Full Title of Article

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Although human's ability to visually understand the structure of the World plays a crucial role in perceiving the World and making appropriate decisions, human perception does not solely rely on vision but amalgamates the information from acoustic, verbal, and visual stimuli. An active area of research has been revolving around designing an efficient framework that adapts to multiple modalities and ideally improves the performance of existing tasks. While numerous frameworks have proved effective on natural datasets like ImageNet, a limited number of studies have been carried out in the biomedical domain. In this work, we extend the available frameworks for natural data to biomedical data by leveraging the abundant, unstructured multi-modal data available as radiology images and reports. We attempt to answer the question, "Does multi-modal learning outperform the self-supervised learning approaches, when our downstream/transfer learning task has only images as an input?". Our experiments indicated that pre-training with a relevant large dataset outperforms the models that are initialized randomly or initialized with ImageNet pre-trained weights. Moreover, in limited labeled data settings, with 1% labeled data, a combination of self-supervised and multi-modal methods consistently performed better with a 2%-5% increase in the AUROC. The code is publicly available online [1].

Keywords: Multi-Modal Learning, Contrastive Learning, Radiology

1. Introduction

Multi-Modal Learning has become an extremely popular area of research in Deep Learning. While a good amount of Multi-Modal research has been conducted on natural images dataset (Radford et al., 2021; Kim et al., 2021; Bardes et al., 2022a; He et al., 2020; Chen et al., 2020a; Grill et al., 2020; Zbontar et al., 2021; Bao et al., 2022; Li et al., 2021; Bardes et al., 2022b; Mu et al., 2022; Caron et al., 2020), very few works have been published in the biomedical domain (Boecking et al., 2022; Tiu et al., 2022; Zhang et al., 2020; Chaitanya et al., 2020; Haghighi et al., 2022; Xiao et al., 2022; Azizi et al., 2021) and the choice between Multi-Modal Learning and Self-Supervised Learning is still unclear, if our ultimate goal is better performance on image tasks. Hence, it is important to compare the performances of Multi-Modal Learning and Self-Supervised Learning in biomedical domain because Multi-Modal possesses a huge potential to automate the diagnosis systems by achieving expert-level performance by leveraging the additional available information (paired clinical reports).

Biomedical data can effortlessly take advantage of Multi-Modal Learning mainly because it occurs Multi-Modal by nature. For example, there is a paired report for a given X-ray scan and the scans can be from multiple contrasts. While many Multi-Modal models

1. Link will be available in camera-ready version.

© 2023 Anonymous.

can be used off-the-shelf (Jaegle et al., 2021; Mu et al., 2022; Radford et al., 2021), they are usually trained on natural image datasets which are ten orders magnitude higher in number than the biomedical data that we can collect with batch size scaled across dozens to hundreds of GPUs/TPUs and their performance evaluation usually does not include any biomedical task benchmarks. While some works have been done on Biomedical Multi-Modal Learning recently (Tiu et al., 2022; Boecking et al., 2022), they mainly focus on evaluating the performance with Mutli-Modal input. To our knowledge, no specific research has been conducted on the comparative studies of Multi-Modal Learning vs. Self-Supervised Learning on the medical image tasks only. Our contributions are summarized as follows:

- investigated whether the existing frameworks can improve vision representation from limited Multi-Modal biomedical data.
- investigated if Multi-Modal Learning is necessary when we can directly perform Self-Supervised Learning for medical images.
- provided a robust benchmark for contrastive learning based self-supervised learning vs. multi-modal learning for Chest X-Ray with code available for further research investigation.

2. Related Works

Recent research on Multi-Modal Learning can be classified into two main categories. One is two-stream frameworks with representative works to be CLIP (Radford et al., 2021) and ViLBERT (Lu et al., 2019). VilBERT passes image and text input separately into two transformer architecture with cross attention in-between the features from two architectures to perform information fusion between two modalities. Another is single-stream frameworks with representative works to be ViLT (Kim et al., 2021). ViLT passes image and text input jointly into a transformer architecture and attempts to minimize Image Text Matching loss by classification, Masked Language Modeling by origin proposal from BERT (Devlin et al., 2019), Word Patch Alignment by optimal transport with inexact proximal point method concurrently.

