Efficiency of Vision-Language Models on Image Caption Generation: Case of Lithuanian Language

Tautvydas Kvietkauskas[1], Pavel Stefanovič[1], Vytas Mulevičius[2], Artūras Nakvosas[2,3]

[1]Faculty of Fundamental Science, Vilnius Gediminas Technical University [2]Neurotechnology [3]Institute of Data Science and Digital Technologies, Vilnius University

Abstract

Vision-language models have gained significant popularity in recent years because they can simultaneously address problems related to image and text analysis. They combine computer vision and natural language processing techniques to perform tasks such as image description, picture-based question-answering systems, and multi-modal search. Recently, these models have become increasingly important in developing advanced applications, such as autonomous vehicles, medical diagnostics, and content management. Many Vision-language models are adapted to the most popular languages, such as English, Spanish, and Chinese, but lack integration with less popular languages, like Lithuanian. This study analysed the effectiveness of various Vision-Language models, such as BLIP, Gemma3, Qwen, and others, using pre-prepared data collected from Lithuanian news portals. Thus, to expand the research data, the Flickr8k dataset was selected, and its captions were translated into Lithuanian. The research dataset consists of photos associated with news articles and their corresponding captions below each image. Given that many models cannot generate captions in Lithuanian, a study was conducted to translate captions from Lithuanian to English. Traditional evaluation metrics, such as BLEU, METEOR, ROUGE, BERTScore and Sentence-BERT were used to evaluate the research results. The results of the experimental investigation show that models trained with languages of smaller countries, such as Lithuania, can be sufficiently accurate.

Background of the research

Experimental studies were conducted using two datasets: the newly collected dataset LT-news-500 and Flickr8k.

- LT-news-500 was collected manually and consisted of 500 images and 500 captions, with one caption per image. The distribution within this dataset was as follows: TV3 – 104, LRT – 65, Lrytas – 55, Delfi – 52, Verslo Žinios – 51, Alfa – 49, 15min – 45, Kas vyksta Kaune – 40, and LNK – 39. The average length of the caption is equal to 4.44 tokens.
- The second dataset, Flickr8k, consists of 8,091 images, each with five captions, resulting in a total of 40,455 entries. The Flickr8k dataset is in English, so all entries were translated into Lithuanian using the Python library deep-translator.

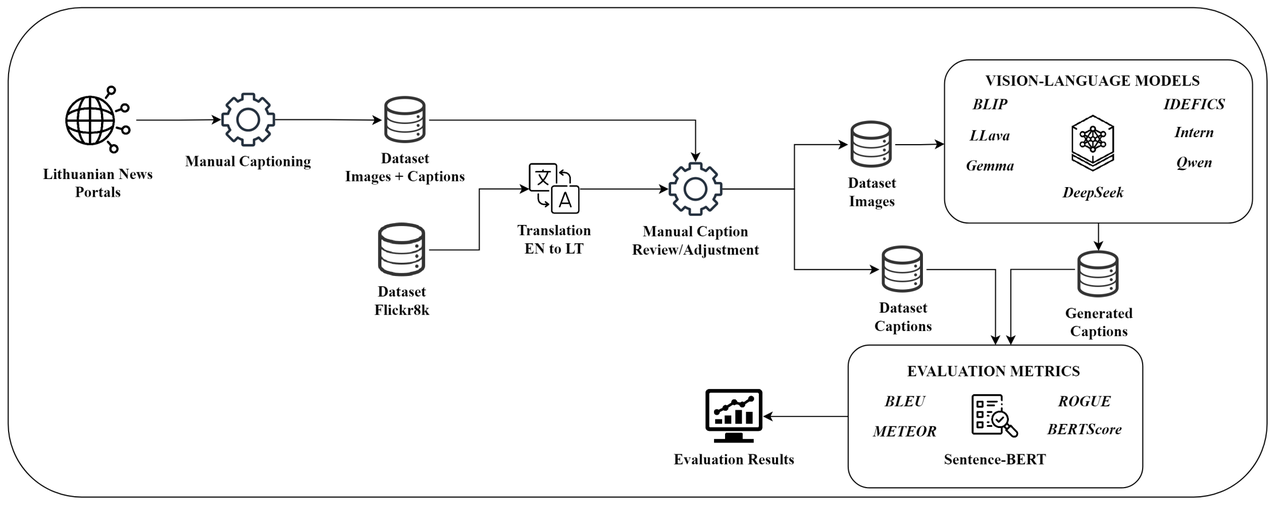

Figure 1. Workflow of the experiment

Results for all models

During the experimental study, open VLM models of different sizes were tested, with parameter sizes up to 40 B. Extremely large models were not used, since their purpose usually involves more in-depth image-to-text tasks, such as OCR. The summarised results using the LT-news-500 dataset are presented in Table 1.

