Class Specific Feature Disentanglement and Text Embeddings for Multi-label Generalized Zero Shot CXR Classification

Dwarikanath Mahapatra[1,2](✉), Antonio Jose Jimeno Yepes[3], Shiba Kuanar[4], Sudipta Roy[5], Behzad Bozorgtabar[6,7], Mauricio Reyes[8], and Zongyuan Ge[9]

[1] Inception Institute of AI (IIAI), Abu Dhabi, UAE
dwarikanath.mahapatra@inceptioniai.org
[2] Faculty of Engineering, Monash University, Melbourne, Australia
[3] Unstructured Technologies, Sacramento, USA
[4] Mayo Clinic, Rochester, USA
[5] Jio Institute, Navi Mumbai, India
[6] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[7] Lausanne University Hospital (CHUV), Lausanne, Switzerland
[8] University of Bern, Bern, Switzerland
[9] AIM for Health Lab, Monash University, Melbourne, Victoria, Australia

Abstract. Robustness of medical image classification models is limited by its exposure to the candidate disease classes. Generalized zero shot learning (GZSL) aims at correctly predicting seen and unseen classes and most current GZSL approaches have focused on the single label case. It is common for chest x-rays to be labelled with multiple disease classes. We propose a novel multi-label GZSL approach using: 1) class specific feature disentanglement and 2) semantic relationship between disease labels distilled from BERT models pre-trained on biomedical literature. We learn a dictionary from distilled text embeddings, and leverage them to synthesize feature vectors that are representative of multi-label samples. Compared to existing methods, our approach does not require class attribute vectors, which are an essential part of GZSL methods for natural images but are not available for medical images. Our approach outperforms state of the art GZSL methods for chest xray images.

Keywords: Multi-label · GZSL · Text Embeddings · chest x-rays · feature synthesis

1 Introduction

Deep learning methods provide state-of-the-art (SOTA) performance for a variety of medical image analysis tasks such as diabetic retinopathy grading [7], and

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_26.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 276–286, 2023.
https://doi.org/10.1007/978-3-031-43895-0_26

chest X-ray diagnosis [10], to name a few. SOTA fully supervised methods have access to all classes as part of the training data whereas most real world clinical applications do not provide access to all classes which leads to unseen classes being wrongly diagnosed as one of the seen classes. Zero-Shot Learning (ZSL) aims to classify unseen test data by learning their plausible representations from seen class features, and in Generalized Zero-Shot Learning (GZSL) the model should accurately classify both seen and unseen classes during test time.

Previous works on GZSL in medical images have focused on the single class scenario where an image is assigned a single disease class [18,21]. However, chest X-ray images have multiple labels and single-label methods do not work well in this setting. Hence we propose a multi-label GZSL approach that takes into account the semantic relationship between the multiple disease labels and learns a highly discriminative feature representation.

GZSL for natural images [6,12,14,22] have the advantage of providing attribute vectors for all classes that enables a model to correlate between attribute vectors and corresponding feature representations of the seen classes. Defining unambiguous attribute vectors for medical images requires deep clinical expertise and time. This is more challenging for the multi-label scenario, where many disease conditions have similar appearances and textures. For example, in lung X-ray diagnosis, many conditions frequently co-occur with labels such as *Atelectasis, Effusion, and Infiltration.* An effective class attribute vector should be able to precisely identify individual labels and differentiate them from other co-occurring disease labels, which is very challenging to define. To overcome the above challenges, we make the following contributions:

1. We propose a novel feature disentanglement method where a given image is decomposed into class-specific and class agnostic features. This enables better feature learning of different classes and subsequently contributes to better feature synthesis in the multi-label scenario.
2. We use text embedding similarities to learn the semantic relationships between different labels. This contributes to more accurate learning of multi-label interactions at a global scale and guide feature generation to synthesize feature vectors that are realistic and preserve the multi-label relationship between disease labels.
3. We solve the GZSL classification problem in terms of cluster assignment. Class specific feature disentanglement performs better for multi-label classification [11] and we use this concept to synthesize unseen class features and subsequently perform classification.

