2019 SIAM International Conference on DATA MINING

Deep Co-Clustering

Dongkuan Xu[1], Wei Cheng[2], Bo Zong[2], Jingchao Ni[2], Dongjin Song[2],
Wenchao Yu[2], Yuncong Chen[2], Haifeng Chen[2], Xiang Zhang[1]

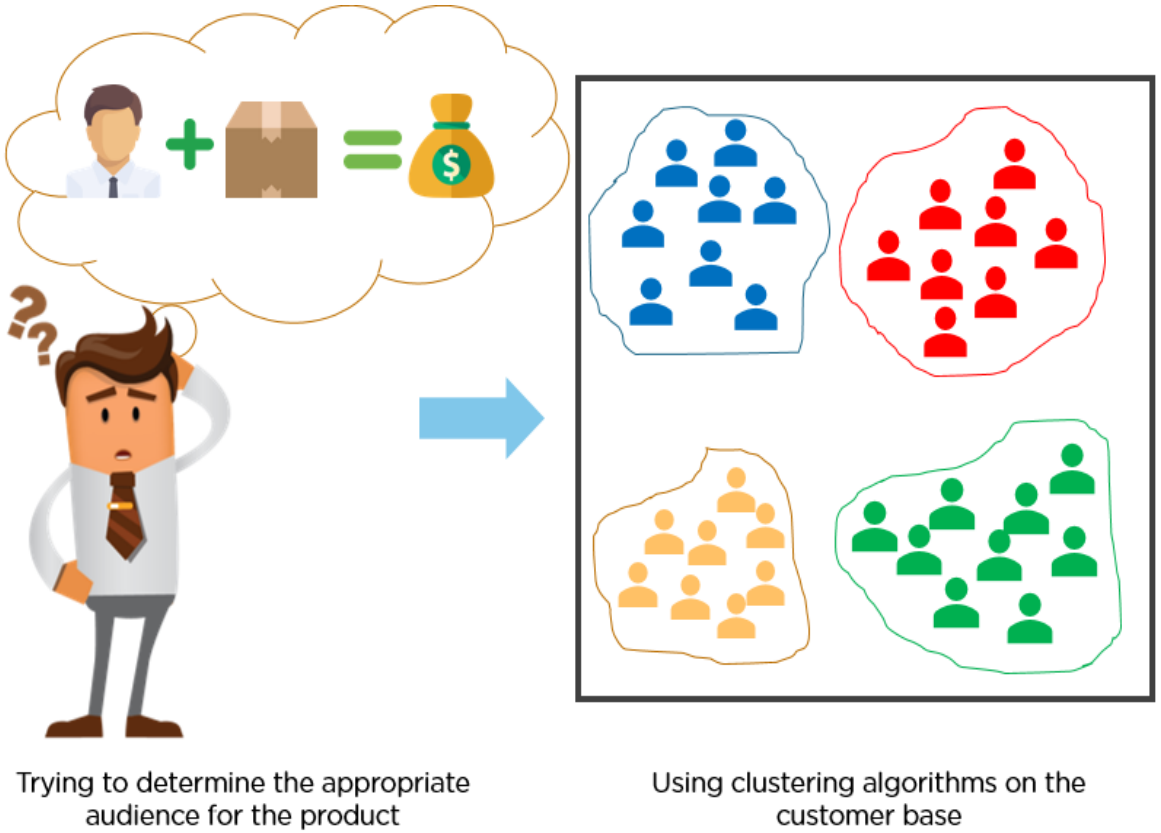
[1]The Pennsylvania State University
[2]NEC Laboratories America

PennState

NEC Laboratories America
Relentless passion for innovation

Clustering

- Group objects into a number of subsets
  - Objects in each subsets are more similar
  - An application



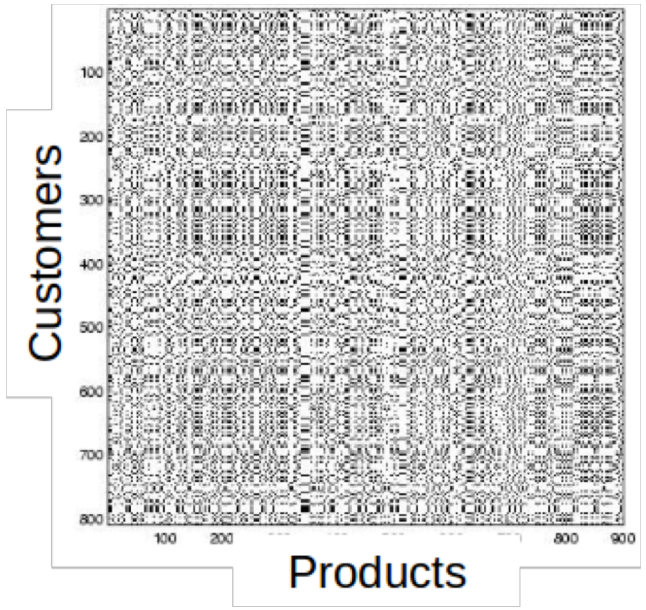
Trying to determine the appropriate audience for the product

