AUC Maximization by Deep Learning for Imbalanced Medical Classification

Andrew Leach, Anil Parthasarathi and Brigham Pettit

Department of Computer Science and Engineering
Texas A&M Unversity, College Station, TX 77843

Abstract

This report presents our approach towards leveraging deep learning to **maximize Area Under the ROC Curve (AUC)** for medical imaging tasks highly constrained by data imbalances. We relied on PyTorch and LibAUC libraries in tandem with a ResNet18 network structure (and some others) to experiment on both BreastMNIST & PneumoniaMNIST classification tasks, exploring a breadth of machine-learning strategies and techniques to generate models with high performance as measured by AUC score.

1 Introduction

The widespread application of deep neural networks has revolutionized many industries. Their capability to efficiently learn complex relationships in a dataset and apply the information to make inferences about new data has transformed the way we work with data in many aspects. One important problem in the medical industry is effectively diagnosing medical phenomena.

Diagnosis can be a long process, dependent on having the right experts available at the right time to review the case — and even then, experts can be wrong. If deep neural networks can be used to accurately diagnose medical conditions, many lives could be saved.

However, much of the medical data is imbalanced — meaning there are vastly more examples of healthy patients than unhealthy ones, or vice versa. This presents a challenge for neural networks learning the training data. By many means of performance evaluation, a model that predicts according to the majority case will be highly accurate (incorrect only for a minority of examples), but highly unuseful (because it will wrongly predict those important minority examples).

AUC maximization is an evaluation method that requires the model to be highly accurate in both the majority and minority cases, penalizing mistakes in both. This paper describes our approach to using various machine learning methods (data augmentation, regularization, optimization, and network structuring) to perform AUC maximization on diagnosis tasks for breast cancer and pneumonia datasets.

2 Methods

2.1 Augmentation

Data augmentation is a useful technique that improves model accuracy by slightly modifying the original data before it is used in training. This is especially useful for creating more clear variation in data, which is key for accurately understanding and defining the underlying patterns. Given the set of image data provided by the

MedMNIST tasks, the augmentations would effectively change the appearance and features of these images. Based on the results we discuss below, we found data augmentation to be a highly effective technique for improving the AUC for both breast cancer and pneumonia data. In fact, augmentation was responsible for some of our best results.

2.2 Overfitting

To tackle the challenge of overfitting, we focused primarily on three main regularization methods: L2 or "Ridge" regularization, L1 or "Lasso" regularization, and dropout regularization. With these methods, we sought to isolate the patterns and features capable of most concisely describing the data across not only the training data, but across multiple testing sets.

L1 and L2 regularization address overfitting by penalizing complexity in a model, encouraging it to fit simpler underlying relationships rather than complex random noise. In contrast, dropout regularization is done in a very different way from L1 and L2 regularization, in the sense that instead of controlling the parameters of the neural network, it alters the structure of the network itself. In a "dropout layer" of a network, the network behaves as it normally would, except the nodes in this layer have a probability to "drop out" — i.e., output zero — during each training session. The dropout method for regularization can train the network to fit the data while reducing overreliance on individual nodes in the network (overfitting), instead learning a more general interpretation of the input data.

2.3 Optimizer

We conducted several experiments to determine which optimization method was best for classifying the medical data per task. In particular, we tried tuning the step size, altering the momentum, and using the Adam optimizer as opposed to stochastic gradient descent (SGD).

Tuning the step size or "learning rate" for SGD affects how quickly the neural network adapts to new training data, and consequently influences the quality and speed of convergence to an optimal solution. The Momentum optimizer modifies the standard SGD algorithm so that the neural network builds "momentum" — a preference to adapt in a consistent direction between iterations. This method can help to reduce the impact of local optima on a model. The Adam optimizer also uses this idea of SGD with momentum, but it includes some other theoretical optimizations to improve performance.

3 Experiments

3.1 Augmentation

The three most effective augmentations we performed were random vertical flips, random rotations with a range of 20 degrees, and random resize cropping. In particular, random vertical flip produced the greatest AUC we achieved for breast cancer scores, for any of our experiments, at 0.913.

Random resized crop, on the other hand, led to one of the highest pneumonia AUC scores we observed, at 0.981. However, it should be said that in all cases besides random resized crop, data augmentation reduced the AUC score for pneumonia as compared to the default or control ResNet18 results.

