MEP-3M: A Large-scale Multi-modal E-Commerce Products Dataset

Delong Chen[1] , **Fan Liu**[1*] , **Xiaoyu Du**[2] , **Ruizhuo Gao**[1] and **Feng Xu**[1]
[1]College of Computer and Information, Hohai University, China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
fanliu@hhu.edu.cn

Abstract

The product categories are vital for the e-commerce platforms due to the core applications on automatic product category assignment, personalized product recommendations, etc. Two key aspects of product classification are multi-modal information and fine-grained understanding. However, recent datasets could hardly support both sides. To address this issue, in this paper, we construct a large-scale **M**ulti-modal **E**-commerce **P**roducts classification dataset MEP-3M, which consists of over 3 million products and 599 fine-grained product categories. Each product is represented with an image-text pair and annotated with hierarchical labels. To our best knowledge, MEP-3M is the first e-commerce products dataset paying attention to the multi-modal and fine-grained aspects concurrently, and its scale achieves the largest in existing E-commerce datasets. We also present the performances of the several methods on this dataset as the baselines, where the best accuracy achieves 90.70%. This dataset is now available at https://github.com/ChenDelong1999/MEP-3M.

1 Introduction

The recent rise of deep learning can be traced back to the creation of ImageNet dataset [Deng *et al.*, 2009] and the revival of deep Convolutional Neural Network (CNN) [Krizhevsky *et al.*, 2012; Li *et al.*, 2021]. Since then, the combination of increasingly complex neural network architectures and increasingly large datasets fundamentally revolutionized the fields of Computer Vision (CV) and Natural Language Processing (NLP). In recent years, the research communities are gradually moving from these single-modal tasks to multi-modal tasks. Large-scale multi-modal datasets, especially vision-language datasets (e.g. Flickr30K [Young *et al.*, 2014], Multi30K [Elliott *et al.*, 2016], MS-COCO [Antol *et al.*, 2015], SBU Captions [Ordonez *et al.*, 2011], WIT [Srinivasan *et al.*, 2021]), have been constructed. These datasets enable us to develop multi-modal models, which learn to utilize the complementary information across different modalities and bring the opportunity to combine the advancements across different fields to further improve the model performance.

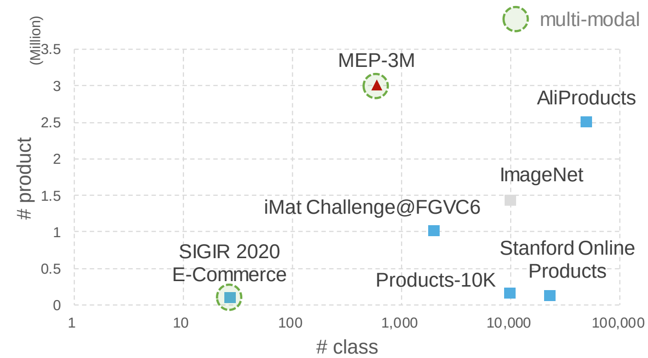

Figure 1: The comparison between our presented dataset and existing public e-commerce product dataset.

Recently, another hot topic in the deep learning field is fine-grained recognition, which aims to discover the subtle differences between different sub-categories, such as birds [Horn *et al.*, 2015], dogs [Sun *et al.*, 2018], cars [Yang *et al.*, 2015], and castles [Anderson *et al.*, 2021]. A lot of fine-grained datasets are created to promote the development of this domain, such as iNaturalist [Horn *et al.*, 2018], Products 10k [Bai *et al.*, 2020], and iMaterialist Fashion [Guo *et al.*, 2019]. Impressively, many e-commerce-related datasets emergence. A possible reason is the construction of this type of dataset can rely on the pre-defined hierarchical categorization information (e.g., Stock Keeping Unit, SKU).

