Abstract

Artificial Neural Networks (ANN) are machine learning models loosely based on the biological structure of the brain. In the past, they have shown success in approximating solutions for multivariate prediction and classification problems, including stock market analysis and the Travelling Salesman Problem (TSP). In this investigation, the goal was to build an extensible, multipurpose ANN for the analysis and classification of images. The neural network was built using an object-oriented approach in Python with an input layer, a hidden layer, and an output layer with variable numbers of neurons per each layer. A logistic sigmoid function was implemented in order to transform the propagated values, and the quasi-Newton Conjugate Gradient algorithm was implemented in order to perform batch gradient descent to minimize the cost of the ANN.

In order to test the viability of this neural network, it was applied for the recognition and classification of optical handwritten digits. An application was built in order to collect supervised pixel maps and images of optical digits, which were then used to train the neural network. Ultimately, the neural network was able to identify test samples of digits with an accuracy of 82.4% across 1560 trials; however, the ANN did show some signs of overfitting. Multiple techniques were employed to overcome overfitting including early stoppage and regularization, although it is hypothesized that an increased training sample size will diminish overfitting and increase accuracy.

Furthermore, this ANN was also applied for the categorical diagnosis of fine needle aspirates (FNA) of breast cancer tumors. Supervised biopsy data was acquired from the UCI Machine Learning Database consisting of 32 attributes compiled from each FNA image. The neural network was able to diagnose test samples as either malignant or benign with an accuracy of 97% across 715 trials with no signs of overfitting. Future endeavors include investigations into more applications as well as more convoluted, deep neural networks.

Artificial Neural Networks (ANN)

- Machine Learning model loosely inspired by biological structure of brain
- Takes inputs, transforms and weights, gives outputs

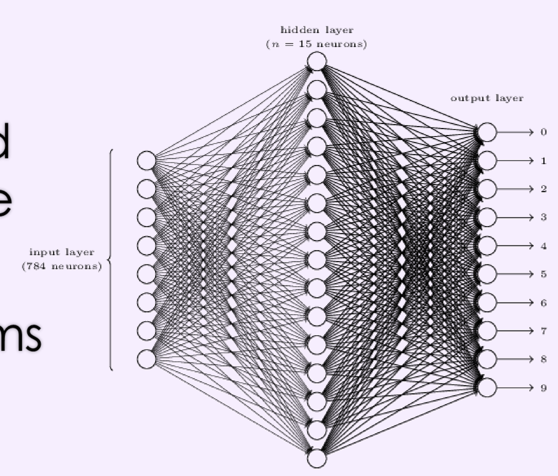


Figure 1. Framework of an Artificial Neural Network (ANN)

See section 1

Goal/Objective

To develop a general-purpose, extensible artificial neural network for the supervised analysis and classification of images. To apply this ANN for Optical Digit Classification and categorical diagnosis of fine needle aspirates of breast cancer tumors.

Network Class Constructor

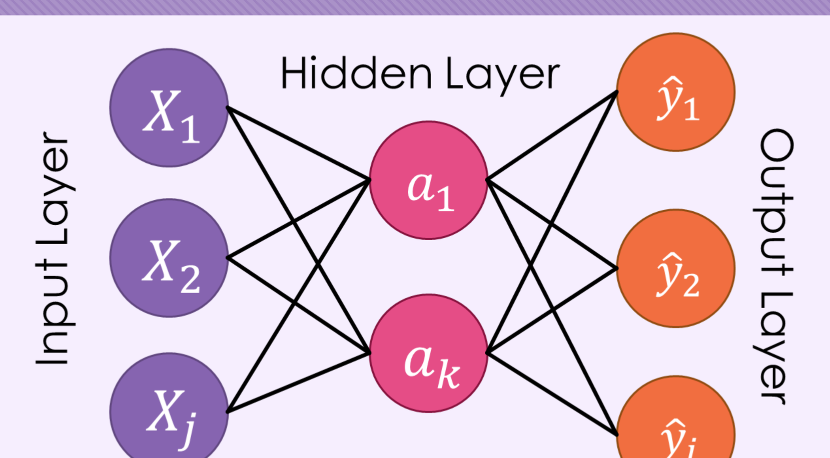
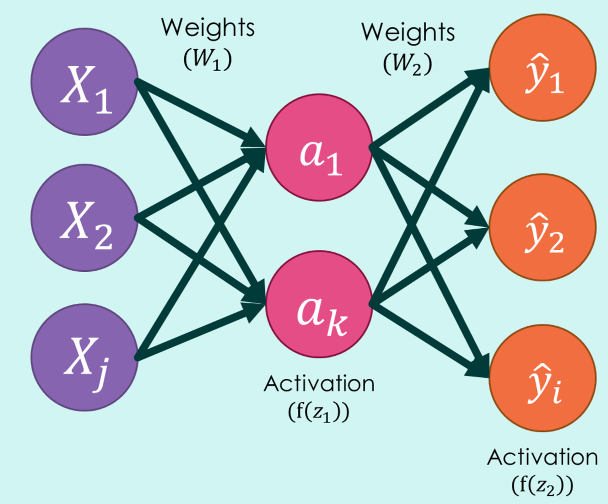


