EasyChair Preprint

№ 9469

Multimodal Classification in E-Commerce: a Systematic Review

Karan Mehta, Yuvraj Maroo, Ria Lele and Pragati Khare

December 14, 2022

Multimodal Classification in E-Commerce: A Systematic Review

Karan Mehta
Student of Computer Science and Business System,
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
karanm45@gmail.com

Yuvraj Maroo
Student of Computer Science and Business System,
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
yuvrajmaroo@gmail.com

Ria Lele
Student of Computer Science and Business System,
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
ria.lele01@yahoo.com

Pragati Khare
Professor of Computer Engineering department,
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
pragati.khare@nmims.edu

Abstract – **The 21st century is the Internet's era of big data. E-commerce has evolved swiftly, and online shopping has become prevalent. E-commerce platforms, including Amazon, and eBay are swamped with numerous product categories. These platforms must categorize the products to assist product management and recommendation but doing so manually can be highly expensive. Machine learning models that can decrease the expense and time of hiring human editors are required due to the enormous volume of new products being uploaded every day and the dynamic nature of the categories. In recent years, ML-based unimodal classification technology, such as SVM and DL, has been widely implemented in the business sector to categorize e-commerce products. Despite the positive findings published thus far, it is thought that the performance of unimodal-based algorithms can be further enhanced by incorporating multimodal product information. Thus, this paper systematically reviews and explores multimodal models used in E-commerce product classification. There are umpteen research articles published in the scientific domain from 2017 to 2022. A comprehensive literature review is missing which can help researchers and E-commerce platforms to understand and utilize the most accurate classification models.**

Keywords: Classification, E-Commerce, Multimodal Architecture, Image Classification, Text Classification

I. INTRODUCTION

E-commerce has revolutionized the retail industry by streamlining the processes of purchasing and delivering products to consumers. The digitization of the fashion industry has resulted in a greater variety of products available to consumers in shorter amounts of time[1]. Customers' interest is piqued using digital technologies in online stores [2].

The categorization of products is a crucial problem for online retailers. Metadata, which includes information on a product's title, picture, color, weight, and other characteristics, is normally assigned manually by the seller for each product. A product is often listed in several categories once it is submitted to an e-commerce website. Product categorization enables e-commerce websites to offer customers a better buying experience, for instance by facilitating quick product catalogue searches or by creating recommendation systems. Internal taxonomies (for business needs), public taxonomies (for items like groceries and office supplies), and the product's shelf (a collection of goods displayed collectively on an e-commerce website) are a few instances of categories. These categories change over time to improve search performance and consider unique events like holidays and sporting occasions [3].

E-commerce websites frequently employ editors and crowdsourcing platforms to categorize products to meet these criteria. However, machine learning techniques for product classification are interesting to reduce the time needed to classify products and the cost associated with doing so because of the large number of new products submitted every day and the dynamic nature of the categories. As a result, there is a growing problem in e-commerce environments related to accurately classifying products. Manual classification can be inefficient and erroneous due to the large number of tags on e-commerce products and the various textual descriptions of product titles [4]. As a result, accurately categorizing things becomes a major problem in e-commerce domains.

In the past, unimodal algorithms have demonstrated promising results as one solution to the classification problem; however, it is believed that their performance can be enhanced by using multimodal product data. This paper is thus focused on reviewing multimodal approaches in e-commerce [5] .

Classifying images is an important part of computer vision and has numerous real-world applications. Image classification is used in many different kinds of information

systems today, from facial identification and object detection to age verification and content restriction for mature audiences [2]. Businesses can benefit from text-classification since it provides insight into the types of keywords that will resonate most strongly with their target audience. This strategy of strategically inserting keywords is supposed to be highly effective because shoppers using any e-commerce platform may easily find the exact article, they're seeking for by entering relevant search terms into the relevant search fields.

The remaining sections of the paper are divided into three parts. Section II describes similar work pertinent to our study. In Section III, we describe the experimental outcomes and discussions. Finally, section IV culminates the paper with conclusions.

