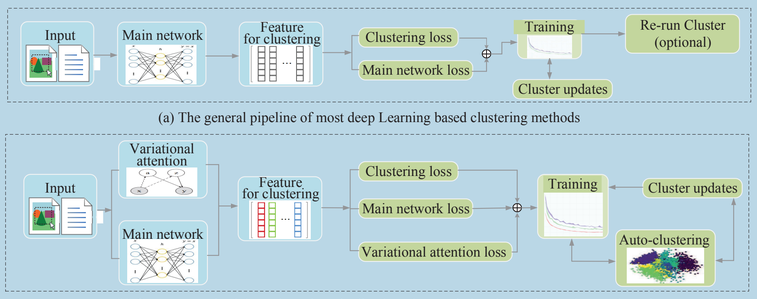
Feature-aware unsupervised learning with joint variational attention and automatic clustering

Ru Wang, Lin Li, Peipei Wang, Xiaohui Tao, Peiyu Liu
ruwang0929@gmail.com, cathylilin@whut.edu.cn, ppwang07@whut.edu.cn, Xiaohui.Tao@usq.edu.au, liupy@sdnu.edu.cn

Background

Most of existing methods remain challenging when handling high-dimensional data and simultaneously exploring the complementarity of deep feature representation and clustering.



(a) The general pipeline of most deep Learning based clustering methods

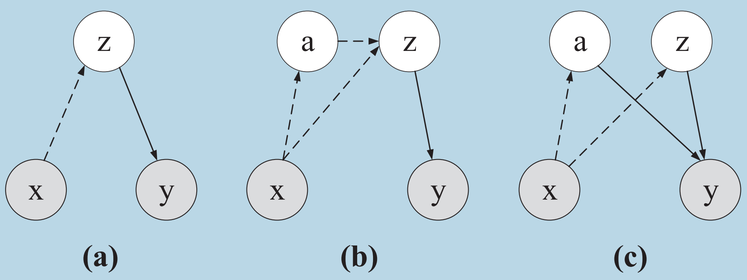
(b) Our unsupervised deep clustering pipeline based on variational attention and auto-clustering

Definition

We consider that the clustering task denotes to divide N samples into K clusters. X=[x_1, x_2, ... , x_n] $\in R^{d\times n}$, d is the dimension of samples and n is the number of samples. We aim to utilize a deep variational encoder-decoder to obtain the representation of each sample x, and via an auto-clustering mechanism to prediction sample's category.

Graphical model

Graphical model representations. (a) variational auto-encoder (VAE),(b) variational encoder-decoder with traditional attention, (c) variational attention encoder-decoder. Dashed lines and solid lines represent the encoding phase and decoding phase respectively.



Datasets

We evaluate the performance of DVAEC on 6 real-world data-sets which includes 3 image datasets and 3 text datasets.

Fashion-MNIST: a dataset consists of 60,000 images labeled as 10 classes.

CIFAR-10: a dataset consists of 60,000 images labeled as 10 classes.

USPS: a dataset contains 9298 grayscale images.

20NEWS: a popular database for text classification or clustering. We use 4 categories.

REUTERS: a dataset has 810,000 English news, following DEC.

StackOverflow: a collection of posts from question and answer site stackoverflow, published as part of a Kaggle challenge .

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 61602353 and 61373148), Hubei Provincial Natural Science Foundation of China (Grant No. 2017CFA012), the National Social Science Foundation under Award (Grant No. 19BYY076), in part Key R & D project of Shandong Province (Grant No. 2019JZZY010129), and Shandong Provincial Social Science Planning Project (Grant No. 18CXWJ01, 18BJYJ04 and 19BJCJ51).

