Michał Znaleźniak, Przemysław Rola, Patryk Kaszuba, Jacek Tabor, Marek Śmieja

CONTRASTIVE LEARNING, DEEP CLUSTERING, HIERARCHICAL CLUSTERING, DATA MINING

Contrastive Hierarchical Clustering

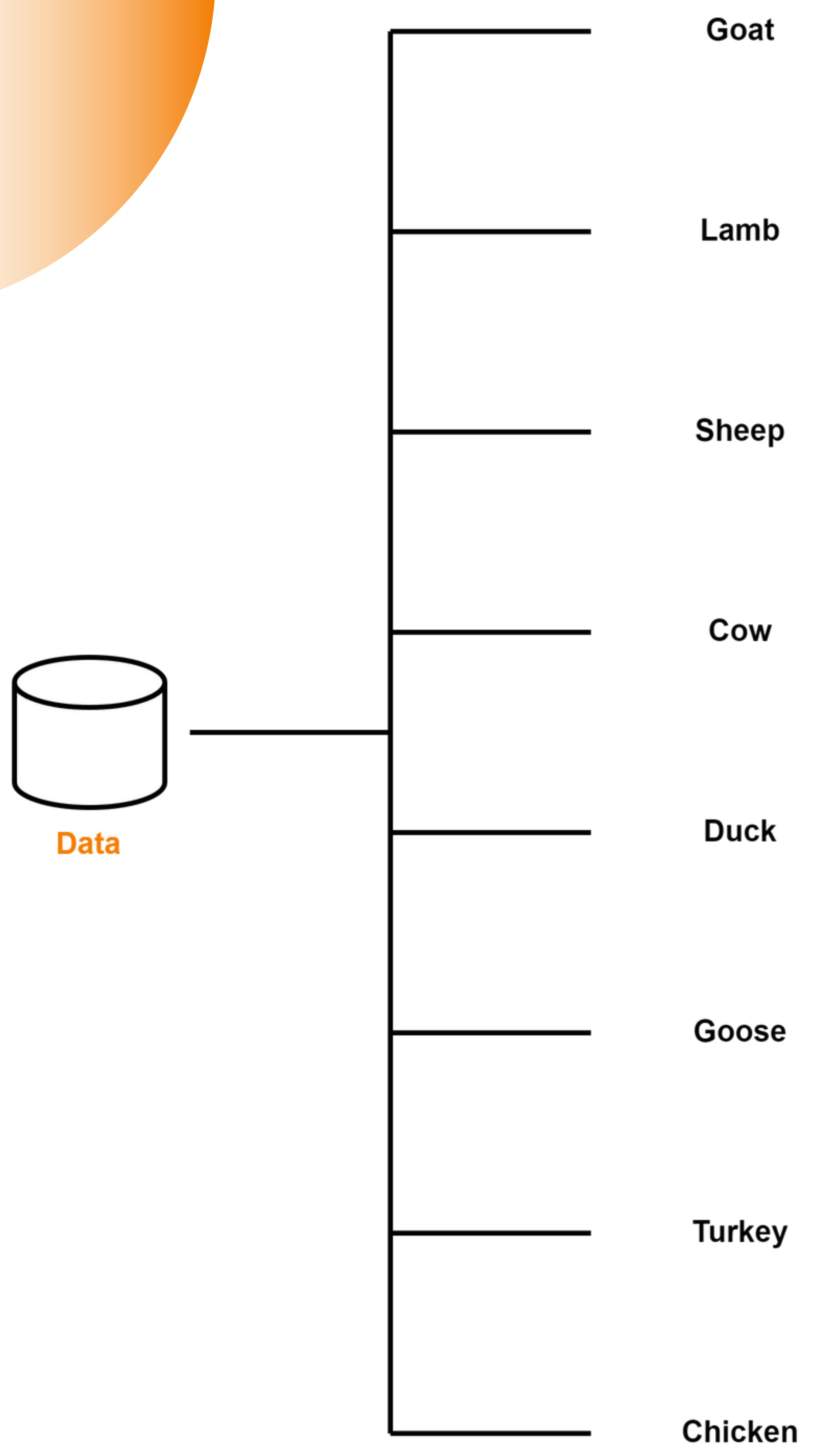
ECML PKDD September 2023

Presented by
Michał Znaleźniak

Table of Contents

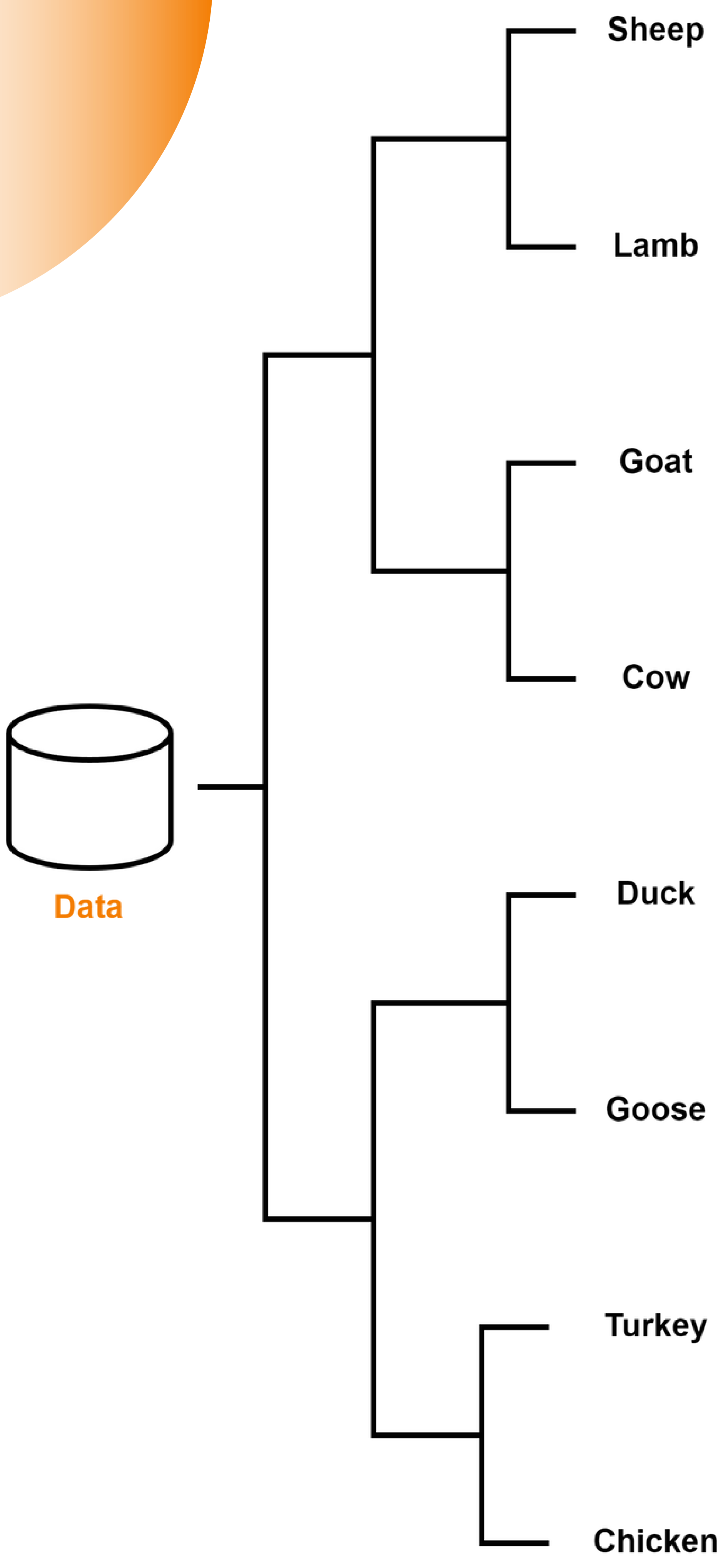
Clustering

- Helps to understand the characteristics of the dataset. It does that by looking for meaningful groups or collections in the dataset.
- Possible to distinguish two broad types:
  - **Flat** clustering
  - **Hierarchical** clustering



Clustering

