1977

UNIVERSITY
OF SKÖVDE

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

Tobias Pettersson, Maria Riveiro, Tuwe Löfström

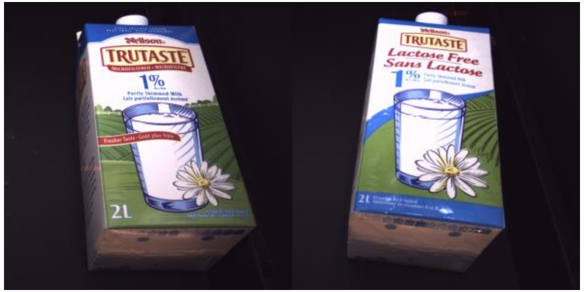
Product Recognition in Grocery Stores

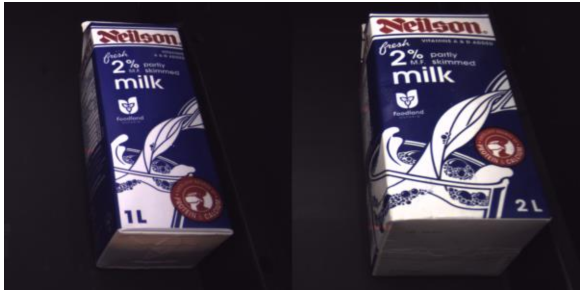
SCO Fraud Detection

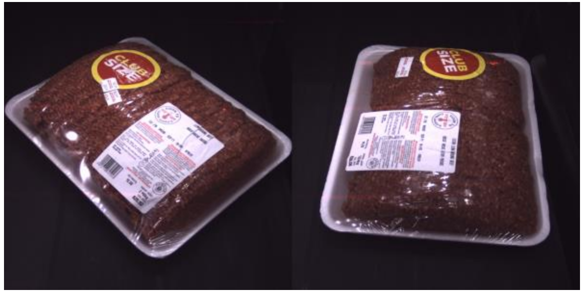
Shelf Availability

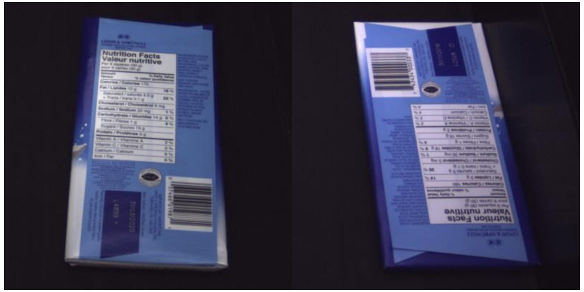
Automatic Checkout

Fine-Grained Product Recognition Problem

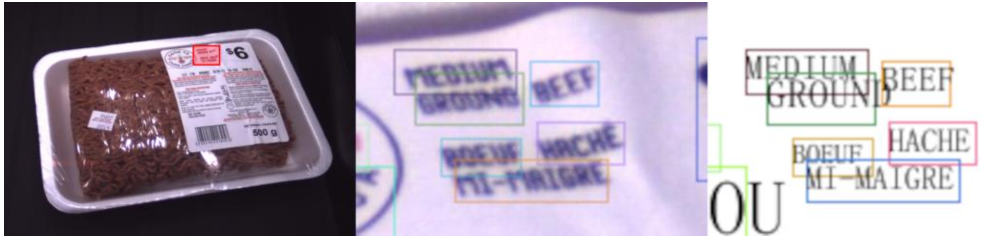




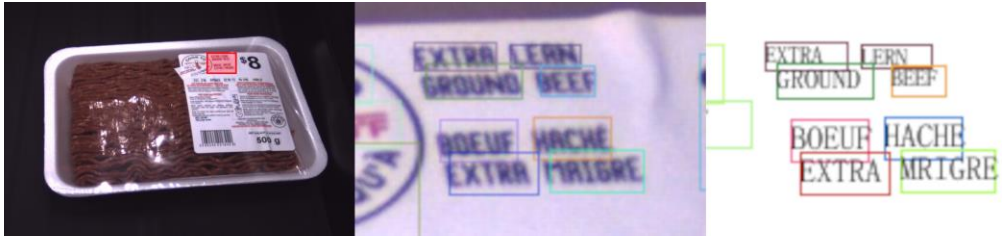




Fine-Grained Product Recognition with Image and OCR



Medium Ground Beef



Extra Lean Ground Beef

Interpretability of Multimodal Product Recognition Models

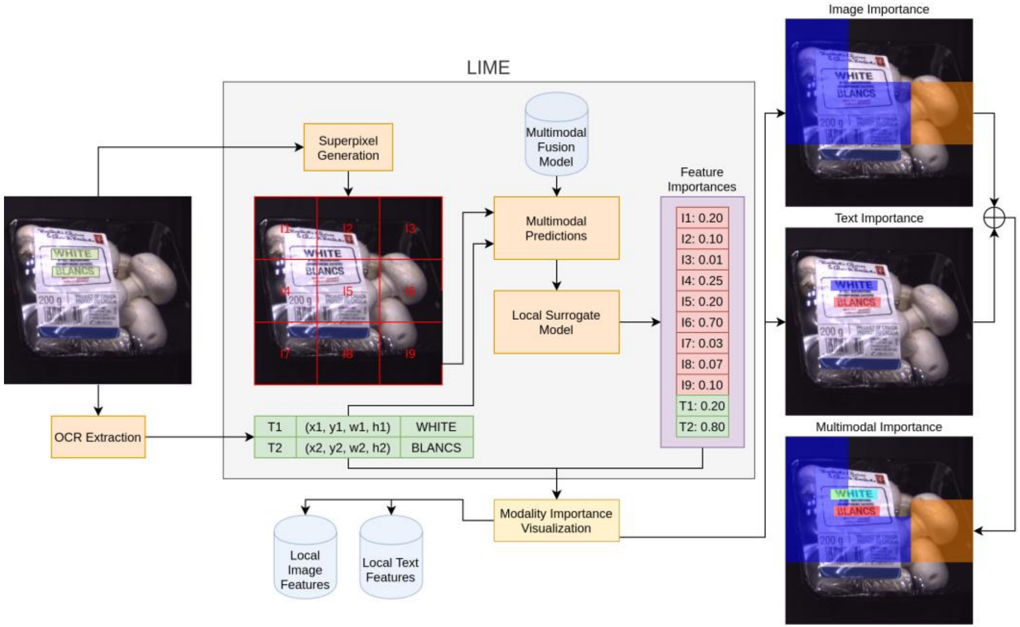
- Complex and difficult to visualize predictions of multimodal product recognition models
- Understanding their behaviour and limitations are key for performing debugging and evaluation before deployment
- Goal: Provide techniques and tools for machine learning experts/developers/stakeholders to debug and assess their multimodal models during development and deployment

Explaining Multimodal Product Recognition Models

We present following:

- A local explanation approach using LIME with multimodal data (Image and OCR) to explain predictions for different samples
- An approach that aggregates the local explanations and provides global explanations for each class
- Demonstrating the utility of our approach using three multimodal models with a fine-grained grocery product dataset

