Semi-supervised Learning for Multi-label Classification[*]

Liyue Shen
Stanford University
liyues@stanford.edu

Ruiyang Song
Stanford University
ruiyangs@stanford.edu

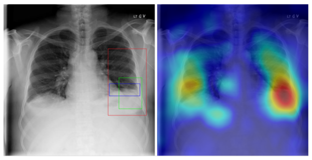
Abstract

In this report we consider the semi-supervised learning problem for multi-label image classification, aiming at effectively taking advantage of both labeled and unlabeled training data in the training process. In particular, we implement and analyze various semi-supervised learning approaches including a support vector machine (SVM) method facilitated by principal component analysis (PCA), and a self-training method that iteratively conducts supervised learning and enlarges the set of training labels on the go. We compare the performances of semi-supervised learning methods with supervised learning benchmarks, and introduce a heuristic performance analysis for the training process. In addition, we analyze the impact of different training parameters for the PCA-SVM and the self-training method on the prediction performance. The algorithms are implemented for the ChestX-ray14 [32] medical image dataset.

1. Introduction

The recent progress in deep learning research has significantly improved the performance of various computer vision tasks for natural images including image classification [16, 29, 30, 12], object detection [9, 27] and instance segmentation [19] benefited from the availability of a large volume of labeled natural image datasets. It is expected that computer vision methods based on supervised learning will also contribute to medical image applications such as early-stage cancer detection and image diagnosis. However, challenges emerge as it is difficult to construct large densely labeled medical datasets since manually annotating medical images requires medical and clinical expertise. In other words, labeled medical data are often much more expensive than unlabeled data.

Radiology report	Keyword	Localization Result

findings include: 1. left basilar atelectasis/consolidation. 2. prominent hilum (mediastinal adenopathy). 3. left pic catheter (tip in atriocaval junction). 4. stable, normal appearing cardiomediastinal silhouette.
impression: small right pleural effusion otherwise stable abnormal study including left basilar infiltrate/atelectasis, prominent hilum, and position of left pic catheter (tip atriocaval junction).
	Effusion;
	Infiltration;
	Atelectasis
		

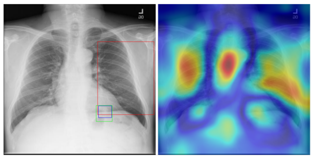
findings: pa and lateral views of the chest demonstrate stable 2.2 cm nodule in left lower lung field posteriorly. the lungs are clear without infiltrate or effusion. cardiomediastinal silhouette is normal size and contour. pulmonary vascularity is normal in caliber and distribution. impression: stable left likely hamartoma.	Nodule; Infiltration	

Figure 1: Two multi-label chest X-ray image samples from ChestX-ray14 dataset [32] with radiology report, disease keywords extraction and localization results.

Two primary approaches have been introduced to tackle the medical image recognition task on a dataset with few labeled samples: semi-supervised learning and transfer learning. Semi-supervised learning considers a prediction problem with only a small number of labeled training data by exploiting the information provided by both labeled and unlabeled data [2]. The labeled data will provide information for joint distribution of samples and labels, while the unlabeled data provides information of the distribution of samples [35]. Multiple approaches have been developed and widely applied in computer vision, and a comprehensive literature review can be found in [35].

One of the notable methods for semi-supervised learning is the self-training technique, which is first introduced in [20]. The main idea of self-training is to iteratively apply a supervised learning algorithm based on the currently available training labels and include the predicted examples with high confidence scores into the updated training set. In this manner, more information of the originally unlabeled data is incorporated into the classifier after each iteration as the training label set is enlarged in each step.

Transfer learning is another method widely used in computer vision scenarios to overcome the challenge of insuf-

[*] This is a joint project with CS331B. The two projects focus on different methods and algorithms to solve the semi-supervised learning problem. In particular, the CS229 project primarily concentrates on the PCA-SVM and the self-training method, while the CS331B project concentrates on the supervised deep learning approaches and the CNN-ladder networks. Some materials and experiment results (e.g. for baseline and comparison purpose) are shared in the two project reports. The projects focus on fundamentally different major approaches and provide independent analyses.

ficient labeled training data. When the unlabeled training dataset is small, transfer learning methods extract information from the trained model of a different large dataset to facilitate the current task. A literature survey on transfer learning is presented in [21].