- Helps to understand the characteristics of the dataset. It does that by looking for meaningful groups or collections in the dataset.
- Possible to distinguish two broad types:
  - **Flat** clustering
  - **Hierarchical** clustering



Clustering

- Helps to understand the characteristics of the dataset. It does that by looking for meaningful groups or collections in the dataset.
- Possible to distinguish two broad types:
  - **Flat** clustering
  - **Hierarchical** clustering

Contrastive Hierarchical Clustering

Observations

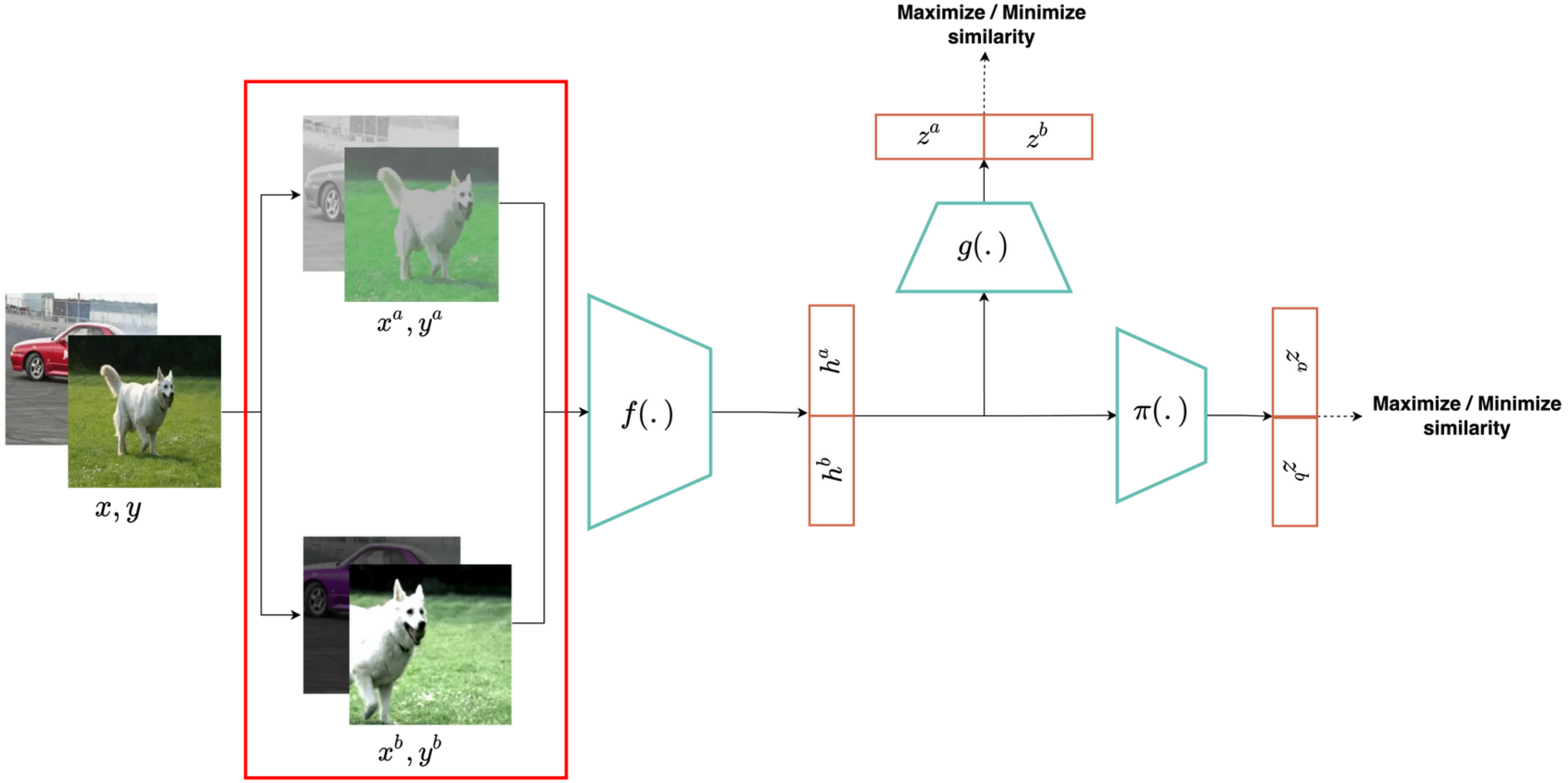
- The information contained in the flat partition is **limited**.
- Deep clustering has been **dominated** by flat models.

