MIRASOL 3B

Multimodal Auto Regressive Model.

Main challenges = Combining Multiple Heterogeneous Modalities
Video Audio and Text

[Video / Audio] ⇒ higher rates than Text
are roughly aligned in time.

(Video / Audio)
They are not synchronized with Text which comes as a global context.
eg: Title / Description.

[Video / Audio Inputs are much larger volumes] as Video length increases
⇓
Needs more compute time for these Modalities.

Mirasol 3B ⇒ ① Autoregressive Component
Time synchronized Modalities
(Audio / Video)

⇓

② Autoregressive Component
[Context modalities which are not necessarily aligned with time but Sequential] non time aligned

To address the long sequence of video / Audio inputs ⇒ partition of video / Audio Sequences in Consecutive Snippets and autoregressively process their representations.

||Combiner Mechanism ... audio / video

|| [illegible] ⟹ [illegible]

information jointly producing compact but expressive representations.

↓

This makes the Model to take 512 frame input without increase in the Model Parameter.

[Note: Architecture for Video Language understanding commonly use a joint transformer [video input + Text tokens] are processed auto regressively]

Inputs:- Input video sequence of N frames

$$V = (v_1^f, v_2^f, \dots v_N^f)$$

Audio waves signal of M timesteps

$$a = (a_1^f, a_2^f, \dots a_M^f)$$

$$t = \{t_1^f, t_2^f, \dots t_P^f\}$$ Text Sequence

Partitioning the video & Audio input:

$$\{\underbrace{v_1^f, v_2^f, \dots v_k^f}_{V_1}\}, \{\underbrace{v_{k+1}^f, v_{k+2}^f, \dots v_{2k}^f}_{V_2}\}, \dots$$

chunks

$$(\underbrace{v^f_{(T-1)k+1}, \dots v_N^f}_{V^T})$$

$\boxed{k = N/T}$

T = No. of Non overlapping Chunks.

[Each & Every chunk] ⟹ Latent features

[Video – Spatiodata representation.
extract sparse 3D tubes. with
Standard 2D Patches.] are processed
using the ViT Encoder

Audio — Represented as Spectrograms

Combiner Module:-

1) Combine video / Audio features at specific snippet at time [joint representatio

2) effe compress the representation from each [audio / Video] snippet, allows the model to scale to longer videos.

$\hat{v} = \{\hat{v_1}, \hat{v_2} \dots \hat{v_N}\}$ => Video

$\hat{a} : \{\hat{a_1}, \hat{a_2}, \dots \hat{a_N}\}$ => audio

Composed of f features of size d (f, d) shape

Composed of s features sized shape {s, d}

Combiner ① $u = \{u_1, u_2, v_3, \dots u_f\}$

Where $\boxed{u_t : \{\hat{v}_t, \hat{a}_t\}}$ and

size $\{n, d\}$ where $n = f + s$

② Then maps u to lower dimensional features per $x = \{x_1, x_2, \dots x_T\}$

Where x_t has the shape $\{m, d\}$

Where $n >> m$

Combines two different Architectures

① Standard Transformer the

② Token Turing Machine.

Time Aligned Video/Audio Autoregressive Mod

Condition (Audio/Video representation from previous time intervals)

x_t ⇒ passed sequentially to auto regressive Model.

$$p(v, a) = \prod_{i=1}^{T} \underbrace{\left[p(v_{t+1}, a_{t+1} \mid h_t) \right]}_{\text{Modality reconstruction Model.}}$$

$$\underbrace{\left[p(h_t \mid x_t) \right]}_{\text{Estimated by latent Causal Model.}} \underbrace{\left[p(x_t \mid v_t, a_t) \right]}_{\text{Combiner.}}$$