This project studies a medical image classification problem based on the ChestX-ray14 [32] dataset that contains over 100 thousand front-view X-ray images with annotations of 14 thoracic diseases. We first propose a deep learning approach with supervised learning. Then we introduce semi-supervised methods using machine learning based on principal component analysis (PCA) and support vector machine (SVM). In addition, we study a self-training approach where in each iteration we perform supervised deep learning algorithms and enlarge the labeled training sets. We compare their performances with the supervised learning benchmarks and the performance of the semi-supervised ladder network approach.

2. Related Work

Deep Learning in Medical Imaging. One of the first successful approaches of applying deep neural networks to biomedical imaging is [4]. Recent works studied the applications including skin cancer classification [8], breast cancer diagnosis [22], brain tumor segmentation [31], and lung nodule detection [3], where deep learning methods have shown good experimental performance. An overview of the recent progress is summarized in [34].

Semi-supervised Learning. Empirical results show that semi-supervised learning improves the performance compared to supervised learning that only exploits labeled data [10]. Primary methods for solving semi-supervised learning include generative models [13], self-training [28], transductive SVMs [1], entropy regularization [10], and graph-based models [36]. [11] and [14] consider a semi-supervised image classification problem with a variational inference algorithm based on deep generative models. Another deep neural network based method is the ladder network [24]. In [18], a self-training support vector machine (SVM) algorithm is studied. In [28], an object detection problem is studied with the self training expecation maximization (EM) method.

3. Methods & Experiments

This project primarily focuses on the thoracic disease classification problem based on X-ray image data, which can be formulated as a multi-label problem since each sample possibly has multiple diseases simultaneously.

In this section, we first briefly introduce the ChestX-ray14 dataset, then describe the methods we apply: the ResNet [12] model which is a supervised baseline, the SVM-PCA method, the self-training approach, and the ladder network for comparison.

3.1. ChestX-ray14 Dataset

The ChestX-ray14 [32] dataset illustrated in Figure 1 is currently the largest chest X-ray database that contains 112,120 frontal-view X-ray images from 32,717 patients with 14 common thoracic disease categories labeled by text mining radiology reports. In ChestX-ray14, 60,412 samples are healthy and 51,708 samples have (possibly multiple) thoracic diseases. Following the experiment setting in [32], we randomly choose 78,484 images (70%) used for training, 11,212 images (10%) for validation and 22,424 images (20%) for testing. In [32], a multi-label classification benchmark is also presented.

3.2. Supervised Learning

3.2.1 ResNet model for transfer learning

To begin with, we apply the deep residual network (ResNet) model for transfer learning and particularly choose the ResNet-18 and ResNet-50 models inspired by [12].

For our multi-label classification setting, we adjust the original ResNet model that is supposed for single-label classification using the multi-label soft margin loss when training the network.

Denote by $\mathcal{C}$ the collection of categories. Let $T \in \{0,1\}^{|\mathcal{C}|}$ be the actual image label and $Y \in \mathbb{R}^{|\mathcal{C}|}$ be the network prediction. The training loss can be formulated as follows:

$$L(Y,T) = -\sum_{i=1}^{|\mathcal{C}|} \left[t_i \log \frac{1}{1+e^{-y_i}} + (1-t_i) \log \frac{e^{-y_i}}{1+e^{-y_i}} \right].$$

The final predicted probability is then derived by applying the sigmoid activation function on the confidence score vector.

We use the ImageNet pre-trained model as the initialization to train the ResNet model with the idea of transfer learning. The experiment results of supervised learning and transfer learning are shown in Table 1. Here, the area-under-curve (AUC) score is applied as the metric. compared with benchmark results [32] and previous work [33, 23], our method gets the state-of-the-art results in average AUC scores as well as most disease classes. We refer to [5] for the implementation of the ResNets.

