Explainable Local and Global Models for Fine-Grained Multimodal Product Recognition

Tobias Pettersson
tobias.pettersson@itab.com
University of Skövde, Jönköping University, ITAB Shop Products AB
Sweden

Maria Riveiro
maria.riveiro@ju.se
Dept Computer Science and Informatics, Jönköping University
Sweden

Tuwe Löfström
tuwe.lofstrom@itab.com
Jönköping University
Sweden

ABSTRACT

Grocery product recognition techniques are emerging in the retail sector and are used to provide automatic checkout counters, reduce self-checkout fraud, and support inventory management. However, recognizing grocery products using machine learning models is challenging due to the vast number of products, their similarities, and changes in appearance. To address these challenges, more complex models are created by adding additional modalities, such as text from product packages. But these complex models pose additional challenges in terms of model interpretability. Machine learning experts and system developers need tools and techniques conveying interpretations to enable the evaluation and improvement of multimodal production recognition models.

In this work, we propose thus an approach to provide local and global explanations that allow us to assess multimodal models for product recognition. We evaluate this approach on a large fine-grained grocery product dataset captured from a real-world environment. To assess the utility of our approach, experiments are conducted for three types of multimodal models.

The results show that our approach provides fine-grained local explanations while being able to aggregate those into global explanations for each type of product. In addition, we observe a disparity between different multimodal models, in what type of features they learn and what modality each model focuses on. This provides valuable insight to further improve the accuracy and robustness of multimodal product recognition models for grocery product recognition.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Multimodal classification, Explainable AI, Grocery product recognition, LIME, Fine-grained recognition, Optical character recognition

Multimodal KDD 2023, August 07, 2023, Long Beach, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Tobias Pettersson, Maria Riveiro, and Tuwe Löfström. 2023. Explainable Local and Global Models for Fine-Grained Multimodal Product Recognition. In *Proceedings of Multimodal KDD 2023 (International Workshop on Multimodal Learning) (Multimodal KDD 2023).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Multimodal models are used in a variety of application domains, from sentiment analysis to product recognition. In the retail sector, multimodal models are increasingly used to automatically recognize grocery products. These models embedded in checkout systems can improve customer flow in stores, improve customer experience, and reduce labour costs and store losses. Product recognition is, however, a challenging task for Machine Learning (ML)-based solutions due to imbalanced datasets with a vast number of categories, continuous updates of new grocery products, and recognition of different products with only subtle details that differentiate them.

But these multimodal models are generally complex and difficult to visualize and interact with. Understanding their behaviour, limitations, and internal interactions are key for performing debugging and evaluation before deployment; this understanding is also crucial for trust calibration, acceptance, and use of such models and the support systems that include them [22].

This work presents an explanatory approach to help ML experts of multimodal models for grocery product recognition to debug and assess their models during development. Our approach builds on existing Explainable AI (XAI) techniques, particularly Local Interpretable Model-agnostic Explanations (LIME) [29] for local explanations and NormLIME [2] for global explanations.

We evaluate our approach using a product recognition dataset collected from a real-world environment. We first build multiple multimodal models using images from 256 products and optical character reading (OCR) text extracted from their packages. Then, we create multimodal local and global explanations using LIME and NormLIME. An explanation prototype software is then built to support the analysis under local and global explanations. We then present the experiments carried out, and finally, we discuss various design choices and overall lessons learned. We carry out our investigations in an industrial setting with domain experts in grocery product recognition.

In summary, the main contributions of this paper are (1) an explanation approach to assess and evaluate multimodal models that includes solutions for aggregating local explanations from image and textual data into global explanations, (2) a demonstration of the utility of the approach by comparing three multimodal fusion

models, and (3) lessons learned and design choices for the implementation and experiments.

2 BACKGROUND AND RELATED WORK

2.1 Recognition of Grocery Products

Existing techniques for automatically recognizing grocery products are mainly based on image-based classifications. Fine-grained image recognition focuses on differentiating between hard-to-distinguish or similar types of products. The authors in [44] categorize fine-grained image recognition into three main paradigms: (1) finding key parts in an image and merging the local feature vector with a global vector representation; (2) learning better feature representation by high-order feature interaction or novel loss functions; (3) 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. Multimodal classification has been explored extensively in recent years; see surveys [6, 8, 37]. In retail, multimodal classification is common in e-commerce applications, see, e.g., [13, 47, 49], where product images combined with textual metadata are used to create more accurate models.

Indeed, recent progress in OCR has enabled extracting textual elements from product packages. The combination of text elements from the product packages with the respective image allows for more accurate and reliable recognition of products using multimodal classifiers [5, 28]. This strategy has been listed as an important research direction to improve the fine-grained recognition of grocery products in two recent surveys [33, 45]. In addition, OCR can be used in document classification [4], product leaflet classification [19] and package identification [3] in logistics.

2.2 Explainable AI and Explanation Methods

AI and ML-based systems are increasingly found in multiple application areas. Given their potential individual and social impact, these systems need to be designed to allow user control and oversight, and avoid the so-called "black-box" problem [26, 29, 31]. This can be achieved by including design aspects that support understandability and transparency. From a technological perspective, it is not straightforward to know what transparency means in these cases or how to achieve it. Still, current research suggests using interpretable ML-models and XAI methods. Consequently, there are now many different approaches to interpretable ML and XAI; see, for example, reviews in [7, 10, 16].