However, recent e-commerce datasets only focus on one aspect from multi-modal or fine-grained without integrating them together. In this paper, we construct a large **M**ulti-modal **E**-commerce **P**roducts classification dataset named MEP-3M, which provides multi-modal and fine-grained data. It is collected from several Chinese large E-commerce platforms and consists of over 3 million image-text pairs of products and 599 classes. As demonstrated in Fig. 1, MEP-3M consists of the largest number of products, even compared with the single-modal E-commerce product datasets. Its scale is far better than the existing multi-modal dataset. The key characteristics of MEP-3M are summarized as follows:

*Contact Author

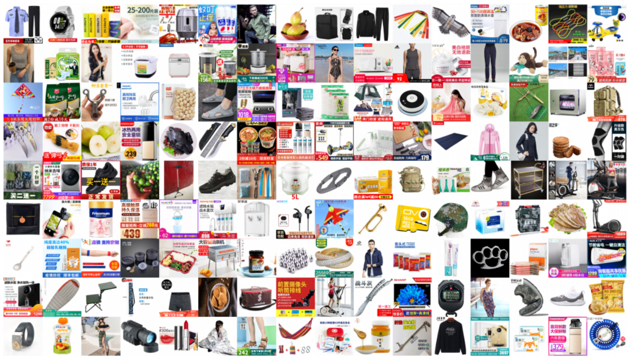

Figure 2: Images randomly selected from the MEP-3M dataset. Our dataset covers a wide range of e-commerce products.

Figure 3: The text cloud (after jieba text segmentation) of product title in the MEP-3M dataset. The text size corresponds to the appearance frequency.

- **Large-scale.** MEP-3M dataset consists of over 3 million product samples in total. Each sample consists of an image-text pair, resulting in 3,012,959 images and 156,069,329 characters. The entire dataset takes approximately 76GB of storage.
- **Hierarchical-categorized.** Three levels of the label are given. There are 14 classes (first level), 599 sub-classes (second level), and 13 sub-classes have further subsubclasses (third level). The illustration for the hierarchical categorization of MEP-3M can be found in Fig. 4.
- **Multi-modal.** Each product has both image and Chinese label and title. Some image samples and the text cloud of the titles are given in Fig. 2 and Fig. 3.
- **Fine-grained.** There are a total of 599 sub-classes, and many of them are fine-grained (e.g., different types of fruit, meat, shoes, clothes, etc.). Many samples are visually similar but belong to different sub-classes, as shown in Fig. 6.
- **Long-tailed.** MEP-3M is highly imbalanced. Some sub-classes in the dataset have more than 90k samples, while some classes have around 30 samples. The distribution is shown in Fig. 5.

Moreover, we also present some baselines on MEP-3M. We test two popular multi-modal learning models: Low-rank Multimodal Fusion (LMF) [Liu *et al.*, 2018] and Tensor Fusion Network (TFN) [Zadeh *et al.*, 2017]. In addition, several single-modal comparisons including LSTM , VGG-19 [Simonyan and Zisserman, 2015] and Inception-V3 [Szegedy *et al.*, 2016] are also involved. The best top-1 accuracy 90.70% is given by the TFN [Zadeh *et al.*, 2017].

2 Related Work

Product classification is a critical issue for an E-commerce platform since it can significantly improve the accuracy and reduce the workload of manual product category assignments. Since the product title usually aims at delivering the product information to users accurately and comprehensively as possible, text-based product classification has drawn more attention in the past years. In contrast, image data is generally harder to collect than text information, but its effectiveness is well demonstrated by a recent study [Zahavy *et al.*, 2018]. Therefore, in this section, we review and compare our presented MEP-3M dataset with several E-commerce product datasets, and mainly focus on image-based ones.

Table 1: Comparison with existing e-commerce datasets.

Dataset	Year	#class	#image	Modality
Stanford	2016	23K	0.120M	image
iMat FGVC6	2019	2K	1.012M	image
SIGIR 2020	2020	27	0.098M	image, text (French)
AliProducts	2020	50K	2.500M	image
Products-10K	2020	10K	0.150M	image
MEP-3M	**2021**	**599**	**3.012M**	**image, text (Chinese)**

In the past several years, different methods have been proposed to improve the performance of product classification, and many product datasets are collected and constructed, but unfortunately, they remained non-public [Zahavy *et al.*, 2018; Tang *et al.*, 2019; Dai *et al.*, 2020; Cao *et al.*, 2020; Gupta *et al.*, 2016; Li and Li, 2019]. On the other hand, there is also some public product dataset that only focuses on a limited subset of products (such as iMaterialist Fashion [Guo *et al.*, 2019]), but classification models on this type of dataset are not applicable for general e-commerce platforms. Meanwhile, there are also some retail groceries datasets such as RPC dataset [Wei *et al.*, 2019], but they differ from e-commerce datasets fundamentally since they are created for training automatic checkout systems. In the following, we briefly review the existing public e-commerce product datasets that aim at general products categories.

