GLAMI-1M: A Multilingual Image-Text Fashion Dataset

Václav Košař[1]
vaclav.kosar@glami.cz
Antonín Hoskovec[1,3]
antonin.hoskovec@glami.cz
Milan Šulc[2]
milan.sulc@rossum.ai
Radek Bartyzal[1]
radek.bartyzal@glami.cz

[1] GLAMI.cz
Křižíkova 148/34, 186 00 Prague 8,
Czech Republic

[2] Rossum.ai
Křižíkova 148/34, 186 00 Prague 8,
Czech Republic

[3] FNSPE, Czech Technical University in Prague,
Břehová 7, 119 15, Prague 1,
Czech Republic

Abstract

We introduce GLAMI-1M: the largest multilingual image-text classification dataset and benchmark. The dataset contains images of fashion products with item descriptions, each in 1 of 13 languages. Categorization into 191 classes has high-quality annotations: all 100k images in the test set and 75% of the 1M training set were human-labeled. The paper presents baselines for image-text classification showing that the dataset presents a challenging fine-grained classification problem: The best scoring EmbraceNet model using both visual and textual features achieves 69.7% accuracy. Experiments with a modified Imagen model show the dataset is also suitable for image generation conditioned on text. The dataset, source code and model checkpoints are published at: https://github.com/glami/glami-1m.

ANKA KEMER Kadın Heybe Çantalı Kemer 16x14 cm
(Turkey)
'womens-belts'

Pánská kotníková obuv Mustang 4107-605-820 modrá
(Czechia)
'mens-boots'

Ženski kopalni plašč DKaren Basic
(Slovenia)
'womens-bathrobes'

Pilgrim Auskarai 'THANKFUL' sidabrinė
(Lithuania)
'womens-earrings'

Figure 1: Examples from GLAMI-1M with their *image*, *name*, *country* and *class*. Available information not displayed: *description*, *label source* and *item-ID*.

1 Introduction

Public datasets are a cornerstone of machine learning research: Cross-evaluation of different methods is possible thanks to public benchmarks with pre-defined training and test data splits. Pushing the envelope in machine learning often relies on considerable amount of training samples. For example, while the existence of Convolutional Neural Networks dates back to the 1980s [11, 23], the deep learning era in computer vision started with the success [21] on the ILSVRC 2012 challenge dataset [39] commonly addressed as ImageNet. At the time of writing this paper, the best results reported[1] on ImageNet were achieved by an image-text model CoCa [59], pre-trained on proprietary large-scale datasets JFT-3B [60] and ALIGN [18] to produce joint image-text representation. Similarly, CMA-CLIP [10] incorporated CLIP [34], an ALIGN model [18] predecessor, trained on proprietary WebImageText to achieve state-of-the-art image-text classification results on Fashion-Gen [38]. These results suggest that image-text models have a great potential to aid image-based classification.

Owing to the success of multilingual models [6, 7] and multimodal models [18, 34], datasets combining both multilingual and multimodal features are increasingly relevant to machine learning research (see Table 1). However, public large scale image-text classification datasets [29, 38, 53, 54] are still of rather limited size and language diversity (see Table 2). Note that, Recipe1M+ is not human annotated, rather its categories are extracted from recipe titles using statistical methods. In particular within the fashion domain, to the best of our knowledge, there is no large diverse multilingual text and image dataset (see Table 3) and machine translation cannot replace human produced text (yet).

In this paper, we introduce GLAMI-1M: the largest multilingual image-text classification dataset and benchmark. The dataset contains images of fashion products with item descriptions from an e-commerce platform. GLAMI-1M is a collection of 1.11M records representing a fashion product with an image, a name and description in one of 13 languages and a category within the GLAMI fashion search engine[2]. Categorization into 191 classes has high-quality annotations: all 100k images in the test set and 75% of the 1M training set were human-labeled.