Lipton and Silva et al. [23, 36] differentiate between models that address transparency (how the model works) and post-hoc explanations (what else the model can tell) [23, 36]. The former refers to interpretable models that facilitate understanding of the mechanism by which the model works [23]. This can be achieved at various levels, at the level of the entire model (e.g., simulatability), at the level of individual components (e.g., parameters and decomposability), or at the level of the training algorithm (algorithmic transparency) [23]. Meanwhile, post-hoc explanations provide helpful information without addressing the model's inner workings [23, 36]. This is achieved through explanations by example, natural language explanations, or factual explanations, e.g., [24, 29]. One of the advantages of post-hoc explanations is that interpretations are provided after-the-fact without sacrificing predictive performance [23]. It is also common to distinguish between *global* and *local* (or instance-level) explanations, roughly equivalent to the former interpretable models and post-hoc explanations.

2.3 LIME and Global Explanations Methods

Among the variety of XAI and post-hoc explanation methods, LIME [29] is one of the most commonly used ones. LIME is a model-agnostic post-hoc method that explains the predictions of any classifier by building local linear models around the predictions of a considered opaque model. LIME (and its variations) can be classified as both an explanation method by simplification, a type of feature importance method, or a local explanation method [7, 39]. Other post-hoc explanation methods are, e.g., SHAP [24], Grad-CAM [35], SmoothGrad [38], Integrated Gradients [40].

Relevant to our case, when classifying images, LIME creates a set of perturbed instances by dividing the input image into interpretable components (contiguous superpixels); each perturbed instance is then run through the model to get a probability value [7]. After that, a simple linear model learns from this dataset, which is locally weighted. Finally, LIME shows the superpixels with the highest positive weights as an explanation [7].

In addition to providing local interpretations, LIME has been used for building global explanations; see extensions in, for instance, SP-LIME [29], global explanation with anchors [30], GALE [41], G-LIME [21] and NormLIME [2]. Regarding the explainability of multimodal models, DIME [25] is a method for fine-grained interpretation of multimodal models by resolving the model into unimodal contributions and multimodal interactions before generating visual explanations for each of them.

3 PROPOSED APPROACH

We present an approach to provide local and global explanations for multimodal product recognition models. First, we describe how an image of a product and extracted OCR are combined to visualize local explanations (Section 3.1). Then we aggregate local explanations, create a global explanation for each type of product and discuss the novel parts of our approach (Section 3.2).

3.1 Local Explanation of Multimodal Product Recognition Models

The overview of our local explanation approach for multimodal product recognition models is presented in Figure 1. We consider product recognition models with an image and a text modality in which the text has been extracted from the product using OCR. The result of the reading is the textual elements of the product and its corresponding bounding boxes. To classify the image and textual data, we use a dual-stream architecture where the image and text data are passed through separate models, and the embedding from these is then combined using a multimodal fusion technique.

We use a local surrogate model to create explanations for each individual sample. These explanations are based on the concept of features. A feature of the text modality in our work is a text element from the OCR reading. For example, if the OCR has read three texts, this gives three textual features. The features of the image modality are based on superpixels. A superpixel is a region of

