Text and Image Multimodal Dataset for Fine-Grained E-Commerce Product Classification

Ajibola Obayemi[1](✉) and Khuong Nguyen[2]

[1] University of Brighton, Brighton, UK
e.a.obayemi@brighton.ac.uk
[2] Royal Holloway, University of London, Egham, UK
Khuong.Nguyen@rhul.ac.uk

1 Introduction

Fine-grained classification is a challenging task that aims to reduce the misclassification errors in the visual classification of similar image samples. Multimodal learning can improve results by combining text and image data. This approach can help minimise misclassification errors caused by intra-class variability and inter-class similarity. For example, consumer products such as grocery products and printer cartridges may have several product variations that are similar enough to make it hard to classify them using SOTA Computer Vision and NLP models [1–3].

In recent research works, multimodal learning has been shown to effectively improve the results of fine-grained classification tasks when compared to the unimodal alternatives [4–6]. A rich and well-curated dataset has significant potential to push forward the research in this area [3–7]. However, such datasets are scarce and difficult to curate, due to the laborious effort in data collection and the challenge of consistency and accuracy of the labels across different modalities.

The main contributions of our paper are: (1) We created a new multimodal dataset with 17,000 image-text pairs. (2) We propose a generalised pipeline for collecting text and image multimodal datasets to simplify the data collection process and encourage more researchers to curate such datasets. (3) We provide the baseline results using a CNN-based unimodal architecture (ResNet-152) and a text and image multimodal architecture (CLIP & MultiModal BiTransformers) to quantitatively demonstrate how the fusion of text and image modalities work to improve the results.

2 Methods

2.1 Data Collection

Over two months, 17,000 images were captured using a Canon DSLR camera, and a photo lightbox (for consistent lighting and image quality). Using the EOS

Q. Z. Sheng et al. (Eds.): ADMA 2024, LNAI 15392, pp. 356–358, 2025.
https://doi.org/10.1007/978-981-96-0850-8

digital SDK software, we automated several aspects of the image capture process, such as lens focus and shutter control. The captured images were saved locally, and their corresponding metadata was stored in a Microsoft SQL Server database. Additionally, to improve the quality and consistency of the dataset, we pre-processed the images by applying centre cropping to remove unnecessary background elements and focus on the object of interest. To create our text & image multimodal dataset, we extracted the text from the product labels using an OCR pipeline (based on a modified version of CRAFT & CRNN + CTC). We generated image and text pairs by combining the images and extracted texts. The dataset is publicly available at https://github.com/multimodal-research/TAIMD-17k.

2.2 Unimodal and Multimodal Architecture

For the unimodal experiment, a pre-trained ResNet-152 model was fine-tuned for image classification. The model was trained on the dataset with early stopping to prevent overfitting, and training was conducted on two NVIDIA RTX 4090 GPUs.

For the multimodal experiment, we used Multimodal BiTransformers (MMBT) for vision-language modelling. CLIP was used for image encoding and BERT for text encoding, and MMBT was trained on the dataset using Hugging Face and PyTorch. Training was also conducted on two NVIDIA RTX 4090 GPUs.

3 Results

We evaluate the results from our experiments using the F1-score, precision and recall metrics and compare the results from the image based unimodal architecture with the results from the image and text based multimodal architecture. We observed that the multimodal model performed better than the unimodal model on our dataset, achieving a 75% precision, 71% recall and 67% F1-score respectively on our dataset (see Table 1).

Table 1. The performance of CNN-based Image Classification and Multimodal Image & Text Classification. Multimodal architecture outperformed unimodal architecture in all metrics.

Method	Precision	Recall	F1-Score
Unimodal (ResNet-152)	0.70	0.61	0.59
Multimodal (MMBT + CLIP)	**0.75**	**0.71**	**0.67**

4 Conclusion

This paper introduces a novel multimodal dataset for fine-grained e-commerce product classification. It comprises 17,000 images categorised into 31 distinct product classes. We further present a generalised pipeline for collecting and annotating text-image multimodal datasets. This pipeline utilises OCR to extract and annotate the data, generating a large collection of image-text pairs.

Furthermore, we compare the unimodal and multimodal architectures applied to our dataset, demonstrating a significant improvement in performance with the multimodal approach. Utilising CLIP and MMBT based architecture, we achieve up to a 10% increase in precision, recall, and F1-score compared to the unimodal architecture. This dataset is publicly available, aiming to contribute to research and understanding of multimodal learning for fine-grained product classification.

The results of our experiments demonstrate that there is potential for improvement in fine-grained product classification using multimodal learning. In the future, We plan to explore co-learning which is a relatively new area of multimodal learning to further improve fine-grained product classification.

References

1. 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). IEEE (2016)
2. Xuan, Q., et al.: Evolving convolutional neural network and its application in fine-grained visual categorization. IEEE Access **6**, 31110–31116 (2018)
3. Zahavy, T., et al.: 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, issue number 1 (2018)
4. Fu, J., et al.: CMA-CLIP: cross-modality attention clip for text-image classification. In: 2022 IEEE International Conference on Image Processing (ICIP). IEEE (2022)
5. Jiang, X., et al.: Delving into multimodal prompting for fine-grained visual classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, issue number 3 (2024)
6. Lu, Z., et al.: StreamSketch: exploring multi-modal interactions in creative live streams. In: Proceedings of the ACM on Human-Computer Interaction, vol. 5, issue number CSCW1, pp. 1–26 (2021)
7. Kim, E., McCoy, K.F.: Multimodal deep learning using images and text for information graphic classification. In: Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (2018)