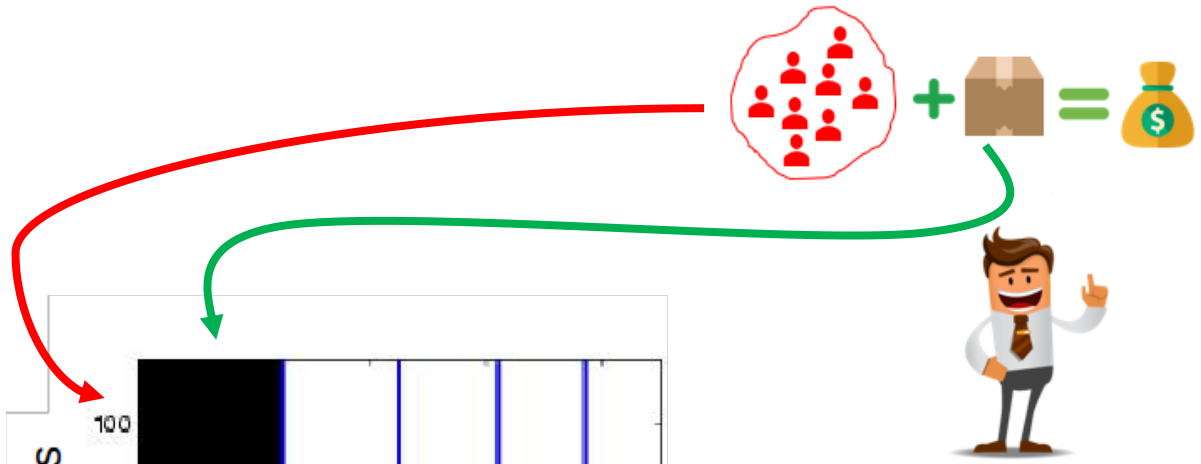
Using clustering algorithms on the customer base

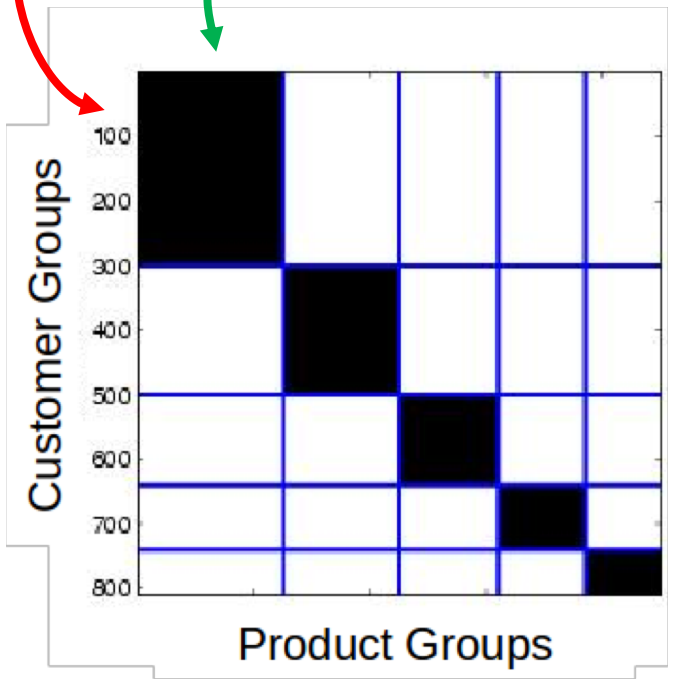
https://www.quora.com/What-is-clustering

Motivation of Co-Clustering

- A large volume of data matrices contain co-cluster structure
- e.g. Customer-product matrices

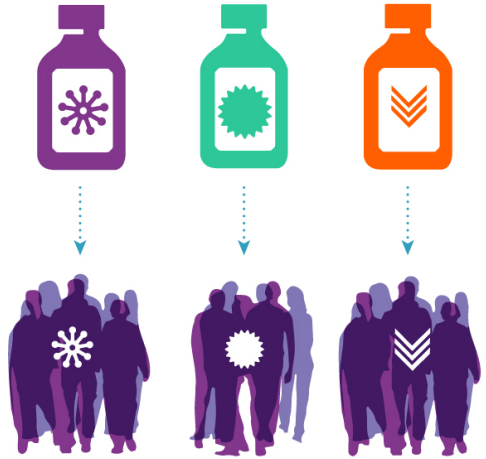






Motivation of Co-Clustering

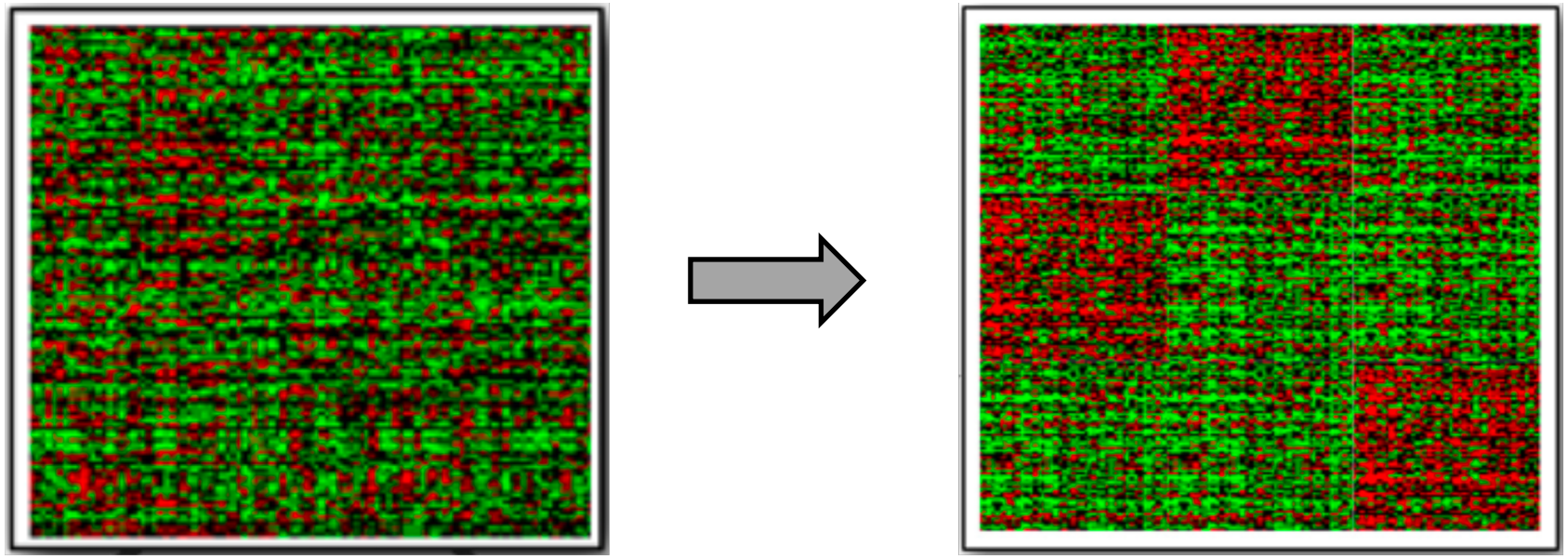
- e.g. Gene-condition matrices
  - Subtypes for a given type of cancer
  - Different patterns of gene expression