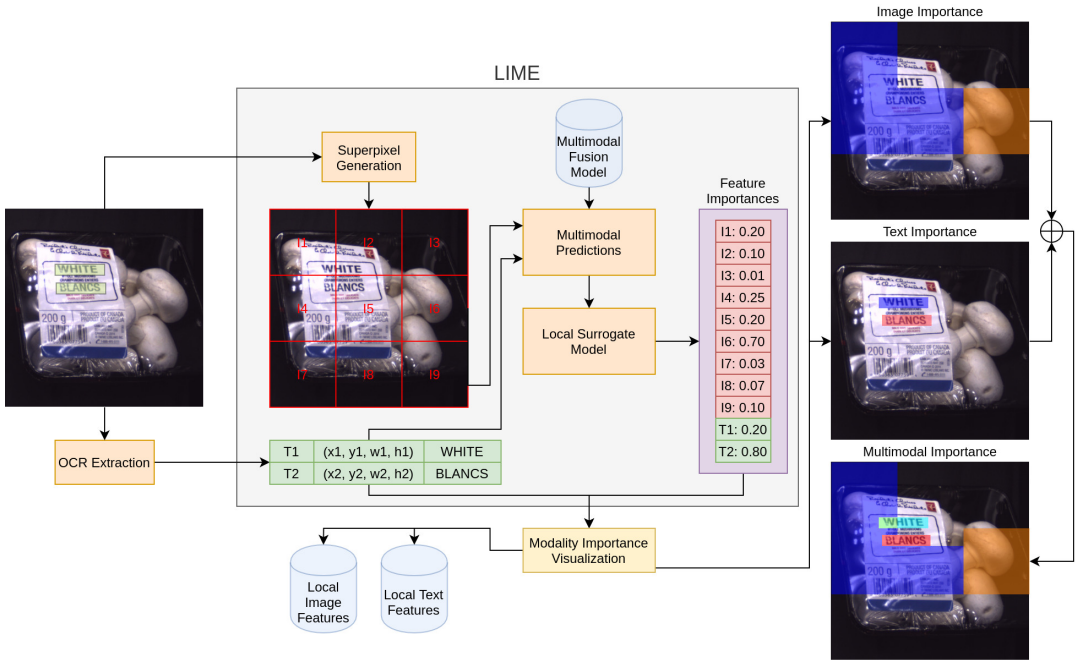


Figure 1: Proposed approach to extract and visualize multimodal local explanations. Samples with image and OCR text on product packages are fitted to a linear surrogate model using LIME. Resulting feature importances from the local surrogate model are then visualized into for the image and text modality. A multimodal visualization is then aggregated by adding the importance of each modality.

pixels with similar visual attributes. Our local surrogate model is trained by creating samples that randomly remove features from the original sample. The multimodal predictions from these perturbed samples are then fitted to a linear surrogate model to obtain local explanations.

From the local explanations, we combine the product with the OCR texts into a multimodal visualization image. We do this by first creating a blended overlayed heatmap over the original product image for each modality. These are constructed by coloring each image superpixel by its feature importance from the local surrogate model. For the text modality, we color it by text feature importance at the bounding box location for the OCR text entry. The multimodal visualization image is then constructed by merging the image and text visualization images (see Figure 1). In our work, we use LIME for creating local explanations; however, any method using local surrogate models can be used within this approach.

3.2 Global Explanation of Multimodal Product Recognition Models

Using local explanations from feature importance explanation methods, such as LIME, can indicate the model's behavior for particular samples. However, the appearance of the products differs significantly depending, for instance, on the orientation of the product. Given this large variability, local explanations are insufficient to assess the overall model behavior, capabilities, and limitations; therefore, global explanations are necessary.

Our approach to generating global explanations from local explanations is described in Figure 2. For each sample, we first calculate all the top local explanations from both modalities with their respective metadata. To create global image explanations, we first calculate the embeddings of each local image feature (superpixel). We then perform class-wise clustering of the embeddings. Grocery products can be of many shapes and different types of appearance, as described above; hence, it is not known how many clusters to expect. Therefore, we use HDBSCAN [9] as a clustering method, which can give an arbitrary number of clusters based on the characteristics of the embeddings. The features of the text explanations are represented as strings; therefore, we use a simple technique for global text explanations. We match the text features using the Levenshtein distance metric, which measures how many characters differ between two strings. With this metric, we can accept misspellings, which OCR readings are prone to have.

To value which of the clusters are the most important for a class, we need to calculate the global importance score of the clusters. However, it is not possible to compare local explanations due to the different feature importance scales for each sample. Therefore, we apply NormLIME [2], a technique to give clusters a global relative importance score compared to other clusters. With these global importance values, we can extract the image superpixels and OCR readings that are most important for a product. We select NormLIME due to its simplicity and performance to provide more faithful explanations compared to LIME [21]. Other similar techniques are LIME-SP [29], Averaged-Importance [41] and Homogeneity [41].