Goals

- Propose a new head for cluster-level representation learning which can generate **hierarchical structure of clusters**.
- Focus on analyzing the **relationship** and **similarites** between **clusters** besides just reporting the metrics.

Contrastive Hierarchical Clustering - Model Architecture

- Transforms any given data example randomly resulting in two correlated views of the same example [1].
- Augmentation list includes:
    - Resized Crop
    - Horizontal Flip
    - Color Jitter
    - Grayscale



Data Augmentation Module

[1] A Simple Framework for Contrastive Learning of Visual Representations

Contrastive Hierarchical Clustering - Model Architecture

- $f(.)$ is a backbone that computes an internal representation.
- We analyzed how backbone architecture impacts the final quality.

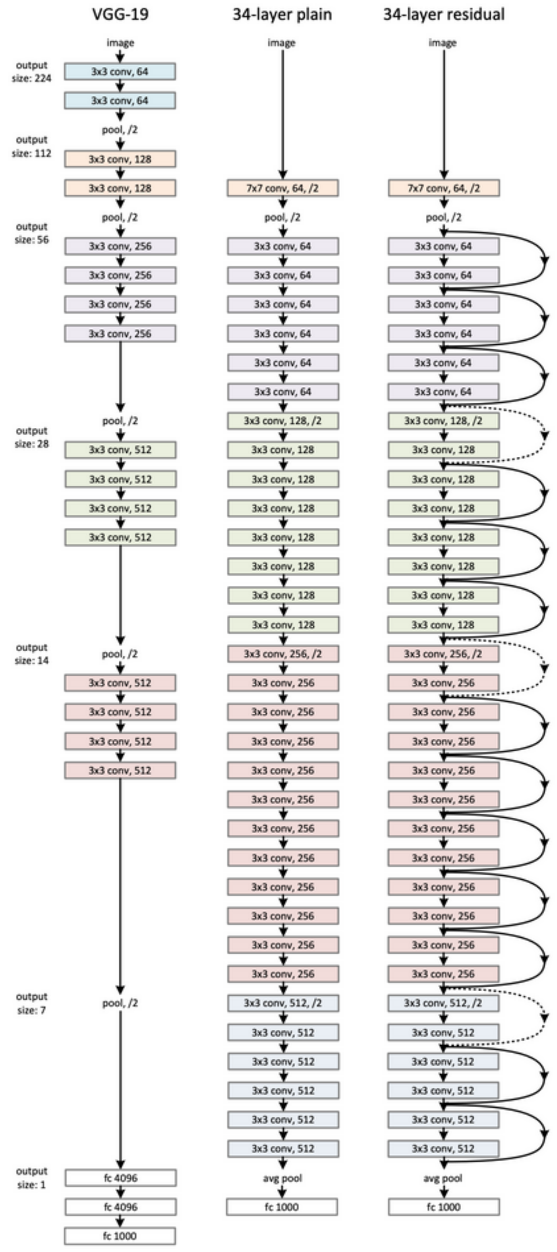
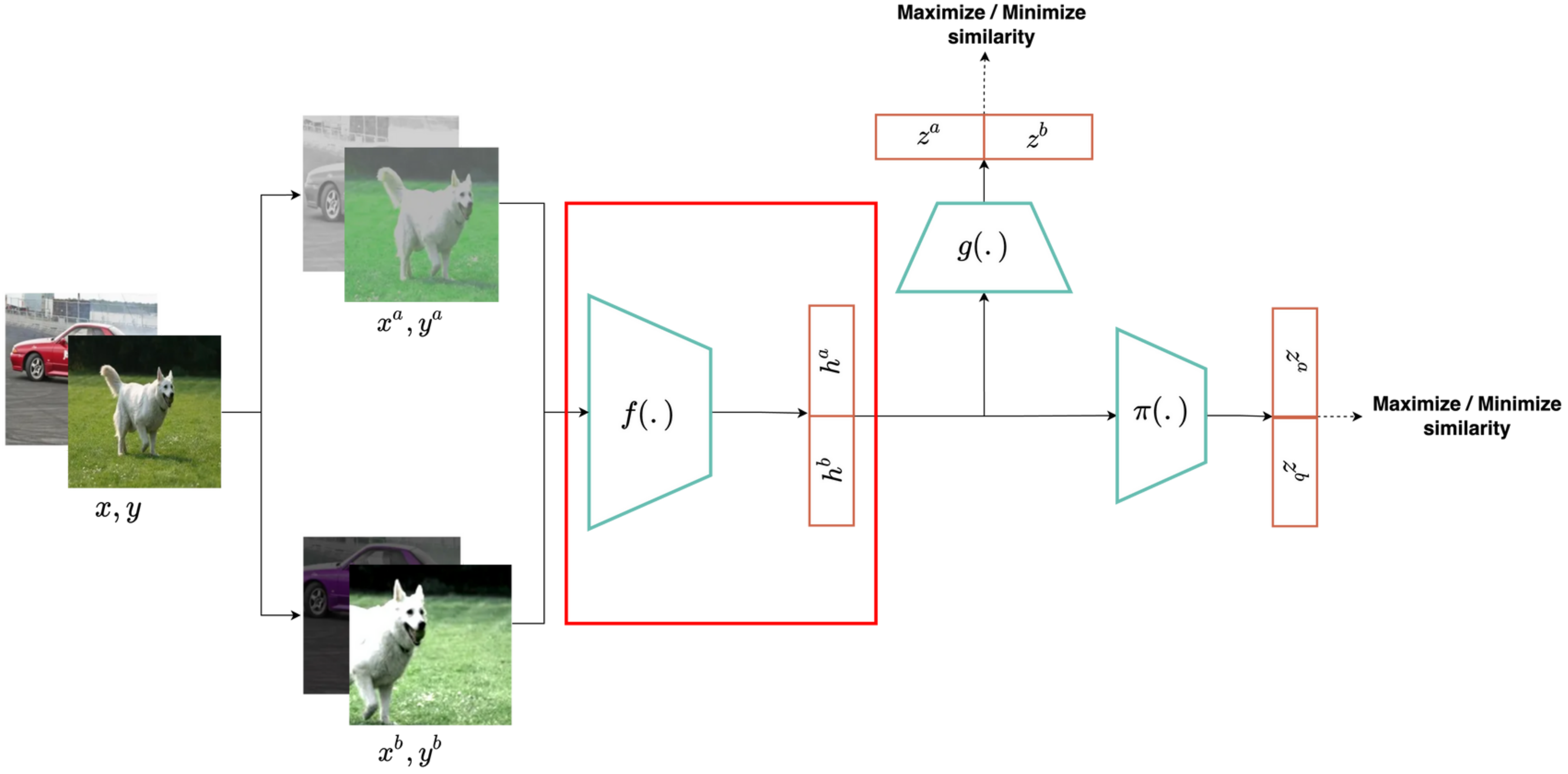


