SmolVLM for Dense Video Captioning

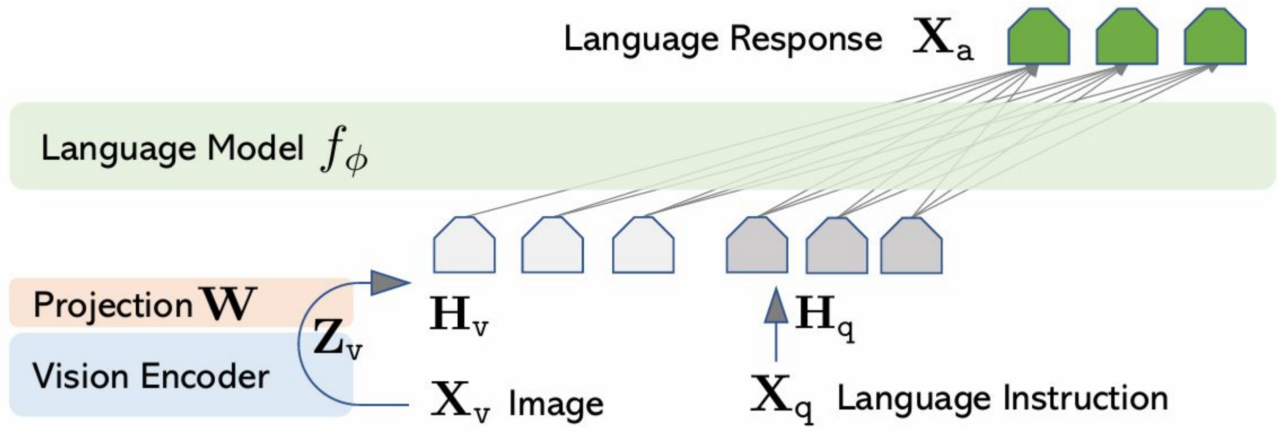
Week 1: Internship Uprade

General Understanding of Visual Language Models (Older and updated)

- Multimodal models
- Usecases
- Benchmarks: MMMU, MMBench
- Leaderboards: Vision Arena does not work
- Most used models
  - Any-to-any models
  - Reasoning models
  - Small
  - Mixture-of-Experts
  - Vision-Language-Action models

LLaVA: Large Language and Vision Assistant

- Textual prompts creating
- Textual encoder, vision encoder, projection matric (LLaVA, LLaVA 1.5 with MLP)



CLIP: Contrastive Language-Image Pre-Training

- Image encoding:
  - ResNET
  - Visual transformer
- Text encoding:
  - Transformer based

Transformers: Attention is all you need

- Attention
- Architecture

SmolVLM

- Efficient Visual Tokenizer (SigLIP)
- Lightweight Language Backbone
- Multimodal Connector
- Pixel Shuffle
- Image splitting, video frame averaging

Temporal Dimension Processing in Videos

- Pixel level difference (L1/L2 difference)
- Feature extraction and embeddings comparison (cosine similarity)
- Frame Voyager: model, based on text-frame matching, ranking score of each frame
  - If we use all of the frames: ‘lost-in-the-middle’, hallucinations
- Scene Boundary Detection: based on LSTMs
- Reinforcement Learning: mask of importance of frames, transformer, trained separately
- Motion Based Filtering: optical flow
- Semi optimal policy: selecting N optimal frames from T instead of T^N space

Temporal Dimension Processing in Videos

- Cross-correlation

$$(F_i \star F_j)[m,n] = \sum_h \sum_w F_i[h,w] \cdot F_j[h+m, w+n]$$

$$\mathrm{Corr}_{ij}[m,n] = \frac{\sum_h \sum_w F_i[h,w] \cdot F_j[h+m, w+n]}{\|F_i\|_F \cdot \|F_j\|_F}$$

$$\|F_i\|_F = \sqrt{\sum_h \sum_w F_i[h,w]^2}$$

$$\mathrm{sim}(F_i, F_j) = \max_{m,n} \left(\mathrm{Corr}_{ij}[m,n]\right)$$

$$\mathrm{diff}(F_i, F_j) = 1 - \max_{m,n} \left(\mathrm{Corr}_{ij}[m,n]\right)$$

$$\tilde{F}_1 = \frac{F_1 - \mu_{F_1}}{\sigma_{F_1} + \varepsilon}, \quad \tilde{F}_2 = \frac{F_2 - \mu_{F_2}}{\sigma_{F_2} + \varepsilon}$$

$$\mathrm{Corr}(F_1, F_2)[m,n] = (\tilde{F}_1 \star \tilde{F}_2)[m,n]$$

$$\mathrm{diff}(F_1, F_2) = \frac{\max_{m,n} \left(\mathrm{Corr}(F_1, F_2)[m,n]\right)}{HW}$$

- 398.04s (128x128)
- CLIP(openai/clip-vit-large-patch14, 512, whole frames)
  - 18.12s