3.3. Semi-supervised Learning

3.3.1 PCA-SVM baseline model

We consider a baseline machine learning model combining PCA and SVM. First, we preprocess the X-ray images of original size 1024×1024 by resizing them into 128×128. We apply the PCA approach to reduce the data dimension of the flattened image vector to a dimension of 2500. Then

Method	**Benchmark** [32]	**DenseNet-LSTM** [33]	**CheXNet** [23]	**Ours - ResNet-18 (Fine-tune)**	**Ours - ResNet-50 (Fine-tune)**
Atelectasis	0.7158	0.772	0.8209	0.8190	**0.8276**
Cardiomegaly	0.8065	0.904	**0.9048**	0.8998	0.9013
Effusion	0.7843	0.859	0.8831	0.8881	**0.8903**
Infiltration	0.6089	0.695	0.7204	0.7165	**0.7229**
Mass	0.7057	0.792	0.8618	0.8534	**0.8690**
Nodule	0.6706	0.717	0.7766	0.7738	**0.7884**
Pneumonia	0.6326	0.713	**0.7632**	0.7593	0.7588
Pneumothorax	0.8055	0.841	0.8932	0.8934	**0.9033**
Consolidation	0.7078	0.788	0.7939	0.8116	**0.8178**
Edema	0.8345	0.882	0.8932	0.9061	**0.9106**
Emphysema	0.8149	0.829	**0.926**	0.9083	0.9198
Fibrosis	0.7688	0.767	0.8044	0.8149	**0.8197**
PT	0.7082	0.765	**0.8138**	0.8007	0.8048
Hernia	0.7667	0.914	**0.9387**	0.8822	0.8700
NoFinding	-	0.762	-	0.7229	**0.7894**
Average	0.7379	0.798	0.8424	0.8377	**0.8432**

Table 1: Per-class AUC scores of ROC curves for multi-label classification on ChestX-ray14 dataset, which present the quantitative performance of different models with or without transfer learning in fully-supervised learning setting.

we use the SVM model to classify the image vectors into normal and abnormal categories. In this way, we train a binary classifier for each of the 14 disease categories.

The number of components selected for the PCA is crucial for the classifier performance. We begin with a low-dimensional PCA and found that the subsequently trained SVM classifier is not able to discriminate images of different classes, and all the prediction outputs are the same. This might be because of the fact that in a low-dimensional space, the images of different classes are mixed together and can not be separated by a SVM model effectively. Besides, we realized that it is important to balance the training samples. Recall that in the semi-supervised learning, the number of training samples is very small. If only a small amount of positive examples are randomly chosen for training, it is hard for the SVM to give a good boundary since the chosen positive examples only represent a partial distribution of the positive samples.

We apply the PCA-SVM method trained on a training set comprising of 2000 labeled samples and evaluate the model in the testing dataset (22,424 images, which is 20% of the whole ChestX-ray database). We use different PCA dimensions ranging from 1000, 2000, and 5000. The testing performance of ROC AUC scores are presented in Table 2. We consult [6] for the implementation of the PCA-SVM algorithm.

In order to explore the impact of the number of labeled training samples and the number of components in the PCA on the testing performance, we compare the performances with the number of training labels ranging from $\{0.1, 1, 2, 5, 10, 15, 20, 25, 50\} \times 10^3$, and the PCA dimension ranging from $\{20, 100, 500, 1000, 1500, 2000, 2500, 5000\}$. The experiment result is shown in Fig. 2. We see that the performance is improved as the number of active dimensions in the PCA step increases. The classifier performs poorly when the PCA dimension is as low as 20 and 100. We also observe that when the number of active PCA components is relatively low at 20, 100, and 500, the AUC scores can decrease as the number of training labels increases. Note that when the PCA dimension is low, a large number of training labels may cause overfitting and jeopardize the testing performance. Approximately 2000 PCA components and 2000 training labels will be sufficient for a good performance, above which increasing the number of training labels and the number of PCA components does not improve the scores significantly.

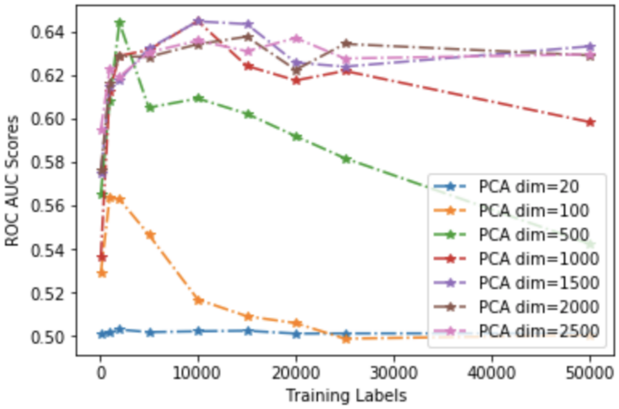


Figure 2: Semi-supervised learning performance of PCA-SVM model on the test dataset for classification of disease class ”Fibrosis” with different number of training labels and various PCA reduction dimensions.