Organizing products from public listings into categories is an important problem in e-commerce platforms. Data from online production systems pose several challenges: dealing with imbalanced long-tailed class distributions [35], prior shift [44, 48], noisy labels in case of rule-based annotations [49, 61] (as opposed to human labels), multimodal inputs [31, 35], multilingual text [31, 35], and utilizing available metadata [52].

Datasets for related tasks and domains are reviewed in section 2. The GLAMI-1M dataset and benchmark is introduced in section 3, including detailed analysis of its content and description of its creation. Baseline methods for image-text classification and text-conditional image generation are introduced in section 4. Additional details about the dataset and the experiments, and baselines for machine translation are provided in the supplementary material.

2 Related Work

Large-scale image and multilingual text datasets are listed in Table 1. GLAMI-1M is the largest multilingual dataset for image-text classification. Larger image-text datasets LAION-

[1] https://paperswithcode.com/sota/image-classification-on-imagenet

[2] at the point of extraction in 2022.

5B [41], WIT [47], FooDI-ML [31] are used for image-text retrieval, and miss standardized class labels. Note that in Table 1, we do not list translations of MS-COCO [25] such as [2, 15, 22, 24, 51, 58] as they are distributed in bilingual form, which does not pass the table's minimum of 3 languages.

Table 1: Publicly available multilingual image-text datasets. Datasets with <3 languages and with <10k images or texts are omitted. The column task gives the most relevant task.

Dataset	Images	Texts	Langs	Domain	Task
LAION-5B [41]	5.85B	5.85B	100+	Web images	image-text retr.
YFCC100M [50]	100M	100M	172	Web images	image-text retr.
WIT [47]	11.5M	37.6M	108	Wiki images	image-text retr.
FooDI-ML [31]	1.5M	9.5M	33	Food, groceries	text-image retr.
GLAMI-1M	968k	1.01M	13	Fashion	classification
MultiSub (I4) [52]	45k	180k	4	subtitles, nouns	fill-in-the-blank
Multi30k [3, 8, 9, 46]	30k	4 x 30k	4	General	machine translation

Large fashion datasets with image and text features are summarized in Table 3. To the best of our knowledge, GLAMI-1M is the largest image-text dataset in terms of items and the most diverse dataset in terms of languages. GLAMI-1M also offers the highest number of categories (191) for classification. The only other multilingual fashion image-text dataset, Fashion-MMT [45], is bilingual and ten times smaller in the number of items.

Other Fashion datasets without text annotations include: DeepFashion2 [12] contains 800k diverse photos with clothing segmentation metadata. Clothing-1M [20] contains 1M product images with majority noisy class (14) labels. MVC [27] dataset of 161k items for view-invariant clothing retrieval, classification (23), colors (13), attribute prediction. ModaNet [63] is a 55k image segmentation dataset. Fashionpedia [19] is a 45k image dataset with fine-grained apparel attribute (294) prediction, segmentation (27 categories, 19 parts), and an ontology. StreetStyle [30] is a 45k image dataset with various attributes including category (7). DeepFashion3D [14] is a 2k image to 3D reconstruction dataset with annotations including 10 categories. Colorful-Fashion [28] is 2k image dataset for segmentation into 23 categories, 13 colors.

3 Dataset Description

We introduce GLAMI-1M: a 13-lingual image-text classification dataset of 1.10M items representing a product and its leaf category within GLAMI production catalog category tree. Each item represents a product listing with: image, texts (name and description) in one of the 13 languages, category label and its label source. Examples from the dataset are in Figure 1 and in the supplementary material.

Table 2: Publicly available image-text classification datasets. Datasets with <30k images or texts are omitted.

Dataset	Images	Texts	Langs	Domain	Class. task	Classes
Recipe1M+ [29]	13M	1M	1	Recipes	single-label	1047
GLAMI-1M	968k	1.01M	13	Fashion	single-label	191
FashionGen [38]	325k	78k	1	Fashion	single-label	121
UPMC Food-101 [53]	100k	100k	1	Food	single-label	101
SNLI-VE [54]	30k	565k	1	General	single-label	3

Table 3: Overview of publicly available fashion product datasets with image and text features. GLAMI-1M is the biggest, most fine-grained, and uniquely multilingual fashion dataset.