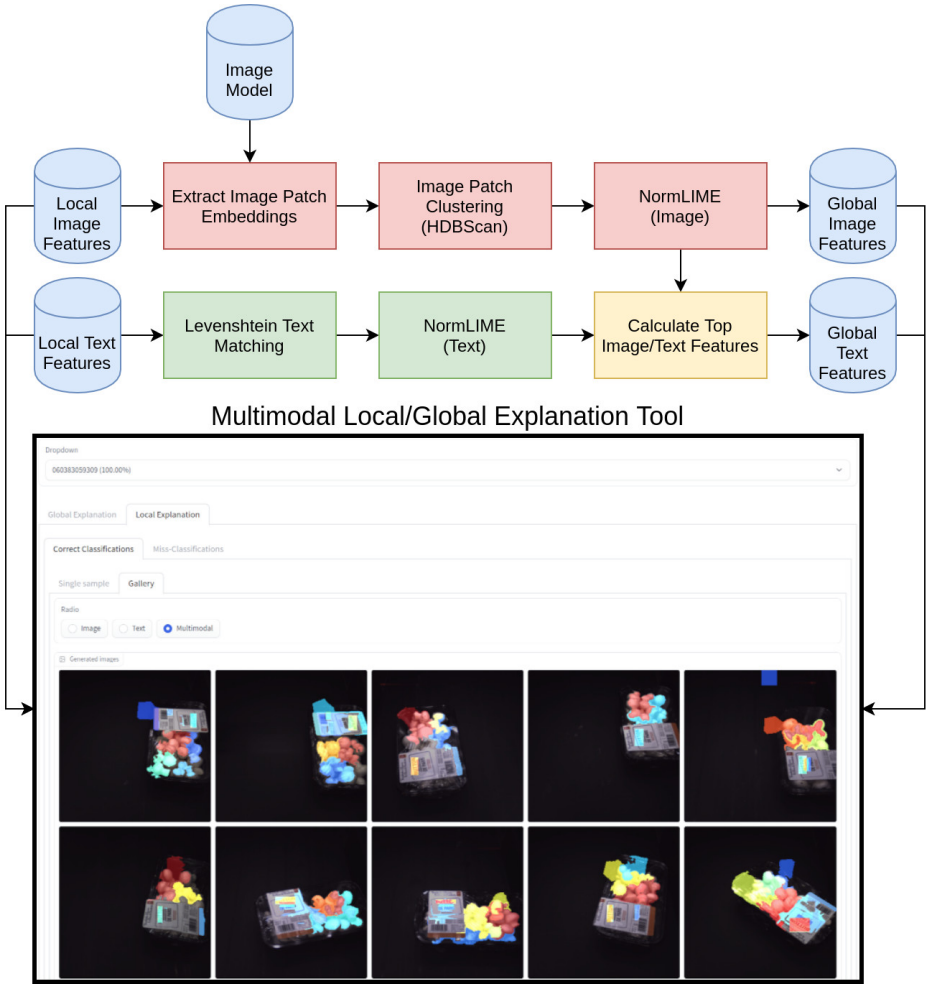


Figure 2: Overview of our approach to extract global multimodal features for image and text. We match and cluster local features from a feature importance explanation method, such as LIME, and calculate the top features from each modality using NormLIME. Results from local and global features are then combined and presented in a multimodal explanation prototype software.

Although generating local and global explanations as an end-to-end solution is possible, our approach combines local and global explanations for different purposes. First, local explanations can be used by engineers to develop and debug multimodal models. Global explanations, in turn, can be employed for model verification before model deployment. Finally, global explanations can be used to discuss results and challenges for different stakeholders in the retail domain.

Our approach is inspired by the global explanation part of G-LIME [21]. We extend and adapt their solution to make it more robust and easy to use, considering the challenges of the industrial case. First, we extend the approach to multimodal data while also providing multimodal visualization. Second, we identify (see Section 4.3) the importance of using image embeddings from supervised training, update the clustering method to address our classification task, identify that language models are unable to discriminatory embeddings for individual OCR texts and propose Levenshtein matching as a solution. Finally, we utilize an explanation prototype software for fast and accurate model evaluation (see Section 4.4).

4 EXPERIMENTAL SETUP

In this section, we describe the experimental setup for the approach with the dataset in Section 4.1, our parameter selection for the explanation method in Section 4.2, the details of our multimodal model selection in Section 4.3, and, finally, our explanation prototype software used in the experiments in Section 4.4.

4.1 Dataset

A significant number of grocery products have similar appearances. For example, different products that have various flavors, a lactose or non-lactose version of a product, similar or identical sides (e.g. the list of ingredients side), and the same type of product with different weights, volumes, or sizes. In these cases, it is very challenging for an image model to classify the products correctly. Combining image data with OCR readings into a multimodal model makes discrimination between products easier than an image-only model.

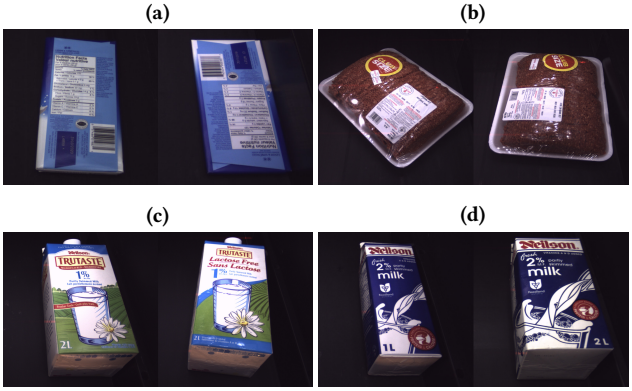


Figure 3: Example of challenging cases from our multimodal fine-grained recognition dataset where products have the same appearance and are only differentiable by text (ingredients side (a), meat packages (b)), have a lactose and non-lactose product variant (c), and have the same type of product with different weight (d).

Therefore, we have collected a product recognition dataset that includes all the above challenges. An example of some of the challenging image samples is presented in Figure 3. There exist several other datasets for grocery product recognition, see for example, [11, 14, 15, 27, 43]. However, none of these includes all of the challenges mentioned above, in particular, the multimodal and fine-grained aspects.

Our dataset is extracted from an automated scanning solution used in a large grocery store, which captures the image and its class using a barcode recognition system. We focus on the following six categories that contain many similar products; chocolate, dairy, meat, milk/cream, mushroom, and toppings. The dataset has 256 classes, each with 100 training and 50 validation samples. Each sample consists of an RGB image with a resolution of 2592x1944 and a text data file containing the OCR reading with its position within the image for the sample. The OCR text of the products is extracted using the Google Vision API[1]. The mean number of OCR reads for each sample in the training set is 27.2 words, with a standard deviation of 21.7. The mean number of OCR reads for the validation set is 25.7 words, with a standard deviation of 20.8.

