Self-attention Multi-view Representation Learning with Diversity-promoting Complementarity

Jian- ei Liu[1], Xi-hao Ding[1], Run-kun Lu[1], Xionglin LUO[1]
1. Department of automation, School of Information Science and Engineering, China Universit of Petroleum, Beijing, 102249
E-mail: liuj ,luo l@cup.edu.cn, 964465871@qq.com, zs lrk@gmail.com

Abstract: Multi-vie learning attempts to generate a model ith a better performance b e ploiting the consensus and/or complementarit among multi-vie data. Ho ever, in terms of complementarit , most e isting approaches onl can nd representations ith single complementarit rather than complementar information ith diversit . In this paper, to utilize both complementarit and consistenc simultaneousl , give free rein to the potential of deep learning in grasping diversit -promoting complementarit for multi-vie representation learning, e propose a novel supervised multi-vie representation learning algorithm, called Self-Attention Multi-Vie net ork ith Diversit -Promoting Complementarit (SAMVDPC), hich e ploits the consistenc b a group of encoders, uses self-attention to nd complementar information entailing diversit . E tensive e periments conducted on eight real- orld datasets have demonstrated the effectiveness of our proposed method, and sho its superiorit over several baseline methods, hich onl consider single complementar information.

Key Words: Multi-vie Learning, Self-attention Mechanism, Complementar Information ith Diversit

1 INTRODUCTION

Aiming to make good use of the information from multi-vie data and improve the generalization performance, multi-vie learning algorithms have made great progress in different tasks, such as classi cation, regression, and clustering, b utilizing conventional machine learning or deep learning to full considering the relationships among multiple vie s [1, 2, 3, 4]. And recentl , [5] anal zes these various algorithms, comes to the conclusion that there are t o fundamental assumptions ensuring their success: consistenc and complementarit principles. The consistenc assumption suggests there is consistent information shared b all vie s, hile the complementarit assumption states each vie of multi-data ma contain some kno ledge that other vie s do not have. Based on these t o assumptions, e revie the literature of multi-vie learning in recent ears, and observe that there are still t o dra backs in man state-of-the-art multi-vie learning algorithms.

First, at present, multi-vie algorithms can be generall categorized into t o t pes: the rst categor aims to e ploit the consistenc , the second one aims to leverage the complementarit among multiple vie s, and each categor onl focuses on consensus or complementarit . In detail, the rst categor usuall tries to e tract the common latent representation on hich all vie s have minimum disagreement, such as canonical correlation anal sis (CCA) class algorithms [6, 7, 8, 9, 10], hich project t o or more vie s into latent subspaces b ma imizing the correlations among projected vie s, matri factorization based methods [11, 12, 13], hich jointl factorize multi-vie data into one common centroid representation b minimizing the overall reconstruction loss of different vie s. And the second categor is to e plicitl preserve complementar information of different vie s, such as co-training st le algorithms [14, 15, 16, 17], hich iterativel train t o classi ers on t o different vie s, and each classi er generates its complementar information to help the other classi er to train in the ne t iteration.

Ho ever, both consistenc and complementarit of multi-vie s data are meaningful, the neglect of each aspect ill result in the loss of valuable information. In order to address this dra back, multi-vie algorithms recentl began to develop the third categor algorithm, hich e ploits the consistenc and complementarit , simultaneousl , such as matri factorization based methods [18, 19], hich nd latent representations composed of common latent factors shared b multiple vie s and the speci c latent factor of each vie . But [18, 19] also inherit the shortcomings of matri factorization, such as the onl learn a linear map relationships, can't re ect the non-linear relationship in the multi-vie dataset, and require feed all data in one time, lack the abilit of dealing ith large scale data.

Second, in terms of complementarit , most of e isting multi-vie learning algorithms onl can nd representations ith single complementarit rather than complementar information ith diversit . Srivastava and Salakhutdinov [20] propose a deep multi-modal RBM to capture the joint distrib..imag
et al. [21] concatenate the nal hidden coding of audio and video modalities as input, then map these inputs to a shared representation la er. Cho et al. [22] directl input multi-vie sequence into RNN encoder to integrate the complementarit of multi-vie data. Su et al. [23] introduce a multi-vie CNN architecture that integrates complementarit among multiple 2D vie s of an object into a

single and compact representation b a vie -pooling la er, hich performs element- ise ma imum operation across the vie s. In general, all the above algorithms focus on one t pe of complementarit among multiple vie s, and the don't consider mining the complementar information ith diversit .