Samples

Cancer Subtypes

Genes

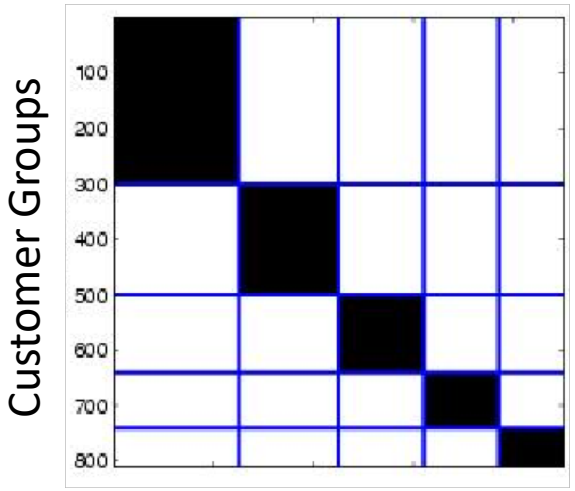


Co-expressed Gene

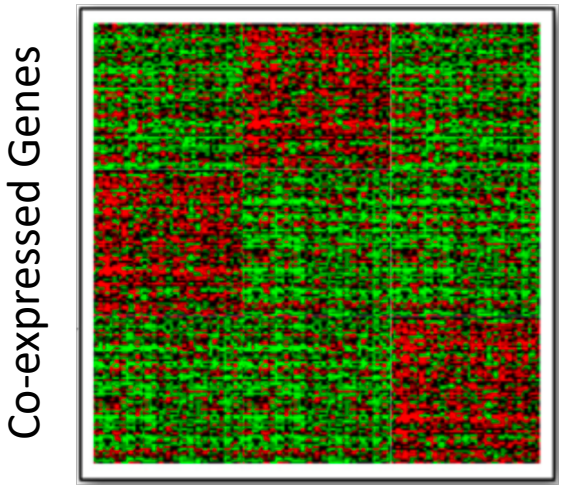
Co-Clustering Problem

- Given a data matrix, simultaneously:
  - Cluster instances into disjoint groups
  - Cluster features into disjoint groups

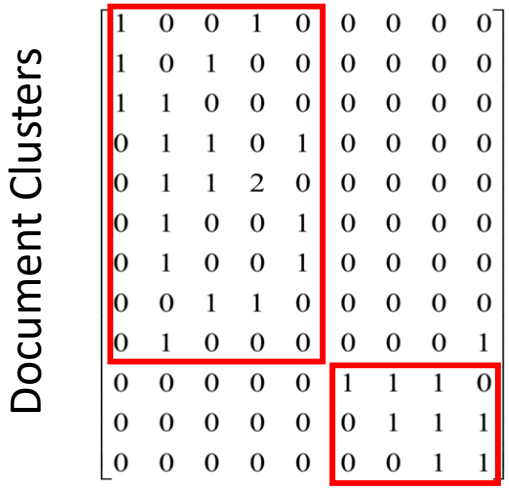
- Key goal is to exploit the “duality” -- co-cluster structure



Product Groups



Cancer Subtypes



Term Clusters

Examples of data matrices containing co-cluster structure

Related Work

- Co-clustering based on information theory
  - Dhillon, Inderjit S., et al. "Information-theoretic co-clustering," in SIGKDD, 2003.
  - Banerjee, Arindam, et al. "A generalized maximum entropy approach to bregman co-clustering and matrix approximation," Journal of Machine Learning Research, 2007.
  - Cheng, Wei, et al. "HICC: an entropy splitting-based framework for hierarchical co-clustering," Knowledge and Information Systems, 2016.

- Co-clustering based on matrix decomposition
  - Cai, Deng, et al. "Non-negative matrix factorization on manifold," in ICDM, 2008.
  - Gu, Quanquan, and Jie Zhou. "Co-clustering on manifolds," in SIGKDD, 2009.
  - Nie, Feiping, et al. "Learning A Structured Optimal Bipartite Graph for Co-Clustering," in NeurIPS. 2017.

- Unsupervised deep learning for clustering
  - Xie, Junyuan, et al. "Unsupervised deep embedding for clustering analysis," in ICML, 2016.
  - Yang, Bo, et al. "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," arXiv preprint arXiv:1610.04794 (2016).
  - Dizaji, Ghasedi, et al. "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in ICCV, 2017.