Two-stream approach means the image and text are input into two separate encoders for Model Learning, where the image and text encoder usually can be any arbitrary architectures. In contrast, one-stream approach means that the image and the text are input into single model, which is usually fixed to be a transformer architecture because the attention mechanism has been proved to be effective on learning from different modalities.

Since self-supervised learning is closely aligned with Multi-Modal Learning, we used it as one of our benchmarks, three of its representative works, SimCLR (Chen et al., 2020a), MOCOV2 (Chen et al., 2020b) and VICREG (Bardes et al., 2022a) are also discussed in section 3. Recently, CheXzero (Tiu et al., 2022) explored the performance of CLIP on radiology chest scan diagnosis by taking in language and vision input together with zero-shot prompting. BioVLP (Boecking et al., 2022) explored the performance of contrastive self-supervised learning with masked language modeling on radiology chest scan diagnosis.

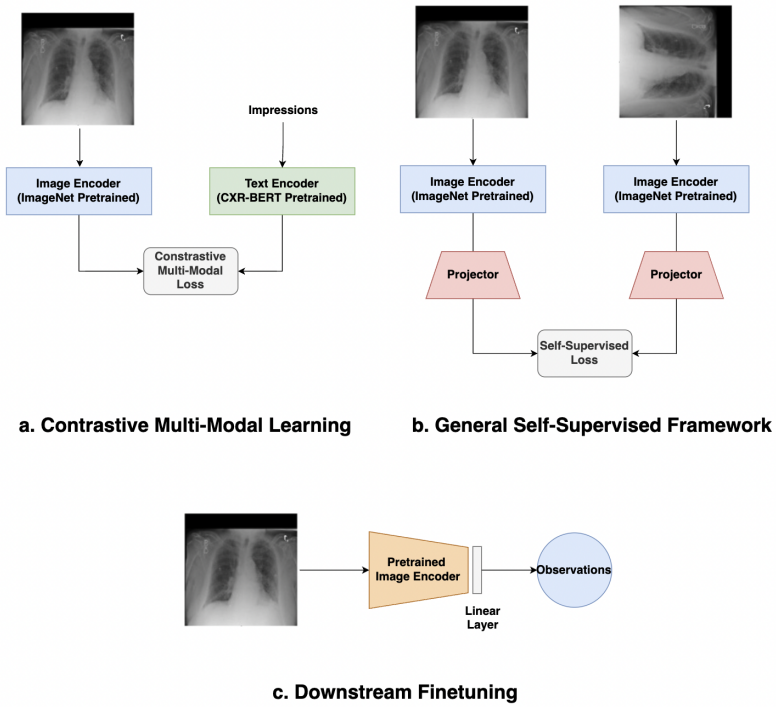


Figure 1: Experimental Process: a) is used for our Multi-Modal Learning training. b) is used for our Self-Supervised Learning training. c) is used for our downstream finetuning. For SLIP, a) and b) are combined during the training.

To the best of our knowledge, our work is the first in the literature to perform a large scale benchmark study for Multi-Modal chest radiographs and compare it with existing self-supervised learning approaches and a combination of self-supervised and Multi-Modal approach.

3. Methods

3.1. Vision-Language Multi-Modal and Vision Self-Supervised Pretraining

We chose CLIP as a benchmarking framework as it is a two-stream approach, which means the two encoders can be set to architectures that are computationally less expensive and faster compute time. For evaluating the pre-trained model quality, we finetune the pre-trained model with additional classification heads. The details are described below, and the diagram of the experimental process is shown in Figure 1

CLIP (Radford et al., 2021): Given images representations u and text representation v, CLIP loss is formulated as

$$\mathcal{L} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} \tag{1}$$

where

$$\mathcal{L}_{i2t} = -\sum_{i\in\mathcal{B}} \log \frac{\exp(\tau u_i v_i)}{\sum_{j\in\mathcal{B}} \exp(\tau u_i v_j)} \tag{2}$$

and

$$\mathcal{L}_{t2i} = -\sum_{j\in\mathcal{B}} \log \frac{\exp(\tau u_i v_i)}{\sum_{i\in\mathcal{B}} \exp(\tau u_i v_j)} \tag{3}$$

where $\mathcal{B}$ is batch size and τ is temperature hyperparameter .

