Paper: MICE: Mixture of Contrastive Experts for Unsupervised image Clustering.[3]

Jiaxin Liu

Group Reading

September 17, 2021

Overview

1. Introduction
   - Instance discrimination task. (MoCo)
2. Mixture of contrastive experts
   - Gating functions and experts.
3. Inference and Learning
   - EM algorithm
   - EM for MICE.
4. Experiments

Introduction

- A set of images: $X = \{x_n\}_{n=1}^{N}$ without the ground-truth labels.
- A unique surrogate label $y_n \in \{1, 2, \ldots, N\}$ for each x_n, $y_n \neq y_j, \forall j \neq n$.
- Two encoder networks:
  - Student Network $f_\theta : x_n \mapsto v_{y_n} \in \mathbb{R}^d$
  - Teacher Network $f_{\theta'} : x_n \mapsto f_n \in \mathbb{R}^d$
- Probability classifier:

$$p(Y|X) = \prod_{n=1}^{N} p(y_n|x_n) = \prod_{i=1}^{N} \frac{\exp(v_{y_n}^\top f_n/\tau)}{\sum_{i=1}^{N} \exp(v_i^\top f_n/\tau)}$$

  where τ is the temperature hyper-parameter.[1]
- InfoNCE Loss[2] (Noise Contrastive Estimation):

$$\log \frac{\exp(v_{y_n}^\top f_n/\tau)}{\exp(v_{y_n}^\top f_n/\tau) + \sum_{i=1}^{v} \exp(q_i^\top f_n/\tau)}$$

  where $q \in \mathbb{R}^{v \times d}$ is a queue storing previous embeddings from $f_{\theta'}$.

Mixture of contrastive experts

Unsupervised clustering: partition a dataset X with N observations into K clusters.

- Cluster label of x_n: $z_n \in \{1, 2, \ldots, K\}$
- Probability classifier:

$$p(Y|X) = \prod_{n=1}^{N} p(y_n|x_n) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(y_n, z_n = k|x_n)^{\mathbb{1}(z_n=k)}$$

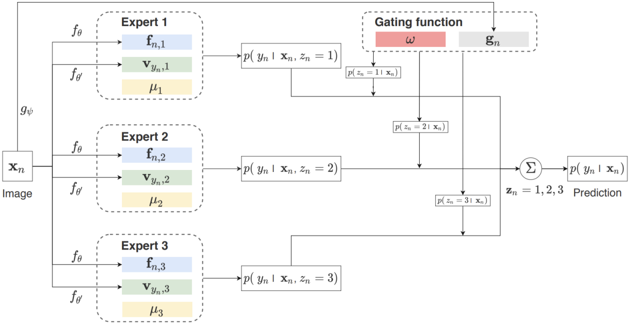
$$= \prod_{n=1}^{N} \prod_{k=1}^{K} p(z_n = k|x_n)^{\mathbb{1}(z_n=k)} p(y_n|x_n, z_n = k)^{\mathbb{1}(z_n=k)}$$

where $\mathbb{1}(\cdot)$ is an indicator function.

Gating functions and experts

$$p(Y|X) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(z_n = k|x_n)^{\mathbb{1}(z_n=k)} p(y_n|x_n, z_n = k)^{\mathbb{1}(z_n=k)}$$

- One expert: $p(y_n|x_n, z_n = k)$
- Gating function: $p(z_n = k|x_n)$



Gating function

Gating function: organizes the instance discrimination task into K simpler subtasks.

- Encoder network $g_\psi : x_n \mapsto g_n \in \mathbb{R}^d$
- Gating function:

$$p(z_n|x_n) = \frac{\exp(\omega_{z_n}^\top g_n/\kappa)}{\sum_{k=1}^K \exp(\omega_k^\top g_n/\kappa)}$$

where κ is the temperature, and $\omega = \{\omega_k\}_{k=1}^K$ is the gating prototypes.

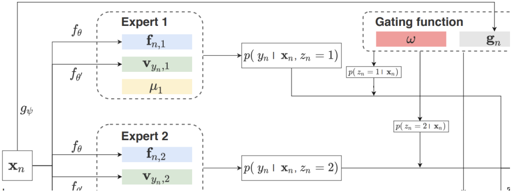


Figure: Gating function.