Image from Deep Residual Learning for Image Recognition



Feature Extractor

[1] A Simple Framework for Contrastive Learning of Visual Representations

Contrastive Hierarchical Clustering - Model Architecture

- $g(.)$ is a projection network (MLP) that projects representation into latent space.
- We minimize / maximize similarity between differently augmented views with **NT-Xent loss [1]** in latent space.

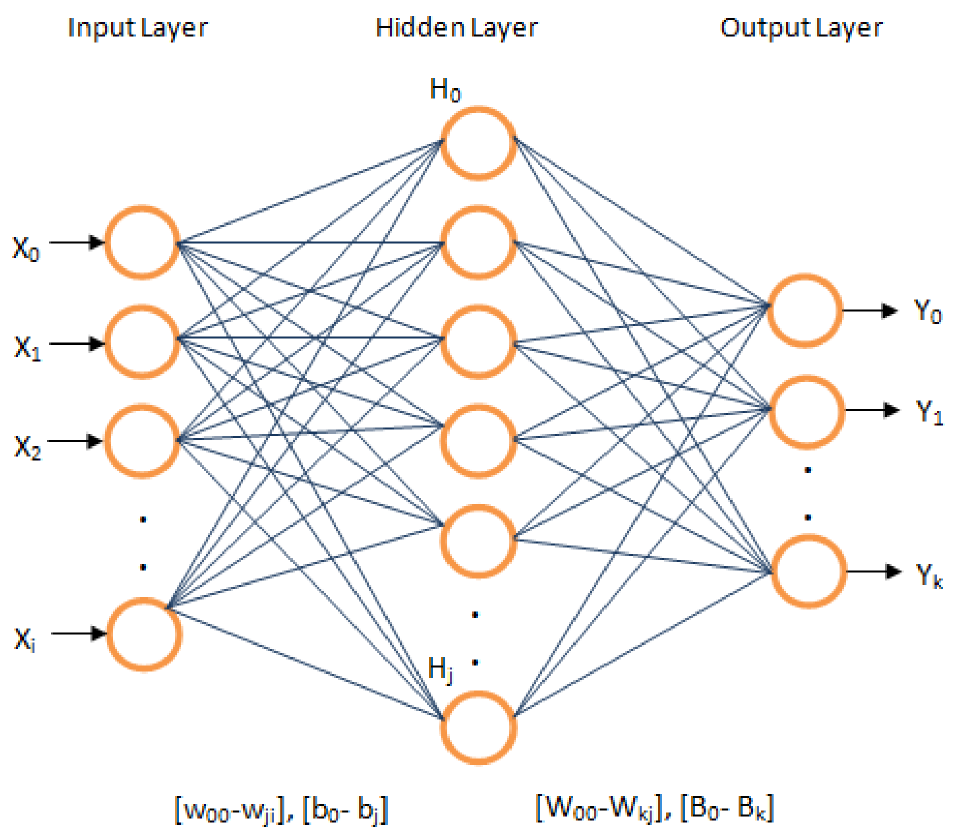
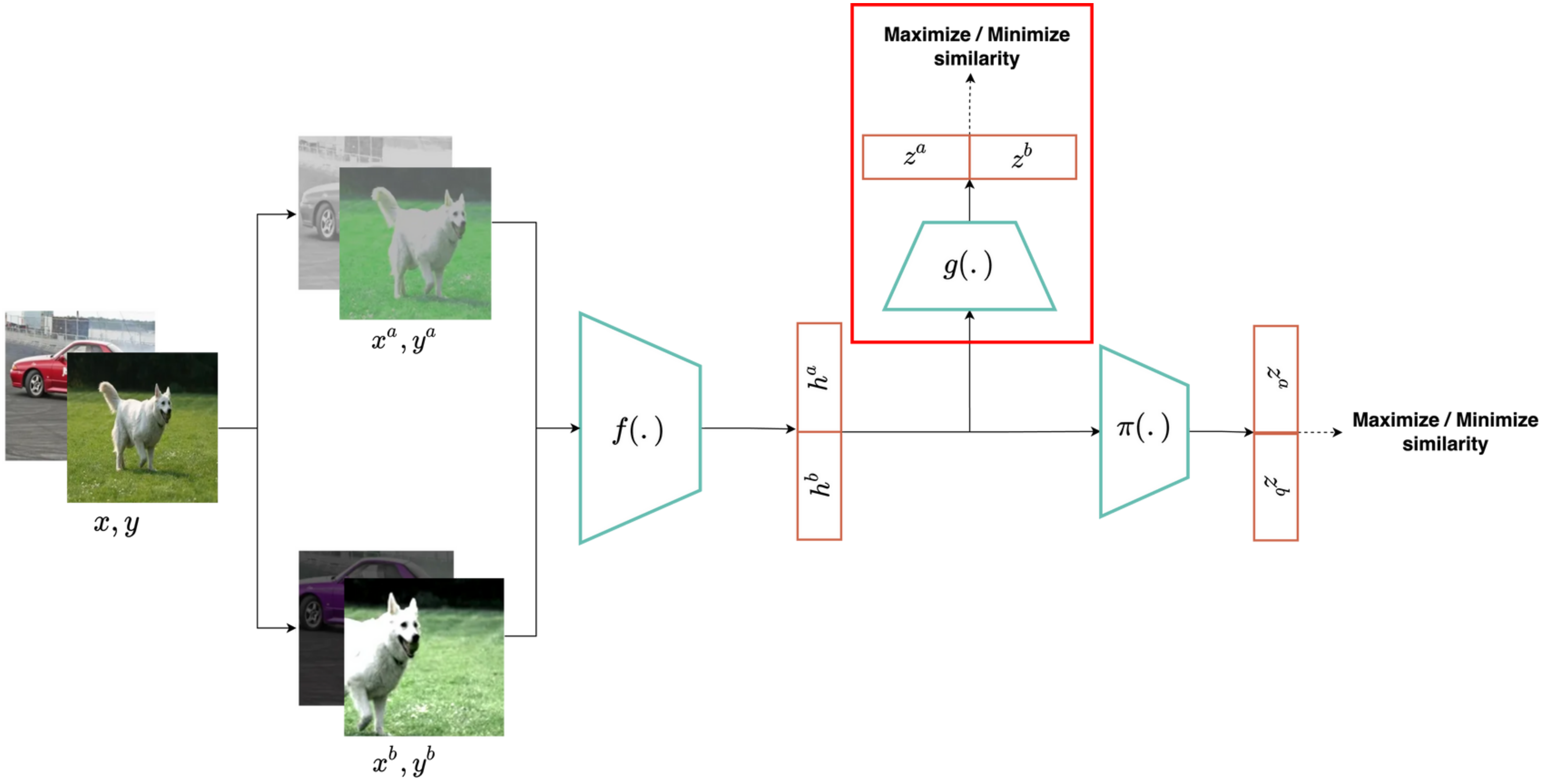