Random rotation with a range of 20 degrees also produced a great AUC score (0.901) for breast cancer classification. While rotations with ranges of 7 and 15 degrees yielded good results, we concluded 20 was the most effective. It is worth mentioning that random rotation required more training iterations to achieve peak performance at around 185 epochs, whereas other augmentations such as vertical flip converged much faster. Additionally, changes in rotation generally seemed to have a more profound impact on AUC for breast cancer classification than pneumonia.

In general, the team saw great performance improvements across several data augmentations, including random vertical flip, random rotations with degrees of 7, 15, and 20, random resized crop, and random erasing. Only two augmentations – random horizontal flip and Gaussian blur – did not increase performance on either task, and as such we deemed them ineffective for these data sets. Interestingly, at this point a trend emerged in both this category and others: many methods that increased AUC for breast cancer tasks tended to decrease the AUC for pneumonia, with one outlier in the augmentation category of a random resized crop experiment. This suggests that the tasks may require unique augmentations suited towards their features, and there is perhaps not a one-size-fits-all approach to classification.

3.2 Overfitting

We started with a simple process with L1 and L2 regularization – iteratively moving across a fixed set of hyperparameters of weight decay, and recording the best AUC scores across a multitude of epochs. The idea was that with varying values of weight decay or λ, we could fine tune the degree of regularization and observe the changes associated with increasing the penalty of model complexity. The values used were 0.001, 0.01, 0.1, and 1.

L2 regularization proved to be a robust method for mitigating overfitting in these scenarios. In essence, Ridge regularization adds a penalty to large coefficients or weights in the loss function, which encourages a simpler model that will perform better in the long term. At its pinnacle, we observed $\lambda = 0.001$ to give great results, with an AUC of 0.89 for breast cancer, and 0.974 for pneumonia.

On the other hand, L1 regularization is more specialized; it is novel in its ability to drive features completely to zero, eliminating them. In contrast, L2 might only reduce features to very small values. This makes for a handy use case in scenarios where you desire to perform feature selection on the predictors in a model. Across epochs, we saw great pneumonia results, with a maximal AUC of 0.96, using $\lambda = 0.0001$. However, it is crucial that we mention the great inconsistencies across epochs, which made us skeptical of L1 as a regularizer. In fact, the regularization grew so aggressively that eventually enough features were determined insignificant to leave us with none. Consequently, we were left with miserably low AUC results, plateauing at around 0.5 in later epochs. It is pretty clear that L1 has its niche applications, but for us it seemed to demand careful judgment of what we valued as predictors, often dropping off rapidly in performance.

To determine if dropout regularization was effective, we used a dropout layer in our next-to-last network layer and tuned the dropout probability to see if any performance improvements could be found. Unfortunately, we found no significant improvement from using dropout regularization.

3.3 Optimizer

To start, we tested whether altering the learning rate used in SGD would have a direct influence on testing AUC. We experimented with learning rates between 0.1 and 0.0001, and it was determined that the default step size of 0.001 was most effective for optimizing AUC, and tuning it further led to decreased performance.

Next, we looked at the effect of modifying our model's "momentum" parameter. The Momentum method is a highly useful optimization technique, but the end result depends on the data in question. This was found to be the case for the two datasets: for the breast data, 0.1 was the best momentum, while for the pneumonia data, a momentum of 0.9 was best. There was a consistent trend of improved performance with increasing momentum for the pneumonia data, while the opposite was true for the breast data, where decreasing momentum consistently improved the results.

We also tested the Adam optimizer for improved performance on the testing data. With no modifications, Adam resulted in a decline in performance, but when paired with L2 weight decay of 0.0001 and step size of 0.0001, it was found to be the best optimizer for the breast data.

3.4 Hybrid

The hybrid experiments we ran combined many of the techniques we used prior. At first we were most interested in simply putting together all of the best results from the aforementioned experiments. Our initial hypothesis was that it would take the best aspects of all of our trials and stack them to create even greater AUC scores.

In some cases, such as combining ResNet50 and a random rotation with 20 degrees, we were able to produce some great results for pneumonia, or even the DenseNet121 & random rotation with 20 degrees for breast results. In particular, the combination of ResNet50 and a random rotation with 20 degrees ended up being our best AUC for pneumonia classification. However, we observed that oftentimes, it wasn't as simple as mashing together the best options. For example, L2 regularization with $\lambda = 0.001$ and random erasing were two of our most promising results on their own as shown below, yet when combined, we had comparatively quite poor results of 0.832 and 0.965 for breast and pneumonia, respectively.

