MEPC: Multi-level Product Category Recognition Image Dataset

Thanh Long Nguyen[1*], Manh Quang Do[2*†], and Ba Nghien Nguyen[1]

[1] Faculty of Information Technology, Hanoi University of Industry
[2] Faculty of Interdisciplinary Digital Technology, Phenikaa University
longnt.vie@gmail.com, quang.domanh@phenikaa-uni.edu.vn,
nguyenbanghien_cntt@haui.edu.vn

Abstract. Multi-level product category prediction is a problem for businesses providing online retail sector systems. Accurate Multi-level prediction supports the system in avoiding the need for sellers to fill in product category information, saving time and reducing the cost of listing products online. This is an open research problem, which always attracts researchers. Deep learning techniques have shown promising results for category recognition problems. A neat and clean dataset is an elementary requirement for building accurate and robust deep-learning models for category prediction. This article introduces a new image dataset of the multi-level product, called MEPC. MEPC dataset has +164.000 images in the processed format available in the dataset. We evaluate the MEPC dataset with popular deep learning models, benchmark results in a top-1 accuracy score of 92.055% with 10 classes and a top-5 accuracy score of 57.36% with 1000 classes. The proposed dataset is good for training, validation, and testing for hierarchical image classification to improve predict multi-level categories in the online retail sector systems. Data and code will be released at *https://huggingface.co/datasets/sherlockvn/MEPC*.

Keywords: Product category prediction, Category prediction, Multi-level classification, Hierarchical image classification

1 Introduction

E-commerce platforms have become more and more popular over the years. The digital transformation 4.0 further stimulated public interest in e-commerce, resulting in a boom in e-commerce businesses [1,2]. As a result, the e-commerce industry has become more competitive, driving firms to make considerable expenditures to improve their platforms. In recent years, hierarchical classification has emerged as a powerful tool in the online retail industry [3], assisting sellers in auto-filling fast product category information. With the speed-up growth

*These authors contributed equally to this work.
†Corresponding author(s). E-mail(s): quang.domanh@phenikaa-uni.edu.vn.

of online products, efficiently hierarchical classifying these products has become crucial for success in the online retail sector. Applying deep learning and machine learning techniques to retail data enhances the seller experience on e-commerce platforms.

Category Prediction (CP), which aims to recognize the intent categories of given texts, is regarded as one of the most fundamental machine-learning tasks in an e-commerce system [4]. For example, this predicted category information will influence product ranking in the search and recommendation system. Different from the traditional classification [5,6]. Category Prediction is formally categorized as a hierarchical classification problem since categories in most e-commerce websites are organized as a hierarchical tree (we consider the situation that the categories are organized as a hierarchical tree, but not a directed acyclic graph). Figure 1. shows a simplified fragment of one category architecture.

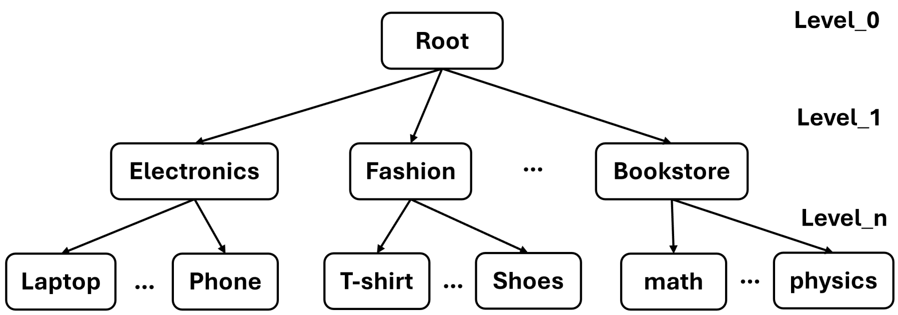

Fig. 1. Visualization of the label hierarchy for hierarchical image classification.

To resolve the problem of category prediction of business then data has always been one of the key points to driving AI-based category recognition research. Different training data may lead to different training results. No matter the target, product image datasets are needed to evaluate the performance of the proposed deep-learning models. In this article, We proposed a new multi-level product category dataset with more than 164.000 images in Table 1. We experimented on many different pre-trained models. The results show that this new dataset is appropriate for improving the model's performance when researching deep learning models to predict multiple categories.

2 Related Work

Category prediction image classification is a classification problem in which hierarchical information related to the classes is given in addition to the image [7]. We list existing known competitions or datasets related to **H**ierarchical **I**mage **C**lassification (HIC) in Table 2. For example, CIFAR-100, ETH Entomologi-

Table 1. Summary of MEPC dataset Statistics.

