Harnessing the Power of SVD: An SVA Module for Enhanced Signal Classification

Lei Zhai[1], Shuyuan Yang[1*], Yitong Li[2], Zhixi Feng[1*], Zhihao Chang[1], Quanwei Gao[1],

[1]Xidian University
[2]Xi'an Jiaotong University
zhailei@stu.xidian.edu.cn, syyang@xidian.edu.cn, zxfeng@xidian.edu.cn

Abstract

Deep learning methods have achieved outstanding performance in various signal tasks. However, due to degraded signals in real electromagnetic environment, it is crucial to seek methods that can improve the representation of signal features. In this paper, a Singular Value decomposition-based Attention, SVA is proposed to explore structure of signal data for adaptively enhancing intrinsic feature. Using a deep neural network as a base model, SVA performs feature semantic subspace learning through a decomposition layer and combines it with an attention layer to achieve adaptive enhancement of signal features. Moreover, we consider the gradient explosion problem brought by SVA and optimize SVA to improve the stability of training. Extensive experimental results demonstrate that applying SVA to a generalized classification model can significantly improve its ability in representations, making its recognition performance competitive with, or even better than, the state-of-the-art task-specific models.

Introduction

Signal classification tasks play an important role in various application scenarios, such as communication signal classification (Ma et al. 2023; Hao et al. 2023), medical detection (Malik et al. 2021; Li et al. 2020b), sound recognition (Zhang et al. 2021c; Tzinis et al. 2020), and so on. With the advancements in artificial intelligence technologies like deep learning, signal classification methods have become more precise and intelligent. Communication signal classification task is a time series classification problem that can be addressed using sequence models like RNN (Chang et al. 2021) and LSTM (Zhou et al. 2020b), which excel at handling temporal information. However, these models may overlook the spatial information of orthogonal signals and suffer from proportional growth of parameters with sequence length due to recurrent connections. Therefore, models that retain long sequence modeling characteristics while adept at processing spatial information are needed.

In order to better imitate human comprehension of complex scenes, attention mechanisms have been introduced to assist models in adaptively focusing on and processing different parts of signals, whether temporal or spatial. Significant progress has been made in attention mechanism modules in existing works, such as SENet (Hu, Shen, and Sun 2018), ECA (Wang et al. 2020), CBAM (Woo et al. 2018), SimAM (Yang et al. 2021), self-attention (Zhao, Jia, and Koltun 2020) , etc. Due to the provision of a universal way to enhance input representation, integrating attention modules into mainstream models (e.g., CNN (Li et al. 2020a), Transformer (Xia et al. 2022)) can lead to further improvements. In particular, for signal classification tasks, CNN models with limited global modeling capabilities require an auxiliary module to extend their view and improve their modeling capabilities for long sequences. Also, different type of attention can provide new representation subspaces for Transformer-based models, thereby enhancing the model's ability to capture complex patterns.

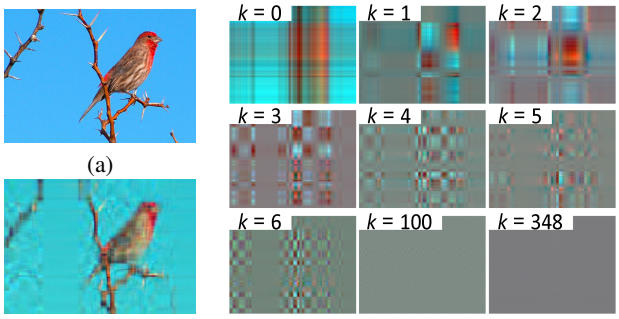


Figure 1: Singular value decomposition and image reconstruction. (a) The original image; (b) Reconstruction result retaining the first 10 singular values; (c) Reconstructed results retaining only the k-th singular value.

However, these attention mechanisms often enhance the focus expression of key features in an explicit manner, which may not be the optimal solution for signal classification. It is well known that communication signals contain a significant amount of noise, or redundancy. When features are explicitly enhanced, it is inevitable to strengthen this part. On the other hand, they ignore the internal structure of the data, especially the different semantic subspaces of the data itself. Signals can be viewed as a superposition of subspaces composed of different semantic information, and the information contained in the internal structure plays an important role in

*Corresponding authors.