ConVIRT (Zhang et al., 2020): ConVIRT followed the same loss as CLIP, but it is more fine-grained toward medical imaging with more specific augmentations and a design to freeze top six layers of text encoder.

SimCLR (Chen et al., 2020a): Given two input representations z_i and z_j in a given batch. The NT-Xent loss by SimCLR is formulate as

$$l_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k\neq i} \exp(sim(z_i, z_k)/\tau)} \tag{4}$$

where τ is temperature hyperparameter, N is number of in batch samples, sim is the similarity metric set to be dot product following the original implementation.

VICREG (Bardes et al., 2022a): Given two input representations Z and Z' obtained by passing two transformed views of same image, VICREG loss is computed as

$$\ell(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')] \tag{5}$$

where $s(Z, Z') = \frac{1}{n}\sum_i \|z_i - z_i'\|_2^2$ is the invariance term for calculating the feature-wised mean-squared euclidean distance with given feature size n. $v(Z) = \frac{1}{d}\sum_{j=-1}^{d} \max(0, \gamma - S(z^j, \epsilon))$ is the variance term for given number of samples d, $S(x, \epsilon) = \sqrt{\mathrm{Var}(x) + \epsilon}$ is the regularized standard deviation with γ as a constant target standard deviation value, ϵ as a scalar to enforce numerical stability. $C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T$ is the covariance term where $\bar{z} = \frac{1}{n}z_i$. λ, μ, ν are hyperparameters to control the contribution of each loss term.

MOCOV2 (Chen et al., 2020b): following original proposal of MOCO that creating a very large memory bank to cache image representations with InfoNCE (van den Oord et al., 2018) in order to introduce more negative samples within single batch. The InfoNCE loss is formulated as

$$\mathcal{L} = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)} \tag{6}$$

where q is the representation for anchor input, k_+ is the representation for positive input, k_i is the representation for all cached input in memory bank and τ is a temperature parameter. The sample representations in the memory bank is gradually updated by a separate encoder that is update by trained encoder with Exponential Moving Average (EMA).

SLIP (Mu et al., 2022): SLIP is trained by jointly minimizing the loss of CLIP on image-text domain and an arbitrary self-supervised learning method on image domain. . Given image and text representations z_i and z_t from CLIP projector, and two projected augmented image representations from some self-supervised learning projector z_1 and z_2, SLIP is trained by minimizing

$$\ell_{slip} = c \cdot \ell_{ssl}(z_1, z_2) + \ell_{clip}(z_i, z_t) \tag{7}$$

where ℓ_{ssl} is the loss for some self-supervised method and c is a hyperparameter to control the contribution of self-supervised loss. In our experiment, we choose the self-supervised learning methods to be SimCLR and VICREG. While SimCLR is experimented in the original SLIP paper, VICREG is not attempted in the original work.

3.2. Downstream Finetuning

Given the pretrained image encoder models from either Self-Supervised or Multi-Modal Learning, the performance of pretrained models are evaluated in three ways. First, finetuning the model with one additional linear layer as classification head on our chosen datasets. Second we use linear probing, in which our pretrained image encoder is frozen and only the final linear layer is trained. Third, we also experiment in a semi-supervised manner, in which we take only 1% and 10% labels of the datasets mentioned in 4.1.