Some great examples of this failing were the usages of Adam as an optimizer with random rotation and varied step size — we saw AUC scores as low as 0.586 and 0.504 in those hybrid experiments, which surprised us. Another surprising failure was combining random rotation 20 degrees with random vertical flip. These were our two highest performing augmentations but when paired together, despite still producing decent results, they were not able to supersede either of their individual best scores.

All in all, we discovered that there is a more complex relationship between the methods used and the resulting models than we initially thought. While certain instances of hybrid testing can produce great results, there is no guarantee that combining two high performing methods will result in an improvement, or even manage to match the effectiveness, of each method individually.

4 Results

Results in bold are notably high AUC scores from our experiments.

Table 1
Neural Network Experiments

Experiment	Best Breast Cancer AUC	Best Pneumonia AUC
ResNet18 (Default)	0.847 (Epoch 189)	0.976 (Epoch 49)
ResNet50	0.819 (Epoch 368)	0.959 (Epoch 97)
DenseNet121	0.732 (Epoch 167)	0.964 (Epoch 14)

Table 2
Regularization Experiments

Experiment	Best Breast Cancer AUC	Best Pneumonia AUC
L2 $\lambda = 0.001$	0.890 (Epoch 142)	0.974 (Epoch 37)
L2 $\lambda = 0.01$	0.883 (Epoch 86)	0.963 (Epoch 6)
L2 $\lambda = 0.1$	0.862 (Epoch 36)	0.942 (Epoch 9)
L2 $\lambda = 1$	0.579 (Epoch 0)	0.500 (Epoch 0)
L1 $\lambda = 0.0001$	0.909 (Epoch 169)	0.960 (Epoch 37)
L1 $\lambda = 0.001$	0.868 (Epoch 28)	0.960 (Epoch 37)
Dropout p = .10	0.895 (Epoch 186)	0.965 (Epoch 9)
Dropout p = .20	0.832 (Epoch 25)	0.970 (Epoch 20)
Dropout p = .20	0.869 (Epoch 89)	0.978 (Epoch 46)
Dropout p = .40	0.870 (Epoch 99)	0.955 (Epoch 6)
Dropout p = .50	0.853 (Epoch 81)	0.966 (Epoch 9)

Table 3
Augmentation Experiments

Experiment	Best Breast Cancer AUC	Best Pneumonia AUC
Random rotation 7 degrees	0.858 (Epoch 184)	0.974 (Epoch 39)
Random rotation 15 degrees	0.884 (Epoch 178)	0.972 (Epoch 14)
Random rotation 20 degrees	0.901 (Epoch 188)	0.972 (Epoch 45)
Random flip horizontally	0.821 (Epoch 97)	0.967 (Epoch 43)
Random flip vertically	**0.913 (Epoch 69)**	0.968 (Epoch 63)
Gaussian blur	0.839 (Epoch 97)	0.968 (Epoch 32)
Random resized crop	0.849 (Epoch 14)	0.981 (Epoch 38)
Random erasing	0.861 (Epoch 17)	0.965 (Epoch 16)
Random resized crop & random rotation 20 degrees	0.768 (Epoch 98)	0.978 (Epoch 42)
Random rotation 20 degrees & random erasing	0.865 (Epoch 130)	0.967 (Epoch 41)
Random flip vertically & random rotation 20 degrees	0.857 (Epoch 199)	0.972 (Epoch 47)

Table 4
Optimizer Experiments

Experiment	Best Breast Cancer AUC	Best Pneumonia AUC
Step size 0.01	0.809 (Epoch 49)	0.951 (Epoch 9)
Momentum 0.9	0.821 (Epoch 74)	0.976 (Epoch 49)
Momentum 0.75	0.871 (Epoch 176)	0.972 (Epoch 35)
Momentum 0.5	0.872 (Epoch 130)	0.964 (Epoch 41)
Momentum 0.1	0.878 (Epoch 191)	0.957 (Epoch 37)
Adam	0.725 (Epoch 28)	0.803 (Epoch 25)

