Data Science Final Project Report

Andrés Ponce
Department of Computer Science
National Cheng Kung University
Tainan, Taiwan
andresponce@ismp.csie.ncku.edu.tw

***Abstract*—Many e-commerce platforms need to search for similar or identical products given some query image. Doing so can increase the platform's ability to recommend interesting products or analyze purchasing trends across product categories. For the eBay eProduct Visual Search Challenge, participants take a set of query images and search a large index set for images of matching products. We first train a model to recognize the hierarchical structure of the different image categories. Then, we use our model's output to produce hashes of the index images and query images using locality sensitive hashing, and locate identical products by comparing these hashes. This paper describes the motviation, approach, and results of the comptetition.**

I. Introduction

E-commerce platforms continue to grow and play a large role in consumer's shopping behavior. Especially with the pandemic, more people relied on such platforms for many of their purchases [1].

Platforms where users directly sell their own products especially benefit from finding images of identical products. When a user searches for a product, he or she expects the results to contain images of the same product. On sites like eBay, identifying identical products can be very useful when aggregating sales of different listings of the same product. Not only do e-commerce platforms rely on such visual search, but also visual search engines such as Google Images, where the user can use an image as a query instead of a search term.

The eBay eProduct Visual Recognition Challenge [2] consists of finding images of the same product from a large index set of images. This challenge is one of fine-grained visual classification, since we are trying to find images of *the same* product. Similar yet non-identical products can differ only by very fine details; likewise, images of identical products can differ only in lighting conditions or other small factors, increasing the difficulty of the task. Despite the possibility for many small changes, our model should be resilient to such invariant factors and be able to distinguish similar products from each other.

II. Competition Description

Current image datasets do not focus enough on super fine-grained object detection. This prompted the authors to create the eProduct dataset focusing on fine-grained visual recognition. The dataset is divided into training, validation, and query sections.

The training set consists of around 1.3 million labeled images modelled after the ImageNet [3] dataset. Each image also comes with three levels of hierarchical labels: its meta class (16 total), level 2 class (17 total), and the leaf class (1,000 total) as well as the product title and a unique identifier. Also like ImageNet, the validation set contains 50,000 images, each containing the same information for each image as the training set.

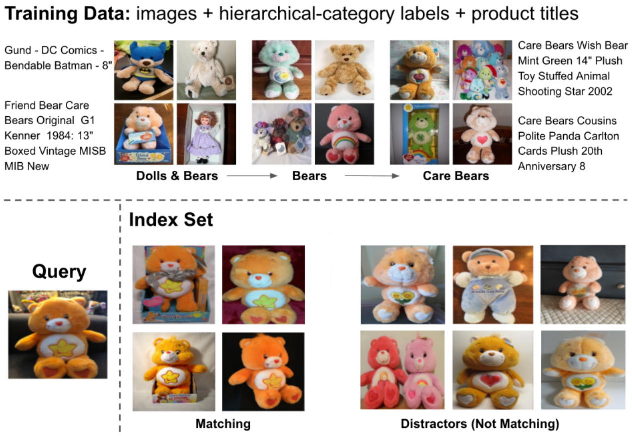
Training Data: images + hierarchical-category labels + product titles
Gund - DC Comics - Bendable Batman - 8"
Friend Bear Care Bears Original G1 Kenner 1984: 13" Boxed Vintage MISB MIB New
Care Bears Wish Bear Mint Green 14" Plush Toy Stuffed Animal Shooting Star 2002
Care Bears Cousins Polite Panda Carlton Cards Plush 20th Anniversary 8
Dolls & Bears
Bears
Care Bears
Index Set
Query
Matching
Distractors (Not Matching)

Fig. 1. eProduct dataset structure. The top section contains a sample training image and the meta, level 2, and leaf categories. Each level contains a more specific type of product, and the leaf category can still contain slightly different, non-matching products. The bottom section shows the query image and the identical products from the index set, along with a set of *distractor* products that do not match with any query image. Source: [2]

The testing set contains 10,000 query images with only the unqiue identifier. Given one of these query images, our task is to search for identical products to this query in a 1.1 million image index set, also provided as part of the testing set. The index set contains *groundtruth sets* which are a match with any of the query images and a *distractor set* which does not match with any query image. Fig. 1 describes the structure of the dataset and the challenge posed by similar products.

III. Related Work

IV. Method Description

The visual search problem can be divided into two major parts: training a deep model to obtain z and calculating the similarity between z and the items in the index set. We train a deep model for the first part and use locality sensitive hashing to find the similarities.

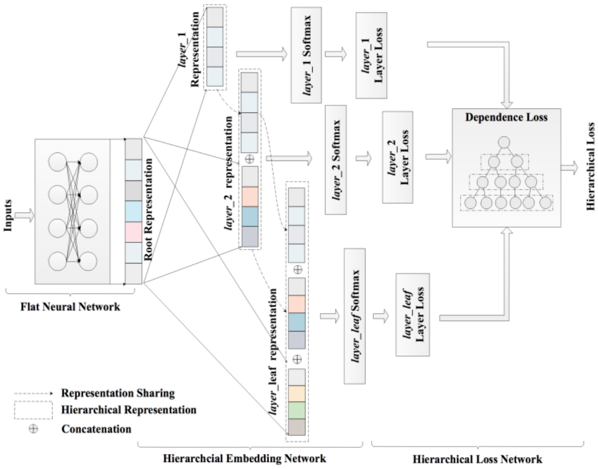

Fig. 2. Deep Hierarchical Embedding's structure . The network first generates a representation for each hierarchical layer which is the concatenation of the prevoius layer's representation and the current layer. Then we calculate the loss function for each layer's prediction and use this to calculate dependence and hierarchical loss. Source: [4]

A. *Deep Learning Model*

The idea behind using a deep model is to obtain a vector z that captures the important information of an image. The contents of the vector should be informed not only by the image contents, but also the hierarchical descriptions given as part of the training data. Since there are three hierarchical levels of information (meta class, level 2, leaf categories), our model should produce similar embeddings for products in similar categories. This means our model should incorporate this information in some way during training.

However, the categories are not independent of each other. For example, if a product belongs to a certain meta category, there is only a subset of child level 2 categories, and similar for child classes. Our model should take these parent-child relationships into account.

We based our work on Deep Hierarchical Classification [4], since this model seemed specifically designed for hierarchical e-commerce classification. The architecture contains a flat neural network(FNN) $f(\theta)$ which generates a representation for each hierarchical layer R'_l. This representation is the concatenation of the previous hierarchical layer's representation $R_l = R_{l-1} \oplus R'_l$ for $l \neq 1$. The first layer's represnetation is not concatenated with anything else.

Our model attempts to predict the three class levels for each image.

B. *Locality Sensitive Hashing*

V. Training and Experiments

References

[1] Petra Jílková and Petra Králová. "Digital consumer behaviour and ecommerce trends during the COVID-19 crisis". In: *International Advances in Economic Research* 27.1 (2021), pp. 83–85.

[2] Jiangbo Yuan, An-Ti Chiang, Wen Tang, et al. *eProduct: A Million-Scale Visual Search Benchmark to Address Product Recognition Challenges*. 2021. DOI: 10.48550/ARXIV.2107.05856. URL: https://arxiv.org/abs/2107.05856.

[3] Jia Deng, Wei Dong, Richard Socher, et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[4] Dehong Gao, Wenjing Yang, Huiling Zhou, et al. "Deep hierarchical classification for category prediction in e-commerce system". In: *arXiv preprint arXiv:2005.06692* (2020).