Experts

Expert: solves the instance discrimination subtask arranged by the gating function.

$$p(y_n|x_n, z_n) = \frac{\Phi(x_n, y_n, z_n)}{Z(x_n, z_n)}$$

$$\Phi(x_n, y_n, z_n) = \exp(v_{y_n,z_n}^\top (f_{n,z_n} + \mu_{z_n})/\tau)$$

- $Z_n(x_n, z_n) = \sum_{i=1}^{N} \Phi(x_n, y_i, z_n)$ is a normalized constant.
- Student network: $f_\theta : x_n \mapsto f_n = \{f_{n,k}\}_{k=1}^{K} \in \mathbb{R}^{K \times d}$
- Teacher network: $f_{\theta'} : x_n \mapsto v_{y_n} = \{v_{y_n,k}\}_{k=1}^{K} \in \mathbb{R}^{K \times d}$
- $\mu = \{\mu_k\}_{k=1}^{K}$ is the cluster prototypes for the experts.

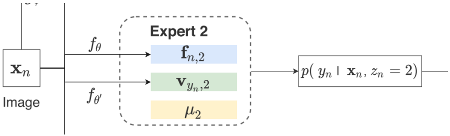


Figure: Expert function.

Expectation Maximization (EM) algorithm

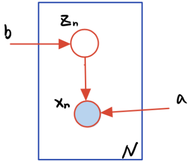


Figure: Graphical Representation for EM algorithm.

Observed variables X, the goal is to maximize the likelihood: $p(X|\theta)$ w.r.t. θ.

- Initial setting for θ^{old}.
- E step: evaluate $p(Z|X, \theta^{old})$.
- M step: evaluate $\theta^{new} = \arg\max_{\theta} \sum_{Z} p(Z|X, \theta^{old}) \ln p(X, Z|\theta)$

Evidence lower bound (ELBO)

$$\log p(Y|X;\theta,\psi,\mu)$$
$$= \mathbb{E}_{q(Z|X,Y)}[\log \frac{p(Y,Z|X;\theta,\psi,\mu)}{q(Z|X,Y)}] + D_{KL}(q(Z|X,Y)\|p(Z|X,Y;\theta,\psi,\mu))$$

cont.

ELBO:

$$L(\theta, \psi, \mu; x_n, y_n) = \mathbb{E}_{q(Z|X,Y)}[\log \frac{p(Y, Z|X; \theta, \psi, \mu)}{q(Z|X, Y)}]$$

E step

$$p(z_n|y_n,x_n;\theta,\psi,\mu)=\frac{p(z_n|x_n;\psi)p(y_n|x_n,z_n;\theta,\mu)}{\sum_{k=1}^{K}p(k|x_n;\psi)p(y_n|x_n,k;\theta,\mu)}$$

Gating function, Expert.

Expert:

$$p(y_n|x_n,z_n)=\frac{\Phi(x_n,y_n,z_n)}{Z(x_n,z_n)}$$

$$\Phi(x_n,y_n,z_n)=\exp(v_{y_n,z_n}^{\top}(f_{n,z_n}+\mu_{z_n})/\tau)$$

$$Z_n(x_n,z_n)=\sum_{i=1}^{N}\Phi(x_n,y_i,z_n)$$

Approximated normalized constant:

$$\hat{Z}(x_n,z_n;\theta,\mu)=\Phi(x_n,y_n,z_n)+\sum_{i=1}^{v}\exp(q_{i,z_n}^{\top}(f_{n,z_n}+\mu_{z_n})/\tau)$$

M step

Stochastic Gradient Ascent to optimize ELBO w.r.t. θ, ψ and μ.

$$\tilde{L}(\theta, \psi, \mu; x_n, y_n) = \mathbb{E}_{q(z_N|x_n, y_n; \theta, \psi, \mu)} \left[\log \frac{\Phi(x_n, y_n, z_n; \theta, \mu)}{Z(x_n, \hat{z_n}; \theta, \mu)} \right] - D_{KL}(q(z_n|x_n, t_n; \theta, \psi, \mu) \| p(z_n|x_n; \psi))$$

cont.

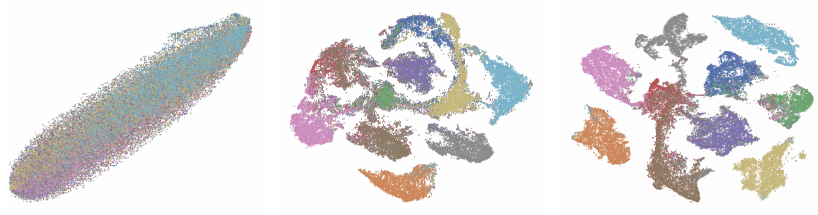
$\arg\max_{\mu_k} = \sum_{n=1}^{N} \hat{q}(z_n = k|x_n, y_n) v_{y_n,k}^{\top} \mu_k / \tau.$ s.t. $\|\mu_k\| = 1.$

