SmolVLM for Dense Video Captioning

Week 2: Internship Update

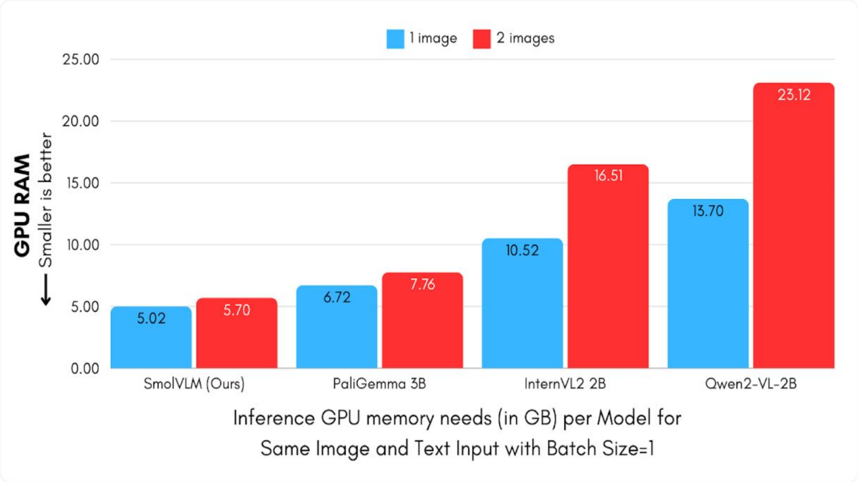
- Paper template
- Historical overview of VLMs and small VLMs
- Attention: cosine similarity = cross correlation = dot produc
- Apollo: state of the art o 7B (December 2024)
    - Video sampling: frames per second better than uniform sampling

Hugging Face (November 2024)

- Models (2.25)
  - SmolVLM-Base
  - SmolVLM-Synthetic
  - SmolVLM-Instruct (Tested)
- Architecture
  - Llama 3.1 8B with SmolLM2 1.7B for the language model,
  - Compressed visual information by reducing it 9 times using pixel shuffle strategy
  - Patches 384x384 because it is divisible by 3
  - Vision encoder is SigLIP with patches 384x384 and inner patches of 14x14

Hugging Face (November 2024)

Model	MMMU (val)	MathVista (testmini)	MMStar (val)	DocVQA (test)	TextVQA (val)	Min GPU RAM required (GB)
SmolVLM	38.8	44.6	42.1	81.6	72.7	5.02
Qwen2-VL 2B	41.1	47.8	47.5	90.1	79.7	13.70
InternVL2 2B	34.3	46.3	49.8	86.9	73.4	10.52
PaliGemma 3B 448px	34.9	28.7	48.3	32.2	56.0	6.72
moondream2	32.4	24.3	40.3	70.5	65.2	3.87
MiniCPM-V-2	38.2	39.8	39.1	71.9	74.1	7.88
MM1.5 1B	35.8	37.2	0.0	81.0	72.5	NaN



Hugging Face (February 2025)

- Video understanding
- Demos
- Models
  - SmolVLM 2.2B Instruc (Tested)
  - SmolVLM 500M Instruct
  - SmolVLM 256M Instruct (Still the smallest)

PySceneDetect

- ContentDetector: HSL/HSV + filtering (filtering is not the best for edge extraction, required threshold changing)
- AdaptiveDetector: HSL + rolling average
- ThresholdDetector: RGB averaging (device dependent)
- HistogramDetector: YUV (histogram of Y component) + correlation
- HashDetector: hash function, difference
- Testing code
- Thresholding algorithm?