Multimodal product data classification

Names

Ateliers d’apprentissage automatique
Université de Strasbourg
January 26, 2024

1 Introduction

The advancement of e-commerce has brought with it a myriad of opportunities and challenges, particularly in the realm of product cataloging and classification. As online marketplaces expand, the task of efficiently categorizing a vast number of products becomes increasingly complex and critical. This challenge is exemplified in the case of Rakuten, a global leader in e-commerce, known for its expansive marketplace that hosts a diverse range of products.

Rakuten, founded in Japan in 1997, revolutionized online shopping with its marketplace concept and has since grown into a major e-commerce platform (Brian, 2022). It boasts a community of over 1.3 billion members (Statista, 2023) and a wide array of services including communications, financial services, and digital content. The Rakuten Institute of Technology (RIT), serving as its research and innovation arm, focuses on areas like computer vision, natural language processing, and human-computer interaction to enhance and innovate within the e-commerce space.

This report addresses a specific challenge posed by Rakuten France: the large-scale multimodal (text and image) classification of products into predefined type codes. In an online marketplace, products are typically accompanied by titles, descriptions, and images. The classification of these products into correct categories is crucial for various aspects of e-commerce, such as personalized recommendations, search optimization, and efficient query processing. However, the task is not straightforward due to the sheer volume of products, the variety of classes, and the common issue of unbalanced data distributions in large catalogs.

The primary objective of this challenge is to develop a model capable of accurately categorizing products based on their textual and visual information. This involves predicting the appropriate product type code for each item in Rakuten France's catalog, a task that requires an intricate understanding of both the textual and visual characteristics of the products. The complexity of this challenge is heightened by the intrinsic variability and potential inconsistencies in product labels and images.

To facilitate this endeavor, Rakuten France has provided a dataset comprising approximately 99,000 product listings in a CSV format, which includes both training and test sets. The dataset contains product titles, detailed descriptions, images, and corresponding product type codes. The benchmark for this challenge is the weighted-F1 score, a metric that balances precision and recall, and is particularly useful in scenarios with uneven class distributions (Humphrey et al., 2022).

In summary, this report delves into the development of a multimodal classification system for Rakuten's extensive product catalog. The focus is on leveraging both textual and visual data to achieve accurate and efficient product categorization, a task essential for enhancing the user experience and operational efficiency in the dynamic world of e-commerce.

2 Data

3 Method

3.1 Data Preparation

3.2 Visualisation

3.3 Feature Extraction

3.4 Modelling

3.4.1 Integrating CLIP for Enhanced Multimodal Product Classification

Central to our approach is the integration of OpenAI's CLIP (Contrastive Language–Image Pretraining) model. CLIP embodies a novel paradigm in artificial intelligence, harmonizing the interpretation of visual and textual data through a multimodal learning framework.

CLIP operates on a dual-encoder structure, comprising an image encoder and a text encoder. This architecture is instrumental in processing and correlating visual and textual inputs. The model's training utilizes a contrastive learning method, where it is exposed to numerous images and their corresponding textual descriptions, drawing from a diverse and extensive internet-sourced dataset. This training enables CLIP to develop a nuanced understanding of the intricate relationships between text and images.

In the realm of product classification, CLIP's integration offers a transformative potential. Our model leverages CLIP's proficiency in associating product images with their textual descriptions, such as titles and detailed narratives. This synergy allows for a more nuanced and accurate classification of products into their respective categories.

3.5 Evaluation

4 Evaluation

5 Conclusion

References

Brian. (2022, July). Top Largest eCommerce Companies In The World (By Revenue) - AovUp (formerly Woosuite). https://aovup.com/stats/ecommerce-companies/

Humphrey, A., Kuberski, W., Bialek, J., Perrakis, N., Cools, W., Nuyttens, N., Elakhrass, H., & Cunha, P. A. C. (2022). Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth. *Monthly Notices of the Royal Astronomical Society: Letters*, *517*(1), L116–L120. https://doi.org/10.1093/mnrasl/slac120

Statista. (2023, August). Rakuten Group's number of member IDs 2014-2023. https://www.statista.com/statistics/223349/rakuten-members/