Performance of the Multilingual LM mBLIP on the WIT dataset

Guseinova, Blicks, Kroiß

Agenda

- Intro
- Background
- Dataset
- Prompting
- Evaluation
- Analysis
- Limitation
- Conclusion

Intro

20 languages

+ {minimal prompt, context prompt, transl+context prompt}

mBlip

generated caption

ROUGE, METEOR, Bertscore ... (original_caption, generated_caption)

Background - Initial Approach

BLIP-2:

- monolingual without it being specified in the paper
- only noticeable by inspecting the dataset
- solution: mBLIP (Geigle et al. 2023)

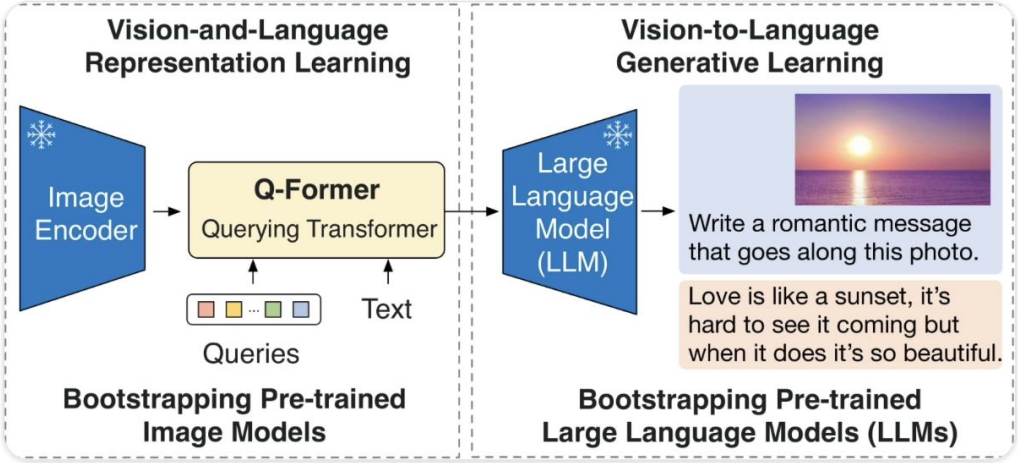
Background - mBLIP

need for mBLIP:

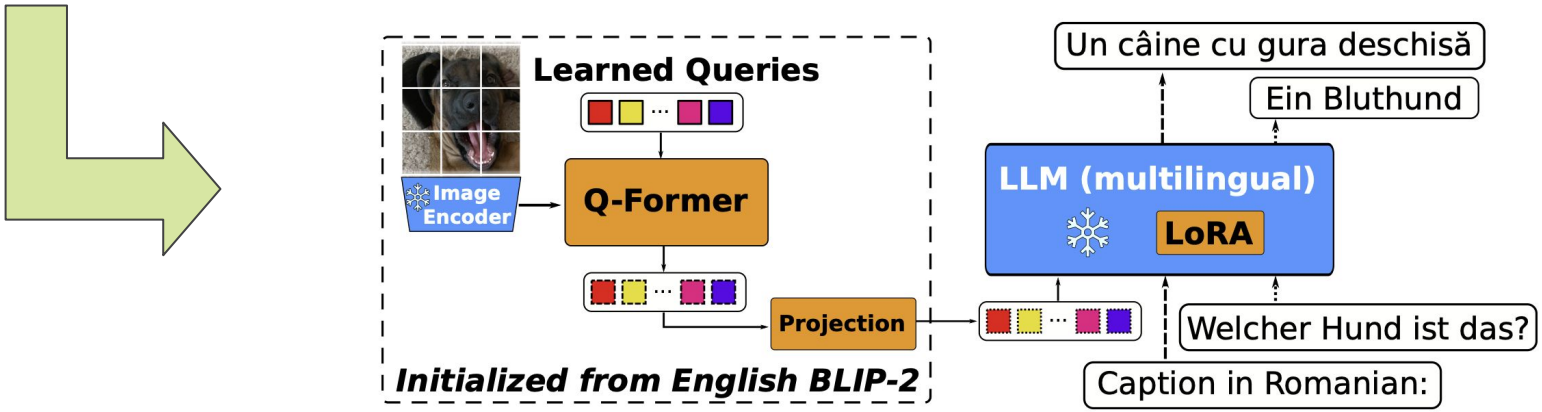
- BLIP and other similar LLMs highly insufficient in regards to multilingual capabilities

architecture:

- Based on the modular BLIP2 structure (VIT, Q-Transformer, LLM)
- can be trained in a few days on consumer-grade hardware because of fewer paramters



BLIP2 and mBLIP architecture as taken from their papers.



Background - mBLIP (cont.)

Training Data:

- MSCOCO dataset
- Translation via NLLB (Costa-jussà et al., 2022)
- 95 languages + English
- mC4 corpora (Xue et al., 2021) and more

Testing

Training Tasks:

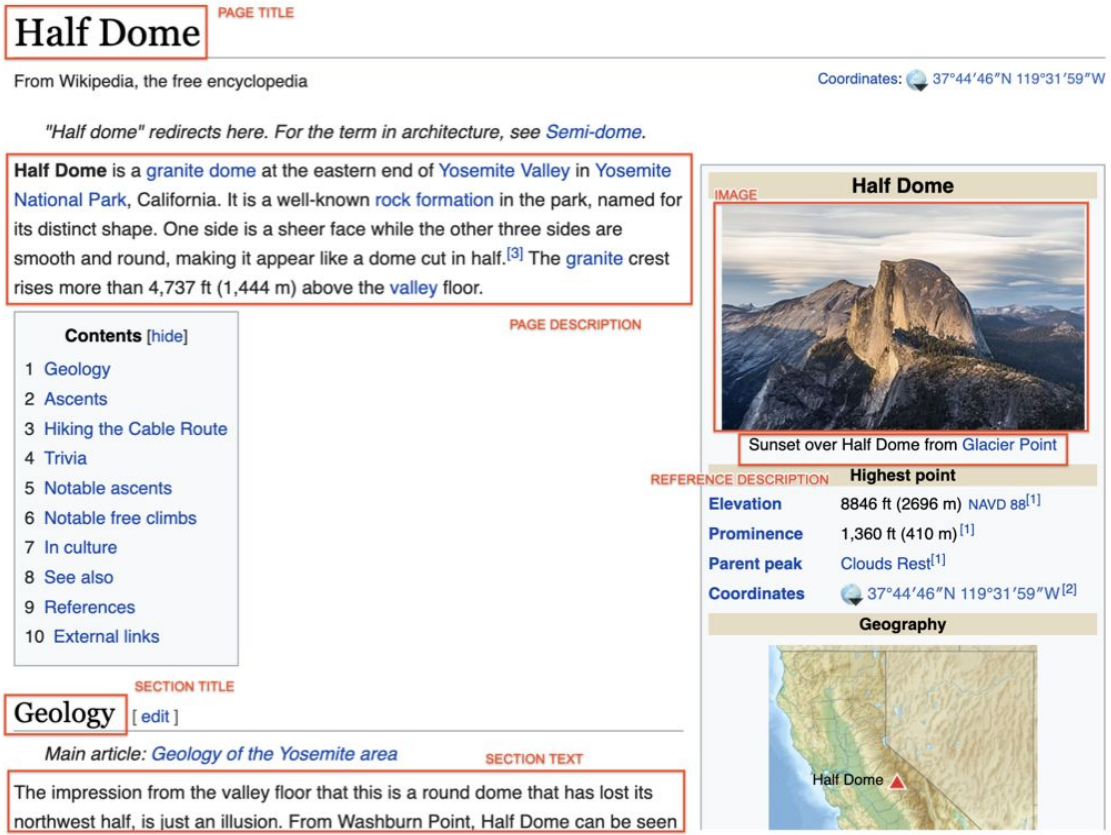
- Image Captioning
- VQA
- Matching

mBLIP evaluation:

- better than state-of-the-art multilingual vision-models and strong English Vision LLMs

Dataset WIT

- language, image, reference description (=original image caption)
- additionally for longer context: page title, section title



Dataset preparation

- clean the data from NaN values
- drop the rows with invalid URLs
- keep only the 20 most frequent languages
- take the equal amount entries for each language – ~800 each

:

	language	image_url	caption_reference_description	page_title	section_title
0	en	https://upload.wikimedia.org/wikipedia/commons...	The second of two images depicting the formal ...	Zolgokh	Tsagaan sar
1	en	https://upload.wikimedia.org/wikipedia/commons...	Local traditional costume.	Vranje	Culture
2	en	https://upload.wikimedia.org/wikipedia/commons...	A 1:25 scale 1960 Ford wagon. Made out of acet...	Hubley Manufacturing Company	Kits and promotionals
3	en	https://upload.wikimedia.org/wikipedia/commons...	Rainbow trout	Püttlach	Nature
4	en	https://upload.wikimedia.org/wikipedia/commons...	Marcel Stephan (Spike*D) at Airbeat One Festiv...	Gestört aber GeiL	Discography
...	...	...	...	...	...
15855	vi	https://upload.wikimedia.org/wikipedia/commons...	María Teresa Vera với Rafael Zequeira vào năm ...	María Teresa Vera	Ý kiến của bà
15856	vi	https://upload.wikimedia.org/wikipedia/commons...	AK-101(Trên) và AK-102(Dưới)	AK-101	Phiên bản cạc-bin AK-102
15857	vi	https://upload.wikimedia.org/wikipedia/commons...	Một loại tương cà chua có chứa cà chua xay nhu...	Tương cà chua	Mô tả
15858	vi	https://upload.wikimedia.org/wikipedia/commons...	Một cái khuy trên cổ áo ve nhọn cho phép người...	Tuxedo	Phụ kiện
15859	vi	http://upload.wikimedia.org/wikipedia/commons/...	Hinter Tierberg	Alpes uranaises	Hình thể

15860 rows × 5 columns

Prompting - general

- 20 of most common languages
  [English, Arabic, Catalan, Czech, German, Spanish, French, Hungarian, Italian, Hebrew, Japanese, Dutch, Polish, Portuguese, Russian, Swedish, Ukrainian, Vietnamese, Chinese (Simplified), Chinese (Traditional)]
- Translation of prompts with NNLB: the same library used in mBlip paper

Prompting - Simple Prompting

- motivation: the only context necessary to push the model towards the required language
- “On the picture” in all 20 languages as the prompt

Prompting - Context Prompting

Minimal Prompting results in simple Image Captions:

“A group of men sitting around a table.”

While Captions from Wikipedia can have far more additional information:

“February, 1958. German scientists repatriated from Sukhumi.”

→ Page and section information as additional context:

“Page Title: {}, Section Title: {}. Caption the image:”

Prompting - Context Prompting

mBLIP able to recognize specification of target language in prompt

→ Additional context prompt with target language instead of translation:

“Page Title: {}, Section Title: {}. Caption the image in [language]:”

→ possible Problem:

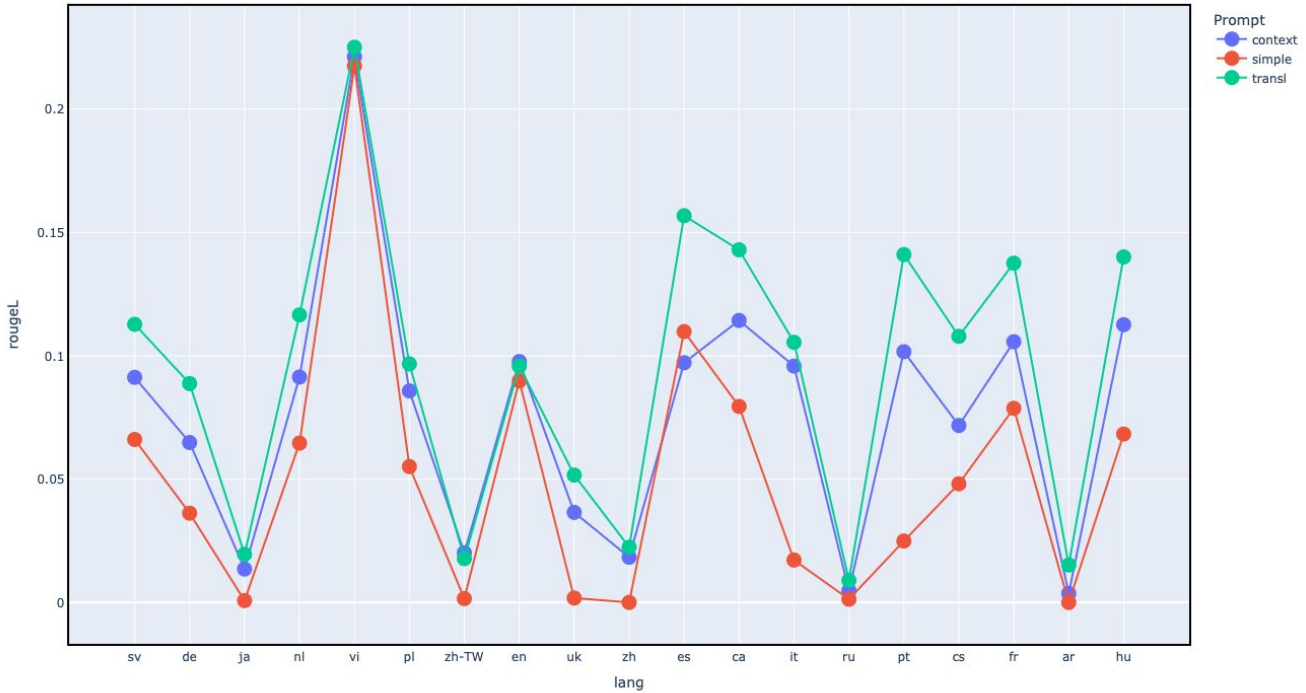
‘language’ mentioned in context → Output in that language