II. LITERATURE REVIEW

This section includes the literature review which defines the different existing algorithms implemented for classification of products in e-commerce using the multimodal approach. Multimodal product classification generally consists of mapping of text-based title description to the image of the product into the correct category [6]. Some e-commerce websites represent categories by numbers such as the Rakuten France multimodal data set [5] [7]. Several different approaches have been utilised for the task at hand.

A method for categorizing fine-grained fashion photographs on deep fashion datasets has been created using a pre-trained Convolutional Neural Network (CNN) and a Siamese network to build a distance function for comparing the similarity of fashion images [2]**.** Transfer learning was used to construct a fashion industry-specific deep learning technology for speedy searches and accurate product classifications. Experiments demonstrated that the XCEPTION deep neural network model beats all others.

Several decision level fusion techniques for multi-modal product classification using neural network classifiers for text and images have been presented. One trains state-of-the-art input-specific deep neural networks for each input source, demonstrating the possibility of fusing them into a multi-modal architecture and training a unique policy network that learns to choose between them [3]**.** On a real-world, large-scale product classification dataset obtained from Walmart.com, the multi-modal network achieves a 2% boost in classification accuracy, which has a considerable impact when deployed in production.

Combining characteristics derived by various neural network models from textual (CamemBERT and FlauBERT) and visual (SE-ResNeXt-50) data using straightforward fusion approaches, a new multi-modal model for commercial product classification is has been built. The technique performs much better than unimodal models. Upon experimenting with several fusing strategies, it was determined that the most effective technique for combining the individual embeddings of a unimodal network is based on the combination of concatenating and averaging the feature vectors [8]. Each modality compensated for the shortcomings of the others, implying that increasing the number of modalities can be an effective way of improving the performance of multi-label and multimodal classification challenges. The greatest result was reached with an F1-Score of 92.67% on the test set employing a 90/10 split of the dataset using the average fusion operation across all fusion layers.

In addition, a multimodal item categorization (MIC) system based purely on the Transformer has been proposed for both text and image processing. On a multimodal product data set acquired from a Japanese ecommerce behemoth, a new picture classification model based on the Transformer was evaluated, and several methods of fusing bi-modal information were explored. The training of the Transformer-based image classifier is four times faster than ResNet-based classifiers, according to experimental results using real-world industry data [1]**.** In addition, it was discovered that a cross-modal attention layer is necessary for the MIC system to obtain performance advantages above text-only and image-only models. The best macro-F values for the three genres Beverages (B), Appliances (A), and Men's Fashion (M) were F1(B) = 0.729, F1(A) = 0.740, and F1(M) = 0.815, respectively, using the Fusion cross-modal model.

Using a Neural Network-trained hierarchical classifier, the method first compares the results to both classifications, i.e traditional methods and proposed method, then selects a set of binary classifiers for each level to lower the likelihood of misclassification at the top levels as that's what inevitably leads to an incorrect label assignment [9] **.** Since all the images used to train the classifier were acquired from products often seen in online retailers, the study's emphasis was on e-commerce.

Overcoming problems with categorization, product recognition, product suggestion, and image-based search led researchers to develop a transfer learning strategy using visual geometry group-19 (VGG-19) and Inception V3. Kaggle's Fashion dataset was used to conduct the experiment [10] . Overall, the fashion industry has a high degree of accuracy when classifying its products, but there are occasional outliers. Overall, the classification accuracy of fashion products is good, but it occasionally exhibits misclassifications of products.

Cross-Modality Attention Contrastive Language-Image Pre-training (CMA-CLIP), a new multi-modal architecture to jointly learn the fine-grained inter-modality relationship was introduced in. It leverages the pretrained CLIP, a two-stream method, to close the inter-modality gap at the global level [11] **.** Then, to capture the fine-grained interaction between text tokens and image patches, we add a sequence-wise attention module, which is a transformer like most one-stream approaches. CMA-CLIP outperforms the state-of-the-art

technique by 5.5% on the Fashion-Gen dataset, achieves competitive performance on the Food101 dataset, and performs on par with the state-of-the-art method on the MM-IMDb dataset.

