
Learning Method Comparison for Small Dataset

Ki Suk Jang
Stanford University
bjang1@stanford.edu

Abstract

Training deep neural network is costly due to high computation cost and lack of large dataset. To mitigate this issue, numerous learning methods, such as transfer learning and meta-learning, were found in academia. This project compares these recently discovered learning methods on classification task of NIH Chest X-rays dataset. According to experiments performed within this project, utilizing transfer learning and advanced meta-learning, such as non-parametric MAML, on small dataset may increase performance, but it is difficult to alleviate the decrease in performance caused by decrease in dataset size and to produce performance that is on par with the large dataset.

1 Introduction

Most of recent industrial problems in applying deep learning comes from data. With advent of neural network, potential benefits of machine learning have been proven both in industry and academia, but many companies still struggle to apply this new technology due to data problem. Not only the companies do not possess enough data that pertains to specific tasks they desire to solve, but also there are high financial and time costs involved in collecting and training with large dataset.

As a result, recent trends in deep learning have been learning methods that demonstrates high performance with a small dataset. Especially, transfer learning and meta-learning have shown competent performance with a small dataset of under 10,000 items in numerous tasks.

For this project, we analyze performances of different learning methods, ranging from traditional convolutional networks to meta-learning, in classifying diseases in chest X-ray images. By analyzing models' performance, we explore their potential implications in solving tasks with small dataset.

2 Related Work

2.1 Transfer Learning for COVID-19 Detection

Basu et al. (1) invents Domain Extension Transfer Learning (DETL) which leverages a model pre-trained on NIH Chest X-rays (2) to classify COVID-19 on chest X-rays. DETL pre-processes NIH Chest X-rays dataset by relabeling X-rays to two classes, *normal* and *disease*, and pre-trains models pre-trained on different architectures, including AlexNet, VGGNet, and ResNet. Then, it constructs a second dataset with four classes, *normal*, *other disease*, *pneumonia*, and *Covid-19*, from multiple datasets, including NIH Chest X-rays dataset, and compares the pre-trained models' performance. The models achieve 82.98%, 90.13%, and 85.98% accuracy with AlexNet, VGGNet, and ResNet respectively, and such performance is promising given that NIH Chest X-rays dataset has labeling accuracy of >90%.

Nevertheless, there are two shortcomings of Basu et al. (1): amalgamation of disease classes and simplicity of evaluation metric. Among NIH Chest X-rays dataset, images with *pneumonia* class account for only 1.2% of the dataset, and images with *COVID-19* class account for less than 5% of the

entire datasets. Therefore, whether DETL’s overall accuracy of 90.13% has implications for potential practicality is unknown. Also, the evaluation metric it uses, accuracy, seems to be impractical in real world, where classifying diseases as normal is detrimental.

2.2 Meta Learning for ORBIT

Unlike Basu et al. (1), Massiceti et al. (3) takes a different approach by utilizing meta learning on small dataset. It compares performance of ProtoNets (5), CNAPs (6), MAML (7), and FineTuner (8) on ORBIT (4) dataset. The paper does a good job in comparing different types of learning algorithms (multi-task learning, optimization-based learning, and non-parametric learning) and providing clear evaluation metrics which are measured on two different types of videos, clean videos and clutter videos. The difference between clean videos and clutter videos is that the former contains objects with more clear background.

However, the best performance described in the work falls short of meeting real world application standard. For clean videos, MAML achieves only 70.58% frame accuracy, and for clutter videos, FineTuner achieves only 53.73% frame accuracy. Also, there seems to lack clear baseline methodology which indicates how well the model is performing compared to previous approaches utilizing large dataset.

3 Dataset and Features

For dataset, we use NIH Chest X-rays (2). The dataset consists of 112,121 chest X-ray images in $1024 \times 1024 \times 1$ dimension, each of which has meta data, including disease types, age, gender, relevant box position. As we endeavor to compare performance of each model, we formulate our objective as a classification task and constraint usage of the dataset to only images and corresponding disease types.

