CS 598 Deep Learning for Healthcare (Spring, 2025)
Project Proposal

Eric Schrock

ejs9@illinois.edu

1 Introduction

1.1 Proposed Project

For my project, I propose to reproduce some or all of the findings of the paper, "Multi-Label Generalized Zero Shot Learning for the Classification of Disease in Chest Radiographs" (Hayat, Lashen, and Shamout 2021). Compute power will likely determine whether I am able to reproduce the full set of findings or merely a subset, given the time remaining in the semester (more about this in Section 3.3). In addition to reproducing original findings, I propose to implement and test at least one ablation or extension of the work done by the authors of the paper (see Section 2.3).

1.2 Problem Statement

Deep learning models for classifying diseases from chest X-ray (CXR) images have had great success, achieving comparable results to human experts. However, collecting and labeling training data for these models is expensive and time-consuming. For rarer diseases, it is often not economical, and sometimes not even possible, to gather the amount of data needed for supervised learning models. When a new disease emerges, the need for massive data collection slows the response. Human radiologists leverage other sources of knowledge to identify diseases they have previously never seen in X-ray form. Could a deep learning model do the same?

Multi-label generalized zero shot learning (ML-GZSL), which uses semantic information to identify classes not present in the set of labeled images used to train the model, has worked well in similar circumstances (Scheirer et al. 2013; Rahman, Khan, and Barnes 2020; Huynh and Elhamifar 2020). However, these prior works have at least two limitations. First, they "extract a fixed visual representation of the image from a pre-trained visual encoder or a detection network" (Hayat, Lashen, and Shamout 2021), which means they cannot be trained end-to-end. Second, "projecting these extracted visual features to the semantic space shrinks the diversity of the visual information, which gives rise to inherent limitations" (Hayat, Lashen, and Shamout 2021), one of those being the hubness problem (Dinu, Lazaridou, and Baroni 2015). Can these limitations be overcome?

2 Methodology

2.1 Specific Approach

The paper in question proposes the CXR-ML-GZSL model, shown in Figure 1. It is comprised of a pre-trained text encoder to convert class labels to a semantic embedding space, a trainable visual encoder to convert X-ray images to a visual embedding space, and mapping models into a shared latent embedding space. The output is a score for each possible class, representing how relevant it is to the input image.

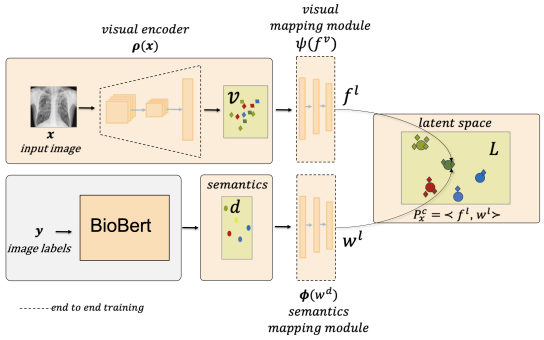

Figure 1: CXR-ML-GZSL (Hayat, Lashen, and Shamout 2021)

Since the visual encoder is trainable, the whole pipeline between the X-ray image and the semantic embedding of the class label is trainable from end-to-end, better tuning the overall model to the task at hand. Additionally, the added latent embedding space is meant to address the limitations of mapping the visual embedding space directly onto the semantic embedding space.

The increased complexity of this model, compared to other ML-GZSL models, requires a multifaceted training objective, captured in a three-part loss function, where γ_1 and γ_2 are the regularization parameters for the second and third components of the loss function.

$$\min_{\phi,\rho,\psi} \mathcal{L} = \mathcal{L}_{rank} + \gamma_1 \mathcal{L}_{align} + \gamma_2 \mathcal{L}_{con}, \tag{1}$$

$\mathcal{L}_{rank}$ adds penalties for any positive ground-truth relevance scores not larger than all negative ground-truth relevance scores by at least a margin of δ. $\mathcal{L}_{align}$ adds penalties

if input images and their labels do not map near each other in the latent embedding space. $\mathcal{L}_{con}$ add penalties if labels do not have similar relationships in both the semantic and latent embedding spaces, as those semantic relationships are key to zero shot learning.

For the pre-trained text encoder, the model uses BioBERT (Lee et al. 2019), which was trained specifically on a biomedical corpora. For the trainable visual encoder, the model uses Densenet-121 (Rajpurkar et al. 2017), as at that time it was the best chest X-ray classifier. For both BioBERT and Densenet-121, the model skips the final classification step in order to get semantic and visual embeddings, respectively, as output instead. The two mapping models are both three layer feed-forward neural networks.

The model was trained over 100 epochs using the Adam optimizer (Kingma and Ba 2017), with multiple repetitions to tune the learning rate, γ_1, and γ_2 hyperparameters. Its performance was measured using the "overall precision, recall and f1 scores for the top k predictions where $k \in \{2, 3\}$ [as well as] the average area under the receiving operating characteristic curve (AUROC) for seen and unseen classes and its harmonic mean, since computing recall for top k predictions may not be a sufficient indicator of class-wise performance" (Hayat, Lashen, and Shamout 2021).

2.2 Novelty, Relevance, and Hypotheses

CXR-ML-GZSL is the first use of ML-GZSL to classify diseases from chest X-ray images. It is also unique from other ML-GZSL models in the following three ways.