3.3.2 Self-training

In this subsection we study the self-training approach that extracts information from both the labeled and unlabeled training data through sequentially training a network with new labels added based on previous iterations. While the general theoretic analysis of the performance of self-training algorithms can be difficult [35], here we try to provide a heuristic analysis.

Suppose the training data are from a set $\mathcal{X}$ and the underlying correct classification labels are from a set $\mathcal{Y}$. We have m_0 labeled training samples in $\mathcal{X}_0 \subset \mathcal{X}$ with labels $\mathcal{Y}_0$ and $m_1 = M - m_0$ unlabeled training samples in $\mathcal{X}_1 \subset \mathcal{X}$. We start with m_0 training labels and implement a supervised learning algorithm with a particular network architecture and derive confidence scores $\mathcal{S}$ for the $M - m_0$ unlabeled samples. Afterwards, we select c unlabeled samples with the highest confidence measures denoted by $\mathcal{S}_c = \{p_1, \dots, p_c\}$ respectively, and incorporate them into the labeled training set together with their predicted labels $\hat{\mathcal{Y}}_c$. This training process continues sequentially as the label set gets enlarged gradually. In this application, we implement a fine tuned ResNet-50 network in each iteration as the supervised learning method. The self training approach is summarized in Algorithm 1.

Suppose the relationship between the samples and the underlying classification result can be described with a mapping $f : \mathcal{X} \mapsto \mathcal{Y}$. If we are given correct labels for all the training data, we will arrive at a learnt mapping $\hat{f}$. However, in the self-training approach, the c picked predictions in each step may bias the learning result from $\hat{f}$ due to possible incorrect predictions, which we roughly view as noise. If we approximate the probability that the selected predictions are correct with their respective confidence scores $p_1, \dots, p_c$, for stochastic gradient descent, each newly selected label will incur noise with mean in the order of $\epsilon_i = (1 - p_i)\alpha$, $i = 1, \dots, c$.

Algorithm 1: Self-training method with ResNet

```
initialize $\mathcal{X}_{\text{label}} := \mathcal{X}_0$, $\mathcal{Y}_{\text{label}} := \mathcal{Y}_0$, $\alpha$, $c$;
repeat
    $(\hat{\mathcal{Y}}_{\text{predict}}, \mathcal{S}_{\text{predict}})$ := ResNet($\mathcal{X}_{\text{label}}, \mathcal{Y}_{\text{label}}, \alpha$);
    $(\mathcal{X}_c, \hat{\mathcal{Y}}_c, \mathcal{S}_c)$ := findmax$_c(\mathcal{S}_{\text{predict}})$;
    $\mathcal{X}_{\text{label}} := \mathcal{X}_{\text{label}} \bigcup \mathcal{X}_c$, $\mathcal{Y}_{\text{label}} := \mathcal{Y}_{\text{label}} \bigcup \hat{\mathcal{Y}}_c$.;
until iterations finished;
```

One of the most important parameters besides the learning rate α is the selection parameter c, i.e. the number of predicted labels added to the training label set in each round. In Figure 3, we present the performances of the self-training method with c ranging in $\{1, 2, 3, 4, 5\}$ with learning rate $\alpha = 0.01$. We see that empirically for this application, $c = 3$ or 4 will be good choices.

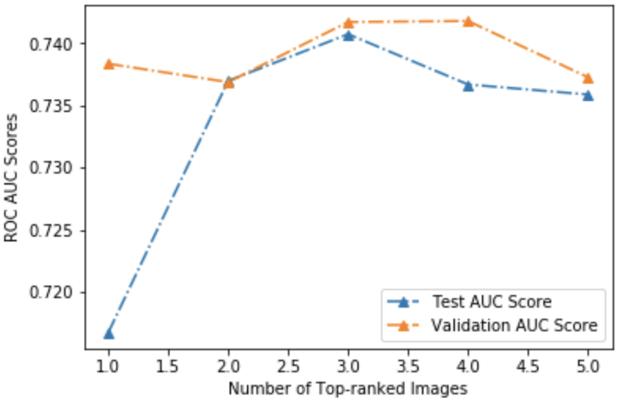


Figure 3: Performance of the self-training approach with different selection parameter c's. The model is based on fine-tuning the ResNet-50 network and trained with 2000 images.