Table 1. Experiment results using LT-news-500 dataset

ID	GROUP	VISION-LANGUAGE MODEL	PARAMETERS	LT support	Prompt language	Avg. BERTScore F1	Avg. Sentence-BERT MRL	Avg. Sentence-BERT MPNET	Gener. Token Mean	Gener. Token Variance	Gener. LT caption
1	BLIP	Salesforce/blip-image-captioning-base	≈0.086B	No	No prompt	0.83	0.2	0.62	8.95	8.16	–
2		Salesforce/blip-image-captioning-large	≈0.21B	No	No prompt	0.83	0.2	0.63	11.62	2.12	–
3		Salesforce/blip-vqa-base	≈0.22B	No	No prompt	**0.80**	**0.15**	**0.49**	**1.71**	**1.31**	–
4		Salesforce/blip-vqa-capfilt-large	≈0.9B	No	No prompt	**0.80**	**0.15**	**0.49**	**1.71**	**1.31**	–
5		Salesforce/blip2-opt-2.7b	≈2.7B	No	No prompt	0.84	0.21	0.64	9.15	5.39	–
6		Salesforce/blip2-opt-6.7b	≈6.7B	No	No prompt	0.84	0.22	0.65	9.25	3.47	–
7		Salesforce/blip2-flan-t5-xl	≈3B	No	No prompt	0.83	0.19	0.61	10.32	7.5	–
8		Salesforce/blip2-flan-t5-xxl	≈12B	No	No prompt	0.83	0.18	0.57	10.14	10.8	–
9	LLava	llava-hf/llava-1.5-7b-hf	≈7B	No	EN	0.83	0.17	0.55	8.51	16.71	–
10		llava-hf/llava-v1.6-mistral-7b-hf	≈7B	No	EN	0.84	0.21	0.67	4.0	1.11	–
11		llava-hf/llava-v1.6-vicuna-13b-hf	≈13B	No	EN	0.84	0.22	0.67	5.13	23.01	–
12	GEMMA	**google/gemma-3-4b-it**	≈4B	**Yes**	**LT**	**0.87**	**0.26**	**0.64**	5.0	1.26	**93.4 %**
					EN	0.84	0.21	0.64	6.78	0.86	–
13		**google/gemma-3-12b-it**	≈12B	**Yes**	**LT**	**0.88**	**0.37**	**0.71**	6.13	2.35	**94.8 %**
					EN	0.84	0.21	0.66	7.3	3.11	–
14		**google/gemma-3-27b-it**	≈27B	**Yes**	**LT**	**0.87**	**0.37**	**0.68**	**7.09**	**3.3**	**95.2 %**
					EN	0.84	0.21	0.64	8.87	0.6	–
15	Qwen	**Ertugrul/Qwen2.5-VL-7B-Captioner-Relaxed**	≈7B	**Yes**	**LT**	**0.86**	**0.24**	**0.61**	**6.6**	**12.12**	**95.4 %**
					EN	0.84	0.21	0.66	9.1	8.39	–
16		**Qwen/Qwen2-VL-7B-Instruct**	≈7B	**Yes**	**LT**	**0,87**	**0.28**	**0.65**	5.3	5.24	**86.2 %**
					EN	0.84	0.23	0.68	6.96	3.96	–
17		**Qwen/Qwen3-VL-2B-Instruct**	≈2B	**Yes**	**LT**	**0.84**	**0.16**	**0.46**	3.28	4.32	**84.8 %**
					EN	0.84	0.23	0.67	4.44	1.34	–
18		**Qwen/Qwen3-VL-32B-Instruct**	≈32B	**Yes**	**LT**	**0.87**	**0.36**	**0.72**	**6.7**	**2.36**	**97.8 %**
					EN	0.87	0.23	0.67	9.27	3.14	–
19	DeepSeek	deepseek-ai/deepseek-vl-7b-chat	≈7B	No	EN	0.84	0.21	0.66	9.13	8.79	–
20	IDEFICS	HuggingFaceM4/idefics2-8b	≈8B	No	EN	0.84	0.20	0.63	9.17	16.1	–
21	Intern	**OpenGVLab/InternVL3_5-8B**	≈8B	**Yes**	**LT**	**0.85**	**0.21**	**0.57**	**5.64**	**219.42**	**87.4 %**
					EN	0.84	0.22	0.66	8.39	2.57	–
22		**OpenGVLab/InternVL3_5-14B**	≈14B	**Yes**	**LT**	**0.86**	**0.24**	**0.60**	**5.16**	**52.15**	**93.6 %**
					EN	0.84	0.22	0.66	8.22	1.54	–
23		**OpenGVLab/InternVL3_5-30B-A3B**	≈30B	**Yes**	**LT**	**0.86**	**0.26**	**0.61**	**4.36**	**31.33**	**86.2 %**
					EN	0.84	0.22	0.67	7.76	2.38	–
24		**OpenGVLab/InternVL3_5-38B**	≈38B	**Yes**	**LT**	**0.87**	**0.29**	**0.67**	**5.5**	**31.66**	**95.2 %**
					EN	0.84	0.23	0.68	8.72	1.72	–

Influence of the model size

To assess the influence of model size, a paired t-test was performed between models of different sizes, with a 95% confidence interval. Experiments were conducted separately for the two prompt languages used in the models, Lithuanian and English. The results are presented in Tables 2 and 3.

