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

Václav Košař
Antonín Hoskovec
Milan Šulc
Radek Bartyzal

ANKA KEMER Kadın Heybe
(Turkey)
'womens-belts'

Pánská kotníková obuv
(Czechia)
'mens-boots'

Ženski kopalni plašč
(Slovenia)
'womens-bathrobes'

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

Kuprinės Guess
(Lithuania)
'womens-backpacks'

Rock Is Dead 1 - Dojčenské
(Slovakia)
'baby-clothing'

The Largest Multilingual Image-text Classification Dataset and Benchmark

- 1.1M fashion items with 968k unique images and 1.01M unique texts
- 191 fine categories (15 shoes types)
- 13 languages (cz, sk, gr, hu, si, lt, lv, hr, bg, ee, tr, ro, es)
- Difficult e-commerce industry problem
- High-quality annotations

Table 1: Examples from GLAMI-1M.

item_id	image_id	geo	name	description	category	category_name	label_source
517876	488425 	gr	Κλειστά παπούτσια TOMS	Κλειστά παπούτσια TOMSΚλειστά παπούτσια TOMS -...	2811	boys-shoes	NaN
989034	863506 	lt	Big Star Woman's Singlet T-shirt 150048 Knitte...	Material: 95%COTTON5%ELASTANE Washing instruct...	53403	womens-tops-tank-tops-and-t-shirts	admin
483208	455633 	gr	BENCH Κάλτσες μαύρο λευκό	Υλικό: Ζέρσεϊ Έξτρα: Κεντημένο λογότυπο, Μαλακ...	132	womens-socks	admin
1009868	876723 	si	Kilpi Ženske športne jakne črna Rosa-W		86531	womens-sport-jackets	custom-tag
586781	544307 	hu	Női blúz ONLY	Új termék címkével.	6	womens-blouses-and-shirts	NaN
1121212	951403 	tr	Nonna Baby Cute Monnet 5 Li Zıbın Seti	Yeni sezon 5 parça zıbın seti,0-3 ay %100 pamu...	39412	baby-clothing	custom-tag

Dataset Comparison

- The largest multilingual image-text classification dataset.
- The second largest image-text classification dataset to Recipe1M+.
- GLAMI-1M has 75% of the training set, and 100% of the test set human labelled.

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

- The largest image-text fashion dataset (1.1M items).
- The finest grained categories (191), e.g.: 15 shoes types.
- The most languages (13).

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

Category Overview

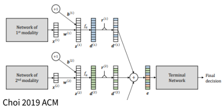
- Long-tailed class distribution across 191 class (~ exponential)

Table 4: 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

EmbraceNet Classification Baseline

- Baseline model EmbraceNet consumes mT5 and ResNeXt-50 embeddings inputs and predicts a class
- Non-human labels helping very little in this setup
- Challenging benchmark since performance only 69.7%



Choi 2019 ACM

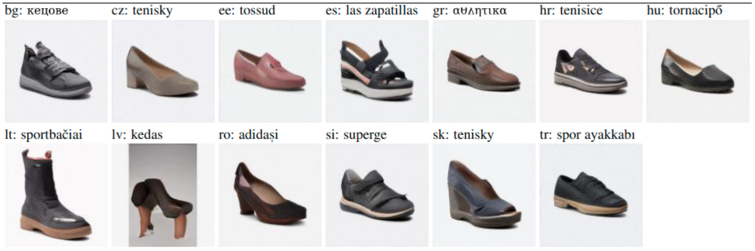
Table 5: Top-k accuracies of EmbraceNet with various input modalities, trained either on all labels (*all*) or human-labeled samples only (*hum.*).

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
CLIP Zero-shot Image + Text	0.323	0.745	-	-

Text-to-Image Generation Baseline

- Single GPU multilingual text conditioned diffusion to 128x128, mT5, 2x UNet models

Figure 1: Images generated by the Imagen-like model for the input "sneakers" translated into all 13 languages, 500 time steps of diffusion.



Conclusion

- A Multilingual alternative to Recipe1M+
- Larger alternative to FashionGen
- Challenging image-text classification benchmark
- Multilingual text-to-image dataset
- Future work: long-tail learning, adaptation to prior shift, learning from a combination of trusted (human) and noisy (rule-based) annotations.

Download GLAMI-1M paper and dataset at: https://github.com/glami/glami-1m