Statistics	Value	
	Train	Val
Number of images	131,292	32,824
Number of 1^{st} level categories	28	
Number of 2^{nd} level categories	193	
Number of 3^{rd} level categories	659	

cal Collection (ETHEC), CUB-200-2011 (CUB), FGVCAircraft (AIR), Stanford Cars (CAR), and Lego-15 as Figure 2.

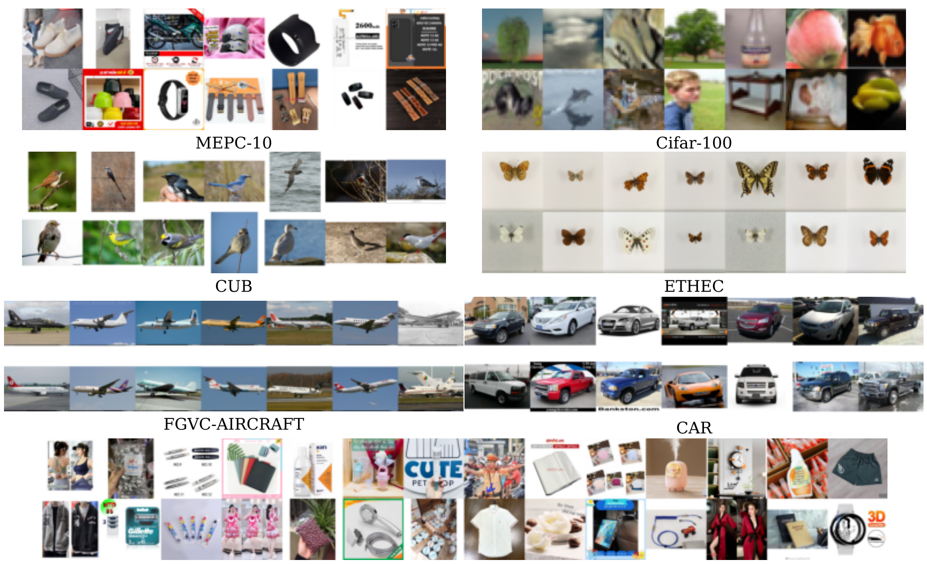

Fig. 2. The summary of random images from the datasets for hierarchical image classification.

The first CIFAR-100 [8] a commonly used benchmark for hierarchical classification, has 20 coarse (super) classes, and each class is associated with five fine classes (i.e., the course class "people" has five fine classes: "baby", "boy", "girl", "man", and "woman"), for a total of 100 fine classes.

The next, for ETH Entomological Collection (ETHEC) dataset [9] comprising images of Lepidoptera specimens with their taxonomy tree. The real-world dataset has variations in terms of the images per category and a significant imbalance in the structure of the taxonomical tree. For CUB-200-2011 (CUB) [10], CUB follows the setting to organize the label hierarchy of birds to 200 species, 122 genera, 37 families, and 13 orders. CUB datasets contain 11,788 images of

Fig. 3. The **categories-cloud** of $1^{nd}/2^{nd}/3^{nd}$ - level keywords of photos. The larger the font size, the more products the corresponding categories.

200 subcategories belonging to birds, 5,994 for training and 5,794 for testing. Each image has detailed annotations: 1 subcategory label, 15 part locations, 312 binary attributes, and 1 bounding box.

Table 2. Table description of hierarchical image datasets.

Name	**Year**	**Public**	**Type**	**Total**	**Classes**	**Levels**
CIFAR-100 [8]	2009	yes	classification	60,000	100	2
ETHEC [9]	2019	yes	classification	47,978	723	4
CUB [10]	2011	yes	classification	11,788	200	4
FGVC-Aircraft [11]	2013	yes	classification	10,200	100	3
CAR [12]	2013	yes	classification	16,185	196	3
Lego-15 [13]	2021	no	classification	4,688	15	3
MEPC-10 (our)	2024	yes	classification	2,192	10	3
MEPC-1000 (our)	2024	yes	classification	164,117	1000	3

The next dataset, FGVC-Aircraft [11] contains 10,200 images of aircraft, with 100 images for each of 102 different aircraft model variants, most of which are airplanes. Each image's main aircraft is annotated with a tight bounding box and a hierarchical airplane model label. Aircraft models are organized in a three-level hierarchy. Stanford Cars (CAR) [12] dataset consists of 196 classes of cars with a total of 16,185 images, taken from the rear. The data is divided into almost a 50-50 train/test split with 8,144 training images and 8,041 testing images. Categories are typically at the level of Make, Model, and Year. CAR has image sizes are 360×240. The last dataset, Lego-15 [13] is a Lego image dataset consisting of 3000 synthetic images and 1688 real images in 15 classes. For synthetic images, there are 200 images in each class. Lego-15 real images, the number of images ranges from 70 to 150 in each class.

