Improving Image Clustering With Multiple Pretrained CNN Feature Extractors

J. Guérin[1], **O. Gibaru**[1], **S. Thiery**[1], **E. Nyiri**[1] & **B. Boots**[2]
[1]LISPEN, Arts et Métiers ParisTech, Lille, France
[2]School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA
Contact: joris.guerin@ensam.eu

1- Image Clustering

- **Inputs :** Set of unlabelled images.
- **Outputs :** Images grouped into clusters.

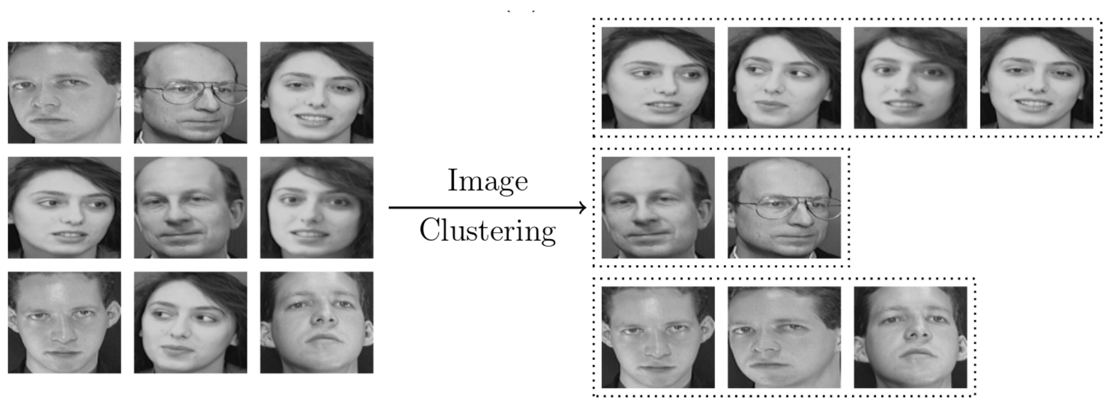

Fig. 1: Example of inputs/outputs of the image clustering problem

Remark : *User defined number of clusters.*

2- Standard Approach

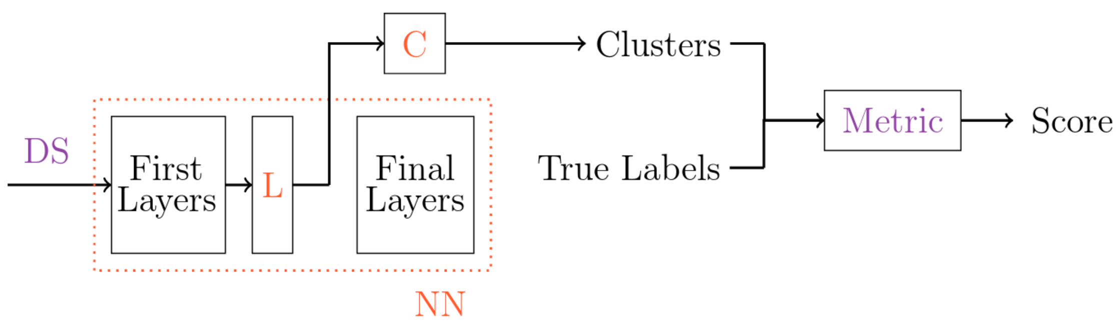

Fig. 2: Standard “pretrained CNN feature extraction + clustering” pipeline

Feature extraction:

- Many pretrained architectures publicly available.
- Choice of architecture (NN) and layer (L) important [1] but often arbitrary.

Clustering:

- Standard clustering methods: K-means (**KM**), Agglomerative Clustering (**AC**)
- Deep end-to-end clustering methods:
  - Partitionning method: **IDEC** [2].
  - Graph-based method: **JULE** [3].

3- Intuition

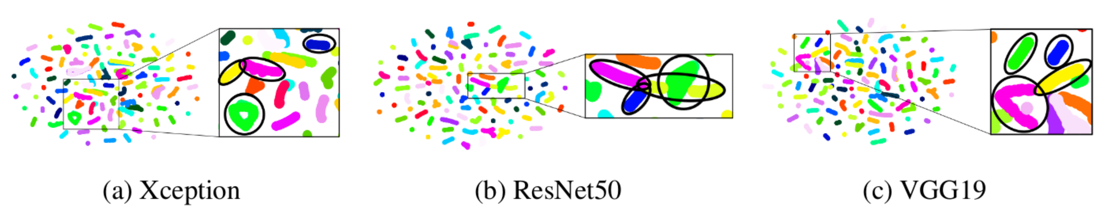

Fig. 3: 2D t-SNE visualization of features extracted by the last layer of three pretrained CNNs for the COIL100 dataset.

