Supplementary Material to 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

1 Supplementary Tables and Figures

Table 1 visualizes dataset examples. Table 2 shows distribution of the *label_source* column. Table 3 results from zero-shot CLIP model baseline. Table 4 results of EmbraceNet with image and text on the input, stratified per-language. Table 9 lists country names corresponding to country codes for all 13 dataset languages.

2 Machine Translation Experiments

Some of the samples appear in the dataset in multiple languages. In these cases the underlying product is the same, but the title and description have been written in a different language often by a different seller. These samples can be identified by the image_id column. They have a different item_id, but identical image_id. The publicly available datasets for machine translation are typically based on Wikipedia, news feeds or scraped websites and are scarcely available for non-English languages [1, 4, 5, 7, 8]. The pair-wise counts of samples in our training and test splits can be found in Table 5 and Table 6 respectively. For some of the language pairs the counts are orders of magnitude higher than what was previously available.

For the baseline we use publicly available, pretrained M2M100 model [3]. M2M100 has been pretrained on the task of machine translation on a superset of the languages in our dataset. We have evaluated the BLEU score of the M2M100 model on the descriptions, on all of the possible pairs of languages in the test split of the dataset Table 7. We report the overall BLEU score of 1.87%. The highest BLEU of 9.96% was achieved in translation from Hungarian to Greek.

In Table 8 are samples of the M2M100 Translations. We show a successful translation from Hungarian to Greek in sample no. 1 and then we show a failure from Czech to Slovak, where the model shortened the text much below the threshold of 32 tokens. Interestingly, this is a mistake that we saw quite often, for some reason the decoded sequence from the model came out much shorter than both the input and targets, this definitely brought the BLEU score down. Another frequent mistake was a change in units of measure, for example we saw the model translate 25 mm as 2.5 mm. On the other hand the dataset frequently contains brand names and if the model did not change them, its BLEU score increased.

This experiment could be reformulated as a multilingual text generation conditioned on the image or the title of the product. However, such models are beyond the scope of this paper.

Table 1: Examples from GLAMI-1M.

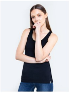
item_id	image_id	geo	name	description	category	category_name	label_source
517876	488425 	gr	Κλειστά παπούτσια TOMS	Κλειστά παπούτσια TOMSΚλειστά παπούτσια TOMS -...	2811	boys-shoes	NaN
989034	863506 	lt	Big Star Woman's Singlet T-shirt 150048 Knitte...	Material: 95%COTTON5%ELASTANE Washing instruct...	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

Table 2: Distribution of values in *label_source* column.

label_source	training set [%]	test set [%]
custom-tag	40.8	54.3
admin	34.3	45.7
NaN	24.4	0.0
combined-tag	0.5	0.0
quality-check	0.1	0.1

Table 3: Top-k accuracies of CLIP zero-shot classification baseline with various input modalities. Image+text variant is classification using unnormalized embedding vector summation of CLIP image and text embeddings. We used prompts "A photo of a category, a type of fashion product" as targets. We used aligned image (ViT-B/32) [6] and multilingual text (XLM-Roberta-Large-Vit-B-32) [2] CLIP embeddings.

Included modality/model	Top-1	Top-5
Text + Image	**0.323**	**0.745**
Image	0.289	0.718
Text	0.265	0.585

Table 4: Top-k accuracies of EmbraceNet with text and image inputs, trained on all labels, stratified per-language. We observe maximal 6% difference in Top-5 accuracy across different countries and 21% in Top-1 accuracy. We speculate that the reason may be variable quality of text embeddings and different distributions of test set samples between the countries.

k	cz	sk	ro	gr	hu	bg	hr	es	lt	si	lv	tr	ee
1	0.592	0.805	0.646	0.589	0.670	0.726	0.739	0.626	0.685	0.683	0.672	0.720	0.659
5	0.905	0.967	0.924	0.911	0.941	0.961	0.956	0.920	0.941	0.946	0.922	0.932	0.928

Table 5: Pairwise counts of samples in multiple languages, training set.

	cz	sk	ro	gr	hu	bg	hr	es	lt	si	lv	tr	ee
cz	0	4669	1249	977	2712	1797	1986	784	1877	1022	723	10	1148
sk	4669	0	2231	635	3882	2366	2526	722	2557	1024	485	31	1485
ro	1249	2231	0	766	1417	1276	1748	416	731	644	238	214	370
gr	977	635	766	0	825	1901	1231	648	1317	444	234	5	461
hu	2712	3882	1417	825	0	5458	2232	1076	3137	1937	613	730	1188
bg	1797	2366	1276	1901	5458	0	3235	1416	6880	2983	1086	548	3297
hr	1986	2526	1748	1231	2232	3235	0	1174	3319	3408	933	0	2095
es	784	722	416	648	1076	1416	1174	0	860	381	710	322	1099
lt	1877	2557	731	1317	3137	6880	3319	860	0	5640	5538	10	9176
si	1022	1024	644	444	1937	2983	3408	381	5640	0	1968	9	1737
lv	723	485	238	234	613	1086	933	710	5538	1968	0	5	5085
tr	10	31	214	5	730	548	0	322	10	9	5	0	4
ee	1148	1485	370	461	1188	3297	2095	1099	9176	1737	5085	4	0

Table 6: Pairwise counts of samples in multiple languages, test set.