Table 2. Comparison of results (p-value) using the Lithuanian language

Group	Model ID Pairs	Evaluation metrics		
		BERTScore F1	SentenceBERT_MRL	SentenceBERT_MPNET
Gemma	12 vs 13	≈0.00	≈0.00	≈0.00
	12 vs 14	≈0.00	≈0.00	≈0.00
	13 vs 14	≈0.00	0.8	≈0.00
Qwen	15 vs 16	≈0.00	≈0.00	≈0.00
	15 vs 17	≈0.00	≈0.00	≈0.00
	15 vs 18	≈0.00	≈0.00	≈0.00
	16 vs 17	≈0.00	≈0.00	≈0.00
	16 vs 18	≈0.00	≈0.00	≈0.00
	17 vs 18	≈0.00	≈0.00	≈0.00
Intern	21 vs 22	≈0.00	≈0.00	≈0.00
	21 vs 23	≈0.00	≈0.00	≈0.00
	21 vs 24	≈0.00	≈0.00	≈0.00
	22 vs 23	0.99	0.03	0.3
	22 vs 24	≈0.00	≈0.00	≈0.00
	23 vs 24	≈0.00	≈0.00	≈0.00

Table 3. Comparison of results (p-value) using the English language

Group	Model ID Pairs	Evaluation metrics		
		BERTScore F1	SentenceBERT_MRL	SentenceBERT_MPNET
Gemma	12 vs 13	0.75	0.28	≈0.00
	12 vs 14	0.01	0.63	0.26
	13 vs 14	0.05	0.49	0.03
Qwen	15 vs 16	≈0.00	≈0.00	≈0.00
	15 vs 17	0.08	≈0.00	0.29
	15 vs 18	0.34	≈0.00	≈0.00
	16 vs 17	≈0.00	0.56	0.02
	16 vs 18	≈0.00	0.46	0.04
	17 vs 18	≈0.00	0.98	0.36
Intern	21 vs 22	0.35	0.61	0.76
	21 vs 23	0.05	0.08	0.22
	21 vs 24	0.18	≈0.00	≈0.00
	22 vs 23	0.24	0.22	0.32
	22 vs 24	0.59	≈0.00	≈0.00
	23 vs 24	0.46	0.20	0.03

Influence of the prompt language

The VLM models, such as Gemma, Qwen, and Intern, that support the Lithuanian language prompts have been compared with results from the English-language prompts to assess their influence on language choice. The results are presented in Table 4.

Table 4. Comparison of results (p-value) between Lithuanian and English

Group	Model ID	Evaluation metrics		
		BERTScore F1	SentenceBERT_MRL	SentenceBERT_MPNET
Gemma	12	≈0.00	≈0.00	0.95
	13	≈0.00	≈0.00	≈0.00
	14	≈0.00	≈0.00	≈0.00
Qwen	15	≈0.00	≈0.00	≈0.00
	16	≈0.00	≈0.00	≈0.00
	17	0.02	≈0.00	≈0.00
	18	≈0.00	≈0.00	≈0.00
Intern	21	≈0.00	0.57	≈0.00
	22	≈0.00	0.01	≈0.00
	23	≈0.00	≈0.00	≈0.00
	24	≈0.00	≈0.00	0.43

Conclusions and Future works

BLEU, METEOR, and ROUGE were used for evaluation in Lithuanian and English but provided no meaningful results, as their scores were near zero, therefore, these metrics are not shown in the tables. Experiments with Lithuanian captions (LT-news-500) were also repeated on the Flickr8k dataset translated into Lithuanian, it shows similar outcomes. Future work includes presenting Flickr8k results and fine-tuning the top three Vision-Language models with Lithuanian data to assess performance improvements.

Acknowledge

The scientific research was carried out using the AI and HPC equipment of the VILNIUS TECH Digital Defense Competence Center, acquired through the project 'Implementation of Mission-Based Science and Innovation Programs' No. 02-002-P-0001 under the theme 'Secure and Inclusive e-Society' DIGI-DEFENSE, funded by the Economic Recovery and Resilience Plan 'New Generation Lithuania'.

Contributors

VILNIUS TECH Faculty of Fundamental Sciences

Vilniaus universitetas

NEUROtechnology