4. Experimental Setup

4.1. Dataset

Our Mutli-Modal frameworks are all trained on MIMIC-CXR (Johnson et al., 2019) dataset, which has 377,110 pairs of a chest X-ray image-report collected from Beth Israel Deaconess Medical Center in Boston, MA. Radiology reports, although an excellent source of information, has additional information. On the other hand, Findings and Impression carry are much more rich in content. Findings present the diagnosis from the given radiology image and Impression presents the complete narrative description from doctors according to multiple radiology images of the same patient. An example of MIMIC-CXR sample with corresponding Findings and Impression can be seen in Figure 2 Our trained models are evaluated on CheXPert (Irvin et al., 2019) dataset with linear probing, fine-tuning, and semi-supervised form. CheXpert dataset consisting of 224,316 chest radiographs of 65,240 patients from Stanford Hospital in Palo Alto, CA. There are 14 observed pathologies labeled by radiologist as positive, negative, or uncertain. For our evaluation, we consider two cases. First, all pathologies are considered. Second, the five pathologies chosen in the orginal paper based on their relevance in the original dataset. Namely, Cardiomegaly, Edema, Consolidation, Atelectasis and Pleural Effusion.

4.2. Backbone Encoders

For all frameworks, we fix the image encoder to be ResNet-50 (He et al., 2016). We experiment with other architectures including DenseNet-121 (Huang et al., 2017) and ViT-base/16 (Dosovitskiy et al., 2021). However, we do not include the results of DenseNet-121 (Huang

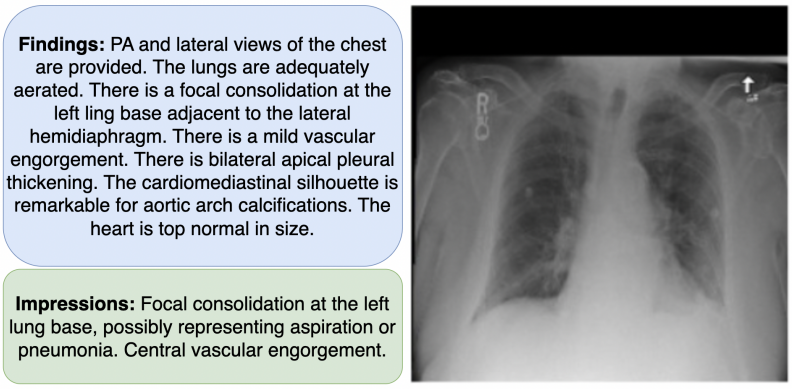


Figure 2: MIMIC-CXR example with corresponding Radiology graph vs. text pair.

et al., 2017) and ViT-base/16 in the paper, we added them for information purposes in 6. For text encoder, we use BERT-base (Devlin et al., 2019). The image encoder is initialized with ImageNet pretrained weights and the text encoder is initialized with CXR-BERT-general pretrained weights (Boecking et al., 2022).

4.3. Feature Projection Details

The image embeddings is originally projected to 2048 dimension and then further projected to 512 dimension for calculating CLIP loss. The text embeddings is originally extracted from last layer of $[CLS]$ token and then projected to 512 dimension as well. For SimCLR and MOCOV2, the image representations are re-projected to 512 dimension and then projected to 128 dimensions. For VICREG, the image representations are re-projected to 8192 dimensions with three linear layers. We used batch normalization (Ioffe and Szegedy, 2015) for all projectors as in-between layers.

4.4. Data Preprocessing

We directly load MIMIC-CXR Radiology images from corresponding DICOM files and the images are resized from varied thousands by thousands dimension to 224×224. Only Impression section is used as text input. Samples with Impressions missing are dropped because Impressions contain most detailed descriptions of the radiology.

4.5. Hyperparameters

We train all methods with learning rate $1e$-4, batch size 256 when backbone is ResNet-50 or batch size 128 when backbone is ViT-base/16 distributed across 4 GPUs (because of RAM limit of GPUs), AdamW optimizer (Loshchilov and Hutter, 2019), cosine scheduling as optimizer scheduler (Loshchilov and Hutter, 2016) , beta2 of AdamW is set to be 0.98 for more stable training, the total epochs of training is set to be 50 with early stopping evaluating on the total loss of validation set. For SLIP, we used same settings as CLIP except that the weight decay is set to be 0.1 while the weight decay of CLIP is 0.5. For fairness of comparisons, we set image augmentations for self-supervised learning to be the

same as Random Resize Crop, Random Horizontal Flip, Color Jitterring, Gaussian Blur and Solarization while the image augmentation for multi-modal learning to be Random Resized Crop, Random Horizontal Flip, Random Affine, Color Jitter and Gaussian Blur. We did not perform any augmentation on text data. See Appendix 6 for detail hyperparameters for each augmentation.

