Less is Enough: Adapting Pre-trained Vision Transformers for Audio-Visual Speaker Verification

R. Gnana Praveen, Jahangir Alam
Computer Research Institute of Montreal
Montreal, Quebec, H3N 1M3, Canada
{gnana-praveen.rajasekhar,jahangir.alam}@crim.ca

Abstract

Speaker Verification has achieved significant progress using advanced deep learning architectures, specialized for speech signals as well as robust loss functions. Recently, fusion of faces and voices received a lot of attention as they offer complementary relationship with each other, outperforming unimodal approaches. In this work, we have investigated the potential of Vision Transformers (ViTs), pre-trained on visual data, for audio-visual speaker verification. To cope with the challenges of large-scale training, we introduce the Latent Audio-Visual Vision Transformer (LAVViT) adapters, where we exploit the existing pre-trained models on visual data by training only the parameters of LAVViT adapters, without fine-tuning the original parameters of the pre-trained models. The LAVViT adapters are injected into every layer of the ViT architecture to effectively fuse the audio and visual modalities using a small set of latent tokens, thereby mitigating the quadratic computational cost of cross-attention across the modalities. The proposed approach further circumvents the need for modality-specific architectures by employing the same ViT architecture with shared pretrained weights for audio and visual modalities. The proposed approach has been evaluated by conducting extensive experiments on the Voxceleb1 dataset and shows promising performance using only a few trainable parameters.

1 Introduction

Speaker verification (SV) has been predominantly explored using speech signals, as the identity of the speaker is often reflected in their voices. Voice-based biometrics is one of the key technologies, widely used to verify the authenticity of a person in several applications such as customer authentication, security applications, etc (1). SV can be classified as text-dependent and text-independent based on whether the constraint is imposed on the lexical content of the utterances or not (2). In this work, we have focused on text-independent SV, where there is no constraint on the text content of speech utterances. Though the performance of SV has been drastically improved over the last few years using advanced architectures such as ECAPA-TDNN (3), xi-vector (4), and transformers (5), relying only on speech signals suffer from poor performance when the speech utterances are corrupted due to external noise, noisy environments, etc. Another widely-explored paradigm to verify the authenticity of the person is based on faces in the vision community. Face verification has also achieved impressive performance with the advancement of deep learning architectures (6) and loss functions (7). Although faces and voices have shown remarkable success individually, they fail to retain their robust performance when the modalities get corrupted. Therefore, audio-visual fusion based on face and voice is gaining a lot of attention as they often complement each other, leading to better performance than that of individual modalities.

4th NeurIPS Efficient Natural Language and Speech Processing Workshop (ENLSP-IV 2024).

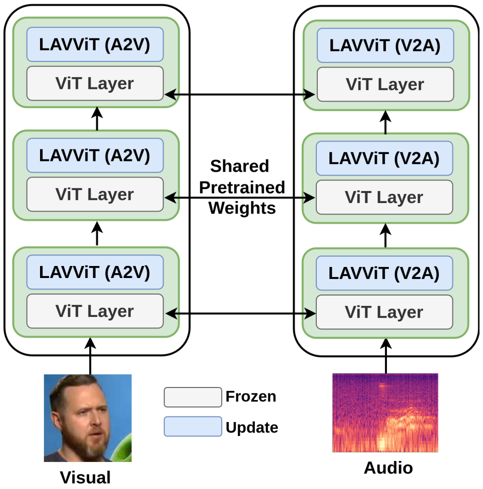

Figure 1: Illustration of the Proposed Approach. LAVViT adapters are injected into every layer of ViT and the pre-trained weights are shared across the modalities, where only the parameters of LAVViT adapters are trained and pre-trained weights are frozen.