Prior Work: GZSL's objective is to recognize images from known and unknown classes. Many works have shown promising results using GANs [23,26], and Intra-Class Compactness Enhancement [12]. Recent works on multi-label zero-shot learning (ML-ZSL) use information propagation [14], attention mechanisms [9] and co-occurrence statistics with weighted combinations of seen classes [19]. ZSL in medical image analysis is a much less explored topic with limited applications such as registration [13], segmentation [1], gleason grading [16] and artifact reduction [4]. [21] used multi-modal images and medical reports for GZSL of

Table 1. Cosine similarity of the labels' BioBERT embeddings

	Atl.	Card.	Cons.	Edema	Eff.	Emph.	Fibr.	Hernia	Inf.	Mass	No Find	Nodule	PT	Pne.	Pneu.
Atelectasis	1.00	0.84	0.93	0.92	0.66	0.99	0.77	0.99	0.93	0.93	0.49	0.70	0.79	0.99	0.89
Cardiomegaly	0.84	1.00	0.97	0.97	0.93	0.88	0.98	0.83	0.95	0.97	0.81	0.96	0.98	0.87	0.60
Consolidation	0.93	0.97	1.00	0.99	0.84	0.95	0.93	0.92	0.99	0.99	0.69	0.88	0.93	0.94	0.72
Edema	0.92	0.97	0.99	1.00	0.86	0.95	0.93	0.91	0.99	0.99	0.70	0.89	0.94	0.94	0.71
Effusion	0.66	0.93	0.84	0.86	1.00	0.71	0.96	0.65	0.84	0.85	0.91	0.98	0.95	0.70	0.40
Emphysema	0.99	0.88	0.95	0.95	0.71	1.00	0.82	0.99	0.95	0.95	0.54	0.75	0.83	0.99	0.86
Fibrosis	0.77	0.98	0.93	0.93	0.96	0.82	1.00	0.76	0.91	0.93	0.87	0.98	0.99	0.80	0.52
Hernia	0.99	0.83	0.92	0.91	0.65	0.99	0.76	1.00	0.92	0.91	0.48	0.70	0.78	0.99	0.91
Infiltration	0.93	0.95	0.99	0.99	0.84	0.95	0.91	0.92	1.00	0.99	0.68	0.87	0.92	0.95	0.73
Mass	0.93	0.97	0.99	0.99	0.85	0.95	0.93	0.91	0.99	1.00	0.70	0.88	0.94	0.95	0.72
No Finding	0.49	0.81	0.69	0.70	0.91	0.54	0.87	0.48	0.68	0.70	1.00	0.91	0.85	0.53	0.23
Nodule	0.70	0.96	0.88	0.89	0.98	0.75	0.98	0.70	0.87	0.88	0.91	1.00	0.97	0.74	0.45
Pleural_Thickening	0.79	0.98	0.93	0.94	0.95	0.83	0.99	0.78	0.92	0.94	0.85	0.97	1.00	0.82	0.54
Pneumonia	0.99	0.87	0.94	0.94	0.70	0.99	0.80	0.99	0.95	0.95	0.53	0.74	0.82	1.00	0.87
Pneumothorax	0.89	0.60	0.72	0.71	0.40	0.86	0.52	0.91	0.73	0.72	0.23	0.45	0.54	0.87	1.00

chest xray (CXR) images while [17,18] used saliency maps and GANs for GZSL using only CXRs.Recently, language models pre-trained on large corpora have also been considered for GZSL of CXRs [8]. However all the above works operate in the single label setting, while we solve the multi-label problem.

2 Method

Method Overview: Given training data with seen classes we: 1) create a dictionary from the text embedddings; 2) disentangle the image into class specific and class agnostic features; 3) use class specific features to generate features of seen and unseen classes using the Mixup approach [28]; 4) for a given test image apply feature disentanglement and feature similarity analysis to identify the different class labels in the image.

Embeddings: We generate embeddings of image class labels using BioBERT [15], a BERT [5]-like pre-trained model. BioBERT [15] is pre-trained on biomedical literature, more specifically the model available from Huggingface[1], which is a base and cased model. We consider a pooled set that produces a single 768 dimension vector for a label. We then calculate the cosine similarity between each of the labels and represent it as a matrix, which we refer to as $Dict_{Text}$ - dictionary for text embeddings, shown in Table 1.