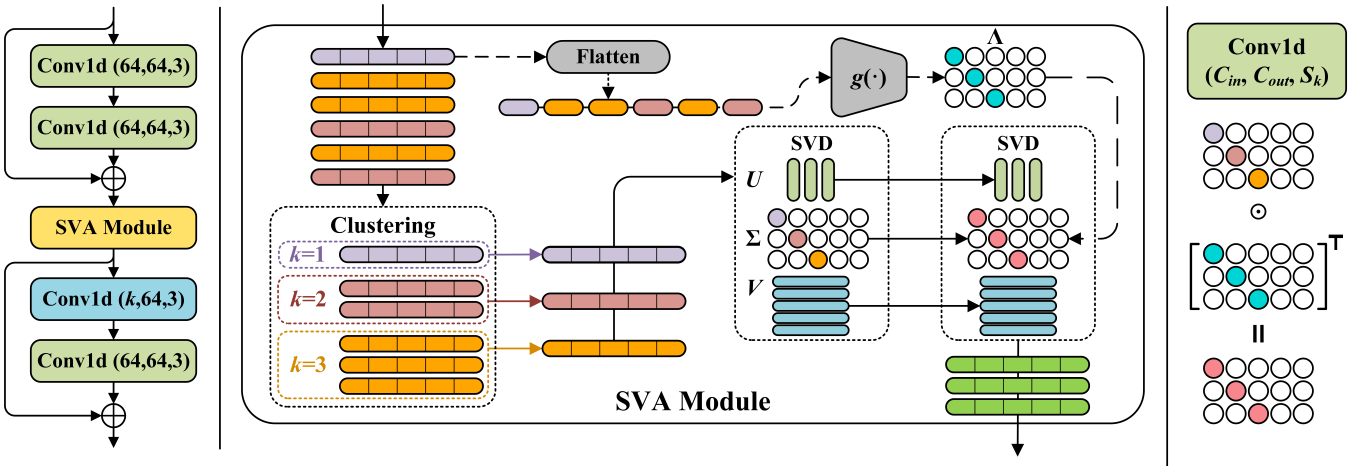


Figure 2: SVA module (middle), embedded network diagram (left) and some module operations (right).In the SVA module, the dashed line is the attention layer and the solid line is the decomposition layer.

signal classification. This motivates us to find an attention module that can model long sequences, attenuate noise, and enhance features based on its own internal structure.

Singular Value Decomposition (SVD) has been widely used in machine learning and signal processing. The factors obtained through the importance-ranked ordering of SVD are beneficial for analyzing the intrinsic structure of the original data. For instance, using an image for visual represetation, Figure 1 illustrates the semantic subspaces obtained through the decomposition of an image. As the importance decreases (corresponding to larger K-th singular values), the information contained in each subspace diminishes. Remarkably, similar images can be reconstructed using only a small portion of the information. Motivated by this, we propose an a SVD-based Attention, SVA. Specifically, different semantic subspaces are obtained through SVD, and the dependencies between different semantic subspaces are learned to conduct differentiated weighting across subspaces. Moreover, the computation of such attention across subspaces helps alleviate the misleading effects of random noise in signals. This internally structured attention mechanism tailored to communication signals facilitates more robust and standardized feature enhancement. Our contributions are summarized as follows:

- A new attention mechanism based on SVD, SVA, is advanced to enhance features by exploring the internal structure of data itself, i.e., the semantic subspace obtained through SVD.
- Due to the drawbacks of embedding SVD into deep learning models, we propose a training method to minimize the time-consuming of SVD as well as the possibility of gradient explosion during training.
- We validate SVA in communication signal classification tasks. By incorporating this module into a generic backbone network, its performance can be significantly improved, becoming competitive with or even surpassing specialized task-specific models. Moreover, results on an EEG signal recognition task demonstrate that SVA can be broadly applied to other time series classification tasks.

Related Work

Signal Classification

Signal classification has been a longstanding and seminal problem in communications. Early methods of signal classification were mainly based on the application of manually crafted feature extraction techniques, such as Likelihood Based (LB) and Feature Based (FB) methods (Wang et al. 2016; Liu et al. 2020). However, in light of the remarkable achievements of deep learning, a wave of deep learning models have found their application in the realm of signal classification. Intuaky, the strategy employed merely transformed signals into a visual format with the objective of leveraging computer vision-based models for classification (Wu et al. 2021; Li et al. 2022; Abdel-Moneim et al. 2022). Nevertheless, contemporary studies have shifted this focus onto the exploration of inherent characteristics of communication signals themselves (Fu et al. 2021; Zheng et al. 2023). For instance, Transformer models that capture long-range dependencies in sequences have demonstrated superior performance compared to CNN and RNN models on signal classification benchmarks (Kong et al. 2023; Jin and Chen 2021; Dong et al. 2022). However, these Transformer-based approaches primarily exploit global information and do not explicitly address the redundant information pervasive in communication signals. Further research into Transformer architectures capable of learning redundant local patterns may yield additional benefits.

Attention Mechanisms

The attention mechanism stems from intellectual endeavors to understand the human visual system, and in recent years, its adoption in deep learning landscapes has become increasingly prolific. Its core lies in the adaptive reweighting of features according to their importance. Initially, combining attention mechanisms with Recurrent Neural Networks (RNNs) was a common approach to predict salient regions

of the input (Kornblith, Shlens, and Le 2019; Shen et al. 2018). This was followed by the development of generating deformable grids to learn the transformation space of input, enabling spatial warping (Parmar et al. 2018; Sandler et al. 2018). The subsequent squeeze and excitation mechanism introduces channel-wise attention on the basis of spatial attention (Chen et al. 2018; Zhang et al. 2021b). Finally, there is the self-attention phase. Self-attention was first proposed for natural language processing (Rabinovich et al. 2007) and soon gained great success after its introduction into computer vision (Zhou et al. 2020a; Carion et al. 2020; Liu et al. 2021). The series of subsequent works focusing on improving the speed, output quality and generalization capability of self-attention based models have demonstrateds enormous potential of models based on attention mechanisms (Tan et al. 2019; Vaswani et al. 2017).

Method

In this section, we outline the proposed SVA module to capture the internal structure of long sequences. First, we introduce the specific approach of SVA. Then, to make it compatible with backpropagation-based optimization frameworks and reduce time complexity, we provide the implementation details of the SVA module. Finally, we give the testing procedure and discuss a special case of the SVA module.

