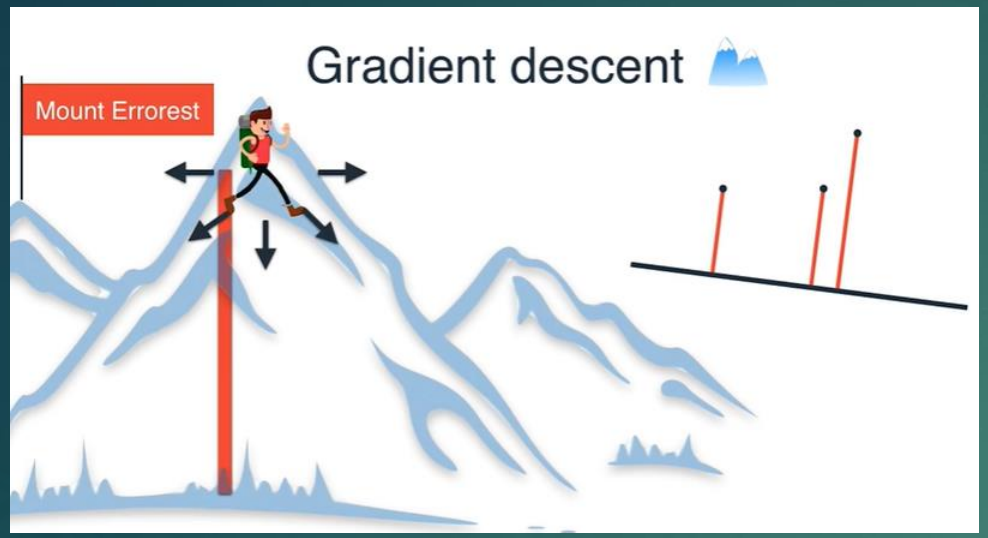
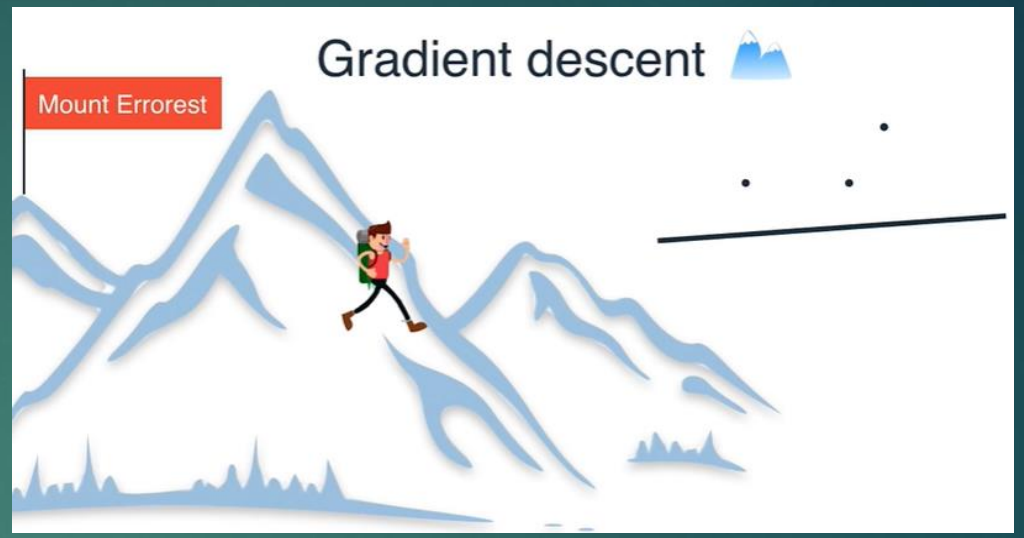
A study on gradient descent algorithms