Figure 2. High-level layout of an ANN design

See section 2.1

Forward Propagation

$$z_1 = XW_1$$
$$a = f(z_1)$$
$$z_2 = aW_2$$
$$\hat{y} = f(z_2)$$
$$\hat{y} = f(f(XW_1)W_2)$$



See section 2.1.1

Figure 3. High-level diagram of forward propagation

Backwards Propagation

- Use differential analysis to tune weights to minimize error of outputs

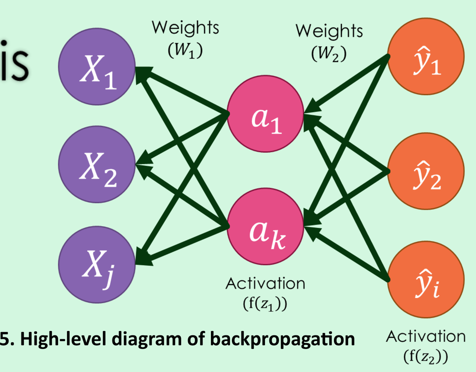


Figure 5. High-level diagram of backpropagation

- Cross Entropy Cost Function

$$J = -\frac{1}{N}\sum y_n \ln(\hat{y}_n) + (1 - y_n)\ln(1 - \hat{y}_n)$$

Compute Average — Compute Cost for Correct Node — Compute Cost for Incorrect Nodes

See section 2.1.2

Digit Recognition

- Use handwritten digits as training data (data collected through PyGame application)
- Crop and compress image to 16x16 pixels
  - Input uses 1-of-N encoding (-1 for White, 1 for Black)
- Output is vector of length 10 (for each digit)

Figure 6. Example raw optical digit image

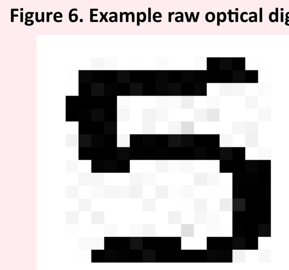


Figure 7. Example optical digit pixel map

See section 2.2.1

Testing

- 20% of data used for testing / 80% used for training

Digit Recognition Results	Breast Tumor Biopsy Results
Overall accuracy of **82.4%** across 1560 trials (some signs of overfitting)	Overall accuracy of **97%** across 715 trials (NO signs of overfitting)

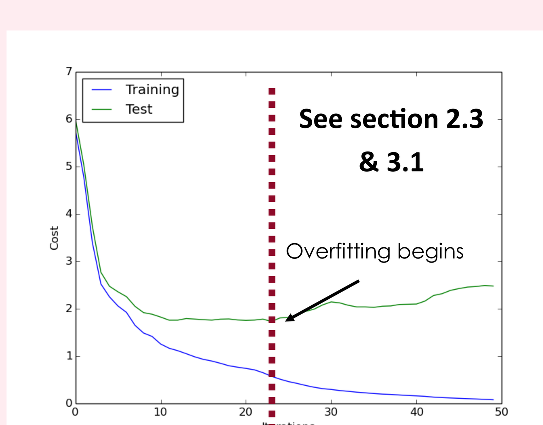


Figure 10. Cost of Training and Test Data vs Iterations of Minimization Algorithm (Optical Digit Recognition)

