Towards Robust and Adaptable Diagnosis of Pneumonia from Chest X-ray Data

Raphaël Attias (287613), Riccardo Cadei (321957)
CS-503 Final Project Report

***Abstract*—Chest radiography is a cost effective and powerful investigator method that conveys crucial respiratory information for pneumonia detection. Artificial intelligence (AI) researchers and radiologists have recently reported AI systems that accurately diagnose pneumonia from a chest X-Ray images using deep neural networks when trained on a sufficient large and homogeneous amount of labelled images. However, the robustness and adaptability of these systems, trained minimizing the empirical risk (ERM), remains far way. In fact, ERM have no way of discard environment specific spurious features and take into account confounders, creating an alarming situation in which the systems appear accurate, but fail when tested in new hospitals. We propose here 2 ideas to address this challenge towards a robust and adaptable diagnosis of pneumonia: (i) discard the spurious feature replacing ERM with a robust training routine (i.e. IRM and v-REx); (ii) replace the straight-forward deep neural networks with a new modular architecture, encoding separately the invariant features (in a self-supervised fashion) and the style confounders. Then we validate the impact of each contribution, one at the time, by 2 experiments on real-word data.**

I. Introduction

According to the World Health Organization (WHO), Pneumonia affects every year children and families worldwide and it is the largest infectious cause of death in children [1]. One key element of its diagnosis is the chest X-rays (CXR) radiography, routinely obtained as standard of care: in fact according to radiologists and general practitioners it appears to contain the most correlated factors with the illness even before observing common symptoms [2]. In Figure 1 we report, in example, a comparison of chest X-rays between a normal lung and a ill lung. However, rapid

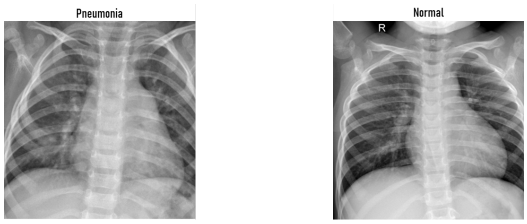

Figure 1. Comparison of chest X-rays between a normal lung and a ill lung. The most correlated factors with pneumonia are hidden in these images.

radiologist prediction is not always available for this illness that requires immediate antibiotic treatment and supportive care, and even among senior radiologists disagreements are common, up to a kappa-score equal to 0.395 [3].

There are already several studies combining deep learning and computer vision to extract this information [4] almost in real-time, based on public dataset collected by several hospitals all around the world. The main assumption of these methods, Empirical Risk Minimization (ERM) based, is that all the collected images, both in training and test set, are independent and identically distributed (*i.i.d.*). However this strong assumption rarely holds in practice, and it is commonly forced by considering only images from the same hospital. In Figure 2 we report an example of these differences comparing the average distribution between the publicly available data from the National Institutes of Health (NIH) [5] and the Guangzhou Women and Children's Medical Center (GMC) [6].

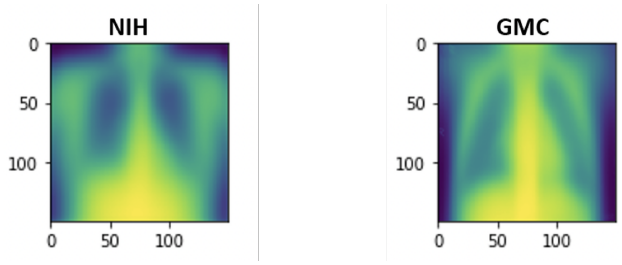

Figure 2. The different average distribution of CXR between the NIH and GMC dataset suggests a strong need to take care of domain shifts.

There are several limitations in this approach:

- the model is over-trained on environment specific spurious features ('shortcuts') and style confounders and it is not able to generalize on new domains (i.e. data from a different hospital);
- there is no specific design for transfer learning on a new domain: all the network has to be retrained and it can even get worse in a few-shots transfer;
- the method does not benefit from having a lot of public data available (mainly unlabelled) from different (heterogeneous) environments.