Dataset	Items	Imgs	Features	Langs
GLAMI-1M	1.11M	968k	image, name, description, class (191)	13
FACAD [57]	130k	993K	image, description, class (78)	1
Fashion-MMT [45]	110k	853k	image, description with noisy translations, class (78), attributes	2
Fashion550k [17]	550k	408k	image (in-the-wild), user comments, garment class, attributes, other metadata	1
Neti-look [26]	350k	355k	image (in-the-wild), comments	1
FashionGen [38]	78k	325k	image, description, class (121)	1
Amazon Fashion Products 2020 [33]	132k	132k+	multiple images, name, other	1
Fashion IQ [13]	50k	50k	image, description, attributes, relative caption	1
Fashion Product Images [1]	44k	44k	image, name, description, class, other	1

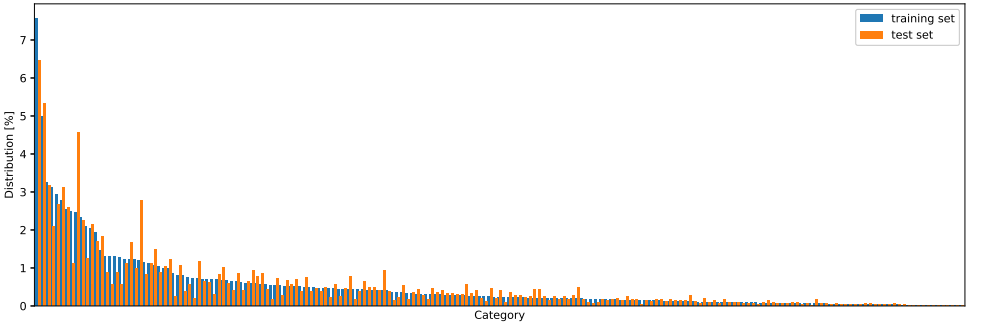

Figure 2: Distribution of samples per category. The distribution is mostly exponential, but steeper along the edges, so we regard this as a long tailed distribution.

Items for the dataset were selected from the GLAMI catalogue in two phases: first, we sampled items with higher-quality human annotations (i.e. based on the label source). 100k of these items were randomly sampled for the test set. Then items with labels from less reliable rule-based (heuristic) labeling systems were sampled proportionally to the catalog category distribution, in order to get a training set of 1M items. Zero overlap between the training and test set images and texts was checked via MD5 hashes and cosine similarity threshold of CLIP embeddings [4, 34]. See the source code and the supplementary material for details. Text was preprocessed by removing backslashes, braces, brackets, semicolons, angle brackets, and replacing line ends, carriage returns and forward slashes with a space.

Table 4 describes the dataset's data columns with the numbers of unique values. The training set may contain several records describing the same item (i.e. records with the same *item_id*) – e.g. because unisex items appear in both men's and women's category variants. The test set contains only consistent human-label annotations without such duplicate records (with same *item_id*). However, up to tens of items still have the same *image_id*, since the same products are sometimes sold by multiple e-shops within the same or different country.

Table 4: GLAMI-1M column descriptions, and unique value count in training and test sets.

Name	Description	# Train.	# Test
item_id	Item integer identifier (Not unique).	992528	116004
image_id	Image integer identifier. Products with duplicate images exists across different geos.	882846	85577
geo	Country code in lower case. It is a strong indicator of language used in the text.	13	13
name	Product name text. Often contains product's brand.	752092	99783
description	Product description text. It describes the product and advertises the product.	656067	90313
category	Integer category id label.	191	191
category_name	Human readable category name label.	191	191
label_source	Source of the class labels indicating label quality: *admin*, *quality-check*, *custom-tag*: human labels *combined-tag*, *NaN*: machine labels – simple rule based systems	5	3

In these cases the items have a different *item_id*. The classes are fine-grained: 15 categories of women shoes and total 191 categories in contrast to FashionGen's 121. The class distribution is long tailed, as shown in Figure 2. The 10 most and 10 least frequent training set categories can be found in Table 5. Figure 3 shows a train-test distribution shift in number of samples per country. The distribution of product name and description lengths is illustrated in Figure 4. For the distribution of label source, please see the supplementary material.

