VATT — For Project

Monday, October 28, 2024 7:55 PM

Audio → Positional Encoding → Audio Embeddings

$(N \times L_a)$ → $(N \times T_a \times d)$

- Sample @ 48kHz
- 128 - sample len patch split
- map each patch to $(d \times 1)$ vector

Text → Positional Encoding → Text Embeddings

$(N \times L_T)$ → $(N \times T_t \times d)$

- Have a max_seq_len
- Tokenize text as one - hot vec
- map each one-hot vec to $(d \times 1)$ vector

Transformer Encoder $(N \times T \times d)$

1) Multiheaded Self Attention
   ↳ K, Q, V $(1 \times d/h)$
2) Layer Norm + Add
3) Feed forward MLP layer
   ↳ 2-layers (ReLU / GELU)
   ↳ $d \to 4d \to d$
4) Layer Norm + Add

2 versions of VATT

1) Modality - specific $(T_a \neq T_t)$
   ↳ one encoder per modality
2) Modality -agnostic $(T_a = T_t)$
   ↳ one encoder is shared

Multimodal Projection Head $(N \times d) \longrightarrow (N \times d_{proj})$

1) Extract the [CLS] of each modality or average pooling
2) Map them with a linear projection to 1D space.
3) This map can be of different types
   ↳ Hierarchical — "Dog bark" — Multiple layers to classify
   ↳ Linear — "Dog" "bark" — One layer to classify

Contrastive Learning

1) NCE (Noise Contrastive Estimation) — Audio + Transcript
2) MIL - NCE (Multiple Instance Learning NCE) — (1) + Labels

Drop Token

↳ Randomly drop some patches and corresponding text patches.
↳ Helps with computational load.
↳ Especially useful for our long sequences

Longer Data Foresight

1) $L_a \uparrow$ $L_t \uparrow$
2) $T_a \uparrow$ $T_t \uparrow$
3) Can experiment with { changes in sample rate vs more aggressive drop token vs changes in model dimensions }
4) Num - layers ↑