Related Work

- Co-clustering based on information theory
  - Dhillon, Inderjit S., et al. "Information-theoretic co-clustering," in SIGKDD, 2003.
  - Banerjee, Arindam, et al. "A generalized maximum entropy approach to bregman co-clustering and matrix approximation,"

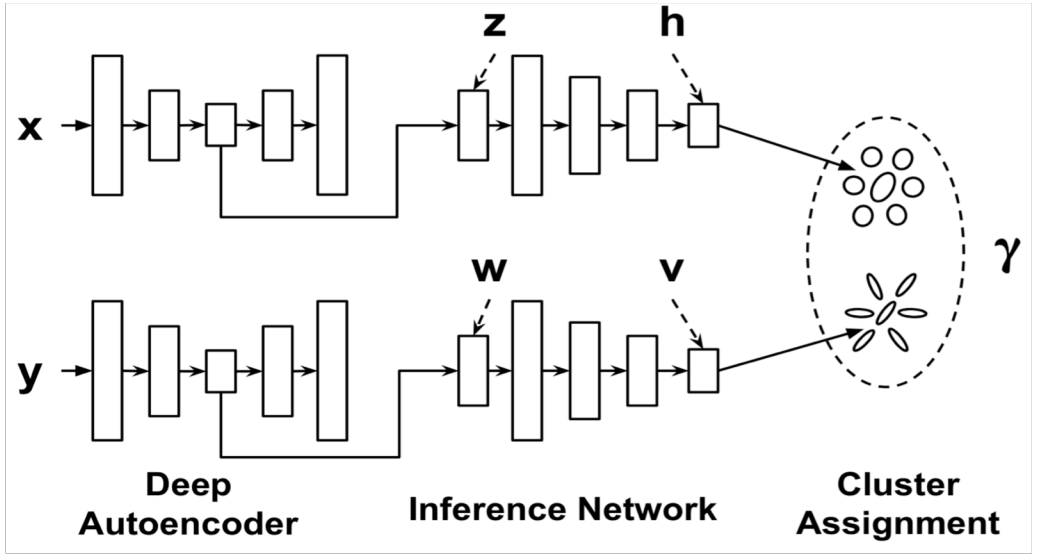
Challenges:

1. Apply deep learning technique to co-clustering
2. Realize feature learning and cluster assignment jointly
3. Learn representations for instances and features simultaneously

  - Xie, Junyuan, et al. "Unsupervised deep embedding for clustering analysis," in ICML, 2016.
  - Yang, Bo, et al. "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," arXiv preprint arXiv:1610.04794 (2016).
  - Dizaji, Ghasedi, et al. "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in ICCV, 2017.

Deep Co-Clustering

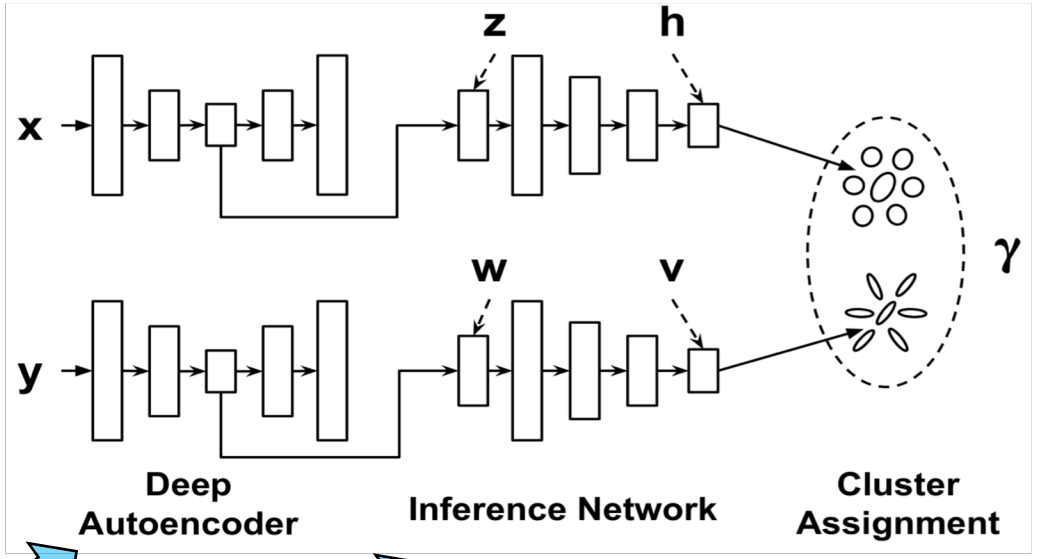
- Model Architecture



- Motivations
  - Autoencoder:
    - Meaningful representation, computation cost
  - Inference network:
    - Initialize cluster assignment
  - Cluster assignment:
    - Update cluster assignment, bridge trainings of instances & features

Deep Co-Clustering

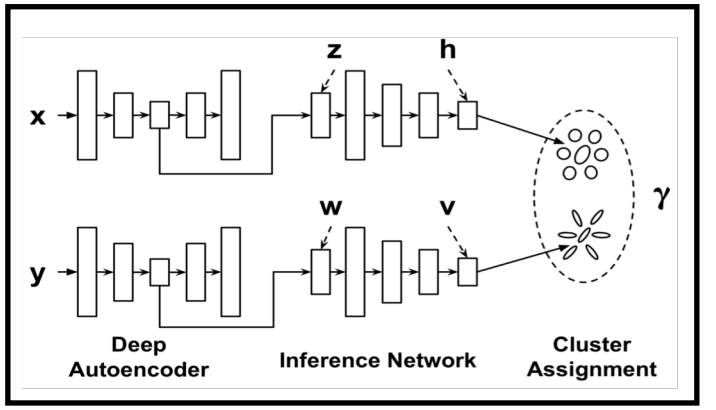
- Model Architecture



- Motivations
  - **Autoencoder:**
    - Meaningful representation, computation cost
  - **Inference network:**
    - Initialize cluster assignment
  - **Cluster assignment:**
    - Update cluster assignment, bridge trainings of instances & features