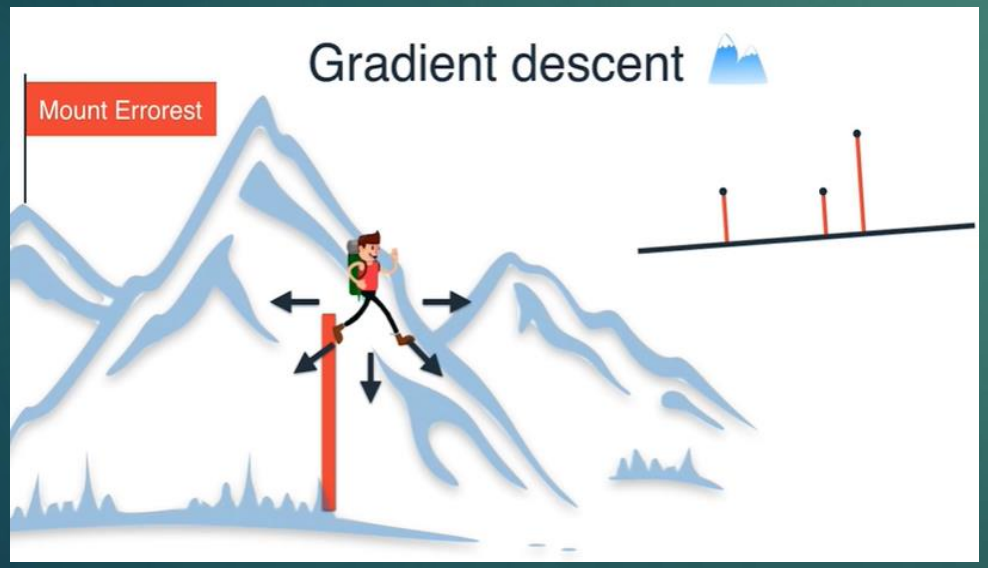
BY: KARTHIKEYA CHALLA

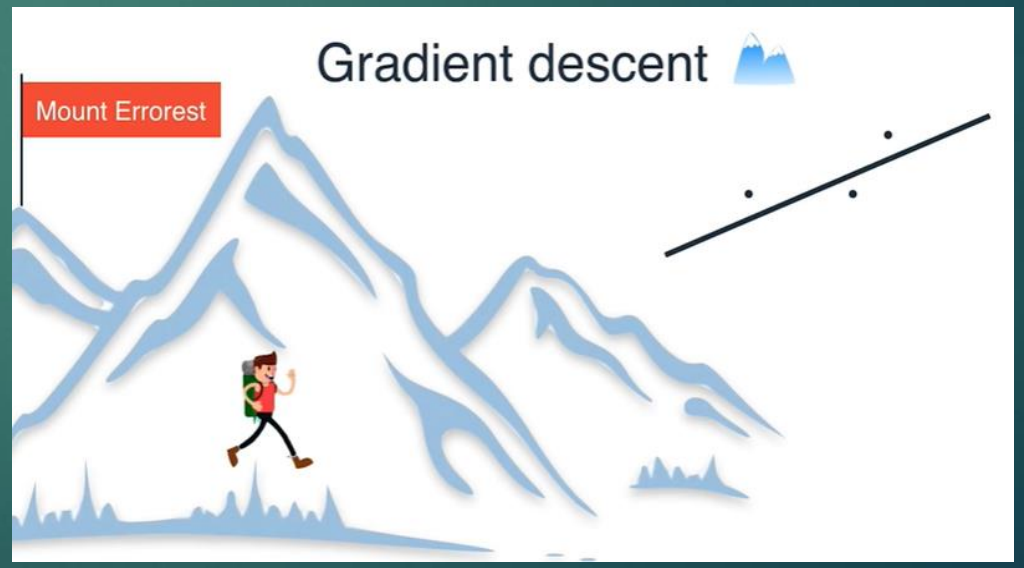
Need for gradient descent

- The aim of a neural network is to reduce the loss
- This can be achieved by updating the learning parameters – Weights and Biases
- Gradient descent is an update rule


Gradient descent
Mount Errorest


Gradient descent
Mount Errorest


Gradient descent
Mount Errorest


Gradient descent
Mount Errorest

Gradient Descent Algorithms

- SGD
- ADAM
- ADAGRAD
- RMSPROP
- ADAMAX
- MOMENTUM

Problem Statement and challenges

- Due to many gradient descent algorithms and large sets of various genres of data it becomes challenging to choose a right optimizer for effective output and high accuracy
- It consumes lots of time and computing power to re-run the models on large data sets in real world
- Thus with focus on image classification problem:
    a) Which GD algorithm is effective?
    b) Pros and Cons of each of the algorithms?