4.6. Computational Cost

We train all models on four V100 GPUs with 32GB RAM for MIMIC-CXR pre-training. For fine-tuning on CheXpert and ChestXRay-14 dataset, we used one V100 GPUs with 16GB RAM. The pre-training takes roughly 1.5 days and fine-tuning takes roughly 0.5 days.

5. Results

We show our result for ResNet-50 and ViT-b16 with benchmark methods on Table 2, 3, 4 and 5. All results are evaluated with AUROC for corresponding pathology. From the table, we can see that even though those more recent methods such as VICREG and MOCOV2 claimed to achieved better performance on ImageNet and some other natural image dataset than SimCLR under the case of small batch size, they are not performing better from our experiments on Chest X-Ray dataset. One of the reason we think this may happen is that those natural images often have very distinct objects with the objects take largest portion of the images. With this kind of image structure, those self-supervised learning algorithms may require large batch size to learn the global distinction of images from the extracted features by getting very large negative samples across batch such as SimCLR. However, in the case of medical imaging, each image is globally very similar while we only focus on small region of the image for pathology diagnostic, which indicate that very large negative samples may not help that much in the case of globally similar images.

Another finding is that we see combining Multi-Modal and Self-Supervised Loss proposed by SLIP does help on performance in case when ResNet-50 is used as backbone model. However, we do not observe improvement on ViT-b16. This contradicts the findings from the original paper, which observed improvement with both ResNet and ViT. However, because ViT is commonly known to be unstable to train, it can also be the case that we did not set the hyperparameters to their optimum.

From Figure **??**, **??** and **??**, we show GradCam (Selvaraju et al., 2017) Visualization for some models we trained for three ground-truth pathology Lung Opacity, Atelectasis and Pleural Effusion. From the heatmap, we can see that all models can identify correct position of pathology while models pretrained with SimCLR or SLIP-SimCLR be slightly more accurate in some cases.

6. Conclusion

Acknowledgments

Acknowledgments withheld.

References

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3458–3468, 2021. doi: 10.1109/ICCV48922.2021.00346.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=bydKs84JEyw.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *NIPS*. OpenReview.net, 2022b.

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

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020b. URL https://arxiv.org/abs/2003.04297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20792–20802, 2022. doi: 10.1109/CVPR52688.2022.02016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. *CoRR*, abs/2103.03206, 2021. URL https://arxiv.org/abs/2103.03206.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, Dec 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21k.html.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=0JLaKwiXSbx.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. URL http://arxiv.org/abs/1608.03983.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, pages 13–23, 2019. URL http://dblp.uni-trier.de/db/conf/nips/nips2019.html#LuBPL19.

Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, volume 13686 of *Lecture Notes in Computer Science*, pages 529–544. Springer, 2022. doi: 10.1007/978-3-031-19809-0_30. URL https://doi.org/10.1007/978-3-031-19809-0_30.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International*

Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, Sep 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00936-9. URL https://doi.org/10.1038/s41551-022-00936-9.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. *arXiv preprint arXiv:2210.12843*, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020. URL https://arxiv.org/abs/2010.00747.

Table 1: CheXPert: With Image Encoder for all Pathology (ResNet-50)

Method	Linear-Probing (%) ↑	Fine-Tuning (%) ↑
ImageNet Init.	-	82.29±2.798
CLIP	-	84.36±3.078
CONVIRT	-	85.20±1.441
SimCLR	-	84.74±0.964
VICREG	-	**85.40**±1.631
MOCOV2	-	84.12±1.085
SLIP-SimCLR-CLIP	-	84.33±0.798
SLIP-VICREG-CLIP	-	82.11±1.341
SLIP-MOCOV2-CLIP	-	82.92±1.239
SLIP-SimCLR-CONVIRT	-	84.39±1.095
SLIP-VICREG-CONVIRT	-	83.05±1.764
SLIP-MOCOV2-CONVIRT	-	83.51±1.376