[1]https://cloud.google.com/vision

4.2 Explanation Method

We have opted to use LIME as our feature importance explanation method. The motivation for this is two-fold; first, multimodal product recognition models are complex, often with different architectures in each stream. The model-agnostic property of LIME makes it easy to evaluate different types of multimodal product recognition models. Second, we want to be able to generate explanations for our entire validation dataset. Although LIME still requires a significant amount of computation power, it is still possible to extract local explanations of our validation dataset using a computer with a high-end graphics card.

Our LIME implementation with multimodal data is based on the implementation from the original authors of LIME[2]. However, we replace Quickshift [42] with Simple Linear Iterative Clustering (SLIC) [1] to generate superpixels. As in [34], we clearly see that SLIC extracts more relevant superpixels than Quickshift. We explored several values for the number of extracted SLIC superpixels and concluded that 100 superpixels captured distinguishing product attributes consistently. Also, we select 2000 as the sampling number for LIME. Larger sampling sizes did not give any significant changes for multimodal explanations. The perturbation of samples is done by removing random OCR text entries and superpixels from the multimodal input data.

We visualize our local explanations by blending an overlay image of the feature importance value with the original sample image. We do this by first normalizing the feature importance values from the image and text modality jointly between $0-1.0$. We then create a visualization image for each modality, where the image superpixels and the bounding boxes of the text features are colored. To easily distinguish the importance of features, we use the JET colormap for colorization. The multimodal explanation image is then constructed by aggregating the contributions from both the image and text visualization. Distinguishing between image and text feature contributions in multimodal visualization is intuitive due to the distinctive colorization of bounding boxes of text features. In addition, we also suppress visualization features that have a normalized feature value of less than 0.2, giving explanation images that are clearer and easier to interpret.

We extract the embeddings from each class's top local image explanations separately for our global explanations approach. We first train a ResNet50 [17] (see Section 4.3) image classifier with our training dataset. Then we use that model to extract embeddings from each sample's top 3 local image explanations. The image embeddings are then clustered using the HDBSCAN algorithm. Similarly to the work presented in [21], we evaluated image embeddings from a ResNet50 pre-trained on ImageNet to extract global explanations. However, this yielded image embeddings from which we could not find good cluster parameters. For global text explanations, we match local text explanations using a Levenshtein distance metric with a match ratio of 0.75. The motivation for this is to allow for the matching of noisy OCR readings. It is noteworthy that we also explored using DistilBERT features from a trained model, but the embeddings from individual OCR words (with misspellings) could not give separable clusters using HDBSCAN.

[2]https://github.com/marcotcr/lime

4.3 Model Selection and Training

We select the baseline models ResNet50 and DistilBERT [32] as the image respective text model backbone of our multimodal models. For our experiments, we evaluate three different multimodal fusion techniques to validate our proposed approach. *Feature Concatenation* [37] is a technique in which the embeddings of each modality are concatenated into a single vector, which is then fed to a neural network. In our experiments, we use the embeddings from the last layer of the image and text model. Similarly, *Score Fusion* [37] merges the predictions of the models and passes them to a one-layer neural network. These two techniques are the two standard approaches to performing multimodal fusion. The third selected mulitmodal fusion technique is *EmbraceNet* [12]. EmbraceNet uses the embedding from each modality and passes through a one-layer neural network that gives each modality the same dimension. These representations are then combined by probabilistically selecting features from each modality during training, guiding it to learn from both modalities. In addition to the three selected multimodal models, we also consider more sophisticated multimodal models from [18, 46, 48]. However, our evaluation of these models has shown that they present worse classification accuracy compared to the best unimodal models on our dataset. While it is interesting to evaluate the reason for this, it is beyond the scope of this paper.

Our multimodal recognition models use pre-trained image and text models from PyTorch[3] (ResNet50) and Huggingface[4] (DistilBERT). We use standard data augmentation techniques during model training, such as vertical/horizontal flipping and image rotation. We use the OCR text as input to the models for the text modality. The image is scaled to a size of 256x256 before being processed by the model, while the text input is set to a maximum length of 256 words. In rare cases where the OCR input exceeds 256 words, the remaining words are removed. All models are trained using the ADAM optimizer with a learning rate $2e^{-5}$ and a weight decay of $1e^{-4}$. In addition, we use a batch size of 16 and cross-entropy as the loss function.

4.4 Explanation Prototype Software

We generate all the data for the local and global explanations and save it to a database. Then a web-based Gradio[5] application written in Python reads all the information from the database and loads the data into the application. This design choice gives the application instant responsiveness, and it is possible to browse between classes and explanations easily. Furthermore, it is also possible to start several instances to perform side-to-side comparisons of different multimodal product recognition systems.

The GUI is simplistic, with a dropdown menu where class can be selected and a tab field where either local or global explanations can be displayed. In the global view, the top global features can be displayed for either the image or text modality. A histogram showing the modality importance ratio is also available. This histogram indicates which type of modality that is the most important for the recognition model when making a prediction. For local explanation, each sample within the class can be explored. We provide two different views. In the *Single Sample* view, the image, text, and multimodal explanation of the selected sample are displayed. In this view, the miss-classifications are also displayed that highlight feature importance for both the current class and the predicted class. This makes it possible to interpret the parts of the image that are important for the wrongly predicted sample. The *Gallery* collects all samples for a class and presents them in a grid. With this view, the user can evaluate whether the class gives consistent explanations throughout the samples. A screenshot of the explanation prototype software can be seen in Figure 2.