- **Stanford Online Products**[1] [Song *et al.*, 2016] is a e-commerce product dataset collected by a group from Stanford University using the web crawling API of eBay.com. Duplicate and irrelevant images in the dataset are filtered out. Each product in this dataset has approximately 5.3 images.
- **iMat Challenge@FGVC6**[2] is the dataset of iMaterialist Challenge on Product Recognition at FGVC6, CVPR 2019, provided by Malong Technologies and FGVC workshop. This dataset has a total number of 2,019

[1]https://github.com/rksltnl/Deep-Metric-Learning-CVPR16
[2]https://www.kaggle.com/c/imaterialist-product-2019/data

product categories, which are organized into a hierarchical structure with four levels.

- **SIGIR 2020 E-Commerce**[3] [Amoualian, 2020] refers to the dataset used by SIGIR 2020 eCom Rakuten Data Challenge. It is a multi-modal dataset, where each sample consists of the image, the title, and the description of a product. Text information is in French.
- **AliProducts**[4] [Cheng *et al.*, 2020] is a large-scale fine-grained SKU-level e-commerce product dataset without human-labelling. It also contains side information, such as hierarchical relationships between classes.
- **Products-10K**[5] [Bai *et al.*, 2020] is a large-scale product recognition dataset covering 10k fine-grained SKU-level products from JD.com. It contains both in-shop photos and customer images. All samples are manually checked to reduce noise.

A detailed comparison of the MEP-3M dataset and the existing public e-commerce datasets is shown in Table 1. Importantly, among the above datasets, only the SIGIR 2020 E-Commerce dataset is multi-modal, and our MEP-3M dataset has much more samples and more categories compared to SIGIR 2020 E-Commerce dataset. Moreover, the text in the SIGIR 2020 E-Commerce dataset is in Franch, while our dataset is in Chinese. Since China has been the world's largest online retail market, the MEP-3M dataset may have more potential application value.

3 The MEP-3M Dataset

The data of MEP-3M is collected from several Chinese online shopping websites. Each sample in MEP-3M is an image-text pair, where the image is a single image randomly selected from the product content page, and the text is the product title. The corresponding first-level class label and second-level sub-class label are also recorded, as shown in Table 2. However, the product labels from the different platforms are not exactly the same, *e.g.* '家居/家具/家装/厨具' (Home, furniture, decoration, and kitchenware) and '厨卫/生活家电/厨具' (Kitchen and bathroom equipment, Household Appliances, and Kitchenware) in the first-level are similar , and both '方便食品' and '方便速食' from the second level indicate the same concept, 'instant foods'. Moreover, the granularity of the classification across different platforms is also different (e.g., '水果' (fruit) v.s. '苹果' (apple), '橙子' (orange) and '芒果' (mango)). Therefore, it's necessary to perform label alignment to merging the data collected from different e-commerce platforms.

3.1 Hierarchical Label Alignment

Our label alignment is based on the analysis of the collected first-level labels (denote as 'class') and second-level labels ('sub-class') . To take the different granularity across different e-commerce platforms into account, we also set the third-level labels ('subsub-class') for some of the sub-classes.