To ght against the above mentioned serious de ciencies, in this paper, e propose a ne multi-vie learning paradigm based on self-attention net ork, called Self-Attention Multi-Vie net ork ith Diversit -Promoting Complementarit (SAMVDPC). Speciall , SAMVDPC rst encodes each vie 's data into a ed-length vector representation to e ploit the consistenc , and then e plores complementar information entailing diversit ith multiple combination forms b self-attention mechanism, nall concatenates all complementar information into a vector representation, hich further be used to make prediction. e ma illustrate this idea using an e ample from a face recognition problem ith t o vie s. iven a group of people, e have collected the face information for each person to form t o-vie dataset. To make classi cation, rst, b building a unique encoder for each vie , SAMVDPC encodes each vie 's data into a ed-length vector representation and outputs $\mathbf{H} = [h_1; h_2] \in R^{2\times H}$. Second, SAMVDPC inputs $\mathbf{H}$ to self-attention mechanism to produce eight matri $\mathbf{W} = [w_1; w_2] \in {}^{2\times 2}$, then outputs t o vectors: $w_1\mathbf{H}$, and $w_2\mathbf{H}$, hich can utilize to combine t o-vie data b different a s for subsequent fusion stage. Finall , SAMVDPC incorporates the concatenated representation $[w_1\mathbf{H},\ w_2\mathbf{H}]$ as input and processes these inputs b a for ard net ork to make prediction.

In summar , our contributions e summarize are sho n as follo s:

(1) e develop a supervised multi-vie deep learning algorithm, hich utilizes both consistenc and complementarit of multiple vie s, here multiple vie s' encoders consider the consistenc , and self-attention mechanism considers the complementarit

(2) Compared to [18, 19], encoders in SAMVDPC can learn nonlinear and hierarchical abstract feature representation for multi-vie data, hich capture the non-linear relationship and real underl ing properties in multi-vie dataset.

(3) SAMVDPC can nd representations ith complementar information possessing diversit rather than single complementarit , and suf cientl re ect the complementarit underl ing multi-vie data.

(4) e have compared SAMVDPC ith other state-of-the-art multi-vie learning algorithms and demonstrated its effectiveness, hat's more, e also build other baselines deep net orks to further anal ze SAMVDPC's performance, hich e plore single complementar b mean-pooling, ma -pooling and eighted summation.

Attention mechanism

In deep neural net orks, attention mechanism [24] has been developed in the conte t of encoder-decoder architectures for Neural Machine Translation (NMT) [22, 25], and rapidl applied to numerous application domains and achieved promising results on several challenging tasks, such as image captioning [26], and summarization [27]. Besides, ith the development of deep learning(de)2re. of attention mechanism,self-attention mechanism, hich

sutionsingle sequecea to conpuea conte n-

self(representation)5-1749(of)5138(the)5138(sequeceg(de859(and)5138(has)

comingenstion, abstracieve summarization, and learning tas-ianepeundent representatio s,and

arc(ers)-287(deeeon)-287(thirn)-286knholleage of attention mecn-

foreboder-decoder the

2and attention mechanism sa follo s:*Attention*

Fig. 1(a), the architecture of SAMVDPC is made up of Encoder-Block, self-attention mapping, and full con-nected la er, and the detailed self-attention mapping pro-cesses are sho n in Fig. 1(b). e describe each of the constituents in the follo ing subsections.

ncoder- loc : As sho n in Fig. 1(a), Encoder-Block is composed ith V same encoders to e tract each vie 's feature. The initial model parameters of each encoder are initialized b the encoder of a corresponding auto-encoder, hich ill be e plained more detailed in sec-tion 4.4. From these encoders, V hidden features ($\mathbf{z}^v \in R^{H\times 1}, v = 1, \cdots, V$) can be obtained and the ill be stacked horizontall and combined into a feature matri : $\mathbf{Z} = \left[\mathbf{z}^1, \cdots \mathbf{z}^V\right], \ \mathbf{z}^v \in R^{H\times 1}$, here H is the number of dimension of hidden feature vector $\mathbf{z}^v$.

