



Multimodal fine-grained grocery product recognition using image and OCR text

Tobias Pettersson¹ · Maria Riveiro² · Tuwe Löfström³

Received: 23 October 2023 / Revised: 23 January 2024 / Accepted: 19 April 2024 / Published online: 7 June 2024
© The Author(s) 2024

Abstract

Automatic recognition of grocery products can be used to improve customer flow at checkouts and reduce labor costs and store losses. Product recognition is, however, a challenging task for machine learning-based solutions due to the large number of products and their variations in appearance. In this work, we tackle the challenge of fine-grained product recognition by first extracting a large dataset from a grocery store containing products that are only differentiable by subtle details. Then, we propose a multimodal product recognition approach that uses product images with extracted OCR text from packages to improve fine-grained recognition of grocery products. We evaluate several image and text models separately and then combine them using different multimodal models of varying complexities. The results show that image and textual information complement each other in multimodal models and enable a classifier with greater recognition performance than unimodal models, especially when the number of training samples is limited. Therefore, this approach is suitable for many different scenarios in which product recognition is used to further improve recognition performance. The dataset can be found at <https://github.com/Tubbias/finegrainocr>.

Keywords Grocery product recognition · Multimodal classification · Fine-grained recognition · Optical character recognition

1 Introduction

The retail industry is currently experiencing a digital transformation. To remain competitive, retailers use new machine learning based technologies that allow them to exploit real-time purchase behavior, achieve more efficient store management, and offer new shopping experiences to their customers. [1] identifies six value drivers for digital transformation that generate new revenue or improve efficiency: innovation, employee productivity, asset utilization, customer

experience, supply chain and logistics, and sustainability. Included in these value drivers are use cases that improve the efficiency of the checkout process and also reduce the number of thefts. In grocery stores, this can be achieved by recognizing products that enable automatic checkout, provide on-shelf availability analysis, and detect fraud at self-checkouts (SCO).

Automatic recognition of products in grocery stores poses many challenges to machine learning. One such challenge relates to the large number of products to recognize. A large supermarket typically has tens of thousands of unique products; for hypermarkets, this number could increase to hundreds of thousands. There is also a large imbalance in sales of products; some products are sold in large quantities each day, while others are only sold a few times a week. Moreover, new products are added and removed from the retailer each week. Another challenge relates to finding robust recognition solutions that can distinguish between products with very subtle visual differences.

In computer vision, indeed, fine-grained image recognition focuses on differentiating between hard to distinguish or similar types of objects. [2] categorizes fine-grained image recognition into three main paradigms: finding key parts in

✉ Tobias Pettersson
tobias.pettersson@itab.com

Maria Riveiro
maria.riveiro@ju.se

Tuwe Löfström
tuwe.lofstrom@ju.se

¹ ITAB Shop Products AB, University of Skövde, Jönköping University, Jönköping, Sweden

² Department of Computer Science and Informatics, Jönköping University, Jönköping, Sweden

³ Department of Computing, Jönköping University, Jönköping, Sweden

an image and merging the local feature vector with a global vector representation, learning better feature representation by high-order feature interaction or novel loss functions, or use of auxiliary data sources. Multimodal classification is part of the last paradigm, using a combination of data from different modalities to improve recognition performance. In retail, this is common in e-commerce applications [3–5], where product images combined with textual metadata are used to create more accurate models. For automatic recognition of grocery products, image-based techniques are the dominant category. However, recent progress in optical character reading (OCR) has made extracting textual elements from product packages possible. Thus, the combination of the text elements from the product packages with the respective image allows for more accurate and reliable recognition of products using multimodal classifiers. This approach has been proposed in two recent surveys (see [6, 7]) and listed as an important research direction to improve fine-grained recognition of grocery products [6, 7]. In fact, recent work has shown that using extracted textual data from packages as input to natural language processing (NLP) models results in a robust text classifier with high accuracy, see [8, 9]. Furthermore, [7] state that no specific datasets exist for fine-grained product recognition. Therefore, a dataset is needed to identify the performance and challenges prevalent for fine-grained recognition of grocery products. In addition, more research is needed on how to construct such multimodal classifiers/models and how image and text information can be fused.

Thus, in this work, we suggest a multimodal product recognition approach that uses image and textual information as input on a fine-grained product recognition dataset. Our approach uses separate image and text models and combines the features with multimodal fusion methods. We test

several deep image models, from small to larger state-of-the-art models. We also use several state-of-the-art text models based on the Transformer architecture [10] to evaluate text recognition performance. Furthermore, experiments are performed with multiple multimodal fusion models to explore the performance of the different fusion methods with respect to model size, image resolution, and text length.

A challenge with existing datasets for automated product recognition is the lack of high-resolution images that include a wide range of classes with subtle distinctions captured in all product orientations. In this regard, to the best of our knowledge, there exists no public product dataset that provides both image and package text information with products that are only distinguishable by subtle details, image samples from all product orientations, detailed text information from packages, and a significant number of classes and samples. Therefore, we collected a dataset with those characteristics from a grocery store used for all our experiments. The created dataset contains a great variety of different products, where most of the products have one or more related product(s), with product package differences consisting of small textual or color details for the related product(s). Figure 1 shows different types of challenging cases for grocery product recognition.

In summary, the main contributions of this paper are as follows:

- A novel fine-grained dataset (FineGrainOCR) for grocery product recognition consisting of image and textual data. The distinguishing marks for this dataset are:
 - Subtle differences between products
 - A wide range of product orientations
 - High-resolution image data

Fig. 1 Example of challenging cases where different grocery products have a similar appearance and are only differentiable by subtle details (ingredients side (a), meat packages (b)), lactose and non-lactose product variant (c), and the same type of product with different weight (d)



- Detailed text information from packages
- A significant number of classes and samples
- A multimodal product recognition approach that combines image and textual data from grocery products using multimodal fusion methods. This approach performs significantly better compared to unimodal models on fine-grained image recognition on a dataset collected from a grocery store. Furthermore, the approach shows state-of-the-art results in a related retail domain.
- Recommendations and trade-off aspects based on our extensive experimental evaluations related to how to implement and deploy multimodal product recognition methods.

The remainder of this paper is structured as follows. Section 2 reviews related work for grocery product recognition and multimodal classification. We then present our dataset in Sect. 3 and details of our proposed approach in Sect. 4. The experimental results are reported in Sect. 5 and a discussion of our results in Sect. 6 in the context of the wider literature. Finally, we conclude our work in Sect. 7.

2 Background

In this section, we present a summary of relevant grocery product recognition research that includes both general and fine-grained methods. To contextualize our own dataset, we provide a detailed description of the available datasets for recognition of grocery products. Furthermore, we give an introduction to multimodal classification techniques and work focusing on multimodal techniques for image and text data. Finally, examples of multimodal image and text models are provided from the retail domain.

2.1 Grocery product recognition

Recognition Techniques Recognition of products is required for many types of automation solutions at grocery stores. This includes automatic checkout systems for the registration of products [11–13], the monitoring of availability and misplacement on store shelves [14–16], frictionless checkout where a camera system with the inclusion of other sensors in a store registers the pick of products by customers [17, 18] and detection of barcode switches and other fraudulent actions at SCOs [8, 19]. A figure showing examples of applications can be seen in Fig. 2.

Early work used local feature-based methods to recognize products, see examples in [20, 21]. In particular, [20] used scale-invariant feature transform (SIFT) [22] keypoints, color histograms, and boosted Haar-like features to recognize products.

In recent years, many methods have used convolutional neural networks (CNN) to improve the recognition of grocery products. These include methods to improve the recognition performance of new models [23–26], improved learning strategies [27–30], use synthetic data for larger training sets [31, 32], and adapt grocery models to new environments [33, 34]. To exemplify the breadth and variety of these solutions, we present details of some of these relevant works below. In [27], the problem of new grocery products introduced in a store is addressed. An object classification system is proposed that incorporates the rejection of unknown products in a checkout counter scenario. A metric learning approach [35] enables more reliable confidence using cosine similarity as a distance measure. In addition, a novel margined unknown loss function further improves the rejection of unknown products. The handling of new grocery products for a product recognition system is proposed by [33]. The method learns the embedding of products using a generative adversarial network (GAN). Furthermore, leveraging the learned embedding, the method can add new classes without any retraining. This is achieved by formulating the task of product recognition as an instance-level recognition problem and applying the k-nearest neighbors algorithm (KNN) to recognize products.

Because some grocery products have only small fine-grained details that differentiate them, several fine-grained methods have been proposed to tackle this challenge in the literature. Such solutions consist of several steps and use complex models; researchers use hierarchical information [36], techniques to find discriminative features [37–40], and others use contextual information [41]. However, the methods for finding discriminating features vary a lot. For example, [37] presents a method to improve the fine-grained recognition of products using destruction and construction learning proposed in [42] with a self-attention mechanism. The model learns to correlate semantic regions in product images and is also able to find specific discriminatory areas. [39] propose a fine-grained classification network where an input image is passed through two classification paths, using object-level and part-level information. The part-level information contains discriminatory regions that are passed through stacked convolutional LSTM before being combined with features from object-level information before a classification layer. [40] propose a framework for product label recognition using textual information. However, the results show challenges associated with reading OCR and correctly classifying product labels. For a more general overview of fine-grained recognition techniques, see [2, 43].

Multimodal techniques, which combine image and text for fine-grained grocery product recognition, have also been proposed in recent years; see, for example, [44, 45]. In [44], the recognition of soda bottles is carried out by combining textual and visual features with a simple attention mechanism.

