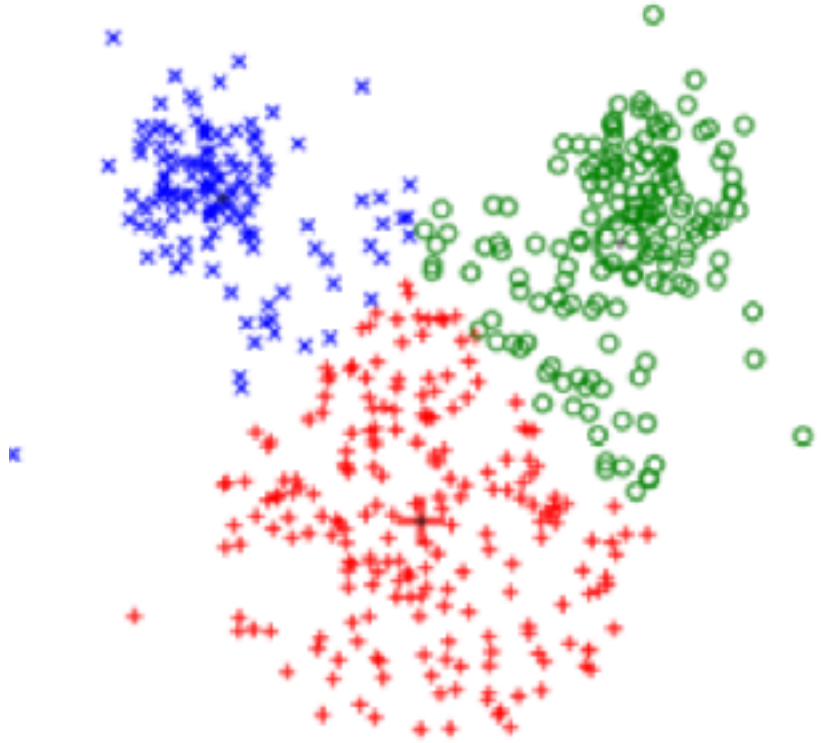
Deep Clustering for Unsupervised Learning of Visual Features

ECCV 2018

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze

Clustering

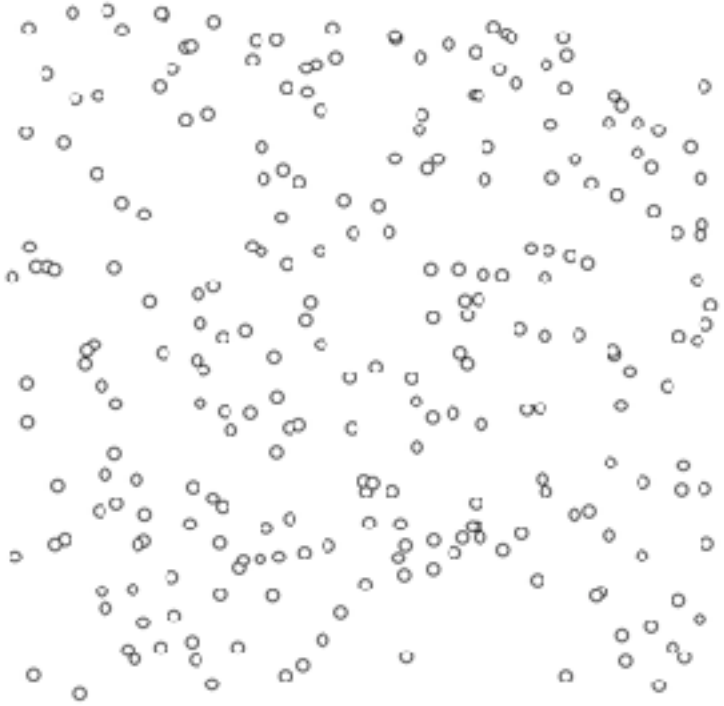
- The goal of clustering is to learn some underlying hidden structure of the data without any ground truth labels.



K-means clustering

K-means Clustering

- K-means algorithm identifies K clusters, and then allocates every data point to the nearest cluster.

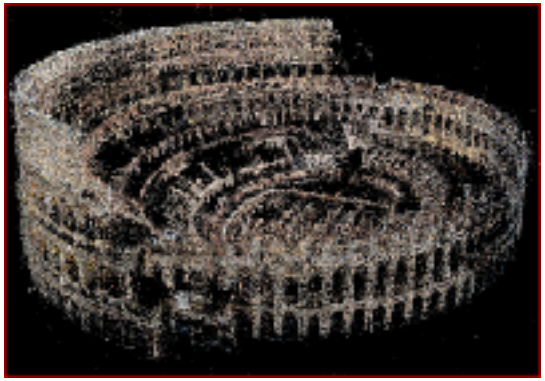


K-means in Computer Vision

- Clustering can be applied to any domain or dataset without the need for manually annotated labels.

Medical Imaging

Satellite Images



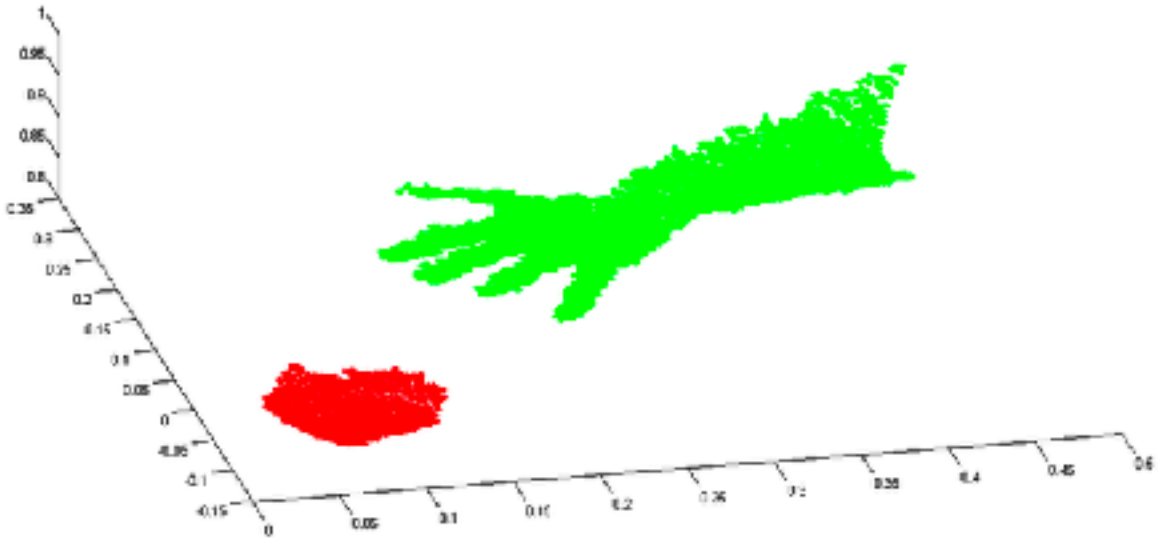
Point Clouds

K-means in Computer Vision

- K-means can be used to compress images by reducing the number of colors.

