Retail-786k: a Large-Scale Dataset for Visual Entity Matching

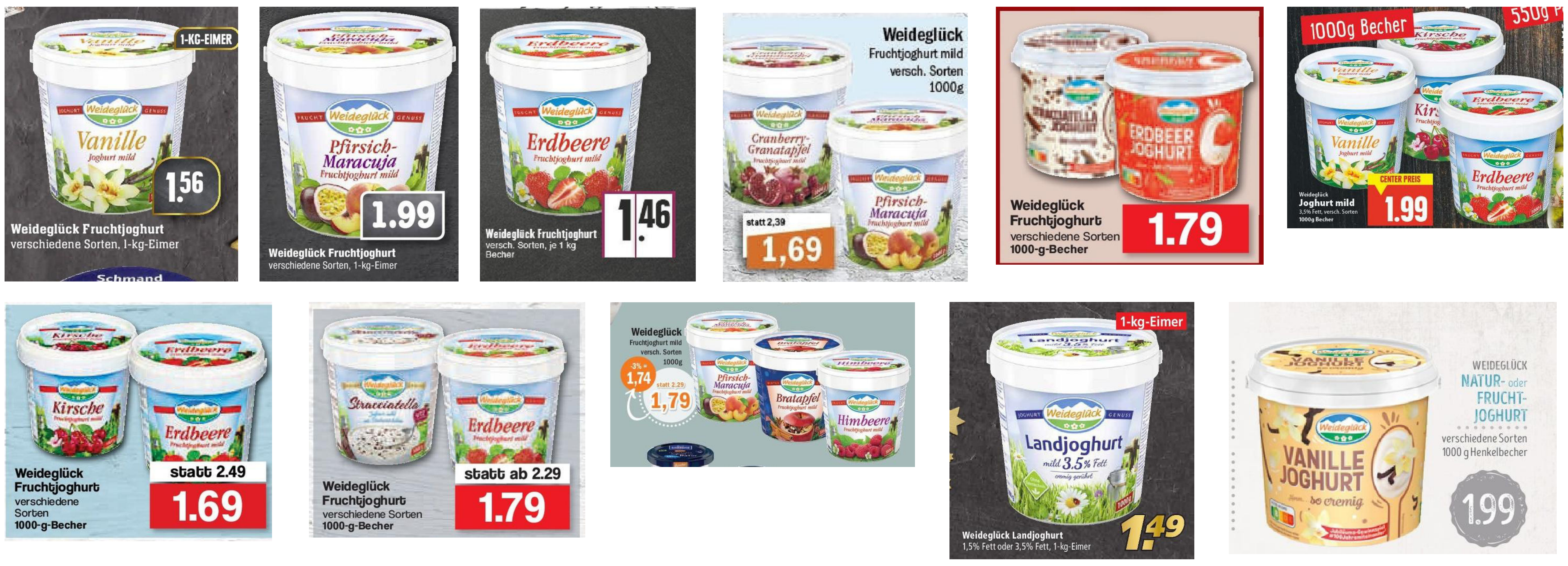
Bianca Lamm[1][2], Janis Keuper[1]

[1]IMLA, Offenburg University, [2]Markant Services International GmbH

paper GitHub dataset

Motivation

In the context of retail products, the term **"Visual Entity Matching"** refers to the task of linking individual product images from diverse sources to a semantic product grouping. On the right there are images illustrated that show different products from the same entity which is defined by the fact that single images are used as "placeholders" by retailers to promote all products of the entity.



Definitions

Products: Every retail item is uniquely identified by the internationally standardized Global Trade Item Number (GTIN).

Entities are defined as semantic groupings of products. An entity represents a group of different GTINs, i.e., contains multiple products which are used as equivalent placeholders.

Visual (Product) Entity Matching (VEM): Extending product matching by the extraction and semantic comparison of image features. The way that many product entities are defined demand algorithms to handle different similarity measures and visual intra-entity variances. In the context of the practical retail task, entities are "things that should be compared" (mostly in price). The actual task is to learn these entities by examples.

Dataset

Data Sources: Starting point was a collection of publicly available full page leaflets in JPG format provided by the company Markant Services International GmbH. Each leaflet page has been manually segmented into product information boxes containing **product image, price, description**, and additionally, logos, price tags, or quality seals. The cropped boxes form our image dataset.

Dataset Properties: The dataset is composed of **3,298 entities** and **786,179 images** in total. Each entity in the training or test set must have at least 10 and 3 samples respectively. In order to prevent data pollution, images from one retailer can only exist in the training or in the test set for each considered entity. The longer edge is fixed to 512 and 256, respectively.