Related Work

- **An Experimental Approach towards the Performance Assessment of Various Optimizers on Convolutional Neural Network (https://ieeexplore.ieee.org/abstract/document/8862686)**
- **Improving Generalization Performance by Switching from Adam to SGD (https://arxiv.org/abs/1712.07628)**
- **Improved Adam Optimizer for Deep Neural Networks (https://ieeexplore.ieee.org/abstract/document/8624183)**

Project Methodology - Aim

- This project does a comparative study on various gradient descent algorithms with focus on image classification

Project Methodology - Datasets

- The datasets are obtained from Tensorflow Datasets including:
  - a) Oxford Flowers
  - b) Cassava Leaf (Also a Kaggle competition)
- The datasets are divided into training, validation and testing sets to perform the analysis

Project Methodology – Transfer Learning

- The transfer learning method is used for fitting the data for classification. The following models are used:

  a) Mobilenet

  b) Inception

  c) Resnet50

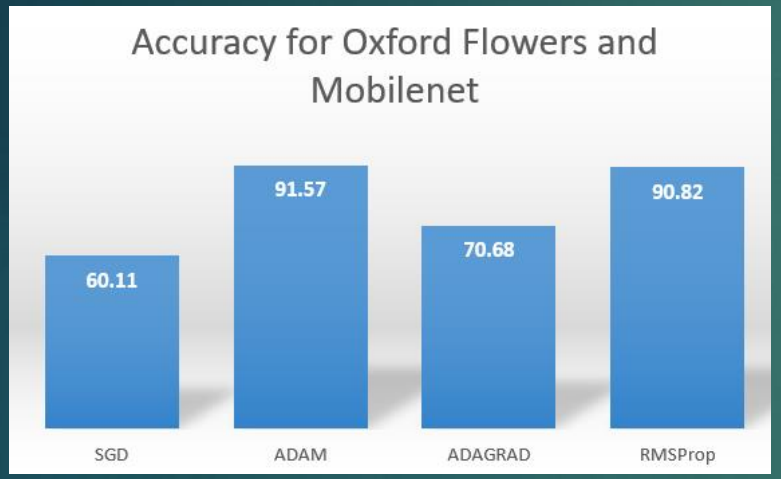
Project Methodology - Optimizers

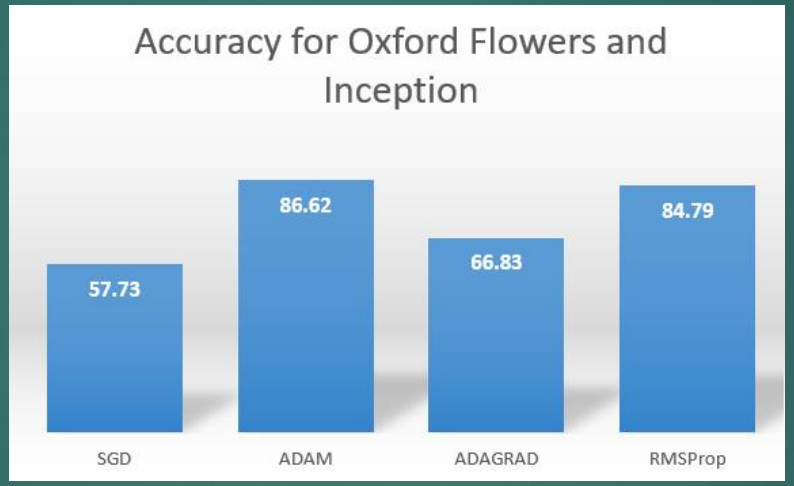
- For compiling the model and data the following optimizers are used:
  a) SGD
  b) ADAM
  c) ADAGRAD
  d) RMSProp