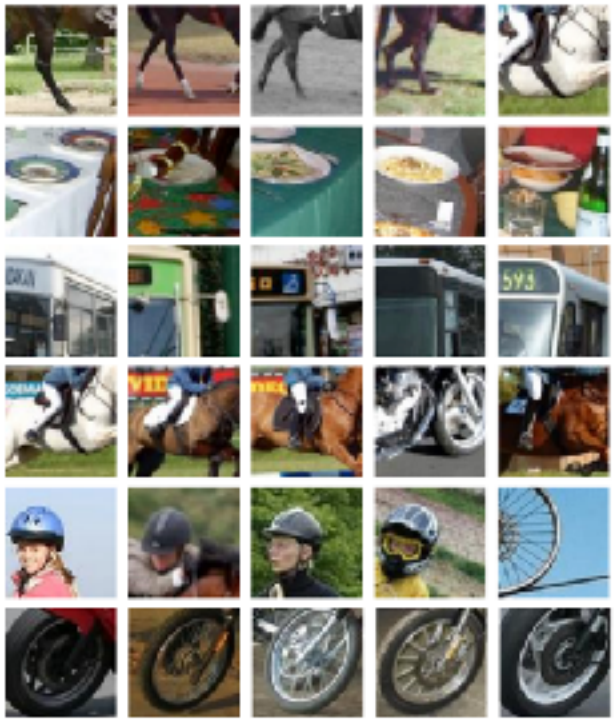
K-means in Computer Vision

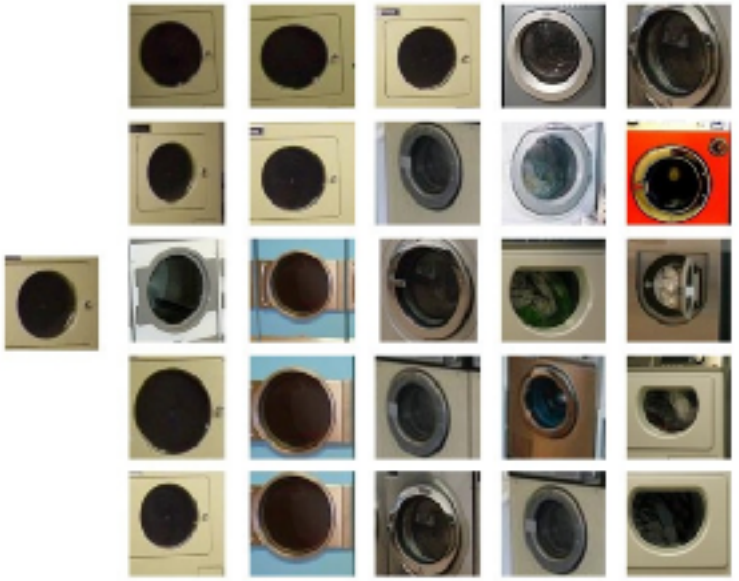
- K-means can be used to segment 3D points into objects/ object parts.



K-means in Computer Vision

- Pre-deep learning, k-means was often used on top of hand-crafted features.

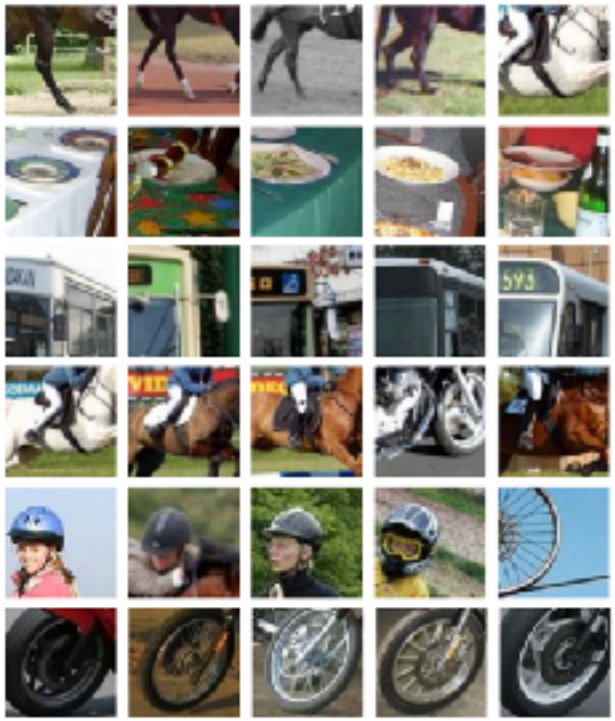


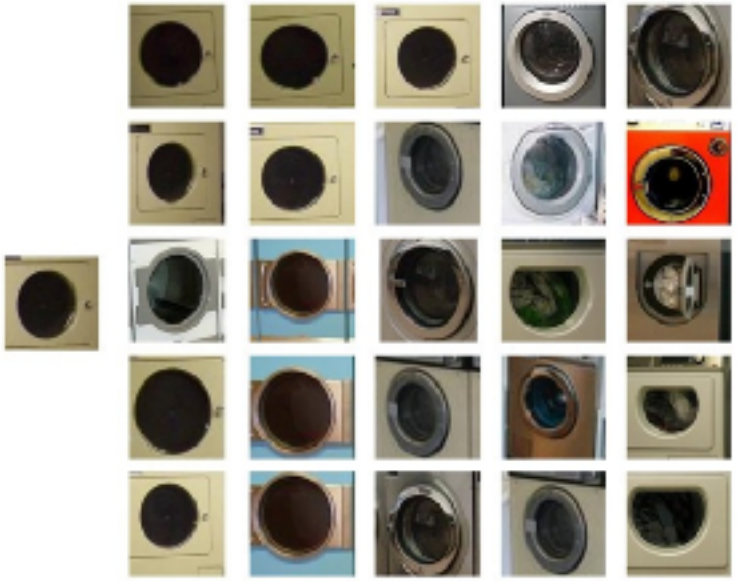


[1] Singh et al., ”Unsupervised Discovery of Mid-Level Discriminative Patches,” ECCV 2012.
[2] Juneja et al. “Blocks that Shout: Distinctive Parts for Scene Classification,” CVPR 2013.

K-means in Computer Vision

- Pre-deep learning, k-means was often used on top of hand-crafted features.

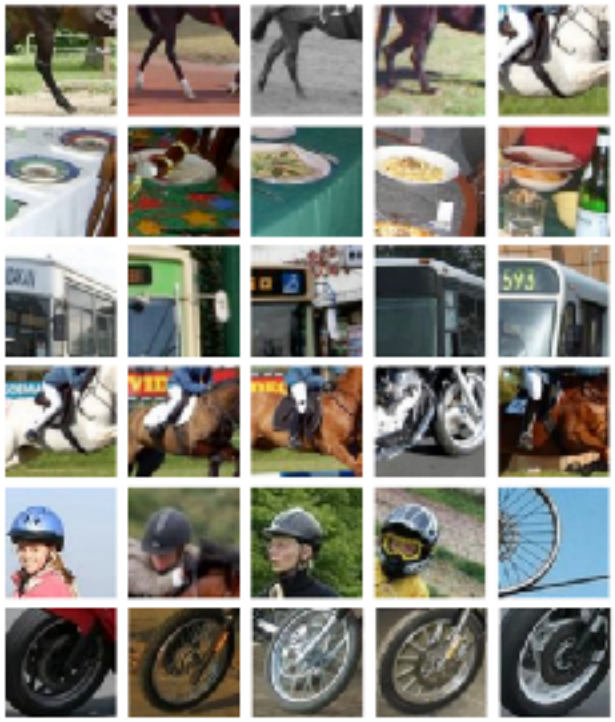


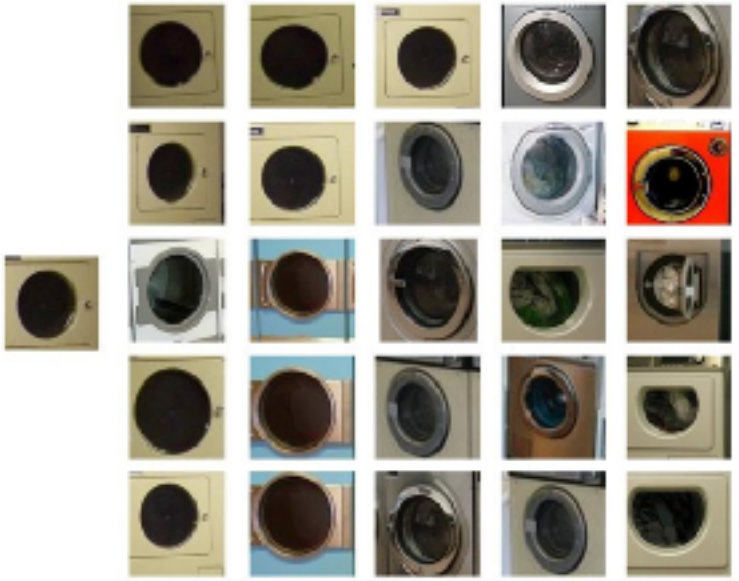


Mostly on top of fixed hand-crafted visual features!

K-means in Computer Vision

- Pre-deep learning, k-means was often used on top of hand-crafted features.





Can we use clustering for unsupervised CNN feature learning?

K-means Objective

- K-means minimize the total squared error between the training samples and their representative cluster centroids.

