Variational Deep Embedding Clustering by Augmented Mutual Information Maximization

Qiang Ji[1], Yanfeng Sun[1,*], Yongli Hu[1] and Baocai Yin[2]
[1]Faculty of Information Technology, Beijing University of Technology, Beijing, China
[2]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China

Abstract

Clustering is a crucial but challenging task in pattern analysis and machine learning. Recent many deep clustering methods combining representation learning with cluster techniques emerged. These deep clustering methods mainly focus on the correlation among samples and ignore the relationship between samples and their representations. In this paper, we propose a novel end-to-end clustering framework, namely variational deep embedding clustering by augmented mutual information maximization (VCAMI). From the perspective of VAE, we prove that minimizing reconstruction loss is equivalent to maximizing the mutual information of the input and its latent representation. This provides a theoretical guarantee for us to directly maximize the mutual information instead of minimizing reconstruction loss. Therefore we proposed the augmented mutual information which highlights the uniqueness of the representations while discovering invariant information among similar samples. Extensive experiments on several challenging image datasets show that the VCAMI achieves good performance.

Model Formulation

Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ be a set of D-dimensional samples, $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N$ be a set of d-dimensional latent representations and y is a discrete variable representing the category. We denote the encoder $p_\theta(\mathbf{z}|\mathbf{x})$ that describes the distribution of the encoded variable. The decoder is defined by $q_\phi(\mathbf{x}|\mathbf{z})$. We let joint distribution $p_\theta(\mathbf{x},\mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{x})p(\mathbf{x})$, $q_\phi(\mathbf{x},\mathbf{z}) = q_\phi(\mathbf{x}|\mathbf{z})q(\mathbf{z})$, where $p_\theta(\mathbf{z}|\mathbf{x})$, $q_\phi(\mathbf{x}|\mathbf{z})$ are Gaussian distributions with trainable network parameters θ, ϕ respectively, $p(\mathbf{x})$ is the evidence distribution of $\mathbf{x}$ and $q(\mathbf{z})$ is usually the standard Gaussian distribution.

$$\mathcal{L}_{VCAMI} = \mathcal{L}_{AMI} + \beta\mathcal{L}_{GMM} + \gamma\mathcal{L}_{REG},$$
$$\mathcal{L}_{AMI} = \mathcal{L}_{GMI} + \lambda\mathcal{L}_{LMI} + \alpha\mathcal{L}_{CMI}.$$

$$\mathcal{L}_{GMI}(\theta,\omega) = \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p(\mathbf{x})p_\theta(\mathbf{z}|\mathbf{x})}[\log\sigma(T_\omega(\mathbf{x},\mathbf{z}))]$$
$$+\mathbb{E}_{(\mathbf{x},\mathbf{z})}\sim p(\mathbf{x})p(\mathbf{z})[\log(1-\sigma(T_\omega(\mathbf{x},\mathbf{z}))].$$

$$\mathcal{L}_{LMI}(\theta,\omega,\psi) = \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p(\mathbf{x})p_\theta(\mathbf{z}|\mathbf{x})}[\log\sigma(T_\psi(\mathbf{x},\mathbf{z}))]$$
$$+\mathbb{E}_{(\mathbf{x},\mathbf{z})}\sim p(\mathbf{x})p(\mathbf{z})[\log(1-\sigma(T_\psi(\mathbf{x},\mathbf{z}))].$$

$$\mathcal{L}_{GMM} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\Big[\sum_y p_{\theta_2}(y|\mathbf{z})\log\frac{p_{\theta_1}(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|y)}\Big]$$

$$\mathcal{L}_{REG} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\Big[D_{\mathrm{KL}}\big(p_{\theta_2}(y|\mathbf{z})\big\|q(y)\big)\Big],$$
$$\mathbf{z}\sim p_{\theta_1}(\mathbf{z}|\mathbf{x}).$$

$$\mathcal{L}_{CMI} = \max_\theta I(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}')).$$

where $\sigma(T(\mathbf{x}))$ is a discriminator network.α, β and γ are constants to balance the contributions of different terms. We summarize the overall training process in Algorithm 1.

Clustering performance

Table 1: Clustering performance of different methods on six challenging datasets. The best results are highlighted in bold. (ACC/NMI).