- It can be trained end-to-end, thanks to the trainable visual encoder.
- It maps both the semantic and visual embedding spaces into a shared latent embedding space, instead of mapping the visual space onto the semantic space, losing less visual information.
- It uses BioBERT, which is trained on a biomedical corpora, resulting semantic embeddings that are tuned for healthcare use cases.

The hypothesis is that these improvements will result in better performance when classifying both seen and unseen diseases in chest X-ray images, compared to two state-of-the-art ML-GZSL models: LESA (Huynh and Elhamifar 2020) and MLZSL (Lee et al. 2018). The results in Figure 2 bear this out.

	k=2			k=3			AUROC		
Method	r@k	p@k	f1@k	r@k	p@k	f1@k	S	U	H
LESA (2020)	0.14	0.10	0.03	0.21	0.11	0.05	0.51	0.50	0.50
MLZSL (2018)	0.20	0.19	0.16	0.30	0.17	0.20	0.72	0.54	0.62
OUR_{e2e}	**0.36**	**0.33**	**0.32**	**0.47**	**0.28**	**0.34**	**0.79**	**0.66**	**0.72**

Figure 2: Performance of LESA, MLZSL, and CXR-ML-ZSL ("OUR_{e2e}") on the ChestX-ray14 database (Wang et al. 2017). Metrics are precision, recall, f1, and AUROC (split by seen, unseen, and the harmonic mean of the two) (Hayat, Lashen, and Shamout 2021).

2.3 Ablations and Extensions

The paper in question already includes two ablation studies, one to quantify the contribution of each of the three components of the loss function and one to quantify the contribution of the trainable visual encoder vs several pre-trained visual encoders. An ablation study I could run would be to test the mapping models with fewer neural network layers.

The paper in question only performed hyperparameter tuning on the learning rate and loss function regularization parameters. It suggests further hyperparameter tuning on the batch size and mapping model dimensions. The δ parameter from the $\mathcal{L}_{rank}$ loss component could also be tuned.

It is possible that Densenet-121 or BioBERT have been superseded by faster and/or more accurate models since the paper in question was published. Testing CXR-ML-GZSL with newer visual and semantic encoders would be valuable extensions.

3 Data Access and Implementation Details

3.1 Dataset

I will use the same ChestX-ray14 database (Wang et al. 2017) used by the paper in question. It is publicly available at: https://nihcc.app.box.com/v/ChestXray-NIHCC/.

3.2 Model and Codebase

The pre-trained weights for CXR-ML-GZSL are publicly available at: https://drive.google.com/file/d/17ioJMW3qNx1Ktmr-hXn-eqp431cm49Rm/view. I plan to use them to run the model without training, as an initial sanity check.

The code for CXR-ML-GZSL is publicly available at: https://github.com/nyuad-cai/CXR-ML-GZSL/. I plan to use it as a reference and skeleton. Per the project instructions, I need to reproduce the data preprocessing, model, training loop, and metrics with the help of an LLM. I will plug these items into the existing skeleton provided by this public code repository.

3.3 Computation Feasibility

CXR-ML-GZSL was trained on an NVIDIA Quadro RTX 6000, which took around 8 hours (Hayat, Lashen, and Shamout 2021). I have access to an NVIDIA GeForce RTX 4050 Laptop locally and an NVIDIA Tesla T4 via Google Colab. Both are much less powerful than the Quadro[1][2].

I have several options. First, I can train CXR-ML-GZSL for a few epochs on the RTX 4050 and Tesla T4 to gauge whether the full training time is doable. If not, I can purchase access to an NVIDIA A100[3] or L4[4] via Google Colab and run the same experiment to gauge whether I can afford a full training run. If not, I will train CXR-ML-GZSL on the RTX 4050 for as many epochs as it can manage.

[1] https://www.topcpu.net/en/gpu-c/quadro-rtx-6000-vs-geforce-rtx-4050-mobile

[2] https://www.topcpu.net/en/gpu-c/quadro-rtx-6000-vs-tesla-t4

[3] https://www.topcpu.net/en/gpu-c/quadro-rtx-6000-vs-a100-pcie

[4] https://www.topcpu.net/en/gpu-c/quadro-rtx-6000-vs-l4

References

Dinu, G.; Lazaridou, A.; and Baroni, M. 2015. Improving zero-shot learning by mitigating the hubness problem. arXiv:1412.6568.

Hayat, N.; Lashen, H.; and Shamout, F. E. 2021. Multi-Label Generalized Zero Shot Learning for the Classification of Disease in Chest Radiographs. arXiv:2107.06563.

Huynh, D.; and Elhamifar, E. 2020. A Shared Multi-Attention Framework for Multi-Label Zero-Shot Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8773–8783.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Lee, C.-W.; Fang, W.; Yeh, C.-K.; and Wang, Y.-C. F. 2018. Multi-Label Zero-Shot Learning with Structured Knowledge Graphs. arXiv:1711.06526.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Rahman, S.; Khan, S.; and Barnes, N. 2020. Deep0Tag: Deep Multiple Instance Learning for Zero-Shot Image Tagging. *Trans. Multi.*, 22(1): 242–255.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; Lungren, M. P.; and Ng, A. Y. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225.

Scheirer, W.; Rocha, A.; Sapkota, A.; and Boult, T. 2013. Toward Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7): 1757–1772.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–3471. IEEE.