Robustness and Transferability are two of the biggest open challenges in theoretical machine learning and even more in applications like diagnosis of diseases. Even state-of-the-art models can rely on confounding factors and spurious 'shortcuts' rather than medical pathology, creating an alarming situation in which the systems appear accurate, but fail when tested in new environments [7].

This study aims to address both the challenges, one at the time, by two main contribution:

1) discarding the spurious feature replacing ERM with a robust training routine (i.e. IRM and v-REx);
2) taking in to account the style confounders (i.e. calibrations) replacing the straight-forward deep neural networks with a new modular architecture, encoding separately the invariant features (in a self-supervised fashion) and the style confounders.

We hope our findings will pave the way towards robust and adaptable models for CXR diagnosis. In Section II we explain how our work is related to the current literature and how it differs; in Section III we describe in detail our two main contributions, in Section IV we describe the experiments on which we test our methods reporting the results, and finally in Section V we summarize the conclusion and limitations.

The source code of our method as well as baselines can be found at `https://github.com/riccardocadei/pneumoniadiagnosis`.

II. Related Work

The COVID-19 pandemic has motivated the research to focus more than ever in the development of systems to predict pneumonia. Deep Learning models have been shown to be powerful tools in predicting pneumonia from CXR images, with sensitivity results often surpassing 85% and reaching up to 98.81% on certain datasets [8], [9]. However, as far to our knowledge, only a recent study [3], considered a dataset composing the data collected from different hospitals. Its focus is to create a robust model even under strong domain shifts between training and test set. This is a crucial aspect, since in real world the data are more and more heterogeneous and there could be a lot of benefits both if we are able to elaborate them all together, either if we are able to tackle this heterogeneity in the test data. Our proposal differs from this study since we aim to create a model not just robust but also adaptable to the new environments. Our modular architecture is strongly inspired by recent advances in Causal Representation Learning, trying to individually extract in a latent representation the invariant features and the confounders (ignoring the spurious features) [10]. Our work differs from this not just for the task (motion forecasting vs image classification), but mainly for the training routine itself. In fact the main autoencoder in charge to extract the invariant features is now trained in a self-supervised fashion through a robust routine and it can take advantage of using a lot heterogeneous unlabelled data (which is quite an advantage since the ground truth is not so precise). This idea was supported by the fact the self-supervised learning was already showed to get surprising performances (99.2% sensitivity) on homogeneous data [11], [12].

III. Method

A. Robustness

A main assumption in Statistical Learning is that the data are all independent and identically distributed (*i.i.d.*) and then, through the Law of Big Number we can approximate the Expected Risk:

$$\mathcal{R}(f) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)} [\ell(f(\boldsymbol{x}), y)] \tag{1}$$

with the Empirical Risk:

$$\mathcal{R}(f) := \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \ell(f(\boldsymbol{x}), y) \tag{2}$$

and minimize it (ERM). Unfortunately in real world applications this assumption doesn't hold and the data are generally split in different environments following different distributions (i.e. CXR images produced by different machines with different calibrations or ambient condition). Shuffling the different environments all together (common practice in Statistics) to enforce the *i.i.d.* condition is a loss of information.

Several methods have been recently proposed to learn the spurious correlations that we want to suppress from the differences among the environments. Two of the most promising robust training routine are Invariant Risk Minimization (IRM) [13] and Risk Extrapolation (v-REx) [14]. Let f the features extractor and g the final classifier of the model on top of g. IRM enforces the model $g \circ f$ to be equally optimal in every environment adding to the vanilla Empirical Risk a constraint on the invariance of the features extractor.

$$\begin{aligned} \min_{f,g} \quad & \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \mathcal{R}^e(f, g) \\ \text{s.t.} \quad & g \in \arg\min_{g^*} \mathcal{R}^e(f, g^*) \quad \forall e \in \mathcal{E}, \end{aligned} \tag{3}$$