Self-attention mapping: The self-attention mechanism takes the hole hidden states matri $\mathbf{Z}$ as input, outputs a matri $\mathbf{A}$, and each ro of $\mathbf{A}$ is a vector of eights $\mathbf{a}_i$:

$$\mathbf{A} = \left[\mathbf{a}_1 \vdots \mathbf{a}_{d\_c}\right] = softmax\left(\mathbf{W}_{s2} \tanh\left(\mathbf{W}_{s1}\mathbf{Z}^T\right)\right), \tag{2}$$

here, $\mathbf{A} \in R^{d\_c\times V}$, $\mathbf{W}_{s1} \in R^{d\_s\times H}$, $\mathbf{W}_{s2} \in R^{d\_c\times d\_s}$, d_s is a h per parameter e can set arbitraril , and the soft-ma () is operated along the second dimension of its input. Inspired b [30], Equation (2) also can be deemed as a 2-la er MLP ithout bias, hose hidden unit numbers are d_s, and parameters are $\{\mathbf{W}_{s2}, \mathbf{W}_{s1}\}$. Finall , e com-pute the d_s eighted sums b multipl ing $\mathbf{A}$ and $\mathbf{Z}$:

$$\mathbf{M} = \mathbf{A}\mathbf{Z}^T, \ \mathbf{M} \in R^{d\_c\times H}. \tag{3}$$

It is orth noting that each ro of $\mathbf{M}$ is a unique nonlinear combinations of multiple vie s data, and the self-attention mechanism outputs formulate d_s kinds of nonlinear com-bination of multiple vie s data. In our e periment part, the value of d_s is set to V.

ML : e concatenate each ro of to produce a multi-vie representation containing multiple combinations of multi-vie data to e tract complementar information entailing diversit . Then e input this representation to 2-la er MLP, and make prediction.

. Ob ective unction and Regulari ation

The embedding matri $\mathbf{M}$ al a s suffer from redundanc problems because the self-attention mechanism often pro-vides similar summation eights for all the d_s hops. In-spired b [ ], e also add regularization to encourage the diversit of summation eights vectors across different hops of attention. Thus, in this paper, our objective func-tion is consist of cross entrop loss and regularization, and can be formulated as follo s:

$$L = cross\_entropy(y, \hat{y}) + \lambda \left\|\mathbf{A}\mathbf{A}^T - \mathbf{I}\right\|_F^2, \tag{4}$$

here λ is regularization parameters, and $\mathbf{I}$ is a unit diago-nal matri .

periment

In this section, e e perimentall evaluate SAMVDPC in classi cation task on eight real orld multi-vie data sets b comparing it to other baseline algorithms, and design a set of e plorator e periments to validate properties of the self-attention mechanism in SAMVDPC, nall anal- se the convergence of our proposed algorithm.

Table 1: Characteristics of the datasets

Data Set	Instances numbers		K	Dimension numbers
Leaves	96	3	6	64 for all
Reuters	1200	5	6	2000 for all
aleFace	256	2	8	2016 for all
BBC	685	4	5	4659/4633/4665/4684
Cornell	195	2	5	1703/585
Te as	187	2	5	1703/561
ashington	230	2	5	1703/690
isconsin	265	2	8	1703/795

