MA-AVT: Modality Alignment for Parameter-Efficient Audio-Visual Transformers

Supplementary Material

A. Overview

In this supplementary material, we provide additional details and experimental results for our proposed MA-AVT. In summary, the following items are presented.

- Details of the mathematical notations used in the main paper.
- Additional implementation details used in model training.
- Additional quantitative results for experimental study.
- Qualitative analysis of the model performance.

B. List of Notations

We provide a detailed list of all symbols, data types, and corresponding feature dimensions in Table 6.

C. Implementation Details

We extract the visual frames with 1 fps in all experiments. We use the same spatial resolution of (224×224) for all images. We use the audio segment length of 1s for AVE dataset, 5s for VGGSound dataset, and 3s for Crema-D dataset. In particular, we use the whole 10s event duration in AVE dataset with 10 audio-visual pairs representing 1s duration each following prior work [15]. The VGGSound, and Crema-D dataset are used with single audio-visual pair from each video as reported in prior work [23]. For audio processing, we use librosa and torchaudio packages with PyTorch. We set the same learning rate of 0.01 across all trainable parameters with a batch size of 256. Local self-attention (LSA) processed learnable tokens are merged with patch tokens after applying position embeddings to patch tokens. We use multi-headed attention operations in LSA modules with 8 attention heads. In blockwise contrastive loss, we use separate projectors after each block to reduce representation variance between audio and visual modalities. We use simple MLP layers with 256 output nodes as projector modules. We note that contrastive loss is only used for training and all of these projector units are discarded in the test phase. Therefore, we ignored the parameter count of these projector units in the reported trainable parameter counts. Only the additional trainable parameters that are used in the testing phase are considered, following prior work [15]. We use temperature 0.07 for calculating contrastive losses in all cases during training.

D. Additional Experimental Study

We present some additional experiments to compare and analyze the performance of MA-AVT.

D.1. Comparison with different ViT backbones

We compare the performance of MA-AVT with different ViT backbones as reported in Table 7. We use LAVISH [15] as the baseline method for parameter-efficient audio-visual transformers. We only report the accuracy on AVE dataset for this analysis. We use ImageNet pretrained frozen transformer encoders for both LAVISH and MA-AVT. In general, MA-AVT achieves significant accuracy improvements over LAVISH with comparable parameter efficiency. To be specific, MBT uses comparatively less trainable parameters than LAVISH with ViT-Tiny and ViT-Large encoders, while achieving $+3.6$, and $+1.5$ higher accuracy, respectively. With ViT-Small, MA-AVT uses marginally higher trainable parameters than LAVISH (**1.81M** vs. **1.65M**) that achieves $+2.7$ higher accuracy. In case of ViT-Base encoders, MA-AVT incorporates slightly more trainable parameters than LAVISH (**7.1M** vs. **4.7M**) for achieving $+2.6$ higher accuracy. For fair comparison, we incorporate more trainable parameters into LAVISH by using deeper convolution in intermediate adapter modules to increase model capacity ($\sim$ **7.3M**). Nevertheless, MA-AVT maintains superior performance over the larger variant of LAVISH with ViT-Base encoders.

D.2. Effects of number of learnable tokens

We analyze the effect of different number of learnable tokens. As a rule of thumb, we maintain the same number of tokens in all three groups, *i.e.* audio, visual, and shared tokens. The results are given in Table 8. We report the accuracy on AVE dataset for this analysis. We use ViT-Base frozen encoders for both audio and visual modalities. The best performance is achieved with 5 tokens in all three groups. With increasing number of tokens, the complexity of the network increases for expensive cross-attention operations in transformer building blocks that results in lower accuracy gain. Hence, we use 5 learnable tokens in each group for all other experiments.

D.3. Comparison of throughput

In Table 9, we compare the throughput of MA-AVT with other competitive transformer based methods, such as MBT [20] and LAVISH [15]. We use the same A5000 GPU

Table 6. Different notations used to describe the operations of MA-AVT. We categorize the notations into three groups.

