On the Combination of Multi-Input and Self-Attention for Sign Language Recognition

Nam Vu Hoai[1], Thuong Vu Van[2], Dat Tran Anh*[3]
Faculty of Information Technology Posts and Telecommunications Institute of Technology Ha Noi, 11398, Viet Nam[1]
Innovation and Entrepreneurship Center Posts and Telecommunications Institute of Technology, Ha Noi, 11398, Viet Nam[2]
Faculty of Information Technology, Thuyloi University, Ha Noi, 11398, Viet Nam[3]

***Abstract*—Sign language recognition can be considered as a branch of human action recognition. The deaf-muted community utilizes upper body gestures to convey sign language words. With the rapid development of intelligent systems based on deep learning models, video-based sign language recognition models can be integrated into services and products to improve the quality of life for the deaf-muted community. However, comprehending the relationship between different words within videos is a complex and challenging task, particularly in understanding sign language actions in videos, further constraining the performance of previous methods. Recent methods have been explored to generate video annotations to address this challenge, such as creating questions and answers for images. An optimistic approach involves fine-tuning autoregressive language models trained using multi-input and self-attention mechanisms to facilitate understanding of sign language in videos. We have introduced a bidirectional transformer language model, MISA (multi-input self-attention), to enhance solutions for VideoQA (video question and answer) without relying on labeled annotations. Specifically, (1) one direction of the model generates descriptions for each frame of the video to learn from the frames and their descriptions, and (2) the other direction generates questions for each frame of the video, then integrates inference with the first aspect to produce questions that effectively identify sign language actions. Our proposed method has outperformed recent techniques in VideoQA by eliminating the need for manual labeling across various datasets, including CSL-Daily, PHOENIX14T, and PVSL (our dataset). Furthermore, it demonstrates competitive performance in low-data environments and operates under supervision.**

***Keywords*—Multi-input; self-attention; deep learning models; video-based sign language; sign language recognition**

I. Introduction

According to the National Disability Survey at the end of 2016 and the beginning of 2017 (VDS2016), Vietnam has approximately 6.2 million persons with disabilities (PWDs), including around 2 million persons with speech and hearing impairments [1]. Hearing and speech are innate faculties possessed by most individuals. However, a significant portion of the population lacks these faculties and faces challenges in interpersonal communication. According to the World Health Organization, an estimated 70 million individuals worldwide are affected by deafness and muteness, with a total of 360 million individuals experiencing some form of hearing impairment, among whom 32 million are children. Deaf-muted children often encounter significant challenges in accessing public services such as education and healthcare. Mainly, educational programs not explicitly designed for deaf-mute children can impede their development compared to the normal ones. Advanced technologies are becoming increasingly prevalent in enhancing our quality of life. The deaf-mute community, especially children, can benefit from these rapid developments. Establishing a sign language recognition model to support the deaf-mute community in learning and communication would be a significant step towards bridging the gap between them and the external world. This sign language recognition model can be integrated into applications to assist them in accessing public services and daily communication. Additionally, it can aid family members in learning sign language to communicate with their deaf-mute relatives [2].

In recent years, multi-input and self-attention mechanisms have garnered significant attention in the computer vision community. Convolutional Neural Networks (CNNs) [3] have been widely applied in image recognition [4], semantic segmentation [5], and object detection [6], [7], achieving high performance across various evaluation metrics. The integration of Multi-input [8] into CNNs has dramatically improved both accuracy and speed, as it enables the model to learn better features. On the other hand, the self-attention mechanism [27] was first introduced as an effective solution to natural language processing tasks. Subsequently, this mechanism was applied to deep learning models for the computer vision domain with promising results. Recently, with the emergence of Vision Transformer [10], the attention mechanism has even achieved higher efficiency than CNN models in some vision tasks. While both approaches have demonstrated significant success independently, they consist of separate architectures for various tasks, with minimal integration for sign language recognition. The multi-input methodology leverages various input perspectives to construct synthesized functions for feature extraction from each input [11], including RGB images, blurred images, and binary images. In contrast, self-attention modules utilize input features to construct attention functions among interconnected pixels [12], prioritizing different regions and capturing more precise feature information within the image. Integrating these two approaches could be a viable solution for the sign language recognition problem. The strength of the combination is that it would significantly enhance the performance of sign language recognition.

This paper aims to explore a more integrated relationship between Multi-input and Self-attention modules in recognizing sign language words. By segmenting the tasks of each module and subsequently amalgamating them into a unified framework, we develop a cohesive model called MISA, which merges Multi-input and Self-attention techniques to enhance efficiency and reduce computational time in addressing sign language recognition challenges. We initially apply the Multi-input module to project the input image and extract a comprehensive set of intermediate features to achieve this. These features are then synthesized and employed within the Self-attention module. Through this integration, the MISA model harnesses the strengths of both modules and proves effective in prediction tasks. Additionally, we construct a PVSL dataset consisting of videos of sign language problems. The videos were collected by setting up a camera system to capture the upper body of individuals while performing sign language gestures.