The above datasets are all used to optimize deep learning models for hierarchical image classification tasks. However, the weakness of the above datasets is that they are not suitable for predicting product categories on e-commerce platforms. In this article, we introduce a new dataset named **M**ulti-level **E**-commerce **P**roduct **C**ategorization image datasets (**MEPC**). **MEPC** is a product hierarchy image dataset focused on product e-commerce for hierarchical image classification tasks.

MEPC dataset has multiple backgrounds to increase the variety of real images. We also evaluate the impact of **MEPC** dataset with EffcientNet, ResNet, VGG, and MobiNet series models. The results benchmark models with a top-1 accuracy score of 92.055% for **MEPC-10** and a top-5 accuracy score of 57.36% for **MEPC-1000**. We hope that the introduced dataset will be of good assistance in fine-tuning the model for image classification in the online retail sector systems.

- Object {7}
  - Motobike {2}
    - Motobike acessory {2}
      - Motobike wet cover {130}
      - Sticker, logo and badge {278}
    - Helmet and helmet acessory {191}
  - Men's shoes {1}
  - Phone and phone accessory {1}

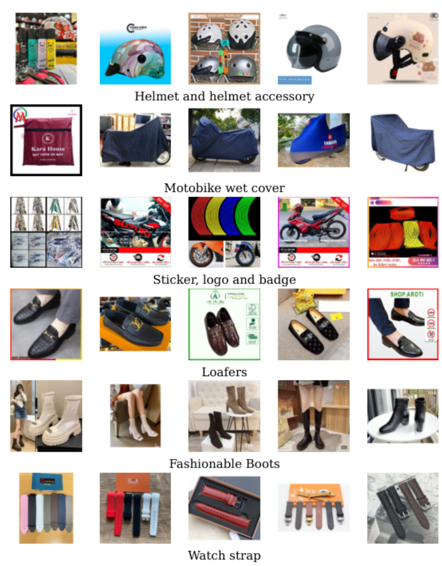

Fig. 4. MEPC dataset overview. An example of random categories of the MEPC dataset. It shows a 2^{nd}-level category "Motorbike accessory" with 2 different 3^{nd}-level categories.

3 Dataset

In this section, we introduce **M**ulti-level **E**-commerce **P**roduct **C**ategorization (MEPC) dataset.

Dataset description All of the images in our **MEPC** dataset are collected from the Internet data collection method introduced in the research article "End-to-End System For Data Crawling, Monitoring, And Analyzation Of E-Commerce Websites" at ICTA2024 [14]. There are nearly +164,000 images in total. For the practical application scenes, the distribution of image amount is imbalanced data, as shown in Figure **4**.

We conducted important statistical analyses better to understand the structure and characteristics of the data. First, we analyzed the number of categories by hierarchy, as illustrated in Figure **6**. This analysis helps us identify the distribution of categories and detect any imbalances among the levels. We used data visualization techniques to highlight this distribution, making identifying categories with fewer samples easier.

Most e-commerce websites organize product displays in a hierarchical tree structure. We can determine which product groups have the most items by counting the number of categories. In Figure **3**, we generated a "categories cloud" at the category level. The larger the font size, the more products the corresponding category has. Based on the figure, it can be seen that the "Women Fashion" category has many products in the MEPC dataset.

Next, we analyzed the image sizes, presented in Figure **7**, to assess the diversity in image dimensions within the dataset. This information helps us optimize the preprocessing steps and ensure that the deep learning models can perform effectively on this dataset.

Dataset Challenge Previously, several datasets [8,10,12,13] have been constructed for HIC. However, they are relatively plain, i.e., scenarios with real-world complexity are not well represented in these benchmarks. Compared with real scene datasets such as (ETHEC [9] and FGVC-Aircraft [11]), MEPC has more diverse products and richer structures. The images in MEPC are much more challenging in that they are taken from website E-Commerce (Fig. **2**) such as background noise, uneven illumination, diverse image size, and complicated connections between parent-childlen Figure **5**. In the next section, we will present the experimental image classification of this dataset using well-known deep-learning models.

4 Methods

In this study, we used the deep learning models ResNet50[15], VGG16[16], MoBiNet [17] and EfficientNet[18], to evaluate a new dataset. These models have been proven to be highly effective in various image recognition and classification tasks. We will provide a detailed account of how we prepare the data, train the models, and evaluate the results.