Also, for the purpose of our project, we use images with a single disease type that are labeled as *No Finding*, *Cardiomegaly*, *Pneumothorax*, *Consolidation*, *Edema*, or *Pneumonia*. As we endeavor to solve a classification task, not multi-class classification, we discard data with multiple disease types. Number of X-ray images with no disease, labeled as *No Finding*, are down-scaled with same proportion, because they are the majority, $> 50\%$, of the dataset and might cause false analysis by introducing dissimilarity in distribution compared to real world data. In addition, since the chest X-ray images are resized to 64×64 dimension for faster training iterations, we select a couple disease types that are still detectable with naked eyes of medical professionals to mitigate the issue of information loss from lower resolution. For instance, as shown in Figure 1, the selected disease types are generally easier to detect even in low resolution. The resulting dataset contains total 41,029 images.

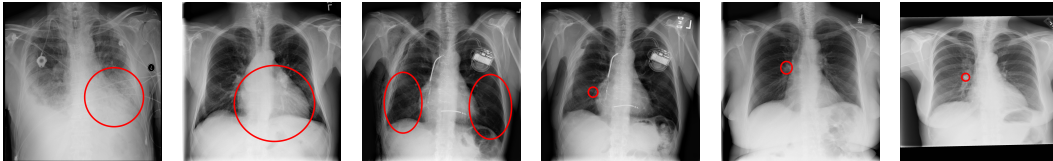


Figure 1: From left to right, X-ray images of 3 diseases that are easier to detect and 3 diseases that are harder to detect with naked human eyes (Pneumonia, Cardiomegaly, Pneumothorax, Emphysema, Fibrosis, Mass)

For the purpose of our project, the model is evaluated on small partial dataset. Models trained with the full dataset act as a baseline while models trained with the partial dataset represent difficulties of training with small dataset. The partial dataset is balanced and consists of 1200 images of selected diseases where number of samples for each class is equal.

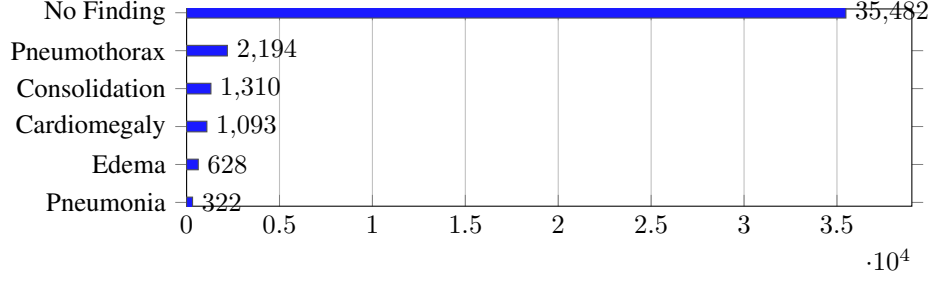


Figure 2: Class distribution of pre-processed NIH Chest X-rays dataset



Figure 3: Convolutional networks architecture: ResNet50 (left) and VGG-16 (right)

4 Methods

4.1 Convolutional Networks

For baseline model, we use convolutional networks, ResNet50 (9) and VGG16 (10) (Figure 3) as they show reasonable performances in Basu et al. (1). ResNet50 is a 50-layer residual network and VGG16 is a 16-layer network that utilizes standard 3×3 kernel and 2×2 max-pooling layers. Last layers of both networks are fully connected layers that output a 6 dimensional vector where each row represents probability of a class.

As a loss function on the full dataset, we use weighted cross entropy loss to mitigate imbalanced dataset problem. Weights are calculated based on the distribution of dataset and normalized.

$$Loss(x, y) = \sum_{n=1}^N \sum_{c=1}^C w_c \log\left(\frac{\exp(x_{n,c})}{\sum_{k=1}^C \exp(x_{n,k})}\right) y_{n,c} \quad (1)$$

$$w_c = \frac{\frac{N}{15 \times n_c}}{\sum_{k=1}^C w_k} \quad (2)$$

where N is number of samples and C is number of classes.

On the other hand, we use cross entropy loss as a loss function for the partial dataset as the distribution of the dataset is balanced.

$$Loss(x, y) = \sum_{n=1}^N \sum_{c=1}^C \log\left(\frac{\exp(x_{n,c})}{\sum_{k=1}^C \exp(x_{n,k})}\right) y_{n,c} \quad (3)$$

4.2 Transfer Learning

For transfer learning, we use ResNet50 pre-trained on ImageNet (11). The model shares same architecture as described above, except that the model is pre-trained on images with $3 \times 256 \times 256$ dimension. For purpose of our project, we customize the model such that last linear layer outputs 6 dimension vector and uses the cross entropy loss described in Eq 3.