Fig. 2 Example of applications areas for product recognition in grocery stores; **a** autonomous checkout systems, **b** shelf monitoring and **c** detection of barcode switch at SCOs



The product text is extracted from each input image by OCR reading. [45] also use OCR to combine textual and visual features. Their approach adopts supervised contrastive learning and achieves a significant improvement in the accuracy of several benchmarks for the recognition of grocery products. However, these techniques are applied to front-facing grocery products with a limited number of classes. Furthermore, the datasets also contain a limited number of text elements, and thereby missing classes with challenging fine-grained textual details. Thus, based on the limited research using image and OCR for grocery product recognition and the research gap identified by [6, 7], we try to bridge such gap by proposing a multimodal product recognition approach. Moreover, we create a fine-grained dataset with a large number of classes, and perform an extensive evaluation that identifies the trade-offs using our multimodal product recognition approach.

Datasets Early work by [20] introduced the first open dataset for the recognition of grocery products. It consists of 120 different types of products, where training data contains product images from a studio environment with only a handful of images for each class. A large test set is also included by extracting images captured by a camera from a real store environment. [21] use the same setup but increase

the number of product classes to 8,350. In [46], this is further extended by adding text descriptions of each product in the dataset, enabling the inclusion of semantic information in model development. [47] use both training and test data from challenging in-store conditions such as different perspectives and lighting. The dataset is aimed at autonomous robots. It consists of 25 classes with 5,000 image samples in total, where each class describes a coarse class, such as, for example, flour, milk, and pasta.

The RPC dataset is aimed at automatic checkout systems [48]. It includes 200 different types of products, with 53,739 single-product images from a studio setup and 30,000 multi-product images from a checkout counter setup. The authors also propose the single-product images to be used for constructing synthetic images by cutting the products from the single images and pasting them into the checkout counter environment. In a similar setting, [49] provide the D2S dataset consisting of semantic segmentation masks for products.

Object detection of products in shelf images poses a significant challenge since there are usually a large number of products per image positioned close to each other or overlapping one another. In the SKU-110K dataset [50], a dense

object detection dataset is proposed for grocery products, which includes 11,762 shelf images, where each image contains around 147 products. The products in the dataset are unlabeled products, hence recognition of specific classes is not possible. In [51], the SKU-110K is expanded by adding multiple rotated instances for each sample. In contrast to the unlabeled products in SKU-110K, the RP2K data set [52] consists of 2,000 products with 500k images of the shelves collected from physical retail stores. The dataset is organized hierarchically with meta-categories for shape and product type. Additionally, details such as brand, flavor, and type are also included, allowing the ability to evaluate on a customized categorization level. In the Locount dataset provided by [53], an instance count is added to each product at each shelf position. This enables evaluations of quantity estimations at each shelf position.

In recent years, datasets have been presented for grocery products that use text as an additional input. [54] present a dataset designed to help people with visual impairment, including textual descriptions acquired from OCR reading of images as metadata. The dataset contains a candidate set that contains 6348 products, with a total of 13,290 images. To evaluate product recognition performance, a test set with 373 samples from 104 randomly selected classes is extracted from a real-world environment using a smartphone camera. United Retail Datasets (Unitail) [55] consist of a product detection and a product recognition dataset. The product recognition dataset uses 1,454 products for the training set, where each product has a front-facing product sample. Manually annotated text regions for each sample allow the dataset to support the tasks of text detection, text recognition, and product matching that utilize text information.

However, despite the large number of aforementioned datasets for grocery product recognition, there is no dataset that addresses the challenges of fine-grained grocery product recognition with the following characteristics:

- Containing challenging products with similar appearance
- Sufficient image resolution to capture fine-grained details
- Product facing in all directions
- Large number of classes
- A substantial amount of training data

2.2 Multimodal classification

By combining several data sources, such as images, video, text, and audio, a better representation of a subject is possible. Multimodal classification is used to fuse the representations from each unimodal data source and to learn a classifier with improved performance. Multimodal data are present in many different types of domains, such as retail [3, 56], agriculture [57, 58], and self-driving cars [59, 60].

[61] describe five challenges for a multimodal setting: learning to construct a representation that uses information from all domains, how to translate from one modality to another, how to align the information, how to fuse the information and finally, how to transfer knowledge between modalities. The description and naming of multimodal fusion methods are inconsistent in the current literature. [62] propose a holistic taxonomy of multimodal classification that we are applying in this work. The taxonomy categorizes the fusion methods into three different types: early fusion, late fusion, and cross-modality fusion. In early fusion, data are aggregated before the learning model is applied. Late fusion combines the extracted features of each modality, which includes the probabilities from each class or the final feature representation from each modality. In cross-modality fusion, data and intermediate features are shared during learning, which can model complex relationships between modalities.

Several methods have been proposed to fuse image and textual features. Encoded textual features to the image domain and passing the resulting image through a CNN have been shown to improve classification performance [63–65]. Furthermore, transformer-based models have been successfully combined for multimodal classification [66]. [67] evaluate different fusion methods using textual features of FastText [68] and image features of a ResNet model [69]. [70] use image and text information for scene understanding. Text from images is extracted via OCR and a pyramidal histogram of characters (PHOC) is used to calculate a Fisher vector that encodes the textual information. The textual features are then fused with the image features of a CNN using an attention mechanism. Examples of multimodal fusion work for the recognition of grocery products are described in Sect. 2.1.

Within the retail industry, e-commerce is another application area where image and text information is often found. Zahavy et al. [71] recognize products using the image of the product and its corresponding title. Two CNNs are used, one for each modality, to extract feature representations. Fusion is performed by training a policy network that learns the preferred modality for classification. [72] learn a simple fusion model composed of a fully connected layer that assigns a weight to each modality. To reduce the risk of neglecting one modality during training, a regularization scheme is proposed using a Kullback-Leibler divergence loss in combination with a cross-entropy loss. Other work uses early fusion methods by concatenating the final feature layers of the unimodal decision results [73, 74]. In [75] and [76], the authors improve the classification results using an attention mechanism that combines the image and textual features. Combining image and text from OCR is restricted not only to grocery product recognition. It has been applied to other domains such as document classification [77], package identification [78] in logistics, and product leaflet classification [79].



Fig. 3 Autonomous checkout system used for data collection. The scanning tunnel where image data is collected is marked with a red rectangle

3 FineGrainOCR—fine grained product dataset with OCR

There are several datasets for the recognition of grocery products, as described in Sect. 2. However, they contain either a limited number of classes with similar appearance, have too low image resolution to capture small distinguishable details or use only front-facing products. Therefore, we created a fine-grained multimodal product recognition dataset. We established that such a dataset would have to fulfill the following requirements:

- Contain high-resolution images
- Present different sides and orientations on packages
- Include a large number of classes where most of them have one or several classes to which they have a strong resemblance.

An autonomous checkout system deployed in a large grocery store has collected images of grocery products for this new dataset. In the system, the products are placed on a moving conveyor belt and then pass through a scanning tunnel. Figure 3 shows the system with the scanning tunnel marked with a red rectangle. Product registration is carried out using barcode scanners placed within the scanning tunnel. Customers in the store have performed the scanning process, and no instructions have been provided on how to place the products, resulting in a wide variety of product orientations. An RGB camera is placed in the scanning tunnel with a resolution of 2592×1944 . The high resolution and close placement of the RGB camera to the products ensure that fine-grained details have been captured.

Our FineGrainOCR dataset has been constructed by extracting a total of 256 classes from the collected data. Classes have been selected from six different categories, namely, chocolate, dairy, meat, milk/cream, mushroom, and toppings. In each category, the major part of all classes have one or several classes that they have a similar appearance to. Figure 4 illustrates some of the categories (rows) and a few examples of different classes in each (columns).

To assess the overall performance of our proposed approach, we also want to investigate how sample size affects product recognition performance. Therefore, a substantial number of samples is extracted for each class, with a target of 500 samples for each. However, the sales of products in grocery stores are highly imbalanced, making it difficult to extract a balanced dataset. Figure 5 illustrates the number of training samples for each class. Although some classes have a limited number of samples, most classes have a number of samples similar to or close to the target value of 500 samples. FineGrainOCR is divided into a training and validation set, where 80% samples are in the training set and 20% of the samples in the validation set for each class.

The product texts of each image sample are extracted using the OCR engine in Google Vision API,¹ which captures many fine-grained words from product images. Figure 6 illustrates parts of the words extracted from a sample in the dataset.

All extracted words from an image sample are concatenated into a single text string, where each word is separated by a newline character. Words are ordered using the default output from the Google Vision API. The text provided to the text models needs to be transformed into a feature vector. This is done by first tokenizing the text and then transforming it into a text embedding.

The transformer-based text models used in our approach expect a fixed number of tokens to be input into the model, with a maximum number of 512 tokens. However, the number of tokens quadratically affects the runtime performance of the text model for the BERT architecture [80], which requires us to use as few tokens as possible without significantly affecting the performance of text recognition. Figure 7 shows a histogram of the length of the tokenized text for the samples in the dataset. Based on this, the maximum number of tokens is selected as 128, 256, and 384 in the experiments. Text samples containing longer sentences are cropped.

4 Proposed approach

As explained earlier, there are many products in a grocery store that have minimal differences and are therefore difficult to distinguish visually. In many of these cases, it can consist of small textual differences (products with different fat content) or small patches with additional information (products without lactose). Figure 8 illustrates two different types of products where small text details "medium ground beef" and "extra lean ground beef" make up the entire difference. In addition, different sides of products can be more or less difficult to classify. For example, the side of the package with ingredients is difficult to differentiate, whereas the front side of a package is typically easier. To handle these chal-

¹ <https://cloud.google.com/vision>.

Fig. 4 Gallery of image samples in the new dataset created for this work. Rows contain example categories of products while columns depict examples for each category



lenges, we suggest using a simple auxiliary input, such as text extracted from OCR on products, to improve the recognition accuracy, thus creating a recognition pipeline that combines both images and text.