Since Problem 3 is a bi-level optimization problem and it can be difficult to solve it in practice, we propose to relax the additional constraint through a gradient norm penalty over the empirical risk $\mathcal{R}^e$ in each training environment (IRM-v1).

$$\min_{f,g} \quad \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} [\mathcal{R}^e(f, g) + \lambda \nabla_g \mathcal{R}^e(f, g)] \tag{4}$$

v-REx proposes instead to add to vanilla Empirical Risk a penalty on its variance among the different environments.

$$\min_{f,g} \quad \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \mathcal{R}^e(f, g) + \mathrm{Var}\{\mathcal{R}^1(f, g), ..., \mathcal{R}^{|\mathcal{E}|}(f, g)\} \tag{5}$$

B. Adaptability

To provide an adaptable model for different environments we propose a novel architecture which both encapsulate the relevant features of the CXR image and also encodes the (style) confounders of the environment. This model is a modular architecture based on three components: the CXR encoder Φ, the style encoder Φ_{style} and the pneumonia / healthy classifier g.

The latent representation encoder Φ has a main motivations: to provide a meaningful representation of the invariant features for pneumonia detection, hopefully independent from the style confounders. We propose to train this module a priori in Self-Supervised way using IRM, and freezing its weights during the remaining parts of the training. In such a way we can take advantage of the huge amount of data available open source even if unlabelled and collected from different environments.

The style encoder Φ_{style} encodes the various features related to the proprieties and defining characteristics of an environment. Given a batch from the said environment, it produces a (unique) single feature vector which is then concatenated with the latent representation of each image in the batch.

The whole embedding is then given in input to the final classifier g. Both the style encoder and the final classifier on top of it are trained together in a supervised way using standard ERM. An alternative approach could be to pretrain also style encoder using an auxiliary loss predicting the environment.

Our hypothesis is that for few-shots transfer learning we don't need to retrain from scratch the whole architecture, but we just need to update the style encoder.

In Figure 3 we report a schematic representation of our modular architecture.

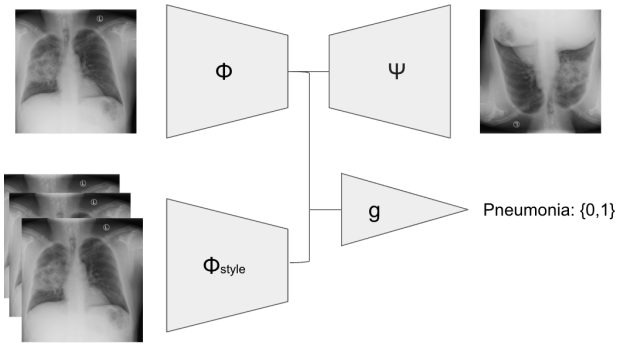

Figure 3. Schematic representation of our architecture composed by three main modules: the representation (invariant) encoder Φ, the style encoder Φ_{style} and the linear classifier g. The features extracted from the two encoders are concatenated and given in input to the final classifier g.

IV. Experiments

We validate the efficacy of our 2 main contribution on the similar experiments on real world data.

A. Experiment 1

The first experiment consists in comparing Invariant Risk Minimization (IRM) and Empirical Risk Minimization (ERM) for Out of Distribution Generalization. The goal is to evaluate the robustness of such methods and we are expecting IRM methods to reduce the possible overfitting that ERM methods would produce on the training set.

We reproduce the experiments setup proposed in Bellot et al. [3] mixing two CXR datasets from NIH and GMC hospitals. The training set is heavily biased toward the NIH dataset, whereas the test set is the other way around with a dataset biased toward the GMC dataset. A robust model should be able to avoid the shortcuts in the training set from the slight differences among the different training environments. We also propose to compare another robust training routine for out of distribution generalization named: Risk Extrapolation (v-REx) [15].

