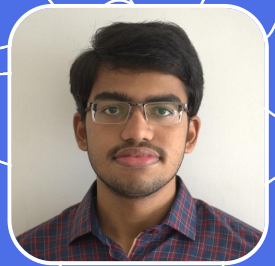

MULTIMODAL AI
SESSION #3
12th July 2024

About Us

Henry Vo
@_lowkeyboi



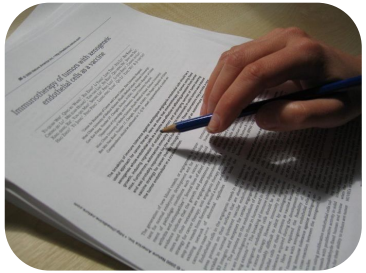
Surya Guthikonda
@suryaguthikonda

What to expect?

#multimodal-ml - Channel
@multimodal - Role

July Edition is Live

NEWSLETTER



Monthly Newsletter
- Industry
- Research
- Learning Resources

First Monday

Paper Reading Session

Guest Speaker Session

Coding Session

Second and Fourth Friday 10 AM ET

And More...

Fine-Tuning Florence-2 for DocVQA

Cohere For AI

Florence-2 101
A Quick Overview

Cohere For AI

Capabilities

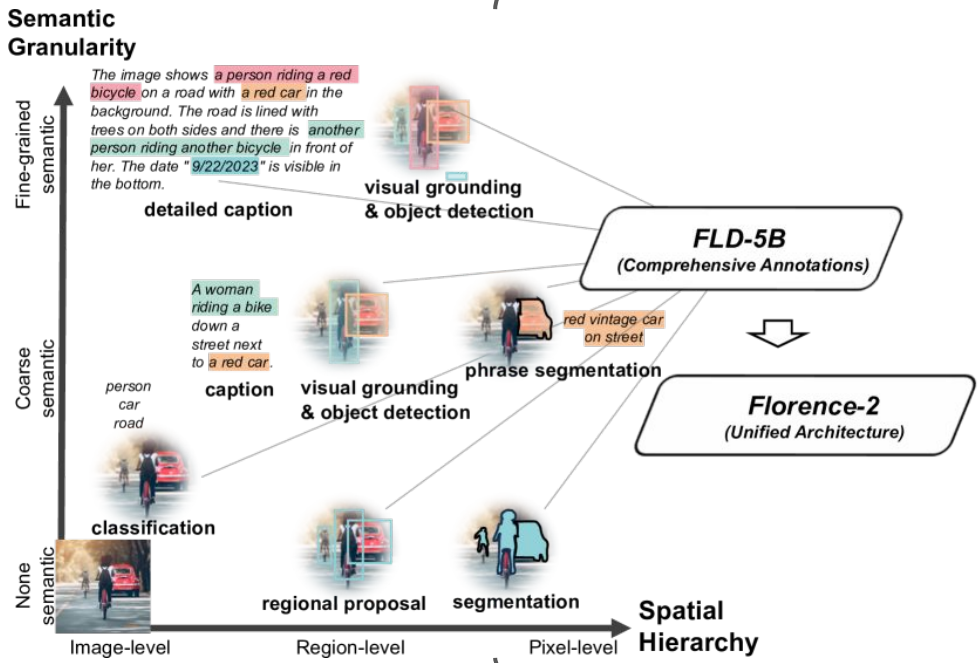
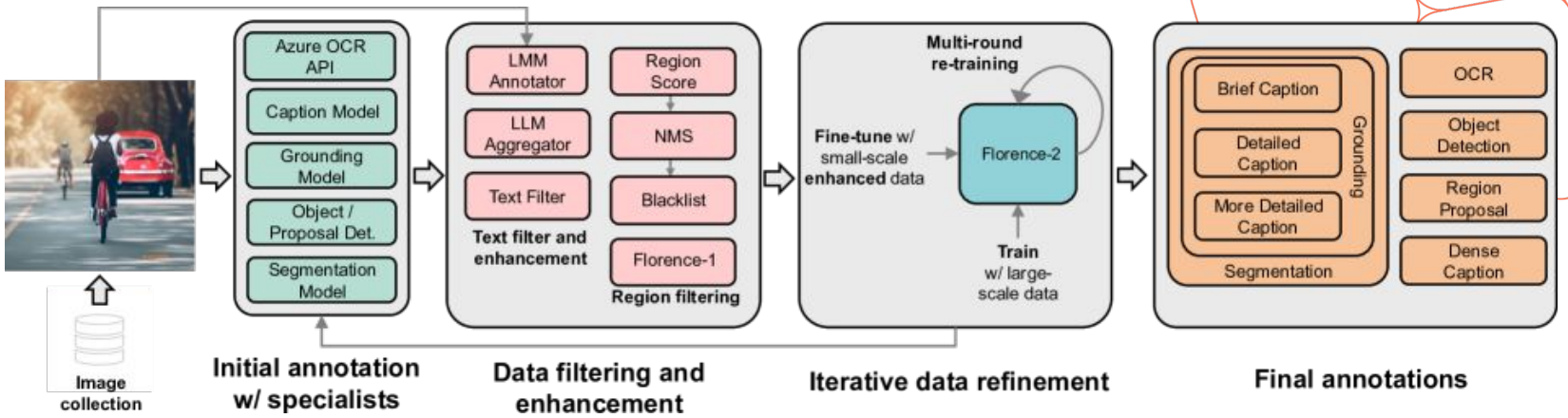