[3]https://pytorch.org/vision/stable/models.html
[4]https://huggingface.co/docs/transformers/model_doc/distilbert
[5]https://gradio.app/

5 EXPERIMENTS

In this section, we validate our approach by performing experiments for local explanations in Section 5.1 and global explanations in Section 5.2.

5.1 Local Explanations

First, we train our image, text, and multimodal models separately and evaluate them on our validation set. The results are presented in Table 1.

Models	Accuracy
DistilBERT	87.1%
ResNet50	93.2%
Score Fusion	93.4%
Feature Concatenation	96.5%
EmbraceNet	96.5%

Table 1: Classification results on the validation set for the unimodal classification models ResNet50 and DistilBERT and the multimodal models Score Fusion, Feature Concatenation, and EmbraceNet.

From the results, we see that both Feature Concatenation and EmbraceNet are able to utilize both textual and image information and provide a significant improvement in classification accuracy by 3.3 percentage points compared to the ResNet50 image model. In contrast, we see that Score Fusion is only capable of increasing the classification accuracy by 0.2 percentage points.

(a) (b) (c)

Figure 4: Example of multimodal visualizations from our approach with samples dominated by text (a) and image (b) explanations, and a sample combining image/text (c) explanations.

Then we run LIME for our multimodal models to calculate the local explanations on the validation set. We use a JET colormap to visualize the feature importance as described in Section 4.2. After

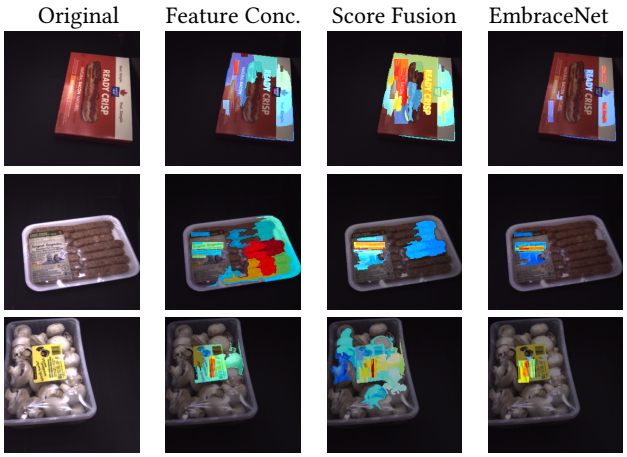


Figure 5: Multimodal local explanations from LIME differs significantly between multimodal classification models. This is illustrated with three exemplary products.

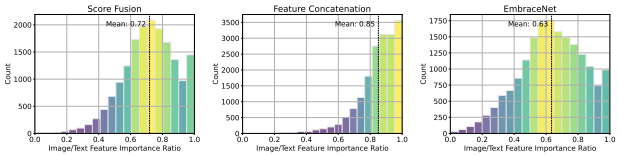
the explanations have been calculated, we extensively evaluate the result utilizing our explanation prototype software. We observe a great variety of explanations depending on the class and samples. Also, we note that the highlighted explanations reflect discriminatory image parts (distinct colors/textures) and OCR text (product-specific words). Figure 4 illustrates three of the main cases that utilize either image, image, and textual information, or textual information when making its prediction.

We also compare each of the multimodal models with our explanation prototype software. We can see that the explanations differ significantly between samples. An example of this is shown in Figure 5 where EmbraceNet bases its prediction on the product text, while Feature Concatenation and Score Fusion focus on the characteristics of the image and text. In general, when we have performed our extensive evaluation, EmbraceNet and Score Fusion combine information from both the image and text modality to a greater extent compared to Feature Concatenation.

In our explanation prototype software, we also summarize the modality feature importance ratio between image and text with our multimodal models. This is done by summarizing how much the feature importance is based on the image and text, respectively, for each sample. This is then aggregated into a histogram for all samples. In Figure 6, we present the image/text feature for the multimodal models. The results confirm our qualitative observations that EmbraceNet and Score Fusion comparison places a lot more attention on the text modality than Feature Concatenation.

5.2 Global Explanations

We then analyze the results of our global approach using our explanation prototype software for the three multimodal models. The top global features are selected for each modality and visualized with an image of the top score local explanation with additional metadata, such as the number of local samples and different text spellings for the text modality. We can observe that the multimodal models consistently extract discriminatory product attributes as



(a) (b) (c)

Figure 6: Histogram of image/text feature importance ratio for Score Fusion (a), Feature Concatenation (b) and EmbraceNet (c)

the top global explanations. These are often specific parts of the product with distinctive colors or textures. For global text explanations, discriminatory words are mostly extracted, such as the name of a meat package. Also, we see that some specific ingredients, such as the amount of sodium and saturated fat in a package, are important. Figure 7 illustrates an example of the top global explanations for image and text, respectively. We argue that this will increase the level of understanding of our trained multimodal models and in turn, increase trust (even if there are trade-offs, explanations from AI systems have shown to increase trust in AI, see e.g. [20]). Furthermore, they also show that they are not overfitted to specific image patterns or noisy OCR readings.