The dataset is primarily shared in a compact 10GB archive with 228x298px images in JPEG format. Larger 800x800px resolution variants are available in a separate archive.

Together with the dataset, we set up a **public benchmark**[3] for multilingual image-text classification. The benchmark's primary score is the test set accuracy. The benchmark allows using pre-trained models and additional training data, if explicitly stated in the method description. Initial results for baseline classification methods are provided in subsection 4.1.

Additionally, we provide baseline generative models for text-conditioned image generation, as described in subsection 4.2, and baseline models and results for machine translation in the supplementary material.

4 Experiments

4.1 Multimodal Classification

Classification is one the fundamental tasks of supervised learning [42]. Multimodal classification models process inputs of several different modalities. In our benchmark the inputs come from three *modalities*: textual (title + description), visual (image) and categorical (label source). The label source could be used as a meta information for training methods like sample weighting [43] or label correction [62], however these experiments are beyond the scope of this paper. For baseline we have chosen EmbraceNet [5], a robust model essentially

[3] accessible from the repository.

Table 5: The 10 most and 10 least represented from the 191 total training set categories.

Category name	# Train.	# Test	Category name	# Train.	# Test
mens-t-shirts-and-tank-tops	75724	7497	mens-bath-robes	211	26
womens-tops-tank-tops-and-t-shirts	50000	6187	mens-handkerchiefs	200	11
mens-sneakers	32385	3668	mens-shoe-laces	187	3
womens-sneakers	31137	2417	mens-umbrellas	179	10
dresses	29350	3084	mens-suspenders	171	19
baby-clothing	27896	3631	broaches	155	17
womens-blouses-and-shirts	25292	3017	mens-chains	122	16
womens-pants	24998	1305	mens-rubber-boots	99	24
bikinis	24712	5286	mens-earrings	88	12
womens-flip-flops	23219	2612	boys-tank-tops	81	14

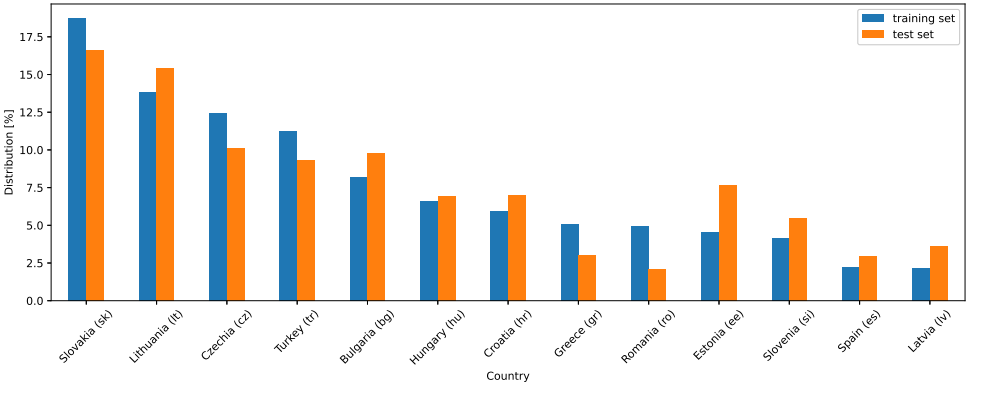

Figure 3: Sample distribution per country (geo), which roughly approximates the language distribution.

capable of taking encoded inputs from any modality and automatically combining them into a single model. In all experiments, the model was trained for two epochs (early stopping) with the Adam optimizer and the internal EmbraceNet dimension set to 512.