In summary, our contributions are as follows:

- **New dataset**: We published PVSL, a new dataset of Vietnamese sign language in the form of videos.
- **Novel model**: We proposed a novel model, MISA, combining two modules, Multi-input and Self-attention.
- **Analysis and evaluation**: We evaluated our model on two public datasets and PVSL.

The remainder of this paper is structured as follows: Section II discusses relevant previous studies. Section III presents our method. Section IV gives the experimental evaluation. Finally, Section V provides some concluding remarks and a brief discussion.

II. Related Works

A. Multi-input Learning

Multi-input aims to process information from images and natural language [13], [14] to train feature sets and learn their representations. This approach has shown promising results across various tasks on multi-source datasets. The success of this approach has also motivated numerous research teams to develop and train multi-input transformer models alongside vision-based models concurrently [15], [16], [17], [18]. However, these studies frequently rely on learning representations of vision-based or natural language-based data through weight updates. Subsequently, a supervised learning model that can be resource-intensive is constructed [19], [20] for dealing with various tasks from videos [21], [22]. On the contrary, our approach entails automatically generating annotations for frames in videos to facilitate comprehension. Moreover, our model can learn global weights, eliminating the need for frequent weight updates during training from multi-input, thus demonstrating the benefit of learning these global weights after pre-training and efficiently training a supervised model for sign language recognition.

B. Learning with Attention Models

The self-attention mechanism has been widely used in recent deep learning models due to its ability to handle long-range dependencies in computer vision tasks [23], [24]. Transformer models, which utilize self-attention, have emerged to solve various computer vision tasks such as image processing and pattern recognition [25], [26]. Numerous attention mechanisms have been proposed to improve the object recognition model's performance in images and videos. As a result, numerous research studies have employed attention modules or leveraged multi-channel information to aggregate image features. In particular, [29], [30], [31] have employed channel-wise attention re-calibration, while the research of [32] have re-calibrated both channel and spatial positions to refine feature maps. [33] has extended the number of convolutional layers with attention map blocks to create distinct independent pipelines. [34] has replaced convolutional operations with self-attention mechanisms in the final stages of the model. Overall, studies have alleviated the local limitations of conventional convolutional networks by incorporating self-attention neural networks.

C. Discussion

In general, studies on multi-input primarily focus on the premise that adding more inputs enhances processing speed and increases model storage memory. Therefore, our research team combined multi-input with an attention model to focus on important input features among a multitude of inputs, thereby improving model accuracy and computational speed.

III. Materials and Methods

A. PVSL Dataset

The PVSL dataset, depicted in Fig. 1, was created to offer the research community a diverse collection of sign language words that are relevant for both research and real-world applications. The dataset is designed to include sign language words commonly used by the deaf-mute community in their daily lives, covering topics such as family communication, educational settings, healthcare, shopping services, and daily communication.

We have involved the participants of the deaf-mute community during dataset collecting periods. The participants include sign language experts, teachers, and students learning sign language at special schools. The participants were asked to perform a set of pre-defined sign language words in front of a camera. They must express sign language words naturally, as they use them daily. Before data collection, we provided training to ensure their understanding through experts and guiding teachers, thus ensuring the accuracy and quality of the PVSL dataset. Video data were collected from 12 participants performing sign language gestures. All participants understood sign language, including five who were deaf-mute. Videos were captured at a resolution 1920x1080 with a frame rate of 30 FPS. The video frames were carefully trimmed at the beginning and end to represent a sign language word accurately. The detailed statistics of the dataset are presented in Table I.

B. Model Description

Our proposed MISA architecture, illustrated in Fig. 2, is designed to combine several parallel language models with a pre-trained image recognition model. The key challenge is to establish a connection between images and text captions to generate a multimodal interpretation that supports sign language recognition. To overcome this challenge, we have integrated two models: an image-to-text projection model and a language model that facilitates sign language recognition. We will now provide a detailed description of our model, outlining the three architectural components: (i) A language model for learning text features, (ii) An image-to-text transfer model, and