Deep Co-Clustering

- Objective Function



Model Architecture

$$\min_{\theta_r,\theta_c,\eta_r,\eta_c} J = J_1 + J_2 + J_3$$

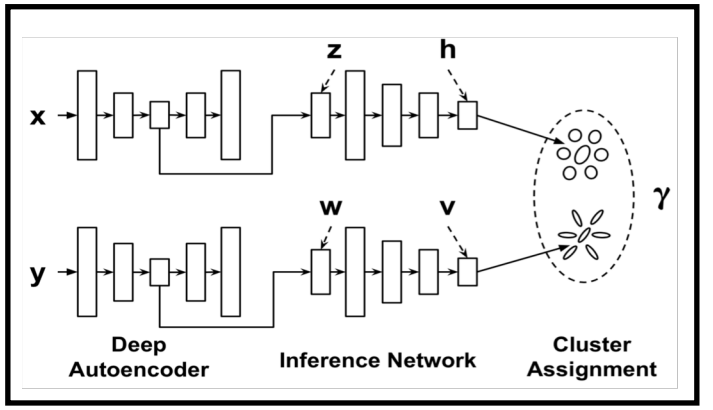
$$J_1 = \frac{\lambda_1}{n}\sum_{i=1}^{n} l(\mathbf{x}_i, g_r(\mathbf{z}_i)) + \lambda_2 P_{ae}(\theta_r) + \lambda_3(-\mathcal{L}_r) + P_{inf}(\Sigma_r)$$

$$J_2 = \frac{\lambda_1}{d}\sum_{j=1}^{d} l(\mathbf{y}_j, g_c(\mathbf{w}_j)) + \lambda_2 P_{ae}(\theta_c) + \lambda_3(-\mathcal{L}_c) + P_{inf}(\Sigma_c)$$

$$J_3 = \lambda_4\left(1 - \frac{I(\hat{X};\hat{Y})}{I(X;X)}\right)$$

Deep Co-Clustering

- Objective Function



Model Architecture

Penalties for autoencoder

$$\min_{\theta_r,\theta_c,\eta_r,\eta_c} J = J_1 + J_2 + J_3$$

$$J_1 = \frac{\lambda_1}{n}\sum_{i=1}^{n} l(\mathbf{x}_i, g_r(\mathbf{z}_i)) + \lambda_2 P_{ae}(\theta_r) + \lambda_3(-\mathcal{L}_r) + P_{inf}(\Sigma_r)$$

$$J_2 = \frac{\lambda_1}{d}\sum_{j=1}^{d} l(\mathbf{y}_j, g_c(\mathbf{w}_j)) + \lambda_2 P_{ae}(\theta_c) + \lambda_3(-\mathcal{L}_c) + P_{inf}(\Sigma_c)$$

$$J_3 = \lambda_4\left(1 - \frac{I(\hat{X};\hat{Y})}{I(X;X)}\right)$$

$$P_{inf}(\Sigma_r) = \sum_{k=1}^{g}\sum_{i=1}^{d_r}\frac{1}{\sum_{r_{ii}}^{k}}$$

Reconstruction Error of Autoencoder

Negative of Variational Lower Bound

Cross Loss Term

- Motivation
  - Mutual information between instance and feature changes least after optimal co-clustering [1][2]
- Joint probability distribution

$$p(X,Y) = \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

e.g. Joint prob. dist. before co-clustering

$$p(\hat{X},\hat{Y}) = \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}$$

e.g. Joint prob. dist. after co-clustering

- Mutual information

$$I(X;Y) = \sum_{\mathbf{x}_i}\sum_{\mathbf{y}_j} p(\mathbf{x}_i,\mathbf{y}_j)\log\frac{p(\mathbf{x}_i,\mathbf{y}_j)}{p(\mathbf{x}_i)p(\mathbf{y}_j)}$$

$$I(\hat{X};\hat{Y}) = \sum_{\hat{\mathbf{x}}_s}\sum_{\hat{\mathbf{y}}_t} p(\hat{\mathbf{x}}_s,\hat{\mathbf{y}}_t)\log\frac{p(\hat{\mathbf{x}}_s,\hat{\mathbf{y}}_t)}{p(\hat{\mathbf{x}}_s),p(\hat{\mathbf{y}}_t)}$$

$$J_3 = \lambda_4(1 - \frac{I(\hat{X};\hat{Y})}{I(X;X)})$$

Cross loss term

[1] I. S. Dhillon, S. Mallela, and D. S. Modha, “Information-theoretic co-clustering,” SIGKDD, 2003.
[2] W. Cheng, X. Zhang, F. Pan, and W. Wang, “Hicc: an entropy splitting-based framework for hierarchical co-clustering,” Knowledge and Information Systems, 2016.