Evaluation - Metrics

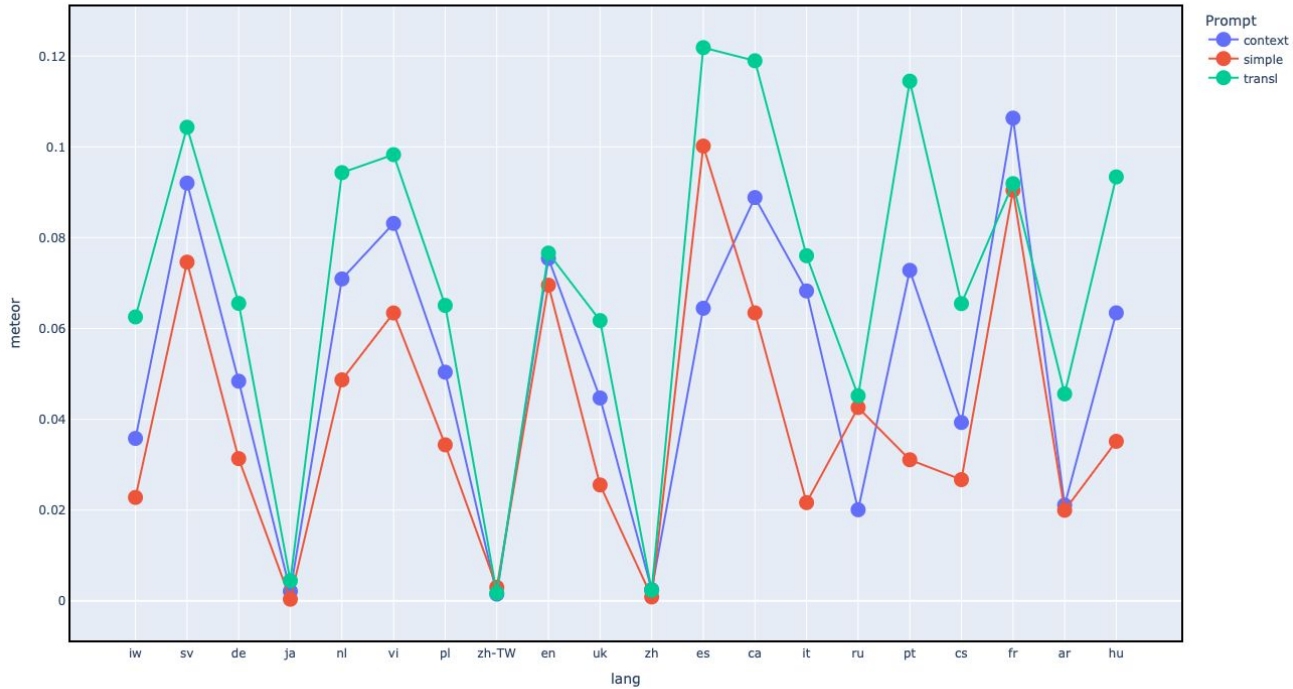
- ROUGE
  - measures n-gram overlap
  - language-independent
  - ROUGE-L: LCS
- Meteor
  - unigrams
  - mean of recall and precision with recall being weighted higher than precision
  - language-independent
- Bertscore
  - calculates similarity score
  - uses pre-trained embeddings from BERT
  - multilingual
  - semantic dimension taken into account

Evaluation - Results
ROUGE-L

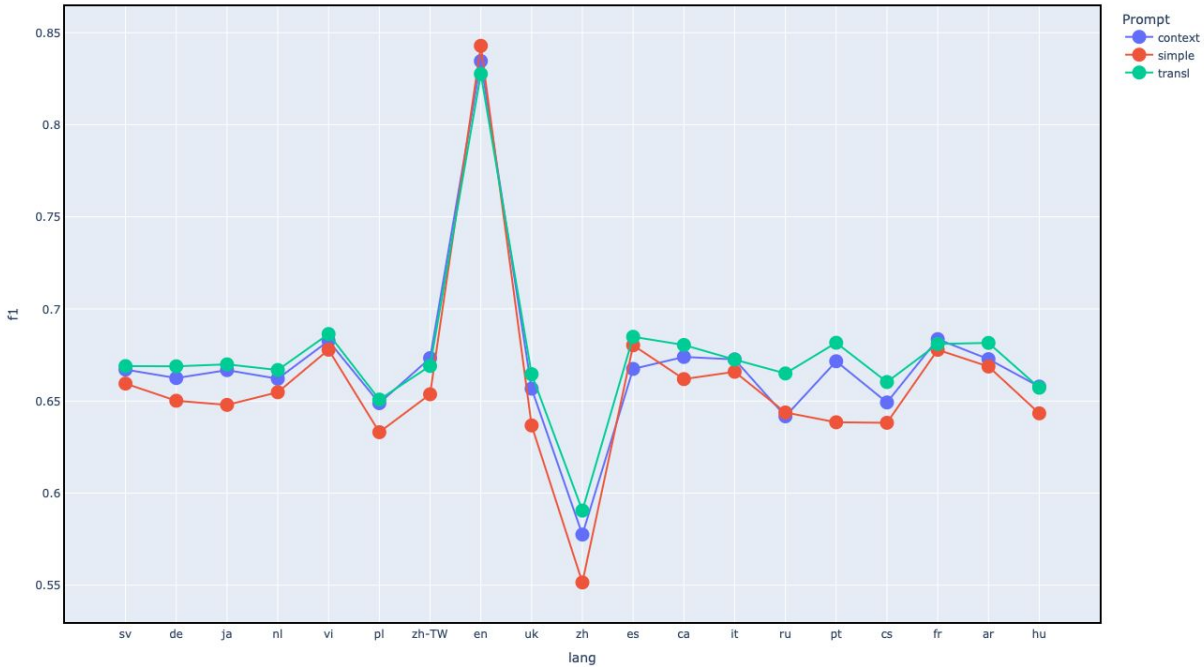
Metrics usually considered excellent at around 0.5