Experiments

Datasets	CIFAR-10			CIFAR-100			STL-10			ImageNet-Dog		
Methods/Metrics (%)	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
k-means (Lloyd, 1982)	8.7	22.9	4.9	8.40	13.0	2.8	12.5	19.2	6.1	5.5	10.5	2.0
SC (Zelnik-Manor & Perona, 2004)	10.3	24.7	8.5	9.0	13.6	2.2	9.8	15.9	4.8	3.8	11.1	1.3
AE† (Bengio et al., 2006)	23.9	31.4	16.9	10.0	16.5	4.8	25.0	30.3	16.1	10.4	18.5	7.3
DAE† (Vincent et al., 2010)	25.1	29.7	16.3	11.1	15.1	4.6	22.4	30.2	15.2	10.4	19.0	7.8
SWWAE† (Zhao et al., 2015)	23.3	28.4	16.4	10.3	14.7	3.9	19.6	27.0	13.6	9.4	15.9	7.6
GAN† (Radford et al., 2015)	26.5	31.5	17.6	12.0	15.3	4.5	21.0	29.8	13.9	12.1	17.4	7.8
VAE† (Kingma & Welling, 2013)	24.5	29.1	16.7	10.8	15.2	4.0	20.0	28.2	14.6	10.7	17.9	7.9
JULE (Yang et al., 2016)	19.2	27.2	13.8	10.3	13.7	3.3	18.2	27.7	16.4	5.4	13.8	2.8
DEC (Xie et al., 2016)	25.7	30.1	16.1	13.6	18.5	5.0	27.6	35.9	18.6	12.2	19.5	7.9
DAC (Chang et al., 2017)	39.6	52.2	30.6	18.5	23.8	8.8	36.6	47.0	25.7	21.9	27.5	11.1
DCCM (Wu et al., 2019)	49.6	62.3	40.8	28.5	32.7	17.3	37.6	48.2	26.2	32.1	38.3	18.2
IIC (Ji et al., 2019)	-	61.7	-	-	25.7	-	-	49.9	-	-	-	-
DHOG (Darlow & Storkey, 2020)	58.5	66.6	49.2	25.8	26.1	11.8	41.3	48.3	27.2	-	-	-
AttentionCluster (Niu et al., 2020)	47.5	61.0	40.2	21.5	28.1	11.6	44.6	58.3	36.3	28.1	32.2	16.3
MMDC (Shiran & Weinshall, 2019)	57.2	70.0	-	25.9	31.2	-	49.8	61.1	-	-	-	-
PICA (Huang et al., 2020)	59.1	69.6	51.2	31.0	33.7	17.1	61.1	71.3	53.1	35.2	35.2	20.1
MoCo (Mean)† (He et al., 2020)	66.0	74.7	59.3	38.8	39.5	24.0	60.5	70.7	53.0	34.2	30.8	18.4
MoCo (Std.)† (He et al., 2020)	0.6	1.7	0.9	0.2	0.1	0.4	0.9	2.0	0.8	0.3	1.7	0.9
MiCE (Mean, **Ours**)	**73.5**	**83.4**	**69.5**	**43.0**	**42.2**	**27.7**	**61.3**	**72.0**	**53.2**	**39.4**	**39.0**	**24.7**
MiCE (Std., **Ours**)	0.2	0.2	0.3	0.5	1.4	0.4	1.2	1.8	2.4	1.8	3.0	2.4
MoCo (Best)† (He et al., 2020)	66.9	77.6	60.8	39.0	39.7	24.2	61.5	72.8	52.4	34.7	33.8	19.7
MiCE (Best, **Ours**)	**73.7**	**83.5**	**69.8**	**43.6**	**44.0**	**28.0**	**63.5**	**75.2**	**57.5**	**42.3**	**43.9**	**28.6**

Figure: Unsupervised clustering performance of different methods.

t-SNE plot



(a) Epoch 1 (12.4%) (b) Epoch 500 (70.2%) (c) Epoch 1000 (83.5%)

Figure: Visualization of the image embddings of MiCE.

	CIFAR-100
(a) No analytical update on $\boldsymbol{\mu}$ in Eq. (13)	21.3
(b) No gradient update on $\boldsymbol{\mu}$	41.0
(c) Initialize $\boldsymbol{\omega}$ with a uniform distribution	41.0
(d) Optimize $\boldsymbol{\omega}$ with gradient	42.0
MiCE (Ours)	42.2

Figure: Ablation study for experts and gating.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (2015).

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. "MiCE: Mixture of Contrastive Experts for Unsupervised Image Clustering". In: *International Conference on Learning Representations*. 2020.