SVA Module

The main idea of SVA is to uncover internal structure in the data, enhancing important patterns while suppressing irrelevant ones to improve the modeling power of the network. Specifically, given the output $X \in R^{n\times m}$ of an intermediate layer in a deep neural network, this feature serves as the input to the SVA module. The SVA module consists of two main components: the decomposition layer and the attention layer. The decomposition layer takes the intermediate feature X as input and performs SVD on it. The attention layer, which is dynamically generated based on the input feature X itself, produces the singular value attention $\Lambda \in R^{n\times m}$. Assuming $n \leq m$, $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then we have

$$X = U\Sigma V^T \tag{1}$$

$$\Lambda = g_\phi(X) \tag{2}$$

$$X' = U(\Lambda \circ \Sigma)V^T \tag{3}$$

where $U \in \mathbb{R}^{n\times n}$ is the left singular matrix of X, $UU^T = I$. $V \in \mathbb{R}^{m\times m}$ is the right singular matrix of X, $VV^T = I$. $\Sigma \in \mathbb{R}^{n\times m}$is the singular value matrix, and it is a diagonal matrix . $\circ$ denotes the Hadamard product. $X' \in \mathbb{R}^{n\times m}$ is the output of the SVA module. $g_\phi(\cdot)$ represents an attention mapping that dynamically generates singular value attention based on the input feature X. The attention layer can be constructed using a mapping, which defined as:

$$\boldsymbol{x} = vec(X) = (x_{11}, \ldots, x_{1m}, \ldots, x_{n1}, \ldots, x_{nm}) \tag{4}$$

$$\lambda = \sigma\left(f_3\left(ReLU\left(f_2\left(f_1\left(\boldsymbol{x}\right)\right)\right)\right)\right) \tag{5}$$

$$\Lambda = diag(\lambda) = diag(\lambda_1, \lambda_2, \ldots, \lambda_n) \tag{6}$$

where $vec(\cdot) : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^{nm}$ is a vectorization operator to flatten the input feature for mapping. f_1, f_2, f_3 are two linear layers and a convolution layer, and σ is the sigmoid function used to derive the strength of singular value attention $\lambda_i, i = 1, 2, \ldots, n$. $diag(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n\times m}$ is a diagonalization operator that turns the attention vector λ into a diagonal matrix Λ.

Figure 2 shows the architecture of the SVA module in detail. In principle, this module can be embedded into any backbone model and applied to different tasks. However, here we primarily integrate this module into the prevalent ResNet architecture based on residual modules and the popular Transformer architecture to uncover internal structure in the intermediate feature representations of different model architectures. Next, we describe the detailed setup in the SVA module.

Back Propagation and Training

When global matrix operations are used in deep networks, they are combined with layers of other processing performed along the way. These steps are architecture-specific and critical. SVD has a strong formulation that allows expressing complex transformations such as matrix functions and algorithms in a numerically stable form. Given a loss function $\mathcal{L}$, when constructing layers that perform global computations in a deep network using SVA, the backward pass through the matrix follows the chain rule (Ionescu, Vantzos, and Sminchisescu 2015):

$$\frac{\partial \mathcal{L}}{\partial X} = U \cdot \left\{ 2 \cdot \Sigma \left(K^T \circ \left(V^T \frac{\partial \mathcal{L}}{\partial V} \right) \right)_{sym} + \left(\frac{\partial \mathcal{L}}{\partial \Sigma} \right)_{diag} \right\} \cdot V^T + \left(\frac{\partial \mathcal{L} \cdot g}{\partial X} \right)_{diag} \tag{7}$$

where

$$K_{ij} = \begin{cases} \frac{1}{\sigma_i^2 - \sigma_j^2}, & i \neq j \\ 0, & i = j \end{cases}, \tag{8}$$

where $\mathcal{L} \cdot g$ denotes the loss function of the singular value attention layer in SVA, and $\frac{\partial \mathcal{L} \cdot g}{\partial X}$ is the partial derivative of loss with respect to the feature X. σ_i is the i-th singular value of the feature X, $i = 1, 2, \ldots, n$. $A_{sym} = \frac{1}{2}(A^T + A)$, and A_{diag} is the all non-diagonal elements of A set to zero. From the derivative expression of the SVA module, it can be seen that the gradient explosion occurs when the singular values σ_i,σ_j are 0 or very small due to the existence of K_{ij}. Therefore, we need to impose some constraints to the SVA module to prevent the feature X from generating a number of singular values tending towards 0.

In order to explore when SVD is prone to producing small singular values, we randomly generate 10,000 matrices of size (n, m) for singular value decomposition. The results are shown in Figure 3. We plot the matrix condition number and decomposition time, which demonstrate that as n increases, the matrix condition number also gradually increases, implying that more diminutive singular values will be generated when continuing SVD on larger matrices. Given an n-dimensional square matrix, the time complexity of its singular value decomposition is $\mathcal{O}(n^3)$. As illustrated in Figure 3, the larger the matrix the more decomposition time is

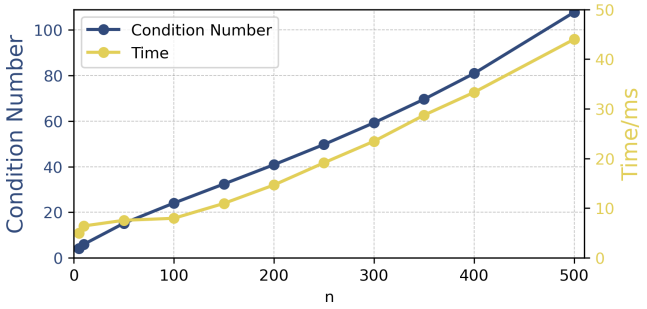


Figure 3: Condition number and decomposition time versus $(n, 1200)$ size matrix.

required. larger matrices require more decomposition time. Therefore, we need to reduce the dimension of the first dimension of the intermediate feature X.

It is well known that the feature maps have redundancy (Quanshi and Chun 2018), that is, the intermediate feature X contains similar features. This fact makes dimension reduction. We can combine similar feature maps, to obtain a feature Y with smaller dimensions. Therefore, we perform clustering on the feature maps with k cluster centers. The mean value of the features belonging to the same cluste ring center is taken to obtain $Y \in \mathbb{R}^{k \times m}$. For ease of training, we perform clustering every G epochs.

