Reproducibility Project for CS598 DL4H in Spring 2023

Quan Nguyen
quanhn2@illinois.edu

Group ID: 11
Paper ID: 27
Presentation link: https://youtu.be/ORotQA0iHN8
Code link: https://github.com/quannhoang/cs598project

1 Introduction

The original research discusses the obstacles faced in implementing effective medical Artificial Intelligence (AI) tools in clinical practice. Despite significant investment in medical AI development aimed at reducing healthcare costs and enhancing efficiency, the transition to clinical deployment has been extremely difficult. Challenges range from regulatory considerations to the opacity of machine learning models.

To tackle these issues, the original authors suggest employing conformal prediction techniques in medical imaging applications. Conformal predictions can better align with clinical decision-making intuition and provide meaningful statistical assurances. The authors assert that conformal prediction sets inherently convey a degree of uncertainty through the number of items within the set, similar to how doctors commonly express uncertainty through comparative sets to diagnose conditions (known as a "differential diagnosis"). Using conformal prediction methods can also encourage quantifying uncertainty to promote greater confidence in medical black-box models and uncover bias in protected patient demographics (Lu et al., 2022).

2 Scope of reproducibility

The authors test four different ways of predicting skin conditions in two groups of people. The result is that the ways all worked similarly well in terms of accuracy and size of the prediction set, but the group methods (GAPS and GRAPS) were slightly better at ruling out skin conditions for some types of skin (types 1 and 2). The paper also looks at how well the methods work for different skin types and find that the group methods had less difference in coverage between skin types, but had larger prediction sets. Finally, the paper also conclude that larger prediction sets are associated with more uncertainty about the diagnosis. The group methods also show better separation between different skin types, especially for skin types 1 and 6.

For the scope of the reproduction paper, it will focus on how the four different ways of predicting skin conditions will vary using the similar prediction model mentioned in the paper.

3 Methodology

The original research uses a pre-trained ResNet18 convolutional neural network as the base model to output the class prediciton probabilities. Those probabilities will then be fed to the conformal prediction algorithm to find the final set of class that satisfy the alpha constraint.

3.1 Model descriptions

The ResNet18 model is a convolutional neural network (CNN) with 18 layers, including a convolutional layer, four residual stages, and a final fully connected layer.

The input to the model is a 3x224x224 image. The first layer is a 7x7 convolutional layer with a stride of 2, which reduces the spatial dimensionality of the image. This is followed by a batch normalization layer, a ReLU activation function, and a 3x3 max pooling layer with a stride of 2.

The residual stages consist of several blocks of convolutional layers, batch normalization layers, and ReLU activation functions. Each block has two 3x3 convolutional layers with a batch normalization layer and ReLU activation function applied after each layer. The output of each block is then added to the input of the block, which helps to prevent vanishing gradients during training. The first residual stage has two blocks with 64 output channels, the second has two blocks with 128 output channels, the third has two blocks with 256 output channels, and the fourth has two blocks with 512 output channels.

Finally, the output of the last residual stage is passed through a 7x7 average pooling layer and then flattened. A dropout layer is applied with a dropout probability of 0.5, and then the output is passed through a fully connected layer with 114 output nodes, which correspond to the 114 classes in the Fitzpatrick dataset. The output of the final layer is then passed through a softmax activation function to produce the predicted class probabilities.

3.2 Data descriptions

Fitzpatrick17k dataset is a comprehensive collection of 16,577 photographs gathered from two dermatology atlases (Groh et al., 2021). This dataset features a hierarchical classification system, consisting of 114 unique skin conditions, which are then organized into three dermatological categories: non-neoplastic lesions (12,080), benign lesions (2,234), and malignant lesions (2,253). To further enhance the breadth of information available, each image is accompanied by a Fitzpatrick skin type label, an ordinal six-point scale used to approximate the amount of melanin pigment present in the skin.

3.3 Hyperparameters

The hyperparameters mentioned below are referenced from the original paper. Some of them are adjusted as the experiment goes due to some computational constraints or ablation goal.

- Train learning rate: 0.001
- Train epochs: 100
- Batch size: 32
- Monte Carlo dropout rate: 0.3
- Monte Carlo number of samples: 10
- Early stopping: yes
- Early stopping tolerance (number of epoch): 10
- Image augmentation: yes
- Image class re-sampling: yes

3.4 Implementation

The reproduction paper does not reuse the original paper's code, but heavily references it.