[3]https://sigir-ecom.github.io/ecom2020/data-task.html

[4]https://tianchi.aliyun.com/competition/entrance/231780

[5]https://www.kaggle.com/c/products-10k

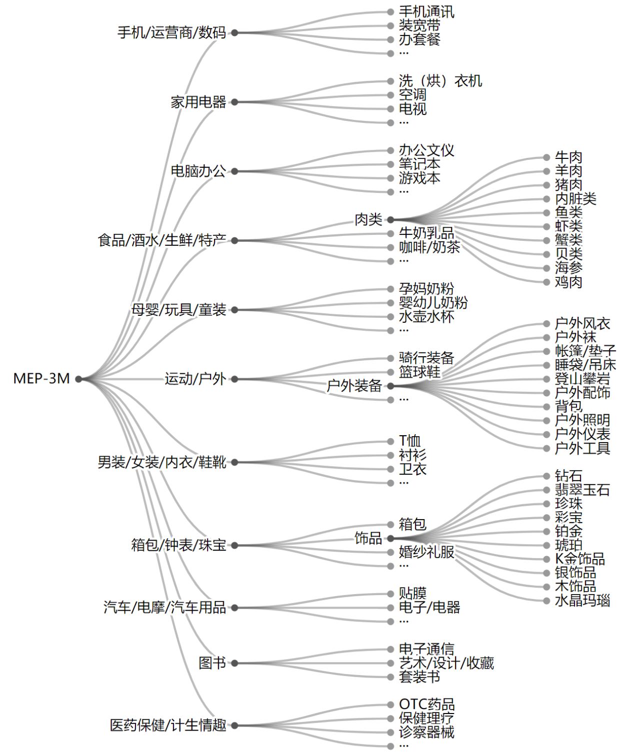

Figure 4: Illustration of the hierarchical structure of MEP-3M. Note that this figure only shows about 5% of the sub-classes of MEP-3M. For more detailed information, please visit the dataset website.

The First-Level Labels (Class)

Due to the number of the first level labels are relatively small (less than 20), we manually align them across different platforms. Classes with similar meanings are merged to a single class, whose new `class_name` is designated to cover the meaning of both sides. Meanwhile, unique classes are preserved as separated classes. Finally, there are a total of 14 different first level classes. A number is assigned to each class as its `class_id`. The `class_id` and the corresponding `class_name` of all the 14 first level classes are shown in Table. 2.

Table 2: The Numbers of samples and sub-classes of the 14 classes.

class_id	class_name	#sample	#sub_class
1	手机/运营商/数码	122,312	21
2	家用电器	240,779	51
3	电脑办公	85,699	17
4	家居/家具/家装/家纺/厨具	534,460	123
5	食品/酒水/生鲜/特产	411,046	53
6	美妆/个护清洁/宠物	139,049	30
7	母婴/玩具/童装	337,425	73
8	运动/户外	346,451	54
9	男装/女装/内衣/鞋靴	536,842	110
10	箱包/钟表/珠宝	86,648	13
11	艺术/礼品鲜花/农资绿植	46,316	14
12	汽车/电摩/汽车用品	66,963	21
13	图书	23,208	5
14	医药保健/计生情趣	35,761	14

The Second-Level Labels (Sub-Class)

The number of sub-classes is far more than the first level, making manually alignment impossible. Therefore, we design an automated alignment approach based on quantitative text analysis. Specifically, the goal of the alignment is to figure out the sub-class pairs that are semantically similar across different e-commerce platforms. We assume these sub-class pairs have the following three characteristics: 1) they belong to the same first level class, 2) their names share a certain degree of similarity, 3) their title contents have similar features on term frequency. For the second and the third characteristics, we respectively calculate label similarity S_{label} and content similarity $S_{content}$ as metrics.

The label similarity S_{label} measures how far the two sub-class names coincide with each other, it is defined as:

$$S_{label} = 2.0 \times M/T \tag{1}$$

, where T indicates the total number of characters in both sub-class names, and M indicates the number of matches. Note that this is 1.0 if the sub-class names are identical, and 0.0 if they have nothing in common.

The content similarity $S_{content}$ is the cosine distance between term-frequency features extract from the title text content of two different sub-classes, it is defined as:

$$S_{content} = \frac{x_1 \cdot x_2}{\|x_1\| \times \|x_2\|} \tag{2}$$

, where x_1 and x_2 are the term-frequency feature vector of title text content. Each element in x_1 and x_2 counts the number of occurrences of a certain term.

The S_{label} are calculated by using python difflib package [6], while the $S_{content}$ is based on python simtext package [7]. In order to improve computational efficiency of $S_{content}$, we use the first 22000 characters of a sub-class product titles, corresponding to approximately 450 products. We iterate over all the sub-class pairs that belongs to the same classes, and filter them according to the criterion of

$$S_{label} \geq 0.50 \text{ AND } S_{content} > 0.75 \tag{3}$$

, where 0.75 is the average $S_{content}$ of $S_{label} = 1.00$ sub-classes. New names are manually assigned for those sub-class pairs that $S_{label} \neq 1.00$. Some examples of the results are listed in Table. 3.