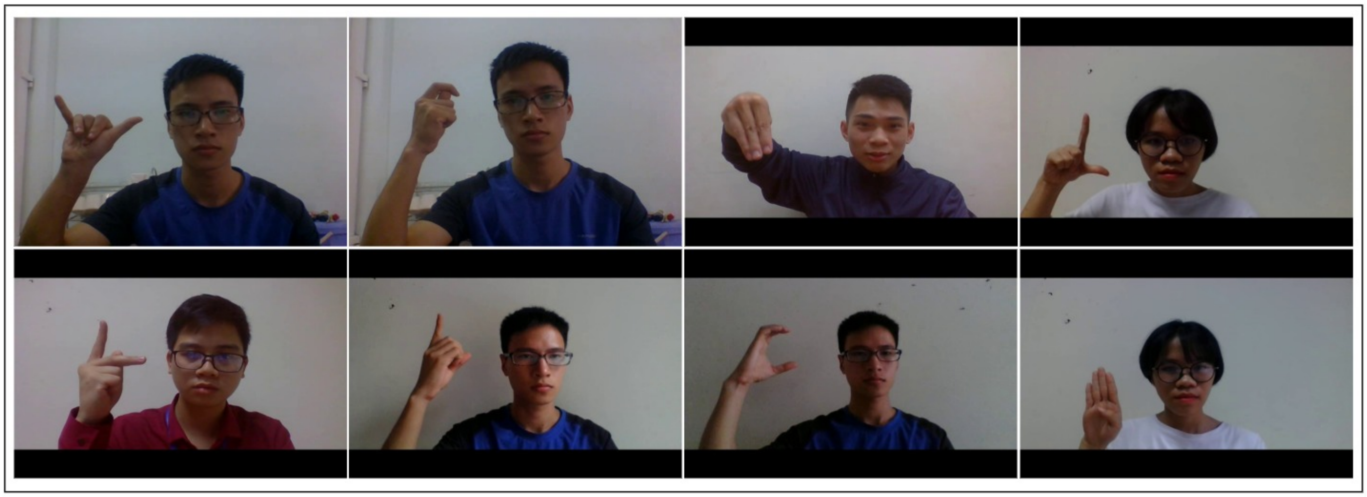

Fig. 1. The illustrative images of the PVSL dataset depict variations in lighting conditions, different backgrounds, and signers with diverse appearances.

TABLE I. Overview of Word-level Datasets in other Languages

Dataset	#Signs	#Videos	#Signers	Type	Sign Language
CSL-Daily [14]	1,066	8,257	9	RGB	Chinese
PHOENIX14T [37]	2,000	20,654	10	RGB	German
PVSL (our dataset)	50	5068	12	RGB	Vietnamese

(iii) A model that merges the two aforementioned components (i) and (ii) into a prediction model.

The language processing model: We use a Transformer-based encoding scheme to encode textual information in this model. To do this, we first tokenize the text into vocabulary units and then into token sequences x. Subsequently, we embed these tokens into a D-dimensional space, which captures contextual information (as shown in Eq. 1). These token embeddings are then mapped with a mask to help classify words based on their distributional properties. This model plays a crucial role in helping us understand information from videos that support sign language recognition.

$$e = WordEmbedding(x) \tag{1}$$

The video processing model: The video is divided into frames, denoted by $f = {f_i}_1^T$. Each frame is then processed by an encoder to generate feature vectors, $v = {v_i}_1^T$, using Eq. 2. We use the ViT encoder [10] with a resolution of 224x224 per frame. Additionally, we incorporate a mapping between images and image descriptions obtained from over 300 million image-text pairs crawled from the internet. The encoder's parameters remain fixed throughout the experimentation process.

$$v_{1:T} = Encoder(f_{1:T}) \tag{2}$$

The integration of Language Processing Model and Video Processing Model: The video features are turned into short answer sentences using a language model. These answers are obtained by mapping video features linearly through an image-to-text projection. The answers are then combined with previous texts and passed through the Transformer encoder to improve sign language recognition results. In the Transformer encoder section, we merge the question with the answer to enhance the accuracy of sign language recognition on the video. To achieve this, the model learns strong multi-modal interactions while maintaining the Transformer's encoding weight sets. We normalize the preceding layer before passing it through the self-attention layer, and each layer is directly fed into the pre-encoder Transformer. As a result, the accuracy of sign language recognition on the video is significantly improved.

C. State Space Model (SSM)

The Structured State-Space Model (SSM), as shown in Fig. 3, is a new type of sequence model in deep learning. It encompasses recurrent neural networks (RNNs), convolutional neural networks (CNNs), and classical state-space models combined with self-attention. These models are inspired by a continuous system that maps a function or one-dimensional sequence, $x_t \in \mathbb{R}$, to $y_t \in \mathbb{R}$, using an unknown hidden state $h_t \in \mathbb{R}^N$.

The structure of SSM independently maps each channel (e.g., D = 5) of the input x to the output y through hidden states of higher dimension h (e.g., N = 4). Previous SSMs avoided realizing this large effective state (DN, multiplied by the batch size B and sequence length L) through intelligent alternative computational paths that require time-invariant parameters that remain constant over time.