As shown in K_{ij}, we know the matrix will generate a gradient explosion if it produces lots number of singular values of 0 when it is subjected to SVD. Therefore, we need the rows of Y to be full rank, i.e., having k distinct singular values. In other words, we need YY^T to be full rank. To do this, we constrain the features Y by an orthogonal loss function:

$$\mathcal{L}_{\text{orth}} = \left\| YY^T \left(YY^T\right)^T - \mathbf{I}_c \right\|_2 \tag{9}$$

where $\mathbf{I}_c$ is a k-dimensional identity matrix. We utilize the L_2 norm to regularize our mode. This loss pushes YY^T towards an orthogonal matrix, which means the row vectors of Y tend to be orthogonal. We balance the regularization loss and the classification loss by $\beta \in (0, 1)$.

Testing Details

In testing, the features entered in the SVA are still subjected to dimensionality reduction. At this point, no clustering is performed, but based on the clustering results obtained in the training phase, the mean of the samples belonging to the same cluster center is taken. This feature is then entered into the SVA module, with the other settings unchanged.

Similarly, K-SVD is a data compression technique that procures the optimal sparse representation of data (Aharon, Elad, and Bruckstein 2006). This technique achieves this objective by employing an iterative optimization of both the dictionary and sparse coding. From this perspective, K-SVD could be interpreted as a specific instantiation of SVA. When attention weights tend toward zero in some feature subspaces, SVA essentially invokes a low-rank approximation of characteristics, paralleling K-SVD. In such circumstances, SVA selectively disregards a portion of the singular values, effectuating the decomposition of the feature matrix into two low-rank matrices, U and V. This strategy facilitates feature noise reduction by discarding redundancies concurrent with the directions of thrown out singular vectors.

Experiments

Experimental Setup

Datasets. Our experiment began with an evaluation of our models on three benchmark datasets: RML2016.10a (O'Shea, Corgan, and Clancy 2016), HisarMod2019 (Tekbıyık et al. 2020), and a large-scale real-world radio dataset (LSRWR) (Ya et al. 2022). The first two are used for signal modulation classification, with signal-to-noise ratios (SNR) ranging from -20dB to 18dB in 2dB increments, containing 11 and 26 modulations respectively. The last dataset consists of ADS-B radio signals from 100 different transmitters, used for Specific Emitter Identification (SEI). For all datasets, we split them into 70% for training, 10% for validation, and 20% for testing. We also applied EEG signals to verify the validity of the model on other time series data: sleep-edf-2002, sleep-edf-2013 (Kemp et al. 2000), and Human Activity Recognition (HAR) (Yilmaz 2017) datasets. More details about the datasets can be found in Appendix B.

Implementation Details. In experiments involving SVA, we set the number of cluster centers $k = 10$, and reclustered at intervals of 20 epochs, $G = 20$. The optimization algorithm used was Adam with a learning rate of 0.001 and batch size of 256. Our objective function was a weighted combination of the cross-entropy loss and orthogonality loss, with the coefficient for orthogonality loss set to $\lambda = 0.5$. Unless explicitly mentioned, the SVA module during training includes the SVA module, clustering component, and orthogonality loss. Each experiment was run on GeForce RTX 4090, and the reported results demonstrate the average performance over 10 independent runs.

Quantitative Evaluation. In our experiments, we focused on implementing the SVA in the 1D ResNet (He et al. 2016), as it is widely used as a backbone for classification, and ViT-Lite (Hassani et al. 2021), a transformer-based classification network known for its compatibility with smaller datasets. Comparative evaluations were conducted between the SVA module and prevalent attention mechanisms such as SENet , CBAM, ECA and SimAM. Concurrently, we benchmarked these models against several outstanding task-specific counterparts, including MCLDNN (Xu et al. 2020) and PETCGDNN (Zhang et al. 2021a) for modulation recognition, and AttnSleep (Eldele et al. 2021) and DeepSleepNet (Supratak et al. 2017) for EEG signal classification.

Overall Performance

Table 1 presents the signal classification performance of three backbones, each supplemented with different attention modules, including SENet, CBAM, ECA, SimAM, and SVA. The tasks in focus include modulation classification (RML2016.10a, HisarMod2019) and SEI (LSRWR). It is easily noticeable that SVA universally improves the performance of most models. In Table 1, the number of model parameters is listed in the first column for each dataset, and the

model	RML2016.10a			HisarMod2019			LSRWR		
	Params/M	MFlops	OA (%)	Params/M	MFlops	OA (%)	Params/M	MFlops	OA (%)
ResNet18	3.849	87.41	58.84	3.849	699.27	73.21	3.849	819.51	55.78
+ SENet	3.936	87.45	58.70	3.936	698.97	74.11	3.936	819.14	73.04
+ CBAM	3.851	87.43	58.90	3.851	699.35	75.08	3.851	819.60	67.00
+ ECA	3.849	87.42	59.46	3.849	699.34	75.29	3.849	819.58	73.24
+ SimAM	3.849	87.41	58.72	3.849	699.27	73.03	3.849	819.51	71.52
+ SVA (Ours)	**3.775**	**83.92**	**61.44**	**3.809**	**671.05**	**76.41**	**3.821**	**786.39**	**73.62**
ResNet34	7.224	178.90	59.40	7.224	1431.18	74.04	7.224	1677.21	66.04
+ SENet	7.382	179.01	59.72	7.382	1430.95	74.60	7.382	1676.91	74.26
+ CBAM	7.226	178.92	59.98	7.226	1431.26	76.35	7.226	1677.30	69.74
+ ECA	7.224	178.91	60.42	7.224	1431.25	75.46	7.224	1677.29	70.58
+ SimAM	7.224	178.90	58.43	7.224	1431.18	75.95	7.224	1677.21	74.68
+ SVA (Ours)	**7.149**	**175.67**	**61.75**	**7.183**	**1405.05**	**84.59**	**7.195**	**1646.54**	**74.93**
ViT-Lite	4.600	532.58	**61.38**	4.600	306.22	76.52	4.600	360.45	76.27
+ SVA (Ours)	**4.344**	**54.46**	61.31	**4.344**	**144.60**	**76.95**	**4.344**	**162.31**	**76.67**

