Model Card for BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

The BLIP-2 multimodal foundation model leverages frozen image encoders and LLMs to support various image-to-text generation tasks in zero-shot manner.

On this model card, you can learn more about how this model was trained, its capabilities, its intended use, and its limitations.

Model Details

Organization
Salesforce Research

Model date
January 30, 2023

Model type
Vision-and-Language model

Input
Image, text, or both

Information about LLM parameters
BLIP-2 OPT-2.7B, BLIP-2 OPT-6.7B, BLIP-2 FlanT5-3B, BLIP-2 FlanT5-11B

Output
Text. The model supports image-to-text generation.

Read the full paper here:
https://arxiv.org/abs/2301.12597

Access the public code here:
https://github.com/salesforce/LAVIS/tree/main/projects/blip2

Citation details:

Title:	BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models
Authors:	Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi
Year:	2023

License
BSD 3-Clause

Questions/Comments
junnan.li@salesforce.com

Intended Use

1. A research prototype for various downstream image-to-text generation tasks, including but not limited to:
   a. Image Captioning;
   b. Visual Question Answering;
   c. Multimodal Conversation.
2. Improvement of other vision-language applications through fine-tuning on another task or other data, e.g., fine-tuning BLIP-2 to generate product descriptions.

Out-of-scope use cases

- It should not be directly used in real-world applications without human supervision.
- It should not be used to collect or create
  - private personal information, such as specific people's names and other profile data; or
  - sensitive data, such as government-issued identification numbers, racial or ethnic origin, political opinions, religious or philosophical beliefs, and health information.
- It should not be used to promote or profit from:
  - violence, hate, and division;
  - environmental destruction;
  - abuse of human rights; or
  - the destruction of people's physical and mental health.

Training Data

The model is trained on 129 million image-text pairs from a variety of sources including COCO, Visual Genome, CC3M, CC12M, SBU, and 115 million images from the LAION400M dataset. The quality of training data has been bootstrapped using the CapFilt method. Please see the paper for details.

Metrics

We evaluate the accuracy performance of BLIP-2 models on a wide range of vision-language downstream tasks, and report the standard evaluation metrics. BLIP-2 achieves state-of-the-art performance on the following image-to-text generation tasks:

- On zero-shot VQA-v2 test-dev set, BLIP-2 FlanT5-11B achieves **65.0% VQA accuracy**.
- On zero-shot image captioning with NoCaps val set, BLIP-2 FlanT5-11B achieves **121.6 CIDEr score**.
- On fine-tuned image captioning with COCO Karpathy test set, BLIP-2 OPT-2.7B achieves **145.8 CIDEr score**.

Note that CIDEr score is a metric that measures the similarity between model-generated captions and human-written reference captions, where a higher score indicates better model performance. Please see the paper for more details on BLIP-2 evaluation.

Ethical Considerations

- **Dataset bias.** The training datasets in our study include image and alt-text pairs automatically collected from the internet. We have performed dataset filtering to remove harmful content. However, the dataset still possibly contains bias including stereotypes based on gender, race, sexual orientation, age, or similar demographic features. As such, the models trained on such data are potentially vulnerable to generating equivalently inappropriate content or replicating inherent biases in the underlying data.

- **Risks and Limitations from LLMs.** BLIP-2 uses off-the-shelf pre-trained language models including OPT and FlanT5. Therefore it inherits the same risks and limitations from these LLMs as written below.

- **Risks and Limitations from FlanT5.** “Language models, including Flan-T5, can potentially be used for language generation in a harmful way. Flan-T5 should not be used directly in any application, without a prior assessment of safety and fairness concerns specific to the application.“ (Chung et al. 2022)

- **Risks and Limitations from OPT.** “Like other large language models for which the diversity (or lack thereof) of training data induces downstream impact on the quality of our model, OPT-175B has limitations in terms of bias and safety. OPT-175B can also have quality issues in terms of generation diversity and hallucination. In general, OPT-175B is not immune from the plethora of issues that plague modern large language models.” (Zhang et al. 2022)

Caveats and Recommendations

- BLIP2 has not been tested in real world applications. It should not be directly deployed in any applications. We recommend that practitioners should first carefully assess the safety and fairness of the model in relation to the specific context within which they’re being deployed. We recommend that practitioners using BLIP-2 in real-world scenarios bear in mind that its generated outputs should be only taken as reference and that domain experts be engaged for further correctness- and security-checking.

- We also recommend that the data be further screened to fine-tune BLIP-2, including sensitive data cleaning and bias mitigation.