D. Loss Function

In the previous section, we discussed training a model for sign language recognition. This is a difficult task because generating answers from videos is not a straightforward

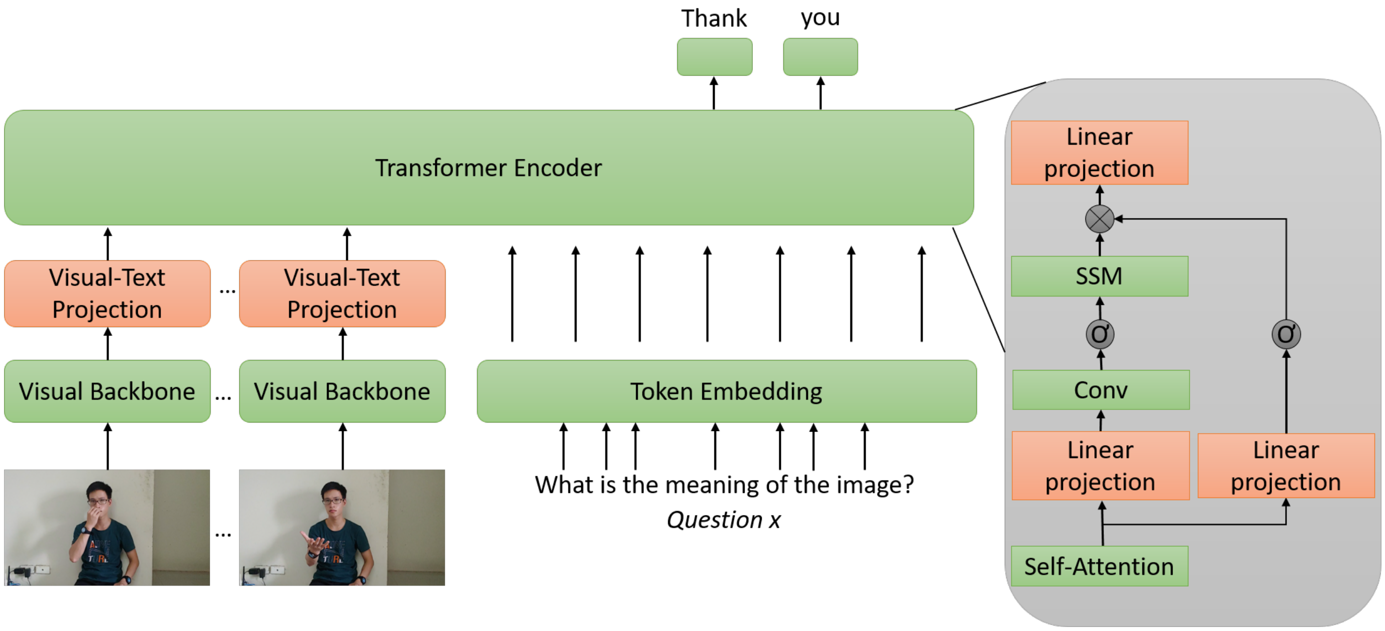

Fig. 2. The framework of the proposed MISA.

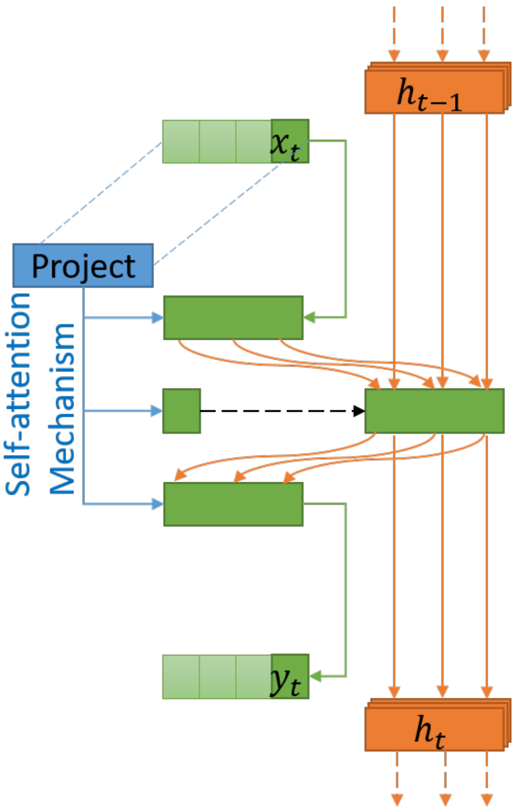

Fig. 3. Structured State Space Models (SSMs).

process, and real-world data can be hard to recognize. To tackle this challenge, we used image-answer pairs from the internet, which are relatively easy to collect and incorporate into training. We trained the model using the parameters of the image-to-text projection model and the combined and coordinated model. To achieve this, we used a language model objective function with a masked image. In this function, x_m represents segments of masked text that need to be predicted, and the model must predict these segments along with the corresponding image content. In terms of computation, we constructed the loss function $L(x, y)$ as follows Eq. (3):

$$L(x,y) = -\frac{1}{N} logp(\hat{x}, y)_m^{x_m} \tag{3}$$

where $\hat{x}$ is the text-encoded sequence from the question, y is the video frame sequence, $p(\hat{x}, y)_m^{x_m}$ is the probability for the m-th token (masked) in x to be x_m, and N is the number of masks in the $\hat{x}$ sequence.