For the encoding of various modalities we have relied on well tested, publicly available, pre-trained models. We have encoded the textual inputs with the *small* variant of the mT5 model [56], which has been pretrained on a superset of the languages in our dataset. We encoded with maximum length of 32 tokens, which resulted in $(32 \times 512 = 16384)$ dimensional embeddings of the concatenated title + description. For the image inputs we have used a pretrained *ResNeXt-50 32x4d* model [55], which after the last max pooling layer gives 2048 dimensional embeddings. We finetuned ResNext, but froze mT5.

To better understand the quality of the input features, we have trained several versions of EmbraceNet by dropping one or multiple modalities from the input and by training on human-labels only or including the noisy labels too, see Table 6. The best top-1 accuracy of 0.697 was achieved with the combination of both text and image and by including the noisy labels, while separately the image features outperformed the textual inputs. We note that we did not tune the probabilities of the fusion process in EmbraceNet [5]. The probability of the

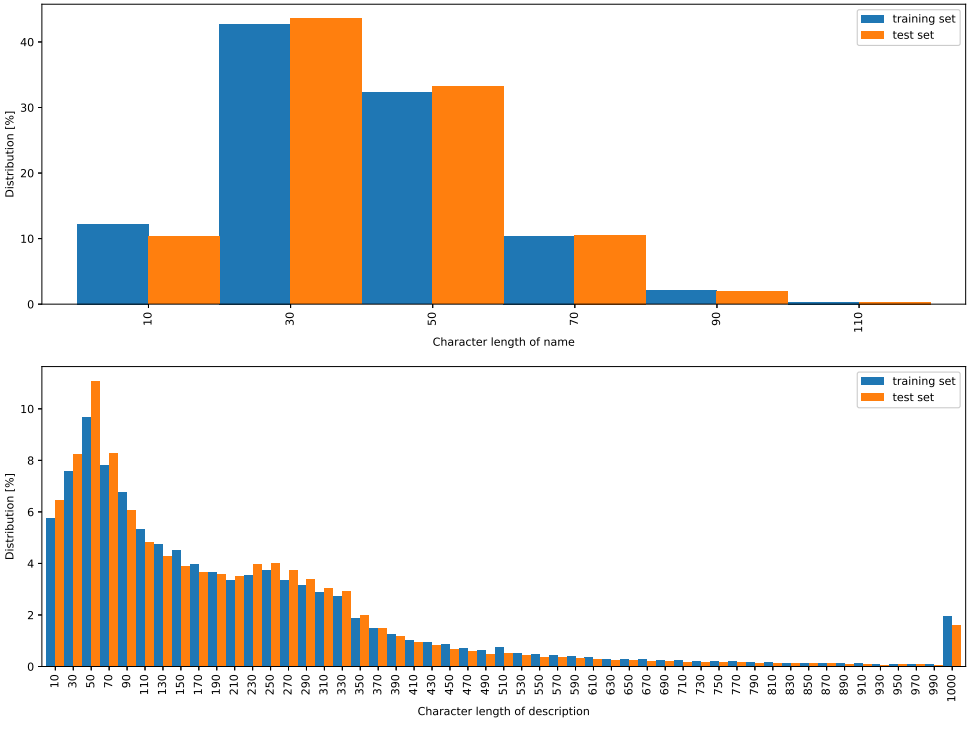

Figure 4: Distribution of *name* (top) and *description* (bottom) length in characters in the training and test set. For *name* the median is 38 characters, for *description* the median is 150. Note that the last bin for description contains all the samples longer than 990 characters (up to 4000).

docking layers of each modality being included was thus $1/2$ for the bi-modal version. To see how EmbraceNet trained on images compares to the ResNeXt-50 32x4d model, we used a pre-trained ResNext and finetuned it on our dataset. Since in this case EmbraceNet essentially replaces the last fully connected layer in ResNext with several layers, thus increasing the number of parameters, the image-only version of EmbraceNet outperformed the original architecture. The presence of the noisy labels only has a small impact on the performance of EmbraceNet.