.1 Datasets

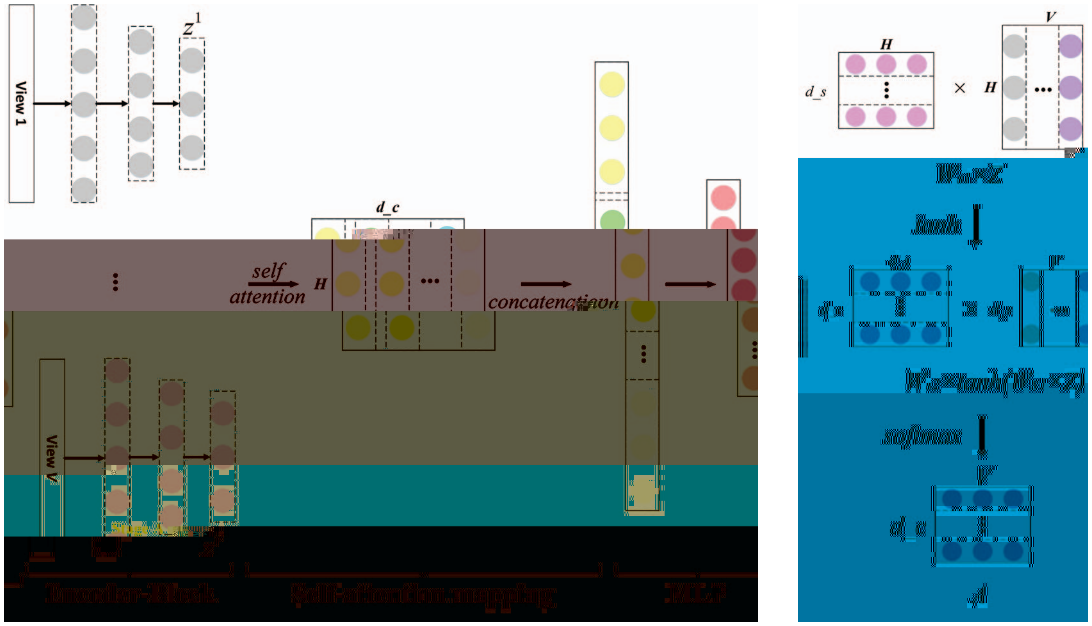
In this paper, e use eight real- orld multi-vie data sets to verif the performance of SAMVDPC, including Leaves, Reuters, aleFace, BBC, Cornell, Te as, ashing-ton, and isconsin datasets. Leaves and aleFace are t o image dataset, Reuters and BBC are t o te t dataset, Cor-nell, Te as, ashington, and isconsin dataset are four subset of data sets selected from eb B data sets, and e-b B are ebpage dataset. The properties of data sets are summarized in Table 1.

. Comparison Algorithms and aseline Models

e evaluate the SAMVDPC performance in classi cation tasks b comparing it ith several state-of-the-art multi-vie learning algorithms based on matri factorization, such as MVNMF [32], multiNMF [12], MVCC [13], DICS [18], and some our designed deep neural net ork baseline models ith three sorts of fusion strategies re-placing ith our self-attention mechanism, including ma -pooling model, mean-pooling model, and eighted sum-mation model. For fair comparison, in terms of matri fac-torization algorithms, e choose the parameters ithin the range that author suggested to obtain good latent represen-tations, and input these representations to NN($k = 1$) for classi cation in terms of deep neural net ork base-line models, e instead of the self-attention mechanism ith ma -pooling, mean-pooling, or eighted summation fusion, maintain the remaining structure unchanged, and remove the regularization in our objective function.

MVNMF is an NMF-based algorithm b merging local geometrical structure information of each vie in a multi-vie feature e traction frame ork. The e tracted feature considered the inner-vie relatedness bet een data, and further can be used to complete various tasks. e select parameters λ_f, μ to 0.01, and 10 as author suggested, re-spectivel .

MultiNMF is an NMF-based multi-vie algorithm, in terms of matri factorization, it requires coef cient ma-trices learnt from different vie s to be softl regularized to ards a common consensus matri , hich re ect the in-formation of multi-vie data and can be used to make clas-



(a) (b)

Figure 1: MVCapsNet Architecture. Fig. 1: (a) is the architecture of SAMVDPC. (b) is concrete self-attention mapping implementation processes.

Table 2: H perparameters on each data set

Data Set	Units number of each encoder layer			Diversity of complementarity	Si e of mini-batch
	l_1	l_2	l_3	d_c	
Leaves	64	32	16	3	4
Reuters	2048	1024	512	2	16
aleFace	1024	512	512	2	32
BBC	1024	512	512	4	32
Cornell	1024	512	128	2	16
Te as	1024	512	128	2	16
ashington	1024	512	128	2	16
isconsin	1024	512	128	2	16