The Third-Level Labels (Subsub-Class)

This part deals with the different granularity of the classification across different e-commerce platforms. Beyond the class and the sub-class labels, we create finer-grained subsub-class labels for a total of 13 sub-classes: '箱包' (bags), '饰品' (accessories), '手机配件' (mobile phone accessories), '男装' (men's clothing), '女装' (women's clothing), '内衣' (underwear), '户外装备' (outdoor equipment), '水果' (fruit), '肉类' (meat), '冲调饮品' (toned drinks), '南北干货' (dry foods), '纸尿裤' (diapers), and '奶瓶奶嘴' (bottle nipples).

In the following, we give an example of products in MEP-3M dataset.

[6]https://docs.python.org/3/library/difflib.html

[7]https://pypi.org/project/simtext

Table 3: Examples of second-level label alignment.

sub-class name	sub-class name	S_{label}	$S_{content}$	new class name
儿童餐具	儿童餐具	1.00	0.929	儿童餐具
孕妈奶粉	孕妈奶粉	1.00	0.898	孕妈奶粉
空调	空调	1.00	0.870	空调
骑行装备	骑行装备	1.00	0.768	骑行装备
洗澡用具	洗澡用具	1.00	0.705	洗澡用具
婴幼奶粉	婴幼儿奶粉	0.89	0.960	婴幼儿奶粉
婴儿湿巾	湿巾	0.67	0.907	湿巾
咖啡/奶茶	咖啡	0.57	0.854	咖啡/奶茶
饮料	饮料饮品	0.67	0.849	饮料饮品
办公文具	办公文仪	0.75	0.756	办公文仪

```
{
    'class_id': '5',
    'class_name': '食品 / 酒水 / 生鲜 / 特产',
    'sub_class_id': '523',
    'sub_class_name': '水果',
    'subsub_class_id': '640',
    'subsub_class_name': '苹果',
    'img_path': 'Images/523/3.jpg',
    'img_resolution': (220, 220, 3),
    'title': '【第 2 件 9.8 , 2 件共发带
        箱 10 斤】脆甜冰糖心红富士苹果 5 斤鲜果时令
        大果新鲜水果陕西洛川一整箱非烟台 5 斤装（净
        重 5 斤）'
}
```

The `class_id` denotes the first level of class label, ranging from 1 to 14. The `sub_class_id` is the second level of class label, ranging from 1 to 599. The `subsub_class_id` corresponds to the third level index, which ranges from 600 to 688. For the sub-class that does not have finer-grained subsub-classes, the `subsub_class_id` and `subsub_class_name` are set to `'FLASE'`.

3.2 Statistics of MEP-3M Dataset

Most images are in a 220×220 resolution, and the others are in 64×50, 75×75, 60×60, 54×54, 100×75, 800×800 and 219×220 resolution. A total of 2,908,596 (96.53%) of the images are in `.jpg` format, while the other 104,363 (3.46%) images are in `.png` format. The text of each level of label and `title` is in simplified Chinese. The length of `title` ranges from 2 characters to more than 100 characters. The average length of it is 49. In total, the dataset consists of 156,069,329 characters in title. The entire dataset takes around 76 GB storage and will be made publicly available for non-commercial research purposes.

The the long-tail distribution of MEP-3M is shown in the left of Fig. 5, while distribution of image size and title length is shown in the right.

The MEP-3M is also a fine-grained dataset. Many images are visually similar but belong to a different class. We select the nearest neighbors in the sample space to demonstrate this point. The clustering is done by calculating the pixel-wise distance. The results are shown in Fig. 6.

