STUDY NOTE ON PALIGEMMA

Linsen Li
Computer Science in Tulane
lli23@tulane.edu

ABSTRACT

This article is to log the implementation of Paligemma[1].

1 Introduction

PaliGemma[1] is a model that excels at interpreting and understanding both images and text. It's used for tasks like generating descriptions of images, answering questions based on visual content, and analyzing complex visuals like infographics and satellite images. The input to PaliGemma can be an image, text, or both, and the output might be a text description, an answer to a question, or information derived from the image. It's designed to handle a wide range of vision-language tasks efficiently, even though it is smaller in size compared to some other advanced models.

The architecture of PaliGemma-3B, as shown in Figure 1, is inspired by the PaLI-3 model and combines the SigLIP visual encoder with the Gemma 2B language model. When PaliGemma-3B processes input, images are first converted into ”soft tokens” by the SigLIP encoder. Simultaneously, any accompanying text, referred to as the ”prefix,” is tokenized by Gemma's tokenizer. These image tokens and text tokens are then combined and fed into the Gemma decoder, which uses full block-attention to generate the final output text, or ”suffix,” in an auto-regressive manner.

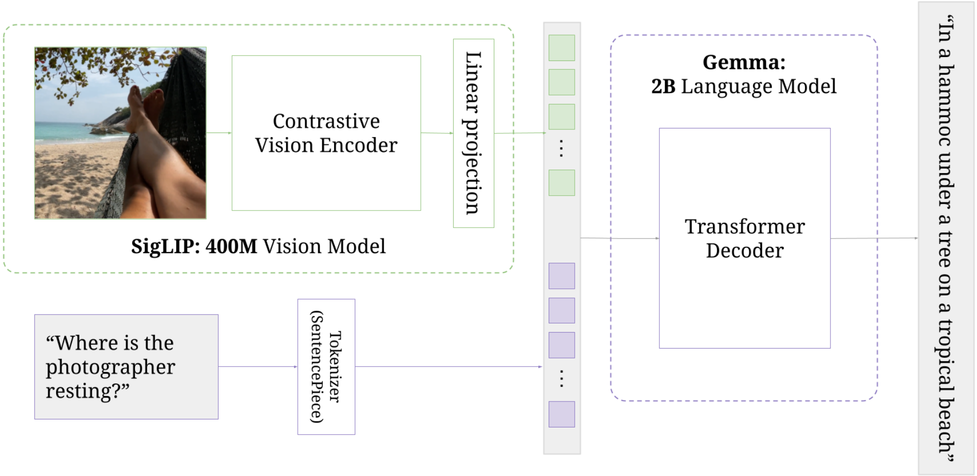
Contrastive Vision Encoder
Linear projection
SigLIP: 400M Vision Model
Gemma: 2B Language Model
Transformer Decoder
"In a hammoc under a tree on a tropical beach"
"Where is the photographer resting?"
Tokenizer (SentencePiece)

Figure 1: Architecture f Paligemma.

2 Related Work

Describe background. Introduce and cite what other people have done for this topic. Discuss the limitations of current approaches.

3 SigLIP

Paligemma uses the SigLIP model[2] as their Contrasive Vision Encoder.

4 Experimental Setup

Describe how you setup the experiments and the questions you try to answer using the experiments.

4.1 Data

Describe datasets you use including where you get the data, data statistics, etc. We encourage you to use public datasets.

4.2 Evaluation Metrics

Introduce evaluation metrics you use; If necessary, use equations.

4.3 Comparison Methods

List other methods you compare with (with citations) and the reasons you choose them.

5 Results

Analyze the results you get. Use tables or figures to show the results.

References

[1] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[2] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.