Table 1: Signal classification performance of three common backbones equipped with attention modules. The indicators include the number of model parameters, the amount of computation and the overall accuracy (OA). The best indicators are bolded.

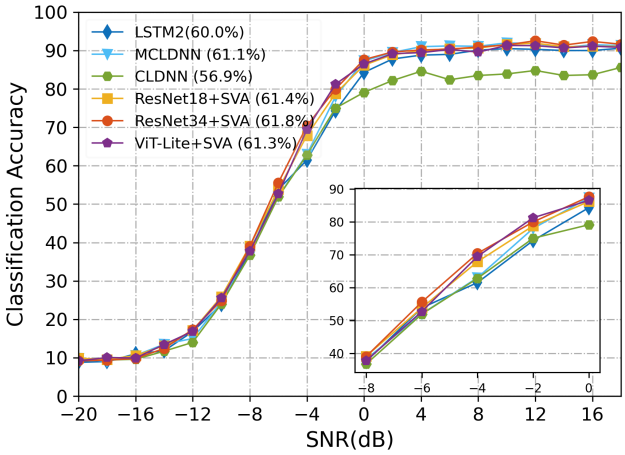


Figure 4: RML2016.10a dataset contrast curves between the general purpose model with SVA and the advanced task-specific model at different SNR.

computational cost in FLOPS during testing is in the second. Furthermore, the models with SVA modules not only reduce parameters and computational cost during testing while achieving superior performance but also demonstrate scaling of parameter counts with data dimensionality, attributable to the attention layers in SVA. Notably, 1D ResNet generally underperforms in signal classification. However, the performance can be improved by integrating attention modules, with SVA conferring distinct enhancements.

Figure 4 presents the comparative curves of generic models equipped with SVA versus task-specific models under different SNR on the RML2016.10a dataset. As evidenced in the figure, models containing SVA demonstrate substantial improvements from -8dB to 0dB, indicating the efficacy of SVA in ameliorating signal degradation. Comparisons of generic models ResNet18, ResNet34, and ViT-Lite equipped with SVA modules against advanced models for signal classification and SEI are appended in Appendix C. The addition of SVA modules enables generic backbones to achieve on-par performance with task-specific models, especially for modulation classification where task-specific models exhibit 0.2%-22.26% enhancement. Compared to task-specific models, the models' SEI performance improves by 3.77%-9.15%. Experimental results of generic models equipped with SVA on EEG data are appended in Appendix D.

Parameters and Model Analysis

Location of SVA. We commenced our examination by benchmarking our model utilizing three standard datasets. As illustrated in Table 2, we took ResNet as a reference point and segregated it into five layers following the model structure, namely $conv1$, $layer1$, $layer2$, $layer3$, $layer4$, and $classification$. To analyze the performance impact of the location of the SVA module at varying positions, we embedded an SVA module subsequent to each of the initial five layers. We performed these experiments on an assortment of models, including ResNet18, ResNet34, and ResNet50. The empirical results indicated that the provision of an SVA

module beyond $layer2$ served to significantly bolster the model's expressive capacity. In contrast to other ResNet versions, the ResNet models that incorporated an SVA module after $layer2$ demonstrated a performance uplift spanning between 1.17% and 2.67%.

Location	- -	SVA Module				
		$conv1$	$layer1$	$layer2$	$layer3$	$layer4$
ResNet18	58.84	60.25	60.26	**61.49**	60.07	58.73
ResNet34	59.40	60.37	60.30	**61.71**	59.97	59.80
ResNet50	59.26	59.33	59.71	**60.41**	60.37	59.78

Table 2: The effect of adding SVA after different layers. The indicator is OA(%), and "- -" is the base value of the model.

Number of Cluster Centers. We embarked on our assessment using three benchmark datasets. Throughout the training phase of our model, we had to determine the number of hyperparameter clustering centers. Notably, the experimentally set value of 10 for this parameter isn't always optimal. Table 3 highlights the variation in model accuracy with 2:2:18 as the clustering center configuration. It can be observed that the accuracy follows an initial increment trend before starting to diminish, as the count of clustering centers rises. A pinnacle in accuracy is reached when we designated the quantity of clustering centers as 6.

k	2	4	6	8	10
OA(%)	61.88	61.85	**62.02**	61.94	61.49
k	12	14	16	18	20
OA (%)	61.56	60.29	60.36	59.97	59.93

Table 3: The effect of the number of clustering centers.

SVA	Clustering	$\mathcal{L}_{orth}$	OA(%)	Training time/μs	Testing time/μs
			58.83	112	71
✓			58.84	1897	1776
	✓		59.51	**113**	**74**
✓		✓	58.86	2009	1792
✓	✓		61.11	509	461
✓	✓	✓	**61.49**	565	506

Table 4: Ablation experiments of each component.