Cross Loss Term

- Motivation
  - Mutual information between instance and feature changes least after optimal co-clustering [1][2]
- Joint probability distribution

$$p(X,Y) = \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

e.g. Joint prob. dist. before co-clustering

$$p(\hat{X},\hat{Y}) = \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix}$$

e.g. Joint prob. dist. after co-clustering

- Mutual information

$$I(X;Y) = \sum_{\mathbf{x}_i}\sum_{\mathbf{y}_j} p(\mathbf{x}_i,\mathbf{y}_j)\log\frac{p(\mathbf{x}_i,\mathbf{y}_j)}{p(\mathbf{x}_i)p(\mathbf{y}_j)}$$

$$I(\hat{X};\hat{Y}) = \sum_{\hat{\mathbf{x}}_s}\sum_{\hat{\mathbf{y}}_t} p(\hat{\mathbf{x}}_s,\hat{\mathbf{y}}_t)\log\frac{p(\hat{\mathbf{x}}_s,\hat{\mathbf{y}}_t)}{p(\hat{\mathbf{x}}_s),p(\hat{\mathbf{y}}_t)}$$

$$J_3 = \lambda_4(1 - \frac{I(\hat{X};\hat{Y})}{I(X;X)})$$

Cross loss term

[1] I. S. Dhillon, S. Mallela, and D. S. Modha, “Information-theoretic co-clustering,” SIGKDD, 2003.
[2] W. Cheng, X. Zhang, F. Pan, and W. Wang, “Hicc: an entropy splitting-based framework for hierarchical co-clustering,” Knowledge and Information Systems, 2016.

Datasets & Baselines

- Datasets

Image datasets

Other datasets

Dataset	#instances	#features	#classes
Coil20	1440	1024	20
Yale	165	1024	15
Fashion-MNIST-test	10000	784	10
Sign-MNIST-test	7172	784	25
Citeseer	3312	3703	6
WebKB4	4199	1000	4
WebKB_cornell	195	1703	5
WebKB_texas	187	1703	5
WebKB_washington	230	1703	5
WebKB_wisconsin	265	1703	5
IMDb_movies_keywords	617	1878	17
IMDb_movies_actors	617	1398	17

Text datasets

- Baseline methods
  - *k*-means [Forgy, *Biometrics'65*]
  - Spectral Co-Clustering [Dhillon, *KDD'01*]
  - Spectral Biclustering [Kluger et al., *Genome Research'03*]
  - CCInfo [Dhillon et al., *KDD'03*]
  - DRCC [Gu and Zhou, *KDD'09*]
  - CCMod [Ailem et al., *CIKM'15*]
  - SCMK [Kang et al., *AAAI'17*]
  - Variants: DCC-INF, DCC-INF, DCC-INF

Clustering Results & Visualization

- Evaluated by clustering accuracy & normalized mutual information (NMI)

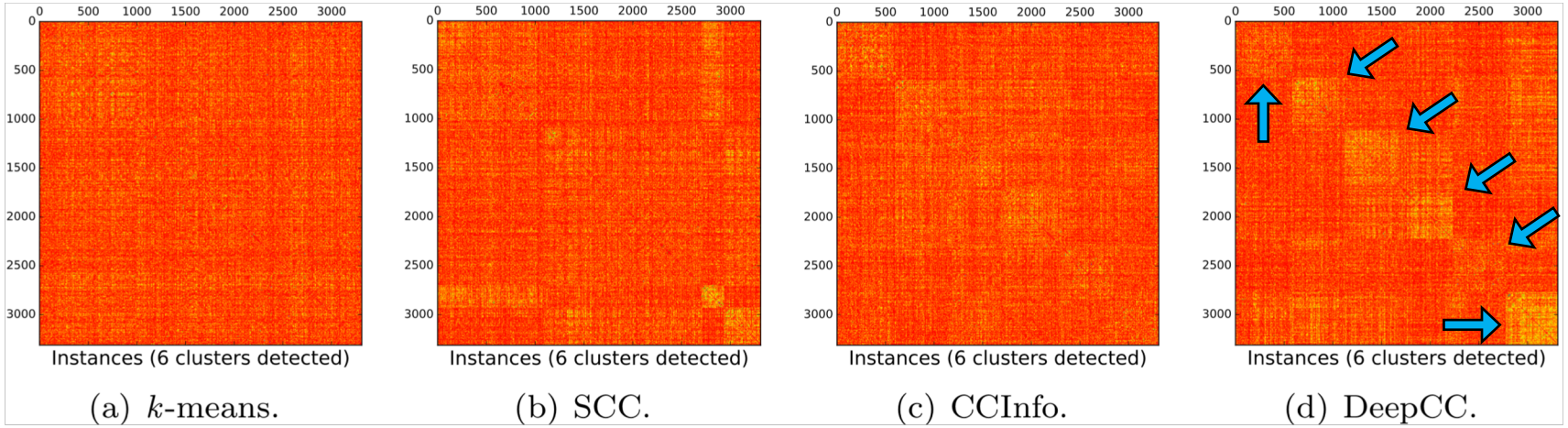
Table 2: Clustering accuracy (%) comparison. The best performance on each dataset is in bold.

