Parallelization of a Neural Network for Breast Cancer Risk Prediction

Dina Galevska

October 2023

Abstract

This project investigates parallelizing a deep neural network (DNN) architecture for breast cancer risk prediction using the joblib library. Partitioning the dataset into subsets and leveraging multiple threads accelerates training. However, the modest dataset size limits parallelization's impact. Nevertheless, integrating an early stopping condition preserves efficiency and accuracy.

Keywords: Deep neural network, parallelization, early stopping condition, accuracy

Introduction

Breast cancer stands as one of the most pressing health concerns worldwide, impacting millions of lives annually. Early detection and accurate risk prediction are paramount in improving patient outcomes and survival rates. Leveraging advancements in machine learning, particularly deep neural networks (DNNs), presents a promising avenue for enhancing the accuracy and efficiency of such predictions. Neural Networks, inspired by the human brain, consist of interconnected artificial neurons. They are capable of processing vast amounts of data to derive meaningful insights. This project explores parallelizing a neural network framework for breast cancer risk prediction to assess the effectiveness of parallelized DNN architectures through meticulous experimentation and empirical analysis.

Related work

Breast cancer, first identified in Egypt, is one of the oldest types of cancer. It can be classified into two categories:normal and abnormal [6]. It is a common cancer in women, its early detection can significantly increase the survival rate of women and be much more effective. This paper [5] is mainly based on the learning transfer process for breast cancer detection.

There are many techniques for breast cancer pattern prediction and classification. Artificial neural network [3], combined neural network and decision tree model are used for classification. Some breast cancer data are classified using multi-layer perceptron neural network, combinatorial neural network, probabilistic neural network, recurrent neural network and support vector machine. Due to the importance of achieving highly accurate classification, [1] artificial neural networks can be used to improve the work of doctors in diagnosing breast cancer. In recent years, several studies have used ML techniques in health domains to detect breast cancer. Since the algorithms provide satisfactory results, some scientists have used them to solve certain challenges.[7] CNN Algorithm was used to predict and diagnose cancer in breast cancer images, which achieved a high accuracy of 88% [12] [11]. In [8], the authors presented breast cancer diagnosis using SVM technique and selected functionalities. The current report of the National Cancer Registry Programs states that "Breast growth accounts for 28-35% of all diseases among ladies in significant urban communities.[10]

Artificial intelligence has huge share in developing models to predict the occurrence and survival in breast cancer. In this paper[4] is presented a parallel Bayesian hyperparameter optimized Stacked ensemble model, developed using stacking of machine learning model, Deep Neural Network (DNN). Bayesian optimization with Gaussian Processes is used to find the best hyperparameters for the machine learning models. The parallelism is coupled with Bayesian optimization to address the high computational time. Machine learning and deep learning algorithms have been used to classify benign and malignant tumors. The breast cancer dataset containing a large number of samples and features was used. The paper highlights various models implemented Support Vector Machine (SVM), Multi-Layer perceptron classifier, Artificial Neural Network(ANN) etc. on the dataset taken from the repository of Kaggle.[9]The aim of this paper is to investigate the feasibility of using ensemble learning and big data fusion for breast cancer risk prediction and categorization. It is mainly based on building a precise model by combining the best features of different learning algorithms on diverse datasets. [2]

These studies highlight the importance of parallelization in optimizing the performance of neural networks for breast cancer prediction tasks.

Solution architecture

Building a Machine Learning Model

The solution I will implement for the problem of predicting Breast Cancer involves a Deep Neural Networks(DNN). Neural Networks are complex structures made of artificial neurons that can take in multiple inputs to produce a single output. They are inspired by the structure and functioning of the human brain. The basic building block of a neural network is the artificial neuron, also known as a perceptron. These neurons are organized into layers, typically consisting of an input layer, one or more hidden layers, and an output layer. The primary job of a Neural Network is to transform the given inputs into a meaningful output. Deep neural networks (Deep Neural Networks) or multi-layered perceptrons (Multi-layered Perceptions) represent a series of logistic regression models layered one above the other in multiple layers (input layer, a certain number of hidden strains and output layer). Since convolutional neural networks (CNNs) are particularly suited to tasks related to computer vision, image recognition, and spatial data, and my dataset contains textual data, I decided to use only general-purpose DNNs that can be applied to a variety of data types. The following figure shows the appearance of a neural network. The architecture of the neural network is as follows: in the input layer there are 7 neurons (perceptrons), in the first hidden layer 16, and in the second 32 neurons which are of course connected to the only neuron from the output layer, which represents the probability that one of the data instances belongs to class 1, because it is a binary classification.