Figure 11. Cost of Training and Test Data vs Iterations of Minimization Algorithm (FNA of Breast Tumors)

Sigmoid Activation Function f(z)

- Differentiable and introduces non-linear attributes

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$f'(z) = \frac{e^{-z}}{(1 + e^{-z})^2}$$

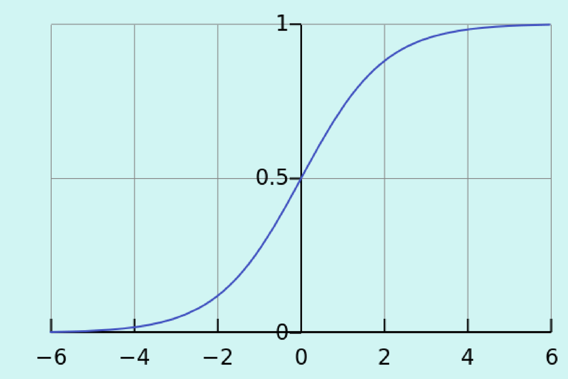


Figure 4. Logistic sigmoid activation function

See section 2.1.1

Batch Gradient Descent with Conjugate Gradient

- Iteratively computes gradient of cost function
- Use gradient to move in negative direction

See section 2.1.2

$$\nabla J = \left(\frac{\partial J}{\partial W_1}, \frac{\partial J}{\partial W_2}\right)$$

$$\frac{\partial J}{\partial W_2} = a^T \cdot \frac{\hat{y} - y}{N}$$

$$\frac{\partial J}{\partial W_1} = X^T \cdot \left[\frac{[\hat{y} - y]}{N} \cdot {W_2}^T \cdot f'(z_2)\right]$$

Diagnosis of Breast Tumors Biopsies

- Input data from Breast biopsies from UCI Machine Learning Database
- Data consists of many attributes of biopsies (radius, texture, perimeter, smoothness, etc)
- Single Output (Benign vs Malignant)

Figure 8. Fine Needle Aspirate of Malignant Tumor

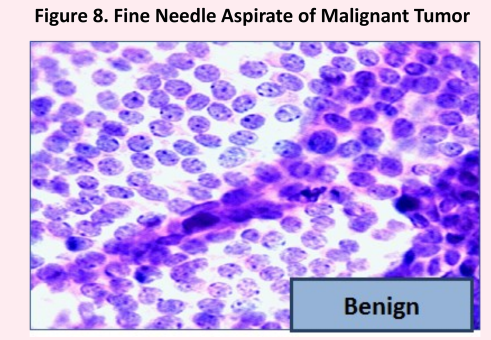


Figure 9. Fine Needle Aspirate of Benign Tumor

See section 2.2.2

Accounting for Overfitting

- Early stoppage
  - Stopped iterations at around 40, computed average for beginning of overfitting
- Need more data (sample size too small compared to number of inputs/outputs)
- Regularization constant

