CS:5430 Machine Learning Project Report

Ross Brown
Department of Computer Science and Engineering
The University of Iowa
Iowa City, Iowa
ross-brown@uiowa.edu

I. ABSTRACT

Selecting parameters for a neural network architecture is a multifaceted issue which can be expensive both in terms of raw computation power needed, and computation time expended. This work aims to develop a more precise and accurate training procedure for the classification of chest radiographs while maintaining a reasonable amount of computation required. The main pathways through which this will be achieved are hyperparameter optimization, in which the selected model architectures will be configured optimally for the presented classification task, data pre-processing techniques, and investigation of multitask learning. The optimization of hyperparameters will be completed using Bayesian optimization, a technique which reduces the number of trials required to be performed for the selection of optimal hyperparameters compared to traditional methods such as grid searching. The data pre-processing and multitask learning effects are examined once a suitable model is found.

II. INTRODUCTION

Machine learning based medical image analysis has the potential to dramatically reduce the overhead costs associated with chest radiograph interpretation, allowing for faster and more affordable treatment to reach a wider range of patients. These approaches also have the potential to increase the rate of successful classifications and greatly reduce the delay between scans being taken and the availability of results if a sufficiently robust model is developed. A major inhibitor of the widespread adoption of these methods is the lack of large, accurately labelled datasets on which to train the models. Additionally, the model must be shown to have a very strong performance when evaluated and compared to radiologists.

The recently released CheXpert dataset [1] provides 224,316 computer labelled chest radiographs for the classification of lung diseases and observations. The natural language processing approach used in the labelling of these data outperformed the Text-Image Embedding network (TieNet) [2] proposed by Wang et al. in the creation of the ChestX-ray14 dataset, a previous state-of-the-art dataset, across all labels considered. There are 14 observations labelled in the CheXpert dataset, but only five diseases will be considered for this work, Cardiomegaly, Edema, Consolidation, Atelectasis, and P. Effusion as per the specifications of the assignment. This data is highly imbalanced toward the negative class and has a large number of uncertainties (see Fig. 2) to be considered by any model hoping to perform well on it. The goal of this work is to create a model capable of classifying the diseases listed with a high precision and recall as measured by the AUC score.

III. RELATED WORK

The classification of chest radiographs using convolutional networks is a difficult task due to the relatively low number of machine-learnable images. Some of the more successful strategies for learning these challenging datasets follow.

The results published in CheXpert for benchmarking on their proprietary dataset used the DenseNet model [3], which was selected after a brief model search. The DenseNet model was trained with default parameters, and with the learning rate fixed for the duration of training. It has been shown previously that variable learning rates increase the speed of convergence [4]. With models

converging after less computation time, it will be much less expensive to implement hyperparameter tuning which is known to increase model performance (e.g. [5]) as well.

DualNet [6] has shown to increase performance on radiograph interpretation when multiple angles are present compared to traditional models by considering both viewing angles simultaneously. DualNet also uses DenseNet architecture and accepts image pairs before training the fully connected layers of the model. The models proposed in DualNet involved pretraining by using weights from models trained on the ImageNet dataset [7]. The DualNet training could likely also be improved through hyperparameter optimization as it again used default parameters for DenseNet.

Hierarchical learning [8] utilized some of the inherent relationships and dependencies between some of the diseases present in the labelled dataset to score very highly when evaluating on the testing set. This method was of particular interest because it improved performance not only by adjusting the model using machine learning principles, but also by implementing domain-specific knowledge of the underlying data distributions.

DeepAUC [9] utilized a novel margin-based surrogate loss in training their models to achieve the highest score currently reported on the CheXpert leaderboard [10]. Additionally, this work was published as a codebase on GitHub [11]. The implementation of Densenet-121 published here is what I used as a model on which to base my Bayesian optimization and create a baseline performance to benchmark against since many of the top models on the leaderboard were utilizing the same architecture.

IV. Methods

This model was evaluated on a different dataset than the CheXpert dataset. The data provided to validate our performance from the extra dataset contained an overwhelming majority of frontal images. As a result, only frontal images were considered while training and evaluating the models on the CheXpert dataset.

Additionally, with the heavy computational costs of Bayesian optimization already consuming large amounts of time and resources during a project which, by nature of being a class project, already has very limited time, the models considered were not pretrained. All benchmarks and final evaluations are reported with models that have not been pretrained on ImageNet. With hyperparameter tuning being the main focus of this project, the computation time required to create a new pretrained ImageNet model of the appropriate size every time a new architecture was evaluated would have been well outside the time allowed for this course.