A weak zero-shot CLIP baseline is available in the supplementary material.

4.2 Text-Conditional Image Generation

The area of image generation conditioned on text has recently attracted much attention [36, 37, 40]. Our dataset can be used for this task. We have trained a "small" version of the Imagen-like model [40] on a single NVIDIA T4 GPU over 72 hours on 884k images and 992k texts. This underscores the position of our dataset in the matter of its size. It lies on the border between extremely large datasets, allowing to push the envelope of the state-of-the-art in machine learning, and datasets compact enough to train a model on a single GPU in days.

We have trained a small cascading Denoising Diffusion Probabilistic Model [16] condi-

Table 6: Top-k accuracies of EmbraceNet with various input modalities, trained either on all labels (*all*) or human-labeled samples only (*hum.*).

Included modality/model	Top-1 (all)	Top-5 (all)	Top-1 (hum.)	Top-5 (hum.)
Text + Image	**0.697**	0.940	0.694	0.932
Image	0.685	**0.948**	0.679	0.943
Text	0.593	0.840	0.613	0.849
Finetuned ResNeXt-50 32x4d	0.631	0.935	0.642	0.933

Figure 5: Cherry-picked images generated by the Imagen-like model with the corresponding country codes, 500 time steps of diffusion. Large images are the generated ones, the two smaller are the closest images from train set based on *ResNeXt-50 32x4d* embeddings.

tioned on text embeddings with two Unet models inside, each with the internal dimension of 128. We used the same text embeddings as in subsection 4.1. The generation happens in two steps, we first generate a 64x64 image from which we upscale to 128x128 pixels. We make this model publicly available, including its weights. See the the source code for more details.

We report a sample of visual results: Figure 5 shows images sampled on the texts from the test set. Another interesting property of the generator is the novelty of the generated images. We have looked up the two closest images using visual embeddings in the training set for each of the generated images and we were unable to find identical looking images

Figure 6: Random samples of images generated by the Imagen-like model with the corresponding country codes, 500 time steps of diffusion.

Figure 7: Images generated by the Imagen-like model for the input "sneakers" translated into all 13 languages, 500 time steps of diffusion.

in the train set. Let us underscore that not only the generated images are not pixel perfect replicas of the train samples, but they are quite far from the training samples even by human standards - not just L2 distance caused by imperceptible noise. We have cherry-picked the samples in this table to show that the model learned to draw product images that appear almost realistic. For an unbiased sample of images generated from the test set, see more examples in the Figure 6. About one third of the images generated by the model appears realistic based on a sample of about 1000 images checked by hand. Furthermore, we have experimented with the text-conditioning and generated images for various phrases translated into all of the languages in the dataset. In about one third of the texts the conditioning failed, in about a third it reflected the correct piece of clothing, but the style was wrong. In other words, high-level category such as "shoes" was correct, but a low-level one such as "sneakers" was often wrong. In about the last third of cases it worked to obtain a realistic sample, see for example Figure 7.

5 Conclusion

The paper introduced GLAMI-1M: the largest publicly available multilingual image-text classification dataset and the largest image-text dataset in the fashion domain. Thanks to its characteristics, the dataset has the potential to accelerate research in several areas of machine learning, including multilingual image-text classification, text-conditional image generation and multilingual machine translation. For example, it can be used as an alternative to Recipe1M+ [29] adding the aspect of multilinguality, or as a larger alternative to FashionGen [33] and other datasets in the fashion domain.

Experiments on multimodal image classification in subsection 4.1 show the dataset presents a challenging problem. Together with the dataset, we introduce a benchmark with baseline results and pre-trained models available, and we invite everyone to evaluate their models in the public leaderboard[3].