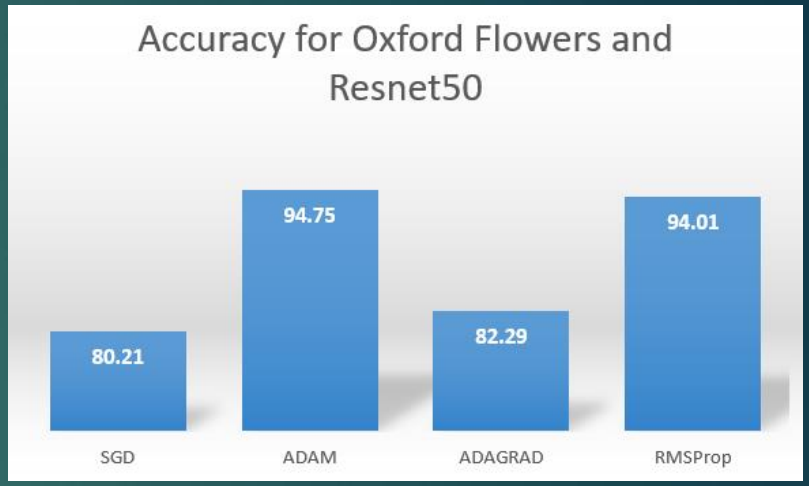
Project Methodology - Results

- The accuracy metric are used for the performance evaluation
- The results for the comparative study are provided in the form of plots and bar graphs