Evaluation - Results
METEOR



Evaluation - Results
Bertscore



Evaluation - Analysis (graphics)

- Tendency:
  - Translate-Prompt better than other languages
    - (However, not necessarily for English)
  - Same pattern across prompts
- Discrepancy:
  - Our model captures the semantic dimension (Bertscore)
  - Bad results: generated texts vastly different from reference texts
    - Not ideal for ROUGE, METEOR, etc.

Analysis: Named Entities in original and generated captions

- spaCy NER pipelines: available for 14/20 languages
- Multi-language, Chinese and Japanese models
- drastic difference between percentage of NE in original and generated captions, decreasing as the context become more helpful
- however, as NER improves in 10 times, the results do not improve as much => the amount of NE is not a decisive factor

	in %	ca	de	en	es	fr	it	nl	pl	pt	ru	sv	uk	ja	zh-TW
multi-language	original	12.92	16.05	14.48	13.07	12.29	14.73	14.94	17.17	13.34	13.68	15.17	14.52	19.06	21.71
	simple	0.82	0.79	1.19	0.49	0.67	0.42	0.46	2.09	5.47	1.01	3.44	0.98	17.34	31.03
	context	4.83	5.54	5.14	8.19	3.4	5.08	3.5	9.29	5.02	13.32	5.76	7.37	21.17	29.39
	transl	9.03	9.91	5.86	9.41	11.21	9.26	6.1	11.98	7.18	13.46	10.67	17.9	14.49	26.64

Analysis: Named Entities in original and generated captions (cont)

japanese	13.68%	chinese	18.54%
	2.43%		6.21%
	15.5%		12.77%
	10.14%		10.71%

Analysis: language of the original vs generated captions

Prompt:

“Page Title: Brandenburg-Prussia, Section Title: Dutch and Scanian Wars. Caption the image:”

Output:

“Een schilderij van een paard en een man die op een paard rijden”

→ Expected English output, but got Dutch

Analysis: language of the original vs generated captions

Use of state-of-art Language Detection Libraries

LangDetect: very wide spread results from ~0.005 to up to ~0.97

Manually check revealed some detection results being wrong

→ low Accuracy maybe because of short captions

Check for English with ASCII Encoding:

```
Simple:     0.0
Context:    0.022486772486772486
Transl:     0.12301587301587301
```

Limitations

Highly specific caption descriptions in the WIT dataset:

> “Great Sleigh Drive (1678):Frederick William pursues Swedish troops across the frozen Curonian Lagoon; fresco by Wilhelm Simmler, ca. 1891” or
>
> “November, 2017”

→ low evaluation scores

- Troubles with accessing URLs - TimeOuts and ConnectionError
- High Power requirements - Long running time
- CIDEr (Consensus-based Image Description Evaluation) metric used in mBlip
  - requires multiple reference captions to be effective

Conclusion

So, how good is the multilingual model? Is there any bias towards english?

- the results heavily rely on the metric used
- for our most reliable metric, Bertscore, the results for English still spike despite multilingual attempts (but less on average, although the metrics are not comparable)

XM-3600 (CIDEr, mBlip)		WIT (Bertscore, transl, F1)	
en	35-avg	en	19-avg
80.17	26.77	82.76	66.72

Future work

- evaluate the data used in mBlip paper with Bertscore to have the possibility to compare reported results with our findings
- filter out generations in the wrong language and perform evaluation again

Thank you for your attention!