Local Explanations with LIME using Multimodal Data



LIME

Superpixel Generation

Multimodal Fusion Model

Multimodal Predictions

Local Surrogate Model

Feature Importances

I1: 0.20
I2: 0.10
I3: 0.01
I4: 0.25
I5: 0.20
I6: 0.70
I7: 0.03
I8: 0.07
I9: 0.10
T1: 0.20
T2: 0.80

OCR Extraction

T1	(x1, y1, w1, h1)	WHITE
T2	(x2, y2, w2, h2)	BLANCS

Modality Importance Visualization

Local Image Features

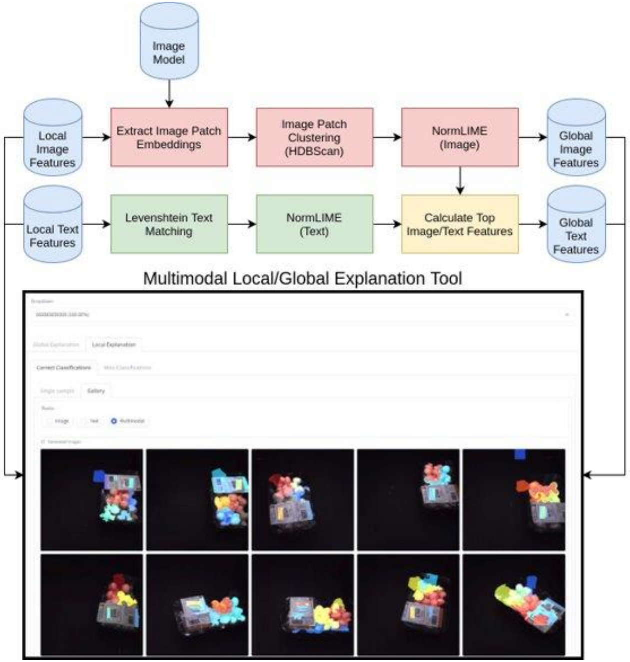
Local Text Features

Image Importance

Text Importance

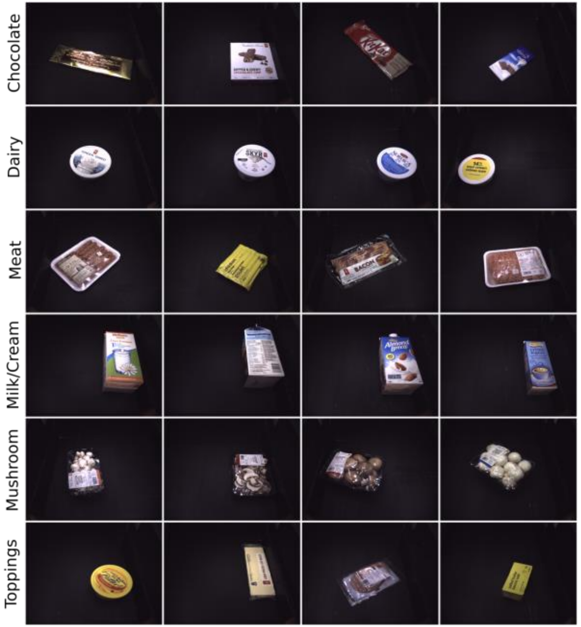
Multimodal Importance

Global Explanations with LIME using Multimodal Data