The details of our approach are explained in the following sections. Section 4.1 presents an overview of the image and text classification pipeline. In Sect. 4.2, we motivate and describe the selected image and text models. In addition, we describe how image and text representations are fused into different multimodal models.

4.1 Overview

The architecture of our approach is shown in Fig. 9. The input image is passed to two separate models, one image-based model and one text-based model. Textual input is extracted using OCR on the input image.² The output of each model

² The OCR extraction has been done as part of the dataset generation, thus the FineGrainOCR dataset is equipped with both image and

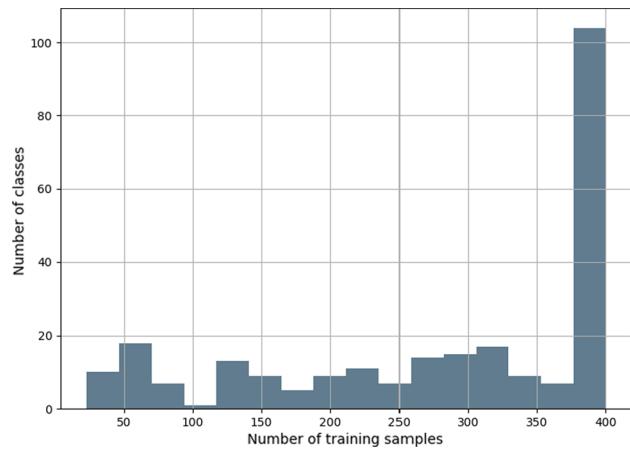


Fig. 5 Histogram of the number of samples for the classes in the Fine-GrainOCR training set

extracted text. It is included here to clarify that it is a separate step necessary to perform prior to feeding the data to the text model.

Fig. 6 Words are recognized by the OCR engine and merged into a single text string. The image exemplifies how a subpart of the recognized words from the OCR from the image are concatenated into a single string

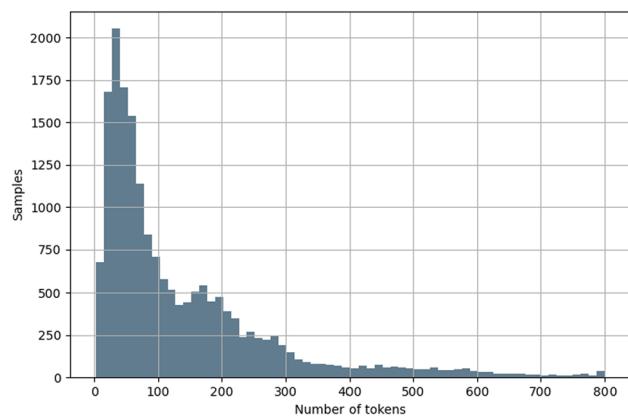
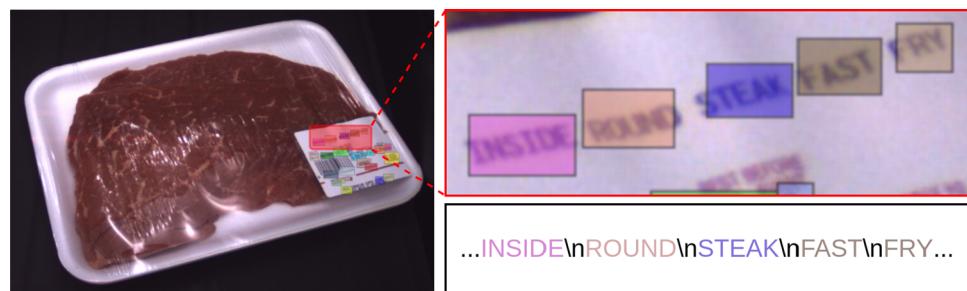


Fig. 7 Histogram of the number of tokens of the text samples in the training dataset

is a feature vector, labeled x_{txt} for the text model and x_{img} for the image model. These feature vectors are passed to a multimodal feature fusion module that combines the feature vectors of each modality and generates a fused representation called x_{fused} . This representation is passed through a classification layer that outputs the class scores y_{score} .

Fig. 8 Example of two different types of packaged meat. While not containing many visual differences, the extracted text patches (red box) in (a) and (b) show that the texts (*medium ground beef* and *extra lean ground beef*) can easily be used for discriminating between the two products



4.2 Model selection

We consider three different types of CNN architectures for the image models: ResNet [69], MobileNetV3 [81], and ConvNext [82]. Each of these has different characteristics, such as being a baseline approach, designed for mobile devices, and a state-of-the-art model. We want to see how these different characteristics affect performance in a multimodal setting. ResNet [69] is the de facto standard as the CNN backbone for many computer vision tasks. Published in 2015, it still offers good performance with recent training procedures [83], while also having a low computational cost. MobileNetV3 [81] is a lightweight CNN designed for mobile devices with a low memory footprint and a low latency requirement. The CNN architecture ConvNext [82] has shown state-of-the-art results in several image recognition tasks. The architecture is based on the ResNet50 design with several techniques and design elements that resemble the recent Vision Transformer architecture [84].

For the recognition of the product package text in grocery products, BERT [85] with a small classification head [9] has shown notably better classification results compared to other methods using GloVe embeddings [86]. Therefore, we use three types of text models based on the Transformer architecture used in BERT; the baseline model BERT [85],

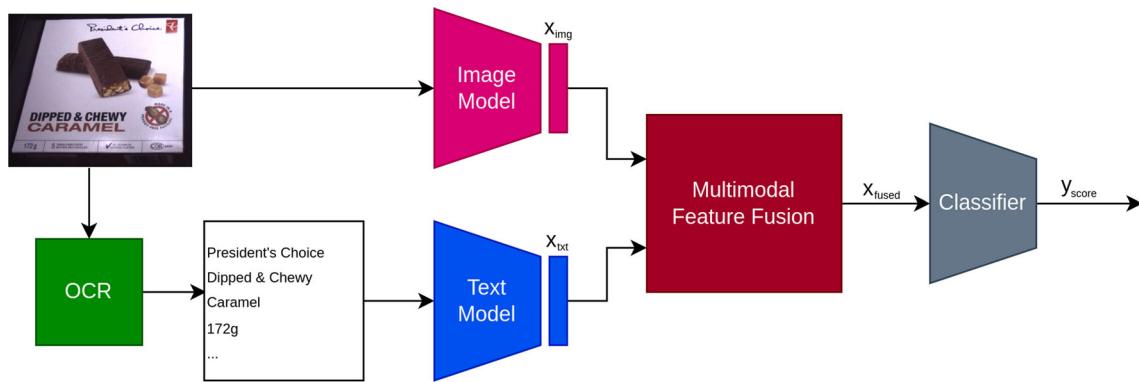
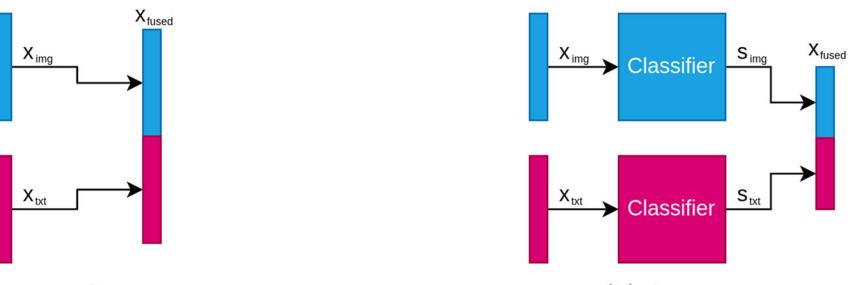


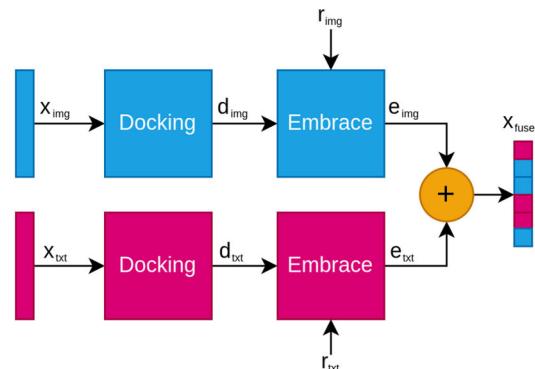
Fig. 9 The proposed multimodal architecture of our approach for grocery product recognition. A two-stream model uses an image and text model to create feature representations x_{img} and x_{txt} for a multimodal fusion module, which produces a new representation x_{fused} and passes it to a classifier

Fig. 10 Multimodal fusion methods used in the proposed approach. Each technique use the features x_{img} and x_{txt} from both modalities, and process it to a final feature representation x_{fused}