Table 5
Hybrid Experiments

Experiment	Best Breast Cancer AUC	Best Pneumonia AUC
Momentum 0.9 & random resized crop	0.851 (Epoch 11)	0.981 (Epoch 38)
Momentum 0.9 & random rotation 20	0.874 (Epoch 198)	0.974 (Epoch 40)
L2 $\lambda = 0.001$ & random rotation 20 degrees	0.888 (Epoch 163)	0.978 (Epoch 43)
L1 $\lambda = 0.001$ & random rotation 20 degrees	0.845 (Epoch 34)	0.938 (Epoch 27)
ResNet50 & random rotation 20 degrees	0.835 (Epoch 259)	**0.982 (Epoch 87)**
DenseNet121 & random rotation 20 degrees	0.895 (Epoch 351)	0.977 (Epoch 92)
L2 $\lambda = 0.0001$ & random erasing	0.880 (Epoch 194)	0.968 (Epoch 41)
L2 $\lambda = 0.001$ & random erasing	0.832 (Epoch 190)	0.965 (Epoch 38)
L1 $\lambda = 0.0001$ & random flip vertically	0.908 (Epoch 175)	0.962 (Epoch 13)
L1 $\lambda = 0.0001$ & random rotation 20 degrees	0.870 (Epoch 179)	0.964 (Epoch 14)
Adam, L2 0.0001 & step size 1e-4	0.886 (Epoch 32)	0.856 (Epoch 19)
Adam, random rotation 20 degrees & step size 1e-4	0.586 (Epoch 0)	0.504 (Epoch 8)
Adam, random flip vertically & step size 1e-4	0.897 (Epoch 168)	0.963 (Epoch 43)
Adam, random flip vertically, step size 1e-4 & L1 $\lambda = 0.0001$	0.876 (Epoch 199)	0.956 (Epoch 36)

5 Conclusion

Our experiments show that, while many of these techniques are useful in their own right, the methods that improve performance on a particular dataset depend largely on the nature of the data and the optimization problem in question.

For AUC maximization on the breast cancer data, we found that L1 regularization with a λ of 0.0001 paired with random erasing as a data transform to be the second-best method with an AUC score of 0.908. Our best AUC score for the breast cancer tasks was 0.913, with a simple augmentation randomly flipping the data vertically. Generally speaking, we feel that there are a lot of surprising results we had across the board for the breast data, and that they usually happened in later epochs. Regularization with L2 usually correlated with higher breast scores, and augmentations also generally showed promise. From this, we might conclude that the breast cancer dataset is one highly prone to overfitting, and these measures taken to reduce overfitting are a step in the right direction towards a more robust model.

For the pneumonia data, several experiments produced high AUC scores — usually around ~0.97. Some such competitors were dropout regularization with p=.25 at 0.978, random resized crop & rotation of 20 degrees also at 0.978, and a momentum of 0.9 with random resized crop with 0.981. The best model for the pneumonia data was ResNet50 with a random rotation of 20 degrees, with a slightly higher 0.982. We can infer that the pneumonia dataset presented a simpler relationship for the models to learn, and correspondingly, our models often exhibited much higher AUC scores on this dataset.

Research in this field is far from complete. If we want to create technologies that can save lives, formulate systems that can diagnose nearly perfectly, further work is needed. AUC maximization for deep neural networks is not a trivial problem, and certainly not one that can be "solved" with our current methods. But with more research, we can continue to build better systems and save more lives.

References

A Deep Learning Library for X-risk optimization. *libauc 1.0.0 documentation.* (2023).
`https://docs.libauc.org/`

Densenet121. *PyTorch.* (2017a).
`https://pytorch.org/vision/main/models/generated/torchvision.models.densenet121.html`

Dropout. *PyTorch.* (2023a).
`https://pytorch.org/docs/stable/generated/torch.nn.Dropout.html`

PyTorch documentation. *PyTorch.* (2023b).
`https://pytorch.org/docs/stable/index.html`

IgnorResnet50. *PyTorch.* (2017b).
`https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html`

Shah, D. (2022, May 6). *The Essential Guide to data augmentation in Deep Learning. V7.*
`https://www.v7labs.com/blog/data-augmentation-guide`

Yadav, H. (2023, May 31). *Dropout in neural networks. Medium.*
`https://towardsdatascience.com/dropout-in-neural-networks-47a162d621d9`