K3M is a revolutionary strategy that incorporates the knowledge modality into multi-modal pretraining to reduce noise and compensate for the absence of picture and text modalities. The layer of modal encoding extracts the characteristics of each modality. Maintaining the separation of the image and text modalities, a novel initial-interactive feature fusion model is created, and a structure aggregation module is created to bring together data from the image, text, and knowledge modalities [12]**.** K3M has been pre-trained on three tasks: masked object modeling (MOM), masked language modeling (MLM), and link prediction modeling (LPM). Experiment results on a real-world E-commerce dataset and a range of product-based downstream tasks demonstrate that when there is modality-noise or modality-missing, K3M outperforms the baseline and state-of-the-art techniques.

In order to capture the relevant information across product picture and title modalities, an unique Two-stream Hybrid Attention Network (HANet) has been suggested. This network makes use of both key-based and keyless attention mechanisms [13]**.** The trials shown that the HANet delivers cutting-edge performance on the challenge of classifying products at the scale of Amazon, with the Two-stream HANet outperforming all benchmark models with an improvement of (+1.22%) thanks to its carefully thought-out structure.

[14] The research offers a deep neural network architecture for e-commerce non-food product classification. Upon implementing Fine-grained Inception V4 Transfer Learning, the hierarchical architecture attained Top-1 precision of 0.61061. It has been discovered that specific networks with hierarchical architecture can be successfully transferred to similar datasets by transferring the network learned from books to a new book dataset. The transferred model outperformed its pre-trained counterpart.

The categorization approach used in the study is based on Convolutional Neural Networks, which simulate human buying habits by merging product image attributes, summarizing the photos into many tiers of parent and child categories, and convolving the neural network from bottom to top [15]**.** The emphasis is on the global features of the parent categories as well as the local characteristics of the child categories. The classification is learned layer by layer, then weighed to determine the final classification. Experiment results show that the accuracy of the results provided by this method has been greatly improved in both the training and validation sets. Both the training set and the validation set experienced an increase in precision of 4.5% and 4.4%, respectively.

Another study presents a weighted multi-modal strategy for product matching that incorporates both images and text into the training and matching process [16]**.** To develop fine-tuned product embedding, the research blends both transformer and ResNet designs into the siamese network. Experiments reveal that our proposed method outperforms single-modal techniques.

gcForest is a novel machine learning technique used for the problem that employs the cascade forest of decision trees and multi grained scanning mechanisms. After preprocessing the product title data with a word analysis technology, the TFIDF algorithm, we conduct a series of experiments with 4000 samples from 35 product categories [4]**.** The results of the experiment indicate that gcForest outperforms SVM with RBF kernel (86.88%), SVM with linear kernel (89.73%), and CNN (86.86%) in terms of classification accuracy.

III. DISCUSSION

This section examines the numerous facets of e-commerce, including picture classification, text classification, multimodal classification, as well as the business and psychological components of e-commerce websites.

E-commerce has altered the operations and procedures of enterprises. Organizations have been able to obtain numerous possibilities and benefits, which they have utilized to achieve a favorable market position and reputation. In terms of business performance, it has been determined that enterprises can greatly boost their performance because to a larger market, better growth opportunities, lower operating costs, less investment needs, fewer risks, and other factors. While numerous businesses have utilized this business strategy and attained a high growth rate, others have been unable to capitalize on the prospects. [17] In addition to a number of benefits, businesses using Ecommerce encounter a number of restrictions and obstacles. As a result, despite the existence of various potential benefits associated with e-commerce, it has been concluded that enterprises must adopt a crucial strategic strategy in order to accomplish these benefits efficiently, one of which is precisely and automatically classifying products on their websites.

Classification problems entail several challenges [18], [19]. It is essential to minimize the chance of running into issue as much as possible. Few of the commonly faced challenges are listed below -

- Low accuracy of model.
- Mislabeled data and unlabeled data.
- Lack of methods to enhance feature level fusion's precision.
- Scalability and computational capacity issues in Bigdata computation for cloud-based categorization.
- Low-resolution images decrease model accuracy.