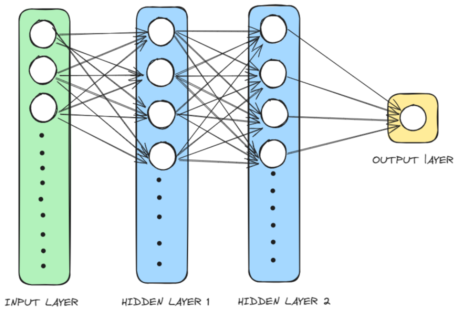

Figure 1: Deep Neural Network

To reduce the dimensionality, I decided on the statistical technique Principal component analysis (PCA). Dimensionality reduction is an important step in data analysis and processing and is used to reduce noise, improve data processing, and preserve the most important information in a data set.

First, the mean value for each of the dimensions in the data set is calculated. This is done for each attribute or feature in the data.

Next, the data is standardized. This includes centering (subtracting the mean) and scaling (dividing by the standard deviation) of the data. The covariance matrix between the different attributes is then calculated. This matrix describes how the attributes affect each other and whether they are correlated.

We choose how many components we want to keep depending on the dimensionality we want to reduce. The final step is to project the original data into the new space composed of the selected components.

To improve the weights of neural networks that are obtained during model training, I decided to use Cross Entropy and Adam optimizer, which are two important concepts.

Cross Entropy is a function used to measure the difference between two probabilities. Cross Entropy is used to measure the error, while the Adam optimizer is used to update the model parameters to reduce this error when training neural networks. The combination of these two concepts helps models learn efficiently.

Parallelization

When training the model, I decided to use the parallelization enabled by the joblib library, so that by splitting the data, the model is allowed to use multiple threads and split the data between them.Each thread works independently of the others, it can be part of a separate process, which allows parallel execution of tasks. The main advantage of this approach is that multiple threads can wait for operations without blocking the entire process, which improves execution efficiency. The use of many threads is chosen in order to utilize the capacities of the system and increase the training speed of the model. The data is divided into several parts, thus creating subsets of data (chunks) that will be trained in parallel. A new model is created for each chunk, and EarlyStopping callback is used to control the training and prevent the model from overfitting. EarlyStopping is a callback function that is used to automatically stop training a model when certain criteria are not met. This is useful for preventing overfitting or terminating training when the model no longer shows improvement on the validation set. Using the joblib library, models are trained in parallel for each chunk. Finally, the results are combined. Parallelization enables maximum utilization of system capacities.

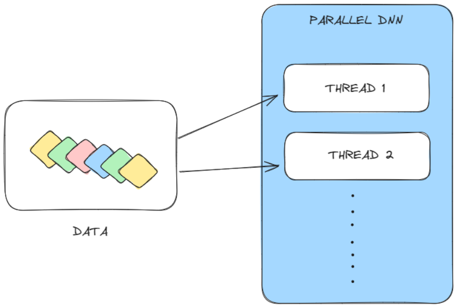

Figure 2: Parallelization

I did another experiment by evaluating the same model (DNN) on GPU. Running the model on the GPU increases parallelization intensity significantly. The GPU has a pronounced ability to parallelize operations within deep neural networks. The GPU, with its many cores and high parallelization capabilities, is ideal for tasks such as training and evaluating models.

Results

To analyze the classification models of different species. The basic tool for analyzing the contents of this model is accuracy, defined as the ratio between the number of correctly determined classes of all data entities from the test set and the number of all data from the test set.

Sequential Neural Network

The results of sequential neural network training are promising and have outstanding accuracy in breast cancer prediction. The model reaches up to 0.9649 accuracy, which indicates excellent results, the error made by the model in training is 0.170 which also indicates that the model is a good learner. It took 18 seconds to train the neural network sequentially, this is also contributed by the Early stopping callback function, the training is stopped early in order to prevent overfitting. I also looked at the PCA algorithm for sequential training of the neural network in order to reduce the dimensionality. The best results in this case were shown for 4 n principal components with an accuracy of 0.9561.

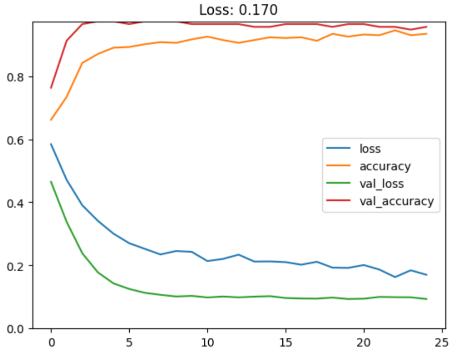

Figure 3: Loss that the model does

Parallelize a NN with joblib

In contrast, multi-threaded parallel training shows less good timing results, this is because adding parallelism is not always a guarantee that training will be faster. Partitioning data and operations in an insufficiently large set may cause additional operations to manage data and memory resources. Unlike the sequential model where I have one complete result for the entire training of the model, in the parallel code I get results for each chunk of training, so this also involves additional processing time. In this case, it took 25 seconds to train the neural network and the accuracy achieved during evaluation is 0.9565, and the training error is 0.218.