4 Baselines

To demonstrate the efficacy possible of the MEP-3M dataset, In this section, we evaluate several baseline models for the

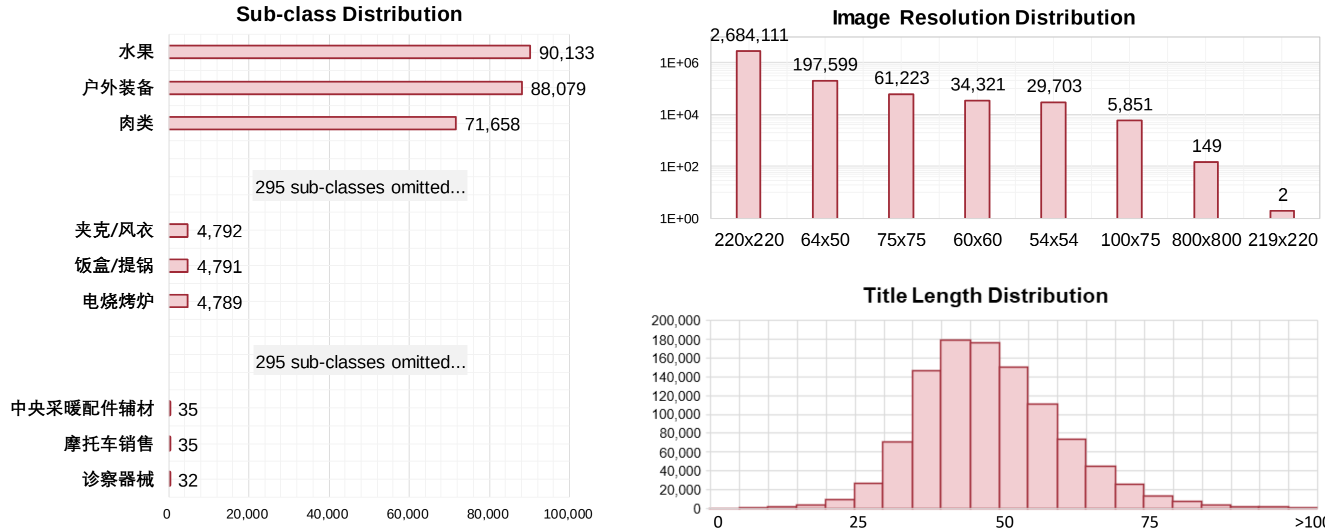

Figure 5: The distribution of sub-classes, image resolution, and title length in MEP-3M dataset.

Figure 6: Two groups of visually similar images that belongs to different classes. The images are obtained by calculating pixel-wise distance.

e-commerce product classification task. We test two popular multi-modal learning models: Low-rank Multimodal Fusion (LMF) [Liu *et al.*, 2018] and Tensor Fusion Network (TFN) [Zadeh *et al.*, 2017]. In addition, we also test several single-modal comparisons (LSTM for text-only, VGG-19 [Simonyan and Zisserman, 2015] and Inception-V3 [Szegedy *et al.*, 2016] for image only) to demonstrate the effectiveness of multi-modal understanding in the e-commerce product classification task.

We divide the full dataset randomly into training and test set at a ratio of 8:2. The model is implemented by TensorFlow, using an Intel i5-9400F CPU and NVIDIA TITAN RTX GPU. All experiments are trained with Adam optimizer, and the initial learning rate is set to 1e-3 and decreases every 2 epochs at a rate of 0.5. The batch size is 64. For LSTM-based text-only model, we first remove meaningless characters from texts with regular expressions and then implement Chinese word segmentation. Word2vec model from the gensim toolkit is used to obtain word embedding. The representation is further passed to the LSTM or BiLSTM model for classification.

The testing accuracies, average precision score (AP) and F1-score of baseline models are shown in Table 4. We can see that the multi-modal methods LMF [Liu *et al.*, 2018] and

Table 4: The classification accuracies of different baseline methods.

Model	Top-1	Top-5	AP	F1-score
VGG-19	76.36%	91.77%	0.3966	0.7275
Inception-v3	79.48%	94.27%	0.6326	0.7493
LSTM	89.13%	98.33%	0.9200	0.8796
Bi-LSTM	90.68%	98.70%	**0.9309**	0.8931
TFN	**90.70%**	**98.74%**	0.9289	0.8899
LMF	89.22%	98.19%	0.8924	**0.9125**

TFN [Zadeh *et al.*, 2017] achieved better results than single-modal methods. It demonstrated the advantage of multi-modal product classification over single-modal-based methods. The best top-1 accuracy of 90.70% is yeild by the TFN [Zadeh *et al.*, 2017].