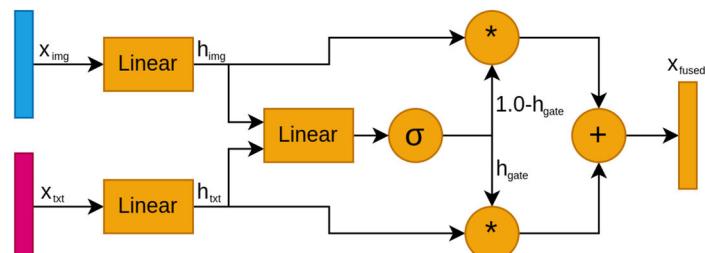


(a) Feature Concatenation

(b) Score Fusion



(c) EmbraceNet



(d) Gated Multimodal Unit

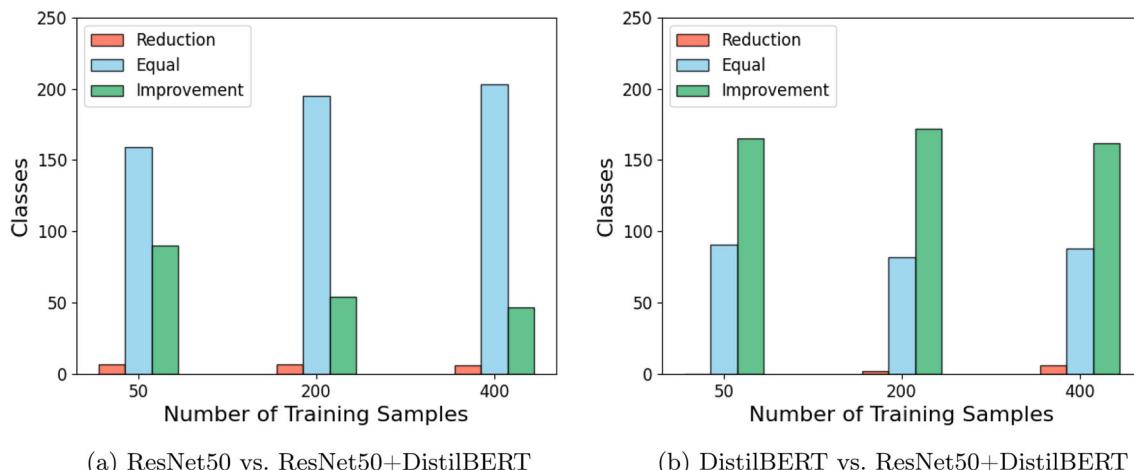


Fig. 11 Histogram of how the per class accuracy is affected by the multimodal model DistilBERT compared to the **a** image model ResNet50 and **b** text model DistilBERT for different number of maximum training samples. A reduction means that the multimodal accuracy for a

class is significantly lower compared to the unimodal model, whereas equal means that they are the same, and improvement means that the multimodal accuracy is higher

the optimized DistilBERT [87] model, and the more accurate DeBERTa [88] model.

Four types of late fusion methods are used to combine the image and text modality. *Feature Concatenation* and *Score Fusion* are two simple and classic methods to combine multiple modalities. Two recent multimodal methods are also used; the *Gated Multimodal Unit (GMU)* [89] learns the influence of each modality and gates the most informative features. *EmbraceNet* [90] is designed for robustness and learns cross-modal correlations by randomly selecting a subset of the features of each modality during model training. Each multimodal fusion method is shown in Fig. 10.

As shown in Fig. 9, the features extracted from the image and text modality are merged into a multimodal feature fusion module; x_{img} denotes the features of the image modality, while x_{txt} is used for the text modality. The resulting representation of the features of the multimodal feature fusion module is denoted as x_{fused} .

Feature Concatenation is a simple feature fusion methods based on concatenating the features of each modality into a single feature vector. It can be expressed as follows:

$$x_{fused} = [x_{txt}, x_{img}]$$

where $[\cdot, \cdot]$ is the concatenation operation.

In *Score Fusion*, the classification results of each modality are used to train a new classifier by concatenating the probabilities of each class. This fusion method has the following form:

$$s_{txt} = \text{Softmax}(W_{txt} \cdot x_{txt} + b_{txt})$$

$$s_{img} = \text{Softmax}(W_{img} \cdot x_{img} + b_{img})$$

$$x_{fused} = [s_{txt}, s_{img}]$$

where s_{txt} and s_{img} are the probabilities of each classifier, W_{txt} and W_{img} are trainable weight matrices, while b_{txt} and b_{img} are trainable bias parameters. Compared to other multimodal fusion methods, a drawback of this method is that it is not end-to-end trainable and requires each modality to be trained separately.

Introduced by Arevalo et al. [89], *Gated Multimodal Unit (GMU)* learns how each modality influences activations. Formally, it can be expressed as:

$$\begin{aligned} h_{txt} &= \tanh(W_{txt} \cdot x_{txt}) \\ h_{img} &= \tanh(W_{img} \cdot x_{img}) \\ h_{gate} &= \text{Sigmoid}(W_{fused} \cdot [x_{txt}, x_{img}]) \\ x_{fused} &= h_{gate} * h_{txt} + (1 - h_{gate}) * h_{img} \end{aligned}$$

where h_{txt} and h_{img} are intermediate feature representations and h_{gate} is a scalar and weighs the contribution from each modality. W_{txt} , W_{img} and W_{fused} are trainable weight matrices.

EmbraceNet [90] is a feature fusion method designed to ensure robustness against loss of data or modalities. It consists of a docking layer that transforms each input modality into the same feature dimension. Then an embracement layer combines the features in a probabilistic fashion. The operations can be expressed as:

$$\begin{aligned} d_{txt} &= W_{txt} \cdot x_{txt} + b_{txt} \\ d_{img} &= W_{img} \cdot x_{img} + b_{img} \\ e_{txt} &= r_{txt} \odot d_{txt} \\ e_{img} &= r_{img} \odot d_{img} \\ x_{fused} &= e_{txt} + e_{img} \end{aligned}$$

where d_{txt} , d_{img} , e_{txt} , and e_{img} are the feature representations after the docking and embracement layer. W_{txt} and W_{img} are trainable weight matrices of the docking layer, and b_{txt} and b_{img} the trainable bias parameters. r_{img} and r_{txt} are influence vectors that match the size of d_{txt} , d_{img} . r_{txt} and r_{img} are vectors of the same size as the docking layers. They are generated jointly by a multinomial distribution function, where each feature value for r_{txt} and r_{img} is set to 1 or 0 for each modality. These vectors are then multiplied by the output from each docking layer, which gives the embracement vectors e_{txt} and e_{img} . This mechanism enhances the contribution from both modalities while also providing additional robustness to a modality, for example, if it is not possible to read OCR from a product image.

The resulting features x_{fused} of each feature fusion method are then combined and passed through a classifier that outputs the resulting class scores.

We also considered using the CLIP model [91] as a standalone multimodal model; however, we discarded it due to its reported poor performance in fine-grained recognition tasks compared to task-specific models.

5 Experiments

5.1 Experimental settings

We perform our experiments with a varying number of training samples of sizes {50, 100, 200, 300, 400}. Training samples are extracted by timestamp (time of registration), ensuring that the same number of samples is always selected for each size. Classes that do not have the same number of training samples as the target number use the maximum number of training samples available. The same validation set is used throughout all experiments, using the validation split consisting of 20% of all samples for each class. The purpose of evaluating the performance of different models is two-fold. First, we want to see how the performance of each modality is affected by the training sample size. Second, we want to analyze the effect of training samples on unimodal models compared to multimodal models.

In grocery stores, the computational requirements depend on the type of application. From in-store shelf monitoring systems that analyze images in minute intervals, while automatic checkout typically needs a response in just a few hundred milliseconds. Therefore, we experiment with different model sizes, image resolutions, and text lengths to find the trade-offs of different model selections and hyperparameters. Our selected image models ResNet, MobileNetV3 and ConvNext evaluated in this work use the PyTorch implementation³ with its pre-trained weights. For text models,

we use the implementations provided by HuggingFace.⁴ For the BERT, DistilBERT and DeBERTa models, we use the pre-trained weights *bert-uncased*, *distilbert-uncased* and *deberta-base* respectively. The default model configurations are used for all text models. Image augmentations are performed when training image and multimodal models. We use a data augmentation pipeline consisting of random vertical/horizontal flipping and random rotation, each with a probability of 50%. When training the text models, no pre-processing of the textual data is performed.

We train our CNN-based image classifiers using stochastic gradient descent (SGD). The learning rate is set at $1e^{-3}$ and the weight decay is $1e^{-3}$ with a step size of 10. Text and multimodal models use the ADAM optimizer with a learning rate $2e^{-5}$ as suggested in [92] and a weight decay of $1e^{-4}$. The standard batch size is set to 32 for the models. For some larger models, a batch size of 16 has been used due to computational limitations. Cross-entropy is used to calculate the loss for all experiments.

5.2 Image classification results

We evaluated the image classification performance using our FineGrainOCR dataset. Three different model architectures are considered: ResNet, MobileNet-V3, and ConvNext. For each model architecture, two versions of the architectures are evaluated to validate whether a larger model size improves the results. The images provided to each image model are downsampled before being passed to the image model. The text extracted from the OCR reading uses the original resolution of the images in the comparison. To see whether this affects the results, the following image sizes are considered 256, 384, and 512; images are quadratic, i.e., 256×256 and so on. The results of product recognition using FineGrainOCR with only the image modality are summarized in Table 1.

As we can see in Table 1, the accuracy of the ResNet50 model is the highest for all image and sample sizes. Despite a significantly larger model size, the ConvNext architecture results are consistently 0.5–1.0 percentage points worse than ResNet50, with only a few combinations where performance is similar. Overall, we see that increasing the size of the model for each of the model architectures yields improved accuracy. The size of the image also affects the accuracy of the model, improving it by one percentage point in most cases. Furthermore, the accuracy improves steadily when more training samples are used and starts to saturate after 200 training samples. It is noteworthy that the ResNet models outperform the ConvNext models in terms of accuracy. A potential reason for this is that the more complex ConvNext needs more samples and classes to effectively learn and generalize.

³ <https://pytorch.org/vision/stable/models.html>.

⁴ <https://huggingface.co/>.