Dataset	k-means	SCC	SBC	CCMod	DRCC	CCInfo	SCMK	DeepCC
Coil20	58.6±2.3	51.7±0.5	66.8±1.1	21.0±2.0	53.2±2.4	60.6±3.4	65.9±0.8	**73.3±1.9**
Yale	41.8±0.7	33.7±0.3	40.0±1.3	21.4±1.4	13.6±0.4	41.8±2.0	46.6±0.5	**53.3±1.4**
Fashion-MNIST-test	54.6±0.4	44.5±0.5	45.8±0.0	28.8±0.0	44.1±1.8	51.8±2.4	-	**62.7±1.6**
Sign-MNIST-test	30.6±0.6	31.8±0.7	18.0±0.0	12.6±0.3	21.3±2.5	33.2±1.4	-	**37.0±1.3**
Citeseer	37.4±0.0	37.4±0.0	40.8±0.1	44.7±5.2	29.5±1.8	43.0±5.3	50.2±0.7	**59.3±2.1**
WebKB4	60.6±0.1	60.6±0.1	47.5±0.1	68.8±3.1	43.6±0.4	68.8±2.5	52.1±0.2	**71.8±2.8**
WebKB_cornell	55.1±2.1	58.9±0.2	54.4±0.6	55.5±2.6	42.6±0.0	56.6±2.7	49.6±0.2	**68.7±1.4**
WebKB_texas	63.9±2.6	59.4±0.2	59.0±0.3	64.5±3.0	55.1±0.0	64.1±3.6	62.0±0.6	**73.8±1.2**
WebKB_washington	65.6±2.7	60.8±0.0	51.7±1.0	68.0±2.7	46.5±0.0	67.7±2.9	65.4±0.4	**75.7±1.9**
WebKB_wisconsin	71.7±3.1	70.2±0.5	72.8±1.4	72.1±3.9	46.1±0.0	72.9±3.1	73.2±0.9	**77.4±1.4**
IMDb_movies_keywords	19.3±0.8	25.2±0.4	24.0±0.2	24.7±2.1	12.6±1.7	23.0±2.0	23.3±1.1	**30.8±1.7**
IMDb_movies_actors	15.4±0.7	20.5±0.4	20.0±0.4	20.0±1.2	14.1±2.8	15.6±0.7	15.8±1.3	**23.8±0.4**

Table 3: NMI (%) comparison. The best performance on each dataset is in bold.

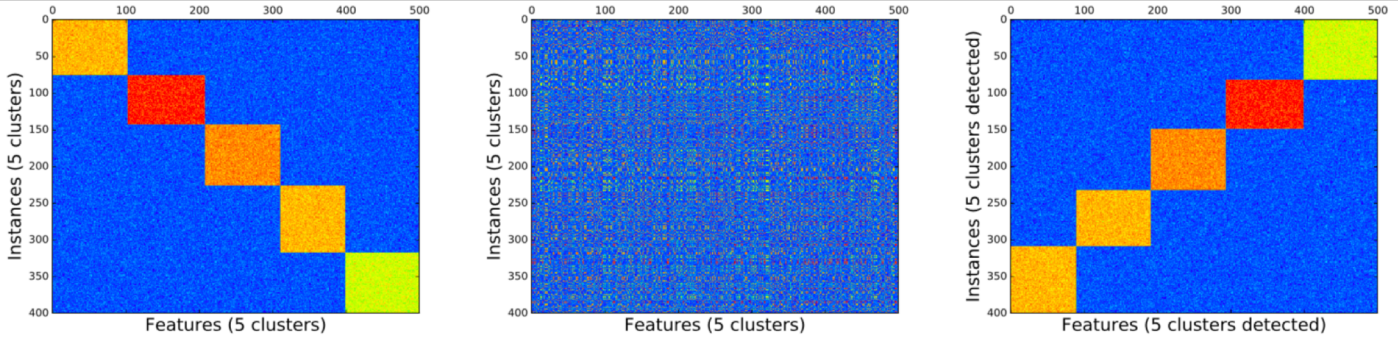
Dataset	k-means	SCC	SBC	CCMod	DRCC	CCInfo	SCMK	DeepCC
Coil20	75.9±2.3	64.9±0.5	73.9±1.1	51.8±1.9	65.6±2.7	72.7±1.5	72.5±0.9	**78.3±2.7**
Yale	48.7±0.7	41.6±0.3	49.8±1.3	24.6±2.3	14.2±1.2	48.5±2.0	49.2±1.2	**55.7±1.1**
Fashion-MNIST-test	52.4±0.4	41.9±0.5	41.3±0.0	45.8±1.4	42.2±1.6	50.6±2.3	-	**60.4±0.7**
Sign-MNIST-test	29.1±0.6	40.1±0.7	17.2±0.0	14.0±1.8	28.2±1.2	43.1±1.0	-	**46.7±0.6**
Citeseer	2.7±0.0	15.2±0.0	17.3±0.1	16.9±1.6	10.5±2.2	17.7±2.2	21.1±1.5	**29.8±1.3**
WebKB4	26.1±0.1	31.1±0.1	13.0±0.1	40.1±1.0	31.9±1.7	39.7±3.6	10.0±2.3	**40.5±0.6**
WebKB_cornell	18.1±2.1	28.8±0.2	21.0±0.6	18.9±3.8	11.6±0.0	20.6±3.1	25.7±0.5	**35.4±0.9**
WebKB_texas	7.0±2.6	12.6±0.2	9.0±0.3	16.9±2.3	10.2±0.0	18.2±4.4	24.0±0.8	**42.9±1.2**
WebKB_washington	33.3±2.7	25.3±0.0	9.5±1.0	28.7±1.4	15.7±0.0	30.7±3.4	30.3±0.2	**45.9±1.3**
WebKB_wisconsin	37.5±3.1	35.4±0.5	38.2±1.4	35.1±2.8	20.4±0.0	39.3±2.7	42.9±0.4	**46.7±1.7**
IMDb_movies_keywords	13.9±0.8	25.5±0.4	20.6±0.2	21.6±1.1	6.9±0.3	18.7±2.3	18.4±0.8	**26.8±1.6**
IMDb_movies_actors	10.5±0.7	19.3±0.4	17.6±0.4	14.5±0.9	9.3±2.5	9.7±1.0	10.6±1.7	**20.6±2.3**