Image from Neural Networks and MLP

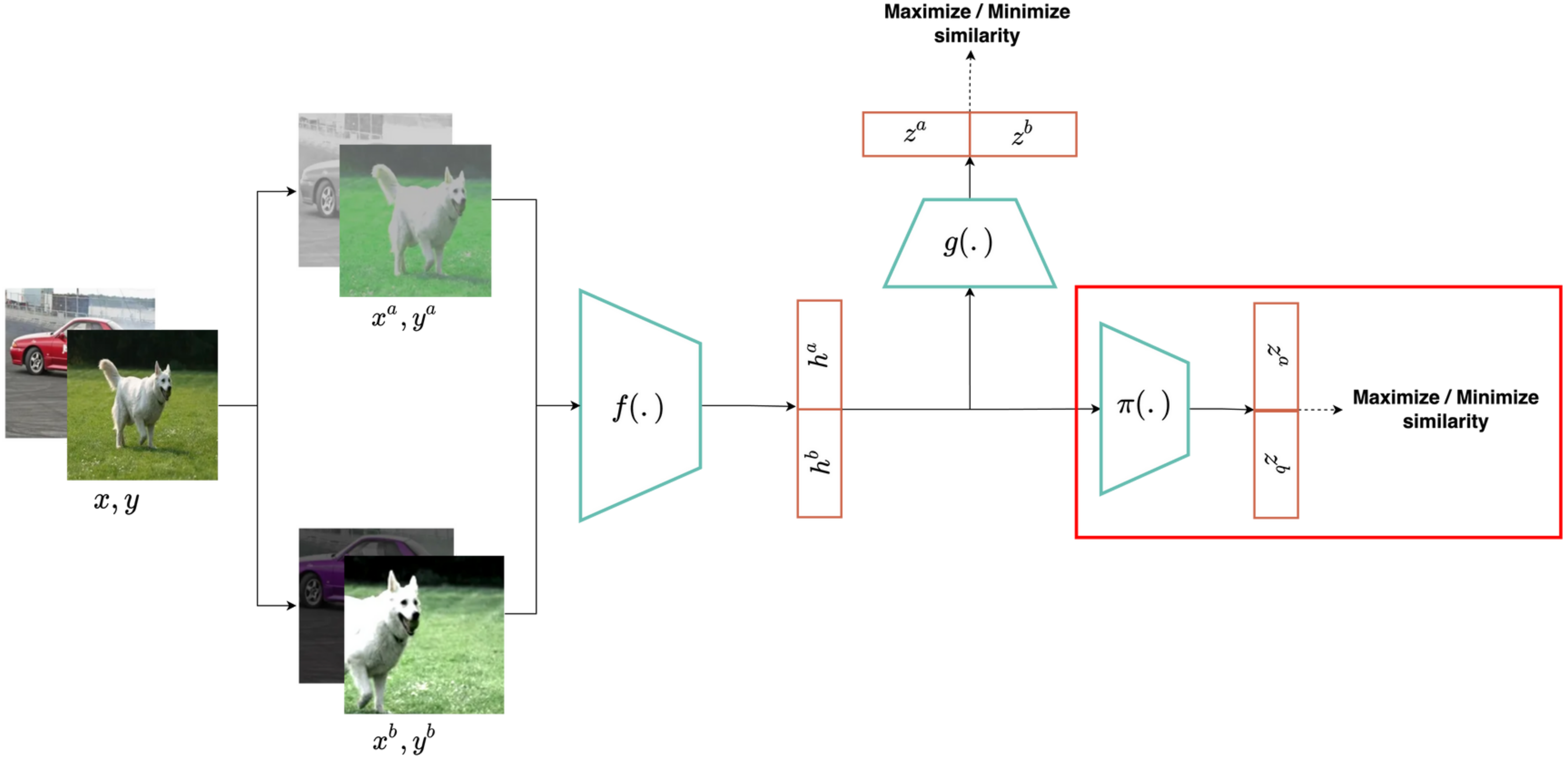


Projection Head

[1] A Simple Framework for Contrastive Learning of Visual Representations

Contrastive Hierarchical Clustering - Model Architecture

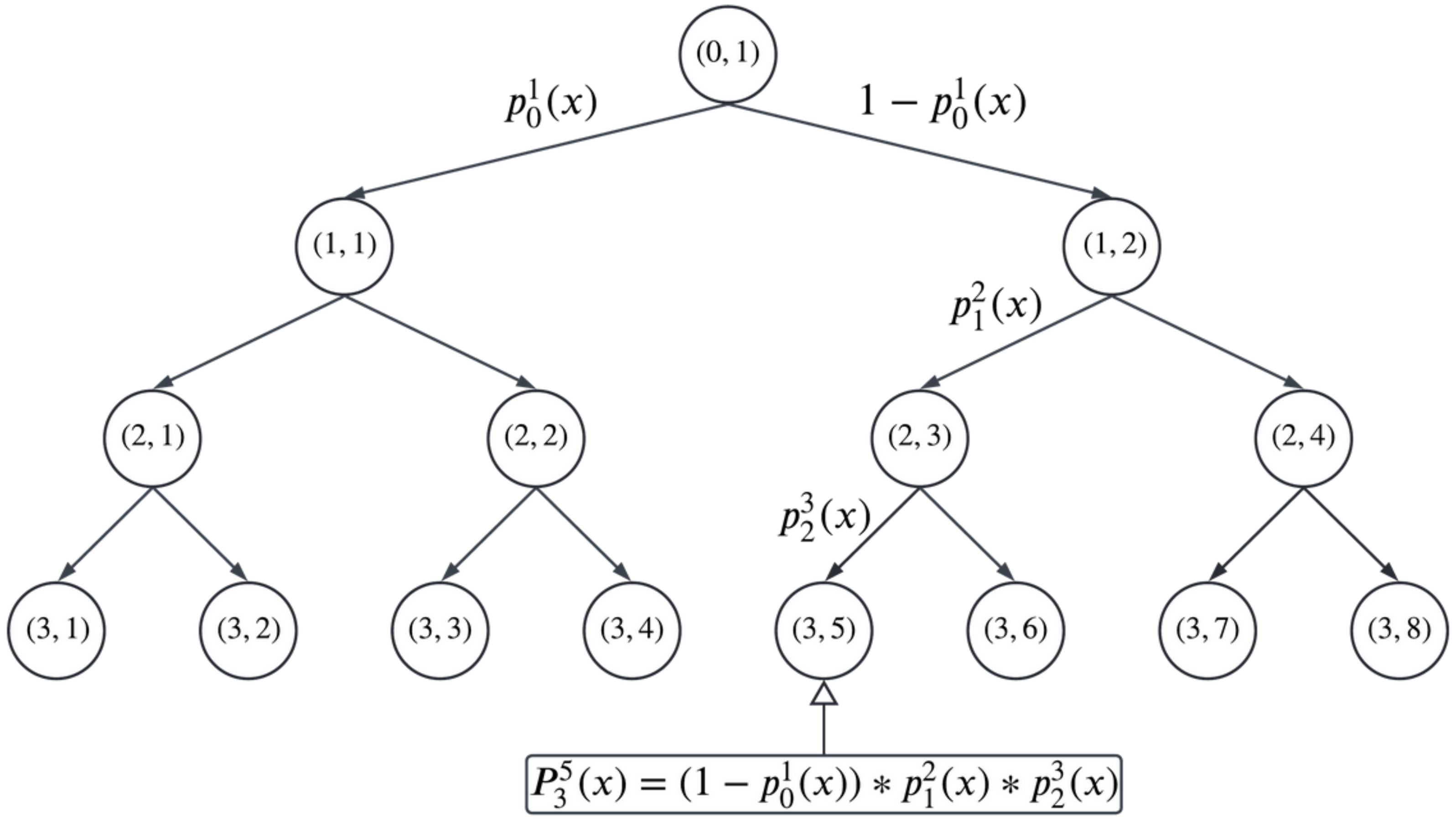
- $\pi(.)$ is one fully connected layer distilled into a **soft decision tree [2]**.
- Assigns data points to clusters by a sequence of decisions.
- Trained with **contrastive hierarchical loss** function which maximizes the likelihood of similar data points being assigned to the same clusters.



Hierarchical Clustering Head

[2] Distilling a Neural Network Into a Soft Decision Tree

Contrastive Hierarchical Clustering - Tree Model



To construct a **decision tree**, we follow the idea behind soft decision trees **[2]**, and model the tree path by a sequence of decisions:

$$\pi(z) = [\sigma(w_1^T z + b_1), \ldots, \sigma(w_K^T z + b_K)]$$

where $\sigma(.)$ is a **sigmoid** function and $w_n \in \mathbb{R}^N$ with $b_n \in \mathbb{R}$ are weights of a linear layer.

With $\pi(.)$ output we can define a probability distribution of assigning data to **clusters** on all levels of the tree:

$$P_t(x) = [P_t^0(x), P_t^1(x), \ldots, P_t^{2^t-1}(x)] \text{ , for } t = [1, T].$$