1) Setup: We have compared 2 models for pneumonia prediction, both based on CNN. The first *baseline* model has been reproduced from [3] and provides a fair comparison with the original paper for the perks of IRM methods. This model is based on three up-sampling Conv2D blocks, with padding 1 and followed by ELU activation function. They are followed by two linear layers to flatten the output and finally we use a sigmoid activation function. The second model is based on a SOTA model called *CheXNet* [16], and is a deeplearning model using DenseNet121 followed by a linear classifier. Both models have been trained from random weights.

The training pipeline is done simultaneously on two data environments, both of same size with in total 1946 samples and with a 50/50 split for the pneumonia and healthy lungs. The first environment *Env 1* has 90% samples from NIH and 10% from GMC, whereas the second environment *Env 2* has 80% NIH and 20% GMC. As discussed previously, the test set is then biased toward GMC, with 10% from NIH and 90% GMC. The total size of the test set is 514 samples and the split remains even with ill and healthy lungs.

2) Results: The results of the various setups is presented in Table I. As expected, for both the models, baseline ERM leads to the best performances in the training set, however it significantly get worse on the test (out of distribution). The main explanation is that, baseline ERM has no reason to discard the spurious correlations (shortcuts) during the training and it overfir on them. In reverse, both IRM and v-REx (try to) extract only the invariant features, and for this reason they get worse performances on the training set for better performances on the test. In particular CheXNet using IRM with $\lambda = 0.1$ leads to the best performances on the test set (Accuracy of 83.269%). Both the robust the robust training routine are very sensitive to the value of the hyper parameter λ which represents somehow the strength of the invariance constraint.

It can also seems surprising that both the robust training routine can lead better performances on the test rather than the training set. Actually the performances should be similar and if they differ it is because probably the data in the main environment in the test set are more expressive or cleaner (easier task) than the the data on the main environment in the training set.

Finally let's remember that pneumonia diagnosis from CXR data is not a deterministic task even for humans and an accuracy greater than 80% is already very significant.

Model	Method	Train (Env 1)	Train (Env 2)	Test
Baseline	ERM	87.564	88.940	61.154
	IRM ($\lambda = 0.1$)	63.512	64.249	61.346
	IRM ($\lambda = 1$)	55.498	58.076	**64.230**
	V-REx ($\lambda = 0.1$)	84.943	86.111	55.961
CheXNet	ERM	77.492	78.909	71.346
	IRM ($\lambda = 0.1$)	72.914	74.845	**83.269**
	IRM ($\lambda = 1$)	68.961	71.913	78.653
	V-REx ($\lambda = 0.1$)	76.104	78.549	80.961

Table I
PERCENTAGE OF ACCURACY FOR PNEUMONIA PREDICTION FROM CHEST X-RAYS IMAGES ON AN BALANCED DATASET. A BASELINE CONVOLUTIONAL NEURAL NETWORK AND CHEXNET ARE TRAINED ON TWO SIMILAR (BUT DIFFERENT) ENVIRONMENTS AND TESTED OUT OF DISTRIBUTION USING DIFFERENT TRAINING ROUTINES. BOTH IRM AND V-REX SIGNIFICANTLY IMPROVE THE ROBUSTNESS OF BOTH THE MODELS.

B. Experiment 2

The second experiment wants to evaluate the goodness of our modular architecture. We consider the same data used in Experiment 1, but rather than combining them in new 'artificial' environments, we propose to train the model in a real world scenario, where the training set is composed by 2 different environments (images collected from 2 different hospitals) and we want to predict well (on the test data) at least on both these environments.

1) Setup: Our modular architecture is built from primarily three components: a CXR representation (invariant) encoder Φ, a style encoder Φ_{style} and a linear classifier g.