2.1 Feature Disentanglement

Our feature disentanglement method is inspired from [20] which decomposes the feature space into shape and texture for domain adaptation applications. We decompose the feature space of the seen class samples into 'class-specific'

[1] https://huggingface.co/dmis-lab/biobert-v1.1.

and ‘class-agnostic’ features. Class specific features encode information specific to the particular class, and have low similarity between different classes. Class agnostic features have high similarity across all classes, and have minimal semantic overlap with class specific features. The disentangled features allow for greater accuracy in identifying the multiple labels in a sample. Figure 1 (a) shows the architecture of our feature disentanglement network (FDN) consisting of L encoder-decoder networks corresponding to the L classes in the training data. The encoders and decoders (generators) are denoted, respectively, as $E_l(\cdot)$ and $G_l(\cdot)$). Similar to a classic autoencoder, the encoder, E_n, produces a latent code z_i for image $x_i \sim p$. Furthermore, we divide the latent code, z_i, into two vectors: class specific component, $z_i^{spec_l}$ for class l, and a class agnostic component, $z_i^{agn_l}$. This is achieved by having two heads instead of one (as in conventional architectures). Both vectors are combined and fed to the decoder, G_n, which reconstructs the original input. The disentanglement network is trained using the following loss:

$$\mathcal{L}_{Disent} = \mathcal{L}_{Rec} + \lambda_1 \mathcal{L}_{spec} + \lambda_2 \mathcal{L}_{agn} + \lambda_3 \mathcal{L}_{agn-spec} \tag{1}$$

Reconstruction Loss: $\mathcal{L}_{Rec}$, is the commonly used image reconstruction loss: $\mathcal{L}_{Rec} = \sum_l \mathbb{E}_{x_i \sim p_l} \left[\left\| x_i^l - G_l(E_l(x_i^l)) \right\| \right]$. It is a sum of the reconstruction losses from the class specific autoencoders. We train different autoencoders for images of each class in order to obtain class specific features and refer to them as ‘Class-specific autoencoders’.

Class Specific Loss: For given class l the class specific component $z_i^{spec_l}$ will have high similarity with samples from the same class and low similarity with the $z_i^{spec_k}$ of other classes k. These two conditions are incorporated as follows:

$$\mathcal{L}_{spec} = \sum_{i,j} \left(\sum_{l} \left(\left(1 - \langle z_i^{spec_l}, z_j^{spec_l} \rangle \right) + \sum_{k \neq l} \langle z_i^{spec_l}, z_j^{spec_k} \rangle \right) \right) \tag{2}$$

where $\langle . \rangle$ denotes cosine similarity. The sum is calculated for all classes indexed by $\sum_l$ and over all samples indexed by i, j.

Class Agnostic Loss: Class agnostic features of different classes have similar semantic content and have high cosine similarity. L_{agn} is defined as

$$\mathcal{L}_{agn} = \sum_{i,j} \sum_{l} \sum_{k \neq l} \left(1 - \langle z_i^{agn_l}, z_j^{agn_k} \rangle \right) \tag{3}$$

We want class specific and class agnostic features of same-class samples to be mutually complementary and have minimal overlap in semantic content, i.e.,

$$\mathcal{L}_{agn-spec} = \sum_{l} \langle z_i^{agn_l}, z_j^{spec_l} \rangle \tag{4}$$

Since the above loss terms are minimized it helps us achieve our stated objectives.

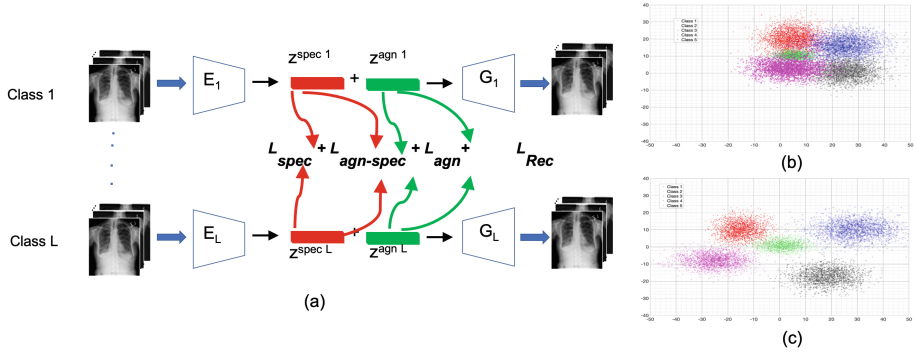

Fig. 1. (a) Architecture of class specific feature disentanglement network. Given training images from different classes of the same domain we disentangle features into class specific and class agnostic using autoencoders. T-sne results comparison between original image features and feature disentanglement output: (b) Original image features; (c) Class specific features. The classes in the tsne plot correspond to Atelectasis, Consolidation, Effusion, Infiltration and Nodule, as per the standard classes used for CheXpert.

