License: Apache 2.0

PaLI

Model Card Authors: Keran Rong , Xi Chen , Xiao Wang

PaLI aims to train a series of increasingly larger (capacity-wise) multimodal models with outstanding performance in vision-only, language-only, and vision + language tasks, by fully reusing existing high-capacity unimodal backbones.

Model Summary

Model Summary

MODEL ARCHITECTURE	OUTPUT(S)
PaLI is a multimodal sequence-to-sequence Transformer model derived from the T5 encoder-decoder architecture. It takes text tokens and ViT dense image embeddings as inputs to an encoder and autoregressively predicts discrete text tokens with a decoder.	Autoregressively generated text tokens and scores.

Model Data

TRAINING DATASET OVERVIEW

The model is pre-trained on the following mixture of datasets, more see our paper.

WebLI
WebLI (Web Language Image) is a web-scale multilingual image-text dataset, built from the public web, including image bytes, image-associated texts (alt-text, OCR, page title), 109 languages and many other features. The dataset is deduped on 68 common vision/vision-language tasks, and has no privileged data with careful RAI considerations. Its data card has been attached in the paper.

PaLM Dataset
High-quality multilingual documents from filtered webpages used in GLaM (Du et al., 2022) and PaLM (Chowdhery et al., 2022).

CC3M-35L
Curated English image-alt_text pairs from webpages (Sharma et al., 2018). We used the GCP Translation API to translate into 34 additional languages.

VQ^{2}A-CC3M-35L
A 100M random subset of VQ2A-CC3M (Changpinyo et al., 2022a), translated into the same additional 34 languages as CC3M-35L, using the GCP translation API.

Open Images
A public annotated image dataset (Krasin et al., 2017). In PaLI pre-training, we derived examples from Open images for English-only object aware VQA tasks and object detection tasks.

Visual Genome
A public annotated dataset (Krishna et al., 2016). In PaLI pre-training, we derived examples from Open images for object detection tasks.

Object365
A public large-scale object detection dataset (Shao et al., 2019). In PaLI pre-training, we derived examples from Open images for object detection tasks.

FINE-TUNING and EVALUATION DATASET

Vision + language tasks
- **Image captioning (English)**: COCO, NoCaps, TextCaps
- **Image captioning (multilingual)**: Crossmodal-3600
- **Visual question answering (English)**: VQAv2, OKVQA, TextVQA, VizWiz-QA
- **Visual question answering (multilingual)**: xGQA, MaXM

Vision-only tasks
- **Image classification (fine-tuning)**: ImageNet, ImageNet-V2, ObjectNet, ReaL
- **Image classification (zero-shot)**: ImageNet, ImageNet-V2, ImageNet-R, ImageNet-A, ImageNet-Sketch, ObjectNet, ReaL, VTAB

Language-only tasks
- **Natural language inference (English)**: SuperGLUE
- **Natural language inference (multilingual)**: XNLI
- **Question Answering (multilingual)**: XQuAD, TyDiQA

Model Creation & Maintenance

MODEL INITIALIZATION

The largest PaLI model is initialized from existing high-capacity unimodal backbones: mT5-XXL and ViT-e (scaled from ViT-G as “ViT-enormous”).

MODEL STATUS

Limited maintenance: The model will not be updated, but any technical issues will be addressed.

Version	1.0
Release Date	2022-09
Update Cadence	N/A. Static

MODEL STATS

The model has 17B parameters: 13B from the language backbone and 4B from the visual backbone.

Parameters	17 billion

Model Usage & Limitations

SENSITIVE USE	KNOWN LIMITATIONS	ETHICAL CONSIDERATIONS & RISKS
The model inherits the safety benefits and safety risks associated with large language models (mT5-XXL) and vision-language models (ViT-G). The model should not be used for downstream applications without prior assessment and mitigation of downstream application-specific security and fairness concerns.	**Additional Notes:** See the “Limitation” section of the paper for details.	• The model is trained on large, often noisy, image-text datasets that are known to contain biases regarding people of different backgrounds. • Inherited risk from the larger language model, including hallucination. • Text and image convey meaning in distinct ways and with distinct limitations. More research is needed to examine questions of efficacy and utility before image-to-text models such as PaLI can be used as communication aids, including for education. We mitigate these risks by: • Removing pornographic images and personally identifiable information from the training data. • Analyzing data to identify potential bias issues resulting from the training data. • Currently not releasing the model checkpoints, source code or dataset to the public.

Usage

APPLICATION	BENEFITS	KNOWN CAVEATS
The model is for research prototype and the current version is not available for the public.	The model achieves new state-of-the-art on various well established vision and language benchmarks.	N/A