Audio-visual fusion has been successfully explored by capturing the complementary relationships using cross-modal attention in several applications such as emotion recognition (8; 9; 10; 11), event localization (12), and action localization (13). Conventionally, audio-visual fusion for SV has been explored using a simple fusion of scores obtained from facial and speaker embeddings (14; 15) or early feature-level fusion (16; 17). Although score-level fusion has achieved better results, the full potential of audio-visual fusion for SV is not fully explored, leaving a lot of room for further improvement. Recently, attention mechanisms have been exploited for effective fusion of audio and visual modalities for SV (18; 19; 20). With the massive success of transformer models (21) in Natural Language Processing (NLP), it is also explored successfully in other domains such as computer vision (22) and speech processing (23). Recently, they have also been explored for SV using speech signals (5). However, leveraging transformer models in the context of audio-visual learning for SV is still at a rudimentary level.

Owing to the remarkable success of Vision Transformers (ViT) (24) for computer vision applications, they have also been found to be promising in achieving better performance for multimodal learning (25). However, one of the challenging aspects of exploiting ViTs for audio-visual learning is the requirement of intense computational power for large-scale training. Moreover, different architectures are often used for extracting the embeddings of the individual modalities, which further increases the burden of additional computational support. Recently, Lin et al (26) investigated the prospect of leveraging ViTs for parameter-efficient audio-visual learning and showed promising performance for applications such as event localization and semantic segmentation. In this work, we have explored the potential of ViTs to effectively extract the audio and visual embeddings, while deploying cross-modal adaptation for audio-visual SV. By representing the speech signals as spectrograms, we have used similar ViT architecture by sharing the weights pre-trained on image data as shown in Fig 1. Moreover, the computational overhead of large-scale training has been drastically reduced by introducing latent tokens in the LAVViT adapters in every layer of ViT architecture. The latent tokens help to compress the large number of audio or visual tokens, thereby reducing the computational overhead of cross attention across the modalities. The major contributions of this work can be summarized as follows.

- To the best of our knowledge, this is the first work to investigate the potential of ViTs in a parameter efficient manner for audio-visual SV.
- The quadratic computational complexity of large-scale training of ViTs has been reduced by introducing latent tokens at every layer of ViT using LAVViT adapters.
- The proposed approach relies on the same ViT architecture with shared pre-trained weights, thereby circumventing the need for modality specific-architectures for audio and visual backbones.

- Extensive experiments are conducted on the voxceleb1 dataset and showed that the proposed model achieves better performance using a few trainable parameters.

2 Related Work

Inspired by the close associations between faces and voices, several approaches have been proposed for cross-modal SV using joint feature embeddings in a shared representation space (27; 28). Traditionally, audio-visual SV has been explored using early feature-level fusion (16; 17) or score-level fusion (14; 15). Sari et al. (29) proposed a multi-view approach to map the audio and visual embeddings to a common space using a shared classifier for cross-modal SV. Shon et al. (18) explored attention models to handle the problem of noisy modalities by emphasizing on the salient modality. Hormonn et al. (19) proposed a multiscale feature-fusion approach to obtain robust feature representations by fusing the features at intermediate layers. Chen et al. (17) explored various fusion strategies based on gating mechanisms and further analyzed the impact of the gating-based fusion under extremely corrupted or missing modalities. Tao et al (30) attempted to exploit the complementary relationships across the modalities to refine the noisy samples using a two-step deep cleansing framework. Most of these methods endeavored to leverage the complementary relationships across the modalities to handle the problem of noisy modalities.

Unlike prior approaches, Liu et al. (31) focused on leveraging the cross-correlation across the modalities using co-learning approach by exploiting the knowledge of one modality to attend to another modality. Sun et al. (32) introduced weight-enhanced attentive statistics pooling to focus on keyframes by deploying cycle consistency loss and gated attention. Praveen et al. (20) explored joint cross attention to effectively capture the intra- and inter-modal relationships simultaneously by introducing joint representation in the cross attention framework. They further improved the performance by incorporating recursive mechanism to refine the attended feature representations (33). Recently, Praveen et al (34) proposed dynamic cross-attention to address the problem of weak complementary relationships and showed further improvement over prior cross-modal attention methods. Even though cross-attention has been explored to capture the complementary relationships, they do not explore the potential of transformer models for audio-visual SV. Another recent approach by Rajasekhar et al. (35) showed that transformers can be explored for effectively capturing the cross-modal relationships, thereby improving the performance of audio-visual SV. Contrary to the prior approaches, we have explored the potential of ViTs, pre-trained on visual data for effectively fusing audio and visual modalities for SV.

