CAVMA reading notes

1. Preparation

(1) audio and image pre-processing and tokenization

- dataset: 10s videos with parallel audios in AudioSet and VGGSound
- audio: AST

  10s audio wave ⟶ 1024 (time) × 128 (frequency) spectrogram
  ⟶ 512 16×16 square patches $\vec{a} = [a^1, \cdots, a^{512}]$

- video: ViT

  frame aggregation strategy: save computation resources

sample 10 RGB frames (1 FPS) —
- training → randomly select 1 RGB frame as input
- inference → average each RGB frame prediction as the video prediction

each RGB frame $\xrightarrow{\text{resize + center crop}}$ 224 × 224
⟶ 196 16×16 square patches $\vec{v} = [v_1, \cdots, v^{196}]$

(2) transformer architecture

standard Transformer

a Transformer layer:

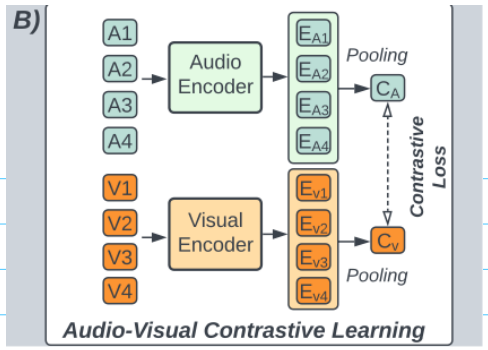
$$\left.\begin{aligned} x' &= MSA(LN_1(x)) + x \\ y &= MLP(LN_2(x')) + x' \end{aligned}\right\} \quad y = Transformer(x; MSA, LN1, LN2, MLP)$$

MSA: multi-headed self attention
LN: layer normalization
MLP: multilayer perceptron

(3) constrastive audio-visual learning (CAV)

audio ⟶ $\vec{a}_i$ ⟶ $c_i^a = MeanPool(E_a(Proj_a(a_i)))$

video ⟶ $\vec{v}_i$ ⟶ $c_i^v = MeanPool(E_v(Proj_v(v_i)))$



$Proj_a : \vec{a}_i \longrightarrow R^{768}$

$Proj_v : \vec{v}_i \longrightarrow R^{768}$

contrastive loss:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)}\right]$$

$s_{i,j} = \|c_i^v\|^T \|c_j^a\|$

τ is the temperature.

(4) single modality masked autoencoder (MAE)

$\vec{x} = [x_1, x_2, \cdots, x^n]$

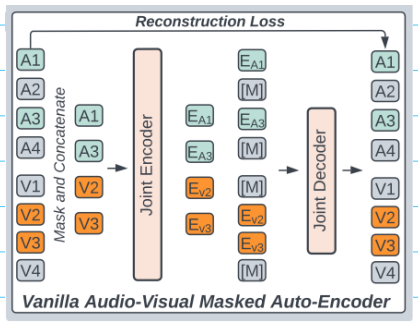
$\vec{x} / x_{mask} \xrightarrow{\text{encoder-decoder model}} \hat{x}_{mask} \longrightarrow MSE(x_{mask}, \hat{x}_{mask})$

(5) vanilla audio-visual masked autoencoder (AV-MAE)

$a' = Proj_a(a) + E_a + E_a^p$

modality type embedding — 2-D sinusoid positional embedding

$v' = Proj_v(v) + E_v + E_v^p$



$x = [a', v']$

mask a portion (75%) of x

Encoder

$x_{unmask} = x \setminus x_{mask}$

↓

audio-visual joint encoder $E_j(\cdot)$

↓

x'_{unmask} — pad with masked token → x'

Decoder

$x' + [E_a', E_v'] + [E_a^{p'}, E_v^{p'}]$

↓

joint audio-visual decoder $D_j(\cdot)$

↓

$\hat{a}, \hat{v} = D_j(x' + [E_a', E_v'] + [E_a^{p'}, E_v^{p'}])$

minimize MSE $\hat{a}, \hat{v}$ and normalized a, v

2. Constrastive audio-visual masked auto-encoder (CAV-MAE)
combine CAV and AV-MAE

mini-batch of N audio-visual pair samples

↓ preprocess

$\{a_i, v_i\}$, $i = 1, \cdots, N$

↓

$a_i^{unmask} = Mask_{0.75}(Proj_a(a_i) + E_a + E_a^p)$

$v_i^{unmask} = Mask_{0.75}(Proj(v_i) + E_v + E_v^p)$

a_i^{unmask} ↓ E_a → a_i'

v_i^{unmask} ↓ E_v → v_i'

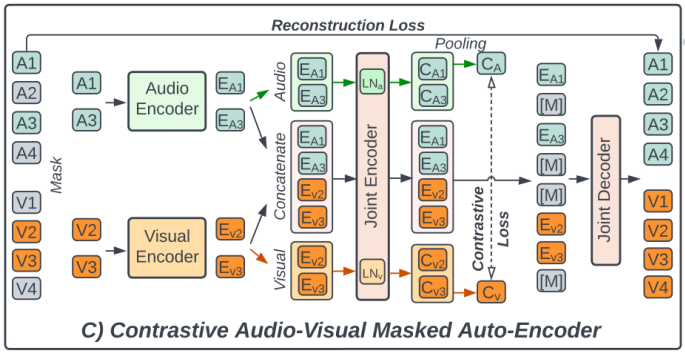
multi-stream forward pass:

$c_i^a = Meanpool(E_j(E_a(a_i^{unmask})); LN1_a, LN2_a))$ → for CL

$c_i^v = Meanpool(E_j(E_v(v_i^{unmask})); LN1_v, LN2_v))$ → for CL

$x_i = E_j([E_a(a_i^{unmask}), E_v(v_i^{unmask})]; LN1_{av}, LN2_{av})$ → for RC

* shared weights for E_j
* different LN layers $LN\{a, v, av\}$



$$\hat{a}, \hat{v} = D_j(x' + [E_a', E_v'] + [E_a^{p'}, E_v^{p'}])$$

Loss:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)}\right]$$

$$L_r = \frac{1}{N}\sum_{i=1}^{n}\left[\frac{\sum(\hat{a}_i^{mask} - norm(a_i^{mask}))^2}{|a_i^{mask}|} + \frac{\sum(\hat{v}_i^{mask} - norm(v_i^{mask}))^2}{|v_i^{mask}|}\right]$$

number of masked patches

$$L_{CAV\text{-}MAE} = L_r + \lambda_c L_c$$

* only keep the encoders for downstream tasks.
  1. single-modality stream output + multi-modal stream output
  2. multi-modal stream output
  1 and 2 perform similarly