The performance of the self-training method is presented in Table 2 and 3. In Table 2, we compare the per-class AUC scores of self-training based on ResNet18 and ResNet50 using 2000 training labels with other approaches. In Table 3, we compare the performances of self-training with 1000, 2000, and 5000 labels respectively with other methods. Both empirical examples show that self-training can be applied to improve the classification performances. It is shown to outperform the CNN ladder network for most classes and is usually better than or close to the benchmark.

3.3.3 CNN-based ladder network

Another approach for semi-supervised learning is the ladder network [24] proposed based on the stacked denoising autoencoders [25, 26], where a noisy encoder feed forward path and a corresponding denoising decoder path are added for learning unlabeled training data to the normal feed forward network through additional lateral connections. The performance of ladder networks for the MNIST dataset [17] and the CIFAR-10 dataset [15] have been evaluated in previous work.

To construct ladder network for chest X-ray image classification, we propose a 11-layer CNN model with increasing feature map channels and decreasing convolution kernels in the feedforward encoder path. And the final layer is a global mean pooling layer to aggregate the spatial image information. The experiment results for CNN-based ladder network is shown in Table 2 for comparison purpose. We refer to the github repository in [7] for the implementation of ladder networks.

Method	Bench-mark [32]	PCASVM (1000)	PCASVM (2000)	PCASVM (5000)	ResNet18	ResNet50	ResNet18ST	ResNet50ST	CNN-Ladder
ATE	0.7158	0.605	0.619	0.629	0.7273	0.7084	**0.7312**	0.7120	0.4941
CARD	0.8065	0.655	0.689	0.704	0.7501	0.7334	0.7453	0.7404	**0.7939**
EFF	0.7843	0.673	0.674	0.686	0.8104	0.8118	**0.8123**	0.8043	0.7395
INFI	0.6089	0.602	0.598	0.609	0.6593	0.6500	**0.6624**	0.6552	0.5803
Mass	0.7057	0.554	0.558	0.579	**0.6758**	0.6729	0.6687	0.6646	0.6108
NOD	0.6706	0.529	0.539	0.551	**0.6511**	0.6386	0.6364	0.6436	0.5312
PNA	0.6326	0.592	0.590	0.612	0.6792	0.6711	0.6823	**0.6861**	0.5907
PTX	0.8055	0.610	0.616	0.625	0.7803	0.7735	0.7755	**0.7821**	0.6630
CON	0.7078	0.667	0.672	0.685	**0.7699**	0.7598	0.7616	0.7645	0.6711
Edema	0.8345	0.754	0.761	0.780	0.8573	0.8559	0.8594	**0.8659**	0.8170
EMPH	0.8149	0.582	0.596	0.616	0.7570	0.7789	**0.7806**	0.7677	0.6879
FIB	0.6158	0.614	0.618	0.632	0.7263	0.7285	0.7307	**0.7377**	0.6158
PT	0.7082	0.575	0.584	0.600	0.6971	0.6982	**0.7088**	0.6964	0.6411
Hernia	0.7667	0.677	0.661	0.649	0.8179	0.8145	0.8219	**0.8497**	0.7007
Avg.	0.7379	0.6206	0.6268	0.6398	0.7399	0.7354	**0.7412**	0.7407	0.6527

Table 2: Per-class AUC scores of different semi-supervised methods with the same 2000 labeled training data for multi-label classification on ChestX-ray14 dataset. "ST" denotes self-learning approach. The per-class scores for the PCA-SVM method with 1000 and 5000 labels are also included for comparison.

Average AUC score with # of used labels	1000	2000	5000	All (78484)
Benchmark [32]	-	-	-	0.7379
DenseNet-LSTM [33]	-	-	-	0.798
CheXNet [23]	-	-	-	0.8424
Baseline (PCA+SVM)	0.6206	0.6268	0.6398	-
ResNet-18 (Fine-tune)	0.7015	0.7399	0.7744	0.8377
ResNet-18 (Fine-tune) + self-training	0.7062	**0.7412**	0.7721	-
ResNet-50 (Fine-tune)	0.7052	0.7354	0.7752	**0.8432**
ResNet-50 (Fine-tune) + self-training	**0.7088**	0.7407	**0.7783**	-

Table 3: Average AUC scores of all 14 thoracic categories for multi-label classification on ChestX-ray14 dataset demonstrating the quantitative performance of different semi-supervised learning approaches including self-training.