	cz	sk	ro	gr	hu	bg	hr	es	lt	si	lv	tr	ee
cz	0	4669	1249	977	2712	1797	1986	784	1877	1022	723	10	1148
sk	4669	0	2231	635	3882	2366	2526	722	2557	1024	485	31	1485
ro	1249	2231	0	766	1417	1276	1748	416	731	644	238	214	370
gr	977	635	766	0	825	1901	1231	648	1317	444	234	5	461
hu	2712	3882	1417	825	0	5458	2232	1076	3137	1937	613	730	1188
bg	1797	2366	1276	1901	5458	0	3235	1416	6880	2983	1086	548	3297
hr	1986	2526	1748	1231	2232	3235	0	1174	3319	3408	933	0	2095
es	784	722	416	648	1076	1416	1174	0	860	381	710	322	1099
lt	1877	2557	731	1317	3137	6880	3319	860	0	5640	5538	10	9176
si	1022	1024	644	444	1937	2983	3408	381	5640	0	1968	9	1737
lv	723	485	238	234	613	1086	933	710	5538	1968	0	5	5085
tr	10	31	214	5	730	548	0	322	10	9	5	0	4
ee	1148	1485	370	461	1188	3297	2095	1099	9176	1737	5085	4	0

Table 7: BLEU scores (in %) of the M2M100 model on all pairs of languages, on left is the source language, on the top is the target language.

	cz	sk	ro	gr	hu	bg	hr	es	lt	si	lv	tr	ee
cz	nan	7.30	0.63	0.79	2.26	0.91	1.38	0.45	2.36	0.47	0.17	0.00	0.07
sk	6.68	nan	1.53	2.57	3.48	1.28	2.27	0.06	2.86	4.49	0.00	0.00	0.35
ro	0.57	2.49	nan	5.87	4.20	0.02	4.58	0.47	2.21	1.33	0.49	2.40	0.17
gr	1.70	4.48	3.86	nan	8.46	2.32	1.59	0.20	6.44	0.78	0.00	nan	0.03
hu	2.35	3.80	4.92	9.96	nan	1.78	2.91	0.56	3.88	0.23	0.00	1.32	0.12
bg	2.67	2.86	1.42	6.00	5.36	nan	7.35	0.34	8.89	4.37	1.98	2.01	0.23
hr	2.18	2.38	3.63	2.02	1.89	0.83	nan	0.36	1.71	2.24	0.09	0.00	0.17
es	0.11	0.07	0.27	0.45	0.35	0.08	0.22	nan	0.00	0.16	0.26	1.64	0.03
lt	2.56	3.64	2.29	3.93	3.30	4.23	2.13	0.00	nan	0.34	0.01	0.00	0.05
si	0.51	0.42	0.81	0.84	1.04	0.28	2.51	0.08	0.29	nan	0.72	nan	0.15
lv	0.06	0.00	0.14	0.26	0.00	0.08	0.00	0.12	0.00	0.16	nan	nan	0.05
tr	0.00	0.00	2.36	nan	0.64	0.88	0.00	2.24	0.00	nan	nan	nan	0.00
ee	0.15	0.32	0.01	0.02	0.09	0.04	0.06	0.00	0.06	0.13	0.10	0.00	nan

Table 8: Examples of M2M100 translations. Sample no. 1 is an example of a successful translation in the highest quality pair of languages based on the BLEU scores.

(Sample no.) Description	Text
(1) EN Translation	Color: white, Collection: Spring Summer 2020, Producer code: MM2T791, Fashion: Regular Fit
(1) HU Source Text	Szín: fehér , Kollekció: Tavaszi Nyár 2020 , Gyártókód: MM2T791, Fazon: Regular Fit
(1) GR Translation	Szín: λευκό, Kollekció: Tavaszi Nyár 2020, Gyártókód: MM2T791, Φά
(1) GR Target Text	Χρώμα: άσπρο, Συλλογή: Άνοιξη Καλοκαίρι 2020, Κωδικός παραγωγού: MM2T791, Μόδα: Regular Fit
(2) EN Translation	A dress with an A-line skirt will liven up your look wherever you go. The delicate and understated look of this dress is completed by the gold zipper on the front. Thanks to its cut, it also conjures up a beautiful figure.
(2) CZ Source Text	Šaty s áčkovou sukní oživí Tvůj vzhled, ať půjdeš kamkoli. Jemný a decentní vzhled těchto šatů doplňuje zlatý zip na přední části. Díky svému střihu Ti navíc vykouzlí krásnou postavu.
(2) SK Translation	Šaty s áčkovou sukní oživí Tvůj vzhled, ať půjdeš kamkoli.
(2) SK Target Text	Šaty s áčkovou sukňou oživia Tvoj vzhľad, nech sa pohneš kamkoľvek. Jemný a decentný vzhľad týchto šiat dopĺňa zlatý zips na prednej časti. Vďaks svojmu strihu Ti navyše vyčarujú krásnu postavu.

Table 9: Country names corresponding to country codes

Country code (geo)	Country name
cz	Czechia
sk	Slovakia
ro	Romania
gr	Greece
si	Slovenia
hu	Hungary
hr	Croatia
es	Spain
lt	Lithuania
lv	Latvia
tr	Turkey
ee	Estonia
bg	Bulgaria

References

[1] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.1. 1

[2] 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

[3] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48, 2021. URL https://arxiv.org/abs/2010.11125. 1

[4] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: extracting large parallel corpora from movie and tv subtitles. In *10th conference on International Language Resources and Evaluation (LREC'16)*, pages 923–929. European Language Resources Association, 2016. 1

[5] Paul Michel and Graham Neubig. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, 2018. 1

[6] 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. 4

[7] Jörg Tiedemann. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt. In *WMT*, 2020. URL https://aclanthology.org/2020.wmt-1.139.pdf. 1

[8] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, 2020. 1