Figure 1 (b) shows the t-sne plots of image features (taken from the fully connected layer of a multi-label DenseNet-121 image classifier) while Fig. 1 (c) shows the plot using class specific features. Plots of original features show overlapping clusters which makes it challenging to have good classification. Clusters obtained using class specific features are well separated with minimal overlap between different clusters. This clearly demonstrates the efficacy of our feature disentanglement method. The features are taken from images belonging to 5 classes from the NIH dataset. We chose 5 classes to clearly demonstrate the output and avoid cluttering.

Feature Generation Network: After disentangling the different seen class samples into their class specific components we create a distribution of each seen class feature. We generate synthetic class specific features of unseen classes using the following approach inspired by Mixup [28]:

$$\boldsymbol{z}^{spec_U} = \sum_l \Lambda_l \boldsymbol{z}_l^{spec_S}; \hat{y} = \sum \boldsymbol{y}_l \tag{5}$$

where $\boldsymbol{z}_k^{spec_U}$ is the class specific synthetic vector for unseen classes $k(\neq l)$, $\boldsymbol{z}_l^{spec_S}$ is a feature sampled from the distribution of seen class l, Λ_l is a random number drawn from a beta distribution. $\hat{y}$ is a one-hot encoded vector and is a sum of the one-hot label vectors of individual classes. Hence we do not need a weight when combining the label vectors. The weights Λ_l are such that $\sum_l \Lambda_l = 1$.

Generating unseen class features through Mixup without additional constraints can generate unrealistic features. We use the dictionary of text embeddings to guide the feature generation process. As synthetic features of the seen and unseen classes are generated we cluster them using the online self supervised learning based SwAV method [3] and calculate the centroids of each cluster. The semantic similarity of the centroid clusters should be such that their cosine similarity values are close to those obtained in Table 1, i.e., we define a loss:

$$\mathcal{L}_{ML-Cluster} = \frac{1}{N^2} \sum_i \sum_j Dict_{Text}(i,j) - Cent_{All}(i,j) \qquad (6)$$

where $Cent_{All}$ refers to the changing matrix of cluster centroid similarities for all seen and unseen classes. N is the total number of classes. The final loss term for **clustering** all class samples is $\mathcal{L}_{Clust} = \mathcal{L}(x_s, x_t) + \lambda_4 \mathcal{L}_{ML-Cluster}$ where $\mathcal{L}(x_s, x_t)$ is the SwAV loss function defined in [3]. We add only those synthetic samples to classifier training data that reduce $\mathcal{L}_{Clust}$. This formulation ensures that the cluster output is well separated semantically and the cluster centroids follow the semantic relationship between all classes in Table 1.

Training, Inference and Implementation: For a given test image we use the pre-trained L class specific autoencoders to get the class specific features. An input 256 × 256 image is passed through the Encoder having 3 convolution layers ($64, 32, 32$ 3×3 filters) each followed by max pooling. The Decoder is symmetric to the Encoder. z^{agn} and z^{spec} are 256-dimension vectors. We then calculate the cosine similarity of the class specific features with the corresponding class centroids. If the cosine similarity is above 0.5 then the sample is assigned to the class. Following standard practice for GZSL, average class accuracies are calculated for the seen (Acc_S) and unseen (Acc_U) classes, and also the harmonic mean $H = \frac{2 \times Acc_U \times Acc_S}{Acc_U + Acc_S}$.

3 Experimental Results

Dataset Description. We demonstrate our method's effectiveness on the following chest xray datasets for multi-label classification tasks: **1.NIH Chest X-ray** Dataset [24]: having 112,120 expert-annotated frontal-view X-rays from 30,805 unique patients and has 14 disease labels. Original images were resized to 224 × 224. Hyperparameter values are $\lambda_1 = 1.1, \lambda_2 = 0.7, \lambda_3 = 0.9, \lambda_4 = 1.2$. **2. CheXpert** Dataset [10]: consisting of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest conditions. Original images were resized to 224 × 224. Hyperparameter values are $\lambda_1 = 1.2, \lambda_2 = 0.8, \lambda_3 = 1.1, \lambda_4 = 1.1$. **3. PadChest** Dataset [2]: consisting of 160,868 from 67,625 patients. Hyperparameter values are $\lambda_1 = 1.3, \lambda_2 = 0.9, \lambda_3 = 0.9, \lambda_4 = 1.3$. A 70/10/20 split at patient level was done to get training, validation and test sets for both datasets.