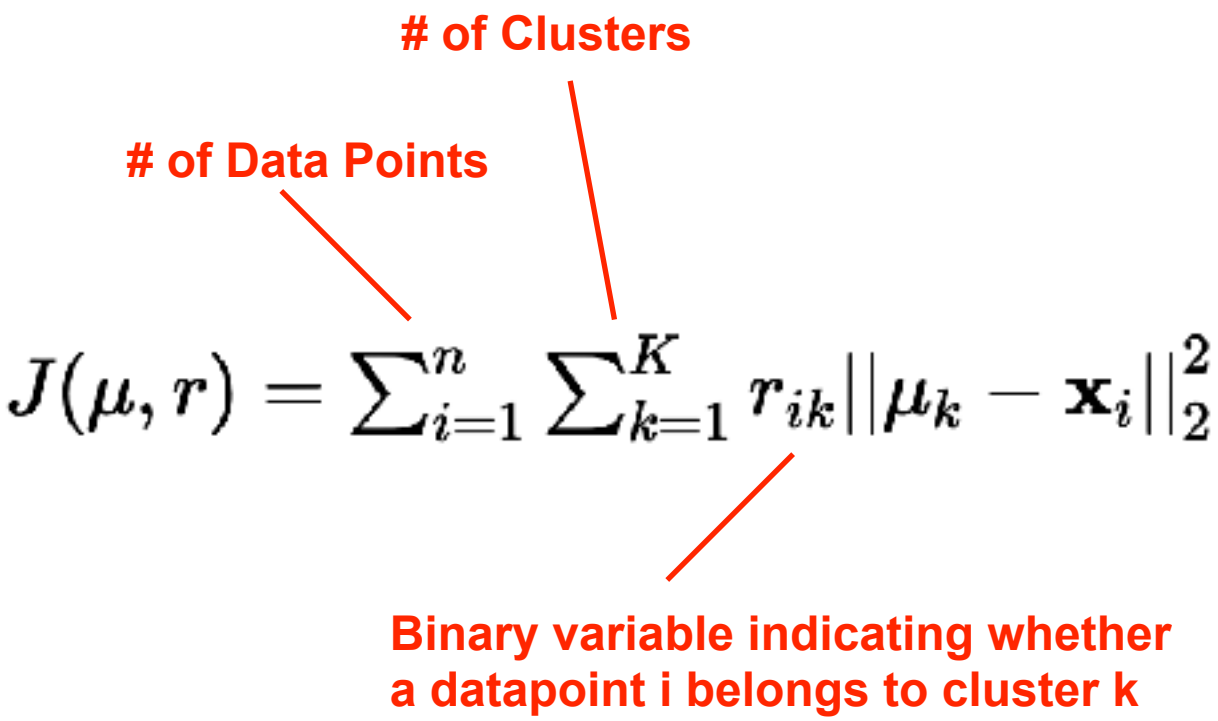
# of Clusters

# of Data Points

$$J(\mu, r) = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} ||\mu_k - \mathbf{x}_i||_2^2$$

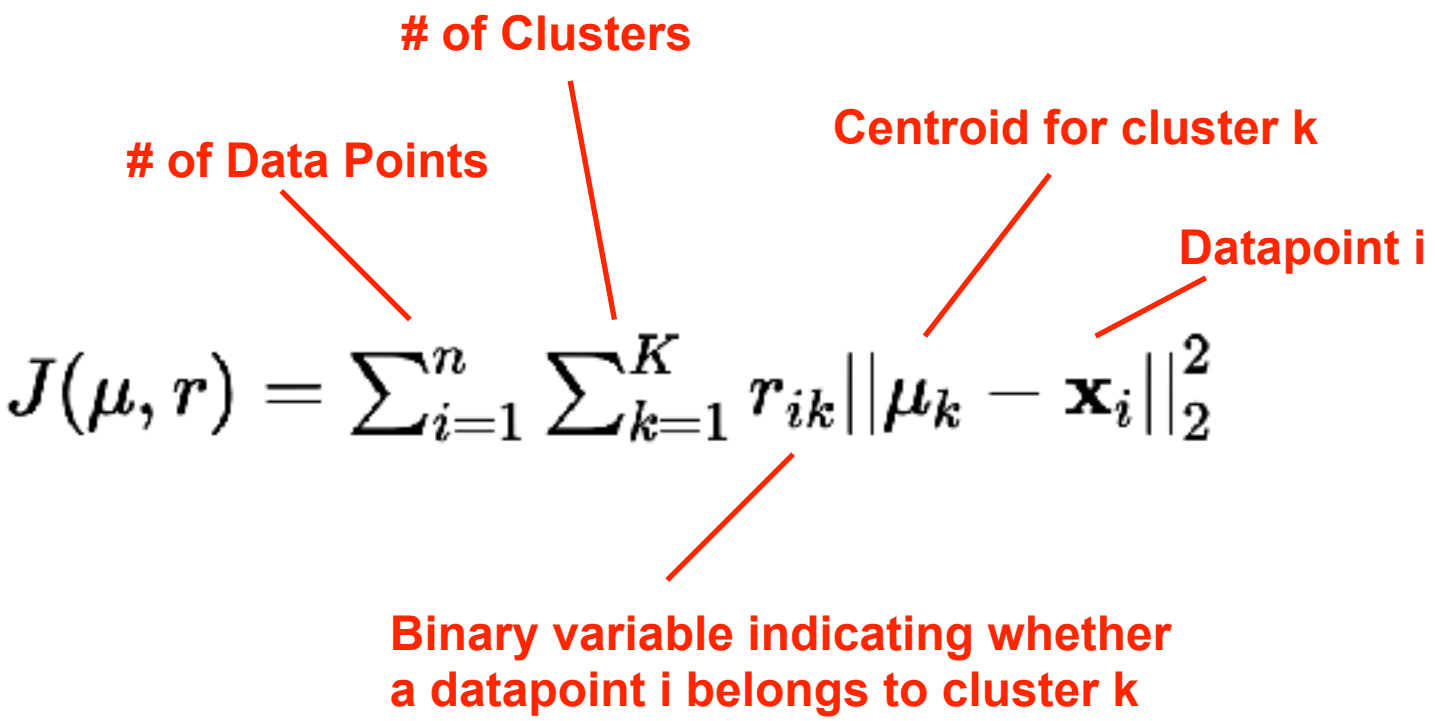
K-means Objective

- K-means minimize the total squared error between the training samples and their representative cluster centroids.



K-means Objective

- K-means minimize the total squared error between the training samples and their representative cluster centroids.



K-means Objective

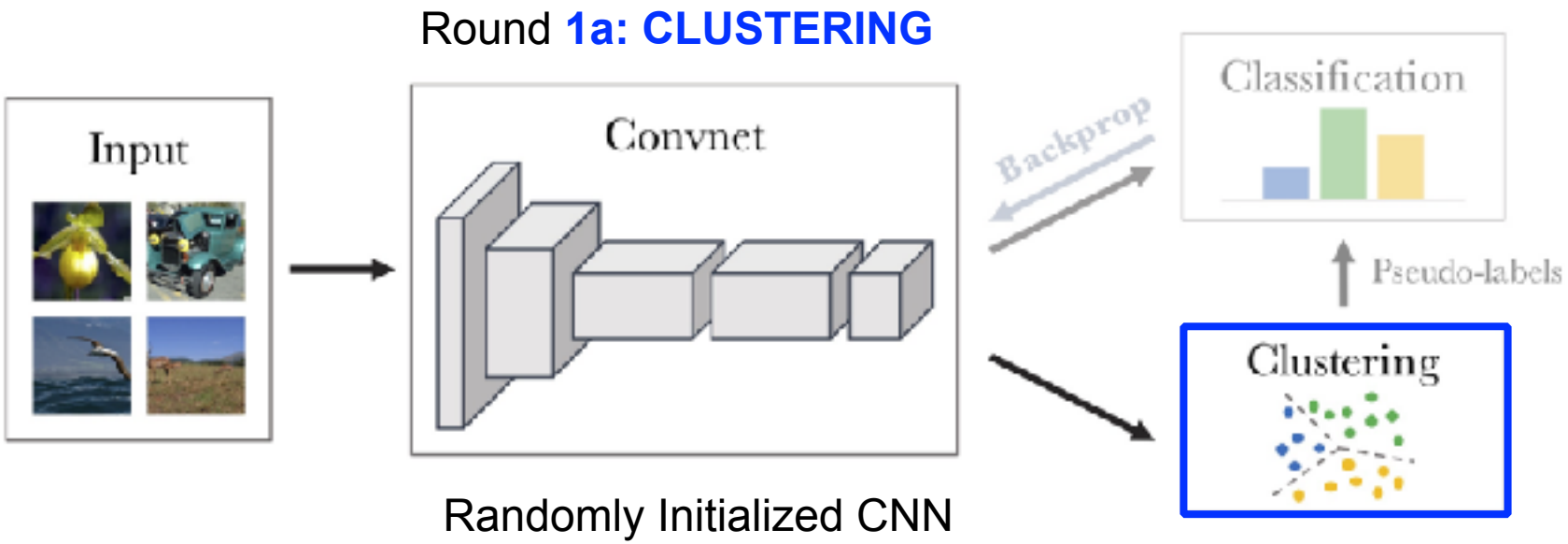
- K-means minimize the total squared error between the training samples and their representative cluster centroids.

$$J(\mu, r) = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} ||\mu_k - \mathbf{x}_i||_2^2$$

Why can't we use the K-means objective jointly with CNN feature learning (i.e., in an end-to-end manner)?