[2] Distilling a Neural Network Into a Soft Decision Tree

Contrastive Hierarchical Clustering - Building structure

Similarity between data points

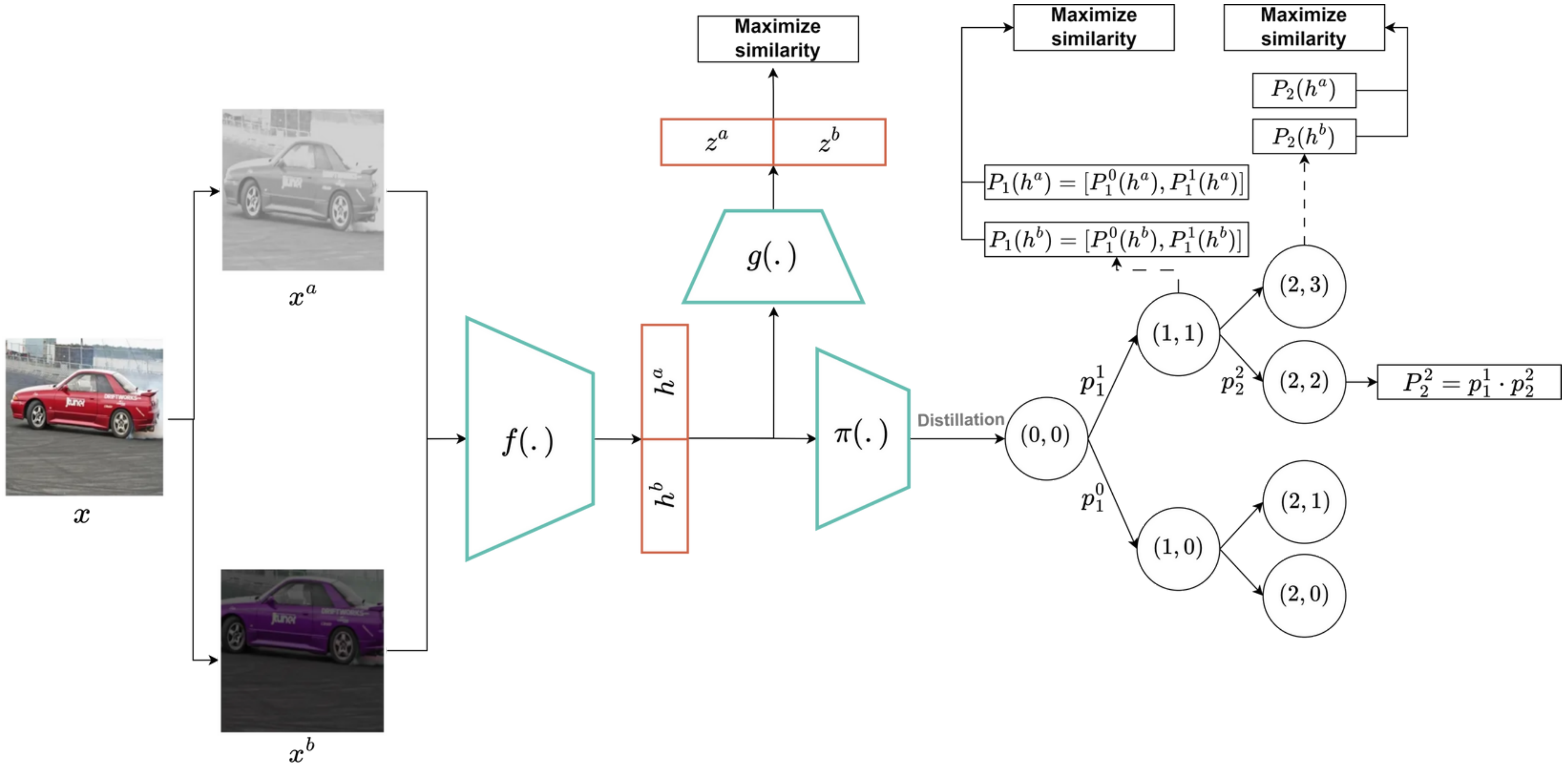
$$s_t(x_1, x_2) = \sqrt{P_t(x_1) \cdot P_t(x_2)} = \sum_{i=0}^{2^t-1} \sqrt{P_t^i(x_1) P_t^i(x_2)}$$

Hierarchical clustering loss

$$CoHiLoss = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{i \neq j} s(x_j, \tilde{x}_i) - \frac{1}{N} \sum_{j=1}^{N} s(x_j, \tilde{x}_j)$$

Training vs Inference

- Tree model in inference mode returns the **index of the most probable path**.
- Tree model in training mode returns the **probability of assigning data to every cluster**.



Contrastive Hierarchical Clustering - Regularization

Regularization

- **(R1)** How to prevent collapsing and how to use sub-trees equally?
  - Minimizing the cross-entropy between the desired distribution [0.5, 0.5] and the actual distribution to choose the left or right path in a given node.
- **(R2)** Improving the representation with **NT-Xent [1]** Loss.

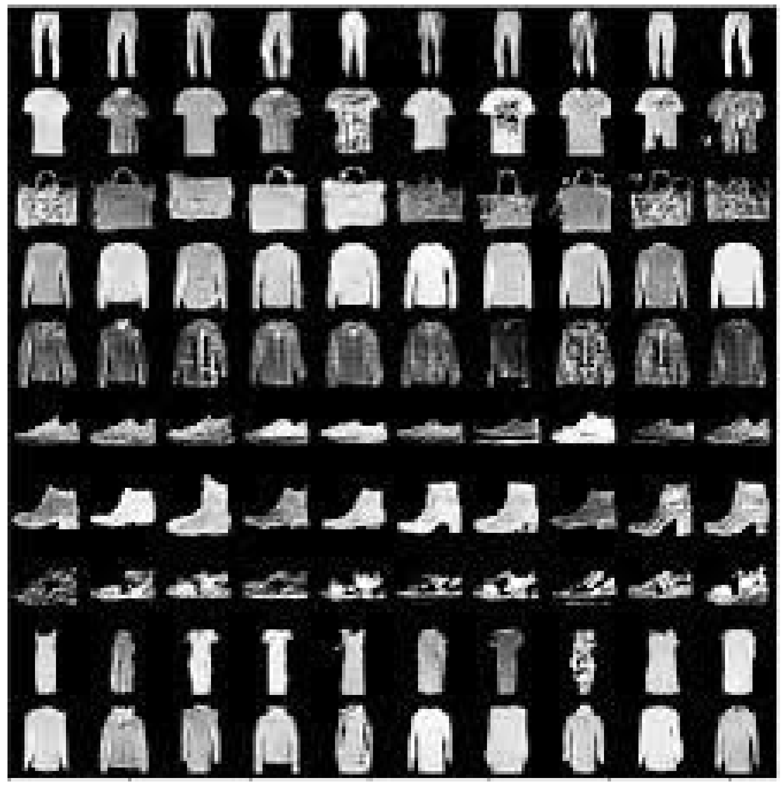
Pruning

- How to match the number of leaves with the number of classes?
  - Namely, we reduce leaves with the **lowest expected fraction** of data points: $P_T^i = \frac{1}{|X|} \sum_{x \in X} P_T^i(x)$

