English (US), in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

A Per-Class Results

Table 4. Per-class AUROC scores using different ResNet-50 initializations. The models are fine-tuned on the full training set (top) and on 10% of the training set (bottom).

100% Training Set

Model	Attentive		MMBT			
Image Pre-Training	Supervised ImageNet	MoCo CheXpert	Random Init.	Supervised ImageNet	AE CheXpert	MoCo CheXpert
Atelectasis	$73.5_{\pm 1.0}$	$72.8_{\pm 0.9}$	$71.8_{\pm 0.6}$	$\mathbf{75.2_{\pm 0.8}}$	$74.2_{\pm 0.7}$	$74.9_{\pm 0.7}$
Cardiomegaly	$77.1_{\pm 0.5}$	$79.1_{\pm 0.5}$	$80.0_{\pm 0.4}$	$81.3_{\pm 0.9}$	$81.6_{\pm 0.4}$	$\mathbf{82.4_{\pm 0.4}}$
Consolidation	$72.3_{\pm 1.1}$	$75.0_{\pm 0.6}$	$71.3_{\pm 0.8}$	$\mathbf{77.2_{\pm 1.8}}$	$74.3_{\pm 0.6}$	$76.5_{\pm 0.6}$
Edema	$82.2_{\pm 0.7}$	$82.8_{\pm 0.5}$	$80.9_{\pm 0.8}$	$83.6_{\pm 1.1}$	$82.7_{\pm 0.4}$	$\mathbf{84.2_{\pm 0.4}}$
Enlarged Card.	$67.0_{\pm 1.6}$	$68.7_{\pm 0.9}$	$68.4_{\pm 0.6}$	$73.3_{\pm 2.1}$	$71.4_{\pm 1.8}$	$\mathbf{75.0_{\pm 1.8}}$
Fracture	$66.7_{\pm 3.0}$	$69.7_{\pm 1.6}$	$68.7_{\pm 1.8}$	$70.0_{\pm 1.2}$	$70.7_{\pm 0.8}$	$\mathbf{72.3_{\pm 0.8}}$
Lung Lesion	$68.9_{\pm 2.2}$	$71.0_{\pm 0.9}$	$69.6_{\pm 0.6}$	$74.5_{\pm 3.0}$	$70.2_{\pm 0.6}$	$\mathbf{76.7_{\pm 0.6}}$
Lung Opacity	$68.9_{\pm 0.4}$	$70.8_{\pm 0.9}$	$66.9_{\pm 0.6}$	$71.8_{\pm 0.5}$	$69.4_{\pm 0.5}$	$\mathbf{72.0_{\pm 0.5}}$
No Findings	$80.4_{\pm 0.4}$	$80.9_{\pm 0.9}$	$79.7_{\pm 0.8}$	$82.5_{\pm 1.2}$	$81.2_{\pm 0.6}$	$\mathbf{82.6_{\pm 0.6}}$
Pleural Effusion	$86.7_{\pm 0.9}$	$86.8_{\pm 0.6}$	$82.6_{\pm 0.3}$	$\mathbf{87.6_{\pm 0.6}}$	$85.0_{\pm 0.2}$	$\mathbf{87.6_{\pm 0.2}}$
Pleural Other	$78.8_{\pm 1.7}$	$80.2_{\pm 2.3}$	$89.4_{\pm 1.2}$	$\mathbf{86.1_{\pm 4.4}}$	$81.6_{\pm 1.9}$	$\mathbf{86.1_{\pm 1.9}}$
Pneumonia	$70.8_{\pm 0.9}$	$74.1_{\pm 1.5}$	$69.7_{\pm 0.9}$	$74.6_{\pm 0.6}$	$71.8_{\pm 0.6}$	$\mathbf{76.7_{\pm 0.6}}$
Pneumothorax	$84.8_{\pm 0.8}$	$86.9_{\pm 0.9}$	$87.4_{\pm 1.2}$	$\mathbf{87.9_{\pm 0.4}}$	$86.9_{\pm 1.0}$	$87.7_{\pm 1.0}$
Support Devices	$90.4_{\pm 0.2}$	$91.1_{\pm 0.2}$	$89.3_{\pm 0.2}$	$\mathbf{91.7_{\pm 0.6}}$	$90.1_{\pm 0.3}$	$\mathbf{91.7_{\pm 0.3}}$
Average	$76.3_{\pm 0.5}$	$77.8_{\pm 0.6}$	$76.1_{\pm 0.3}$	$79.8_{\pm 1.1}$	$77.9_{\pm 0.4}$	$\mathbf{80.4_{\pm 0.3}}$