Results for Oxford Flowers

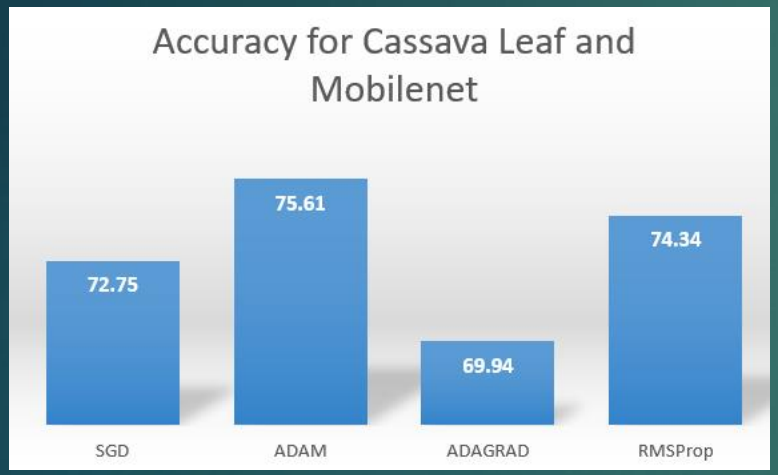


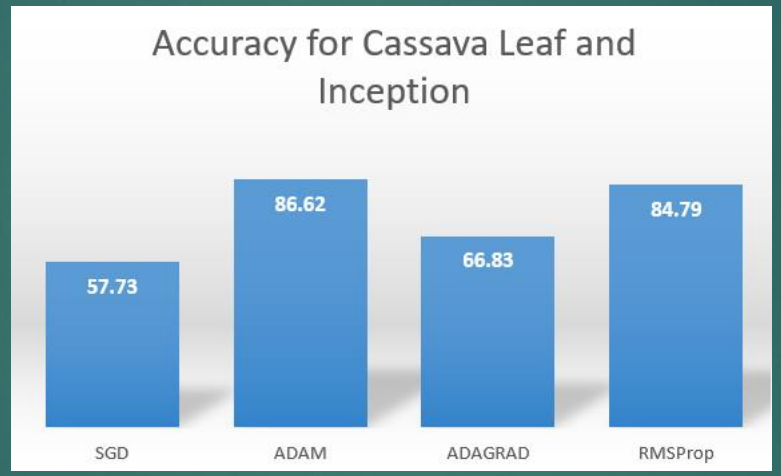


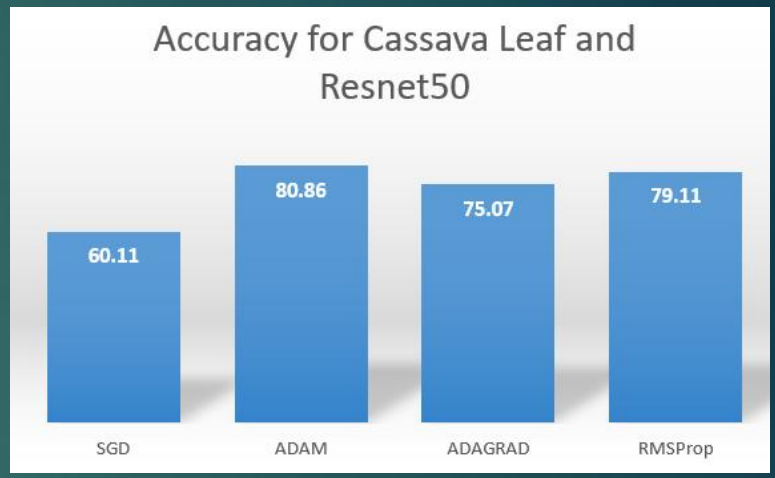


We can observe from the bar graphs above that the ADAM is a better algorithm for image classification example.