- Visualization of clustering results on Citeseer



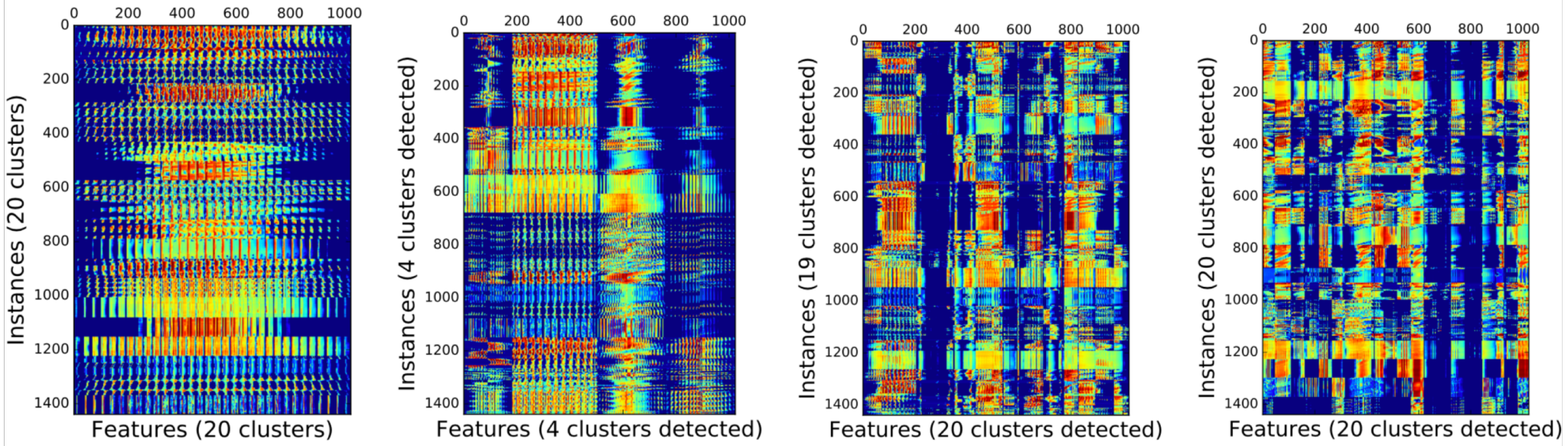
(a) k-means. (b) SCC. (c) CCInfo. (d) DeepCC.

Visualization of Co-Clustering Results



(a) Synthetic data. (b) Data are shuffled. (c) Co-clustering result.

Figure 3: The co-clustering result of DeepCC on the synthetic data.

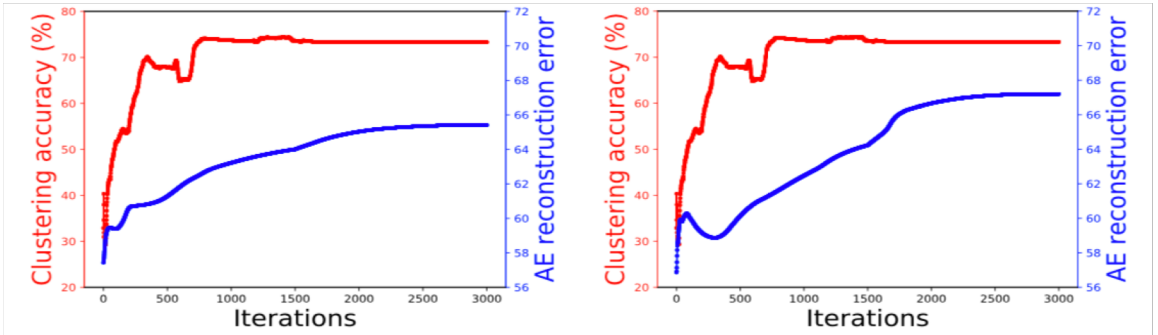


(a) The Coil20 dataset. (b) CCMod. (c) SCC. (d) DeepCC.

Figure 4: Visualization of the co-clustering results on the Coil20 dataset.

Insights of Effectiveness

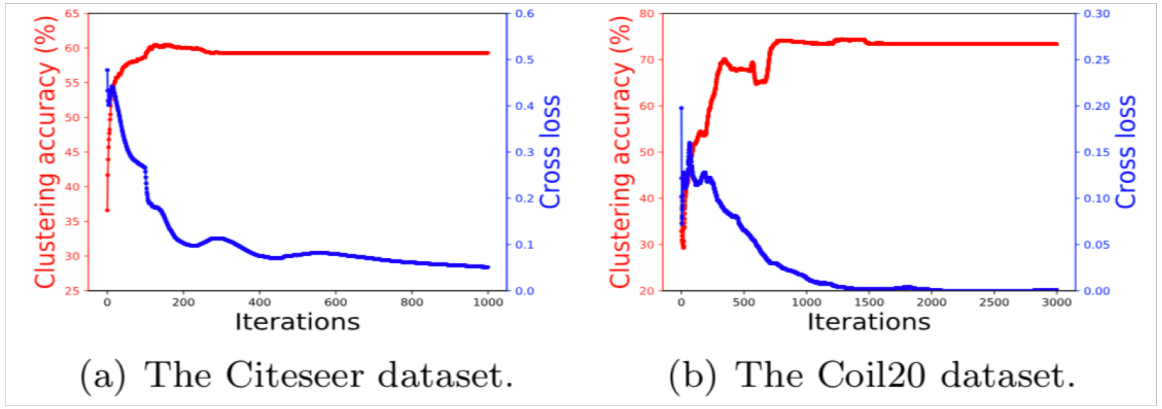
- Advantage of end-to-end training is demonstrated



(a) On the instances of Coil20. (b) On the features of Coil20.

Reconstruction error of autoencoder v.s. clustering accuracy

- Effectiveness of cross loss is demonstrated



(a) The Citeseer dataset. (b) The Coil20 dataset.

Cross loss v.s. clustering accuracy

Summary

- ✓ A deep co-clustering model
- ✓ End-to-end training
- ✓ Experimental evaluations

Thanks!

Q & A