Original paper code:
https://github.com/clu5/AAAI-22
Reproduction paper code:
https://github.com/quannhoang/cs598project

3.5 Computational requirements

Originally, the computational requirement was estimated to be suitable using regular Google Colab only. As the experiment started, the actual requirement for the ResNet18 model training with the Fitzpatrick17k dataset is much more challenging. The compute units provided by reguar Google Colab is not enough to train the whole dataset, and often timed out mid-training.

The reproduction experiments ended up using Google Colab, but having its kernel connected to a Google Compute Engine virtual machine with GPU capability. The compute units and notebook session are not limited using this approach. The virtual machine family is n1-highmem-2, with 2 CPUs, 16 Gb of Memmory and 16 Gb of GPU.

4 Results

The reproduction experiment is still in the phase of training the ResNet18 CNN model. The training process poses many challenges that makes the model training difficult to converge.

- Medium to long training time: each epoch takes approximately 7-12 minutes to complete. One whole training session takes 3-14 hours to complete. Whenever some hyperparameters are changed, a new training session need to be performed again.
- Classes imbalance: there are 114 total classes in the dataset. The average classes count is 145, while the most common class is psoriasis which has 653 samples, compared to the least, pilomatricoma, which has only 53 samples.
- The imbalance in the class probably results in the models' validation accuracy difficulty to converge. While the training accuracy rate consistently reaches 80-90%, the validation accuracy hardly get over 50%.
- Training the model without the pre-trained weight from Pytorch library appears to be cumbersome. After more than 30 epochs the training early stops with validation accuracy at barely over 20%
- Limited computational capability: the original paper uses Python Pillow (PIL) image handling library to read and load images into memory. The library uses too much memory and the dataset even with very small batch size

does not fit into the 16 Gb of the GPU. The reproduction paper turns to Pytorch Read Image for more effecient memory management.

- Mote Carlo drop out rate of 50% with 30 samples imposes significant more training to the process (approximately 40% more). The reproduction paper uses only 30% with 10 samples to reduce the training time.

Each of the below training session trains on 70% and validates on 15% of the data. The rest is for testing which the experiments has not reached yet. Most of the sessions also use pre-trained weights from Pytorch ResNet18, except for attempt 2

Training attempts summary

1. Attempt 1: batch size 64, augmentation on all images.

2. Attempt 2: train from scratch without pre-trained weights, augmentation on all images.

3. Attempt 3 : batch size 16, Mote Carlo dropout rate of 50% with 30 samples, augmentation on train images only.

4. Attempt 4 : batch size 16, Mote Carlo dropout rate of 30% with 10 samples, augmentation on train images only, classes balanced on whole datasets.

5. Attempt 5 : batch size 16, Mote Carlo dropout rate of 30% with 10 samples, augmentation on train images only, classes balanced on train set only.

6. Attempt 6 : batch size 32, Mote Carlo dropout rate of 30% with 10 samples, augmentation on all images, classes balanced on whole dataset.

Attempt	Train Accuracy	Val Accuracy	Time (hours)
1	91%	49%	4
2	21%	19%	7
3	58%	29%	6
4	90%	64%	13.5
5	83%	26%	3
6	91%	74%	6

Table 1: Training attempts summary

4.1 Result 1

Using marginal coverage and cardinality as performance metrics, the Naive conformal prediction method perform almost exactly the same with Adaptive Prediction Set (APS).

For marginal coverage, both ranges from 79% to 71% for ascending alpha values.

For cardinality, the range is from 2.72 to 1.13 for the same alphas order.

4.2 Result 2

When ruling in critical skin conditions, using conformal prediction results in better accuracy for all alphas less than 0.4, compared to the averaged critical skin condition prediction accuracy by the reproduced model. This is applied to both Naive and APS approaches (Figure 1).

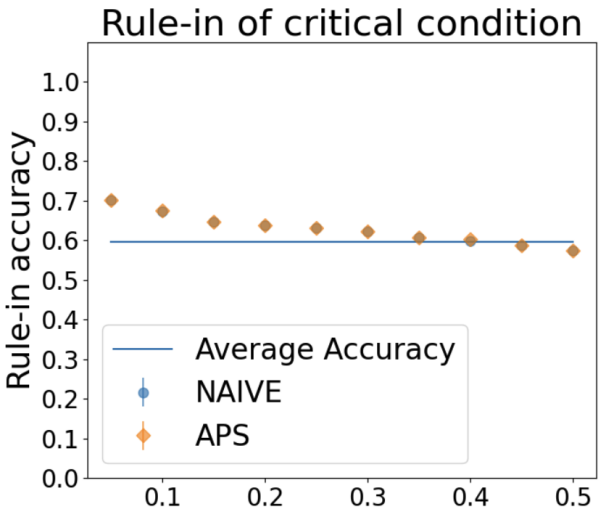

Figure 1: Rule-in of critical condition

For ruling out cases, conformal prediction performs significantly better than traditional method for all alphas. Conformal accuracy are consistently over 90% compared to the average accuracy of 71.16% (Figure 2).

