A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels

Marcus Klasson, Cheng Zhang, and Hedvig Kjellström

Introduction

- We present a novel dataset with natural images of grocery items.
- Dataset can be used for evaluating models on image classification.
- Includes corresponding information about each product from an online shopping website.

Motivation

- **Assistive vision devices** for visually impaired people utilizes barcodes to recognize products.
- **Problems with barcodes:**
  - Barcodes can be hard to find for a visually impaired person.
  - Some items do not have barcodes, e.g. fruits and vegetables.
- **Goal:** Assistive vision systems relying on natural image information.
- Recognizing groceries from natural images requires system to distinguish between **similar** and **misplaced** items.

Similar groceries.

Misplaced groceries.

We need proper data that illustrates these scenarios in order to achieve the goal above.

References

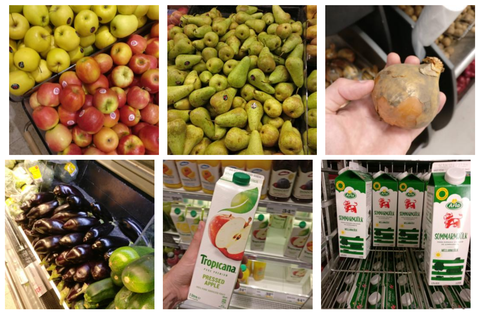
[1] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[2] N. Zhang, J. Donahue, R. B. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European Conference on Computer Vision*, 2014.

[3] W. Wang, H. Lee, and K. Livescu, "Deep variational canonical correlation analysis," *CoRR*, vol. abs/1610.03454, 2016.

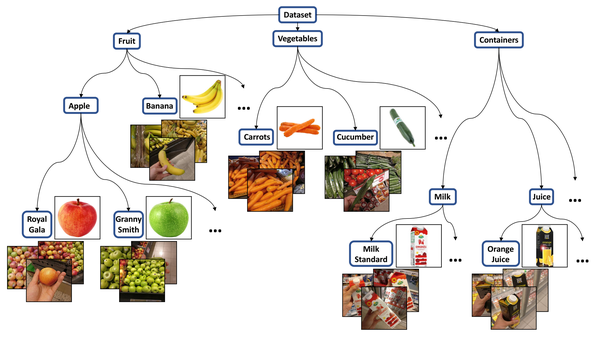
Dataset

- We collected natural images of fruits, vegetables, and carton items, such as dairy and juice products.
- All images were taken with a smartphone camera in grocery store environments.
- Consists of 5125 natural images from 81 different fine-grained classes.
- Dataset was split into a training and test set with respect to grocery store locations.



Hierarchical Structure

The 81 classes are divided into 46 coarse-grained classes to create a hierarchical structure.



Classification Results

We benchmark our dataset by using off-the-shelf CNN features [1] and fine-tuned CNNs [2].

	AlexNet	VGG16	DenseNet-169
Off-the-shelf	69.2	62.1	72.5
Fine-tuned	69.3	73.8	**84.0**

Fine-grained classification accuracy (%).

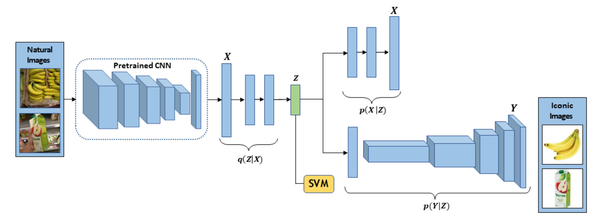
Iconic Images and Product Descriptions

Each fine-grained class has an iconic image and a product description of the grocery item.

- **Granny Smith** is a green apple with white, firm pulp and a clear acidity in the flavor.
- **Yellow onion** is a good flavoring. Fits well both cold and hot in cooking.
- **Tropicana Apple** is a ready to drink juice with pulp pressed on apples. Not from concentrate. Mildly pasteurized.
- **Bananas** are good as snacks. The **banana** is temperature sensitive, very sensitive to dehydration, ethylene and cold damage. Store in room temperature or cool, never in a refrigerator.

Utilizing Iconic Images

We propose a multi-view deep generative model, **Variational Autoencoder Canonical Correlation Analysis** (VAE-CCA), that utilizes the iconic images [3].



VAE-CCA architecture.

VAE-CCA achieves more meaningful latent representations compared to a standard VAE.

	VAE+SVM	VAE-CCA+SVM
DenseNet-169	79.1	**80.4**

Fine-grained classification accuracy (%) with SVM and a fine-tuned DenseNet-169 as feature extractor.

We translate iconic images from unseen natural images with VAE-CCA to demonstrate the quality of the representation as well as enhancing the interpretability of the method.

Natural images translated into iconic images through the iconic image decoder of VAE-CCA.

Conclusion

- **Future directions:** utilize product descriptions, explore other architectures and multi-view models, collect more data...
- **Dataset available at** `https://github.com/marcusklasson/GroceryStoreDataset`.
- **Paper available at** `https://arxiv.org/abs/1901.00711`.