Comparison Methods: We compare our method's performance with multiple GZSL methods - single label and multi-label techniques - employing different feature generation approaches such as CVAE or GANs. Our method is denoted as ML-GZSL (**M**ulti **L**abel **GZSL**). Our benchmark is a fully supervised learning (FSL) based method of [27] which is the top ranked method for [10], where the ranking is based on AUC. It builds upon a DenseNet-121 trained for multi-label classification.

3.1 Generalized Zero Shot Learning Results

Classification results for medical images in Table 2 show our proposed method significantly outperforms all competing GZSL methods. Note that we use the cluster centroids in place of attribute vectors for these feature synthesis methods. This significant difference in performance can be explained by the fact that the complex architectures that worked for natural images will not be equally effective for medical images which have less information. Absence of attribute vectors for medical images is another contributing factor. The class attributes provide a rich source of information about natural images which can be leveraged using existing architectures. Since those are not available for medical images these methods do not perform equally well. Different combinations of 7 seen and unseen classes are taken, and for each combination we run our model 5 times and the final reported numbers are the average of different combinations.

Table 2. GZSL Results For chest xray Images in Multi-Label setting: Average per-class classification accuracy (%) and harmonic mean accuracy (H) of generalized zero-shot learning when test samples are from seen or unseen classes. Results demonstrate the superior performance of our proposed method.

Method	NIH X-ray				CheXpert				PadChest			
	S	U	H	p	S	U	H	p	S	U	H	p
	Single Label GZSL Methods											
f-VAEGAN [26]	82.9	80.0	81.4	0.002	88.5	87.6	88.0	0.001	81.0	78.4	79.7	0.001
SDGN [25]	84.4	81.1	82.7	0.003	89.8	88.3	89.0	0.003	82.3	80.0	81.1	0.004
Feng [6]	84.7	81.4	83.0	0.0012	90.2	88.6	89.4	0.0017	82.5	80.2	81.3	0.0021
Kong [12]	84.8	81.2	82.9	0.0031	90.0	88.7	89.3	0.0034	82.7	80.5	81.6	0.0029
Su [22]	84.5	81.4	82.9	0.004	90.3	88.6	89.4	0.0045	82.3	79.8	81.03	0.0041
	Multi Label GZSL Methods											
Hayat [8]	79.1	69.2	73.8	0.005	81.2	79.8	80.5	0.0056	77.3	68.1	72.4	0.006
Lee [14]	85.1	81.3	83.1	0.008	87.4	85.7	86.5	0.0075	82.9	78.4	80.6	0.008
Huynh [9]	84.7	80.8	82.7	0.0065	86.9	85.1	86.0	0.0071	82.5	77.3	79.8	0.0073
	Proposed Method And Benchmarks											
ML-GZSL	86.2	85.0	85.6	–	90.8	90.2	90.5	–	88.2	86.1	87.1	–
FSL(Multi Label)	86.0	85.1	85.5	0.061	90.8	90.5	90.6	0.068	88.4	86.5	87.4	0.058
Mahapatra [18]	84.3	83.2	83.7	0.014	88.9	88.5	88.7	0.01	86.2	84.1	85.1	0.02

ML-GZSL's performance is almost equal to that of the benchmark fully supervised method FSL. Although GZSL methods generally perform inferior to FSL methods, our use of class specific features significantly improves performance. Additionally, the use of semantic relation between text embeddings significantly improves the performance due to better feature synthesis. The average accuracy is obtained by first calculating True Positive, False Positive, True Negative, False Negative values and using these values to get the global accuracy. Furthermore the AUC(and F1) values for CheXpert data are as follows: FSL-93.0(91.7), ML-GZSL- 92.8(91.6), [18]-91.9(90.0), [8]-84.3(82.4).