Table 3: Accurac of different methods

Method	ACC(%)							
	Leaves	**ale ace**	**Reuters**	**C**	**Cornell**	**Te as**	**Washington**	**Wisconsin**
NM	95.0±0	50.0±2.5	40.8±1.2	38.0±1.5	41.0±1.8	57.9±1.8	69.6±2.2	52.8±1.4
MultiNM	95.0±0	64.2±4.2	52.7±0.2	73.1±0.2	49.7±7.7	68.7±3.4	59.3±2.6	50.3±3.5
M CC	100±	33.3±6.9	54.4±1.9	[illegible]	60.8±5.0	64.7±5.5	62.8±3.8	64.3±2.7
DICS	97.9±2.5	89.1±3.2	70.3±4.0	90.2±2.4	[illegible]	81.6±4.0	[illegible]	85.1±4.5
MA - ooling	100±0	90.0±4.6	71.2±3.4	80.5±7.6	71.3±8.7	74.7±5.2	67.5±8.2	86.2±7.7
M AN- ooling	100±0	90.6±5.3	71.2±4.3	83.3±6.7	70.9±5.5	76.6±4.0	70.0±4.9	84.8±5.2
Weighted Sum	100±0	92.9±4.8	[illegible]	87.2±4.5	72.5±13	76.3±4.9	66.9±7.8	[illegible]
AM D C	**100±0**	[illegible]	70.0±5.2	93.5±2.4	72.2±4.9	[illegible]	75.0±6.1	84.0±5.2

si cation. e select the values of regularization parameter λ are 10^{-3}, 10^{-2}, 10^{-1}, and 1.

MVCC is a novel multi-vie method based on concept factorization ith local manifold regularization, hich also drives a common consensus representation for multiple vie s. e set parameter α to 100, and both select the values of parameters β and γ are 50, 100, 200, 500, and 1000.

DICS is an NMF-based multi-vie learning algorithm, b e ploring the discriminative and nondiscriminating information e isting in common and vie -speci c parts among different vie s via joint non-negative matri factorization, and produce discriminative and non-discriminative feature from all subspaces. And then, discriminative and nondiscriminative features are further used to produce classi cation results. e select parameters α and β ithin a small range of $[0, 1]$, and set parameter γ to 1.

Due to no publicl available multi-vie clustering algorithm based on deep neural net ork, e generate three baseline models based on deep neural net ork. These baseline models are e plorator models to validate properties of the self-attention mechanism in SAMVDPC, the separatel use ma pooling, mean pooling, and eighted summation to fusion all multiple vie s representations produced b Encoder-Block, nd a fusion representation ith single complementarit , and then input the fusion representation to full connected la er to make prediction. And for fair comparison, e use the same settings for baseline models as hat e did in SAMVDPC.

. Con guration and Tric s

In this subsection, e specif the con guration of SAMVDPC. In Encoder-Block, structures of all encoders are the same, each encoder has one input la er and three hidden la ers l_1, l_2, and l_3, the number of number in each hidden la er decrease as the la ers of encoder deepens, and the activation function of all hidden la ers is ReLU. In self-attention mapping, self-attention MLP has a hidden la er ith 300 units d_s, and e al a s choose the matri embedding to have V ro s (d_c). In MLP, e use a 2-la er ReLU output MLP ith 512 hidden states to output the classi cation result. For objective function, e usuall set λ to 0.0001. For the con guration of three baseline models, e use ma -pooling, mean-pooling, or eighted summation to take replace of the self-attention mechanism in SAMVDPC, and set λ in objective function to 0. The h per parameters on each data sets are summarized in Table 2.

ith regard to the initialization of SAMVDPC eights, In Encoder-Block, e pre-train V auto-encoders through minimizing the reconstruction error of each vie , and then use the pre-trained parameters of auto-encoders to initialize the corresponding encoder's eight of Encoder-Block. In self-attention MLP and MLP, Xavier is used as the eight initialization method [ ].

In training process, the optimizer algorithm e used is Adam, the learning rate is al a s initialized to 10^{-3}, 10^{-4}, 10^{-5}, and ill decreases graduall ith the development of training process. To avoid over tting, all la ers in Encoder-Block and MLP are regularized b dropout regularization in training process, and dropout rate as set to 0.5.