Results for Cassava Leaf

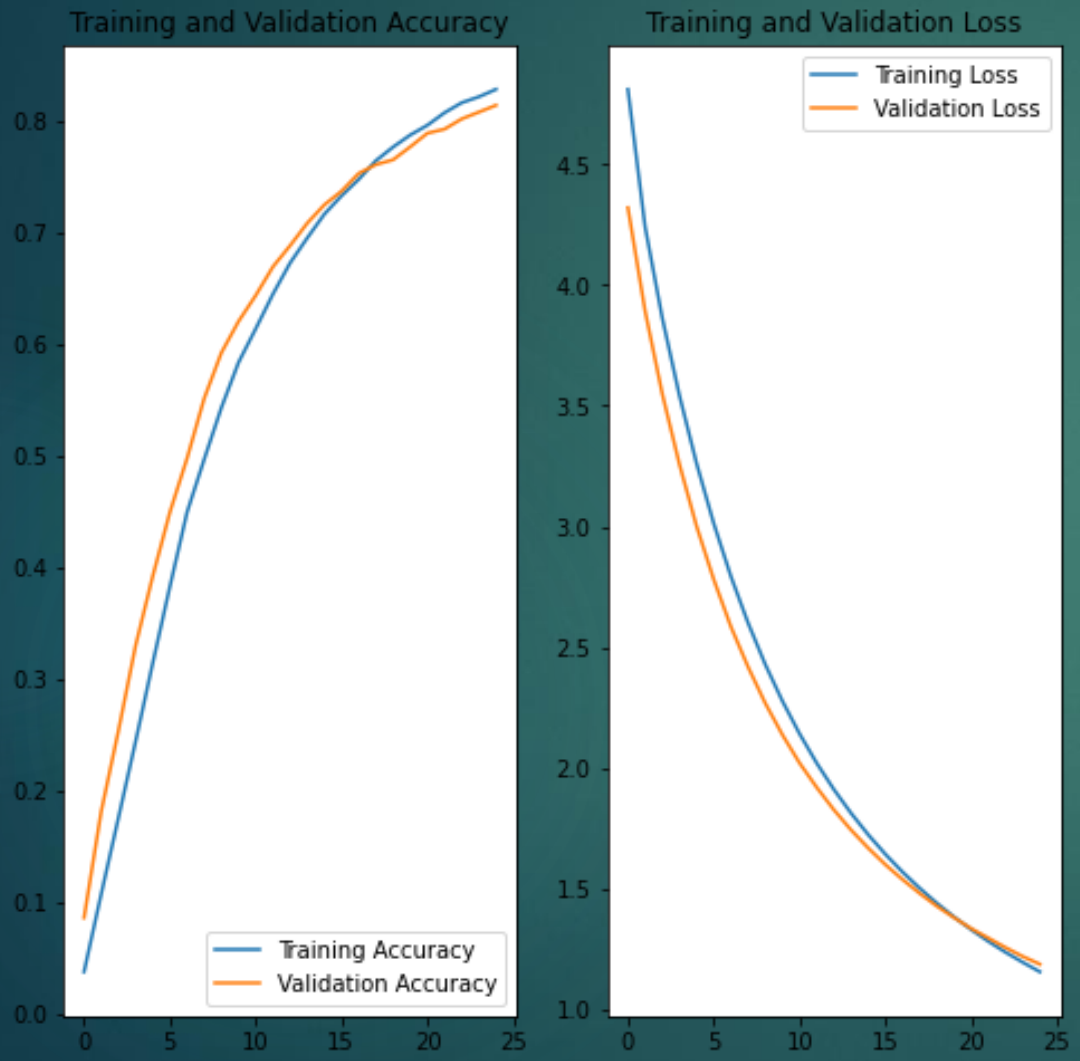






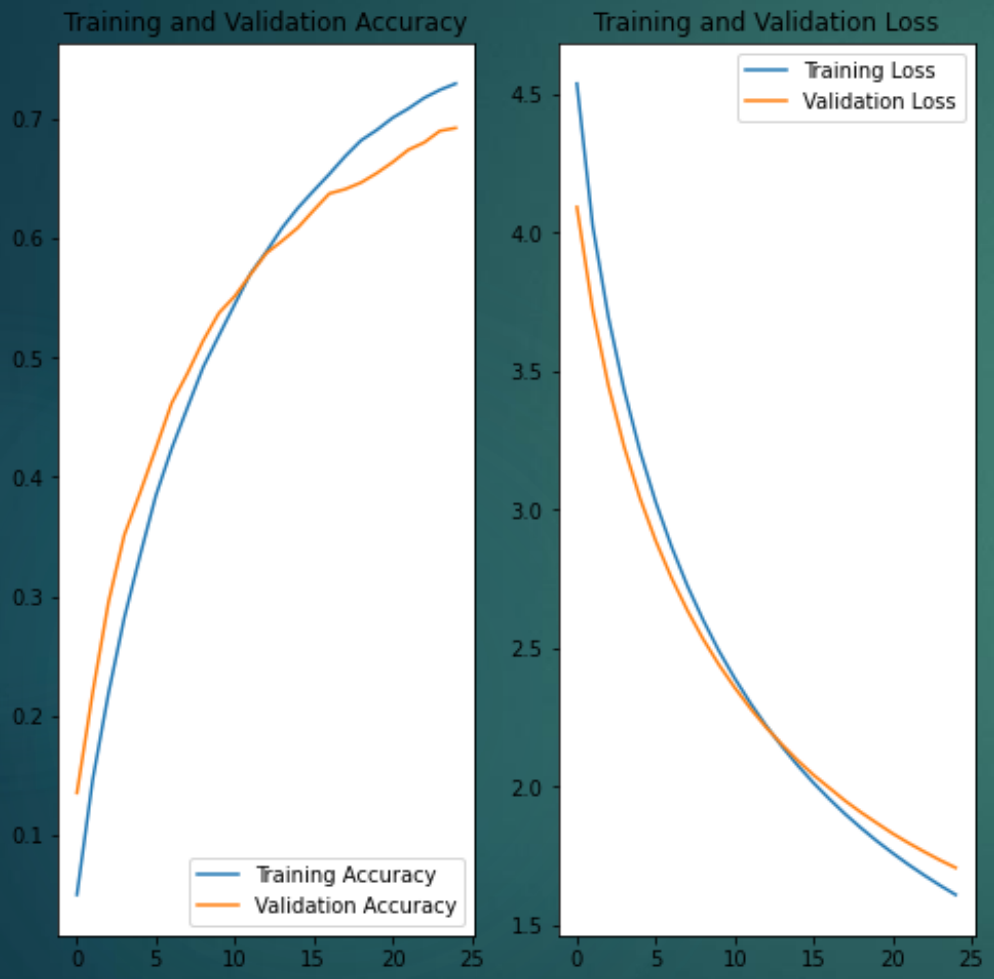
We can observe from the bar graphs above that the ADAM is a better algorithm for image classification example.