Group	Description	Notation	Datatype	Dimension
Data Parameters	complete dataset	$\mathcal{D}$	-	-
	total audio-visual pairs	N	scalar	$\mathbb{R}^0$
	image frame	v	vector	$\mathbb{R}^{3\times H\times W}$
	audio spectrogram	a	vector	$\mathbb{R}^{F\times T}$
	audio patch tokens	P_a^0	vector	$\mathbb{R}^{m\times d}$
	visual patch tokens	P_v^0	vector	$\mathbb{R}^{n\times d}$
	visual token embedding after k block	E_v^k	vector	$\mathbb{R}^{(m+n_v+n_s+2)\times d}$
	audio token embedding after k block	E_a^k	vector	$\mathbb{R}^{(n+n_a+n_s+2)\times d}$
Model Parameters	learnable audio prompts	z_a	vector	$\mathbb{R}^{n_a\times d}$
	learnable visual prompts	z_v	vector	$\mathbb{R}^{n_v\times d}$
	learnable shared multimodal prompts	z_s	vector	$\mathbb{R}^{n_s\times d}$
	learnable foreground class prompt	z_f	vector	$\mathbb{R}^{1\times d}$
	learnable background class prompt	z_b	vector	$\mathbb{R}^{1\times d}$
	audio prompt LSA unit	$A_a(\cdot)$	operator	-
	visual prompt LSA unit	$A_v(\cdot)$	operator	-
	shared prompt LSA unit	$A_s(\cdot)$	operator	-
Loss	foreground-background loss	$\mathcal{L}_{bf}$	scalar	$\mathbb{R}^0$
	contrastive loss at k^{th} block	$\mathcal{L}_{cnt}^k$	scalar	$\mathbb{R}^0$
	audio-to-visual contrastive loss	$\mathcal{L}_{v\rightarrow a}$	scalar	$\mathbb{R}^0$
	visual-to-audio contrastive loss	$\mathcal{L}_{a\rightarrow v}$	scalar	$\mathbb{R}^0$
	total loss	$\mathcal{L}$	scalar	$\mathbb{R}^0$
	background label	y_b	vector	$\mathbb{R}^{1\times 1}$
	foreground label	y_f	vector	$\mathbb{R}^{1\times C}$
	total number of classes	C	scalar	$\mathbb{R}^0$

Table 7. **Comparison with different ViT encoders.** We use ImageNet pretrained frozen (F) ViT encoders in all cases. * denotes our improved implementation. We only report the accuracy on AVE dataset. In all ViT backbones, MA-AVT achieves significant accuracy improvements over LAVISH with comparable parameter-efficiency.

Models	**Image Encoder**	**Audio Encoder**	**Total Params (M)**	**Trainable Params (M)**	**Accuracy(%)**
LAVISH [15]	ViT-Tiny (F)	ViT-Tiny (F)	10.39	0.65	61.8
MA-AVT (ours)	ViT-Tiny (F)	ViT-Tiny (F)	10.20	0.46	**65.6**
LAVISH [15]	ViT-Small (F)	ViT-Small (F)	31.72	1.65	71.4
MA-AVT (ours)	ViT-Small (F)	ViT-Small (F)	31.88	1.81	**74.1**
LAVISH [15]	ViT-Base (F)	ViT-Base (F)	107.2	4.7	75.3
LAVISH* [15]	ViT-Base (F)	ViT-Base (F)	110.4	7.3	75.8
MA-AVT (ours)	ViT-Base (F)	ViT-Base (F)	110.2	7.1	**77.9**
LAVISH [15]	ViT-Large (F)	ViT-Large (F)	340.1	14.5	78.1
MA-AVT (ours)	ViT-Large (F)	ViT-Large (F)	338.4	12.6	**79.6**

for measuring the throughput. We note that MA-AVT gains 2x and 1.5x speedup over LAVISH for ViT-Large and ViT-Base encoders, respectively. In LAVISH, residual adapters are used in intermediate blocks for cross-modal fusion and modality alignment. However, this residual operation reduces the throughput in practice. In contrast, the learnable tokens in MA-AVT directly operate on cross-attention layers of frozen transformer rather than relying on residual adapters that results in almost 2x speedup. However, the integrated local self-attention (LSA) modules increase the computation burden compared to fully-tuned MBT which operates without such additional blocks. Moreover, additional tokens in MA-AVT increase the computational cost in each cross-attention layer. As a result, MBT is faster than MA-AVT for both ViT-Base and ViT-Large encoders. Nevertheless, MBT is parameter hungry since it relies on fully-tuning and pre-training of the large transformer encoders. Therefore, MA-AVT provides considerable speedup

Table 8. **Effects of number of tokens.** We use the same number of tokens in all three groups. Only accuracy in AVE dataset is reported. The best accuracy is achieved with 5 learnable tokens.

Number of Tokens	Accuracy(%)
2	76.3
5	**77.9**
10	77.5
15	77.1

Table 9. **Throughput comparison.** We use ViT-Base and ViT-Large to compare throughput. We use the same A5000 GPU to measure the throughput. MA-AVT is nearly 2x faster than LAVISH. However, MBT is the fastest among three.

Model	**Image Encoder**	**Audio Encoder**	**Samples/Sec.**
MA-AVT (ours)	ViT-Base	ViT-Base	250
LAVISH [2023]	ViT-Base	ViT-Base	125
MBT [2021]	ViT-Base	ViT-Base	**400**
MA-AVT (ours)	ViT-Large	ViT-Large	85
LAVISH [2023]	ViT-Large	ViT-Large	55
MBT [2021]	ViT-Large	ViT-Large	**218**

with higher accuracy while maintaining parameter efficiency comparable to LAVISH (Table 7).