Table 1 Accuracy (in percentage) from image classification on our FineGrainOCR using the image sizes {256, 384, 512} and max training samples for each class {50, 100, 200, 300, 400}.

| Max #Samples Image Size Model | 50 | | 100 | | | 200 | | | 300 | | | 400 | | | |
|-------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | 256 | 384 | 512 | 256 | 384 | 512 | 256 | 384 | 512 | 256 | 384 | 512 | 256 | 384 | 512 |
| MobileNetV3-Small | 86.5 | 87.1 | 86.5 | 89.5 | 90.7 | 90.3 | 92.1 | 92.5 | 92.9 | 92.6 | 93.0 | 93.2 | 92.8 | 93.6 | 93.7 |
| MobileNetV3-Large | 88.2 | 89.8 | 90.2 | 92.0 | 93.3 | 93.7 | 93.5 | 94.9 | 95.2 | 94.2 | 95.3 | 95.7 | 94.4 | 95.4 | 95.8 |
| Resnet18 | 89.1 | 88.7 | 88.8 | 92.2 | 91.8 | 91.7 | 94.4 | 94.7 | 94.2 | 95.2 | 95.7 | 95.6 | 95.2 | 96.0 | 96.1 |
| Resnet50 | 92.0 | 90.9 | 93.0 | 95.1 | 94.7 | 96.5 | 95.9 | 96.6 | 96.4 | 96.3 | 96.8 | 97.2 | 96.6 | 96.9 | 97.1 |
| ConvNext-Tiny | 91.6 | 89.5 | 92.1 | 94.0 | 94.4 | 94.5 | 95.0 | 95.6 | 95.9 | 95.7 | 95.9 | 96.1 | 95.5 | 95.9 | 96.4 |
| ConvNext-Large | 90.1 | 90.8 | 93.0 | 93.0 | 94.3 | 95.4 | 94.9 | 95.5 | 96.4 | 95.6 | 95.6 | 96.5 | 95.5 | 95.8 | 96.3 |

The best results from each image size and max sample combination are marked in bold. Furthermore, we italic the best result for each max sample combination

Exploratory experiments with different data augmentation techniques and hyperparameters selections have been tested with these architectures. However, no deviations from the results presented in Table 1 were observed. We also performed experiments with two vision transformer networks, namely, ViT [84] and Swin [93]. However, under identical experimental conditions, these models achieved an accuracy that was 1.0 – 2.0 percentage points lower than the results observed in the ResNet and ConvNext models.

5.3 Text classification results

We evaluate the accuracy of the text classification module using the BERT, DistilBERT, and DeBERTa models. For each model, we test the following maximum sizes of the sequence length: 128, 256, and 384, as described in Sect. 3. The results of the text classification can be seen in Table 2.

From the results, it can be observed that a maximum sequence length of 128 gives an accuracy that is typically 2–3 percentage points lower than with a maximum sequence length of 256 or 384 for most models. Furthermore, an increasing number of training samples improve the accuracy of text classification. However, the gain is drastically reduced after 200 training samples. BERT shows the best overall performance for sequence length 128, and DeBERTa for sequences 256 and 384. However, the difference compared to DistilBERT is just a few sub-percentage points.

5.4 Multimodal fusion results

To evaluate multimodal recognition performance, we use insights from the image and text classification results presented in Sects. 5.2 and 5.3. We observe that ResNet50 is the model with the highest classification performance, while the lowest performing image classifier is MobileNetV3-Small. We select those for further evaluation to investigate the effect of the image classifier selection.

Image size is a factor that affects image recognition performance. Therefore, we select two image size values, 256 and 512, for further evaluation. All text models have similar performance when compared against each other on the same sequence length and sample size. Therefore, we select all of them to see how each complements the image model. We choose a sequence length of 256, which performs significantly better than a sequence length of 128. It also performs similarly to that of a sequence length of 384, while requiring less computational power. The sampling size is also evaluated for the multimodal models to see how each modality complements the other and how it evolves when the sampling size increases.

Analysis of potential multimodal performance To see how this potential multimodal classifier is affected by all these parameters, an oracle classifier is applied. The oracle classifier always selects the best prediction result from either the image or text recognition models, and the result can be seen as the potential performance of an optimal classifier. The result in parentheses is the gain compared to the best unimodal classifier with the same sample size and image size. We evaluate different combinations of the selected image and text models, the image size, and the sampling size. The results of the oracle classifier can be seen in Table 3.

The results of the oracle evaluation show the potential of a multimodal classifier to improve the classification performance. In addition, the number of training samples does not have a large effect on the accuracy of the oracle classification. The performance difference between 50 and 400 training samples is between 1.3 and 2.6 percentage points depending on the multimodal configurations. Furthermore, the increased image size affects only MobileNetV3-Small and ResNet50 with a few subpercentage points, and not always in a positive direction. There is no deviation from using different text models, indicating that they are learning to distinguish similar classes.

Multimodal classification Based on the results of the oracle classification, we continue to use the MobileNetV3-Small

Table 2 Accuracy (in percentage) from text classification on our FineGrainOCR with sequence lengths {128, 256, 384} and max training samples for each class {50, 100, 200, 300, 400}

| Max #Samples Max Seq. Length Model | 50 | | | 100 | | | 200 | | | 300 | | | 400 | | |
|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | 128 | 256 | 384 | 128 | 256 | 384 | 128 | 256 | 384 | 128 | 256 | 384 | 128 | 256 | 384 |
| BERT | 81.4 | 85.3 | 85.1 | 85.8 | 87.6 | 88.6 | 87.5 | 89.4 | 89.8 | 88.1 | 90.2 | 90.3 | 88.5 | 90.4 | 90.7 |
| DistilBERT | 82.1 | 84.2 | 84.9 | 85.3 | 87.4 | 88.3 | 87.3 | 89.1 | 89.5 | 87.7 | 90.1 | 90.4 | 88.2 | 90.1 | 90.6 |
| DeBERTa | 82.3 | 82.7 | 85.6 | 85.5 | 87.6 | 87.9 | 87.2 | 90.0 | 90.3 | 88.0 | 90.4 | 90.9 | 88.4 | 90.7 | 90.9 |

The best results from each sequence length and max sample combination are highlighted in bold. In addition, the best max sample combination is italic

Table 3 Ideal oracle classification accuracy (in percentage) for different combinations of the image models MobileNetV3-Small and ResNet50, and the text models BERT, DistilBERT and DeBERTa

| Image model | Text model | Image Size | Max #Samples | | |
|-------------------|-------------------|------------|--------------|-------------|-------------|
| | | | 50 | 200 | 400 |
| MobileNetV3-Small | BERT | 256 | 96.3 (+9.8) | 97.9 (+5.8) | 98.1 (+5.2) |
| | DistilBERT | | 96.0 (+9.5) | 97.9 (+5.8) | 98.1 (+5.2) |
| | DeBERTa | | 95.8 (+9.3) | 97.8 (+5.7) | 98.1 (+5.2) |
| | ResNet50 | BERT | 97.2 (+5.2) | 98.5 (+2.6) | 98.6 (+2.0) |
| | | DistilBERT | 96.8 (+4.8) | 98.4 (+2.5) | 98.6 (+2.0) |
| | | DeBERTa | 96.8 (+4.8) | 98.3 (+2.4) | 98.5 (+1.9) |
| | MobileNetV3-Small | BERT | 96.2 (+9.7) | 98.1 (+5.2) | 98.1 (+4.4) |
| | | DistilBERT | 95.9 (+9.4) | 98.1 (+5.2) | 98.1 (+4.4) |
| | | DeBERTa | 95.6 (+9.1) | 98.0 (+5.1) | 98.2 (+4.5) |
| | ResNet50 | BERT | 97.5 (+4.5) | 98.6 (+2.2) | 98.8 (+1.7) |
| | | DistilBERT | 97.4 (+4.4) | 98.6 (+2.2) | 98.8 (+1.7) |
| | | DeBERTa | 97.3 (+4.3) | 98.5 (+2.1) | 98.8 (+1.7) |

In parentheses, we have the potential improvement compared to the best single model accuracy

and ResNet50 models for multimodal evaluation. The results suggest that a multimodal model with MobileNetV3-Small can achieve competitive performance with a multimodal model with ResNet50, although MobileNetV3-Small has significantly lower accuracy individually. A larger image size requires an increase in the computational power of the image models, and the oracle classifier showed minor sub-percentage point improvements using a larger image size. Therefore, the image size for the image model is set to 256. DistilBERT is selected as the text model as it is the smallest text model and because of similar results from the text models. All fusion models except score fusion are trained end-to-end, initialized with pre-trained features from the provided frameworks. For the score fusion model, the image and text modality is first trained separately. Resulting in separate image and text classifiers. In the second step, the score fusion model is trained on the probabilities of each modality as input. The results of the multimodal classification can be seen in Table 4.

The multimodal results clearly show the improvements in fusing image and textual features. For both image models with the DistilBERT model, all fusion models increase the classification results compared to DistilBERT. The number of training samples affects the classification accuracy of the

multimodal model, increasing the accuracy by several percentage points on 50 training samples, while the increment is reduced when more training samples are used. The feature concatenation fusion method is overall better compared to the other fusion methods. Unlike the oracle classifier, there is also a large gap in classification accuracy when using image models, showing that ResNet50 provides superior performance when used for multimodal classification.

Experiments with more complex feature fusion models have also been carried out. The co-attention and cross-attention techniques proposed in Zhang et al. [94] did not improve the results compared to the selected fusion method. Furthermore, we investigate whether using a long short-term memory (LSTM) network in the output of the text model, as suggested by Gallo et al. [95], yields improved classification results; however, this did not happen. A more thorough discussion on the results of the fusion models is presented in Sect. 6.