Parallelize the training of a NN with GPU

Evolving the same GPU model with parallelization enabled again, the training time was reduced to 14 seconds, which is better compared to sequential training. The model reached an accuracy of 0.9737, which is an excellent result and shows that the parallelization of the training successfully contributed to the improvement of the model's performance.

	Accuracy	Loss
DNN	0.9649	0.170
PCA	0.9561	0.197

Table 1: Sequential Neural Network

	Accuracy	Loss
DNN with multiple threads	0.9565	0.218
DNN on GPU	0.9737	0.153

Table 2: Parallel Neural Network

In this paper, I have listed the mean times for sequential and parallel neural network training. The time comparison chart is shown in Figure 5 below.

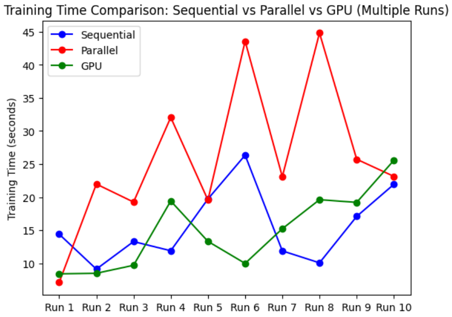

Figure 4: Time Comparssion

Conclusion

In conclusion, although the parallel training approach did not exhibit significant improvements for this model due to the dataset's limited size, it underscores the necessity of considering various optimization techniques. For many data sets, parallelization would be superior, but in this case, while partitioning alone may not yield substantial benefits, the implementation of an early stopping condition during training emerges as a critical factor in achieving commendable results in terms of both time efficiency and accuracy. This underscores the need for a comprehensive approach to refining models, considering factors like dataset details, algorithm complexity, and convergence methods, to optimize breast cancer risk prediction.

References

[1] Hussein A Abbass. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial intelligence in Medicine*, 25(3):265–281, 2002.

[2] Varshali Jaiswal, Praneet Saurabh, Umesh Kumar Lilhore, Mayank Pathak, Sarita Simaiya, and Surjeet Dalal. A breast cancer risk predication and classification model with ensemble learning and big data fusion. *Decision Analytics Journal*, 8:100298, 2023.

[3] Murat Karabatak and M Cevdet Ince. An expert system for detection of breast cancer based on association rules and neural network. *Expert systems with Applications*, 36(2):3465–3469, 2009.

[4] Parampreet Kaur, Ashima Singh, and Inderveer Chana. Bsense: A parallel bayesian hyperparameter optimized stacked ensemble model for breast cancer survival prediction. *Journal of Computational Science*, 60:101570, 2022.

[5] Aditya Khamparia, Subrato Bharati, Prajoy Podder, Deepak Gupta, Ashish Khanna, Thai Kim Phung, and Dang NH Thanh. Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidimensional systems and signal processing*, 32:747–765, 2021.

[6] Mazin Abed Mohammed, Belal Al-Khateeb, Ahmed Noori Rashid, Dheyaa Ahmed Ibrahim, Mohd Khanapi Abd Ghani, and Salama A Mostafa. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. *Computers & Electrical Engineering*, 70:871–882, 2018.

[7] Ting Pang, Jeannie Hsiu Ding Wong, Wei Lin Ng, and Chee Seng Chan. Deep learning radiomics in breast cancer with different modalities: Overview and future. *Expert Systems with Applications*, 158:113501, 2020.

[8] Md Mamunur Rahman, Yasaman Ghasemi, Eniola Suley, Yuan Zhou, Shouyi Wang, and Jamie Rogers. Machine learning based computer aided diagnosis of breast cancer utilizing anthropometric and clinical features. *Irbm*, 42(4):215–226, 2021.

[9] Monika Tiwari, Rashi Bharuka, Praditi Shah, and Reena Lokare. Breast cancer prediction using deep learning and machine learning techniques. *Available at SSRN 3558786*, 2020.

[10] R Vijayarajeswari, P Parthasarathy, S Vivekanandan, and A Alavudeen Basha. Classification of mammogram for early detection of breast cancer using svm classifier and hough transform. *Measurement*, 146:800–805, 2019.

[11] UN Wisesty, TR Mengko, and A Purwarianti. Gene mutation detection for breast cancer disease: A review. In *IOP Conference Series: Materials Science and Engineering*, volume 830, page 032051. IOP Publishing, 2020.

[12] Xiaomin Zhou, Chen Li, Md Mamunur Rahaman, Yudong Yao, Shiliang Ai, Changhao Sun, Qian Wang, Yong Zhang, Mo Li, Xiaoyan Li, et al. A comprehensive review for breast histopathology image analysis using classical and deep neural networks. *IEEE Access*, 8:90931–90956, 2020.