Analyzing Cross-Language Efficiency of BLIP-2 Through the WIT Dataset

Erweiterungsmodul, SS2024

Luca Blicks, Anita Guseinova, Julius Kroiß

Motivation

- LLMs tend to underperform on the languages that are not English
- case study: Don't Trust ChatGPT when Your Question is not in English: A Study of Multilingual Abilities and Types of LLMs (Zhang et al., 2023)

Task	**En**	Fr	De	Es	Ja	Zh
MR	**0.90**	0.80	0.78	0.80	0.82	0.78
CSR	**0.68**	0.58	0.52	0.54	0.48	0.52
KA	**0.96**	**0.96**	0.94	0.94	0.80	0.68

Table 1: Accuracy for TE tasks: math reasoning (MR), commonsense reasoning (CSR), and knowledge access (KA).

Research questions

- How good is BLIP-2’s multilingual performance on image captioning tasks?
- Is multilingual BLIP-2’s performance on the same task better and where exactly is it better?
- In particular, is there any bias towards the English language?
- ITA in case there is a vast difference in performance, is it possible to fine-tune the model to achieve better results?

Dataset

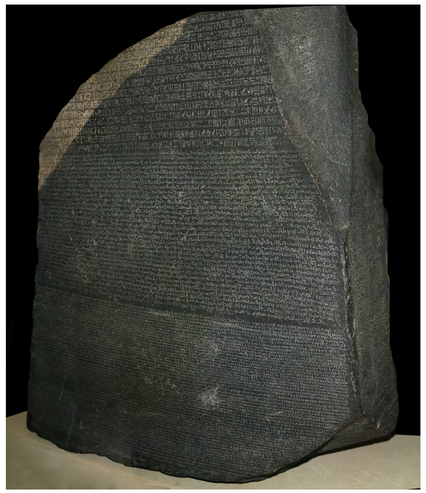
- Wikipedia-based Image Text Dataset (WIT). WIT is composed of a curated set of 37.6 million entity rich image-text examples with 11.5 million unique images across 108 Wikipedia languages.
  - *language*: language code depicting wikipedia language of the page
  - *image_url*: URL to wikipedia image
  - *caption_reference_description*: the caption that is visible on the wiki page directly below the image

Task, input and output

- Task: multilingual image captioning
- input: image with a prompt to have it described
- output: description of the image
- evaluation: ROUGE score between the generated and existing caption

Toy example for English

original caption:
The Rosetta Stone in the British Museum.



+ prompt: “Describe what is on the picture”

BLIP-2

ROUGE(“A beautiful stone”, “The Rosetta Stone in the British Museum”)

Thank you for your attention!