Even though unimodal learning has long dominated the field

of machine learning and classification, multimodal algorithms are gaining popularity. Multimodal based architectures have the potential to outperform the conventional unimodal models, according to many of the examined research publications. Big data, class imbalance, and instance level complexity, three of the most challenging issues in classification, haven't yet been adequately addressed in this context. Additionally, we have observed situations in which unimodal datasets might be used to solve multimodal issues and make use of these cutting-edge learning techniques.

Categorization of images has several obstacles. Several difficulties include variance across photographs of the same class, lack of scale variation (image of the object with numerous sizes), and perspective variation where an object may be oriented/rotated in multiple dimensions regarding how the object is shot and recorded in image. Certain objects in a picture are very similar to the backdrop, making it difficult for the image classification system to identify them.

Typically, the function of automated text classification is to assign documents to specific groups using machine learning algorithms. In general, one of the most important tactics is organizing and utilizing the massive amounts of information available in unstructured text format. Text categorization is one area of language processing and text mining research that has gotten a lot of interest. A document is merely a collection of words that have been taken from their more exact context, such as their placement within a phrase or document, according to standard text classification. When term frequency data is present, the vector space solely makes use of the wider document context. As a result, the semantics of words—which may be deduced from their placement and relationships to other words in a sentence—are frequently ignored. However, semantic relationships between words and documents are crucial because approaches that capture semantics typically do better in categorization.

The effectiveness of an E-commerce on how well the products are categorized in the database. The more logically grouped your products are, the faster your customers will be able to find what they're looking for and make informed purchasing decisions.

Classifying products into one of four broad groups, according to factors such as consumer preferences, pricing, and differentiation from competing brands, simplifies the task of finding a suitable purchase. Marketing managers can better cater to their target demographic by categorizing products according to their demands.

Business practices and consumer habits from the previous decade have been affected by the rapid development of internet technology and the advancements in smart gadgets. Electronic Commerce, the concept of supplying and accessing goods and services via online, such as selling goods and services, executing financial transactions, and even scheduling an appointment, has the potential to flourish with the proliferation of internet users.

Offering great customer service, a low price, high-quality product photos and descriptions, a discount code, a money-back guarantee, and free shipping or returns aren't always enough to convince people to make a purchase. E-commerce platforms that have a deeper understanding of user psychology can broaden their reach and increase their revenue. Knowledge of cognitive bias and other psychological characteristics related to consumer behavior can be helpful in running an internet business in addition to smart marketing and a betting on a client base approach. Always keep in mind that the "why" behind any conversion-rate improvement strategy is rooted in some sort of psychological theory. The true secret to eCommerce success is educating customers on their part in the buying process and tending to the website in accordance with psychological principles. If you break down the steps involved in making a choice, you'll find it's very easy to not just exceed the ever-rising expectations of each individual customer, but also get more out of your ecommerce spending.

Collective knowledge of various elements of human emotions, business strategy, technical intricacies, and sector economy can greatly benefit any firm.

IV. CONCLUSION

In this study, we have conducted an extensive literature assessment on a number of other research publications in an effort to identify the multi-modal classification algorithm that is the most effective. When it comes to solving classification issues, one of the most effective strategies is the model that is based on machine learning. In order to increase the amount of revenue generated from online sales; it is vital to correctly categorize your products. In addition, it is helpful to improve not only the experience of the consumer but also the experience of the seller in terms of product classification and search. Incorrectly labelled data points are another significant obstacle that must be overcome before classification can be performed.

The shopping experience is enhanced for users when products are found in the right category as the conversion rate can be improved and so does user retention. Future studies could (1) investigate simultaneously modelling words and images in one Transformer model like FashionBERT, (2) employ self-training to overcome the limitation imposed by the small amount of annotated image data available to the image model. There is also a requirement for a DNN model that would classify products in small quantities. Furthermore, the accuracy can be increased by progressively subdividing categories. Finally, including and using the high-resolution image in the search, as well as for testing and training, can

boost the search's precision.

REFERENCES

[1] L. Chen, H. Chou, Y. Xia, and H. Miyake, *Multimodal Item Categorization Fully Based on Transformer*. 2021. doi: 10.18653/v1/2021.ecnlp-1.13.