IV. Experiments

A. Experimental Setup

MISA model: We employed a parallel language model with 370 million parameters, Mamba [35], trained with the MLM objective on a 160G text corpus and tokenized using SentencePiece [36] with a vocabulary size of V = 128,000. The input of the MISA model consists of a question about sign language recognition and a video. The task is to find the correct answer from a vast vocabulary set A comprising approximately 2,000 answers. The answers are all concise, meaning that most answers consist of only one or two words of sign language recognition. A token [CLS] and a token [SEP] are added respectively at the beginning and end of each text sequence. Meanwhile, [MASK] represents the sign language word being sought. We design the following prompt:

"[CLS] Question: <Question>? Answer: The action of sign language is <Answer Candidate> [MASK]. Subtitles: <Subtitles> [SEP]"

Datasets:To conduct our training, we utilized the publicly available WebVid10M dataset [42], comprising 10 million pairs of video-text, where video annotations are derived from available alternative descriptions. Additionally, we generated 20 thousand pairs of video-text from two datasets: CSL-Daily [14] and PHOENIX14T [37] leveraging Image Captioning technology [38]. We evaluated the outcomes on a subsequent dataset encompassing various text and video domains, namely PVSL (our dataset).

Evaluation Metric: We use Word Error Rate (WER) as the evaluation metric, as shown in Eq. 4. Note that the lower WER, the better accuracy.

$$WER = \frac{sub + ins + del}{ref} \tag{4}$$

In which, sub represents substitution, ins represents insertion, and del represents deletion. These operations are essential for transforming the predicted sentence into the reference sentence. Hence, ref denotes the reference sentence.

B. Experimental Results

Our empirical study in this subsection is designed to answer three key research questions (RQs).

- RQ1. How does the MISA model improve the performance of sign language recognition compared to current state-of-the-art methods?
- RQ2. How does each scenario in MISA contribute to correct deep learning?
- RQ3. How can deep learning (DL) be visualized, including t-SNE plots of features and distribution plots of predicted scores from the MISA model?

1) Comparison With State-of-the-Art Approaches (RQ1): Table II compares our MISA model and other state-of-the-art methods comprehensively. We evaluated five different methods for the sign language recognition task on videos. Our observations indicate that MISA outperforms other state-of-the-art methods across all three datasets. This superiority is achieved through the attention mechanism within the MISA model, which focuses on the different body gestures of participants. The ability to reduce noise between frames in the video through a propagation or selective forgetting mechanism along the sequence length also contributes to this outperformance. Additionally, MISA demonstrates faster processing speed compared to competitive methods due to its rapid inference capability (processing speed up to five times faster than the traditional Transformer model) and linear scalability with sequence length. The MISA model exhibits significant performance improvements on real-world datasets, even on longer sequences, without incurring additional training costs.

2) Applicability to Fringe Scenarios (RQ2): We initialize the parameter set from a pre-trained language model and fine-tune it with the scenarios outlined in Table III. We have observed that leveraging pre-trained weights from previous successful language models plays a crucial role in our proposed architecture. The model initialized solely for video recognition of sign language (line 1 - the first scenario) exhibits inferior performance compared to the model initialized with combined weights (lines 2 and 4). Notably, the model trained in the second scenario, combining the language and video processing models, outperforms the variant in the third scenario and falls slightly behind the fourth scenario. This observation suggests that integrating video and text as input for the model can yield significant effectiveness. Additionally, the combination in the third scenario (line 3) demonstrates favorable outcomes when integrating the video processing model with the state space model. Ultimately, our proposed MISA model (line 4 - the fourth scenario) illustrates that amalgamating multi-input and self-attention in the state space is the most effective approach for video sign language recognition.

3) Qualitative Study (RQ3): In order to showcase the effectiveness of our method, we used t-SNE [43] to create a visualization of the recognition results obtained from the MISA model on the PVSL test set. The original data from the test set was processed through the MISA model to create a new data dimension. The feature vector size, in our case, was 12288. Next, the data was passed through the MISA model corresponding to the 50 primary training labels, after which t-SNE was used to project and visualize the reduced features in a 2D space. The resulting Fig. 4 provides strong evidence of the superior performance of the combined features with our MISA model.

V. Conclusions

This paper introduces the MISA model, a framework for extending the language model that combines multi-data and self-attention in the state space model (SSM). We trained this model on our self-collected dataset PVSL and data collected from multiple sources. We aimed to address the sign language recognition problem for the deaf-mute community in the context of Video Question Answering (VideoQA). We also conducted an in-depth analysis to demonstrate the effectiveness of our MISA model, which enhances accuracy on three popular sign language datasets.