First, we resized the images to 224x224 and normalized the image data. Next, we used the Adam [19] optimizer with a learning rate of $1e^{-3}$. When visualizing

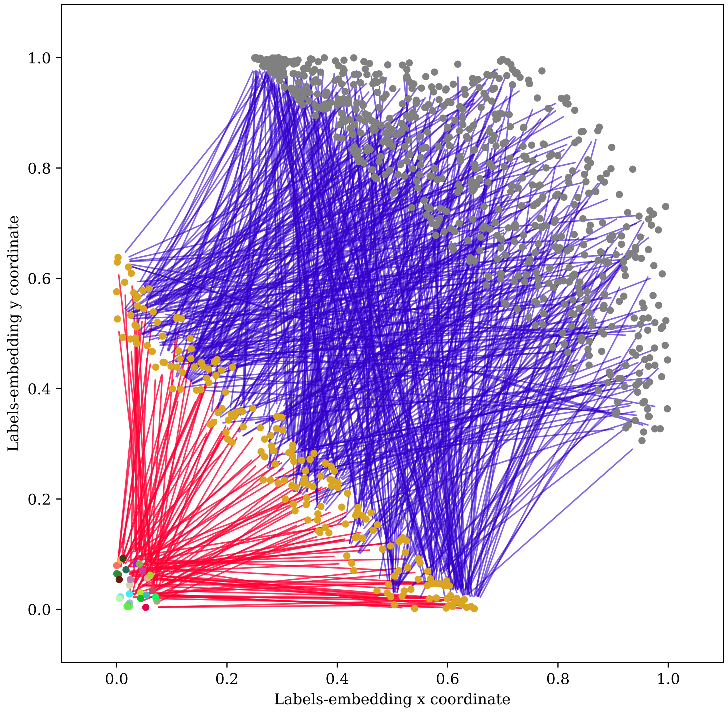

Fig. 5. Label-only embeddings visualizing label connections of the MEPC dataset, with multi-colored labels for level 1, yellow for level 2, gray for level 3, red connections from level 1 to level 2, and blue connections from level 2 to level 3.

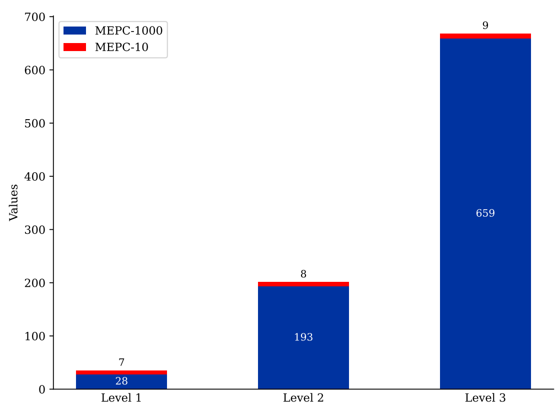

Fig. 6. Statistics of the number of multi-level categories in the two datasets MEPC-10 and MEPC-1000.

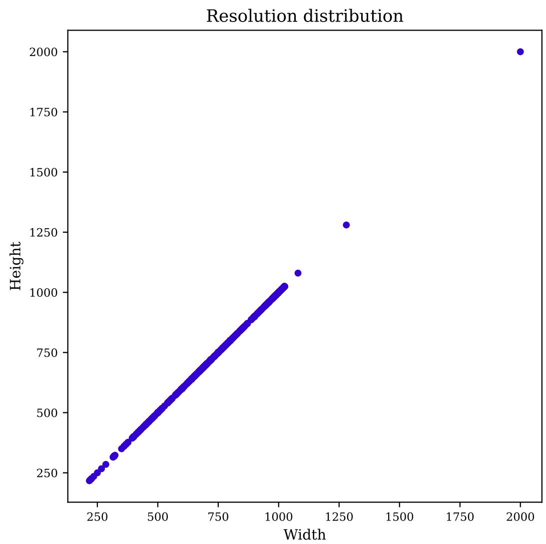

Fig. 7. Statistics of image sizes of MEPC dataset. It can be seen that the images of the dataset are square with various aspect ratios.

Fig. 8. Random visualization of classes in MEPC dataset. It is easy to see that the MEPC data is imbalanced.

the random number of images in each class as shown in Figure 8, we noticed that the MEPC data was imbalanced among the classes. Therefore, we used the focal loss [20] function to penalize the model whenever it incorrectly predicted classes with fewer samples. We used k-fold and metrics such as Top-1 Accuracy for the MEPC-10 dataset to evaluate performance. The experimental results are described in Table 3. In addition, We split the MEPC-10 and MEPC-1000 datasets into 80% training and 20% evaluation sets and used the YOLOv8 [18] model for image classification as follows in Table 4.