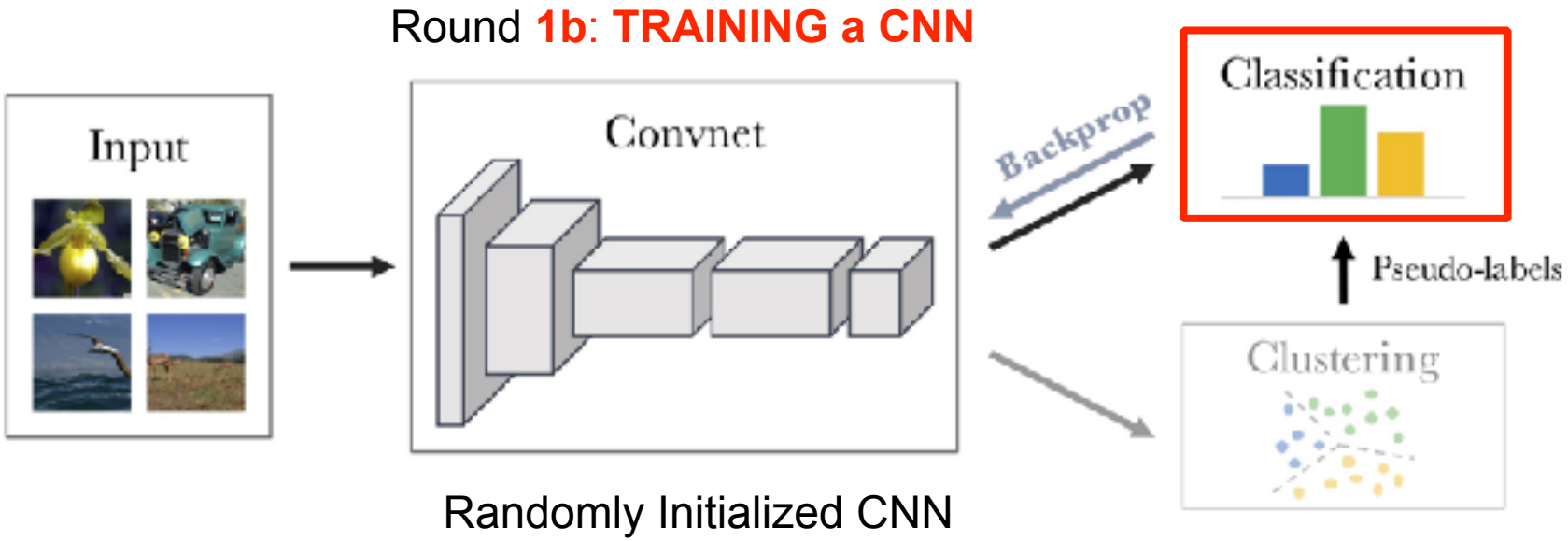
Alternating Training

- The method (1) iteratively clusters deep features and then (2) re-trains the CNN using cluster assignments as pseudo-labels.



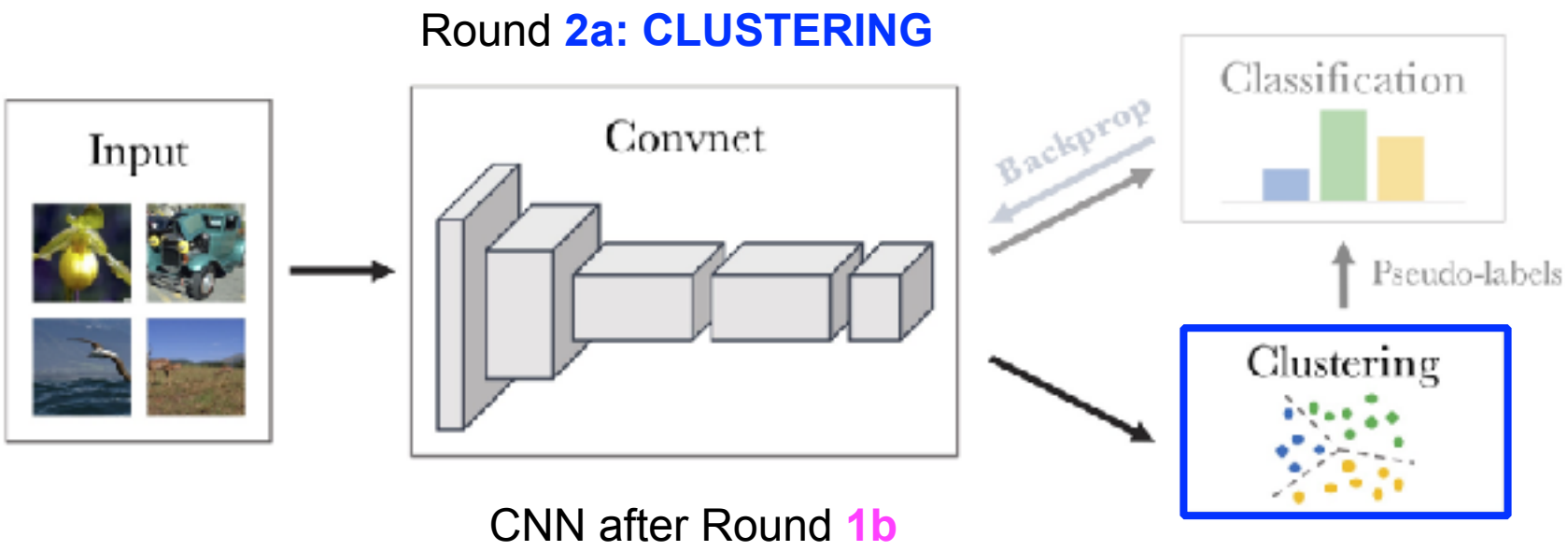
Alternating Training

- The method (1) iteratively clusters deep features and then (2) re-trains the CNN using cluster assignments as pseudo-labels.



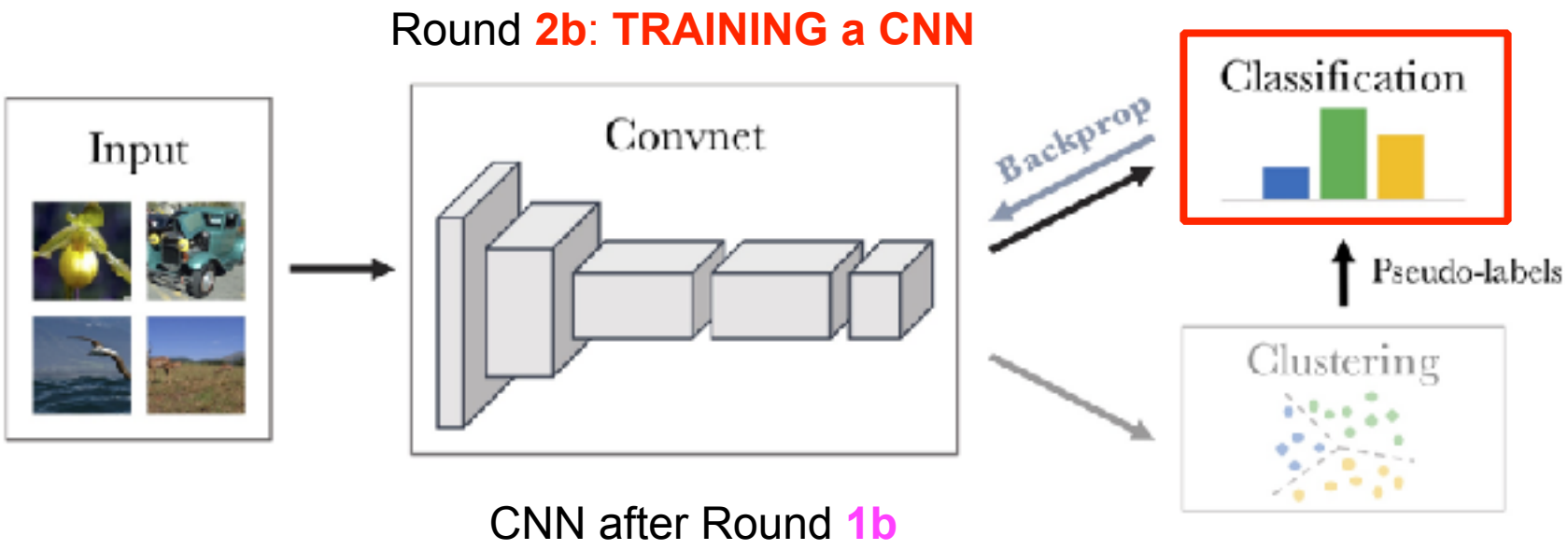
Alternating Training

- The method (1) iteratively clusters deep features and then (2) re-trains the CNN using cluster assignments as pseudo-labels.