3 Proposed Approach

3.1 Audio-Visual Inputs

Given an input video sequence, we extract the audio and visual streams. For the visual modality, we randomly sample an image of size $H \times W \times 3$ in the video sequence, where H and W denote the height and width of the sampled image respectively. For audio modality, we compute the spectrogram of the audio stream with size $M \times C$ where M and C denote the spatial dimensions of the spectrogram. Since we use a similar ViT architecture for both audio and visual inputs, the number of channels of the spectrogram is inflated from 1 to 3 to match the channels of visual input. As per the convention of ViT (24), the sampled image and the spectrogram of the visual and audio inputs are divided into N and K non-overlapping patches respectively. These patches are further flattened, and fed to a linear projection layer to obtain the audio and visual tokens, where $\boldsymbol{X}_a^{(0)} \in K \times d$ and $\boldsymbol{X}_v^{(0)} \in N \times d$ represent the audio and visual tokens respectively and d denote the dimension of the tokens.

3.2 LAVViT Adapters

This is the core module of the proposed approach, which helps to drastically reduce the computational overhead of ViTs using the latent tokens and few trainable parameters. ViTs with adapters have been recently attaining significant attention as they are found to be promising in achieving parameter-efficient training with few trainable parameters (36; 37). In this work, we explore the potential of LAVViT adapters for audio-visual SV. The proposed LAVViT adapter is composed of two modules: (i) Latent Self Attention (LSA) and (ii) Cross Modal attention (CMA) as shown in Fig 2. In the LSA

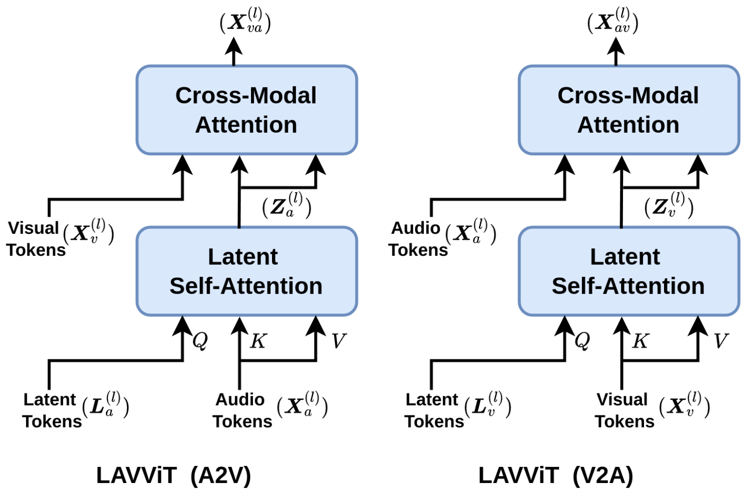
$(\boldsymbol{X}_{va}^{(l)})$
Cross-Modal Attention
$(\boldsymbol{Z}_a^{(l)})$
Visual Tokens $(\boldsymbol{X}_v^{(l)})$
Latent Self-Attention
Q
K
V
Latent Tokens $(\boldsymbol{L}_a^{(l)})$
Audio Tokens $(\boldsymbol{X}_a^{(l)})$
LAVViT (A2V)
$(\boldsymbol{X}_{av}^{(l)})$
Cross-Modal Attention
$(\boldsymbol{Z}_v^{(l)})$
Audio Tokens $(\boldsymbol{X}_a^{(l)})$
Latent Self-Attention
Q
K
V
Latent Tokens $(\boldsymbol{L}_v^{(l)})$
Visual Tokens $(\boldsymbol{X}_v^{(l)})$
LAVViT (V2A)

Figure 2: LAVViT adapter module of the proposed approach. Separate LAVViT adapters are used for audio and visual modalities.