Additional experiments on text-conditional image generation and multi-language machine translation are described in subsection 4.2 and in the supplementary material respectively. The experiments illustrate the dataset's usefulness for other tasks than classification. Pre-trained models and code for the tasks are also shared with the paper.

Other relevant problems left for future work include long-tail learning, adaptation to prior shift, learning from a combination of trusted (human) and noisy (rule-based) annotations.

References

[1] Param Aggarwal. Fashion product images dataset, version 1, 2019, March. URL https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset. 4

[2] Scaiella Antonio, Danilo Croce, and Roberto Basili. Large scale datasets for image and video captioning in italian. *Italian Journal of Computational Linguistics*, 2019. 3

[3] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, 2018. 3

[4] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.739. 4

[5] Jun-Ho Choi and Jong-Seok Lee. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019. URL https://arxiv.org/pdf/1904.09078.pdf. 5, 6

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020. URL https://arxiv.org/pdf/1911.02116.pdf. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. URL https://aclanthology.org/N19-1423.pdf. 2

[8] D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30k: Multilingual english-german image descriptions. *Association for Computational Linguistics*, pages 70–74, 2016. 3

[9] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September 2017. 3

[10] Jinmiao Fu, Shaoyuan Xu, Huidong Liu, Yang Liu, Ning Xie, Chien-Chih Wang, Jia Liu, Yi Sun, and Bryan Wang. Cma-clip: Cross-modality attention clip for text-image classification. In *IEEE ICIP 2022*, 2022. URL https://www.amazon.science/publications/cma-clip-cross-modality-attention-clip-for-text-image-classif 2

[11] K FUKUSHIMA. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36 (4):193–202, 1980. 2

[12] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019. 3

[13] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven J. Rennie, and Rogério Schmidt Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11302–11312, 2021. 4

[14] Zhu Heming, Cao Yu, Jin Hang, Chen Weikai, Du Dong, Wang Zhangye, Cui Shuguang, and Han Xiaoguang. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision – ECCV 2020*, pages 512–530. Springer International Publishing, 2020. ISBN 978-3-030-58452-8. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123460494.pdf. 3

[15] Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. Multimodal pivots for image caption translation. *ArXiv*, abs/1601.03916, 2016. URL https://aclanthology.org/P16-1227.pdf. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf. 7

[17] Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa. Multi-Label Fashion Image Classification with Minimal Human Supervision. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2017. 4

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[19] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[20] Takuhiro Kaneko, Y. Ushiku, and Tatsuya Harada. Label-noise robust generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2471, 2019. 3

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[22] Quan Hoang Lam, Quang Duy – Le, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Uit-viic: A dataset for the first evaluation on vietnamese image captioning. In *ICCCI*, 2020. 3

[23] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 2

[24] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21:2347–2360, 2019. 3

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. URL https://arxiv.org/abs/1405.0312. 3

[26] Wen Hua Lin, Kuan-Ting Chen, HungYueh Chiang, and Winston H. Hsu. Netizen-style commenting on fashion photos: Dataset and diversity measures. *Companion Proceedings of the The Web Conference 2018*, 2018. 4

[27] Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016. 3

[28] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16:253–265, 2014. 3

[29] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. URL http://pic2recipe.csail.mit.edu/. 2, 3, 9

[30] Kevin Matzen, Kavita Bala, and Noah Snavely. Streetstyle: Exploring world-wide clothing styles from millions of photos. *ArXiv*, abs/1706.01869, 2017. URL https://arxiv.org/abs/1706.01869. 3

[31] David Amat Olóndriz, Ponç Palau Puigdevall, and Adrià Salvador Palau. Foodi-ml: a large multi-language dataset of food, drinks and groceries images and descriptions, 2021. URL https://arxiv.org/abs/2110.02035. 2, 3

[32] Lukáš Picek, Milan Šulc, Jiří Matas, Thomas S. Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias Frøslev. Danish fungi 2020 - not just another image recognition dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1525–1535, January 2022. 2