Pros and Cons of SGD



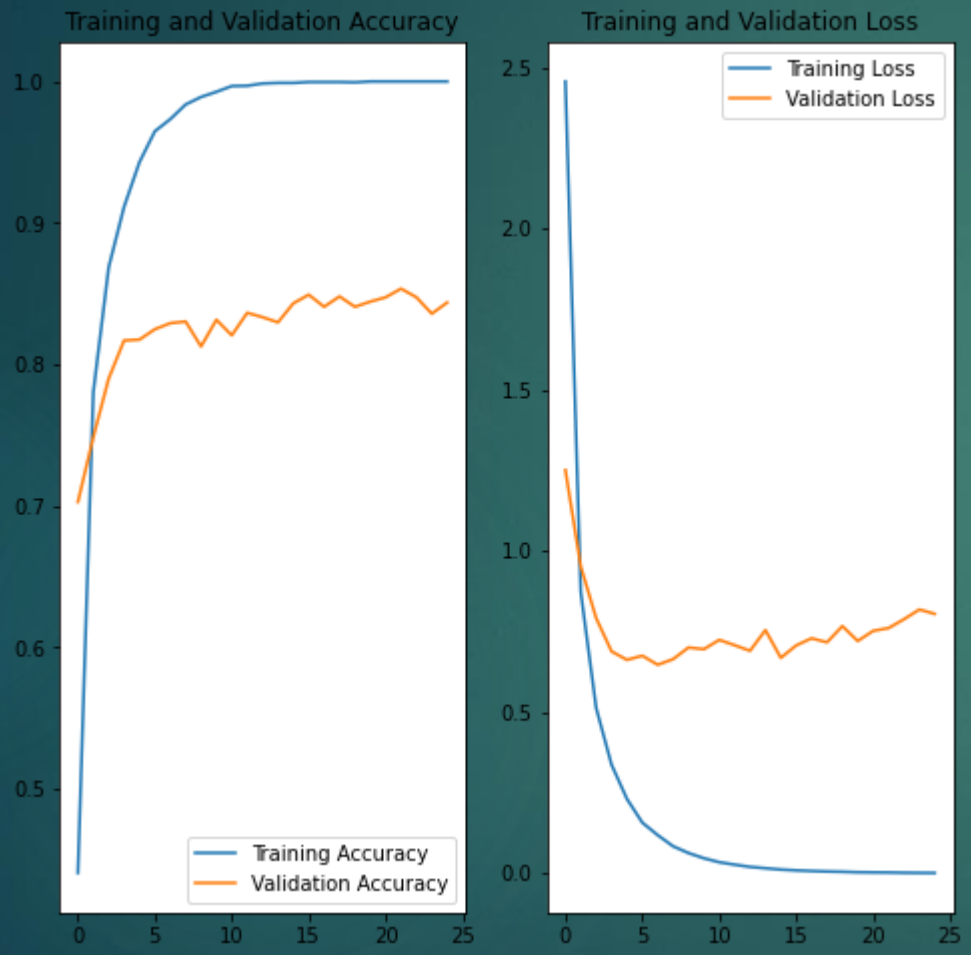
- SGD updates the parameters for each observation, hence it is faster.
- But it is only a stochastic approximation from a random subset of data and hence may not be accurate
- Because of the approximation it may get stuck at the local minima and will need higher number of epochs increasing the learning time
- For lower number of epochs the accuracy of the model will be poor