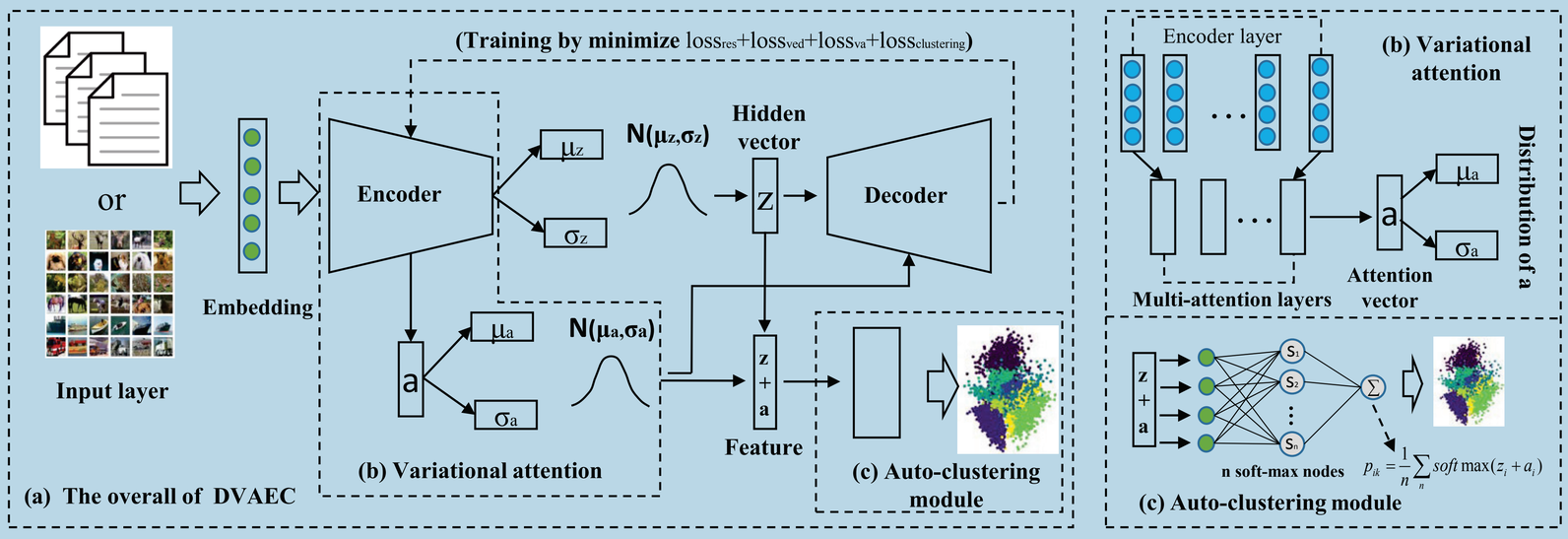
OUR DVAEC

The overall architecture of our DVAEC is a deep encoder-decoder framework that merges a variational attention mechanism and an auto-clustering mechanism.

Input layer: It is to reshape the data as a input of the model. **Encoder layer**: Encoding layer consists of multi-convolutional layers(CNN).**Calculation layer**: It calculates the mean and variance of the encoder's feature. We propose two posterior distributions for encoder, the recognition model p(z|x)$\sim$N(μ_z,σ_z) as a probabilistic encoder and p(a|x)$\sim$N(μ_a,σ_a) as a attention probabilistic encoder from attention vector. **Variational attention layer**: This layer calculates the attention vector between each embedded encoder layer, and then calculates the mean and variance of the attention vector. The attention vector a_{lj} is obtained by a soft-max function from the encoder layer's weight vector e_{lj}. p(a|x, μ_a,σ_a) is sampling from a_i. **Hidden layer**: The hidden layer is the hidden space (z) of the model. **Feature-aware auto-clustering module**: We design a feature-aware auto-clustering module to learn similarity calculation and predict category directly instead of traditional sample similarity calculation process. **Decoder layer**: The decoder layer is the corresponding encoder layer, mainly to reconstruct the input according to the deconvolutional networks. We model the posterior q(x'|z)and q(x'|a) as the general model by (z) and (a). Both q(x'|z)and q(x'|a) are the normal distribution. **Output layer**: The output layer of encoder-decoder framework is to reconstruct the input of the encoder. By minimizing the difference between output x' and the input x to train the generation of the model.

The overall training objective of DVAEC consists of four parts: $loss_{ved} + loss_{va} + loss_{ckustering} + loss_{res}$. We show the objective is to minimize in Eq.(1).

$$J(w,\phi) = j_{res}(w,\phi,x') + \lambda KL(q(x'|z)||p(z|x)) + (1-\lambda)\,KL(q(x'|a)||p(a|x)) + KL(q_{ik}||p_{ik}) \tag{1}$$



Experiments

To verify the effectiveness of the our DVAEC, in this section, we first introduce the experiments settings, and then analyze the experimental results compared with several popular methods. First, we compare four methods for feature representation, and there is K-means clustering algorithm for clustering 4 features (hidden feature of AE, VAE, the hidden feature of DVAEC(z) and hidden of DVAEC add attention vector(z+a)). Besides, we compare with 4 baselines models on the overall performance of the model (DEC, IDEC, VaDE, SpectralNet).

Methods \ Datasets	Fashion		CIFAR-10		USPS		20news		REUTERS		StackOverflow	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means+AE	47.5	51.5	21.8	10.1	65.3	62.8	33.76	0.71	53.3	52.7	40.1	39.2
K-means+VAE	50.3	51.4	28.0	23.4	70.3	61.1	45.3	23.1	73.5	70.7	42.7	40.5
K-means+our(z)	52.0	51.6	31.2	32.6	71.8	68.4	51.8	47.1	77.2	74.3	45.3	44.7
K-means+our($z+a$)	53.4	53.6	32.3	31.7	73.3	72.5	51.9	49.0	78.6	76.3	46.1	45.2
DEC [4]	51.6	54.6	26.3	25.7	74.1	74.3	50.1	44.4	75.6	70.4	46.3	45.6
IDEC [10]	52.9	55.7	25.1	24.7	76.2	78.5	53.6	44.5	-	-	47.9	46.2
VaDE [8]	58.2	57.3	44.6	45.2	76.9	71.2	67.4	63.5	79.2	72.7	47.1	46.5
SpectralNet [21]	60.1	58.7	48.3	**46.7**	82.5	**80.4**	73.4	71.5	82.1	80.0	52.7	50.4
DVAEC (ours)	**64.7**	**61.4**	**50.6**	42.6	**84.5**	79.3	**75.1**	**70.0**	**83.5**	**81.7**	**56.4**	**53.7**

Parameter sensitivity analysis. Impact of the parameter λ on the REUTERS dataset (text). It shows that the variation of ACC (a) and NMI (b) with epoch sizes. The main contrast is the influence of parameter on our model when processing text data. It is limited that only lambda changes, and the remaining parameter settings are consistent.