Analysis of multimodal per-class improvements We evaluate how the multimodal model affects the classification accuracy of each class compared to the unimodal models. The image and text combination ResNet50+DistilBERT with feature concatenation as the fusion method had the highest multimodal accuracy. Hence, we select this one for evalua-

Table 4 Multimodal classification accuracy (in percentage) of the image model MobileNetV3-Small and ResNet50 model with the text model DistilBERT, using the fusion models Feature Concatenation, Score Fusion, GMU and EmbraceNet

| Image/Text models | Fusion model | Max #Samples | | |
|--------------------|--------------|--------------------|--------------------|--------------------|
| | | 50 | 200 | 400 |
| DistilBERT | N/A | 84.2 | 89.1 | 90.1 |
| MobileNetV3-Small | N/A | 86.5 | 92.1 | 92.8 |
| ResNet50 | N/A | 92.0 | 95.9 | 96.6 |
| | Feature Conc | 90.7 (+4.2) | 95.5 (+3.4) | 95.9 (+3.1) |
| MobileNetV3-Small+ | Score fusion | 87.9 (+1.4) | 93.3 (+1.2) | 94.2 (+1.4) |
| DistilBERT | GMU | 89.5 (+3.0) | 94.5 (+2.4) | 95.6 (+2.8) |
| | EmbraceNet | 84.9 (-1.6) | 92.8 (+0.7) | 94.6 (+1.8) |
| | Feature conc | 95.1 (+3.1) | 97.2 (+1.3) | 97.6 (+1.0) |
| ResNet50+ | Score fusion | 94.7 (+2.7) | 97.2 (+1.3) | 97.5 (+0.9) |
| DistilBERT | GMU | 94.9 (+2.9) | 96.9 (+1.0) | 97.2 (+0.6) |
| | EmbraceNet | 94.4 (+2.4) | 97.0 (+1.1) | 97.2 (+0.6) |

These are compared with the best unimodal accuracy of DistilBERT, MobileNetV3-Small and ResNet50. The best multimodal fusion result for each image-text model for the maximum sampling size is marked in bold, while the best overall model for each maximum sampling size is italic. In parentheses is the potential improvement compared to the best single model accuracy of the image-text combination

Table 5 The bottom 5 classes with lowest accuracy for the image model ResNet50, text model DistilBERT and the multimodal classifier combining these models with the feature concatenation as the fusion method

| (a) ResNet50 Class | Id | Accuracy (%) | (b) DistilBERT Class | Id | Accuracy | (c) ResNet50+DistilBERT Class | Id | Accuracy |
|---|-----|--------------|---|-----|----------|---|-----|----------|
|  | 224 | 33.3 |  | 163 | 20.0 |  | 249 | 58.0 |
|  | 249 | 46.0 |  | 55 | 21.0 |  | 135 | 70.7 |
|  | 69 | 50.0 |  | 250 | 34.3 |  | 118 | 72.2 |
|  | 228 | 51.5 |  | 248 | 37.0 |  | 224 | 72.2 |
|  | 54 | 54.4 |  | 49 | 38.0 |  | 225 | 72.2 |

tion. Sampling sizes of 50, 200, and 400 are used, where the purpose is to see how many classes are affected by the multimodal model, positively and negatively. We perform a paired t-test to see which classes have reduced, equal, or improved performance. We train 10 different image, text, and multimodal models for this. For each class, we extract the total

number of correct predictions for the class on both the unimodal and multimodal models. Then, a null hypothesis test is performed, checking if the mean difference between the unimodal and multimodal models is different. A p -value of 1% is used to test the null hypothesis. As we are interested in a qualitative measure of the multimodal effect on the different

classes, this p -value ensures that the results of the 256 classes are reliable. The results of the test can be seen in Fig. 11.

The results show that the class accuracy for the multimodal model is equal to or improved compared to that of the image and text models. The reason for less improved classes for the multimodal model compared to the image model is that the performance of the image model is closer to that of the multimodal model when more training samples are available.

Analysis of low accuracy classification cases To see how the most challenging products are affected by a multimodal model, we show the bottom 5 products with the lowest classification accuracy for image, text, and multimodal models as shown in Table 5. Similarly to previous experiments, we use the ResNet50+DistilBERT combination with feature concatenation as the multimodal model. A training sample size of 200 is selected and the maximum sequence length and image size are set to 256 (similar results are achieved with other settings). The results show that the multimodal model is much more accurate for the classes with the lowest classification performance, improving the accuracy per product by a minimum of 12 percentage points and up to more than 50 percentage points. Inspection of the image classes with the poorest performance showed that each had classes that looked similar and only subtle text details distinguished the products. For the text model, poor performance had multiple reasons, such as lack of discriminative text information, poor/partial OCR reading, and changes in product appearance in the test set. For most of the worst classes in the multimodal model, poor performance is due to the changing appearance of the product. Although the text information is similar, the packages have a different visual appearance. The other classes with poor performance are almost identical, and the OCR reading is unable to extract discriminatory parts in many cases.

Evaluation on Product Leaflet dataset As described in Sect. 3, no previous grocery product recognition datasets contain challenging products for fine-grained recognition. To compare our approach to other methods, we therefore evaluate our approach on a dataset in a similar field. The Product Leaflet Dataset [79] dataset is a fine-grained classification dataset with product leaflet images. The challenge with the dataset is that there is a great variation in the appearance of the images in a class. However, the textual information is rich and informative. The dataset has 832 classes with 33280 training images and 8320 test images. For each class, there are 40 training samples and 10 test samples.

We evaluate our approach on this dataset with the data extraction method described in Sect. 3. ResNet50 and DistilBERT are selected for unimodal comparisons. For multimodal comparison, we select Score Fusion and Feature Concatenation. We use the same experimental setup as described in the article, with an image size of 256 for the

image classifier and an image size of 512 for OCR extraction and text classification. The results can be seen in Fig. 6.

We see that our approach substantially improves the accuracy on both unimodal and multimodal models, setting a new state-of-the-art baseline. In particular, the performance of the DistilBERT model which surpasses the results of the weight proba fusion method of the previous approach. In addition to the above mentioned results, our approach also won the Kaggle competition associated with the article.⁵ We achieved a mean F1-score of 94.67% on the private leaderboard, while the second runner-up had an F1-score of 93.00%.

6 Discussion and recommendations

In this section, we discuss the significance of our results and provide recommendations for recognising grocery products using image and textual data.

In our work, we can see that fusing image and text information from product packages improves the recognition performance of grocery products. This has been done using a dataset from a grocery store that captures products with fine-grained details. To show that textual information provides additional information, we have shown that an increase in image resolution of image models does not improve performance notably. Instead, by using textual information, the recognition performance is significantly increased; hence, the image models in our experiments are not able to capture fine-grained details. An alternative approach is to use only fine-grained image-based recognition methods. However, these methods often require additional annotated data or complex training methods [2, 43]. Another aspect is that two products might be easy to distinguish on all sides except one, making it only necessary to use the fine-grained recognition techniques on that side. Our approach handles this by learning which modality to focus on, depending on the input data.

It is also shown that the performance of the different text models is negligible. However, the size of each model differs significantly. For example, DistilBERT has 40% less parameters than BERT. This indicates that the complexity of the text models could be further reduced. The default settings are used in the experiments, but several parameters are possible to configure to further reduce the complexity of the models, for example, the number of attention heads, the size of the encoder layers, and the number of hidden layers. Another approach is to use even simpler text models, such as the MobileBERT model [96]. Reducing the maximum sequence lengths from 256 to 128 showed a minor change in accuracy of just a few percentage points. Even if this work focuses on

⁵ <https://www.kaggle.com/competitions/retail-products-classification-2023/overview>.

Table 6 Recognition results on Product Leaflet Dataset

| Model | Method | Accuracy | Top3 | Top5 |
|----------------------------|----------------|----------|-------|-------|
| ResNet50 [79] | Image | 92.5% | 96.1% | 96.9% |
| SGD-Classifier [79] | Text | 91.5% | 96.2% | 96.7% |
| Combined [79] | Weighted proba | 96.4% | 99.2% | 99.3% |
| ResNet50 (ours) | Image | 94.4% | 98.1% | 98.4% |
| DistilBERT (ours) | Text | 97.3% | 99.4% | 99.5% |
| ResNet50+DistilBERT (ours) | Score Fusion | 97.8% | 99.5% | 99.6% |
| ResNet50+DistilBERT (ours) | Feature Conc | 98.1% | 99.8% | 99.9% |

products with detailed text elements, this also suggests that our approach can be used in setups where only limited text details are distinguishable. This makes this solution viable in product recognition solutions in SCOs where the products are moved closer or farther to the camera by the customer when scanning products. We also showed that our approach can be utilized in other domains with image and OCR data by achieving state-of-the-art results on the Product Leaflet dataset. In this work, no filtering or data augmentation techniques have been applied to the text models; this is an area to explore in the future.

We also see that the less complex fusion method performs better in our multimodal models. This is in line with previous results reported by [72]; the authors argue that patterns learned in pre-training are forgotten due to high error in the backpropagation in the initial training iterations. To alleviate this, they suggested training multimodal models with small learning rates using ADAM as the optimizer, which is also applied in this work. Another observation is that both domains complement each other well when a small number of training samples are available. When the number of training samples increases, the difference between the best unimodal model and the multimodal model decreases. Even if it decreases, there is still a performance gap that shows that the image and text domains complement each other.

Although not better than the simple feature concatenation fusion model, EmbraceNet has several properties to work in a practical environment. For multiple types of products, there is no textual information, for example, for fruits and vegetables. This could be handled seamlessly by the EmbraceNet model by only activating the visual features for recognition. In the other fusion models, an image-based model and a multimodal model have to be trained separately to achieve the same purpose. In [9], it is shown that text-based grocery product recognition models are robust to domain adaptation, while the performance of image-based models degrades significantly. This domain adaptation aspect is a practical problem when a system is installed in a new grocery store or if the camera type or camera position changes in an existing one. Using the EmbraceNet model, only the text model could be used at the beginning and the image model could be

activated when sufficient training data have been acquired in the new environment.