Image Properties and Variance

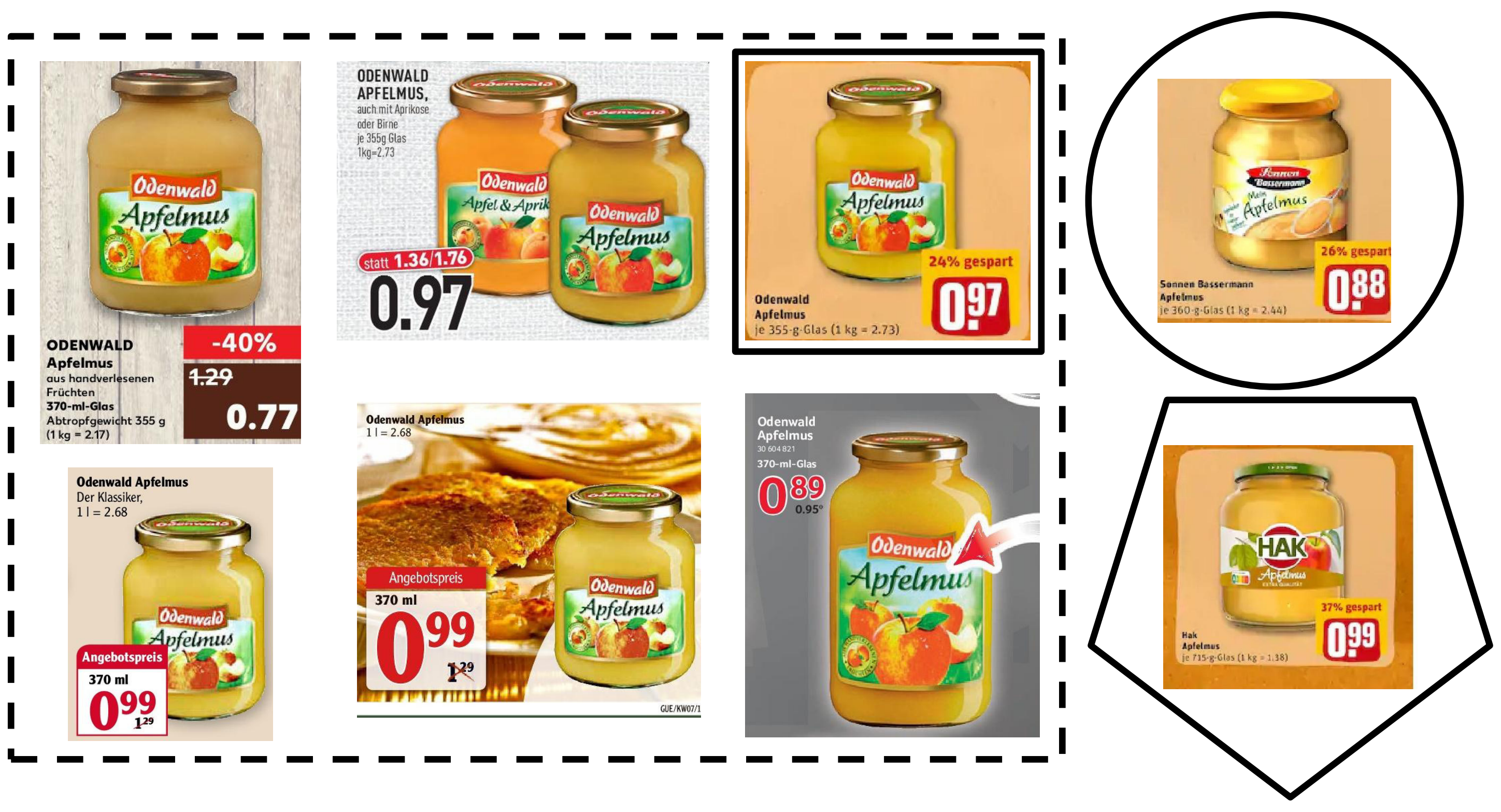
Images within an entity can be similar and diverse at the same time. Nevertheless, images from different entities can also have strong visually similarities.

High intra-entity variances:

- Each retailer has its leaflet design. Thus, the image background can be very different.
- Products of an entity differ usually in flavor. Therefore, the number and/or the choice of the product placeholders of the promotion image varied.
- The imaging perspectives, i.e., the point of view, of a product image diverse.
- The color distribution of products often varies strongly between different leaflets.

Low extra-entity variances:

An image (solid rectangle) of an entity (dashed rectangle) can have a strong similarity to images of other entities (solid circle and polygon).