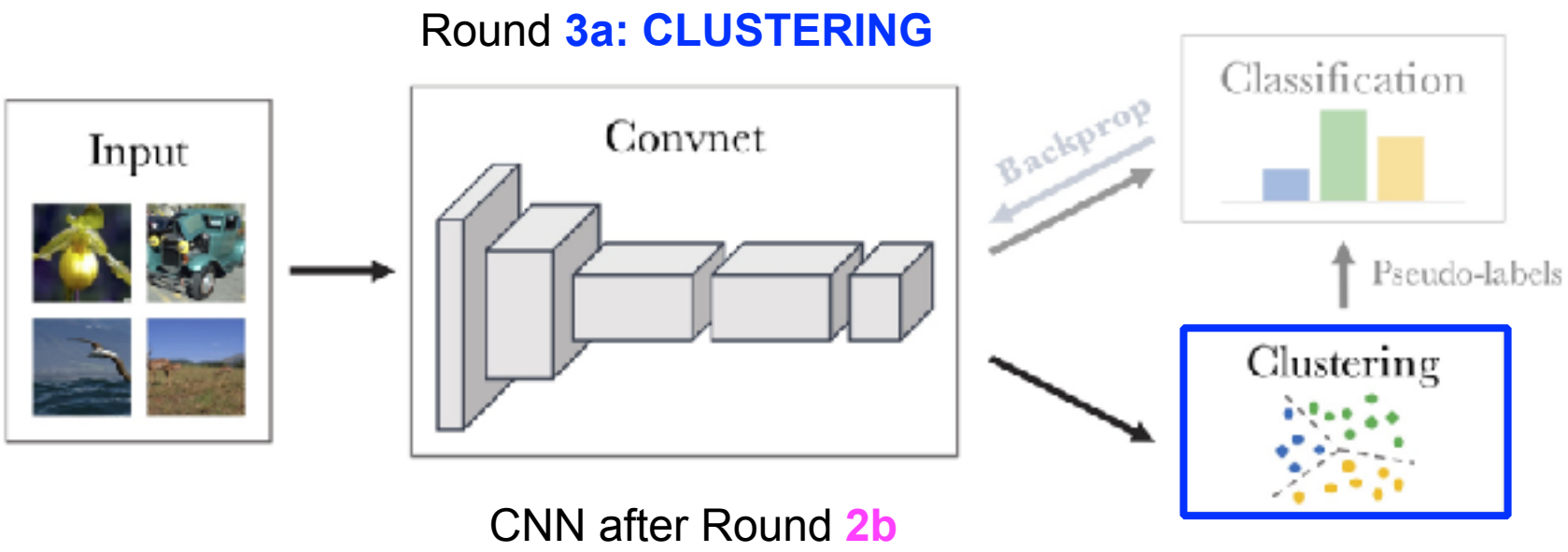
Alternating Training

- The method (1) iteratively clusters deep features and then (2) re-trains the CNN using cluster assignments as pseudo-labels.



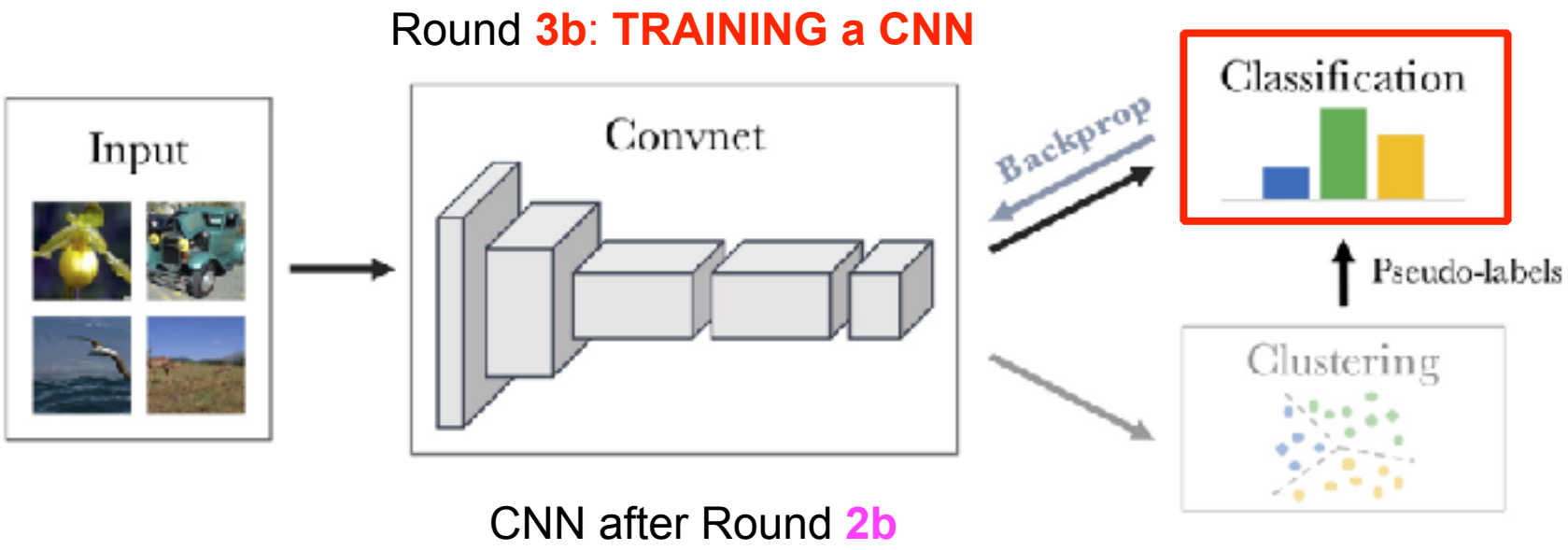
Alternating Training

- The method (1) iteratively clusters deep features and then (2) re-trains the CNN using cluster assignments as pseudo-labels.



Alternating Training

- The method (1) iteratively clusters deep features and then (2) re-trains the CNN using cluster assignments as pseudo-labels.



Classification using Pseudo Labels

- The CNN is trained using a standard cross-entropy loss.
- Cluster IDs are used as pseudo ground truth labels.

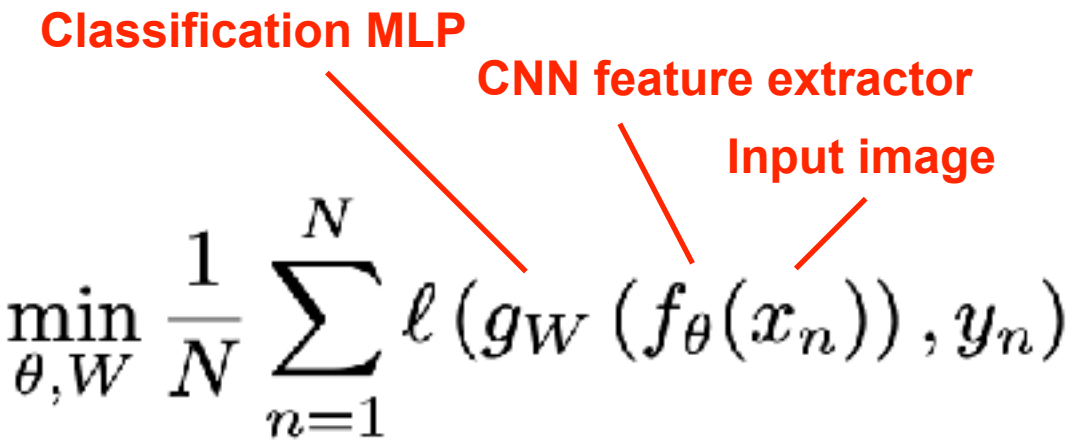
CNN feature extractor

Input image

$$\min_{\theta,W} \frac{1}{N} \sum_{n=1}^{N} \ell\left(g_W\left(f_\theta(x_n)\right), y_n\right)$$