In practice we decide to not train from scratch the representation encoder Φ, using instead a pretrained model. We propose to use SimCLR [17], which was trained on a self supervised routine on ImageNet and has already been successfully used as a latent representation encoder in a variety of medical tasks [18]. More precisely this encoder is based on ResNet model with depth 50 and width 2, with weights available on the official project page. The latent representation is encoded as a feature vector of size 4096. It has already been reported that training from scratch on medical tasks had marginal improvements compared to using the pre-trained weights [18] so, for this reason and as the task is already computational heavy, we choose to start from these pre-trained weights. However since this model was pretrained in a Self-Supervised way, using ERM on a contrastive loss, and we strongly suspect that a pretraining using IRM could improve a lot the performances of the whole architecture (disentangling the main latent representation with the syle latent representation).

The style encoder Φ_{style} is a convolutional neural network composed of three up sampling blocks. The up sampling blocks are followed by ReLU activation and maxpool. The goal with the style encoder is to encapsulate any new environment with little to no training. We hope that the latent representation given by Φ is distinct enough from the encoded style by Φ_{style} such that the concatenated feature vectors contains all the meaningful information to provide accurate pneumonia prediction.

Finally the classifier g is a 3 layered linear network with ReLU activation. The dataset used are both from GMC and NIH datasets, using the provided list of samples from their train and test set. For computational purposes we kept the number of samples per environment and stage to 2500 and with batch size 5. In order to get comparable performances with the previous experiment we chose to balance the data to 50:50 (even if in practice the ratio is around more than 1:99). Adam optimizer was used, with learning rate of 0.001.

2) Results: The final test accuracy for pneumonia detection can be found on Table II for both the GMC and NIH balanced dataset. We observe satisfactory test accuracy for GMC of 84.736%, which is a qualitative increase of accuracy compared to the SOTA model CheXNet seen in Table I. However the test accuracy on NIH is effectively lower than GMC with only 65.169%. This poor performance on these balanced environments can be due to many reasons, mainly because we haven't retrain the representation encoder Φ using a IRM and there is no reason it could filter just the invariant features (disentangling from the style encore Φ_{style}). A second interpretation is that the NIH dataset may me a more difficult environment for pneumonia detection, which would explain also part of the better results for NIH biased environments in the first experiment. While this approach need additional and reflection on the training pipeline, it still shows promising results that are in the range of actual SOTA model for this task.

Model	Method (Φ_{style}, g)	Train		Test	
		GMC	NIH	GMC	NIH
Ours	ERM	93.639	79.857	84.736	65.169

Table II
PERCENTAGE OF ACCURACY FOR PNEUMONIA PREDICTION FROM CHEST X-RAYS IMAGES ON AN BALANCED DATASET. OUR MODULAR ARCHITECTURE IS TRAINED ON BOTH 2 DIFFERENT ENVIRONMENTS AND TESTED ON THE SAME ENVIRONMENTS. EVEN IF THE MODEL CAN GENERALIZE QUITE WELL ON GMC ENVIRONMENT, IT IS NOT PROPERLY THE CASE ON NIH ENVIRONMENT. THE MAIN LIMITATION IS THAT WE ARE USING A PRETRAIN REPRESENTATION ENCODER (RATHER THAN TRAINING IT FROM SCRATCH) WHICH IS NOT INVARIANT.