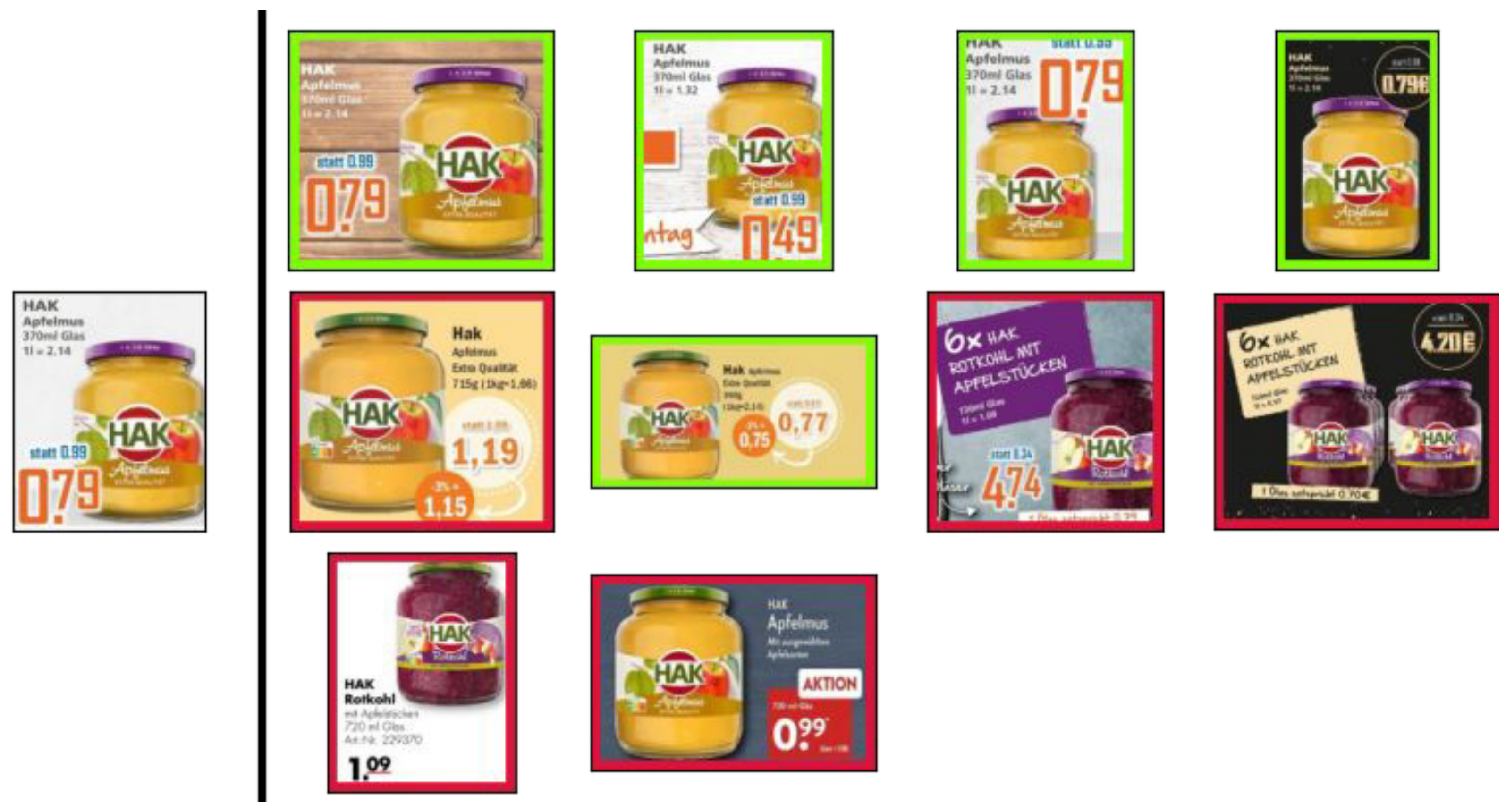
Baseline Results

VEM as Classification:

We take our dataset as basis for a static classification problem, i.e., the entities are fixed. We chose three different state-of-the-art image classification models: ResNet50, ViT, and ConvNeXt. Overall, the ConvNeXt model attains the best test set **accuracy of 0.855** and a **F1 score of 0.832**.

VEM as Image Retrieval:

For most real applications of VEM, the entities will rapidly change over time. For our baseline experiment considered VEM as image retrieval problem we used the ROADMAP approach [1]. The metric used is mAP@R [2]. On the test set of our dataset, we achieved a **mAP@R score of 72.23%**. Further, the **R@10 score is 56.34%** and the **Recall@1 score is 44.85%**.



Limitations & Outlook

Limitations:

- Manual miss-annotations can still occur despite checks.
- Missing textual data like price, discount, and product descriptions which is implicitly contained in our images are not yet extracted.
- Entity labeling: the entities in our dataset have been defined based on the very specific application of price monitoring in printed advertisements. Hence, it is not clear how good results on our dataset will transfer to other problems.

Outlook:

The extraction of textual information in the image (price, brand, weight unit) could further enhance the database toward multi-modal solutions.

References

[1] Elias Ramzi, Nicolas Thome, Clément Rambour, Nicolas Audebert, and Xavier Bitot. Robust and decomposable average precision for image retrieval. Advances in Neural Information Processing Systems, 34:23569–23581, 2021.

[2] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In European Conference on Computer Vision, pages 681–699. Springer, 2020.