From the analysis of challenging classes, some products had low recognition performance because the OCR text information could not be extracted sufficiently. To handle these cases, the recommendation is to look at improving the camera and lighting to be able to capture fine-grained images, which can be used to separate classes.

The additional computational cost of incorporating OCR for product recognition is significant. This has been identified as a concern for the applicability of these types of models in [97]. But in many cases, it is not relevant due to non-real-time requirements, for example, in shelf monitoring analysis where the interval of capturing images is several minutes. When barcode switches are identified at SCOs, only one input image is needed to identify when the product is scanned at the barcode scanner. In streaming applications like frictionless checkout, our approach can be adapted to specialize in products that are hard-to-classify. Specifically, the EmbraceNet architecture can be utilized to classify the results on the image-modality first. If it indicates a pre-known hard-to-classify product, the textual information from OCR could be applied to further improve the confidence in the classification. While the above applies to systems with either in-store machine learning servers and cloud computing, edge devices might pose additional challenges. It is not within the scope of this paper to perform a complete analysis of edge device performance. However, using suggestions on how to adapt the text model architecture or using even more lightweight versions of BERT in combination with previous work that uses similar text models on edge devices [98], we claim that our approach is also feasible to adapt and use on edge devices.

To summarize our results, we find that:

- Combining image and text from OCR significantly increases classification performance. Especially with a low number of training samples, and the difference is still considerable with a large number of training samples.
- The size of the input image and the length of the text sequence had a small impact on the multimodal classification results. This indicates that our approach generalizes

well, and thus, it can be used in systems with less image resolution.

- Model selection matters for the image domain, while simpler text models saw only a small drop in performance. This suggests that simpler text models can be used, further reducing computational complexity.
- Feature fusion gave the best multimodal classification results, outscoring more complex fusion models. However, we also recommend EmbraceNet, due to its practical design for easy deployment.

Our results show that this technology can be applied to many product recognition applications and further improve recognition performance.

7 Conclusions and future work

In this work, we present the dataset FineGrainOCR for fine-grained recognition of grocery products. For this dataset, we propose an approach that uses image and OCR-extracted product text to improve recognition performance. Several experiments are performed that combine image and text models with multimodal fusion methods. Several trade-offs are considered, such as the number of training samples, image resolution, and length of text. The results show that the combination of image and text information using multimodal techniques is superior compared to unimodal models and significantly improves the results on our dataset with a limited number of training samples. In addition, we see that our approach generalizes well and shows state-of-the-results in another retail domain. In future work, we intend to evaluate our approach in a more real-world open retail environment, such as an SCO. In this environment, multiple products are normally present and overlap each other, requiring a more complex method to handle such overlaps. Additionally, we will tackle the challenge of performing OCR reading and multimodal classification in real-time with limited computational power, applying the recommendations proposed in this paper. Finally, we will employ explainable AI techniques to interpret what types of information image and text models are learning compared to the multimodal model.

Acknowledgements The authors would like to thank ITAB Shop Products AB and Smart Industry Sweden (KKS-2020-0044) for their support. The machine learning training was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Funding Open access funding provided by University of Skövde.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Barbier, J., France, N., Kirby, S., McDonald, B., O'Connell, K., Riegel, M.: A Roadmap to Digital Value in the Retail Industry. <https://pos.report/whitepapers/a-roadmap-to-digital-value-in-the-retail-industry>
2. Wei, X.-S., Song, Y.-Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J., Belongie, S.: Fine-grained image analysis with deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(12), 8927–8948 (2021)
3. Zhou, X., Pan, P., Zheng, Y., Xu, Y., Jin, R.: Large scale long-tailed product recognition system at alibaba. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3353–3356 (2020)
4. Zuo, Z., Wang, L., Momma, M., Wang, W., Ni, Y., Lin, J., Sun, Y.: A flexible large-scale similar product identification system in e-commerce. In: KDD Workshop on Industrial Recommendation Systems (2020)
5. Feng, F., Niu, T., Li, R., Wang, X., Jiang, H.: Learning visual features from product title for image retrieval. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4723–4727 (2020)
6. Santra, B., Mukherjee, D.P.: A comprehensive survey on computer vision based approaches for automatic identification of products in retail store. *Image Vis. Comput.* **86**, 45–63 (2019)
7. Wei, Y., Tran, S., Xu, S., Kang, B., Springer, M.: Deep learning for retail product recognition: challenges and techniques. *Comput. Intell. Neurosci.* **20**, 20 (2020). <https://doi.org/10.1155/2020/8875910>
8. Ouchikh, R., Pettersson, T., Löfström, T.: Product verification using OCR classification and Mondrian conformal prediction. *Expert Syst. Appl.* **188**, 115942 (2022)
9. Pettersson, T., Ouchikh, R., Lofstrom, T.: Nlp cross-domain recognition of retail products. In: 2022 7th International Conference on Machine Learning Technologies (ICMLT), pp. 237–243 (2022)
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems **30** (2017)
11. Sun, Y., Luo, T., Zuo, Z.: Context-guided feature enhancement network for automatic check-out. *Neural Comput. Appl.* **34**(1), 593–606 (2022)
12. Li, C., Du, D., Zhang, L., Luo, T., Wu, Y., Tian, Q., Wen, L., Lyu, S.: Data priming network for automatic check-out. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2152–2160 (2019)
13. Chen, H., Zhou, Y., Li, J., Wei, X.-S., Xiao, L.: Self-supervised multi-category counting networks for automatic check-out. *IEEE Trans. Image Process.* **31**, 3004–3016 (2022)

14. Yilmazer, R., Birant, D.: Shelf auditing based on image classification using semi-supervised deep learning to increase on-shelf availability in grocery stores. *Sensors* **21**(2), 327 (2021)
15. Falcão, J., Ruiz, C., Pan, S., Noh, H.Y., Zhang, P.: Faim: vision and weight sensing fusion framework for autonomous inventory monitoring in convenience stores. *Front. Built Environ.* **6**, 568372 (2020)
16. Higa, K., Iwamoto, K.: Robust shelf monitoring using supervised learning for improving on-shelf availability in retail stores. *Sensors* **19**(12), 2722 (2019)
17. Falcão, J.D., Ruiz, C., Bannis, A., Noh, H.Y., Zhang, P.: Isacs: in-store autonomous checkout system for retail. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **5**(3), 1–26 (2021)
18. Liu, X., Jiang, Y., Kim, K.-H., Govindan, R.: Grab: Fast and accurate sensor processing for cashier-free shopping (2020)
19. Beck, A.: Self-Checkout in Retail: Measuring the Loss. ECR Community Shrinkage and On-shelf Availability Group, Brussels (2018)
20. Merler, M., Galleguillos, C., Belongie, S.: Recognizing groceries in situ using *in vitro* training data. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007). IEEE
21. George, M., Floerkemeier, C.: Recognizing products: a per-exemplar multi-label image classification approach. In: European Conference on Computer Vision, pp. 440–455 (2014). Springer
22. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
23. Srivastava, M.M.: Bag of tricks for retail product image classification. In: Image Analysis and Recognition: 17th International Conference, ICIAR 2020, Póvoa de Varzim, Portugal, June 24–26, 2020, Proceedings, Part I 17, pp. 71–82 (2020). Springer
24. Franco, A., Maltoni, D., Papi, S.: Grocery product detection and recognition. *Expert Syst. Appl.* **81**, 163–176 (2017)
25. Tonioni, A., Serra, E., Di Stefano, L.: A deep learning pipeline for product recognition on store shelves. In: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), pp. 25–31 (2018). IEEE
26. Hameed, K., Chai, D., Rassau, A.: Class distribution-aware adaptive margins and cluster embedding for classification of fruit and vegetables at supermarket self-checkouts. *Neurocomputing* **461**, 292–309 (2021)
27. Shiraishi, S., Kikuchi, K., Iwamoto, K.: Classification system with capability to reject unknowns. In: 2019 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6 (2019). IEEE
28. Yang, Y., Sheng, L., Jiang, X., Wang, H., Xu, D., Cao, X.: Increaco: incrementally learned automatic check-out with photo-realistic exemplar augmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 626–634 (2021)
29. Ciocca, G., Napoletano, P., Locatelli, S.G.: Multi-task learning for supervised and unsupervised classification of grocery images. In: International Conference on Pattern Recognition, pp. 325–338 (2021). Springer
30. Advani, S., Smith, B., Tanabe, Y., Irick, K., Cotter, M., Sampson, J., Narayanan, V.: Visual co-occurrence network: using context for large-scale object recognition in retail. In: 2015 13th IEEE Symposium on Embedded Systems for Real-time Multimedia (ESTIMedia), pp. 1–10 (2015). IEEE
31. Wei, Y., Xu, S., Kang, B., Hoque, S.: Generating training images with different angles by GAN for improving grocery product image recognition. *Neurocomputing* **488**, 694–705 (2022)
32. Wei, Y., Xu, S., Tran, S., Kang, B.: Data augmentation with generative adversarial networks for grocery product image recognition. In: 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 963–968 (2020). IEEE
33. Tonioni, A., Di Stefano, L.: Domain invariant hierarchical embedding for grocery products recognition. *Comput. Vis. Image Underst.* **182**, 81–92 (2019)
34. Zhang, L., Du, D., Li, C., Wu, Y., Luo, T.: Iterative knowledge distillation for automatic check-out. *IEEE Trans. Multimed.* **23**, 4158–4170 (2020)
35. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. *IEEE Signal Process. Lett.* **25**(7), 926–930 (2018)
36. Hameed, K., Chai, D., Rassau, A.: A sample weight and adaboost CNN-based coarse to fine classification of fruit and vegetables at a supermarket self-checkout. *Appl. Sci.* **10**(23), 8667 (2020)
37. Wang, W., Cui, Y., Li, G., Jiang, C., Deng, S.: A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. *Neural Comput. Appl.* **32**(18), 14613–14622 (2020)
38. Geng, W., Han, F., Lin, J., Zhu, L., Bai, J., Wang, S., He, L., Xiao, Q., Lai, Z.: Fine-grained grocery product recognition by one-shot learning. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 1706–1714 (2018)
39. Santra, B., Shaw, A.K., Mukherjee, D.P.: Part-based annotation-free fine-grained classification of images of retail products. *Pattern Recogn.* **121**, 108257 (2022)
40. Georgieva, P., Zhang, P.: Optical character recognition for autonomous stores. In: 2020 IEEE 10th International Conference on Intelligent Systems (IS), pp. 69–75 (2020). IEEE
41. Baz, I., Yoruk, E., Cetin, M.: Context-aware hybrid classification system for fine-grained retail product recognition. In: 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), pp. 1–5 (2016). IEEE
42. Chen, Y., Bai, Y., Zhang, W., Mei, T.: Destruction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5157–5166 (2019)
43. Zhao, B., Feng, J., Wu, X., Yan, S.: A survey on deep learning-based fine-grained object classification and semantic segmentation. *Int. J. Autom. Comput.* **14**(2), 119–135 (2017)
44. Shaker, Z., Feng, X., Tahir, M.A.A.: Using text and visual cues for fine-grained classification. *Int. J. Adv. Netw. Monit. Controls* **6**(3), 42–49 (2021)
45. Azizi, S., Fang, U., Adibi, S., Li, J.: Supervised contrastive learning for product classification. In: International Conference on Advanced Data Mining and Applications, pp. 341–355 (2022). Springer
46. Klasson, M., Zhang, C., Kjellström, H.: A hierarchical grocery store image dataset with visual and semantic labels. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 491–500 (2019). IEEE
47. Jund, P., Abdo, N., Eitel, A., Burgard, W.: The freiburg groceries dataset (2016)
48. Wei, X.-S., Cui, Q., Yang, L., Wang, P., Liu, L.: Rpc: a large-scale retail product checkout dataset (2019)
49. Follmann, P., Bottger, T., Hartinger, P., Konig, R., Ulrich, M.: Mvtac d2s: densely segmented supermarket dataset. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 569–585 (2018)
50. Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., Hassner, T.: Precise detection in densely packed scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5227–5236 (2019)
51. Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., Xu, C.: Dynamic refinement network for oriented and densely packed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11207–11216 (2020)
52. Peng, J., Xiao, C., Li, Y.: Rp2k: A large-scale retail product dataset for fine-grained image classification (2020)