To validate whether our global explanations have a significant effect on the classification accuracy of the multimodal models, we perform a classification experiment that evaluates the effect of our global explanations. Utilizing the deletion metric described in [21], we first select the image samples that include either the top 3 images or text global features for each class. Then we classify these samples by the original sample, removal of the top global features, and finally, removal of a random feature. We perform the multimodal classification of each modality separately. In samples that contain global features for both image and text, we run both of these cases separately with one modality adjusting its input. If a sample contains more than one global feature for a modality, we remove the input from all of them. The removal of text features is done by removing the OCR text from the text input, while the image features blacken out the superpixels in the input image. We remove a random non-global text feature from the input data in the random deletion case. For the image modality, we blacken out a nearby image region by the size of the top image feature superpixel. The result of our experiment can be seen in Table 2.

Modality	Samples	Fusion model	Accuracy		
			No deletion	Top expl. deletion	Random expl. deletion
Image	16774	Feature Concat.	99.97%	77.59%	99.33%
	16398	Score Fusion	99.97%	93.42%	99.20%
	16951	EmbraceNet	99.41%	85.78%	99.06%
Text	9453	Feature Concat.	99.97%	96.01%	99.46%
	8053	Score Fusion	99.99%	83.89%	98.99%
	10396	EmbraceNet	99.77%	93.18%	99.44%

Table 2: Multimodal classification with samples containing global features, without global features and with samples deleting non-top features.

(a) Global image explanation for a class (left) with its supporting local image explanations from different samples (right)

(b) Global text explanation with metadata including most common text, different types of spellings, and number of explanation samples.

Figure 7: Example of top global image (a) and text (b) explanations

We are unable to compare the multimodal model side-by-side because different samples contain global explanations. However, we can draw some general conclusions. First, we can see that the removal of globally important features significantly reduces the accuracy of the model. In comparison, removing random features only slightly affects the accuracy, at most reducing accuracy by one percentage point. Second, the multimodal models are affected differently by removing different modalities. This is also in line with our results from the analysis of local explanations in Section 5.1.

6 CONCLUDING DISCUSSION

We present and demonstrate an approach for interpreting multimodal product recognition models using image and OCR text, empowering ML experts and system developers to create more accurate and robust models. Our approach supports the extraction of local explanations for a particular sample while also generating global explanations for each type of product.

We evaluate the suggested approach with a fine-grained grocery store dataset. We perform experiments with three different multimodal models, which validate that our approach extracts convincing local and global explanations for both the image and text modality. The experiments also show that multimodal product recognition models focus on different parts and modalities of the multimodal data. Furthermore, we validate that the robustness of the models differs when removing global features from the image or text modality.

Besides being a great utility for debugging different multimodal models, the approach can also be used to analyze different learning techniques or hyperparameters for a specific multimodal model. A limitation of our work is the computational power needed to extract global explanations for large datasets, yet, this can be significantly reduced by selecting specific classes with low accuracy.

Our approach is not limited to LIME and works for other feature-importance explanation methods. Furthermore, it is not limited to the recognition of grocery products. Any recognition system that uses images and OCR text can benefit from our proposals, for example, in document classification and package identification in logistics.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34, 11 (2012), 2274–2282.

[2] Isaac Ahern, Adam Noack, Luis Guzman-Nateras, Dejing Dou, Boyang Li, and Jun Huan. 2019. NormLime: A new feature importance metric for explaining deep neural networks. *arXiv preprint arXiv:1909.04200* (2019).

[3] Amazon. [n. d.]. *How Amazon Robotics is working on new ways to eliminate the need for barcodes.* Accessed: 2023-06-30.

[4] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. 2020. Multimodal deep networks for text and image-based document classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 427–443.

[5] Sahel Azizi, Uno Fang, Sasan Adibi, and Jianxin Li. 2022. Supervised Contrastive Learning for Product Classification. In *Advanced Data Mining and Applications: 17th International Conference, ADMA 2021, Sydney, NSW, Australia, February 2–4, 2022, Proceedings, Part II*. Springer, 341–355.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 arXiv:1910.10045

[8] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2021. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer* (2021), 1–32.

[9] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17*. Springer, 160–172.

[10] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (jul 2019), 832.

[11] Fangyi Chen, Han Zhang, Zaiwang Li, Jiachen Dou, Shentong Mo, Hao Chen, Yongxin Zhang, Uzair Ahmed, Chenchen Zhu, and Marios Savvides. 2022. Unitail: Detecting, Reading, and Matching in Retail Scene. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer, 705–722.

[12] Jun-Ho Choi and Jong-Seok Lee. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion* 51 (2019), 259–270.

[13] Fangxiang Feng, Tianrui Niu, Ruifan Li, Xiaojie Wang, and Huixing Jiang. 2020. Learning Visual Features from Product Title for Image Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4723–4727.