3.2 Ablation Studies

Table 3 shows results for ablation studies. We exclude each of the three loss terms related to feature disentanglement - $\mathcal{L}_{agn}$,$\mathcal{L}_{spec}$ and $\mathcal{L}_{agn-spec}$- and report the results as ML-GZSL$_{w\mathcal{L}_{agn}}$, ML-GZSL$_{w\mathcal{L}_{spec}}$, and ML-GZSL$_{w\mathcal{L}_{agn-spec}}$. We also compare with the results of using image features obtained from a CNN based feature extractor (ResNet50 trained on Imagenet), which we denote as 'pre-train'. We observe that the class specific features has the greatest influence on the results and excluding it, ML-GZSL$_{w\mathcal{L}_{spec}}$, results in significant performance degradation compared to ML-GZSL. ML-GZSL$_{w\mathcal{L}_{agn-spec}}$ and ML-GZSL$_{w\mathcal{L}_{agn}}$ also show significantly lower performance. These results highlight the importance of the class specific features and at the same time illustrate class agnostic features have an important influence on the method's performance.

We also investigate the influence of $\mathcal{L}_{ML-Cluster}$ (Eq. 6) in the clustering process. The numbers in Table 3 show that ML-GZSL$_{w\mathcal{L}_{ML-Cluster}}$ (which is essentially the original SwAV algorithm) performs much worse. This proves the significant contribution of the text embedding dictionary in our multi-label GZSL framework.

Table 3. Ablation Results: Average per-class classification accuracy (%) and harmonic mean accuracy (H) of generalized zero-shot learning when test samples are from seen (Setting S) or unseen (Setting U) classes.

Method	NIH X-ray				CheXpert				PadChest			
	S	U	H	p	S	U	H	p	S	U	H	p
	Our Proposed Method											
ML-GZSL	86.2	85.0	85.6	-	90.8	90.2	90.5	-	88.2	86.1	87.1	-
	Feature Disentanglement Effects											
$w\mathcal{L}_{agn-spec}$	83.8	81.9	82.8	0.012	88.6	86.3	87.4	0.009	85.5	82.0	83.7	0.014
pre-train	83.4	82.0	82.7	0.017	88.2	85.3	86.7	0.009	85.1	81.7	83.4	0.011
$w\mathcal{L}_{agn}$	84.5	82.1	83.3	0.008	89.1	86.9	88.0	0.0094	86.5	83.4	84.9	0.011
$w\mathcal{L}_{spec}$	84.0	82.2	83.1	0.02	88.8	86.2	87.5	0.018	86.1	83.0	84.5	0.014
	Effect of Text Dictionary											
$w\mathcal{L}_{ML-Cluster}$	82.6	80.7	81.6	0.009	87.0	84.5	85.7	0.011	84.2	80.8	82.5	0.015

Hyperparameter Selection: The λ's were varied between $[0.4-1.5]$ in steps of 0.05 and the performance on a separate test set of $10,000$ images were monitored. We optimize Eq. 1 by setting $\lambda_2 = \lambda3 = \lambda_4 = 1$, and select the optimum value of λ_1. After fixing λ_1 we determine optimal λ_2, and subsequently λ_3, λ_4.

Realism of Synthetic Features. We reconstruct the xray images from the synthetic feature vectors using the feature disentanglement autoencoders' decoder part. We select 1000 such synthetic images from 14 classes of the NIH dataset and ask two trained radiologists, having 12 and 14 years experience in examining chest xray images for abnormalities, to identify whether the images are realistic or not. Each radiologist was blinded to the other's answers.

Results for ML-GZSL show one radiologist (RAD 1) identified 912/1000 (91.2%) images as realistic while RAD 2 identified 919 (91.9%) generated images as realistic. Both of them had a high agreement with 890 common images (89.0%) identified as realistic. Considering both RAD 1 and RAD 2 feedback, a total of 941 (94.1%) unique images were identified as realistic and 59/1000 (5.9%) images were not identified as realistic by any of the experts. ML-GZSL showed the highest agreement between RAD 1 and RAD 2.

4 Conclusion

Our experiments demonstrate that our approach of multi label GZSL is more accurate than using conventional approaches that solve the single-label scenario. We propose a novel feature disentanglement approach that obtains class specific and class agnostic features from the training images. Additionally, the relationship between text embeddings of disease labels is used to create a dictionary that guides clustering and feature synthesis. Classification results on multiple publicly available chest xray datasets demonstrate the improved performance obtained by using class specific features. The synthetic features obtained by our method are realistic since a major percentage of the corresponding reconstructed images are validated as realistic by trained clinicians.