Ablation Experiments on Various Components. Initially, we scrutinized our model employing three standard datasets. During the SVA training process, we incorporated a clustering component and an orthogonal loss term ($\mathcal{L}_{orth}$). Table 4 illustrates the ablation studies relating to these three components, in conjunction with the time necessitated for training/testing a singular sample. In comparison to the pristine network, the addition of all components indicated an uplift in performance, pronouncedly the inclusion of the SVA and the clustering element. The incorporation of the clustering component served a dual role: augmenting performance and diminishing the time required for training/testing samples. Upon adding a clustering component to the SVA, the training/testing time for samples witnessed a reduction of 73.2%. Concurrently, as per a model equipped with all three components, a performance boost of 2.6% was observed against the result excluding the clustering module alone, resulting in a time reduction of 71.8% for training/testing samples. The orthogonal loss component offered a slight increase in model performance, suggesting the indispensability of such a clustering module in enhancing the model.

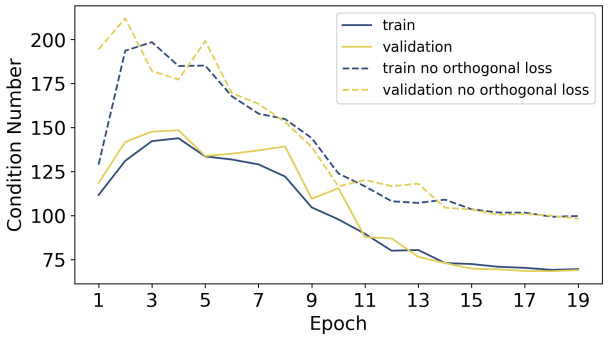
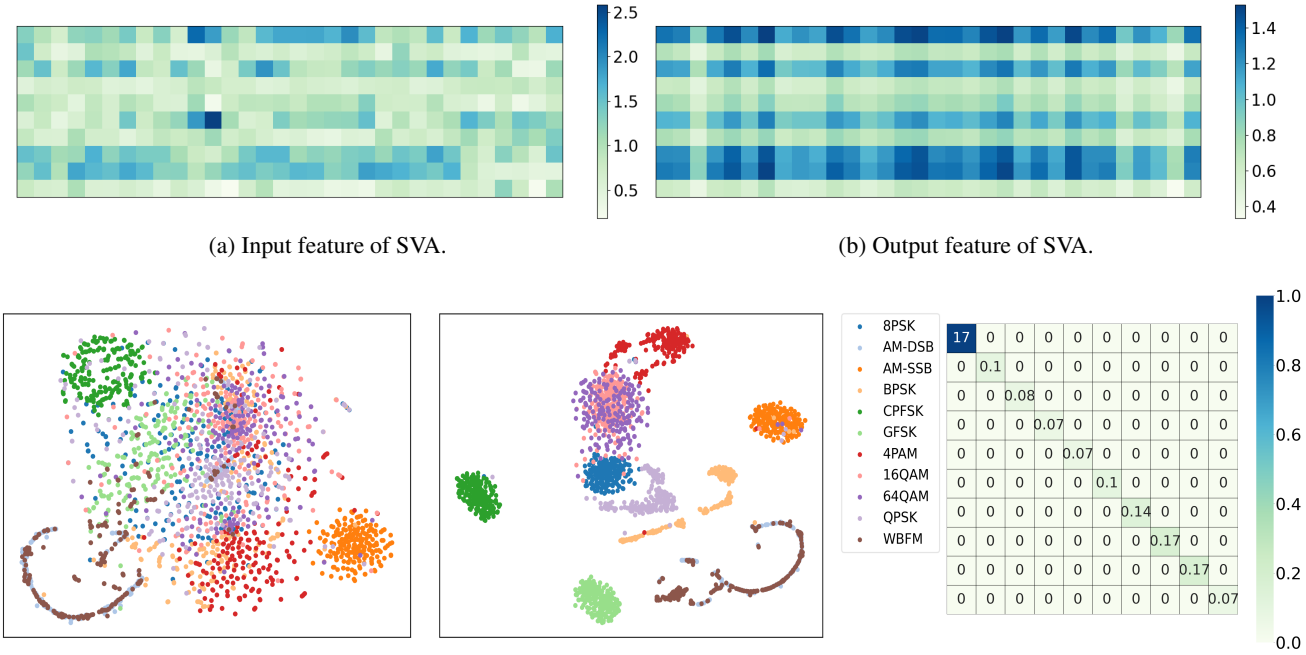


Figure 5: The number of conditions of the matrix for SVD varies with training. The figure only shows the matrix containing 10 clustering centers, and the original model cannot be visually displayed because of the large number of matrix conditions added to the SVD layer. The dashed line contains SVA module and clustering module, and the solid line contains SVA, clustering and orthogonal loss $\mathcal{L}_{orth}$.

The ill-conditioned nature of SVD input attributes can potentially detriment the model's training process. Thus, higher condition numbers serve as indicators of an escalating matrix-ill-condition state. Figure 5 delineates the fluctuation curve of a matrix condition number undergoing SVD during training. With the progression of training epochs, the condition number of the SVA, along with its respective constituents, steadily diminishes. Specifically, the introduction of an orthogonal loss component yields an additional decrement in the condition number, whereas the condition number tied exclusively to the SVD within the same layer exhibits a propensity to overflow. Such findings indicate the indispensable role of the orthogonal loss component for ensuring a seamless experimental process. Figure 7 depicts statistical findings of failed training attempts (marked by a gradient explosion or matrix ill-conditioning) during the training phase of ResNet18 with incorporated SVD and SVA. Here, the stars signify the experiments involving the SVA and its components, while circles denote the SVD-included experiments. Observably, the introduction of the SVA and its components diminishes the failure rate considerably, with fewer



(a) Input feature of SVA. (b) Output feature of SVA.