10% Training Set

Model	Attentive		MMBT			
Image Pre-Training	Supervised ImageNet	MoCo CheXpert	Random Init.	Supervised ImageNet	AE CheXpert	MoCo CheXpert
Atelectasis	$66.9_{\pm 1.5}$	$69.3_{\pm 1.3}$	$64.6_{\pm 0.4}$	$65.5_{\pm 0.9}$	$67.2_{\pm 0.4}$	$\mathbf{71.4_{\pm 1.3}}$
Cardiomegaly	$67.3_{\pm 0.5}$	$72.4_{\pm 0.8}$	$71.8_{\pm 0.5}$	$70.7_{\pm 0.7}$	$74.0_{\pm 1.3}$	$\mathbf{77.0_{\pm 0.9}}$
Consolidation	$61.3_{\pm 0.3}$	$68.0_{\pm 0.8}$	$66.3_{\pm 1.1}$	$64.0_{\pm 1.3}$	$67.7_{\pm 0.8}$	$\mathbf{71.3_{\pm 1.0}}$
Edema	$76.1_{\pm 1.0}$	$78.4_{\pm 1.4}$	$74.5_{\pm 0.6}$	$76.5_{\pm 1.1}$	$77.1_{\pm 0.8}$	$\mathbf{80.7_{\pm 1.3}}$
Enlarged Card.	$58.5_{\pm 2.5}$	$63.3_{\pm 1.8}$	$62.6_{\pm 3.3}$	$62.8_{\pm 1.8}$	$61.5_{\pm 5.0}$	$\mathbf{67.8_{\pm 3.0}}$
Fracture	$52.2_{\pm 3.4}$	$51.8_{\pm 3.0}$	$61.6_{\pm 4.5}$	$58.7_{\pm 4.0}$	$60.1_{\pm 2.4}$	$\mathbf{62.4_{\pm 2.6}}$
Lung Lesion	$56.7_{\pm 1.0}$	$64.5_{\pm 2.9}$	$64.8_{\pm 2.2}$	$60.7_{\pm 2.0}$	$65.8_{\pm 2.0}$	$\mathbf{67.2_{\pm 1.4}}$
Lung Opacity	$60.4_{\pm 0.7}$	$65.7_{\pm 0.6}$	$61.7_{\pm 0.8}$	$62.2_{\pm 1.7}$	$62.2_{\pm 0.8}$	$\mathbf{67.2_{\pm 1.1}}$
No Findings	$72.1_{\pm 1.4}$	$75.2_{\pm 1.2}$	$74.5_{\pm 0.8}$	$74.1_{\pm 0.6}$	$75.8_{\pm 0.9}$	$\mathbf{78.7_{\pm 0.6}}$
Pleural Effusion	$80.0_{\pm 0.9}$	$82.6_{\pm 0.7}$	$73.0_{\pm 0.8}$	$79.2_{\pm 0.9}$	$76.8_{\pm 0.4}$	$\mathbf{84.7_{\pm 0.3}}$
Pleural Other	$60.0_{\pm 1.3}$	$62.2_{\pm 3.6}$	$61.2_{\pm 1.2}$	$65.0_{\pm 6.7}$	$62.1_{\pm 3.2}$	$\mathbf{67.6_{\pm 3.3}}$
Pneumonia	$57.6_{\pm 1.5}$	$64.1_{\pm 1.3}$	$66.4_{\pm 1.5}$	$60.4_{\pm 1.4}$	$66.0_{\pm 0.7}$	$\mathbf{68.6_{\pm 2.0}}$
Pneumothorax	$68.4_{\pm 3.7}$	$78.7_{\pm 2.8}$	$84.6_{\pm 2.0}$	$79.9_{\pm 2.0}$	$\mathbf{85.0_{\pm 0.6}}$	$84.7_{\pm 0.8}$
Support Devices	$78.0_{\pm 1.9}$	$86.5_{\pm 2.0}$	$87.1_{\pm 0.6}$	$86.0_{\pm 0.9}$	$86.9_{\pm 0.5}$	$\mathbf{88.8_{\pm 1.2}}$
Average	$65.1_{\pm 1.1}$	$70.2_{\pm 0.5}$	$69.6_{\pm 0.7}$	$69.0_{\pm 0.4}$	$70.5_{\pm 0.4}$	$\mathbf{74.1_{\pm 0.7}}$