Image Level Understanding

Region/Pixel Level Recognition

Fine-grained Visual Semantic Alignment

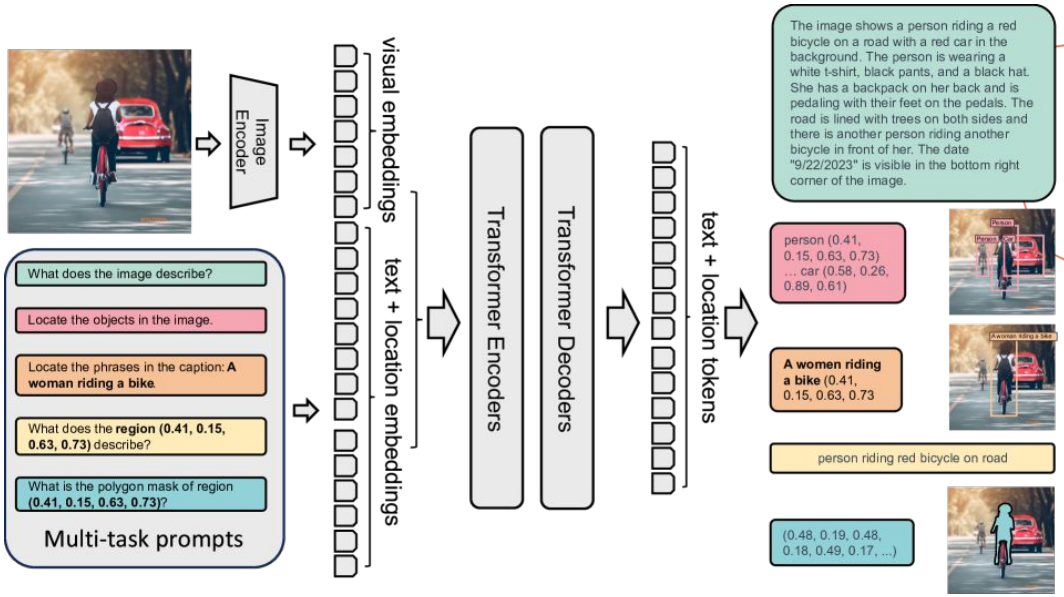
Data Engine



126M Images | **500M** Text Annotations | **1.3B** Region-Text Annotations | **3.6B** Text-Phrase-Region Annotations

Dataset Sneak-Peek

Model Architecture



Model	Image Encoder (DaViT)				Encoder-Decoder (Transformer)			
	dimensions	blocks	heads/groups	#params	encoder layers	decoder layers	dimensions	#params
Florence-2-B	[128, 256, 512, 1024]	[1, 1, 9, 1]	[4, 8, 16, 32]	90M	6	6	768	140M
Florence-2-L	[256, 512, 1024, 2048]	[1, 1, 9, 1]	[8, 16, 32, 64]	360M	12	12	1024	410M

Supported Tasks

Task	Annotation Type	Prompt Input	Output
Caption	Text	Image, text	Text
Detailed caption	Text	Image, text	Text
More detailed caption	Text	Image, text	Text
Region proposal	Region	Image, text	Region
Object detection	Region-Text	Image, text	Text, region
Dense region caption	Region-Text	Image, text	Text, region
Phrase grounding	Text-Phrase-Region	Image, text	Text, region
Referring expression comprehension	Region-Text	Image, text	Text, region
Open vocabulary detection	Region-Text	Image, text	Text, region
Referring segmentation	Region-Text	Image, text	Text, region
Region to text	Region-Text	Image, text, region	Text
Text detection and recognition	Region-Text	Image, text	Text, region

12 Tasks

Available Models

Model	Model size	Model Description
Florence-2-base[HF]	0.23B	Pretrained model with FLD-5B
Florence-2-large[HF]	0.77B	Pretrained model with FLD-5B
Florence-2-base-ft[HF]	0.23B	Finetuned model on a colletion of downstream tasks
Florence-2-large-ft[HF]	0.77B	Finetuned model on a colletion of downstream tasks

Florence-2 In Action

Inference : Notebook

Fine Tuning: Notebook

Resources

- Github Repository from Microsoft
- HuggingFace Blog
- Fine-tuned Model on DocVQA
- Demo


THANK YOU :)