4. Conclusion

In this paper, we study the problem of multi-label image classification based on the frontal chest X-ray image dataset ChestXray14. We conduct a machine learning approach based on PCA-SVM, and a self-training method based on ResNet in each iteration, and compare them with the performances of several benchmarks and the CNN networks. In particular, we studied the impact of PCA dimension in the PCA-SVM method and the influence of the selection parameter in self-training approach. We find that in the particular classification problem considered in this report, self-training usually exhibits good performance scores comparing to other benchmarks and methods. Besides, we introduce an intuitive understanding of the performance analysis of self-training. Future works include the rigorous theoretic study of the performance of self-training, and improvements to the network structures for better capturing information from unlabeled data.

5. Contribution

Shen is responsible for the introduction and literature review of computer vision and medical imaging with deep learning, and Song is responsible for the introduction and survey of semi-supervised learning and transfer learning. The two authors worked jointly on implementing and analyzing the machine learning and self-training methods. Shen also contributed to the supervised baselines and the ladder network as a joint work with her CS331B project.

References

[1] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information processing systems*, pages 368–374, 1999. 2

[2] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 1

[3] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen. Computer-

aided diagnosis with deep learning architecture: Applications to breast lesions in us images and pulmonary nodules in ct scans. *Nature*, (6):24454 EP –, 2016. 2

[4] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. *Nips*, pages 1–9, 2012. 2

[5] Contributors. *Github repository*. https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py. 2

[6] Contributors. *Github repository*. https://github.com/bjlkeng/sandbox/tree/master/notebooks/vae-semi_supervised_learning. 3

[7] CuriousAI. *Github repository*. https://github.com/CuriousAI/ladder. 4

[8] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 2

[9] R. Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[10] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 2

[11] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 902–909. IEEE, 2010. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 1, 2

[13] A. Holub, M. Welling, and P. Perona. Exploiting unlabelled data for hybrid object classification. In *Proc. Neural Information Processing Systems, Workshop Inter-Class Transfer*, volume 7, 2005. 2

[14] D. Kingma, D. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014. 2

[15] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). 4

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998. 4

[18] Y. Li, C. Guan, H. Li, and Z. Chin. A self-training semi-supervised svm algorithm and its application in an eeg-based brain computer interface speller system. *Pattern Recognition Letters*, 29(9):1285–1294, 2008. 2

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, Nov. 2015. 1

[20] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93. ACM, 2000. 1

[21] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 2

[22] R. Platania, S. Shams, S. Yang, J. Zhang, K. Lee, and S.-J. Park. Automated breast cancer diagnosis using deep learning and region of interest detection (bc-droid). In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, pages 536–543, New York, NY, USA, 2017. ACM. 2

[23] 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. *CoRR*, abs/1711.05225, 2017. 2, 3, 5

[24] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko. Semi-supervised learning with ladder network. *CoRR*, abs/1507.02672, 2015. 2, 4

[25] R. T. Rasmus, A. and H. Valpola. Denoising autoencoder with modulated lateral connections learns invariant representations of natural images. *CoRR*, 2016. 4

[26] R. T. Rasmus, A. and H. Valpola. Lateral connections in denoising autoencoders support supervised learning. *CoRR*, 2016. 4

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1

[28] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. 2005. 2

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015. 1

[31] K. Tseng, Y. Lin, W. H. Hsu, and C. Huang. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. *CVPR*, 2017. 2

[32] L. L. Z. L. M. B. R. S. Xiaosong Wang*, Yifan Peng*. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017. 1, 2, 3, 5

[33] P. E. D. D. C. B. B. D. Yao, Li and K. Lyman. Learning to diagnose from scratch by exploiting dependen- cies among labels. *arXiv preprint*, 2017. 2, 3, 5

[34] S. Zhou, H. Greenspan, and D. Shen. *Deep Learning for Medical Image Analysis*. Elsevier Science, 2017. 2

[35] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. 1, 4

[36] X. Zhu, J. Lafferty, and R. Rosenfeld. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, language technologies institute, school of computer science, 2005. 2