A Visual RAG Pipeline for Few-Shot Fine-Grained Product Classification

Supplementary Material

Additional analyses can be found in the supplementary Sections 7 to 10. Section 7 shows an example of OCR-extracted text used by the Fine-Grained Classification (FGC) task solved by Text Classification, see Section 5.1.2. The structured output of a VLM response is depicted in Section 8. Further illustrations of erroneous predictions yielded by the Visual RAG pipeline are presented in Section 9. Section 10 includes examples of defective product segmentation masks. The segmentation masks are necessary for the initial step *Preprocessing* in the Visual RAG pipeline.

7. OCR-extracted Text

Figure 8 shows an image with the corresponding OCR-extracted product description text by using PyTesseract [12].

Diamant Gelierzucker 1:1 1-kg-Packung

Figure 8. Image and its OCR-extracted product description text.

8. Structured Output of VLM Response

The following is the structured output of the VLM *GPT-4o-mini_2024-07-18* for the query image shown in Figure 2:

```
brand='Lorenz'
price=0.99
regular_price=1.87
relative_discount=47
absolute_discount=None
product_category=['Saltletts Sticks']
GTINs=['04018077683015',
       '04018077686719']
weight_number=250.0
weight_unit=Gramm
different_sorts=yes
```

9. Examples of False Predictions

Figure 9 shows an image with the GT value and prediction for the target *GTINs*. The false prediction is based on the extraction of text from the image. Figures 10a and 10b show images for which false predictions for target *price* are resulted from the Visual RAG pipeline. Regarding the false predictions for the target *regular price*, Figures 11a and 11b display in each case an image.

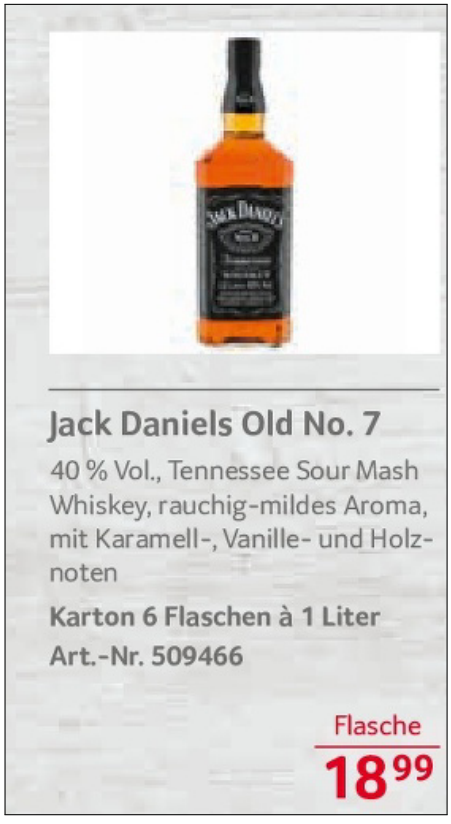
	GTINs
GT	04008100140301
prediction	24000952

Figure 9. Illustration of an image plus GT and prediction values for the target *GTINs*. The false prediction is based on the extraction of text from the image.

10. Defect Product Segmentation Examples

Figure 12 shows an image without the second segmented product. The result shows that only one product image is segmented. The second printed product image is not found. Figures 13 and 14 show images in which only a part of the actual product is segmented. In Figure 13 only the bottles without the beer crate is segmented. In Figure 14, only a part of the product image is the segmentation mask.

(a) Multiple price information.



(b) Price information per bottle.

Figure 10. Illustration of images for which the price predictions are false. The left image contains two information for the price: 0.99€ and a reduced price of 0.96€. The printed price information in the right image is the price per bottle. But the advertisement promotes a carton containing six bottles.

(a) Recommended retail price.

(b) Reference weight price.

Figure 11. Illustration of images for which the regular price predictions are false. Figure 11a shows the print of the recommended retail price. The reference weight price of 1kg is 3.58€ for the advertisement image in Figure 11b.

Figure 12. Missing second printed product image for the segmentation mask.

Figure 13. Segmentation mask includes the bottles without the beer crate.

Figure 14. Only a part of the product image is segmented.