53. Cai, Y., Wen, L., Zhang, L., Du, D., Wang, W.: Rethinking object detection in retail stores. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 947–954 (2021)
54. Georgiadis, K., Kordopatis-Zilos, G., Kalaganis, F., Migkotzidis, P., Chatzilari, E., Panakidou, V., Pantouvakis, K., Tortopidis, S., Papadopoulos, S., Nikolopoulos, S.: Products-6k: a large-scale groceries product recognition dataset. In: The 14th PErvasive Technologies Related to Assistive Environments Conference, pp. 1–7 (2021)
55. Chen, F., Zhang, H., Li, Z., Dou, J., Mo, S., Chen, H., Zhang, Y., Ahmed, U., Zhu, C., Savvides, M.: Unitail: detecting, reading, and matching in retail scene. In: European Conference on Computer Vision, pp. 705–722 (2022). Springer
56. Paolanti, M., Sturari, M., Mancini, A., Zingaretti, P., Frontoni, E.: Mobile robot for retail surveying and inventory using visual and textual analysis of monocular pictures based on deep learning. In: 2017 European conference on mobile robots (ECMR), pp. 1–6 (2017). IEEE
57. Garillo-Manliguez, C.A., Chiang, J.Y.: Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation. *Sensors* **21**(4), 1288 (2021)
58. Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B.: Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sens. Environ.* **237**, 111599 (2020)
59. Melotti, G., Premebida, C., Gonçalves, N.M.D.S., Nunes, U.J., Faria, D.R.: Multimodal cnn pedestrian classification: a study on combining lidar and camera data. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3138–3143 (2018). IEEE
60. Gupta, A., Anpalagan, A., Guan, L., Khwaja, A.S.: Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues. *Array* **10**, 100057 (2021)
61. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(2), 423–443 (2018)
62. Sleeman, W.C., IV, Kapoor, R., Ghosh, P.: Multimodal classification: current landscape, taxonomy and future directions. *ACM Comput. Surv.* **55**(7), 1–31 (2022)
63. Gallo, I., Calefati, A., Nawaz, S., Janjua, M.K.: Image and encoded text fusion for multi-modal classification. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7 (2018). IEEE
64. Nawaz, S., Calefati, A., Janjua, M.K., Anwaar, M.U., Gallo, I.: Learning fused representations for large-scale multimodal classification. *IEEE Sens. Lett.* **3**(1), 1–4 (2018)
65. Gallo, I., Nawaz, S., Landro, N., Grassainst, R.L.: Visual word embedding for text classification. In: International Conference on Pattern Recognition, pp. 339–352 (2021). Springer
66. Chordia, V., BG, V.K.: Large scale multimodal classification using an ensemble of transformer models and co-attention. arXiv preprint [arXiv:2011.11735](https://arxiv.org/abs/2011.11735) (2020)
67. Kiela, D., Grave, E., Joulin, A., Mikolov, T.: Efficient large-scale multi-modal classification 32(1) (2018)
68. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification (2016)
69. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
70. Mafla, A., Dey, S., Biten, A.F., Gomez, L., Karatzas, D.: Fine-grained image classification and retrieval by combining visual and locally pooled textual features. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2950–2959 (2020)
71. Zahavy, T., Krishnan, A., Magnani, A., Mannor, S.: Is a picture worth a thousand words? a deep multi-modal architecture for product classification in e-commerce. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
72. Comble, A., Dutt, A., Montalvo, P., Salah, A.: Multimodal attribute extraction for e-commerce. arXiv preprint [arXiv:2203.03441](https://arxiv.org/abs/2203.03441) (2022)
73. Bi, Y., Wang, S., Fan, Z.: A multimodal late fusion model for e-commerce product classification (2020)
74. Åberg, L.: Multimodal classification of second-hand e-commerce ads (2018)
75. Chen, L., Chou, H., Xia, Y., Miyake, H.: Multimodal item categorization fully based on transformer. In: Proceedings of The 4th Workshop on e-Commerce and NLP, pp. 111–115 (2021)
76. Chen, Q., Shi, Z., Zuo, Z., Fu, J., Sun, Y.: Two-stream hybrid attention network for multimodal classification. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 359–363 (2021). IEEE
77. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. In: Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I, pp. 427–443 (2020). Springer
78. Amazon: How Amazon Robotics Is Working on New Ways to Eliminate the Need for Barcodes. Accessed 30 June 2023
79. Ladwig, D., Lamm, B., Keuper, J.: Fine-grained product classification on leaflet advertisements. arXiv preprint [arXiv:2305.03706](https://arxiv.org/abs/2305.03706) (2023)
80. Pati, S., Aga, S., Jayasena, N., Sinclair, M.D.: Demystifying bert: Implications for accelerator design (2021)
81. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
82. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s, 11976–11986 (2022)
83. Wightman, R., Touvron, H., Jégou, H.: Resnet strikes back: an improved training procedure in timm (2021)
84. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)
85. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
86. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
87. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019)
88. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention (2020)
89. Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint [arXiv:1702.01992](https://arxiv.org/abs/1702.01992) (2017)
90. Choi, J.-H., Lee, J.-S.: Embracenet: a robust deep learning architecture for multimodal classification. *Inf. Fusion* **51**, 259–270 (2019)
91. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR
92. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18, pp. 194–206 (2019). Springer

93. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
94. Zhang, D., Nayak, R., Bashar, M.A.: Exploring fusion strategies in deep learning models for multi-modal classification. In: Australasian Conference on Data Mining, pp. 102–117 (2021). Springer
95. Gallo, I., Ria, G., Landro, N., La Grassa, R.: Image and text fusion for UPMC food-101 using bert and CNNs. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6 (2020). IEEE
96. Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., Zhou, D.: Mobilebert: a compact task-agnostic bert for resource-limited devices (2020)
97. Guimarães, V., Nascimento, J., Viana, P., Carvalho, P.: A review of recent advances and challenges in grocery label detection and recognition. *Appl. Sci.* **13**(5), 2871 (2023)
98. Pandelea, V., Ragusa, E., Apicella, T., Gastaldo, P., Cambria, E.: Emotion recognition on edge devices: training and deployment. *Sensors* **21**(13), 4496 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Tobias Pettersson received M.Sc. degree in Computer Science from Linköping University in 2012. He is currently pursuing a Ph.D. at the Virtual Production Development department, University of Skövde, Sweden, as a member of the Smart Industry Sweden graduate school. His main research interests include learning, computer vision, natural language processing, and retail technology.



able AI and visual analytics.



Maria Riveiro received an M.Sc. degree in Telecommunication Engineering from the University of Vigo, Spain, and a Ph.D. degree in Computer Science from Örebro University, Sweden, 2004 and 2011, respectively. Currently, she is a Professor of Computer Science at the Department of Computer Science and Informatics at the School of Engineering, Jönköping University, with the Human-Centered Technology Group. Her main research interests are human-centered explainable AI and visual analytics.

Tuwe Löfström is an assistant professor Computer Science within Jönköping AI Lab (JAIL) at Jönköping University. His current research focus is on algorithms and methods for uncertainty quantification, trustworthiness and explanations within ML and AI.