module, we attempt to compress the audio or visual tokens to a small number of latent tokens in order to reduce the computational cost. These latent tokens help to efficiently represent the audio or visual tokens for effective cross-modal attention across the modalities. The latent tokens are randomly initialized and they are not shared across the modalities as well as across the ViT layers. By training only the parameters of LAVViT adapters in each layer of ViT, we are able to efficiently condense the information at every layer using a few trainable parameters, yet exploiting the pretrained frozen models. Let M denote the number of latent tokens, which is chosen as significantly smaller than that of the total number of audio or visual tokens (in our experiments, we use $M = 2$). The compressed tokens of the LSA for audio and visual modalities can be expressed as:

$$\boldsymbol{Z}_a^{(l)} = LSA(\boldsymbol{L}_a^{(l)}, \boldsymbol{X}_a^{(l)}) \tag{1}$$

$$\boldsymbol{Z}_v^{(l)} = LSA(\boldsymbol{L}_v^{(l)}, \boldsymbol{X}_v^{(l)}) \tag{2}$$

where l denotes the l^{th} layer, $L_a^{(l)}$ and $L_v^{(l)}$ represents the latent tokens of audio and visual modalities, $Z_a^{(l)}$ and $Z_v^{(l)}$ represents the compressed tokens of audio and visual modalities respectively.

The output of the LSA module helps to focus only on the most relevant information by compressing into the latent tokens. Now the compressed tokens of the LSA module are fed to the CMA along with the tokens of other modality to compute cross attention across the modalities, which is given by

$$\boldsymbol{X}_{av}^{(l)} = CMA(\boldsymbol{Z}_v^{(l)}, \boldsymbol{X}_a^{(l)}) \tag{3}$$

$$\boldsymbol{X}_{va}^{(l)} = CMA(\boldsymbol{Z}_a^{(l)}, \boldsymbol{X}_v^{(l)}) \tag{4}$$

where $\boldsymbol{X}_{va}^{(l)}$ and $\boldsymbol{X}_{av}^{(l)}$ are the cross-attended features of visual and audio modalities, respectively, where audio modality is used to attend to visual modality and vice-versa. By using the compressed latent audio tokens to attend to visual modality, the quadratic computational cost of the cross attention module is eliminated, while still retaining the benefits of cross attention. Note that separate LAVViT adapters are used for each modality, where the knowledge of one modality is exploited to attend to the other modality.

3.3 ViT with LAVViT Adapters

3.3.1 Standard ViT layer

For the sake of completeness, we briefly review the ViT layer of the conventional ViT architecture (24) as a preliminary to the proposed approach. Traditionally, the standard ViT layer has two modules: (i) Multi-head Self-Attention (MSA) and (ii) Multilayer perceptron (MLP), where each module is

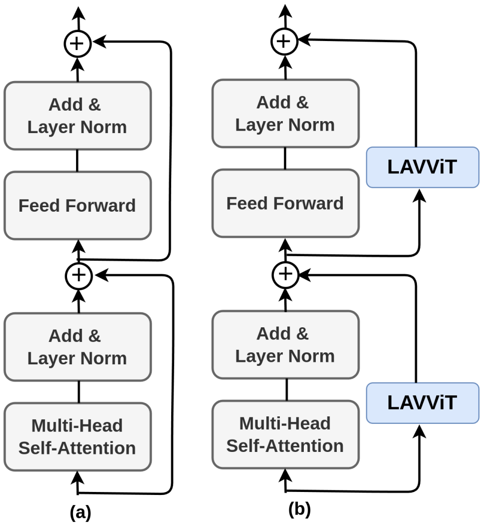

Figure 3: (a) Standard ViT Layer (b) ViT Layer with the integration of LAVViT adapters.

followed by layer norm and residual connections as shown in Fig 3. Given the audio ($\boldsymbol{X}_a^{(l)}$) and visual inputs ($\boldsymbol{X}_v^{(l)}$), the output of the MSA can be expressed as :

$$\boldsymbol{Y}_a^{(l)} = \boldsymbol{X}_a^{(l)} + LN(MSA(\boldsymbol{X}_a^{(l)})) \tag{5}$$

$$\boldsymbol{Y}_v^{(l)} = \boldsymbol{X}_v^{(l)} + LN(MSA(\boldsymbol{X}_v^{(l)})) \tag{6}$$

where LN denotes layer normalization, $\boldsymbol{Y}_a^{(l)}$ and $\boldsymbol{Y}_v^{(l)}$ denotes the output of MSA for audio and visual modalities, respectively.