[33] PromptCloud. Amazon fashion products 2020, version 1, 2021, March. URL https://www.kaggle.com/datasets/promptcloud/amazon-fashion-products-2020. 4

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language

supervision. In *ICML*, 2021. URL https://arxiv.org/pdf/2103.00020.pdf. 2, 4

[35] Kshetrajna Raghavan. Using rich image and text data to categorize products at scale, 2021. URL https://shopify.engineering/using-rich-image-text-data-categorize-products. 2

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125. 7

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. URL https://arxiv.org/pdf/2112.10752.pdf. 7

[38] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge, 2018. URL https://arxiv.org/abs/1806.08317. 2, 3, 4, 9

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. 2

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv*, 2022. URL https://arxiv.org/abs/2205.11487. 7

[41] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://openreview.net/pdf?id=M3Y74vmsMcY. 3

[42] Pratap Chandra Sen, Mahimarnab Hajra, and Mitadru Ghosh. Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics*, pages 99–111. Springer, 2020. 5

[43] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 5

[44] Tomáš Šipka, Milan Šulc, and Jiří Matas. The hitchhiker's guide to prior-shift adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1524, 2022. URL https://arxiv.org/pdf/2106.11695.pdf. 2

[45] Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. Product-oriented machine translation with cross-modal cross-lingual pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 3, 4

[46] Lucia Specia, Stella Frank, Loïc Barrault, Fethi Bougares, and Desmond Elliott. Wmt18: Multimodal machine translation on multi30k, 2018. URL https://www.statmt.org/wmt18/multimodal-task.html. 3

[47] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. URL http://arxiv.org/pdf/2103.01913. 3

[48] Milan Šulc and Jiří Matas. Improving cnn classifiers by estimating test-time priors. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3220–3226, 2019. URL https://arxiv.org/pdf/1805.08235.pdf. 2

[49] Chong Sun, Narasimhan Rampalli, Frank Yang, and AnHai Doan. Chimera: Large-scale classification using machine learning, rules, and crowdsourcing. *Proc. VLDB Endow.*, 7:1529–1540, 2014. URL https://pages.cs.wisc.edu/~anhai/papers/chimera-vldb14.pdf. 2

[50] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59:64–73, 2016. URL https://arxiv.org/abs/1503.01817. 3

[51] Emiel van Miltenburg, Ákos Kádár, Ruud Koolen, and Emiel J. Krahmer. Didec: The dutch image description and eye-tracking corpus. In *COLING*, 2018. URL https://aclanthology.org/C18-1310/. 3

[52] Josiah Wang, Josiel Figueiredo, and Lucia Specia. MultiSubs: A large-scale multimodal and multilingual dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6776–6785, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.730. 3

[53] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, 2015. URL https://hal.archives-ouvertes.fr/hal-01196959/file/CEA_ICME2015.pdf. 2, 3

[54] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706, 2019. URL https://arxiv.org/abs/1901.06706. 2, 3

[55] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. URL https:

//openaccess.thecvf.com/content_cvpr_2017/papers/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.pdf. 6

[56] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41. 6

[57] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *ECCV*, 2020. 4

[58] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2066. URL https://aclanthology.org/P17-2066. 3

[59] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. doi: 10.48550/ARXIV.2205.01917. URL https://openreview.net/forum?id=Ee277P3AYC. 2

[60] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1204–1213, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Zhai_Scaling_Vision_Transformers_CVPR_2022_paper.pdf. 2

[61] Zizhao Zhang, Han Zhang, Sercan Ö. Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9291–9300, 2020. URL https://arxiv.org/pdf/1910.00701.pdf. 2

[62] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan T. Dumais. Meta label correction for noisy label learning. In *AAAI*, 2021. 5

[63] Shuai Zheng, F. Yang, Mohammad Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. *Proceedings of the 26th ACM international conference on Multimedia*, 2018. 3