Pros and Cons of ADAGRAD



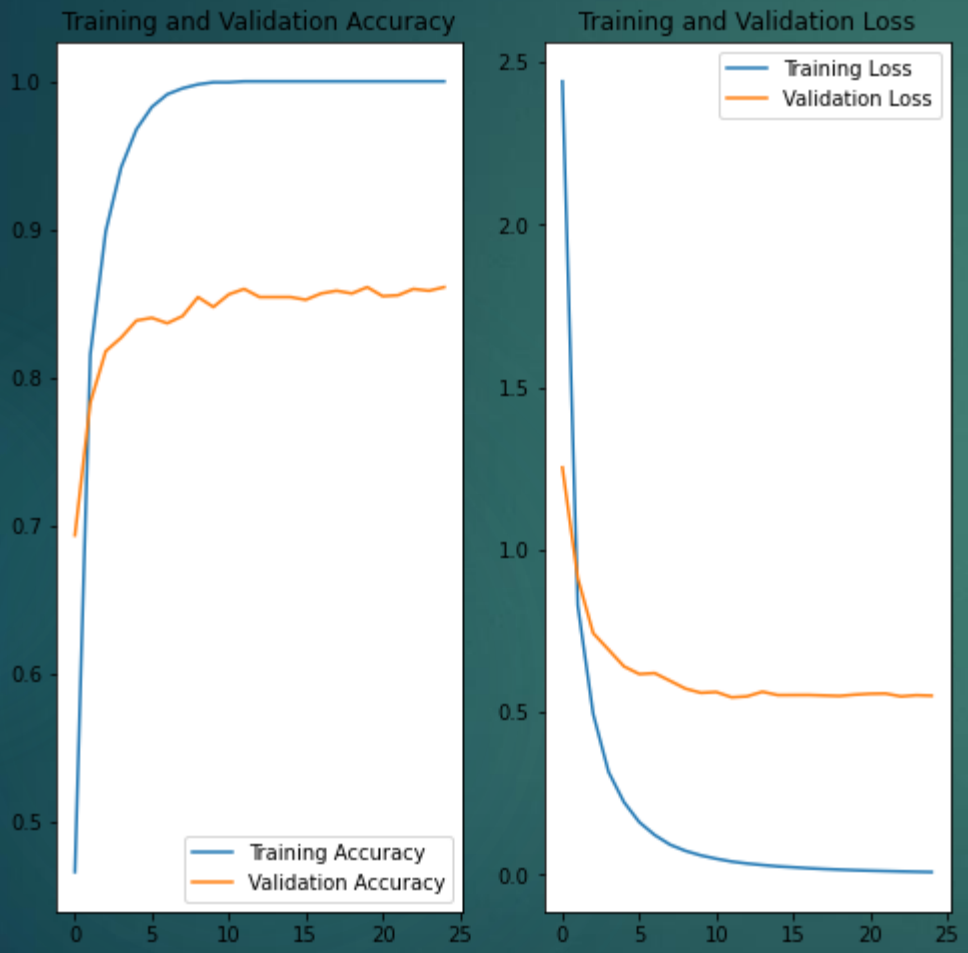
- ADAGRAD updates the parameters with an adaptive learning rate
- It is a running sum of squares of gradients and hence effective learning rate will be either high or low based on the accumulated sum
- Thus the learning rate decays very aggressively causing poor accuracy

Pros and Cons of RMSProp



- RMSProp overcomes the decaying learning rate issue of ADAGRAD as it just accumulates a fraction of squares of weights
- This allows the algorithm to keep learning and hence improves the accuracy
- But the issue is there are too many oscillations which can be confusing

Pros and Cons of ADAM



- ADAM brings in the adaptive momentum and adaptive learning rate methodologies which overcomes the issues of SGD and ADAGRAD
- The algorithm converges fast, keeps learning at faster rate hence improving speed and accuracy
- Also reduces the number of oscillations

Conclusion

- ADAM optimization algorithm with fewer oscillations moves deterministically in the right direction improving the speed by converging faster and better accuracy
- Therefore from the experiment results and graph, we can safely assume that ADAM would be the best choice of optimizer for the image classification application irrespective of the model being used

THANK YOU ☺