Table 3. Benmark K-fold for MEPC-10 dataset (K=5).

Name	MEPC-10^{Top-1}(%)	#**Params**
MoBiNet [17]	79.041% $\pm$ 2.978	4.3M
MoBiNetV2 [21]	72.694% $\pm$ 1.766	3.5M
MoBiNetV3-S [22]	89.224% $\pm$ 1.185	2.5M
MoBiNetV3-L [22]	92.055% $\pm$ 0.602	5.4M
ResNet50 [15]	89.817% $\pm$ 0.897	25.6M
VGG16 [16]	90.457% $\pm$ 1.203	138.4M
EfficientNetB4 [18]	91.37% $\pm$ 1.68	19.5M

Table 4. Experimental results with YOLOv8 Model

Name	**YOLOv8m (17M params)**	
	Top-1 Accuracy (%)	Top-5 Accuracy (%)
MEPC-10	90.87%	-
MEPC-1000	34.41%	57.36%

Based on Tables 3 and 4, we look at the MEPC-10 dataset has a Top-1 Acc of 92.055%. Therefore, MEPC-10 works well for improving deep learning models or for evaluating new deep learning model architectures. However, the MEPC-1000 dataset has a Top-5 Acc of 57.36% so this dataset will be a big challenge for the hierarchical image classification problem.

5 Conclusion

We have released an MEPC dataset for HIC tasks, where the images present numerous challenges such as background noise, uneven lighting, diverse image sizes, complex parent-child linkages, etc. This dataset includes a variety of product images, focusing on predicting product categories for e-commerce systems.

In our study, we also provided an overview of the dataset, visualized the data from various perspectives, and tested this dataset with well-known deep learning models. In the future, we will collect additional semantic descriptions of the images from the real-world or large language models to optimize and improve the classification performance of this dataset.

References

1. M. I. Barat and M. M. Haque, "Small business boom in ecommerce: an in-depth research exploration," *International Journal of Business Management and Finance Research*, vol. 2, pp. 1–14, 01 2024. 1
2. A. Azam and A. M. Ansari, "The emerging role of e-commerce in today's business: A conceptual study," *Asian Journal of Management and Commerce*, vol. 05, pp. 428–439, 05 2024. 1
3. Y. Wei, S. Tran, S. Xu, B. Kang, and M. Springer, "Deep learning for retail product recognition: Challenges and techniques," *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, p. 8875910, 2020. 1
4. A. Cevahir and K. Murakami, "Large-scale multi-class and hierarchical product categorization for an E-commerce giant," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (Y. Matsumoto and R. Prasad, eds.), (Osaka, Japan), pp. 525–535, The COLING 2016 Organizing Committee, Dec. 2016. 2
5. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2324, 1998. 2
6. L. S. Larkey and W. B. Croft, "Combining classifiers in text categorization," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, (New York, NY, USA), p. 289–297, Association for Computing Machinery, 1996. 2
7. X. Zhu and M. Bain, "B-cnn: Branch convolutional neural network for hierarchical classification," 2017. 2
8. A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., 2009. 3, 4, 6
9. A. Dhall, "Eth entomological collection (ethec) dataset [palearctic macrolepidoptera, spring 2019]," 2019. Please cite the dataset and our work if you use it or report results based on it: "Learning Representations For Images With Hierarchical Labels" (https://arxiv.org/abs/2004.00909) and "Hierarchical Image Classification using Entailment Cone Embeddings" (https://arxiv.org/abs/2004.03459). 3, 4, 6
10. C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *The Caltech-UCSD Birds-200-2011 Dataset.* 7 2011. 3, 4, 6
11. S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013. 4, 6
12. J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. 4, 6
13. L. He, D. Song, and L. Zheng, "Hierarchical image classification with a literally toy dataset," 2021. 4, 6
14. e. a. Manh Quang Do, "End-to-end system for data crawling, monitoring, and analyzation of e-commerce websites," 11 2024. 6

15. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. 6, 9
16. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. 6, 9
17. H. Phan, D. Huynh, Y. He, M. Savvides, and Z. Shen, "Mobinet: A mobile binary network for image classification," 2019. 6, 9
18. M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. 6, 9
19. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. 6
20. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018. 9
21. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019. 9
22. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," 2019. 9