References

1. Bian, C., Yuan, C., Ma, K., Yu, S., Wei, D., Zheng, Y.: Domain adaptation meets zero-shot learning: an annotation-efficient approach to multi-modality medical image segmentation. IEEE Trans. Med. Imaging **41**(5), 1043–1056 (2022)
2. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: PadChest: A large chest x-ray image dataset with multi-label annotated reports. Med. Image Anal. **66**, 101797 (2020)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 9912–9924. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf
4. Chen, Y., et al.: Zero-shot medical image artifact reduction. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 862–866 (2020). https://doi.org/10.1109/ISBI45749.2020.9098566
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint: arXiv:1810.04805 (2018)

6. Feng, Y., Huang, X., Yang, P., Yu, J., Sang, J.: Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9336–9345 (2022)
7. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016). https://doi.org/10.1001/jama.2016.17216
8. Hayat, N., Lashen, H., Shamout, F.: Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In: Proceeding of the Machine Learning for Healthcare Conference, pp. 461–477 (2021)
9. Huynh, D., Elhamifar, E.: A shared multi-attention framework for multi-label zero-shot learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8773–8783 (2020). https://doi.org/10.1109/CVPR42600.2020.00880
10. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint: arXiv:1901.07031 (2017)
11. Jia, J., He, F., Gao, N., Chen, X., Huang, K.: Learning disentangled label representations for multi-label classification (2022). https://doi.org/10.48550/arXiv.2212.01461
12. Kong, X., et al.: En-compactness: self-distillation embedding and contrastive generation for generalized zero-shot learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9296–9305 (2022). https://doi.org/10.1109/CVPR52688.2022.00909
13. Kori, A., Krishnamurthi, G.: Zero shot learning for multi-modal real time image registration. arXiv preprint: arXiv:1908.06213 (2019)
14. Lee, C.W., Fang, W., Yeh, C.K., Wang, Y.C.F.: Multi-label zero-shot learning with structured knowledge graphs. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1576–1585 (2018). https://doi.org/10.1109/CVPR.2018.00170
15. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
16. Mahapatra, D., Bozorgtabar, B., Kuanar, S., Ge, Z.: Self-supervised multimodal generalized zero shot learning for Gleason grading. In: Albarqouni, S., et al. (eds.) DART/FAIR -2021. LNCS, vol. 12968, pp. 46–56. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87722-4_5
17. Mahapatra, D., Bozorgtabar, B., Ge, Z.: Medical image classification using generalized zero shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 3344–3353 (2021)
18. Mahapatra, D., Ge, Z., Reyes, M.: Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. IEEE Trans. Med. Imaging **41**(9), 2443–2456 (2022). https://doi.org/10.1109/TMI.2022.3163232
19. Mensink, T., Gavves, E., Snoek, C.G.: COSTA: co-occurrence statistics for zero-shot classification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2441–2448 (2014). https://doi.org/10.1109/CVPR.2014.313
20. Park, T., et al.: Swapping autoencoder for deep image manipulation. In: Advances in Neural Information Processing Systems (2020)
21. Paul, A., et al.: Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training. IEEE Trans. Med. Imaging **40**, 2642–2655 (2021). https://doi.org/10.1109/TMI.2021.3054817

22. Su, H., Li, J., Chen, Z., Zhu, L., Lu, K.: Distinguishing unseen from seen for generalized zero-shot learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7875–7884 (2022). https://doi.org/10.1109/CVPR52688.2022.00773
23. Verma, V., Arora, G., Mishra, A., Rai, P.: Generalized zero-shot learning via synthesized examples. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4281–4289 (2018)
24. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the CVPR (2017)
25. Wu, J., Zhang, T., Zha, Z.J., Luo, J., Zhang, Y., Wu, F.: Self-supervised domain-aware generative network for generalized zero-shot learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12767–12776 (2020)
26. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: F-VAEGAN-D2: a feature generating framework for any-shot learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10275–10284 (2019)
27. Yuan, Z., Yan, Y., Sonka, M., Yang, T.: Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3020–3029 (2021)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=r1Ddp1-Rb