The output of the MSA is further fed to the MLP module, which is given by

$$\boldsymbol{X}_a^{(l+1)} = \boldsymbol{Y}_a^{(l)} + LN(MLP(\boldsymbol{Y}_a^{(l)})) \tag{7}$$

$$\boldsymbol{X}_v^{(l+1)} = \boldsymbol{Y}_v^{(l)} + LN(MLP(\boldsymbol{Y}_v^{(l)})) \tag{8}$$

Similar to the transformer encoder, these two modules of MSA and MLP are repeated in each ViT layer.

3.3.2 ViT layer with LAVViT adapters

To effectively integrate the LAVViT adapters into the ViT layer, we have injected LAVViT adapters in both the modules of MSA and MLP of each ViT layer. The audio or visual inputs are fed to the LAVViT adapters and the output of the LAVViT adapters are then added to the output of MSA module, which is given by

$$\boldsymbol{Y}_a^{(l)} = \boldsymbol{X}_a^{(l)} + LN(MSA(\boldsymbol{X}_a^{(l)})) + LAV(\boldsymbol{X}_a^{(l)}, \boldsymbol{X}_v^{(l)}) \tag{9}$$

$$\boldsymbol{Y}_v^{(l)} = \boldsymbol{X}_v^{(l)} + LN(MSA(\boldsymbol{X}_v^{(l)})) + LAV(\boldsymbol{X}_v^{(l)}, \boldsymbol{X}_a^{(l)}) \tag{10}$$

Similarly, another LAVViT adapter is integrated into the MLP module of each ViT layer, which is expressed as

$$\boldsymbol{X}_a^{(l+1)} = \boldsymbol{Y}_a^{(l)} + LN(MLP(\boldsymbol{Y}_a^{(l)})) + LAV(\boldsymbol{Y}_a^{(l)}, \boldsymbol{Y}_v^{(l)}) \tag{11}$$

$$\boldsymbol{X}_v^{(l+1)} = \boldsymbol{Y}_v^{(l)} + LN(MLP(\boldsymbol{Y}_v^{(l)})) + LAV(\boldsymbol{Y}_a^{(l)}, \boldsymbol{Y}_v^{(l)}) \tag{12}$$

Separate LAVViT adapters are deployed for the audio and visual modalities in each ViT layer. By integrating LAVViT adapters in both MSA and MLP modules of each ViT layer, we are able to effectively capture the information using a few trainable parameters.

3.4 Overview of the proposed approach

The audio and visual inputs are fed to parallel ViT architectures pre-trained on visual data independent of each other. By deploying the same architecture with shared pre-trained weights, the need to train the modalities separately with modality-specific backbones can be avoided. We have further deployed bidirectional attention mechanisms where we have used separate LAVViT adapters for audio and visual modalities. This helps in effectively deploying cross-attention to leverage the knowledge of one modality to attend to another modality in a bidirectional fashion. By freezing the pre-trained weights and training only the parameters of the LAVViT adapters injected into each ViT layer, it helps to eliminate the large-scale training of fine-tuning all the parameters of ViT. Finally, the trainable parameters of the LAVViT adapters are optimized using Additive Angular Margin Softmax (AAMSoftmax) function (38).

4 Results and Discussion

4.1 Dataset

We have evaluated the proposed approach on the Voxceleb1 dataset (39), which consists of 1,48,642 videos, captured under challenging environments from Youtube videos. The videos are obtained from 1251 speakers, which is gender balanced with 55% of speakers being male. Each video spans a duration of 4 to 145 seconds, and the videos are chosen from a wide range of ethnicities, accents, professions, and ages. For our experiments, we divided the Voxceleb1 development set of 1211 speakers by randomly selecting 1150 speakers for training and 61 speakers for validation. The results of the proposed approach have been reported on both validation and test sets.