(c) t-SNE visualization of SVA input feature. (d) t-SNE visualization of SVA output feature. (e) Singular value attention weight.

Figure 6: Visual analysis of features and weights.

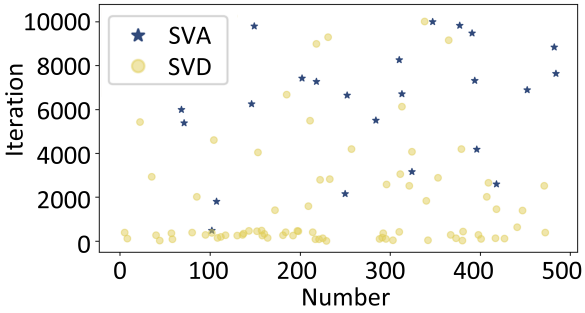


Figure 7: When SVD and SVA layers are added to ResNet18, the experimental statistical results of gradient explosion or ill-conditioned input matrix appear.

errors materializing during the first epoch. Contrariwise, for experiments merely executing SVD, a marked rise in failures ensues, prone to flopping in the initiating iterations. These observations reinforce the significance of the adept design of clustering and orthogonal loss modules.

Visual Analysis of Weights and Features. Figure 6 offers a visualization of output/input attributes and attention weights pertaining to SVA. Figure 6a and 6b depict the visualization of features that are inputted and outputted in SVA, with the x-axis symbolizing feature length dimension and y-axis corresponding to the feature channel dimension. Conversely, Figures 6c and 6d are t-SNE (van der Maaten and Hinton 2008) diagrams of the input and output features of SVA, representing their respective distribution profiles. Figure 6e furthermore provides a visualization of singular value attention weights, where the segments with a weight of zero are omitted. As discernable from Figure 6a and 6b, post-processing by SVA results in features concentrating more prominently on distinct channels, thus endorsing a more streamlined and efficient feature representation. Similarly, Figures 6c and 6d showcase that features outputted by SVA exhibit higher divergence across distinct categories, intrinsically closer for the same category, and a denser distribution. This points to the adaptive feature distribution modification accomplished by SVA, thereby significantly bolstering feature representation of the neural network. In Figure 6d, all the substantial values in the feature matrix are amassed at the upper-left corner, not arranged in descending order of magnitude. This implies that during classification execution, the model's feature attention span synchronized with the crucial feature yield from SVD, regardless of the sequencing discordance, thereby affirming the necessity of attention in SVA.

Conclusion

To address the issue of signal degradation in real-world electromagnetic environments, this paper proposes a novel attention module that excavates the structural information within signal data based on singular value decomposition. To validate its efficacy, we incorporated it into three simple generic backbone networks to achieve better performance. Experiments on multiple datasets demonstrate that adding this module helps improve backbone models and achieves competitive results to task-specific models. We hope the design of this attention module will facilitate future research on interpretable features for time series data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, Nos.U22B2018, 62276205, and 2023 Innovation Fund of Xidian University, No. YJSJ23001.

References

Abdel-Moneim, M. A.; Al-Makhlasawy, R. M.; Abdel-Salam Bauomy, N.; El-Rabaie, E.-S. M.; El-Shafai, W.; Farghal, A. E.; and Abd El-Samie, F. E. 2022. An efficient modulation classification method using signal constellation diagrams with convolutional neural networks, Gabor filtering, and thresholding. *Transactions on Emerging Telecommunications Technologies*, 33(5): e4459.

Aharon, M.; Elad, M.; and Bruckstein, A. 2006. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11): 4311–4322.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chang, S.; Huang, S.; Zhang, R.; Feng, Z.; and Liu, L. 2021. Multitask-learning-based deep neural network for automatic modulation classification. *IEEE internet of things journal*, 9(3): 2192–2206.

Chen, L.-C.; Collins, M.; Zhu, Y.; Papandreou, G.; Zoph, B.; Schroff, F.; Adam, H.; and Shlens, J. 2018. Searching for efficient multi-scale architectures for dense image prediction. *Advances in neural information processing systems*, 31.

Dong, B.; Liu, Y.; Gui, G.; Fu, X.; Dong, H.; Adebisi, B.; Gacanin, H.; and Sari, H. 2022. A Lightweight Decentralized-Learning-Based Automatic Modulation Classification Method for Resource-Constrained Edge Devices. *IEEE Internet of Things Journal*, 9(24): 24708–24720.

Eldele, E.; Chen, Z.; Liu, C.; Wu, M.; Kwoh, C.-K.; Li, X.; and Guan, C. 2021. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29: 809–818.

Fu, X.; Gui, G.; Wang, Y.; Ohtsuki, T.; Adebisi, B.; Gacanin, H.; and Adachi, F. 2021. Lightweight automatic modulation classification based on decentralized learning. *IEEE Transactions on Cognitive Communications and Networking*, 8(1): 57–70.

Hao, X.; Feng, Z.; Yang, S.; Wang, M.; and Jiao, L. 2023. Automatic Modulation Classification via Meta-Learning. *IEEE Internet of Things Journal*.

Hassani, A.; Walton, S.; Shah, N.; Abuduweili, A.; Li, J.; and Shi, H. 2021. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Ionescu, C.; Vantzos, O.; and Sminchisescu, C. 2015. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE international conference on computer vision*, 2965–2973.

Jin, C.-c.; and Chen, X. 2021. An end-to-end framework combining time–frequency expert knowledge and modified transformer networks for vibration signal classification. *Expert Systems with Applications*, 171: 114570.