4.3 Meta Learning

For meta learning, we use MAML (7). Our MAML model uses 4 layers of convolutional network with Batch Normalization and ReLu activation. The last layers of the model are a fully connected network

that outputs 6 dimensional vector representing probabilities of each class. For updating gradients, the model utilizes optimization-based adaptation where inner loops are updated with following minimize function:

$$\min_{\theta} \sum_{task_i} L(\theta - \alpha \nabla_{\theta} L(\theta, D_i^{tr}), D_i^{tr}) \quad (4)$$

where θ is parameter of the model, L is the cross entropy loss described in Eq 3, and D is dataset. Then, outer loops are updated with following update function:

$$\theta \leftarrow \nabla_{\theta} L(\theta_i, D_i^{test}) \quad (5)$$

In addition we implement MAML_Non_parametric, a blackbox-based, non-parametric MAML (12). The mode uses LSTM to encode support set and a fully connected layer to encode query set, and it optimizes following loss function,

$$Loss = L(\hat{y}^{ts}, \sum_{x_k, y_k \in D^{tr}} f_{\theta}(x^{ts}, x_k) y_k) \quad (6)$$

where L is the cross entropy loss described in Eq 3.

4.4 Evaluation

For evaluation, we measure four metrics: overall accuracy, precision, recall, and F1 score. Precision, recall, and F1 score are averaged over the classes. Notice that in real world scenario, missing a disease detection is more detrimental than falsely labeling normal as disease. Thus, for our study, recall is more significant than precision metrics.

5 Results

5.1 Experiments

For convolutional networks, ResNet50 and VGG16, the models are trained on Adam optimizer with 0.001 learning rate and batch size of 32. While VGG16 is trained only on the balanced partial dataset. ResNet50 is trained on the full dataset, partial dataset proportional to the full dataset, and the balanced partial dataset to compare models' performance on different datasets.

For transfer learning, images of partial balanced dataset, which are in $1 \times 64 \times 64$ dimension, are scaled to $3 \times 256 \times 256$ dimension before being fed into ResNet50 pretrained on ImageNet. Similar to the convolutional networks, the model was trained on Adam optimizer with 0.001 learning rate and batch size of 32.

For meta-learning algorithms, MAML is trained on the partial balanced dataset with inner learning rate of 0.4, outer learning rate of 0.001, and batch size of 16. Out of 6 classes in the partial dataset, 4 were used for training, 5 for validation, and 6 for testing. Also, the models were experimented on 3-way 1-shot and 3-way 3-shot with 15 queries. For MAML_Non_Parametric, it is trained on the partial balanced dataset with learning rate of 0.001, batch size of 16, 3-way 1-shot with 15 queries.

5.2 Analysis

As shown in Figure 4, models trained on imbalanced dataset, ResNet50 (Full) and ResNet50(Partial, Proportional), achieve high accuracy while precision, recall, and F1 are relatively lower. One potential reason could be large proportion of *No Finding* within the dataset; the models might be optimized for *No Finding* class and demonstrate low performance in other classes as shown in Figure 5.

Also, models with deeper network tend to achieve higher performance. Compared to VGG16, which completely fails to learn classification, ResNet50 on the same dataset achieves performance almost equal to the performance by ResNet50 on the full dataset. Also, given the fact that MAML is trained

Model	Accuracy	Precision	Recall	F1
ResNet50 (Full)	0.57	0.31	0.39	0.30
ResNet50 (Partial, Proportional)	0.63	0.23	0.18	0.20
ResNet50 (Partial, Balanced)	0.48	0.49	0.45	0.43
VGG16	0.16	0.02	0.17	0.05
ResNet50-TL	0.34	0.27	0.32	0.26
MAML, 3-way 1-shot	0.43	0.39	0.43	0.39
MAML, 3-way 3-shot	0.49	0.51	0.49	0.46
MAML_Non_Parametric	0.46	0.48	0.47	0.42

Figure 4: Model performance: accuracy, precision, recall, and F1 metrics

Model	No Finding	Cardiomeg.	Pneumotho.	Consolid.	Edema	Pneumonia
ResNet50 (Proportional)	0.68	0.00	0.21	0.00	0.00	0.00
ResNet50 (Balanced)	0.49	0.46	0.63	0.27	0.77	0.05

Figure 5: Recall metric per class for ResNet50 models trained on partial dataset

on 3-way classification, its performance is low compared to ResNet50 as it utilizes much shallow network of 4 layers.