$$\frac{\lambda}{2n}\sum {W_j}^2$$

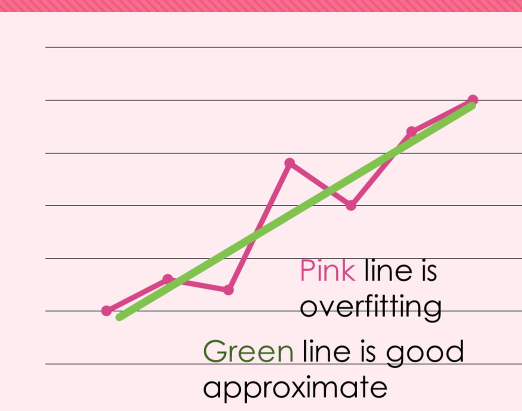


Figure 12. Overfitting illustrated

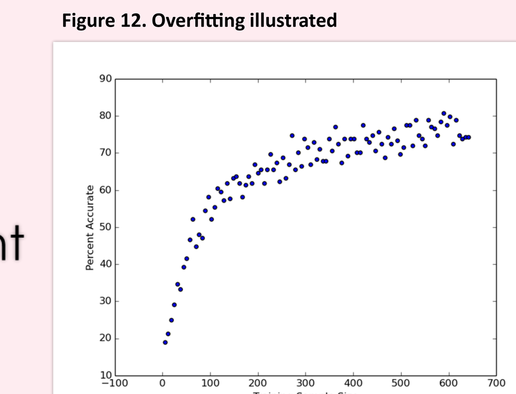


Figure 13. Training Sample Size vs Percent Accuracy

See section 2.4 & 3.1

Determining Hyperparameters

- % Accuracy vs Hidden Layer Size
- Constant iterations and samples
- Change is negligible after approximately 10

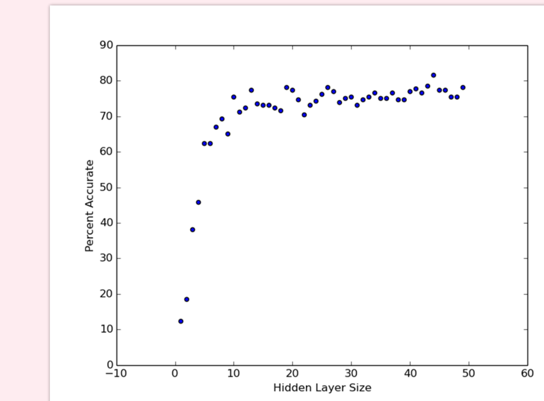


Figure 14. Percent Accuracy vs Hidden Layer Size

See section 2.4 & 3.1

Discussion

Overall, the Artificial Neural Network framework built in this project illustrated a large degree of success. For one, it was highly extensible in that it proves its capability to be applied to any number of supervised image classification problems. This is proven by its successful application to both optical digit recognition as well as categorical diagnosis of fine needle aspirate biopsies of breast cancer tumors.

In the application to optical digit recognition, testing against part of the unsupervised sample yielded an accuracy of approximately 82.4%. This is a fairly accurate result; however, any lack of accuracy may likely be attributed to the neural network's conjugate gradient minimization algorithm overfitting the training data, as it is evident in Figure 10. The strategies employed to avoid this issue include early stoppage and adding a regularization constant term to the cost function. Although these showed some success in alleviating the effects of overfitting, it is likely that overcoming overfitting will drastically decrease and accuracy will increase as the training sample size increases. This is illustrated by Figure 13, which shows that percent accuracy and sample size show a logarithmically growing proportionality.

Furthermore, in the application for breast tumor biopsy diagnosis, testing against part of the unsupervised sample yielded an accuracy of 97%. As it is evident in Figure 11 there are relatively no signs of overfitting. This is a very accurate result; although miniscule, any lack of accuracy may likely be attributed to outliers in the test data set not accounted for when training, incomplete minimization of the cost function, and image inconsistencies. Similarly, an increase in the training data's sample size would likely increase the accuracy of the neural network.

Future implications of this project include investigation into deeper, more convoluted neural networks. In this project, only one hidden layer was implemented into the neural network, but it may be useful to recursively extend the current backpropagation algorithm to include a larger number of hidden layers. Additionally, it may be useful to collect more training data or apply this ANN to more image classification problems in order to fine-tune it even further. Furthermore, it may also be useful to extend this neural network to continuous or unsupervised problems in the future and optimize approximate solutions to stock market analysis and the Travelling Salesman Problem (and other NP problems).

Successes

- High % accuracy for both applications
- Efficient gradient descent algorithm
- Neural network is extensible

See section 4

Developing

- More training data (esp. for ODR)
- More applications
- Bootstrap aggregation of data set (esp. for ODR)
- Investigation into more convoluted/deep neural networks

Practical Applications

- Optical Character Recognition (OCR)
- Stock Market Prediction Analysis
- Travelling Salesman Problem (TSP)
- Other multivariate fuzzy predictions
- Image Classification (Digits, Medical Images, etc.)

See section 1