4.2 Evaluation Metrics

We have considered Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) as evaluation metrics to measure the performance of the proposed approach, whcih has been widely explored in the literature of speaker verification (16; 33). EER provides the error rate at which the false acceptance rate (FAR) and false rejection rate (FRR) are equal. So lower the EER, higher will be the performance and reliability of the system. On the other hand, minDCF provides the control over the costs associated with false positives (imposters) and false negatives (missed detections) (40). By using these two metrics, we are able to comprehensively evaluate the performance of our model.

4.3 Ablation Study

It is worth mentioning that all the experiments are conducted only on the Voxceleb1 dataset. In order to understand the impact of the latent tokens, we have conducted experiments without the latent tokens in the LAVViT adapters i.e., performing cross attention on all the original audio and visual tokens. In the LAVViT adapters, we have removed the LSA modules and computed the cross-attention across the original tokens of the modalities. Note that the latent tokens helps in reducing the quadratic computational cost of cross attention by compressing the audio or visual tokens not the parameters. We can observe that directly applying the cross attention across the original audio and visual tokens drastically increases the computational complexity due to the large number of audio or visual tokens. We can also observe that the performance of the system without latent tokens slightly degrades compared to that of the one with latent tokens as shown in Table 1. This can be attributed to the fact that the latent tokens are able to act as an attention bottleneck, focusing on the most relevant information using a few trainable parameters.

We have further evaluated the proposed approach using multiple variants of the ViT transformers (ViT and Swin architectures) for both the scenarios of with and without latent tokens. We have considered two variants of Swin transformers (41) (Swin-V2-B and Swin-V2-L) and two variants of ViT transformers (ViT-B-16 and ViT-L-16). Since Swin transformers introduce the inductive bias of images into the ViT architecture, we can observe Swin transformers outperform the standard ViT architecture. Among the two variants of Swin transformers, Swin-V2-L exhibits the best performance as it employs more number of parameters than that of Swin-V2-L.

Table 1: Performance of the proposed approach w/ and w/o latent tokens on the validation set. We also present the results with various architectures of ViTs in case of both w/ and w/o latent tokens. Here w and w/o denotes with and without respectively.

ViT Architecture	**Total Params (M)**	**Trainable Params (M)**	**Validation Set**	
			EER ↓	**minDCF ↓**
LAVViT (w/o latent tokens)				
ViT-B-16	107.2	2.1	2.105	0.140
ViT-L-16	340.1	6.7	1.894	0.129
Swin-V2-B	114.2	2.3	1.903	0.121
Swin-V2-L	238.8	4.6	1.826	0.119
LAVViT (w/ latent tokens)				
ViT-B-16	107.2	4.7	1.993	0.134
ViT-L-16	340.1	14.5	1.841	0.122
Swin-V2-B	114.2	5.0	1.735	0.115
Swin-V2-L	238.8	10.1	1.649	0.104

Table 2: Performance of the proposed approach in comparison to state-of-the-art models on the validation and Vox1-O sets.

Fusion Method	**Validation Set**		**Vox1-O Set**	
	EER ↓	**minDCF ↓**	**EER ↓**	**minDCF ↓**
Visual	3.720	0.298	3.779	0.274
Audio	2.553	0.253	2.529	0.228
Deep Cleanse (30)	2.476	0.203	2.409	0.198
JCA (20)	2.173	0.126	2.214	0.129
DCA (34)	2.138	0.119	2.172	0.121
RJCA (33)	1.851	0.112	1.975	0.116
LAVViT (Ours)	**1.649**	**0.104**	**1.785**	**0.110**