However, our study has some limitations. First, MISA is quite large, making it impractical for deployment on mobile devices. Second, our model is unable to handle videos with multiple individuals performing sign language. In the future, we aim to enhance the model's efficiency based on unsupervised learning and implement dimensionality reduction methods for video data, which will enable better learning and higher-quality results.

References

[1] T. V. Nguyen, "Women with physical disabilities in northern Vitetnam: the lived experience of pregnancy, childbirth, and maternal healthcare," pp. 1–324, 2021, [Online]. Available: https://eprints.qut.edu.au/207988/1/Thi Vinh_Nguyen_Thesis.pdf

[2] Vu, Hoai-Nam, Trung Hoang, Cong Tran, and Cuong Pham. "Sign Language Recognition With Self-Learning Fusion Model." IEEE Sensors Journal (2023).

TABLE II. EVALUATION OF THE S2T NETWORK COMBINATIONS ON WER (THE LOWER THE BETTER)

Methods	Datasets		
	CSL-Daily [14]	PHOENIX14T [37]	PVSL
FCN [39]	33.2	25.1	26.5
CNN+LSTM+HMM [40]	-	26.5	27.8
Joint-SLRT [14]	32.0	24.5	23.3
Cornet [41]	30.1	20.5	20.2
MISA (our method)	**28.5**	**19.4**	**19.8**

TABLE III. FOUR SCENARIOS WITH DIFFERENT NETWORKS ON WER (THE LOWER THE BETTER)

Scenarios			Datasets		
Language processing model	Video processing model	Connecting model	CSL-Daily [14]	PHOENIX14T [37]	PVSL
no	yes	no	35.0	30.7	31.5
yes	yes	no	30.2	23.1	22.3
no	yes	yes	32.1	25.5	24.2
yes	yes	yes	**28.5**	**19.4**	**19.8**