4.3 Additional results not present in the original paper

Training the prediction model from scratch approach unexpectedly took longer time and it's training and validation accuracy could not converge. The model was suspected to require much more fine tuning and experiment to converge when not using pre-trained weights.

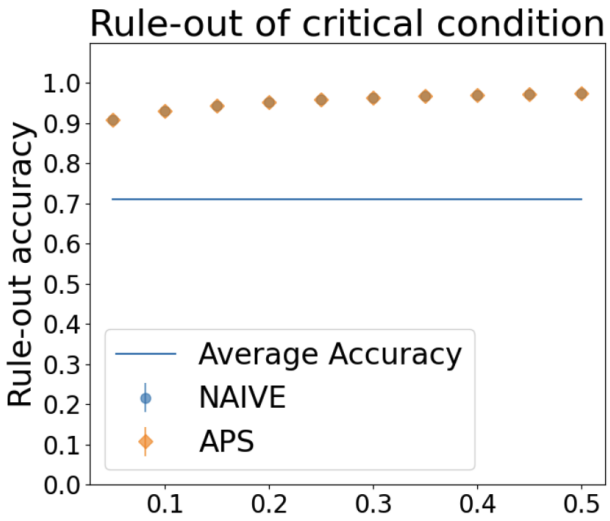

Figure 2: Rule-out of critical condition

5 Discussion

The reproduced research confirms that conformal prediction methods appear to perform better in both ruling in and ruling out critical skin conditions for the Skin Lesion Dataset. Conformal prediction methods are generally promising as approaches to increase clinical usability and trustworthiness in medical AI (Lu et al., 2022).

In this reproduced research, there have been many obstacles in replicating the model training process, as well as experimenting conformal methods of Regularized Adaptive Prediction Set(RAPS), and the group variants (Group APS and GRAPS). There have been many attempts made to study and analyze the original research in using Naive and APS conformal methods. However, the reproduced research is considered incomplete.

5.1 What was easy

Conformal prediction using Naive and APS approach. The tutorial for learning conformal prediction linked by the original paper was extremely helpful in learning and understanding the basic for the conformal methods. Conformal prediction basic is easy-to-use, easy-to-understand, and it applicable to any of the dataset's distribution type. It also works independently of the prediction model used.

5.2 What was difficult

Firstly, the process, as well as training, validating, testing metrics of the prediction model was not discussed transparently in the original paper, as well as the source code. This results in a lack of direction on how good the reproduced model need to be in order to have the same results. There was no clear goal in training the model, and it took a huge chunk of time in the training process to fine tune and experiment the reproduced model. Even after achieving the best model accuracy of all the trials, conformal prediction results were way off compared to the original. Conformal prediction outputs prediction sets with an accompanied confident level, based on the set prediction probabilities for each class (Anastasios and Stephen). Thus, its prediction sets outcome are heavily depended on the prediction model, in this case, the CNN model. The reproduced model has difference performance as well as difference predicted probabilities compared to the unknown performance of the original model. In short, it was difficult in replicating the original paper results because of the lack of information on the original model training process.

Secondly, the original paper experimented and compared much more method such as Group APS, Regularized APS, Group Regularrized APS. The reproduced paper did not replicated these method due to the lack of the mathematical and statistical knowledge required for implementation, even though the source code is available. The original paper did not discuss in details of these methods, but cited the paper that researched those methods in details. Although many attempts were made to read and understand the cited papers, it was not feasible considering the fore-mentioned lack of required knowledge as well as time constraints.

5.3 Recommendations for reproducibility

In order have a success in reproducing the original research, firstly one must spend more time figuring out clearly a the goal, or a set of metrics that was used to train the original prediction model. That would guarantee similar conformal prediction metrics later used to bench mark different methods. Secondly, the important of the mathematics and statistics behind the RAPS method is also important. Without RAPS figured out, majority of the important analysis would not be able to happen.

6 Communication with original authors

It would be so much more helpful for new learner, who has introduction-knowledge level about deep learning and statistics, to have a more transparent process of the training, validating and testing the

prediction model. Moreover, it would also be useful have the experimented methods explained in more details

References

Angelopoulos Anastasios and Bates Stephen. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Conformal Prediction: A Gentle Introduction*.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. 2022. Fair conformal predictors for applications in medical imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12008–12016.