4.4 Comparision to state-of-the-art

To have a fair comparison with the state-of-the-art (SOTA), we have used the same train and validation set partitions as provided by the authors of (34). Similar to that of (34), we have used the deep audio-visual cleansing framework of (30) to refine the noisy samples of the Voxceleb1 dataset. The results of the proposed approach have been compared to the SOTA approaches on the validation and Vox1-O test sets as depicted in Table 2. First, we evaluated the performance with the individual modalities and found that the audio modality outperforms the visual modality. By deploying a simple score-level fusion as in (30), the performance of the fusion model performs better than that of individual modalities. (20) and (33) improved the performance of the system by employing sophisticated cross attention models using Resnet-18 (42) for visual and ECAPA-TDNN (43) for audio modality as backbones, respectively. (34) further improved the performance of cross-attention models by addressing the problem of weak complementary relationships. In most of the SOTA approaches, separate architectures are used for training the backbones, whereas the proposed approach does not require separate modality-specific backbones. Unlike existing methods, we rely on the ViTs for a unified training framework of audio and visual modalities, yet efficiently training with a few trainable parameters using the LAVViT adapters. We can observe that the proposed approach demonstrates better performance than SOTA models by leveraging the potential of ViT transformers.

5 Conclusion

In this work, we have explored the potential of Vision Transformers (ViTs) for audio-visual speaker verification (SV) using Latent Audio-Visual Vision Transformer (LAVViT) adapters. By introducing latent tokens in the LAVViT adapters, which are injected into every layer of the ViT, we demonstrate that ViTs can be efficiently trained using a few trainable parameters. The latent tokens in each LAVViT module help us to drastically reduce the original number of audio and visual tokens, thereby avoiding the quadratic computational cost of the cross-modal attention. By using the same architecture with shared pre-trained weights, the proposed approach circumvents the need to train separate modality-specific backbones. Extensive experiments were conducted to demonstrate that exploring ViTs in a parameter-efficient way using LAVViT adapters is a promising line of research. The proposed approach can be further enhanced by training with Voxceleb2 dataset as it can improve the generalization ability of the model. We can also further investigate the integration of sophisticated cross-attention models into the LAVViT adapters to further improve system performance.

Acknowledgments and Disclosure of Funding

The authors wish to acknowledge the funding from the Government of Canada's New Frontiers in Research Fund (NFRF) through grant NFRFR-2021-00338.

References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] Y. Tu, W. Lin, and M.-W. Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99038–99049, 2022.

[3] Z. Zhao, Z. Li, W. Wang, and P. Zhang, "Pcf: Ecapa-tdnn with progressive channel fusion for speaker verification," in *IEEE ICASSP*, pp. 1–5, 2023.

[4] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 1385–1389, 2021.

[5] J. Peng, O. Plchot, T. Stafylakis, L. Mosner, L. Burget, and J. H. Černocký, "Improving Speaker Verification with Self-Pretrained Transformer Models," in *Proc. INTERSPEECH 2023*, pp. 5361–5365, 2023.

[6] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[7] G.-S. J. Hsu, H.-Y. Wu, and M. H. Yap, "A comprehensive study on loss functions for cross-factor face recognition," in *IEEE CVPRW*, pp. 3604–3611, 2020.

[8] R. G. Praveen, P. Cardinal, and E. Granger, "Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2023.

[9] R. G. Praveen, E. Granger, and P. Cardinal, "Recursive joint attention for audio-visual fusion in regression based emotion recognition," in *IEEE ICASSP*, pp. 1–5, 2023.

[10] R. G. Praveen, E. Granger, and P. Cardinal, "Cross attentional audio-visual fusion for dimensional emotion recognition," in *IEEE FG*, 2021.

[11] R. G. Praveen, W. C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A. L. Koerich, S. Bacon, P. Cardinal, and E. Granger, "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2486–2495, June 2022.

[12] B. Duan, H. Tang, W. Wang, Z. Zong, G. Yang, and Y. Yan, "Audio-visual event localization via recursive fusion by joint co-attention," in *IEEE WACV*, pp. 4012–4021, 2021.