Table 2: ChestXRay14: With Image Encoder for all Pathology (ResNet-50)

Method	Linear-Probing (%) ↑	Fine-Tuning (%) ↑
ImageNet Init.	-	76.58±1.230
CLIP	-	77.92±0.726
CONVIRT	-	78.94±0.652
SimCLR	-	77.62±0.992
VICREG	-	78.28±0.797
MOCOV2	-	78.50±0.297
SLIP-SimCLR-CLIP	-	78.14±1.513
SLIP-VICREG-CLIP	-	77.82±0.389
SLIP-MOCOV2-CLIP	-	78.56±0.802
SLIP-SimCLR-CONVIRT	-	78.42±0.750
SLIP-VICREG-CONVIRT	-	78.94±1.300
SLIP-MOCOV2-CONVIRT	-	**79.00**±0.420

Table 3: ChexPert: With Image Encoder and by Five More Important Pathology (ResNet-50)

Method	Cardiomegaly (%) ↑	Edema (%) ↑	Consolidation (%) ↑	Atelectasis (%) ↑	Pleural Effusion (%) ↑	Pneumonia (%) ↑	Average (%) ↑
ImageNet Init.	80.664.199	92.83±0.781	85.33±2.962	81.81±2.791	92.60±0.764	67.11±8.088	
CLIP	81.73±2.816	92.36±1.185	85.28±0.670	81.70±1.425	91.37±1.116	73.71±7.260	
SimCLR	84.09±1.760	92.61±1.168	**86.69**±1.554	83.35±2.112	92.80±7.865	78.14±4.122	
VICREG	**84.65**±2.293	**93.28**±0.258	85.18±0.730	**84.20**±1.288	**92.91**±0.876	72.34±7.092	
MOCOV2	82.14±3.005	93.05±0.550	86.41±2.044	82.37±2.336	92.35±0.824	79.54±2.155	
SLIP-SimCLR	83.73±3.065	92.80±0.850	86.36±1.649	83.05±2.224	92.74±0.696	79.82±2.756	
SLIP-VICREG	81.61±3.443	92.69±0.678	86.66±1.666	82.44±2.440	92.85±0.696	**79.93**±2.207	
SLIP-MOCOV2	79.63±2.585	91.63±0.704	86.50±1.631	81.65±3.297	92.26±0.404	71.22±4.164	

Table 4: CheXPert: With Image Encoder for all Pathology (ViT-base/16)

Method	Linear-Probing (%) ↑	Fine-Tuning (%) ↑
ImageNet Init.	72.35	78.61
CLIP	55.24	69.47
SimCLR	**75.77**	81.10
VICREG	75.67	79.85
MOCOV2	75.49	**81.14**
SLIP-SimCLR	72.11	78.33
SLIP-VICREG	72.49	79.26
SLIP-MOCOV2	73.15	78.56

Table 5: ChexPert: With Image Encoder and by Five More Important Pathology (ViT-base/16)

