# MiCE: Mixture of Contrastive Experts for Unsupervised Image Clustering

Tsung Wei Tsai, Chongxuan Li, Jun Zhu

Department of Computer Science and Technology, Tsinghua University, China

ICLR 2021


ICLR

# Current difficulties in Deep Clustering

- *Discriminative* representation learning that capture the *semantic* similarity between images
- Cumbersome techniques to avoid cluster *degeneracy*
  - Pre-training
  - K-mean initialization
  - Extra regularization terms
  - Combining multiple clustering-related loss

# A Unified Probabilistic Clustering Framework

## Contrastive Learning

- Instance discrimination task
- Discriminative representations

+

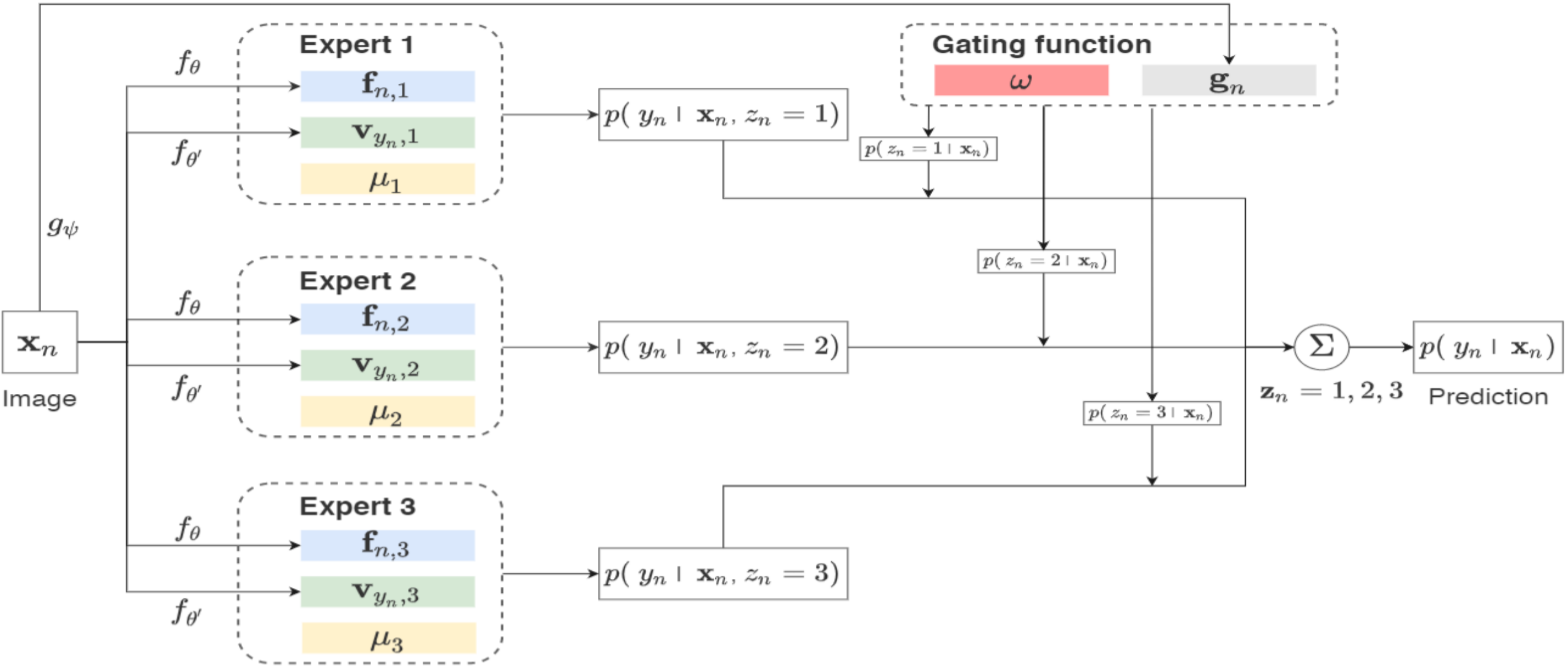
## Latent Mixture Model

- Capture semantic structure
- Divide-and-conquer

||

## MiCE = **1** Gating function + **K** experts

$$p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \underbrace{p(z_n = k|\mathbf{x}_n)}_{\text{Gating function}}^{\mathbb{1}(z_n=k)} \underbrace{p(y_n|\mathbf{x}_n, z_n = k)}_{\text{Expert}}^{\mathbb{1}(z_n=k)}$$




