Udacity ML Nanodegree Capstone

Project Proposal

Domain Background

This project idea comes from the Shopee - Price Match Guarantee kaggle competition. The domain is describe in the competition like this:

“Retail companies use a variety of methods to assure customers that their products are the cheapest. Among them is product matching, which allows a company to offer products at rates that are competitive to the same product sold by another retailer. To perform these matches automatically requires a thorough machine learning approach.

The applications go far beyond Shopee or other retailers [to supporting] more accurate product categorization and uncover[ing] marketplace spam. Customers will benefit from more accurate listings of the same or similar products as they shop. Perhaps most importantly, this will aid ... shoppers in ... [the] hunt for the very best deals.”[1]

Problem Statement

This is a classification problem where the model takes in a product posting (in the form of an image and it’s related features) and predicts a set of labels (in the form of a list of predicted matching product postings).

The problem is described in the competition like this:

“Two different images of similar wares may represent the same product or two completely different items. Retailers want to avoid misrepresentations and other issues that could come from conflating two dissimilar products. Currently, a combination of

[1] Shopee - Price Match Guarantee - https://www.kaggle.com/c/shopee-product-matching

deep learning and traditional machine learning analyzes image and text information to compare similarity. But major differences in images, titles, and product descriptions prevent these methods from being entirely effective. In this competition, you'll apply your machine learning skills to build a model that predicts which items are the same products."[2]

Datasets and Input

The dataset[3] is provided by kaggle for the competition. Though not stated explicitly, it is assumed to have been provided by Shopee for the purpose of the competition.

It contains ~34,000 rows with an associated product image file for each row. Each row contains four columns:

- posting_id - the ID code for the posting
- image - the image id/md5sum
- image_phash - a perceptual hash of the image
- title - the product description for the posting
- label_group - ID code for all postings that map to the same product

Some observations about the dataset:

- The ~34,000 rows have only ~11,000 different label groups
- The groups contain from 2 to 34 members, with a long tail distribution as the group size increases.
- The ~34,000 rows have only ~29,000 different image phashes
- The groups of matching image phashes contain from 1 to 26 members, with a long tail distribution as the group size increases.
- The ~34,000 rows have only ~33,000 different titles
- The groups of matching titles contain from 1 to 9 members, with a long tail distribution as the group size increases.
- The images come in a wide variety of dimensions and include both square and rectangular examples. This indicates the likely necessity of either some sort of

[2] Shopee - Price Match Guarantee - https://www.kaggle.com/c/shopee-product-matching
[3] Shopee - Price Match Guarantee - https://www.kaggle.com/c/shopee-product-matching/data

size regularization before feeding into the model or selecting a model which can accommodate irregularly sized inputs.

Solution Statement

The solution will rely heavily on Convolutional Neural Networks (CNNs), as they are the go to algorithm for image processing tasks.

The targeted solution can be broken down into two steps. First, design and apply a Convolutional Neural Network (CNN) to extract a set of features for each image. Second, apply a nearest neighbors algorithm to the extracted feature sets to produce a list of match predictions. Group sizes are capped at 50 in the competition, so there will be no need to predict more than 50 matches.

Benchmark Model

The score for the lowest “Silver” entry on the kaggle leaderboard will be used as a benchmark (currently an F1 score of 0.586). It is directly relevant to the problem and provides an objective comparison, and is clearly defined and measurable.

Evaluation Metrics

Like the competition on which this project is derived, the solution will be evaluated based on its mean F1 score. The F1 score is appropriate as it takes into account both precision and recall, forcing the solution to predict matches correctly and efficiently.

Project Design

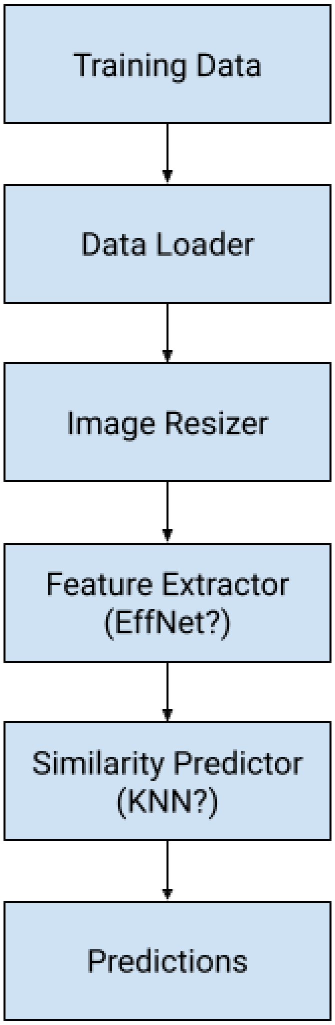
The project will proceed in three iterative steps.

First, the data will be retrieved, explored, and processed to ensure understanding and consistency of the dataset. Images will likely be regularized.

Second, the data will be split into a testing and a training dataset. Because the posts are potentially associated with each other, matching posts will likely need to be split into the same testing or training dataset.

Third, potential models will be designed, applied, evaluated, and adjusted until an acceptable result is derived. This may involve testing or combining multiple, off the shelf, pretrained models into the final solution.

The final workflow could look something like this:


Training Data
Data Loader
Image Resizer
Feature Extractor (EffNet?)
Similarity Predictor (KNN?)
Predictions