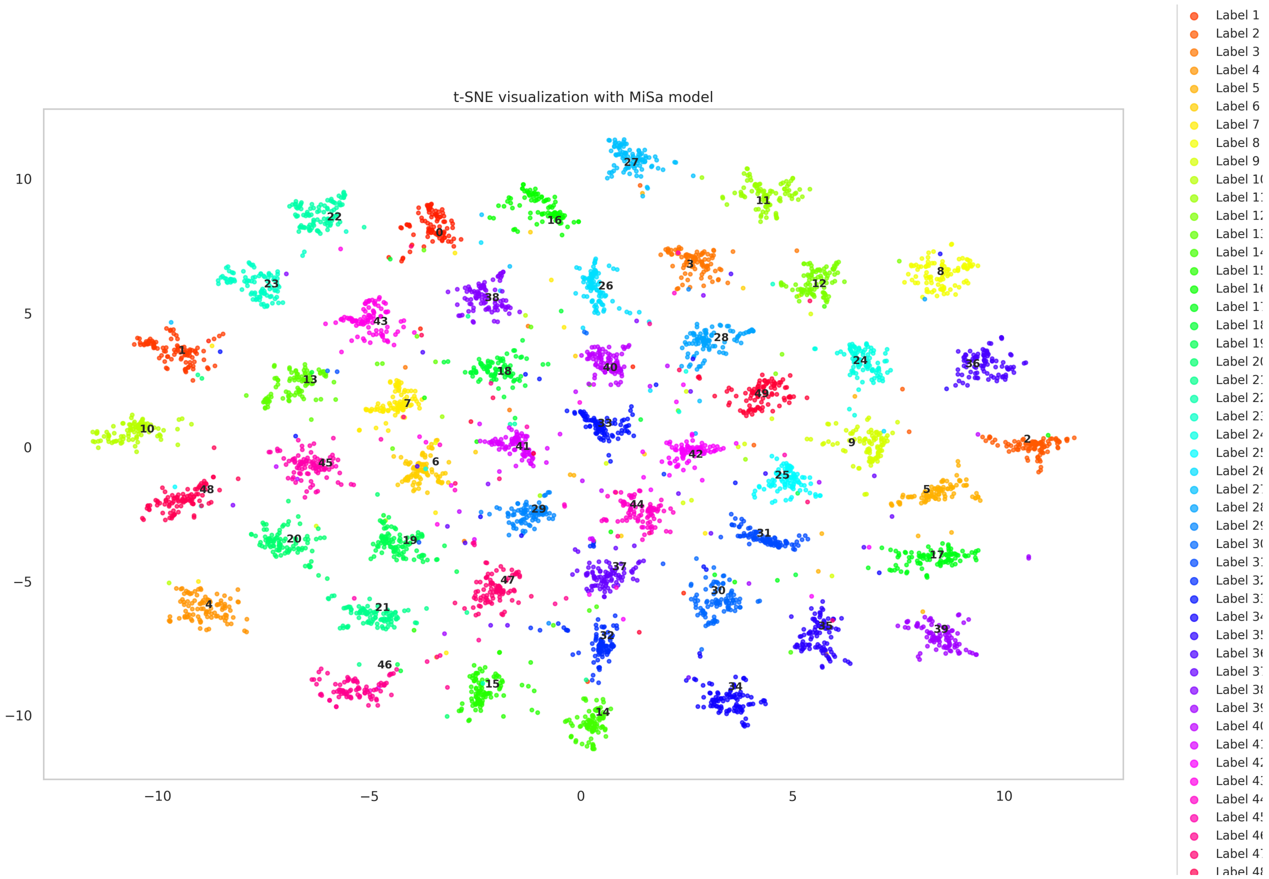

Fig. 4. Feature visualization for MISA architectures.

[3] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," ISPRS J. Photogramm. Remote Sens., vol. 173, no. July 2020, pp. 24–49, 2021, doi: 10.1016/j.isprsjprs.2020.12.010.

[4] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, "Adversarial examples improve image recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 816–825, 2020, doi: 10.1109/CVPR42600.2020.00090.

[5] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," Proc. IEEE Int. Conf. Comput. Vis., pp. 7242–7252, 2021, doi: 10.1109/ICCV48922.2021.00717.

[6] H. Chen et al., "Pre-trained image processing transformer," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 12294–12305, 2021, doi: 10.1109/CVPR46437.2021.01212.

[7] Hoai, Nam Vu, Nguyen Manh Dung, and Soonghwan Ro. "Sinkhole detection by deep learning and data association." In 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), pp. 211-213. IEEE, 2019.

[8] M. Ferianc and M. Rodrigues, "MIMMO : Multi-Input Massive Multi-Output Neural Network," pp. 4564–4569.

[9] B. Yang, L. Wang, D. F. Wong, S. Shi, and Z. Tu, "Context-aware Self-Attention Networks for Natural Language Processing," Neurocomputing, vol. 458, pp. 157–169, 2021, doi: 10.1016/j.neucom.2021.06.009.

[10] H. Fan et al., "Multiscale Vision Transformers," Proc. IEEE Int. Conf. Comput. Vis., pp. 6804–6815, 2021, doi: 10.1109/ICCV48922.2021.00675.

[11] J. Fang, J. Yang, A. Khader and L. Xiao, "MIMO-SST: Multi-Input Multi-Output Spatial-Spectral Transformer for Hyperspectral and Multispectral Image Fusion," in IEEE Transactions on Geoscience and Remote Sensing, doi: 10.1109/TGRS.2024.3361553.

[12] X. Pan et al., "On the Integration of Self-Attention and Convolution," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2022–June, pp. 805–815, 2022, doi: 10.1109/CVPR52688.2022.00089.

[13] Kumar, A., Sachdeva, N. Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimedia Systems 28, 2027–2041 (2022). https://doi.org/10.1007/s00530-020-00672-7

[14] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving Sign Language Translation with Monolingual Data by Sign Back-Translation," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1316–1325, 2021, doi: 10.1109/CVPR46437.2021.00137.

[15] S. Karthick, M. Ramesh Babu, S. Gomathi, D. Kirubakaran, I. Cephas and M. R. Faridha Banu, "Analysis of Multi Input Transformer Coupled Bidirectional DC-AC Converter for Hybrid System," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 145-153, doi: 10.1109/ICOEI53556.2022.9777236.

[16] Xie, J., Li, J., Zhu, M., Wang, Q. (2023). Multi-step Air Quality Index Forecasting Based on Parallel Multi-input Transformers. In: Lu, H., Blumenstein, M., Cho, SB., Liu, CL., Yagi, Y., Kamiya, T. (eds) Pattern Recognition. ACPR 2023. Lecture Notes in Computer Science, vol 14408. Springer, Cham. https://doi.org/10.1007/978-3-031-47665-5_5

[17] Y. Chen, P. Wang, Y. Elasser and M. Chen, "Multicell Reconfigurable Multi-Input Multi-Output Energy Router Architecture," in IEEE Transactions on Power Electronics, vol. 35, no. 12, pp. 13210-13224, Dec. 2020, doi: 10.1109/TPEL.2020.2996199.

[18] L. Yang, Z. Zhu, X. Lin, J. Nong, and Y. Liang, "Long-Range Grouping Transformer for Multi-View 3D Reconstruction," 2023, [Online]. Available: http://arxiv.org/abs/2308.08724

[19] J. A. Prenner and R. Robbes, "Making the Most of Small Software Engineering Datasets With Modern Machine Learning," in IEEE Transactions on Software Engineering, vol. 48, no. 12, pp. 5050-5067, 1 Dec. 2022, doi: 10.1109/TSE.2021.3135465.