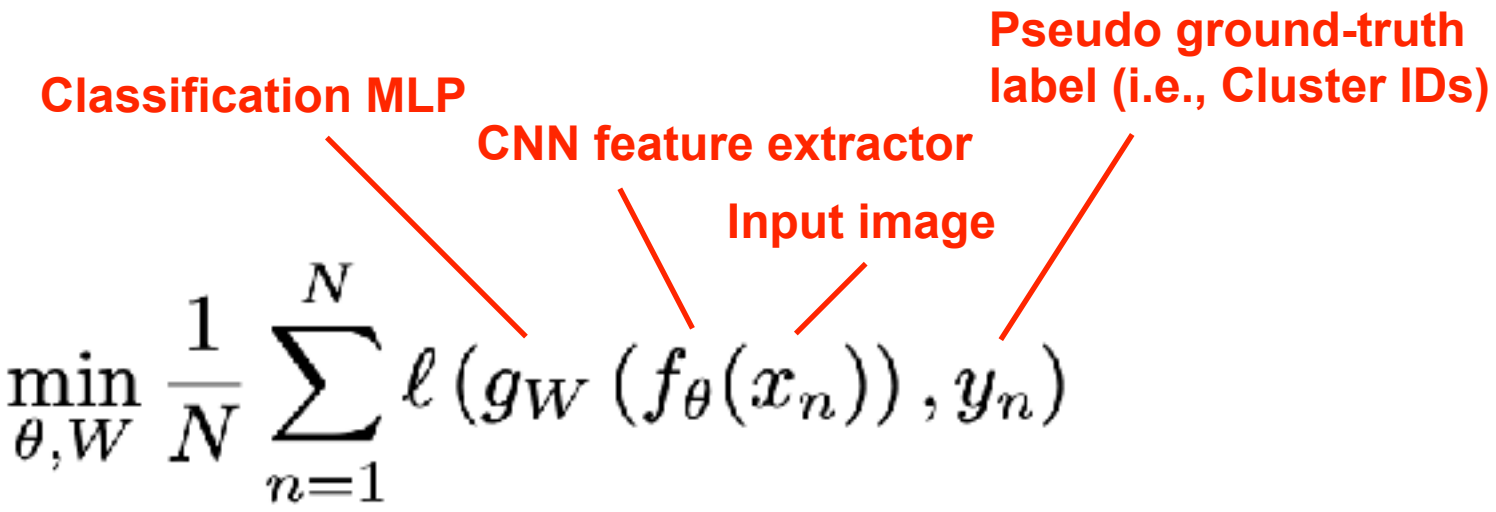
Classification using Pseudo Labels

- The CNN is trained using a standard cross-entropy loss.
- Cluster IDs are used as pseudo ground truth labels.



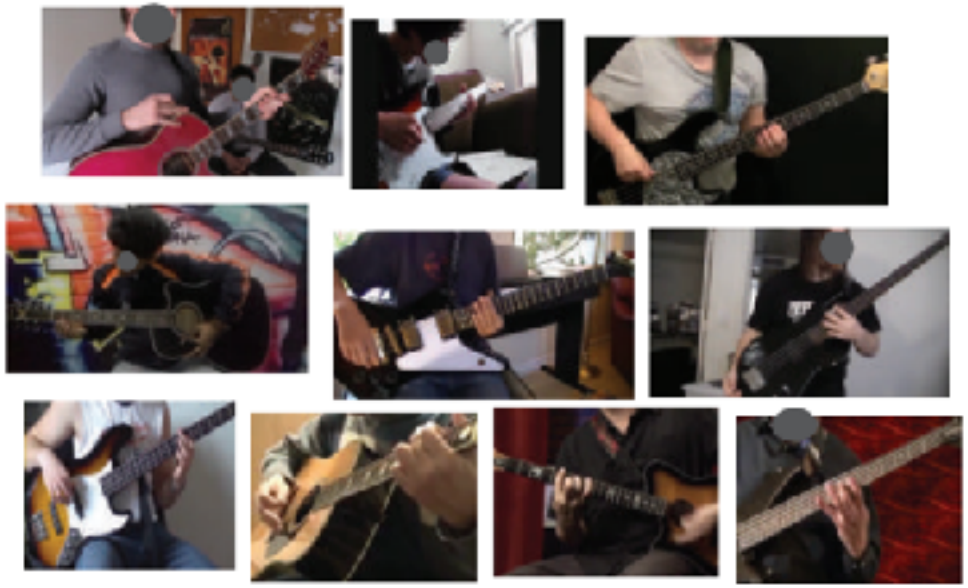
Classification using Pseudo Labels

- The CNN is trained using a standard cross-entropy loss.
- Cluster IDs are used as pseudo ground truth labels.



Why Does This Work?

- What do the clusters actually represent?



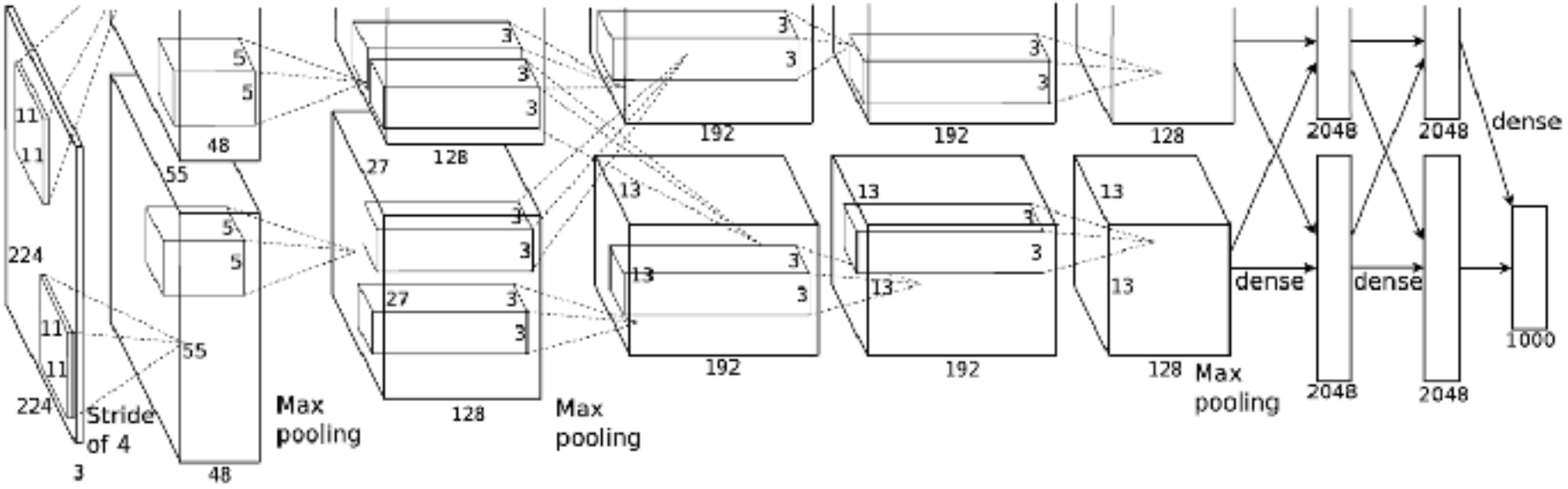
A guitar cluster

A scuba diving/snorkeling cluster

Alwassel et al., ”Self-Supervised Learning by Cross-Modal Audio-Video Clustering,” ICCV 2019

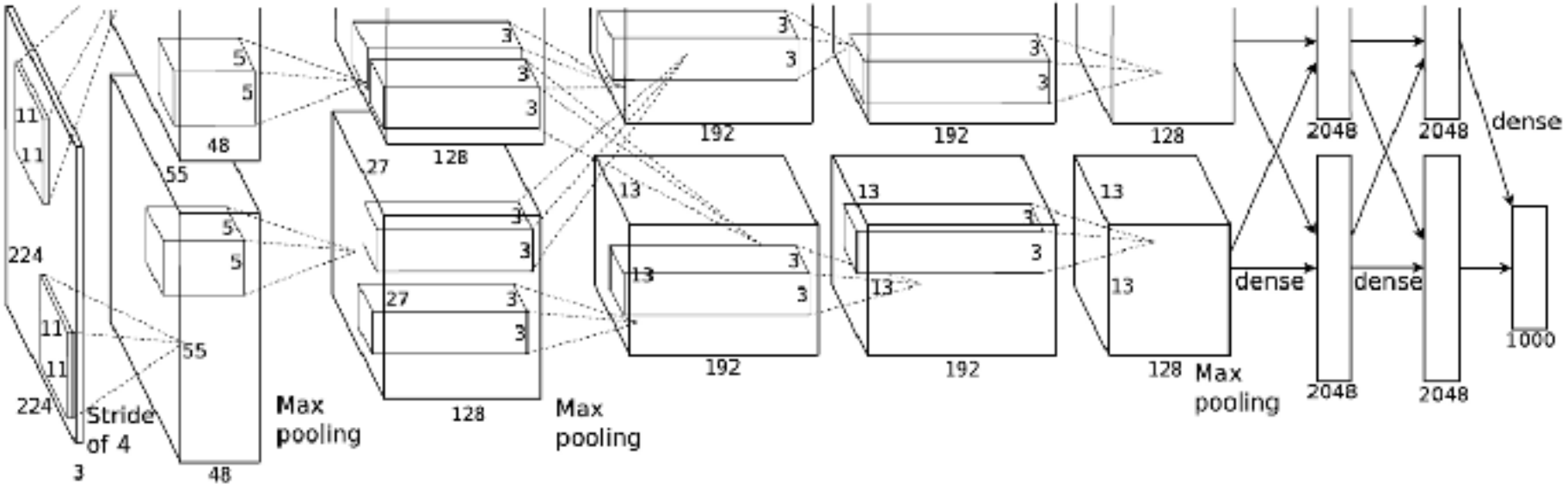
Why Does This Work?

