Project 3: Predictive Modeling for Image Classification

Team 6

Luke Arceneaux, Kejun Liu, Srushti Divyesh Sanghavi,
Qingyang Tang, Wenchang Zhu, Xinyu Zhu

Problem statement

- Our client is interested in creating an mobile AI program that accurately classifies the images. The portability of this AI program (holding storage and memory cost) and the computational efficiency (test running time cost) are of great concern to our client.

Task

- Baseline model
  - Logistic model with 24% accuracy
- Imperfect dataset
  - 50k images with noisy labels
  - Only 10k clean labels

- Model I
  - More sophisticated than the baseline model
  - Treat noisy labels as clean labels
- Model II
  - Same predictive model as Model I
  - But address label noise issues

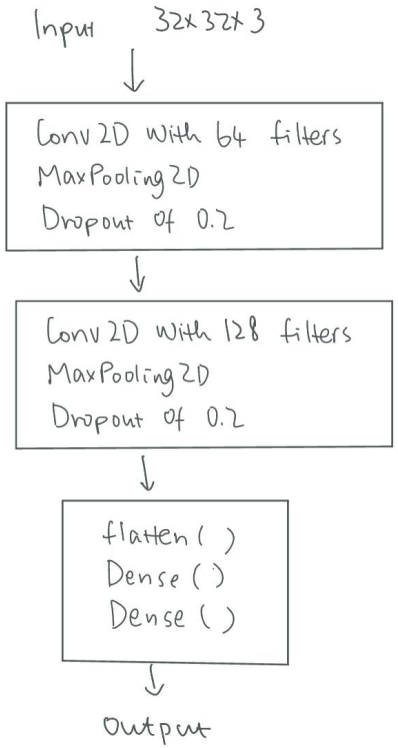
Exploration Models

- Exploration

Models	Accuracy
Decision Tree	14%
Random Forest	26%
KNN	21%
Naive Bayes	25%
CNN	46%

Model I - CNN

- Data
  - First 10k images with clean labels as **test data**
  - For the remaining 40k images with noisy labels, we take 80% as **training data** and 20% as **validation data**
- Model
  - Mainly 2 CNN layers
  - Batch size=32; epoch =8; optimizer: adam; loss: sparse categorical cross entropy
- Performance
  - Accuracy: around **47.59%**
  - Training time: around **4 minutes**



Model II - Label Cleaning

- Use pretrained VGG16 model + additional layers on 8k clean labels, test on 2k
  - Accuracy: 56.25%
- Train again on the entire 10k clean labels for later use
- Generate predicted labels for all 50k images
  - Originally, 3968 / 10k correct labels
  - Now, 6613 / 10k correct labels
- Transfer Learning
  - Take relevant parts of models and apply it to a new but similar problem
  - Advantage: saving of resources and improved efficiency when training new models

Model II: Use CNN on cleaned dataset

- Data
  - Training: 40k generated labels from label cleaning model
  - Test: 10k clean labels
- Performance
  - Accuracy: **58.27%**
  - Training time: **5.04 mins**

Summary

- Performance improvement & running cost tradeoff
  - Decent improvement in accuracy against the baseline model
  - Longer running time but acceptable given the improvement in accuracy

	Baseline	Model I	Model II
Accuracy	24%	47.56%	58.27%
Running time	< 10s	4 mins	5 mins