[20] Magumba, M.A., Nabende, P. Evaluation of different machine learning approaches and input text representations for multilingual classification of tweets for disease surveillance in the social web. J Big Data 8, 139 (2021). https://doi.org/10.1186/s40537-021-00528-5

[21] S. Oprea et al., "A Review on Deep Learning Techniques for Video Prediction," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 2806-2826, 1 June 2022, doi: 10.1109/TPAMI.2020.3045007.

[22] Liu, H., Ruan, Z., Zhao, P. et al. Video super-resolution based on deep learning: a comprehensive survey. Artif Intell Rev 55, 5981–6035 (2022). https://doi.org/10.1007/s10462-022-10147-y

[23] X. Zhang, Y. Hu, H. Wang, X. Cao, and B. Zhang, "Long-range attention network for multi-view stereo," Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vision, WACV 2021, vol. c, pp. 3781–3790, 2021, doi: 10.1109/WACV48630.2021.00383.

[24] D. M. Argaw, J.-Y. Lee, M. Woodson, I. S. Kweon, and F. C. Heilbron, "Long-range Multimodal Pretraining for Movie Understanding," pp. 13392–13403, 2023, [Online]. Available: http://arxiv.org/abs/2308.09775

[25] Acheampong, F.A., Nunoo-Mensah, H. & Chen, W. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif Intell Rev 54, 5789–5829 (2021). https://doi.org/10.1007/s10462-021-09958-2

[26] I. Misra, R. Girdhar, and A. Joulin, "An End-to-End Transformer Model for 3D Object Detection," Proc. IEEE Int. Conf. Comput. Vis., pp. 2886–2897, 2021, doi: 10.1109/ICCV48922.2021.00290.

[27] Z. Yang, Y. Wei, and Y. Yang, "Associating Objects with Transformers for Video Object Segmentation," Adv. Neural Inf. Process. Syst., vol. 4, no. NeurIPS, pp. 2491–2502, 2021.

[28] S. Khan et al., "Transformers in Vision," ACM Comput. Surv., vol. 54, no. 10, pp. 1–41, 2022.

[29] A. Behera, Z. Wharton, Y. Liu, M. Ghahremani, S. Kumar and N. Bessis, "Regional Attention Network (RAN) for Head Pose and Fine-Grained Gesture Recognition," in IEEE Transactions on Affective Computing, vol. 14, no. 1, pp. 549-562, 1 Jan.-March 2023, doi: 10.1109/TAFFC.2020.3031841.

[30] J. Miao, Y. Wu and Y. Yang, "Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 9, pp. 4624-4634, Sept. 2022, doi: 10.1109/TNNLS.2021.3059515.

[31] X. Li et al., "Disagreement Matters: Exploring Internal Diversification for Redundant Attention in Generic Facial Action Analysis," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2023.3286838.

[32] P. Fang, J. Zhou, S. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," Proc. IEEE Int. Conf. Comput. Vis., vol. 2019–October, pp. 8029–8038, 2019, doi: 10.1109/ICCV.2019.00812.

[33] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural Sign Language Translation," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 7784–7793, 2018, doi: 10.1109/CVPR.2018.00812.

[34] F. B. Slimane and M. Bouguessa, "Context Matters: Self-Attention for Sign Language Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 7884-7891, doi: 10.1109/ICPR48806.2021.9412916.

[35] Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." arXiv preprint arXiv:2312.00752 (2023).

[36] C. Mugisha and I. Paik, "Optimization of Biomedical Language Model with Optuna and a Sentencepiece Tokenization for NER," 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 3859-3861, doi: 10.1109/BIBM55620.2022.9994919.

[37] B. Zhou et al., "Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining," pp. 20871–20881, 2023, [Online]. Available: http://arxiv.org/abs/2307.14768

[38] Y. Zhou, Z. Hu, D. Liu, H. Ben, and M. Wang, "Compact Bidirectional Transformer for Image Captioning," 2022, [Online]. Available: http://arxiv.org/abs/2201.01984

[39] Cheng, K.L., Yang, Z., Chen, Q., Tai, YW. (2020). Fully Convolutional Networks for Continuous Sign Language Recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12369. Springer, Cham. https://doi.org/10.1007/978-3-030-58586-0_41

[40] H. Zhang, Z. Guo, Y. Yang, X. Liu, and D. Hu, "C 2 ST: Cross-modal Contextualized Sequence Transduction for Continuous Sign Language Recognition," pp. 21053–21062.

[41] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous Sign Language Recognition with Correlation Network," pp. 2529–2539, 2023, doi: 10.1109/cvpr52729.2023.00249.

[42] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Zero-Shot Video Question Answering via Frozen Bidirectional Language Models," Adv. Neural Inf. Process. Syst., vol. 35, no. NeurIPS, pp. 1–18, 2022.

[43] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, pp. 2579–2605, Nov. 2008.