Gating function

$$p(z_n|\mathbf{x}_n) = \frac{\exp(\boldsymbol{\omega}_{z_n}^\top \mathbf{g}_n/\kappa)}{\sum_{k=1}^{K}\exp(\boldsymbol{\omega}_k^\top \mathbf{g}_n/\kappa)},$$

Expert

$$p(y_n|\mathbf{x}_n, z_n) = \frac{\Phi(\mathbf{x}_n, y_n, z_n)}{Z(\mathbf{x}_n, z_n)},$$

$$\Phi(\mathbf{x}_n, y_n, z_n) = \exp\left(\mathbf{v}_{y_n,z_n}^\top (\mathbf{f}_{n,z_n} + \boldsymbol{\mu}_{z_n})/\tau\right),$$

# MiCE – Inference & Learning

- Parameters to update
  - DNNs: $\boldsymbol{\theta}, \boldsymbol{\psi}$
  - Prototypes: $\boldsymbol{\mu}$

$$p(y_n|\mathbf{x}_n, z_n) = \frac{\Phi(\mathbf{x}_n, y_n, z_n)}{Z(\mathbf{x}_n, z_n)},$$

$Z(\mathbf{x}_n, z_n)$ → Intractable

- A variant of EM algorithm
  - **E step:** approximate posterior inference, construct ELBO
  - **M step:** update above parameters to maximize ELBO

# Evidence Lower Bound (ELBO)

- MiCE optimize the ELBO of the log conditional likelihood

$$\log p(y_n|\mathbf{x}_n) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\mu}; \mathbf{x}_n, y_n)$$
$$:= \mathbb{E}_{q(z_n|\mathbf{x}_n, y_n)}[\log p(y_n|\mathbf{x}_n, z_n; \boldsymbol{\theta}, \boldsymbol{\mu})] - D_{\mathrm{KL}}(q(z_n|\mathbf{x}_n, y_n)\|p(z_n|\mathbf{x}_n; \boldsymbol{\psi}))$$

- q(.) is the variational distribution

- 1st term: possibly relief the ***degeneracy*** issue
- 2nd term: refine gating network to consider info in experts
- ***Only a single object function***

# Theoretical Analysis

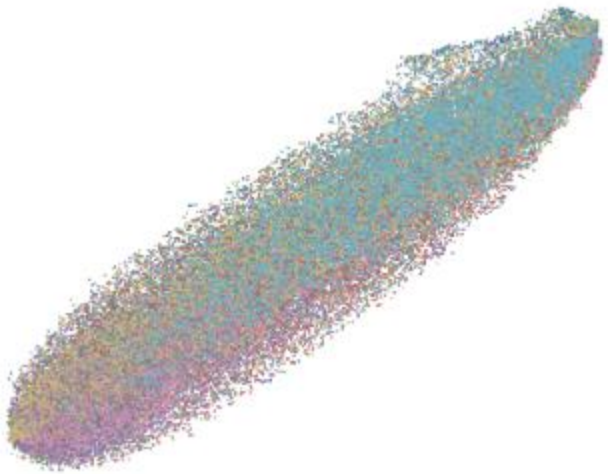
1. Relationship to a two-stage approach
   - MoCo + spherical *k*-means is a special case of MiCE

2. The objective function of MiCE converges under the proposed EM algorithm

# Experiments – comparing to SOTA

| 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** |

# Visualization



(a) Epoch 1 (12.4%)

(b) Epoch 500 (70.2%)

(c) Epoch 1000 (83.5%)

*Thank you for listening*

- Primary contact: peter83112414@gmail.com
- Github: https://github.com/TsungWeiTsai/MiCE