- Initialize AlexNet with random weights.
- Train a 2-layer MLP on its frozen last layer features for Imagenet classification.



Why Does This Work?

- Initialize AlexNet with random weights.
- Train a 2-layer MLP on its frozen last layer features for Imagenet classification.



The performance of such a classifier is 12%, which is far above the 0.1% chance accuracy.

Why Does This Work?

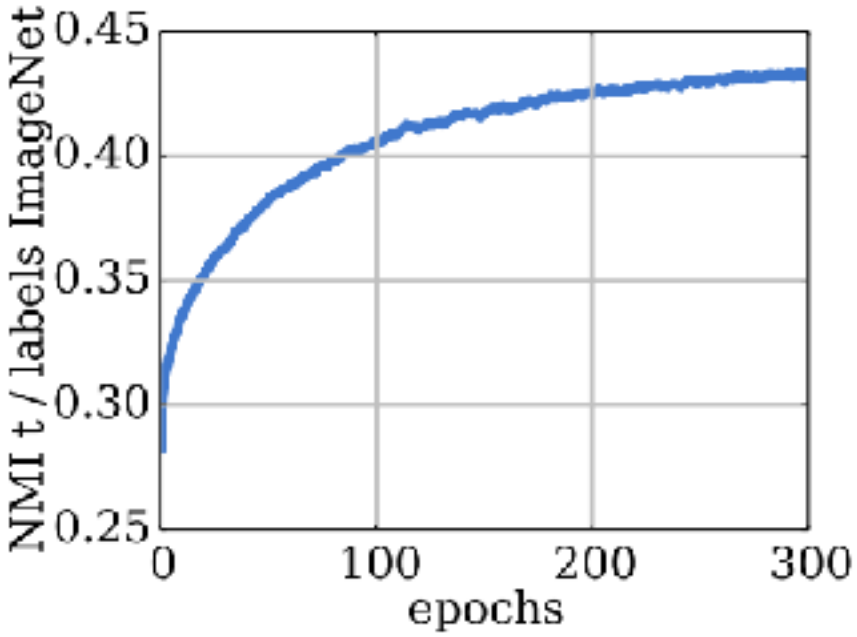
- Convolution has a strong inductive prior, which is useful even when the convolutional kernel is initialized randomly.

Input Image

Feature Maps of a Randomly Initialized CNN

Quality of Clusters

- Normalized Mutual Information is used to assess the quality of the obtained clusters.



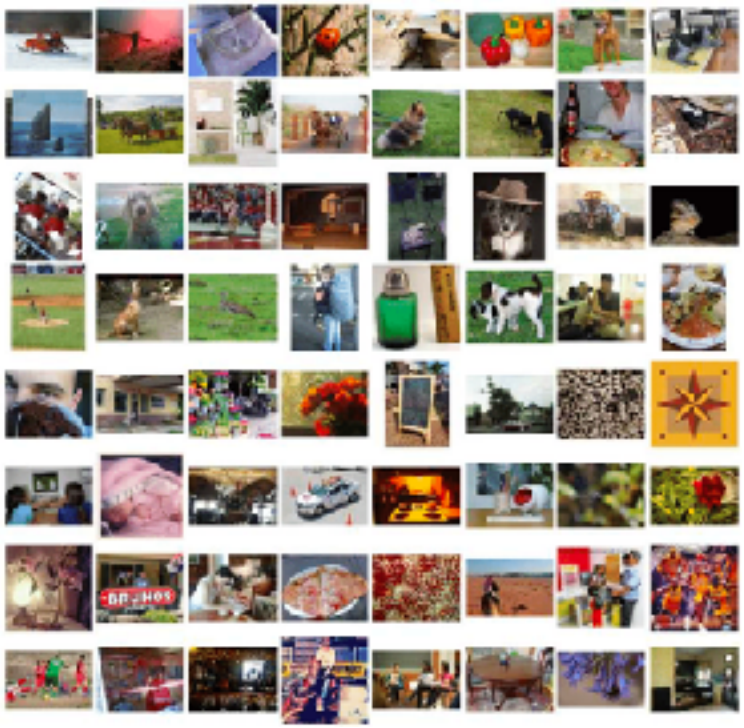
(a) Clustering quality

Experimental Setup

- Self-supervised pretraining is conducted on the ImageNet dataset (without using any GT labels).

Experimental Setup

- A linear classifier is trained on top of frozen CNN features and evaluated on the Imagenet and Places datasets.



a) Imagenet

b) Places

Linear Evaluation on Imagenet and Places

- A linear classifier is trained on top of frozen CNN features and evaluated on the Imagenet and Places datasets.

	ImageNet					Places				
Method	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels	–	–	–	–	–	22.1	35.1	40.2	43.3	44.6
ImageNet labels	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
Pathak *et al.* [38]	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
Doersch *et al.* [25]	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Zhang *et al.* [28]	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
Donahue *et al.* [20]	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
Noroozi and Favaro [26]	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Noroozi *et al.* [45]	18.0	**30.6**	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
Zhang *et al.* [43]	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
DeepCluster	12.9	29.2	**38.2**	**39.8**	**36.1**	18.6	30.8	**37.0**	**37.5**	**33.1**

Substantial boost over state-of-the-art

Linear Evaluation on Imagenet and Places

- A linear classifier is trained on top of frozen CNN features and evaluated on the Imagenet and Places datasets.

	ImageNet					Places				
Method	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels	–	–	–	–	–	22.1	35.1	40.2	43.3	44.6
ImageNet labels	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
Pathak *et al.* [38]	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
Doersch *et al.* [25]	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Zhang *et al.* [28]	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
Donahue *et al.* [20]	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
Noroozi and Favaro [26]	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Noroozi *et al.* [45]	18.0	**30.6**	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
Zhang *et al.* [43]	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
DeepCluster	12.9	29.2	**38.2**	**39.8**	**36.1**	18.6	30.8	**37.0**	**37.5**	**33.1**

Still lagging significantly behind a supervised baseline

Linear Evaluation on Imagenet and Places

- A linear classifier is trained on top of frozen CNN features and evaluated on the Imagenet and Places datasets.

	ImageNet					Places				
Method	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels	–	–	–	–	–	22.1	35.1	40.2	43.3	44.6
ImageNet labels	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
Pathak *et al.* [38]	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
Doersch *et al.* [25]	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Zhang *et al.* [28]	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
Donahue *et al.* [20]	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
Noroozi and Favaro [26]	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Noroozi *et al.* [45]	18.0	**30.6**	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
Zhang *et al.* [43]	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
DeepCluster	12.9	29.2	**38.2**	**39.8**	**36.1**	18.6	30.8	**37.0**	**37.5**	**33.1**

Comparable performance with supervised Imagenet features

Transfer to Detection & Segmentation

- Comparison of the proposed approach to state-of-the-art on classification, detection and segmentation on Pascal VOC.

	Classification		Detection		Segmentation	
Method	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
ImageNet labels	78.9	79.9	–	56.8	–	48.0
Random-rgb	33.2	57.0	22.2	44.5	15.2	30.1
Random-sobel	29.0	61.9	18.9	47.9	13.0	32.0
Pathak *et al.* [38]	34.6	56.5	–	44.5	–	29.7
Donahue *et al.* [20]*	52.3	60.1	–	46.9	–	35.2
Pathak *et al.* [27]	–	61.0	–	52.2	–	–
Owens *et al.* [44]*	52.3	61.3	–	–	–	–
Wang and Gupta [29]*	55.6	63.1	32.8†	47.2	26.0†	35.4†
Doersch *et al.* [25]*	55.1	65.3	–	51.1	–	–
Bojanowski and Joulin [19]*	56.7	65.3	33.7†	49.4	26.7†	37.1†
Zhang *et al.* [28]*	61.5	65.9	43.4†	46.9	35.8†	35.6
Zhang *et al.* [43]*	63.0	67.1	–	46.7	–	36.0
Noroozi and Favaro [26]	–	67.6	–	53.2	–	37.6
Noroozi *et al.* [45]	–	67.7	–	51.4	–	36.6
DeepCluster	**70.4**	**73.7**	**51.4**	**55.4**	**43.2**	**45.1**

DeepCluster outperforms all prior approaches in all evaluation regimes.