5 Conclusion

In this paper, we constructed a large-scale multi-modal e-commerce products classification dataset named MEP-3M, which contains over 3 million image-text pairs of products and covers 599 fine-grained product categories. MEP-3M is the largest in existing E-commerce datasets to our best knowledge. Moreover, several baseline models are implemented to give a brief evaluation of the dataset. We believe that the MEP-3M dataset has great potential for facilitating related research since it is simultaneously large-scale, hierarchical-categorized, multi-modal, fine-grained, and long-tailed.

Acknowledgments

This work was partially funded by Natural Science Foundation of Jiangsu Province under Grant No. BK20191298, Fundamental Research Funds for the Central Universities under Grant No. B200202175.

References

[Amoualian, 2020] Hesam Amoualian. SIGIR 2020 e-commerce workshop data challenge overview. 2020.

[Anderson *et al.*, 2021] C. E. Anderson, Al Farrell, and Brigham Young. Have fun storming the castle(s)! In *WACV*, 2021.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society, 2015.

[Bai *et al.*, 2020] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10K: A large-scale product recognition dataset. *CoRR*, abs/2008.10545, 2020.

[Cao *et al.*, 2020] Zhihao Cao, Shaomin Mu, and Mengping Dong. Two-attribute e-commerce image classification based on a convolutional neural network. *Vis. Comput.*, 36(8):1619–1634, 2020.

[Cheng *et al.*, 2020] Lele Cheng, Xiangzeng Zhou, Liming Zhao, Dangwei Li, Hong Shang, Yun Zheng, Pan Pan, and Yinghui Xu. Weakly supervised learning with side information for noisy labeled images. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 306–321, Cham, 2020. Springer International Publishing.

[Dai *et al.*, 2020] Jin Dai, Tianyu Wang, and Shaowei Wang. A deep forest method for classifying e-commerce products by using title information. In *International Conference on Computing, Networking and Communications, ICNC 2020, Big Island, HI, USA, February 17-20, 2020*, pages 1–5. IEEE, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[Elliott *et al.*, 2016] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics, 2016.

[Guo *et al.*, 2019] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R. Scott, and Serge J. Belongie. The imaterialist fashion attribute dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 3113–3116. IEEE, 2019.

[Gupta *et al.*, 2016] Vivek Gupta, Harish Karnick, Ashendra Bansal, and Pradhuman Jhala. Product classification in e-commerce using distributional semantics. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 536–546. ACL, 2016.

[Horn *et al.*, 2015] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge J. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 595–604. IEEE Computer Society, 2015.

[Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8769–8778. IEEE Computer Society, 2018.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[Li and Li, 2019] Guo Li and Na Li. Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network. *Electron. Commer. Res.*, 19(4):779–800, 2019.

[Li *et al.*, 2021] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2021.

[Liu *et al.*, 2018] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2247–2256. Association for Computational Linguistics, 2018.

[Ordonez *et al.*, 2011] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Song *et al.*, 2016] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4004–4012. IEEE Computer Society, 2016.

[Srinivasan *et al.*, 2021] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. *CoRR*, abs/2103.01913, 2021.

[Sun *et al.*, 2018] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 834–850. Springer, 2018.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016.

[Tang *et al.*, 2019] Yina Tang, Fedor Borisyuk, Siddarth Malreddy, Yixuan Li, Yiqun Liu, and Sergey Kirshner. MSURU: large scale e-commerce image classification with weakly supervised search data. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2518–2526. ACM, 2019.

[Wei *et al.*, 2019] Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, and Lingqiao Liu. RPC: A large-scale retail product checkout dataset. *CoRR*, abs/1901.07249, 2019.

[Yang *et al.*, 2015] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3973–3981. IEEE Computer Society, 2015.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1103–1114. Association for Computational Linguistics, 2017.

[Zahavy *et al.*, 2018] Tom Zahavy, Abhinandan Krishnan, Alessandro Magnani, and Shie Mannor. Is a picture worth a thousand words? A deep multi-modal architecture for product classification in e-commerce. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7873–7881. AAAI Press, 2018.