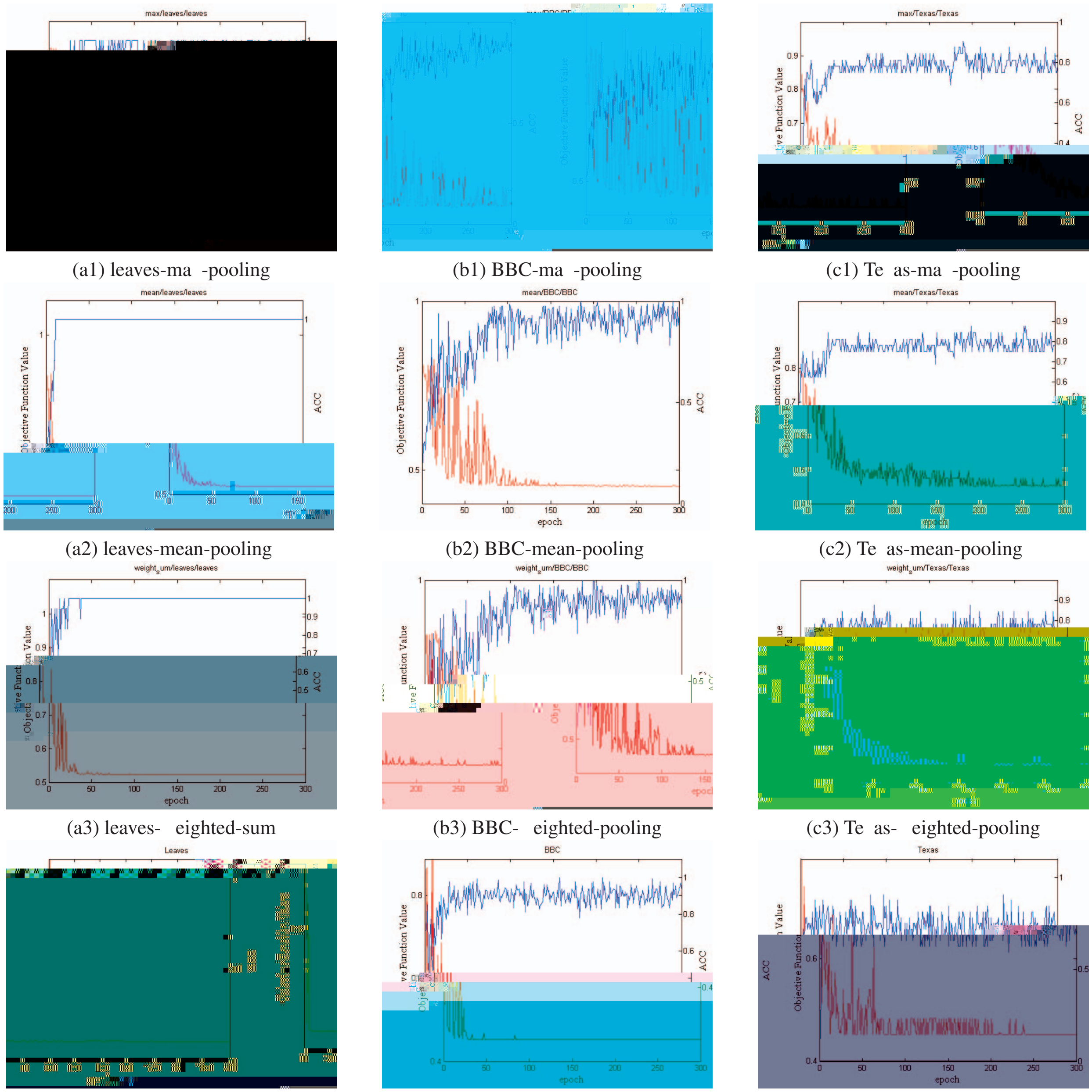
. Result

All datasets divided into training, veri cation and testing data in a ratio of 0.6:0.2:0.2. For SAMVDPC comparison algorithms, and baseline models, e rst run each model on each dataset to select h per parameters that has the best accurac and generalization performance. And then based on these h per parameters, e run all algorithms 10 times on each dataset and report the mean values and standard deviation of accuracies.