[14] Patrick Follmann, Tobias Bottger, Philipp Hartinger, Rebecca Konig, and Markus Ulrich. 2018. MVTec D2S: densely segmented supermarket dataset. In *Proceedings of the European conference on computer vision (ECCV)*. 569–585.

[15] Kostas Georgiadis, Giorgos Kordopatis-Zilos, Fotis Kalaganis, Panagiotis Migkotzidis, Elisavet Chatzilari, Valasia Panakidou, Kyriakos Pantouvakis, Savvas Tortopidis, Symeon Papadopoulos, Spiros Nikolopoulos, et al. 2021. Products-6K: a large-scale groceries product recognition dataset. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*. 1–7.

[16] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey Of Methods For Explaining Black Box Models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–42.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[19] Daniel Ladwig, Bianca Lamm, and Janis Keuper. 2023. Fine-Grained Product Classification on Leaflet Advertisements. *arXiv preprint arXiv:2305.03706* (2023).

[20] Yeonjoo Lee, Miyeon Ha, Sujeong Kwon, Yealin Shim, and Jinwoo Kim. 2019. Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI. *Computers in Human Behavior* 101 (2019), 180–196.

[21] Xuhong Li, Haoyi Xiong, Xingjian Li, Xiao Zhang, Ji Liu, Haiyan Jiang, Zeyu Chen, and Dejing Dou. 2022. G-LIME: Statistical Learning for Local Interpretations of Deep Neural Networks using Global Priors. *Artificial Intelligence* (2022), 103823.

[22] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. MultiViz: Towards Visualizing and Understanding Multimodal Models. arXiv:2207.00056 [cs.LG]

[23] Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 61, 10 (2018), 35–43.

[24] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[25] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations. *arXiv preprint arXiv:2203.02013* (2022).

[26] Thomas Muhlbacher, Harald Piringer, Samuel Gratzl, Michael Sedlmair, and Marc Streit. 2014. Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (dec 2014), 1643–1652.

[27] Jingtian Peng, Chang Xiao, and Yifan Li. 2020. RP2K: A large-scale retail product dataset for fine-grained image classification. *arXiv preprint arXiv:2006.12634* (2020).

[28] Tobias Pettersson, Rachid Oucheikh, and Tuwe Lofstrom. 2022. NLP Cross-Domain Recognition of Retail Products. In *2022 7th International Conference on Machine Learning Technologies (ICMLT)*. 237–243.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you? Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[31] Cynthia Rudin. 2014. Algorithms for interpretable machine learning. In *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*. 1519–1519.

[32] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[33] Bikash Santra and Dipti Prasad Mukherjee. 2019. A comprehensive survey on computer vision based approaches for automatic identification of products in retail store. *Image and Vision Computing* 86 (2019), 45–63.

[34] Ludwig Schallner, Johannes Rabold, Oliver Scholz, and Ute Schmid. 2020. Effect of superpixel aggregation on explanations in LIME–a case study with biological data. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*. Springer, 147–158.

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[36] Vivian S Silva, André Freitas, and Siegfried Handschuh. 2019. On the Semantic Interpretability of Artificial Intelligence Models. *arXiv preprint arXiv:1907.04105* (2019).

[37] William C Sleeman IV, Rishabh Kapoor, and Preetam Ghosh. 2022. Multimodal classification: Current landscape, taxonomy and future directions. *Comput. Surveys* 55, 7 (2022), 1–31.

[38] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[39] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *ACM International Conference Proceeding Series* (2022), 2239–2250.

[40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[41] Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global aggregations of local explanations for black box models. *arXiv preprint arXiv:1907.03039* (2019).

[42] Andrea Vedaldi and Stefano Soatto. 2008. Quick shift and kernel methods for mode seeking. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*. Springer, 705–718.

[43] XS Wei, Q Cui, L Yang, P Wang, and L Liu. [n. d.]. RPC: A large-scale retail product checkout dataset. arXiv 2019. *arXiv preprint arXiv:1901.07249* ([n. d.]).

[44] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. 2021. Fine-Grained Image Analysis with Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[45] Yuchen Wei, Son Tran, Shuxiang Xu, Byeong Kang, and Matthew Springer. 2020. Deep learning for retail product recognition: Challenges and techniques. *Computational intelligence and neuroscience* 2020 (2020).

[46] Duoyi Zhang, Richi Nayak, and Md Abul Bashar. 2021. Exploring Fusion Strategies in Deep Learning Models for Multi-Modal Classification. In *Data Mining: 19th Australasian Conference on Data Mining, AusDM 2021, Brisbane, QLD, Australia, December 14-15, 2021, Proceedings*. Springer, 102–117.

[47] Xiangzeng Zhou, Pan Pan, Yun Zheng, Yinghui Xu, and Rong Jin. 2020. Large scale long-tailed product recognition system at alibaba. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3353–3356.

[48] Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. *arXiv preprint arXiv:2009.07162* (2020).

[49] Zhen Zuo, Lixi Wang, Michinari Momma, Wenbo Wang, Yikai Ni, Jianfeng Lin, and Yi Sun. 2020. A flexible large-scale similar product identification system in e-commerce. In *KDD Workshop on Industrial Recommendation Systems*.