V. Conclusion and Limitations

In this study we have highlighted the limits of standard Empirical Risk Minimization for robust pneuomonia diagnosis and we proposed two main contributions not just to enforce the diagnosis robustness of a model Out of Distribution but also to make it Adaptable for few-shots trasfer learning. In particular we replaced Empirical Risk Minimazation with Invariant Risk Minimazion and Risk Extrapolation: two robust training routine which enforce the invariance of the latent representation (with respect to different environments) in add to the risk minimization. In add to this we proposed a modular architecture to disentangle the latent space filtering the invariant features, discarding the spurious features but still taking in to account the style confounders. We proposed to train a priori the (invariance) encoder in a Self Supervised fashion using IRM. In such a way it can take advantage of the huge open source datasets of CXR available online even if unlabelled (or not perfectly labelled) and collected from different environments (i.e. hospitals). In practice unfortunately, we hadn't the resources to train this encoder from scratch and we used instead a pre-trained model still in a self-supervised way minimizing a contrastive loss but using Empirical Risk Minimization. There is no reason it should just filter the invariant features and unfortunately this is our main limitation in the Experiment 2. A natural extension of our Experiment 2 is then to pre-train from scratch also this encoder using Invariant Risk Minimization. An other natural extension for Experiment 2 is to pretrain individually also the style encoder using an auxiliary task (environment classification). Finally once we will be able to properly solve the Experiment 2 (good prediction accuracy on both the test environments already seen during the training) we propose to add a third experiment regarding few-shots transfer learning, verifying if our modular architecture is properly able to quickly adapt to new domains. In particular we propose to compare it with CheXNet (ERM), evaluating also if there are benefits in updating the weights only of the style encoder (our hypothesis) rather than the whole (huge) architecture.

VI. Individual contributions

Riccardo took care of the study of the literature and he is the main contributor of the methods and experiments design. Both Riccardo and Raphael coded the different models and algorithms. Raphael took care of collecting and preprocessing the data, setting up two different machines (Azure Cluster and a GPU in local) and running there the majority of the experiments.

References

[1] World Health Organization, https://www.who.int/news-room/fact-sheets/detail/pneumonia, 2019.

[2] A. M. Speets, A. W. Hoes, Y. van der Graaf, S. Kalmijn, A. P. E. Sachs, and W. P. T. M. Mali, "Chest radiography and pneumonia in primary care: diagnostic yield and consequences for patient management," *European Respiratory Journal*, vol. 28, no. 5, pp. 933–938, 2006. [Online]. Available: https://erj.ersjournals.com/content/28/5/933

[3] A. Bellot and M. van der Schaar, "Accounting for unobserved confounding in domain generalization," *arXiv preprint arXiv:2007.10653*, 2020.

[4] W. C. Daniel S. Kermany, Michael Goldbaum, "Identifying medical diagnoses and treatable diseases by image-based deep learning: Cell," https://www.cell.com/cell/fulltext/S0092-8674(18)30154-5, 2019, (Accessed on 10/15/2021).

[5] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, "Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE CVPR*, vol. 7, 2017.

[6] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[7] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, pp. 1–10, 2021.

[8] "Pneumonia detection in chest x-ray images using an ensemble of deep learning models," https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0256630, (Accessed on 10/15/2021).

[9] "Deep learning approaches for detecting pneumonia in covid-19 patients by analyzing chest x-ray images," https://www.hindawi.com/journals/mpe/2021/9929274/, (Accessed on 10/15/2021).

[10] Y. Liu, R. Cadei, J. Schweizer, S. Bahmani, and A. Alahi, "Towards robust and adaptive motion forecasting: A causal representation perspective," *arXiv preprint arXiv:2111.14820*, 2021.

[11] I.-Y. L. C. Park, J.; Kwak, "A deep learning model with self-supervised learning and attention mechanism for covid-19 diagnosis using chest x-ray images," file:///Users/raphael-attias/Downloads/electronics-10-01996-v2.pdf, 2021, (Accessed on 10/15/2021).

[12] J. P. P. D. Matej Gazda, Jakub Gazda, "Self-supervised deep convolutional neural network for chest x-ray classification," https://arxiv.org/pdf/2103.03055.pdf, 2020, (Accessed on 10/15/2021).

[13] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[14] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5815–5826.

[15] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," 2021.

[16] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[18] L. Chaves, A. Bissoto, E. Valle, and S. Avila, "An evaluation of self-supervised pre-training for skin-lesion analysis," 2021.