5243 Project 3: Image Predictive Analytics

Group8: Zhejing Shi, Jiapeng Xu, Yudan Zhang & John Podias

Task

We are tasked with building 2 additional image classification models that could improve upon a clients existing logistic regression model and tackle the issue of label noise

Priorities of model:

- Portability (holding storage and memory cost)
- Computational efficiency (running time cost)
- Predictive performance

Baseline Model

- Simple Logistic Regression model that treats the noisy labels as clean labels
- Accuracy: 20%
- Quick run time

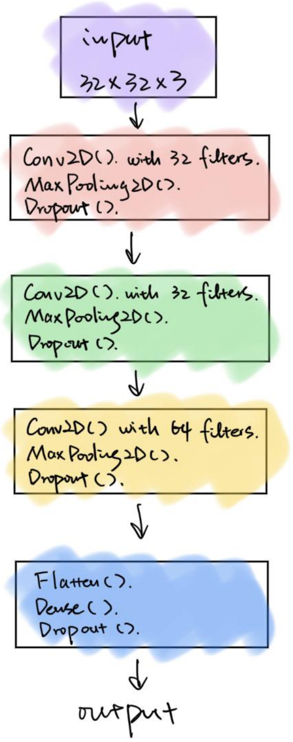
Logic Procedure for Model I

Dataset: 10k clean labels + 40k noisy labels = 50k images

- Assumes the input 50k dataset are all clean (ie, assuming no noisy labels).
- Mainly have 3 CNN layers + other layers such as flatten and dropout

The accuracy for model I is about **20**%.

Training time: 5 mins



Logic Procedure for Model II

Dataset: 10k clean labels & 40k noisy labels

Model II uses effective strategies such as **transfer learning (**Mobile Net**)** and l**abel correction** to handle label noise

<u>Step 1</u>: Split 10k images with clean label into train (70%) & test (30%) then fit the Model by the training dataset.

<u>Step 2 (Label updating):</u> Use trained model from step 1 to predict the labels of 40k images with noisy labels

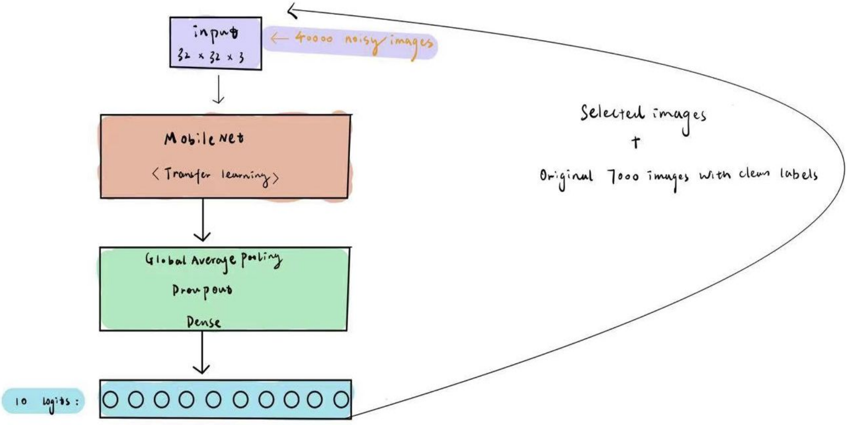
<u>Step 3 (Image selection):</u> Each image will get 10 logits (classes) from step 2. We only select an image if one of its logit > 0.85 (setting parameter) as a way to be confident in our predicted labels

<u>Step 4:</u> Combine the selected images with the original training dataset (70%) to refit the model

Model accuracy : **70**%.

Runtime: about 20 minutes

Label updating and image selection



Only if one of the logit ≥ 0.8, The input images and corresponding updated labels will be selected

Hyperparameters of Model II

Optimizer: Adam

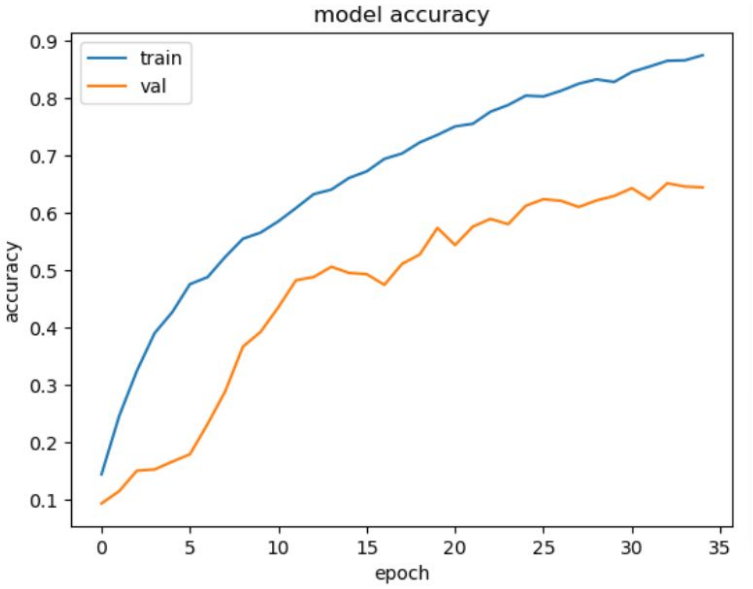
Learning rate: 0.0001

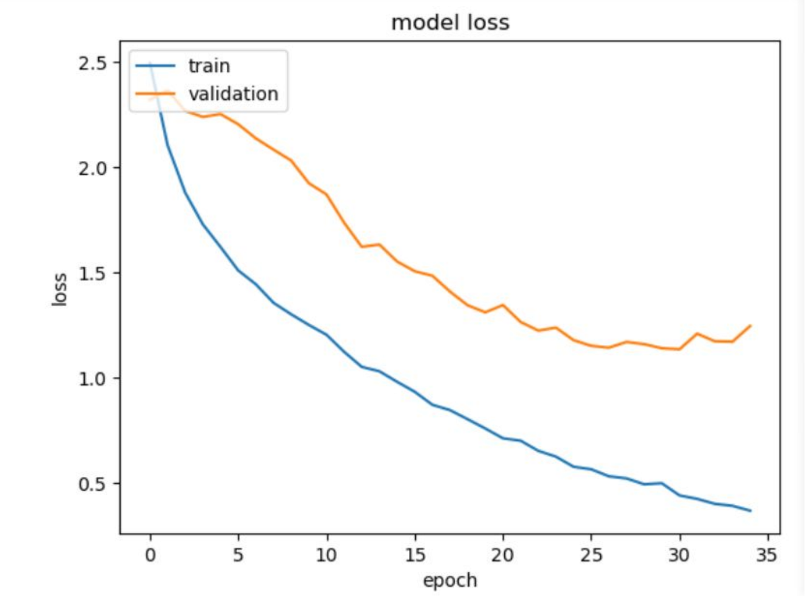
Batch size: 32

Epoch: 60

Early stop patience : 4

Model II Training and Validation for Accuracy and Loss





Model II Notes:

1. Cannot guarantee how many images will be selected for new training set in model II
    a. The setting parameter 0.85 can change. After testing, we can get a reasonable number of images from the 40k images (eg, 22546/40000) by setting it to 0.85.
2. The sum of the 10 logits from model II equals to 1

Final Recommendation

- Model II provides a large increase in accuracy (70%) against the client’s baseline model and would be a better option
- The runtime is slightly longer than the logistic regression model, but we will accept that for the increase in accuracy
- Interesting strategies used for label noise:
  a. Transfer learning
  b. Label correction