[2] S. Bhoir and S. Patil, "Transfer Learning with Deep Neural Networks for Image Classification in the E-commerce Industry," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, 2022, pp. 1–8. doi: 10.1109/I2CT54291.2022.9824903.

[3] T. Zahavy, A. Magnani, A. Krishnan, and S. Mannor, "Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-commerce," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Nov. 2016, doi: 10.1609/aaai.v32i1.11419.

[4] J. Dai, T. Wang, and S. Wang, "A Deep Forest Method for Classifying E-Commerce Products by Using Title Information," in *2020 International Conference on Computing, Networking and Communications (ICNC)*, 2020, pp. 1–5. doi: 10.1109/ICNC47757.2020.9049751.

[5] Y. Bi, S. Wang, and Z. Fan, *A Multimodal Late Fusion Model for E-Commerce Product Classification*. 2020.

[6] S. C. Guntuku, J. T. Zhou, S. Roy, W. Lin, and I. W. Tsang, "Understanding Deep Representations Learned in Modeling Users Likes," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3762–3774, 2016, doi: 10.1109/TIP.2016.2576278.

[7] H. Amoualian, P. Goswami, P. Das, P. Montalvo, L. Ach, and N. Dean, "An E-Commerce Dataset in French for Multi-modal Product Categorization and Cross-Modal Retrieval," 2021, pp. 18–31. doi: 10.1007/978-3-030-72113-8_2.

[8] T. Misikir Tashu, S. Fattouh, P. Kiss, and T. Horvath, *Multimodal E-Commerce Product Classification Using Hierarchical Fusion*. 2022. doi: 10.48550/arXiv.2207.03305.

[9] M. G. Vieira and J. Moreira, "Classification of E-Commerce-Related Images Using Hierarchical Classification with Deep Neural Networks," in *2017 Workshop of Computer Vision (WVC)*, 2017, pp. 114–119. doi: 10.1109/WVC.2017.00027.

[10] B. K. Jha, S. G. G, and V. K. R, "E-Commerce Product Image Classification using Transfer Learning," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 904–912. doi: 10.1109/ICCMC51019.2021.9418371.

[11] H. Liu *et al.*, *CMA-CLIP: Cross-Modality Attention CLIP for Image-Text Classification*. 2021.

[12] Y. Zhu *et al.*, *Knowledge Perceived Multi-modal Pretraining in E-commerce*. 2021. doi: 10.1145/3474085.3475648.

[13] Q. Chen, Z. Shi, Z. Zuo, J. Fu, and Y. Sun, "Two-Stream Hybrid Attention Network for Multimodal Classification," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 359–363. doi: 10.1109/ICIP42928.2021.9506177.

[14] C. S. Islam and M. Alauddin, "A Novel Idea of Classification of E-commerce Products Using Deep Convolutional Neural Network," in *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, 2018, pp. 342–347. doi: 10.1109/CEEICT.2018.8628161.

[15] X. Jin, X. Du, X. Han, H. Sun, and J. Li, "Fine Classification Method of Product Image Based on Multi-Level Convolutional Neural Networks," in *2021 2nd Asia Symposium on Signal Processing (ASSP)*, 2021, pp. 113–117. doi: 10.1109/ASSP54407.2021.00025.

[16] K. Gupte, L. Pang, H. Vuyyuri, and S. Pasumarty, "Multimodal Product Matching and Category Mapping: Text+Image based Deep Neural Network," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 4500–4505. doi: 10.1109/BigData52589.2021.9671384.

[17] A. Andonov, G. Dimitrov, and V. Totev, "Impact of E-commerce on Business Performance," *TEM Journal*, vol. 10, pp. 1558–1564, Nov. 2021, doi: 10.18421/TEM104-09.

[18] A. Chaudhuri *et al.*, "A Smart System for Selection of Optimal Product Images in E-Commerce," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 1728–1736. doi: 10.1109/BigData.2018.8622259.

[19] X. Zhang, "Content-Based E-Commerce Image Classification Research," *IEEE Access*, vol. 8, pp. 160213–160220, 2020, doi: 10.1109/ACCESS.2020.3018877.