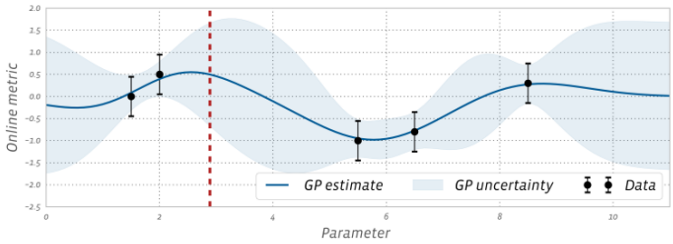

Fig. 1. Example of a Gaussian Process' estimation of model metric and uncertainty [12].

A. Multitask Learning

One of the investigated pathways for improvement was the use of multitask models rather than a collection of single-task models, one for each prediction task. This likely allowed the model to learn dependencies between each of the prediction tasks to improve performance by exploiting some of the same relationships that the hierarchical learning models discussed above did. The multitask model was trained to predict all 5 scored observations in the final output layer, and the single task models were trained to predict only one observation at a time.

B. Hyperparameter Tuning

Hyperparameter tuning was implemented to perform an architecture search for the Densenet model selected using Bayesian optimization. This created, trained, and evaluated a wide range of models that provided insights into the optimal parameter space for the model for chest radiograph classification. Notably, Bayesian optimization reduces search time significantly compared to methods such as grid searches by implementing a probabilistic surrogate model to generate the trials to be run based on the uncertainty and likelihood of improved performance.

I utilized the Ax environment [12] for implementing Bayesian optimization in this project. A

Gaussian process is used in Ax to generate the surrogate model based on the results of previous trials as is shown in Fig.1. Optimizing in this way reduces the number of trials required for optimal parameter space exploration as the non-linear relationships between each hyperparameter and the performance of the model can be estimated to ensure poorly performing parameter space does not consume much of the computation time. This also allows for parameter spaces in which the model is performing unusually well to be explored with finer hyperparameter adjustments.

The model was optimized by training on the CheXpert training dataset of roughly 224,000 images and then evaluating the model on the CheXpert evaluation dataset of roughly 200 images. The performance of the model on the evaluation dataset was the score set to be maximized during parameter tuning. With the testing data being selected from a different dataset, it is possible this optimization selected for architectures that were more appropriate for the CheXpert predictions than the final evaluation predictions, but with both datasets originating from the same domain it is unlikely this had any significant impact on the final evaluation performance.

C. Image Augmentation

Image augmentation was also investigated for potential improvement, once again utilizing Bayesian optimization to find augmentations which provided the most improvement to the model. The images were augmented by rotating (degrees +/-), translating, and scaling (+/-). A comparison was made between models trained with images receiving no augmentation, models with augmentation optimized on the CheXpert validation dataset, and models augmented with the default augmentations of the LibAUC codebase.

D. Uncertainty Labels

The CheXpert dataset has a large number of uncertain labels present as it was derived from free-text radiology reports and for many mentions of the observations it was difficult to determine if the mention was for the positive or negative class. The distribution of these uncertainties is shown in Fig. 2. Many methods for handling these uncertain labels were proposed in the initial publication of the CheXpert dataset. The results reported showed

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Fig. 2. Data distribution in the CheXpert dataset [1].

that considering the uncertain labels to be positive indicators (U-Ones) or negative indicators (U-Zeros) provided the highest score on the Atelectasis and Edema tasks and had an AUC score differing from the highest score by only 0.015, 0.005, and 0.002 respectively for Cardiomegaly, Consolidation, and Pleural Effusion. With these two methods preforming this well across the board, I selected them as the two uncertainty handling methods I would investigate in this work.

V. Results and Conclusion

A. Hyperparameter Optimization

The models selected through Bayesian optimization showed improvement across all 5 tasks considered when compared to their default counterparts (Fig. 4). Interestingly Edema showed only a small increase in performance, but all other tasks showed significant increases.

One of the more surprising results of the optimization was the Beta2 parameter used in the Adam optimizer. Traditionally Adam optimizers have a very high Beta2 parameter at 0.999, but the optimal parameter space for Beta2 was much closer to 0.8 for this classification task. Additionally, as shown in Fig. 3, the search was stopped at 0.8 by limitations I imposed upon the model selection. If this project had a longer duration, it would have been interesting to run another optimization with an expanded search space for the Beta parameters.