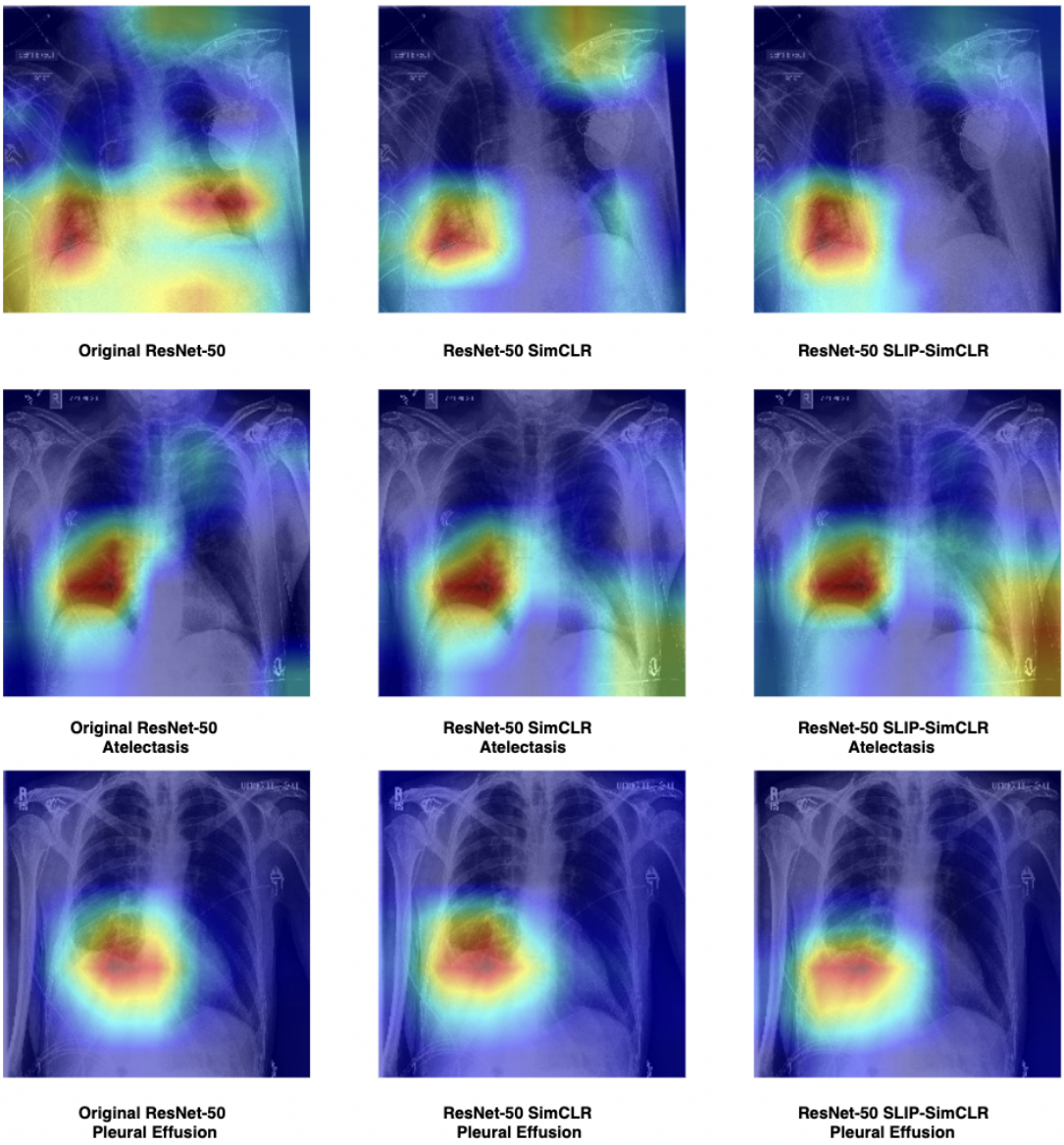
Method	Cardiomegaly (%) ↑	Edema (%) ↑	Consolidation (%) ↑	Atelectasis (%) ↑	Pleural Effusion (%) ↑	Average (%) ↑
ImageNet Init.	80.87	92.35	84.50	78.53	93.03	85.91
CLIP	74.86	88.33	81.51	77.56	86.90	81.83
SimCLR	80.88	91.98	82.85	79.20	**93.81**	85.74
VICREG	**83.20**	91.85	**89.98**	**84.34**	93.50	**88.37**
MOCOV2	82.73	91.76	88.34	81.75	94.70	87.86
SLIP-SimCLR	79.91	90.22	89.01	80.17	92.00	86.26
SLIP-VICREG	79.55	**92.77**	86.21	81.08	92.49	86.42
SLIP-MOCOV2	79.43	91.25	87.56	82.01	92.15	86.48