Dataset	MNIST		CIFAR-10		CIFAR-100		STL-10		ImageNet-10		Imagenet-dog	
Metrics	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC
Kmeans	0.501	0.572	0.087	0.229	0.084	0.130	0.125	0.192	0.119	0.241	0.055	0.105
SC	0.662	0.695	0.103	0.247	0.090	0.136	0.098	0.159	0.151	0.274	0.038	0.111
AE	0.725	0.812	0.239	0.314	0.100	0.165	0.250	0.303	0.210	0.317	0.104	0.185
JULE	0.913	0.964	0.192	0.272	0.103	0.137	0.182	0.277	0.175	0.300	0.054	0.138
DEC	0.771	0.843	0.257	0.301	0.136	0.185	0.276	0.359	0.282	0.381	0.122	0.195
VAE	0.876	0.945	0.245	0.291	0.108	0.152	0.200	0.282	0.193	0.334	0.107	0.179
DEPICT	0.917	0.965	0.237	0.279	0.094	0.137	0.229	0.312	0.242	0.363	0.128	0.219
GAN	0.763	0.736	0.265	0.315	0.121	0.151	0.212	0.298	0.225	0.346	0.121	0.174
DeCNN	0.757	0.817	0.240	0.282	0.092	0.133	0.227	0.299	0.186	0.313	0.098	0.175
DAC	0.935	0.977	0.396	0.522	0.185	0.238	0.366	0.470	0.394	0.527	0.219	0.275
ICC	-	0.992	-	0.617	-	0.257	-	0.596	-	-	-	-
DCCM	0.951	0.982	0.496	0.623	0.285	0.327	0.376	0.482	0.608	0.710	0.321	0.383
Ours	**0.987**	**0.995**	**0.521**	**0.654**	**0.301**	**0.338**	**0.391**	**0.512**	**0.636**	**0.746**	**0.375**	**0.391**

Optimization

We conclude that VCAMI objective function monotonically decreases under the optimization in Algorithm. 1.

Algorithm 1 Variational Deep Embedding Clustering by Augmented Mutual Information Maximization

Input: Unlabelled dataset $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$
Parameter: Class number K, α, β and γ
Output: Cluster assignment and embedded representations

```
while epoch ⩽ Maxiter do
    for batch x in X do
        Generate x' via data augmentation;
        Computing L_GMI;
        Computing L_LMI;
        Computing L_CMI;
        Computing L_GMM;
        Computing L_REG;
        Computing L_VCAMI;
        Update model parameters by backpropagation;
    end for
end while
return cluster assignment
```

Conclusion

In this paper, we developed an end-to-end clustering framework, i.e., variational deep embedding clustering by augmented mutual information maximization (VCAMI). To extract the useful representations, we proposed the augmented mutual information, which combines the mutual information variational estimation of continuous variables, the mutual information exact computation of discrete variables, and the data augmentation techniques. While achieving excellent clustering performance, the VCAMI improves the robustness and avoids the degenerate solutions. Extensive experiments on several challenging image datasets show that VCAMI achieves significant improvement over the state-of-the-art methods.

Acknowledgements

This research was supported by National Natural Science Foundation of China under Grant 61772048, 61632006,61672071, U1811463, U19B2039, Beijing Talents Project(2017A24).

Results Analysis

Table 2: Ablation study of VCAMI on the CIFAR-10 datasets.

Dataset	CIFAR-10		
Metrics	NMI	ACC	ARI
$\mathcal{L}_{DMI}$	0.213	0.294	0.184
$\mathcal{L}_{CMI}$	0.397	0.576	0.373
$\mathcal{L}_{AMI}$	0.497	0.632	0.401
$\mathcal{L}_{VCAMI}$	**0.517**	**0.654**	**0.426**

Dataset	ImageNet-10		
Metrics	NMI	ACC	ARI
$\mathcal{L}_{DMI}$	0.273	0.361	0.251
$\mathcal{L}_{CMI}$	0.537	0.641	0.497
$\mathcal{L}_{AMI}$	0.611	0.726	0.553
$\mathcal{L}_{VCAMI}$	**0.632**	**0.742**	**0.573**

In table 1, several tendencies can be observed from the clustering results with further analysis. First, the performance of the clustering methods based on deep learning is generally superior to the traditional methods (e.g., K-means, SC). Secondly, the performance of DCCM and VCAMI using the data enhancement technique is better than that of other algorithms. It implies that introducing data augmentation technology into unsupervised clustering can help the model to be optimized more reasonably and to avoid degenerate solutions. More importantly, both DCCM and VCAMI are dedicated to finding discriminative representations by maximizing triplet-level mutual information. Different from DCCM, VCAMI combines the mutual information estimation between continuous variables and the exact mutual information computation between discrete variables to efficiently obtain the unique and invariant information of the representations.

In table 2, we observe that extracting invariant information is more helpful for the clustering task by maximizing mutual information between similar samples. Combining $\mathcal{L}_{DMI}$ and $\mathcal{L}_{CMI}$, we can see that $\mathcal{L}_{AMI}$ significantly boosts the clustering performance. The only difference between $\mathcal{L}_{AMI}$ and $\mathcal{L}_{VCAMI}$ lies in whether the Gaussian mixture distribution constraint is imposed on the representation or not. We can see that the Gaussian mixture distribution constraint is helpful for the model to extract the representations with cluster-like structures.