Another interesting result is that there were multiple different configurations of the number of layers in each pooling block in the DenseNet that provided similar AUC scores. There was no single area of

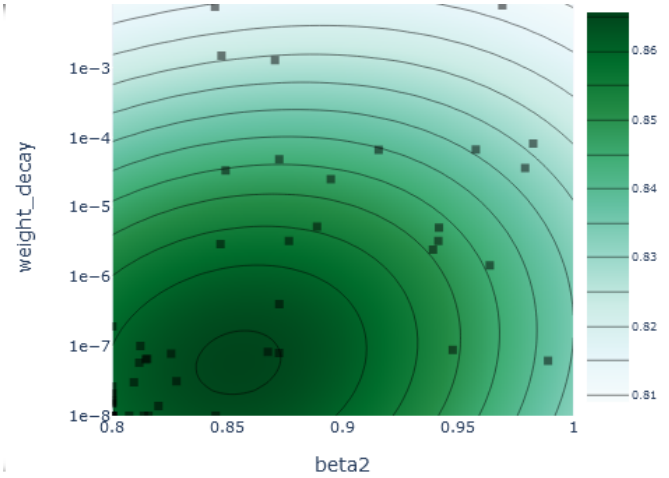

Fig. 3. Contour plot showing the parameters optimized. Darker regions are higher AUC scores and each trial is represented by a dot.

Task	Multitask optimized	single task optimized	multitask default	single task default
Cardiomegaly	**0.9000**	0.8808	0.6931	0.6497
Edema	**0.9699**	0.9184	0.9644	0.7352
Consolidation	0.8753	**0.8986**	0.6384	0.5172
Atelectasis	**0.7257**	0.6372	0.5789	0.5603
Pleural Effusion	**0.9387**	0.9356	0.7404	0.5211
Average:	**0.8819**	0.8541	0.7230	0.5967

Fig. 4. Results of optimization and single task compared to multitask models.

parameter space providing higher scores than the others.

B. Multitask Learning

For both optimized and default architectures results were collected as multitask models and as single task models. The single task optimized models were trained using an architecture optimized while evaluating the multitask model, so it is possible the single task results might be slightly higher if the models were instead optimized for single task performance. Notably though, the default architecture multitask model outperformed the default architecture single task models across every scored observation (Fig. 4). As a result, it seemed prudent to only optimize the multitask model architecture with the short duration of this project in mind. The multitask optimized model also outperformed its single task counterpart nearly invariably. The only task in which it did not was Consolidation, with an AUC difference of 0.0233.

Task	Default augmentation	Optimized augmentation	No augmentation
Cardiomegaly	0.9000	0.8978	0.9190
Edema	**0.9699**	0.8718	0.8440
Consolidation	0.8753	0.8466	0.8671
Atelectasis	0.7257	0.7271	0.7244
Pleural Effusion	0.9387	0.9353	0.9402
Average:	0.8819	0.8557	0.8589

Fig. 5. Image augmentation methods comparison.

With the results of the multitask evaluation in mind, a model was trained which predicted all 14 tasks. The averaged AUC was 0.61 indicating the additional tasks were having a hugely adverse effect on the model performance. To save on valuable computation time, this was not investigated further. It is possible there would be benefit to any or all of these tasks being included for training for the score of main 5 tasks. It would likely need a much larger model to gain the full benefit of this though as there would almost certainly be more features of high importance for some of the other tasks.

C. Image Augmentation

Image augmentation was investigated as another means of improving performance, though with very limited success. An optimization was run over the rotation, translation, and scaling of the images and compared to the default augmentations suggested in LibAUC and un-augmented images. The results are shown in Fig. 5, interestingly there was very little performance difference between any of the tasks for any of the methods except Edema. The default augmentations scored well above the optimized and un-augmented models. The likely cause of this is that the optimizations were performed on the AUC score calculated on the CheXpert validation set, not the extra validation set used for final scoring. The model likely over-selected for augmentations allowing the features present in the CheXpert set to be properly distinguished by the model. Due to the extra data being from a different dataset, it is possible that those augmentations that led to improvements on CheXpert validation made little to no difference on the performance on the extra validation data because of different feature distributions.