[13] J.-T. Lee, M. Jain, H. Park, and S. Yun, "Cross-attentional audio-visual fusion for weakly-supervised action localization," in *International Conference on Learning Representations*, 2021.

[14] Seyed, C. Greenberg, E. Singer, D. Olson, L. Mason, and J. Hernandez-Cordero, "The 2019 nist audio-visual speaker recognition evaluation," The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, -1, 2020-05-18 2020.

[15] J. Alam, G. Boulianne, L. Burget, M. Dahmane, M. S. Diez, O. Glembek, M. Lalonde, A. D. Lozano, P. Matějka, P. Mizera, L. Mošner, C. Noiseux, J. Monteiro, O. Novotný, O. Plchot, A. J. Rohdin, A. Silnova, J. Slavíček, T. Stafylakis, P.-L. St-Charles, S. Wang, and H. Zeinali, "Analysis of abc submission to nist sre 2019 cmn and vast challenge," vol. 2020, pp. 289–295, 2020.

[16] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb," in *Proc. Interspeech*, pp. 2252–2256, 2020.

[17] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1079–1092, 2021.

[18] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *IEEE ICASSP*, pp. 3995–3999, 2019.

[19] S. Hörmann, A. Moiz, M. Knoche, and G. Rigoll, "Attention fusion for audio-visual person verification using multi-scale features," in *IEEE FG*, pp. 281–285, 2020.

[20] G. P. Rajasekhar and J. Alam, "Audio-visual speaker verification via joint cross-attention," in *Speech and Computer:25th International Conference, SPECOM*, p. 18–31, 2023.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, 2022.

[23] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir, "Transformers in speech processing: A survey," 2023.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[25] A. Nagrani, S. Yang, A. Arnab, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *NIPS*, 2021.

[26] Y.-B. Lin, Y.-L. Sung, J. Lei, M. Bansal, and G. Bertasius, "Vision transformers are parameter-efficient audio-visual learners," in *IEEE CVPR*, 2023.

[27] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *IEEE CVPR*, pp. 8427–8436, 2018.

[28] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. of ECCV*, 2018.

[29] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *IEEE ICASSP*, pp. 6194–6198, 2021.

[30] R. Tao, K. A. Lee, Z. Shi, and H. Li, "Speaker recognition with two-step multi-modal deep cleansing," in *IEEE ICASSP*, pp. 1–5, 2023.

[31] M. Liu, K. A. Lee, L. Wang, H. Zhang, C. Zeng, and J. Dang, "Cross-modal audio-visual co-learning for text-independent speaker verification," in *IEEE ICASSP*, pp. 1–5, 2023.

[32] P. Sun, S. Zhang, Z. Liu, Y. Yuan, T. Zhang, H. Zhang, and P. Hu, "A Method of Audio-Visual Person Verification by Mining Connections between Time Series," in *Proc. INTERSPEECH 2023*, pp. 3227–3231, 2023.

[33] R. G. Praveen and J. Alam, "Audio-visual person verification based on recursive fusion of joint cross-attention," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5, 2024.

[34] R. G. Praveen and J. Alam, “Dynamic cross attention for audio-visual person verification,” in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5, 2024.

[35] R. Gnana Praveen and J. Alam, “Cross-modal transformers for audio-visual person verification,” in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pp. 240–246, 2024.

[36] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, “Vision transformer adapter for dense predictions,” in *The Eleventh International Conference on Learning Representations*, 2023.

[37] D. Yin, Y. Yang, Z. Wang, H. Yu, K. Wei, and X. Sun, “1% vs 100%: Parameter-efficient low rank adapter for dense predictions,” in *IEEE CVPR*, pp. 20116–20126, 2023.

[38] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *IEEE/CVF Conference on CVPR*, pp. 4685–4694, 2019.

[39] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in *INTERSPEECH*, 2017.

[40] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.

[41] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, “Swin transformer v2: Scaling up capacity and resolution,” in *IEEE CVPR*, 2022.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE CVPR*, pp. 770–778, 2016.

[43] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Proc. Interspeech*, pp. 3830–3834, 2020.