Advanced Data Analytics W24
Assignment 2

February 15, 2024

1 Submission

We highly encourage you to use Google Colab for this assignment!

You need to submit your analysis as an executable Python Jupyter Notebook file, named "ADA-1234-Assn2.ipynb", where 1234 stands for the last 4 digits of your student ID.

You should use Markdown cells in Jupyter Notebook to highlight answer to each question. Your submission will be evaluated based on the performance of your model on the provided testing dataset, and the quality of your presentation in the ipynb file.

An "I uploaded the wrong file" excuse will result in a mark of zero.

2 Background

E-commerce has changed the business world and the way we purchase items. One common problem faced by the ecommerce owner/platform is to tag appropriate labels with images of products so that those products can pop out when a relevant query is entered. Unfortunately, human annotation is expensive and error-prone. Thus, an automated solution that can identify the tags of images is needed.

The goal of this assignment is to create a classifier that can identify the label of a fashion product image using CNN and iterative improve the performance of your model.

You can reuse code from online resources, but you can not copy answers from other students in the same class. This is an independent homework.

3 Dataset

To complete this assignment, you must download images from `https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small`, and follow the given train.csv and test.csv file to setup the train and test dataset.

4 Part 1: Image Classification using CNN (40 points)

In the first part, your task is to create a basic CNN model that can identify if an image is related to the following 13 subcategories: Topwear, Bottomwear, Innerwear, Bags, Watches, Jewellery, Eyewear, Wallets, Shoes, Sandal, Makeup, Fragrance, Others. More specially, you are given a dataset containing 44,441 fashion product images. You are also given two files, i.e., train.csv and test.csv containing the meta-data (image id, label, productDisplayName) related to each image.

We do not have specific requirements on the structure of the CNN model. But you need to describe the motivation of your baseline structure in a markdown cell above the model class.

After you complete the required experiment for part 1, use a markdown cell to present your conclusion on the performance of your baseline approach on **train, validation, and test data set.** All your statements should be supported by experiment results.

5 Part 2: Improved Image Classification (60 points)

Your task in the second part is to create two enhanced models, leveraging the following two methods:

- Tuning one hyper-parameter and explain why this is worth to tune.
- Data augmentation, i.e., generating more images for training.

After you complete the required experiment for part 2, use a markdown cell to present your conclusion on the performance of two improved models on **train, validation, and test data set.** All your statements should be supported by experiment results.