All the classi cation results of eight multi-vie datasets are summarized in Table 3, and the best result on each dataset is highlighted in boldface. As e can see, the proposed SAMVDPC achieves best accurac on Te as and ale-Face datasets, and are comparable ith other algorithms on the else datasets. The promising result ma reason from four aspects: (1) DICS, baseline models, and SAMVDPC are all algorithms e ploiting the consistenc and complementarit , simultaneousl , and compared to NMF, multiNMF, and MVCC, the all achieve better performance on all datasets (2) compared to matri factorization algorithms, the Encoder-Block in both baseline models and SAMVDPC can e tract features in a a of effectivel fetching consistent information and grasping the underl ing common properties of multi-vie datasets (3) compared to baseline models, the complementar ith diversit e ploited b self-attention mechanism contains more information than ma -pooling, mean-pooling, and eighted summation.

. Convergence Analysis of Training rocess

In order to empiricall investigate the convergence propert of SAMVDPC, e plot the iterative curves of objective function and the corresponding classi cation accuracies on three t pical data sets, Leaves, BBC, and Te as in Fig. 2. From Fig. 2, e can observe that: (1) the objective function values drop sharpl and mean hile the classi cation accuracies increase rapidl ithin the previous rounds of iterative process, and then the objective function and the accurac curves begin to decrease/gro mildl , nall converge to a value or uctuate around a constant (2) ith respect to convergence speed, the objective function values of SAMVDPC converge in the least iterations, in contrast, ma -pooling corresponds to the most iterations, because ma -pooling operation is loss compression process and the backpropagation process doesn't make full use of information from multiple vie s data (3) in respect of convergence result, the objective function of SAMVDPC can nall converge to a ed value on ever dataset, but the objective function of baseline models al a s nall uctuate around a constant, hat's more, compared to baseline models, e can nd that the classi cation accurac curves of SAMVDPC often uctuate ithin a narro range. In conclusion, compared to baseline models, SAMVDPC get a better performance on the iterative curves of objective function and the corresponding classi cation accurac .



(a1) leaves-ma -pooling (b1) BBC-ma -pooling (c1) Te as-ma -pooling

(a2) leaves-mean-pooling (b2) BBC-mean-pooling (c2) Te as-mean-pooling

(a3) leaves- eighted-sum (b3) BBC- eighted-pooling (c3) Te as- eighted-pooling

(a4) leaves-self-attention (b4) BBC-self-attention (c4) Te as-self-attention

Figure 2: The convergence propert of SAMVDP

/ Conclusion and uture Wor

In this paper, e propose a novel multi-vie net ork,

[2] Li, B. eng, .J. ha, D. Tao, L. ang, and C. Xu. 2011. Dif cult guided image retrieval using linear multivie embedding. In 19th ACM Multimedia Proceedings, 1169-1172.

[3] C. an, R. Pan, and J. Li. 2011. Bi- eighting domain adaptation for cross-language te t classi cation. In 22nd IJCAI Proceedings, 1535-1540.

[4] B. Xie, Mu, D. Tao, and . 2011. Huang. m-sne: Multi-vie stochastic neighbor embedding. IEEE Trans. S stems, Man, and C bernetics, 41(4):1088-1096.

[5] Ch. Xu, D. Tao, and C. Xu. 2013. A Surve on Multi-vie Learning. arXiv preprint arXiv, 1304.5634.

[6] . Chaudhuri, S. M. akade, . Livescu, and . Sridharan. 2009. Multi-vie clustering via canonical correlation anal -sis. In 26th ICML Proceedings, 129-136.

[7] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Sha e-Ta lor, and S. Szedm'k. 2005. T o vie learning: SVM-2 , Theor and Practice. In 18th NIPS Proceedings, 355-362.

[8] D. R. Hardoon, S. Szedm'k, and J. Sha e-Ta lor. 2004. Canonical correlation anal sis: an overvie ith application to learning methods. Neural Computation, 16(12):2639-2664.

[9] M. an, S. Shan, H. hang, S. Lao, and X Chen. 2016. Multi-vie discriminant anal sis. IEEE Trans. Pattern Anal. Mach. Intell, 38(1):188-194.

[10] A. Sharma, A. umar, H. Daum', D. and . Jacobs. 2012. eneralized multivie anal sis: a discriminative latent space. In 20th CVPR Proceedings, 2160-2167.