For transfer-learning, it performs relatively poorly compared to ResNet50 trained on the partial balanced dataset. One reason why the transfer learning with pre-trained ResNet50 model shows low performance is that the pre-trained model utilizes ImageNet dataset which causes model to learn and focus on shapes and colors. However, NIH Chest X-rays dataset requires models to focus on other details, such as size and patterns, so utilizing the pre-trained model might be unhelpful.

On the other hand, increasing number of shots and taking non-parametric approach both improve performance in meta-learning. For instance, training on non-parametric MAML increases accuracy, precision, recall, and f1 performance by 7%, 23%, 9%, and 8%, respectively.

6 Future Works

Given the fact that only 90% of images in NIH Chest X-rays dataset are correctly labelled, accuracy of the experimented models seems reasonable. However, in real-world medical practices, high recall metric is a must-have, so <50% recall indicates that there are lots of rooms for improvements.

One potential way to improve performance is to increase expressivity. As described in 5.2, when training with small dataset, models with deeper network and/or more sophisticated approaches achieve higher performance. Thus, we have noticed that models with shallow networks have difficulty in fully incorporating information contained within input images. To resolve this problem, we can experiment a couple different approaches, including but not limited to increasing image resolutions, utilizing models pre-trained with similar input images and tasks, and implementing deeper networks for meta-learning.

Improving performance of models on not just partial NIH Chest X-rays dataset, but also all kinds of small dataset is critical to deployment of machine learning algorithms. As implementation cost of deep neural networks is significantly high and will continue to remain high, alternative learning methodologies, such as training models with small dataset, are key to advancement of machine learning technology. As machine learning is often referenced as electricity in terms of its industrial impact, more research and allocation of resources on easily deployable machine learning algorithms are needed.

7 Contributions

All works including design, implementation, training, evaluating, and writing the entire paper are done by Ki Suk Jang.

8 Work done for CS230 & CS330

Training/testing of Resnet50, VGG16, and transfer learning are done for CS230 while training/testing meta learning is done for CS330.

References

- [1] Sanhita Basu, Sushmita Mitra, and Nilanjan Saha (2020) *Deep Learning for Screening COVID-19 using Chest X-Ray Images* <https://arxiv.org/pdf/2004.10507.pdf>
- [2] Intramural Research Program of the NCI Clinical Center and National Library of Medicine *NIH Chest X-rays* Over 112,000 Chest X-ray images from more than 30,000 unique patients. <https://www.kaggle.com/nih-chest-xrays/data>
- [3] Daniel Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Catja Hofman, and Simone Stumpf (2021) *ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition*. <https://arxiv.org/pdf/2104.03841.pdf>
- [4] *ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition* <https://github.com/microsoft/ORBIT-Dataset>
- [5] Jake Snell, Kevin Swersky, and Richard Zemel (2017) *Prototypical networks for few-shot learning*. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)
- [6] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E. Turner (2019) *Fast and flexible multitask classification using conditional neural adaptive processes*. In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS)
- [7] Chelsea Finn, Pieter Abbeel, and Sergey Levine (2017) *Modelagnostic meta-learning for fast adaptation of deep networks*. In Proceedings of the 34th International Conference on Machine Learning (ICML)
- [8] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola (2020) *Rethinking few-shot image classification: A good embedding is all you need?* In Proceedings of the European Conference on Computer Vision (ECCV)
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015) *Deep Residual Learning for Image Recognition* <https://arxiv.org/abs/1512.03385>
- [10] Karen Simonyan and Andrew Zisserman (2015) *Very Deep Convolutional Networks for Large-Scale Image Recognition* <https://arxiv.org/abs/1409.1556>
- [11] *TORCHVISION.MODELS* PyTorch <https://pytorch.org/vision/stable/models.html>
- [12] Qiurui Chen (2020) *Non-parametric meta-learning* Towards Data Science <https://towardsdatascience.com/non-parametric-meta-learning-bd391cd31700>