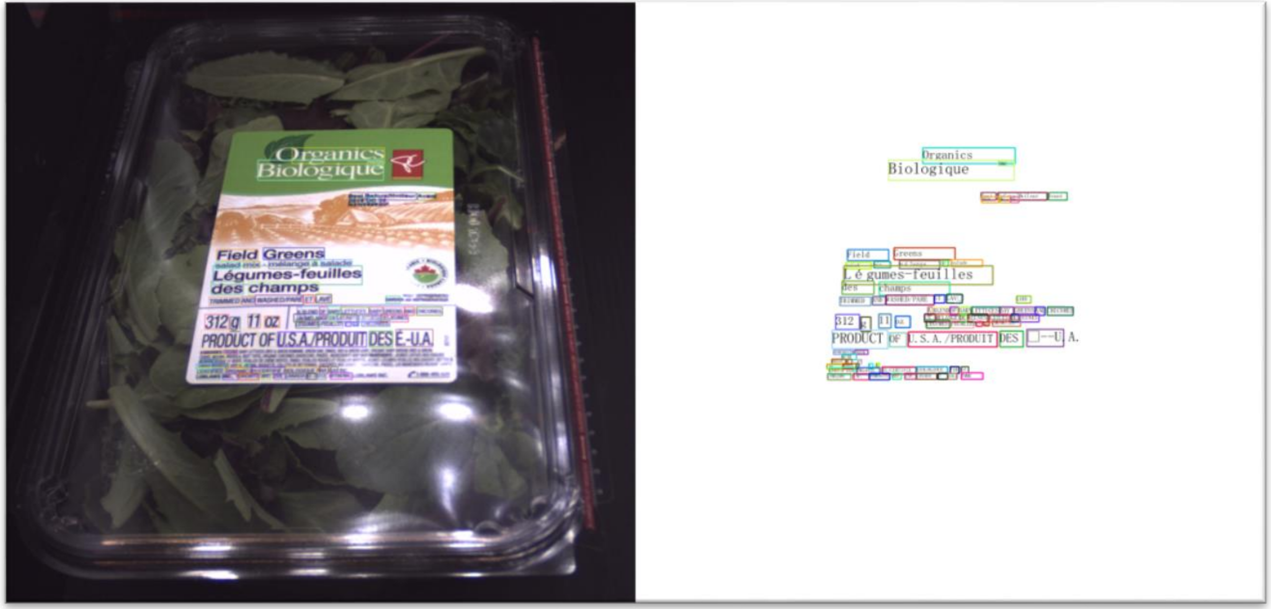


Experimental Setup - Dataset

- 256 classes, each with 100 training and 50 validation samples
- Real-world environment

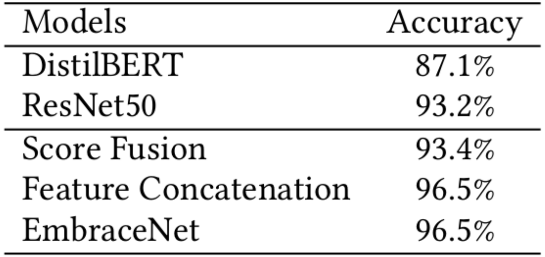


Experimental Setup - Dataset



Experimental Setup - Models

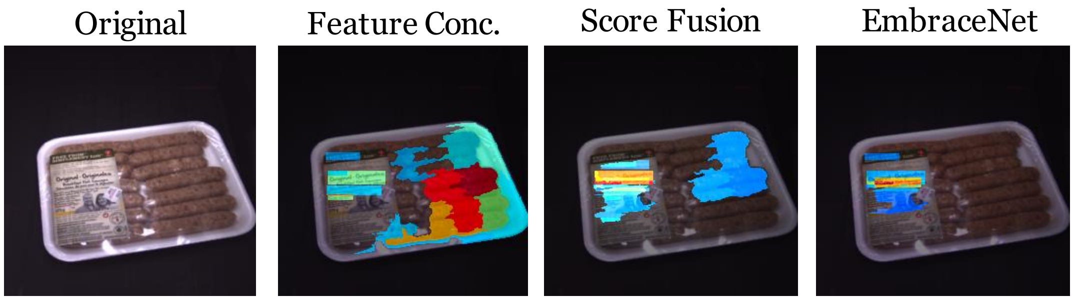
- Unimodal models: ResNet50 and DistilBERT
- Multimodal models: Score Fusion, Feature Concatenation, EmbraceNet



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

Classification Results

Results - Local Explanations



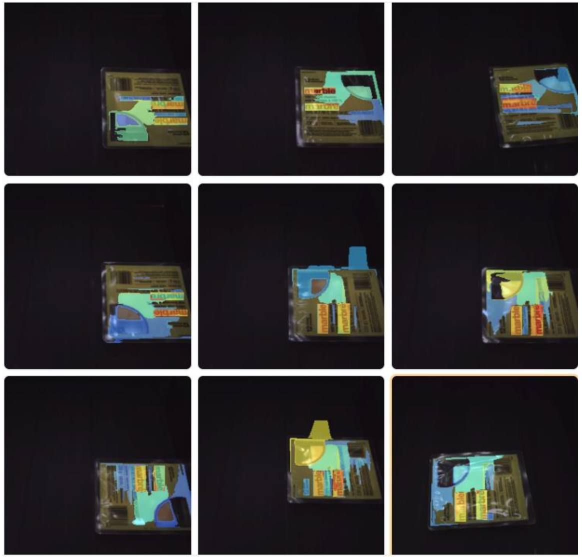
Original

Feature Conc.