[11] . uan, L. hang, J. Peng, and J Fan. 2015. Multi-vie concept learning for data representation. IEEE Trans. no l. Data Eng., 27(11): 3016-3028.

[12] J. ao, J. Han, J. Liu, and C. ang. 2013. Multi-vie clustering via joint nonnegative matri factorization. In 13th SDM Proceedings, 252-260.

[13] H. ang, ang, and T. Li. 2016. Multi-vie clustering via concept factorization ith local manifold regularization. In 16th ICDM Proceedings, 1245-1250

[14] A. Blum, T. M. Mitchell. 1998. Combining labeled and unlabeled data ith co-training. In 16th COLT Proceedings, 92-100.

[15] A. umar, and H. Daum'. 2011. A co-training approach for multi-vie spectral clustering. In 28th ICML Proceedings, 393-400.

[16] . ang, . hou. 2010. A ne anal sis of co-training. In 27th ICML Proceedings, 1135-1142.

[17] M. hang, . hou. 2011. CoTrade: con dent co-training ith data editing. IEEE Trans. S stems, Man, and C bernetics, 41(6):1612-1626.

[18] . hang, . in, P. Li, . ang, and J. Shao. 2018. Multi-vie discriminative learning via joint non-negative matri factorization. In 23rd DASFAA Proceedings, 542-557.

[19] A. P. Singh, . J. ordon. 2008. Relational learning via collective matri factorization. In 14th DD Proceedings, 650-658.

[20] N. Srivastava, R. Salakhutdinov. 2014. Multimodal learning ith deep Boltzmann machines. Journal of Machine Learning Research, 15(1):2949-2980.

[21] J. Ngiam, A. hosla, M. im, J. Nam, H. Lee, and A. Ng: 2011. Multimodal Deep Learning. In 1rth DD Proceedings, 689-696.

[22] . Cho, B. V. Merrienboer, . 'l ehre, D. Bahdanau, F. Bougares, H. Sch enk, and Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In 2014 EMNLP Proceedings, 1724-1734.

[23] H. Su, S. Maji, E. alogerakis, and E. . Learned-Miller. 2015. Multi-vie Convolutional Neural Net orks for 3D Shape Recognition. In 2015 ICCV Proceedings, 945-953.

[24] D. Bahdanau, . Cho, and Bengio. 2015. Neural Machine Translation b Jointl Learning to Align and Translate. In 3rd ICLR Proceedings.

[25] I. Sutskever, O. Vin als, and . V. Le. 2014. Sequence to Sequence Learning ith Neural Net orks. In 28t hNIPS Proceedings, 3104-3112.

[26] . Xu, J. Ba, R. iros, . Cho, A. C. Courville, R. Salakhutdinov, R. S. emel, and Bengio. 2015. Sho , Attend and Tell: Neural Image Caption eneration ith Visual Attention. In 32nd ICML Proceedings, 2048-2057.

[27] A. M. Rush, S. Chopra, and J. eston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In 2015 EMNLP Proceedings, 379-389.

[28] J. Cheng, L. Dong, and M. Lapata. 2015. Long Short-Term Memor -Net orks for Machine Reading. In 2016 EMNLP Proceedings, 551-561.

[29] A. P. Parikh, O. T ckstr m, D. Das, and J. Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In 2016 EMNLP Proceedings, 2249-2255.

[30] . Lin, M. Feng, C. N. d. Santos, M. u, B. Xiang, B. hou, and Bengio. 2017. A Structured Self-Attentive Sentence Embedding. In 5th ICLR Proceedings.

[31] A. Vas ani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. omez, L. aiser, and I. Polosukhin. 2017. Attention is All ou Need. In 31t h NIPS Proceedings, 6000-6010.

[32] . ang, X. ong, H. Fu, M. Li, and hang. 2015. Feature e traction via multi-vie non-negative matri factorization ith local graph regularization. In 2015 ICIP Proceedings, 3500-3504.

[33] X. lorot, Bengio: 2010. Understanding the dif cult of training deep feedfor ard neural net orks. In 13th AISTATS Proceedings, 249-256.

[34] C. Finn, P. Abbeel, and S. Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep net orks. In 34th ICML Proceedings, 1126-1135.