Methods		ChestXray14 (AUROC)		
	LP	1%	10%	All
Random Init.		58.48 ± 1.55	69.50 ± 0.73	77.42 ± 0.39
ImageNet Init.		64.18 ± 1.35	68.68 ± 0.28	76.58 ± 1.23
CLIP		62.20 ± 3.20	68.62 ± 1.89	77.92 ± 0.72
ConVIRT		61.20 ± 1.97	69.85 ± 0.40	78.94 ± 0.65
MoCoV2		63.18 ± 1.44	69.48 ± 0.84	78.50 ± 0.29
SimCLR		61.38 ± 1.08	68.45 ± 1.64	77.62 ± 0.99
VICReg		61.46 ± 1.79	67.50 ± 1.57	78.28 ± 0.79
SLIP-MoCoV2-CLIP		62.92 ± 2.56	69.16 ± 0.59	78.50 ± 0.72
SLIP-MoCoV2-ConVIRT		63.68 ± 0.89	68.90 ± 1.11	79.00 ± 0.42
SLIP-SimCLR-CLIP		63.14 ± 3.19	70.45 ± 1.30	78.14 ± 1.51
SLIP-SimCLR-ConVIRT		60.97 ± 1.28	69.08 ± 3.05	78.42 ± 0.75
SLIP-VICReg-CLIP		61.26 ± 2.63	68.14 ± 1.36	77.82 ± 0.39
SLIP-VICReg-ConVIRT		61.72 ± 2.48	67.92 ± 1.55	78.94 ± 1.30
SLIP-VICReg-ConVIRT (Ours)	O	60.58 ± 1.00	70.52 ± 0.96	79.10 ± 0.50

Table 6: This is the caption

Methods	CheXpert (AUROC)			
	LP	1%	10%	All
Random Init.	62.91 ± 0.62	62.88 ± 1.13	70.36 ± 0.98	78.28 ± 1.10
ImageNet Init.	67.12 ± 0.24	69.33 ± 0.15	77.98 ± 0.28	82.29 ± 2.79
CLIP	78.17 ± 0.23	78.19 ± 1.80	82.90 ± 1.46	84.36 ± 3.07
ConVIRT	79.92 ± 0.11	76.96 ± 2.34	82.89 ± 2.40	85.20 ± 1.44
MoCoV2	71.37 ± 0.17	74.33 ± 1.53	79.58 ± 1.30	84.12 ± 1.08
SimCLR	68.70 ± 0.39	72.83 ± 0.91	78.33 ± 0.88	84.74 ± 0.96
VICReg	71.05 ± 0.48	72.16 ± 1.83	78.51 ± 1.18	85.40 ± 1.63
SLIP-MoCoV2-CLIP	76.89 ± 0.16	77.61 ± 1.57	80.29 ± 0.71	82.92 ± 1.24
SLIP-MoCoV2-ConVIRT	77.02 ± 0.14	77.93 ± 1.43	83.45 ± 1.79	83.51 ± 1.37
SLIP-SimCLR-CLIP	78.02 ± 0.13	77.84 ± 0.99	81.67 ± 2.76	84.33 ± 0.79
SLIP-SimCLR-ConVIRT	75.97 ± 0.10	76.68 ± 1.61	82.57 ± 1.55	84.39 ± 1.09
SLIP-VICReg-CLIP	69.01 ± 0.27	75.06 ± 1.64	81.17 ± 1.61	82.11 ± 1.34
SLIP-VICReg-ConVIRT	71.72 ± 0.36	74.93 ± 2.27	81.17 ± 2.34	83.05 ± 1.76
SLIP-VICReg-ConVIRT (Ours)	74.65 ± 0.18	76.55 ± 1.53	83.05 ± 1.58	85.51 ± 0.28

Table 7: This is the caption-2


Original ResNet-50
ResNet-50 SimCLR
ResNet-50 SLIP-SimCLR
Original ResNet-50
Atelectasis
ResNet-50 SimCLR
Atelectasis
ResNet-50 SLIP-SimCLR
Atelectasis
Original ResNet-50
Pleural Effusion
ResNet-50 SimCLR
Pleural Effusion
ResNet-50 SLIP-SimCLR
Pleural Effusion