Kemp, B.; Zwinderman, A. H.; Tuk, B.; Kamphuisen, H. A.; and Oberye, J. J. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9): 1185–1194.

Kong, W.; Jiao, X.; Xu, Y.; Zhang, B.; and Yang, Q. 2023. A Transformer-based Contrastive Semi-Supervised Learning Framework for Automatic Modulation Recognition. *IEEE Transactions on Cognitive Communications and Networking*.

Kornblith, S.; Shlens, J.; and Le, Q. V. 2019. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2661–2671.

Li, J.; Jin, K.; Zhou, D.; Kubota, N.; and Ju, Z. 2020a. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing*, 411: 340–350.

Li, L.; Dong, Z.; Zhu, Z.; and Jiang, Q. 2022. Deep-learning hopping capture model for automatic modulation classification of wireless communication signals. *IEEE Transactions on Aerospace and Electronic Systems*, 59(2): 772–783.

Li, R.; Johansen, J. S.; Ahmed, H.; Ilyevsky, T. V.; Wilbur, R. B.; Bharadwaj, H. M.; and Siskind, J. M. 2020b. The perils and pitfalls of block design for EEG classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 316–333.

Liu, M.; Qu, N.; Tang, J.; Chen, Y.; Song, H.; and Gong, F. 2020. Signal estimation in cognitive satellite networks for satellite-based industrial internet of things. *IEEE Transactions on Industrial Informatics*, 17(3): 2062–2071.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, J.; Hu, M.; Wang, T.; Yang, Z.; Wan, L.; and Qiu, T. 2023. Automatic Modulation Classification in Impulsive Noise: Hyperbolic-Tangent Cyclic Spectrum and Multibranch Attention Shuffle Network. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–13.

Malik, J.; Devecioglu, O. C.; Kiranyaz, S.; Ince, T.; and Gabbouj, M. 2021. Real-time patient-specific ECG classification by 1D self-operational neural networks. *IEEE Transactions on Biomedical Engineering*, 69(5): 1788–1801.

O'Shea, T. J.; Corgan, J.; and Clancy, T. C. 2016. Convolutional radio modulation recognition networks. In *Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17*, 213–226. Springer.

Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; and Tran, D. 2018. Image transformer. In *International conference on machine learning*, 4055–4064. PMLR.

Quanshi, Z.; and Chun, Z. S. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1): 27–39.

Rabinovich, A.; Vedaldi, A.; Galleguillos, C.; Wiewiora, E.; and Belongie, S. 2007. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, 1–8. IEEE.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Supratak, A.; Dong, H.; Wu, C.; and Guo, Y. 2017. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11): 1998–2008.

Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; and Le, Q. V. 2019. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2820–2828.

Tekbıyık, K.; Ekti, A. R.; Görçin, A.; Kurt, G. K.; and Keçeci, C. 2020. Robust and fast automatic modulation classification with CNN under multipath fading channels. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 1–6. IEEE.

Tzinis, E.; Wisdom, S.; Hershey, J. R.; Jansen, A.; and Ellis, D. P. 2020. Improving universal sound separation using sound classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 96–100. IEEE.

van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. journal of Machine Learning Research 9. *Nov (2008)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, F.; Dobre, O. A.; Chan, C.; and Zhang, J. 2016. Fold-based Kolmogorov–Smirnov modulation classifier. *IEEE Signal Processing Letters*, 23(7): 1003–1007.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, X.; Zhang, J.; Hou, C.; Liu, G.; Zhang, J.; and Liu, J. 2021. Signal modulation recognition based on convolutional autoencoder and time-frequency analysis. In *2021 8th International Conference on Dependable Systems and Their Applications (DSA)*, 664–668. IEEE.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4794–4803.

Xu, J.; Luo, C.; Parr, G.; and Luo, Y. 2020. A spatiotemporal multi-channel learning framework for automatic modulation recognition. *IEEE Wireless Communications Letters*, 9(10): 1629–1632.

Ya, T.; Yun, L.; Haoran, Z.; Zhang, J.; Yu, W.; Guan, G.; and Shiwen, M. 2022. Large-scale real-world radio signal recognition with deep learning. *Chinese Journal of Aeronautics*, 35(9): 35–48.

Yang, L.; Zhang, R.-Y.; Li, L.; and Xie, X. 2021. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*, 11863–11874. PMLR.

Yilmaz, Y. 2017. Online nonparametric anomaly detection based on geometric entropy minimization. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 3010–3014. IEEE.

Zhang, F.; Luo, C.; Xu, J.; and Luo, Y. 2021a. An efficient deep learning model for automatic modulation recognition based on parameter estimation and transformation. *IEEE Communications Letters*, 25(10): 3287–3290.

Zhang, Y.; Li, K.; Li, K.; and Fu, Y. 2021b. MR image super-resolution with squeeze and excitation reasoning attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13425–13434.

Zhang, Z.; Xu, S.; Zhang, S.; Qiao, T.; and Cao, S. 2021c. Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing*, 453: 896–903.

Zhao, H.; Jia, J.; and Koltun, V. 2020. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10076–10085.

Zheng, S.; Zhou, X.; Zhang, L.; Qi, P.; Qiu, K.; Zhu, J.; and Yang, X. 2023. Towards Next-Generation Signal Intelligence: A Hybrid Knowledge and Data-Driven Deep Learning Framework for Radio Signal Classification. *IEEE Transactions on Cognitive Communications and Networking*.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020a. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13041–13049.

Zhou, Q.; Jing, X.; He, Y.; Cui, Y.; Kadoch, M.; and Cheriet, M. 2020b. LSTM-based automatic modulation classification. In *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 1–4. IEEE.