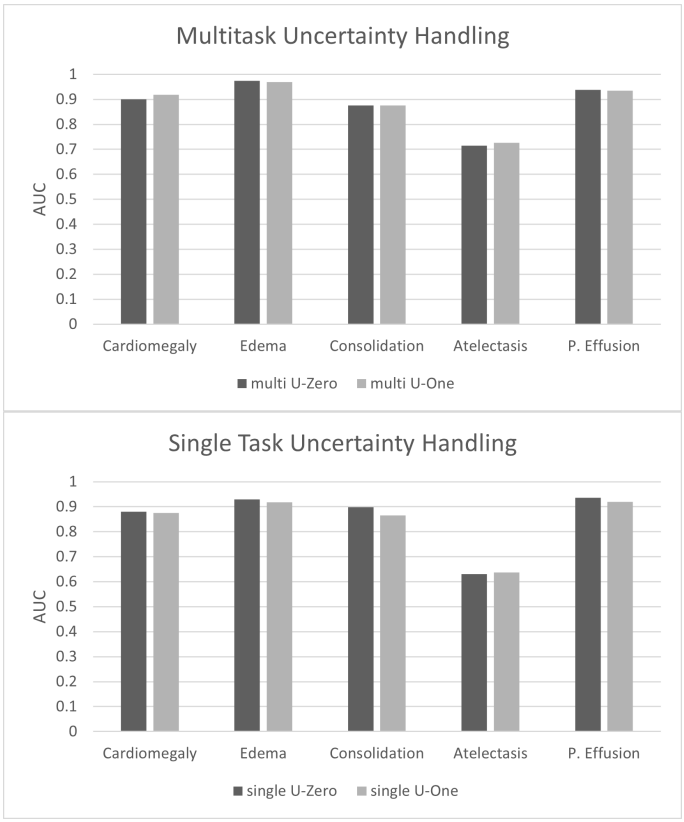

Fig. 6. Results of testing U-Ones and U-Zeros on optimized single task and multitask models

D. Uncertainty Labels

The two strategies discussed for handling uncertainties in the data, U-Ones and U-zeros were evaluated on the extra validation dataset using optimized multitask and single task models. The results are shown in Fig. 6 and surprisingly made little to no difference. No absolute difference in AUC score was greater than 0.05, and the average absolute difference was 0.01. This was a very surprising result. I had expected the uncertainty handling to be a major factor in model performance. It seems that a sufficiently robust model is able to learn to predict the classes appropriately regardless of how the uncertain data is labelled in the training. One possible explanation for this is that the uncertain labels are likely not present in cases where there are features clearly indicating a certain lung classification. As a result, the uncertain data is probably less important than the data that is clearly indicative of a positive or negative classification.

VI. FURTHER WORK

Due to the limited time available for this work, there are many different areas that could be explored for future improvements.

Implementing pretraining on optimized models, or even adding pretraining to the optimization workflow, would almost certainly show an increase in performance when compared to the randomly initialized models currently used.

New loss functions or optimizers beyond the binary cross entropy loss and Adam optimizer used might improve the models as well.

Earlier stopping of models during training could also boost performance. All optimizations and evaluations were done with models only stopping to evaluate every 1000 batches. With such long training intervals, there are probably model configurations not being saved that could outperform the ones eventually reached. This large window was simply chosen because of the timeframe of the project since additional computational costs are entailed in evaluation. There might also be some benefit to early stopping and saving based on single task AUC scores for multitask models since the multitask models typically outperform the single task models. This would be a way to increase each tasks multitask performance beyond what was already achieved.

The multitask effect could easily be explored further. There may be some learnable relationships between some of the tasks not currently included and the scored tasks that would yield a better model.

Lastly, uncertainty handling methods beyond U-Zeros and U-Ones could also potentially improve the results of this model.

REFERENCES

[1] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).

[2] Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9049-9058).

[3] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[4] Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. Neural networks, 1(4), 295-307.

[5] Cox, D., Pinto, N. (2011, March). Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG) (pp. 8-15). IEEE.

[6] Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., Xu-Wilson, M. (2018). Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. arXiv preprint arXiv:1804.07839.

[7] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

[8] Yuan, Z., Yan, Y., Sonka, M., Yang, T. (2020). Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. arXiv preprint arXiv:2012.03173.

[9] Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., Nguyen, H. Q. (2021). Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing, 437, 186-194.

[10] Stanford ML Group. Chexpert: A large chest x-ray dataset and competition. https://stanfordmlgroup.github.io/competitions/chexpert/

[11] https://github.com/Optimization-AI/LibAUC

[12] Bakshy, E., Dworkin, L., Karrer, B., Kashin, K., Letham, B., Murthy, A., Singh, S. (2018). AE: A domain-agnostic platform for adaptive experimentation. In Conference on Neural Information Processing Systems (pp. 1-8).