- Many possible ways to solve ImageNet.
- Different CNNs might contain complementary information

⟶ Ensemble methods

4- Proposed approach

- Use all available pretrained nets to generate different views of the original data. This creates a multi-view clustering (MVC) problem.
- Use a multi-input MLP to solve MVC within any existing deep clustering frameworks.

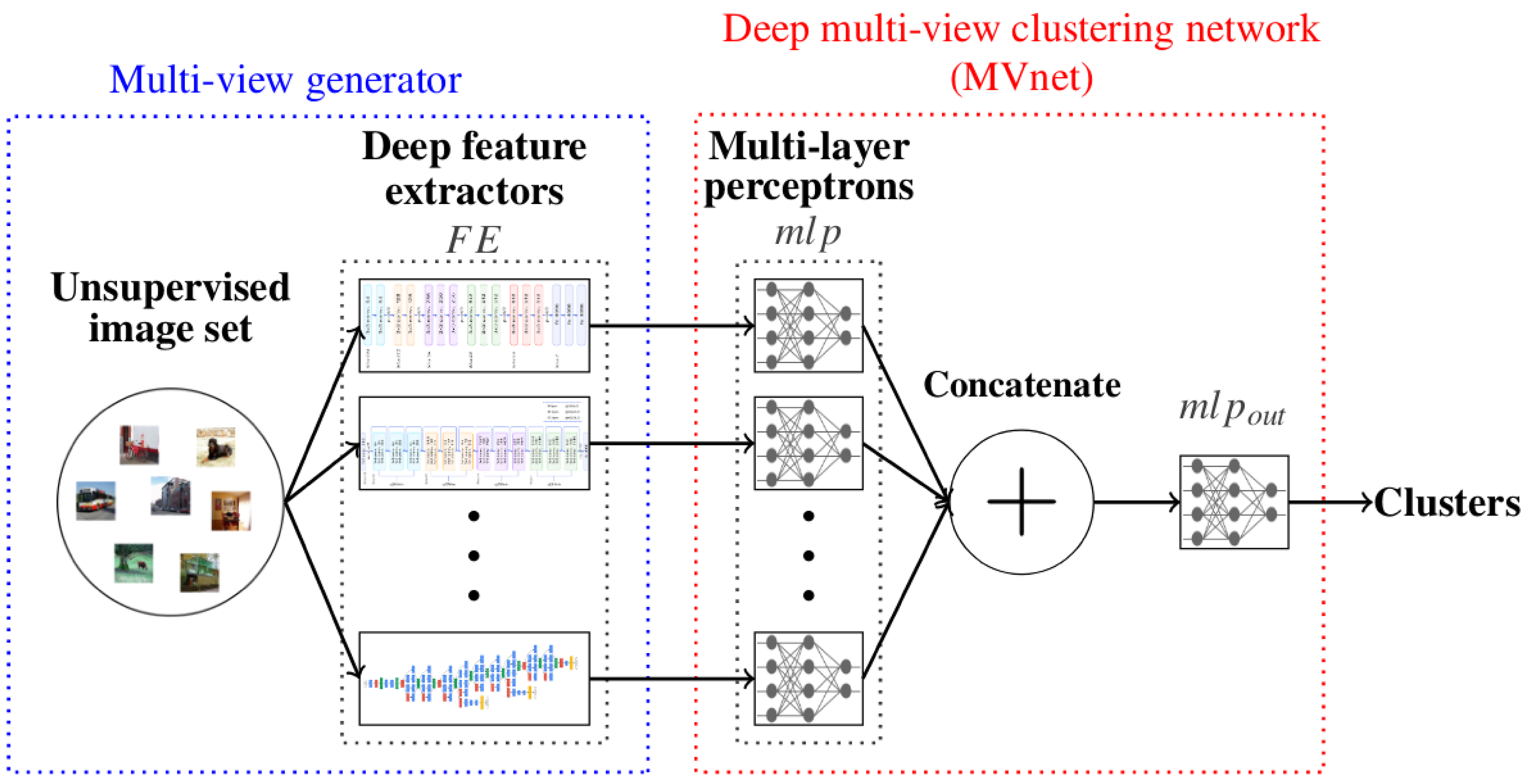

Fig. 4: Proposed multi-view generation + deep multi-view clustering (DMVC) approach to solve the Image Clustering problem.

5- Datasets description

Table 1: Datasets used for our experiments.

Dataset	COIL100	UMist	VOC2007
# Images	7200	575	2841
# Classes	100	20	20
Image Size	128x128	112x92	Variable

6- Experimental results

Table 2: Comparison of clustering performance (NMI) of DMVC against different MV clustering methods and different fixed CNN features.