Score Fusion

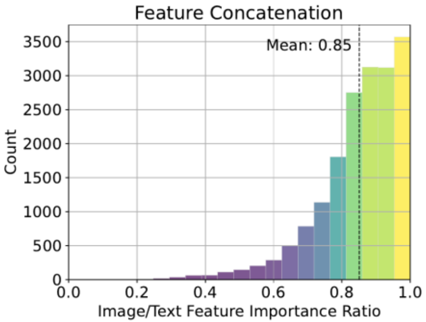
EmbraceNet

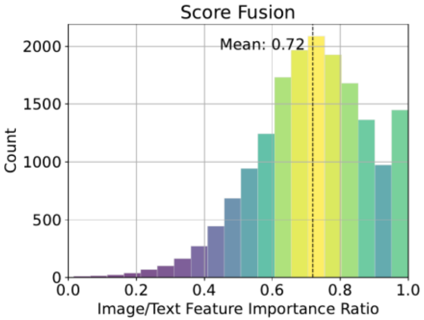
Visualization of local explanations for multimodal models

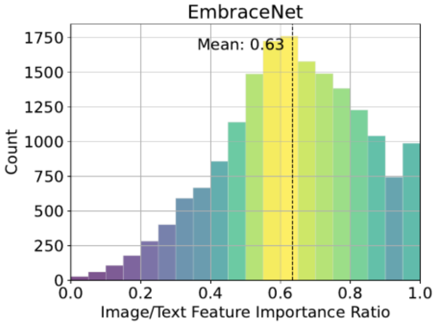


Screenshot from Explanation Tool

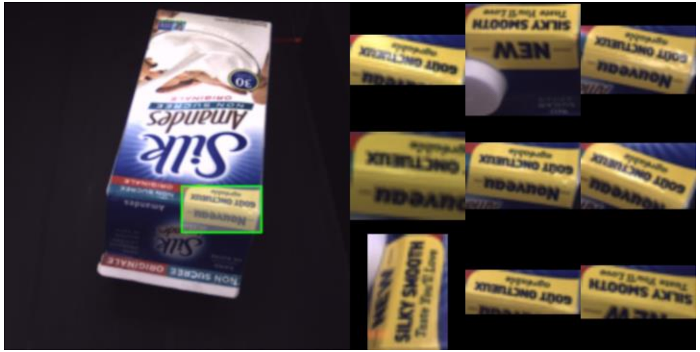
Results - Local Explanations



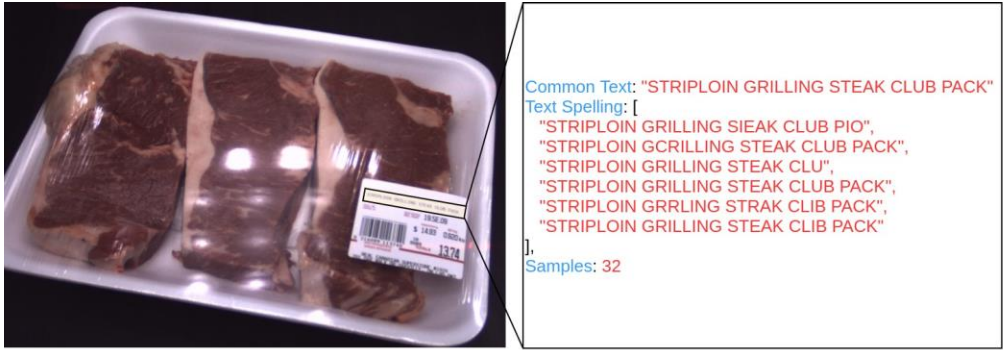




Results - Global Explanations



Top global image feature



Common Text: "STRIPLOIN GRILLING STEAK CLUB PACK"
Text Spelling: [
"STRIPLOIN GRILLING SIEAK CLUB PIO",
"STRIPLOIN GCRILLING STEAK CLUB PACK",
"STRIPLOIN GRILLING STEAK CLU",
"STRIPLOIN GRILLING STEAK CLUB PACK",
"STRIPLOIN GRRLING STRAK CLIB PACK",
"STRIPLOIN GRILLING STEAK CLIB PACK"
],
Samples: 32

Top global text feature

Summary and Future Work

- Summary
  - Present an approach for local and global explanations for product recognition models using image and OCR data
  - Shown the utility of our approach by comparing three different multimodal models
  - Applicability in other domains: Online retail, document classification
- Future work
  - Dataset will be available
  - Reduce computational requirements
  - User study

1977

UNIVERSITY
OF SKÖVDE

Thank You!