Transfer to Detection & Segmentation

- Comparison of the proposed approach to state-of-the-art on classification, detection and segmentation on Pascal VOC.

	Classification		Detection		Segmentation	
Method	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
ImageNet labels	78.9	79.9	–	56.8	–	48.0
Random-rgb	33.2	57.0	22.2	44.5	15.2	30.1
Random-sobel	29.0	61.9	18.9	47.9	13.0	32.0
Pathak *et al.* [38]	34.6	56.5	–	44.5	–	29.7
Donahue *et al.* [20]*	52.3	60.1	–	46.9	–	35.2
Pathak *et al.* [27]	–	61.0	–	52.2	–	–
Owens *et al.* [44]*	52.3	61.3	–	–	–	–
Wang and Gupta [29]*	55.6	63.1	32.8†	47.2	26.0†	35.4†
Doersch *et al.* [25]*	55.1	65.3	–	51.1	–	–
Bojanowski and Joulin [19]*	56.7	65.3	33.7†	49.4	26.7†	37.1†
Zhang *et al.* [28]*	61.5	65.9	43.4†	46.9	35.8†	35.6
Zhang *et al.* [43]*	63.0	67.1	–	46.7	–	36.0
Noroozi and Favaro [26]	–	67.6	–	53.2	–	37.6
Noroozi *et al.* [45]	–	67.7	–	51.4	–	36.6
DeepCluster	**70.4**	**73.7**	**51.4**	**55.4**	**43.2**	**45.1**

Fine-tuned random baselines perform almost as well as prior SSL methods

Transfer to Detection & Segmentation

- Comparison of the proposed approach to state-of-the-art on classification, detection and segmentation on Pascal VOC.

	Classification		Detection		Segmentation	
Method	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
ImageNet labels	78.9	79.9	–	56.8	–	48.0
Random-rgb	33.2	57.0	22.2	44.5	15.2	30.1
Random-sobel	29.0	61.9	18.9	47.9	13.0	32.0
Pathak *et al.* [38]	34.6	56.5	–	44.5	–	29.7
Donahue *et al.* [20]*	52.3	60.1	–	46.9	–	35.2
Pathak *et al.* [27]	–	61.0	–	52.2	–	–
Owens *et al.* [44]*	52.3	61.3	–	–	–	–
Wang and Gupta [29]*	55.6	63.1	32.8†	47.2	26.0†	35.4†
Doersch *et al.* [25]*	55.1	65.3	–	51.1	–	–
Bojanowski and Joulin [19]*	56.7	65.3	33.7†	49.4	26.7†	37.1†
Zhang *et al.* [28]*	61.5	65.9	43.4†	46.9	35.8†	35.6
Zhang *et al.* [43]*	63.0	67.1	–	46.7	–	36.0
Noroozi and Favaro [26]	–	67.6	–	53.2	–	37.6
Noroozi *et al.* [45]	–	67.7	–	51.4	–	36.6
DeepCluster	**70.4**	**73.7**	**51.4**	**55.4**	**43.2**	**45.1**

DeepCluster outperforms other methods by a large margin if only fc6-8 layers are trained.

Transfer to Detection & Segmentation

- Comparison of the proposed approach to state-of-the-art on classification, detection and segmentation on Pascal VOC.

	Classification		Detection		Segmentation	
Method	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
ImageNet labels	78.9	79.9	–	56.8	–	48.0
Random-rgb	33.2	57.0	22.2	44.5	15.2	30.1
Random-sobel	29.0	61.9	18.9	47.9	13.0	32.0
Pathak *et al.* [38]	34.6	56.5	–	44.5	–	29.7
Donahue *et al.* [20]*	52.3	60.1	–	46.9	–	35.2
Pathak *et al.* [27]	–	61.0	–	52.2	–	–
Owens *et al.* [44]*	52.3	61.3	–	–	–	–
Wang and Gupta [29]*	55.6	63.1	32.8†	47.2	26.0†	35.4†
Doersch *et al.* [25]*	55.1	65.3	–	51.1	–	–
Bojanowski and Joulin [19]*	56.7	65.3	33.7†	49.4	26.7†	37.1†
Zhang *et al.* [28]*	61.5	65.9	43.4†	46.9	35.8†	35.6
Zhang *et al.* [43]*	63.0	67.1	–	46.7	–	36.0
Noroozi and Favaro [26]	–	67.6	–	53.2	–	37.6
Noroozi *et al.* [45]	–	67.7	–	51.4	–	36.6
DeepCluster	**70.4**	**73.7**	**51.4**	**55.4**	**43.2**	**45.1**

The gap between DeepCluster and supervised baselines is smaller on detection tasks.

Training on Uncurated Data

- Impact of the training set on the performance of DeepCluster measured on the Pascal VOC transfer tasks.

		Classification		Detection		Segmentation	
Method	Training set	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
Best competitor	ImageNet	63.0	67.7	43.4†	53.2	35.8†	37.7
DeepCluster	ImageNet	72.0	73.7	51.4	55.4	43.2	45.1
DeepCluster	YFCC100M	67.3	69.3	45.6	53.0	39.2	42.2

Training on YFCC100M produces slightly lower performance.

Training on Uncurated Data

- Impact of the training set on the performance of DeepCluster measured on the Pascal VOC transfer tasks.

		Classification		Detection		Segmentation	
Method	Training set	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
Best competitor	ImageNet	63.0	67.7	43.4†	53.2	35.8†	37.7
DeepCluster	ImageNet	72.0	73.7	51.4	55.4	43.2	45.1
DeepCluster	YFCC100M	67.3	69.3	45.6	53.0	39.2	42.2

However, DeepCluster still outperforms previous SOTA by a big margin.

Using Different Architecture

- Pascal VOC 2007 object detection with the AlexNet and VGG16 architectures.

Method	AlexNet	VGG-16
ImageNet labels	56.8	67.3
Random	47.8	39.7
Doersch *et al.* [25]	51.1	61.5
Wang and Gupta [29]	47.2	60.2
Wang *et al.* [46]	–	63.2
DeepCluster	**55.4**	**65.9**

Deeper architecture leads to better results.

Using Different Architecture

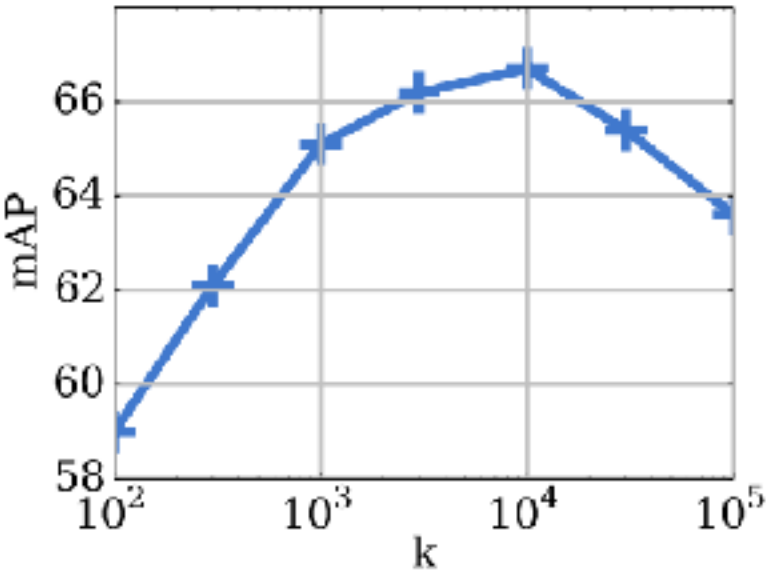
- Pascal VOC 2007 object detection with the AlexNet and VGG16 architectures.

Method	AlexNet	VGG-16
ImageNet labels	56.8	67.3
Random	47.8	39.7
Doersch *et al.* [25]	51.1	61.5
Wang and Gupta [29]	47.2	60.2
Wang *et al.* [46]	–	63.2
DeepCluster	**55.4**	**65.9**

Only ~1.4% lower than the fully supervised baseline.

Ablation on the Number of Clusters

- Ablation on the number of clusters.
- The evaluation is doe on the downstream object detection task on the PASCAL VOC dataset.



(c) Influence of k

Summary

- One of the first approaches that integrates unsupervised clustering with CNNs at large-scale.
- Large performance boost over previous SOTA and closing the gap with the supervised baselines on detection tasks.
- Thorough evaluation on multiple downstream tasks.
- The idea is simple but the implementation / actual training might be quite cumbersome.