	VOC2007		COIL100		UMist	
	JULE	IDEC	JULE	IDEC	JULE	IDEC
VGG16	0.687	0.666	0.989	0.963	0.920	0.771
VGG19	0.684	0.677	0.994	0.963	0.933	0.742
InceptionV3	0.768	0.760	0.984	0.957	0.823	0.705
Xception	0.759	0.779	0.986	0.955	0.829	0.707
ResNet50	0.679	0.691	**0.997**	0.973	0.919	0.784
CC	0.718	0.587	0.995	0.886	0.855	0.699
MVEC	0.785	0.782	**0.996**	0.977	**0.963**	0.797
DMVC-fix	**0.792**	0.730	**0.996**	0.973	**0.963**	0.737
DMVC	**0.810**	-	0.995	-	**0.971**	-

- Using several pretrained CNNs enables to
  - Improve image clustering,
  - Avoid the feature extractor selection problem.
- Multi-view clustering can be improved by adopting end-to-end training.
- These two ideas can be combined to obtain state-of-the-art results at image clustering.

7- Learned representations

Table 3: Comparison of clustering performance (NMI) of a simple method (KMeans) applied to different representations of the dataset.

	VOC2007	COIL100	UMist
InceptionV3	0.624	0.932	0.680
InceptionV3 + JULE	0.754	0.938	0.775
DMVC-fix	0.759	0.961	0.895
DMVC	**0.786**	**0.964**	**0.973**

- DMVC enables to get a single unified feature representation despite the initial split into multiple views.
- This new feature representation separates the original data better and is more compact.

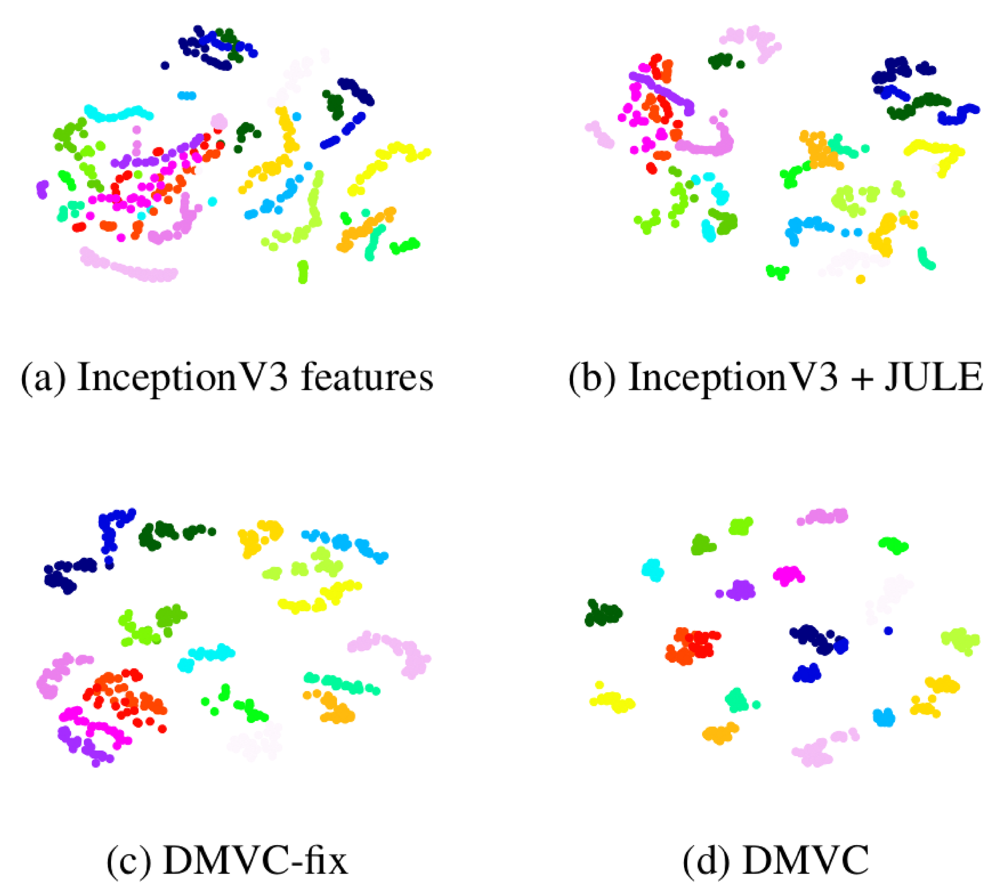

Fig. 5: 2D t-SNE visualization of the features extracted from the UMist dataset at different stages of the DMVC framework.

8- References

[1] J. Guérin, O. Gibaru, S. Thiery, and E. Nyiri, “Cnn features are also great at unsupervised classification,” *arXiv preprint arXiv:1707.01700*, 2017.
[2] X. Guo, L. Gao, X. Liu, and J. Yin, “Improved deep embedded clustering with local structure preservation,” in *International Joint Conference on Artificial Intelligence (IJCAI-17)*, 2017, pp. 1753–1759.
[3] J. Yang, D. Parikh, and D. Batra, “Joint unsupervised learning of deep representations and image clusters,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5147–5156.