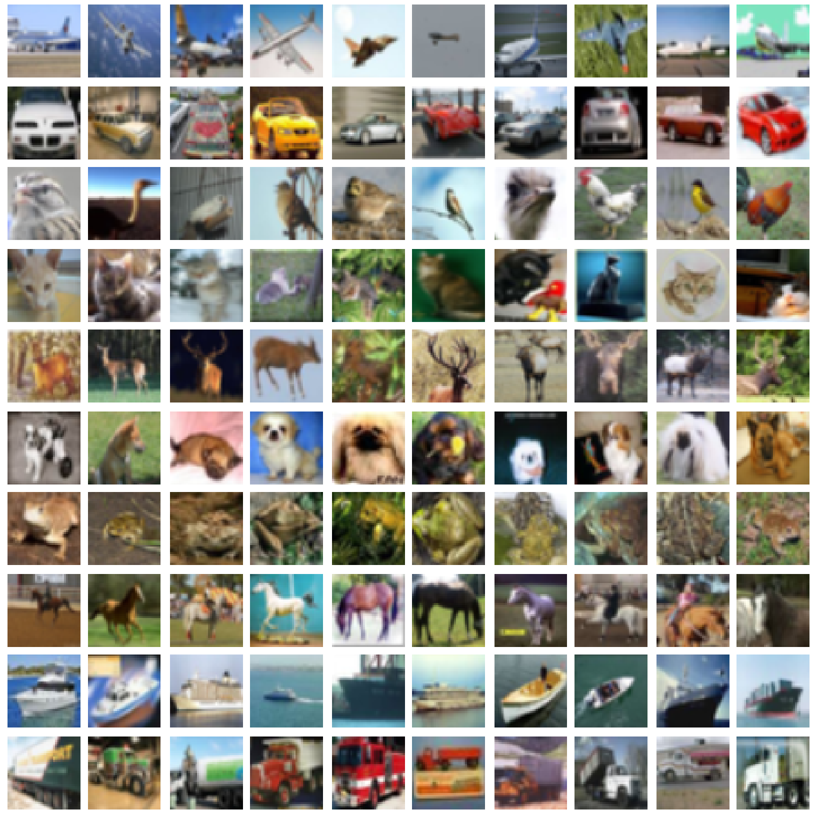
[1] A Simple Framework for Contrastive Learning of Visual Representations

MNIST

F-MNIST

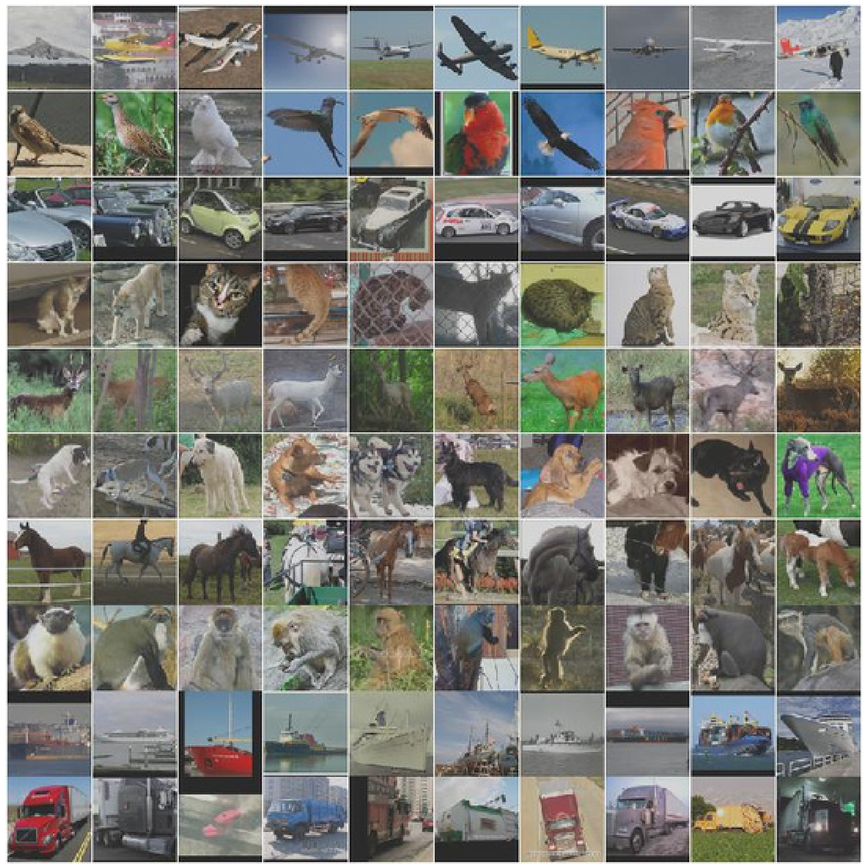


CIFAR10



CIFAR100

STL10



ImageNet10

Results

Comparison with flat clustering methods on datasets of color images

Dataset	CIFAR-10			CIFAR-100			STL-10			ImageNet-10			ImageNet-Dogs		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
K-means (Mac)	0.087	0.229	0.049	0.084	0.130	0.028	0.125	0.192	0.061	0.119	0.241	0.057	0.055	0.105	0.020
SC (Zelnik-Manor & Perona)	0.103	0.247	0.085	0.090	0.136	0.022	0.098	0.159	0.048	0.151	0.274	0.076	0.038	0.111	0.013
AC (Gowda & Krishna, 1978)	0.105	0.228	0.065	0.098	0.138	0.034	0.239	0.332	0.140	0.138	0.242	0.067	0.037	0.139	0.021
NMF (Cai)	0.081	0.190	0.034	0.079	0.118	0.026	0.096	0.180	0.046	0.132	0.230	0.065	0.044	0.118	0.016
AE (Bengio et al.)	0.239	0.314	0.169	0.100	0.165	0.048	0.250	0.303	0.161	0.210	0.317	0.152	0.104	0.185	0.073
DAE (Vincent et al., 2010)	0.251	0.297	0.163	0.111	0.151	0.046	0.224	0.302	0.152	0.206	0.304	0.138	0.104	0.190	0.078
DCGAN (Radford et al., 2015)	0.265	0.315	0.176	0.120	0.151	0.045	0.210	0.298	0.139	0.225	0.346	0.157	0.121	0.174	0.078
DeCNN (Zeiler et al., 2010)	0.240	0.282	0.174	0.092	0.133	0.038	0.227	0.299	0.162	0.186	0.313	0.142	0.098	0.175	0.073
VAE (Kingma & Welling, 2013)	0.245	0.291	0.167	0.108	0.152	0.040	0.200	0.282	0.146	0.193	0.334	0.168	0.107	0.179	0.079
JULE (Yang et al., 2016)	0.192	0.272	0.138	0.103	0.137	0.033	0.182	0.277	0.164	0.175	0.300	0.138	0.054	0.138	0.028
DEC (Xie et al., 2016)	0.257	0.301	0.161	0.136	0.185	0.050	0.276	0.359	0.186	0.282	0.381	0.203	0.122	0.195	0.079
DAC (Chang et al., 2017)	0.396	0.522	0.306	0.185	0.238	0.088	0.366	0.470	0.257	0.394	0.527	0.302	0.219	0.275	0.111
DCCM (Wu et al., 2019)	0.496	0.623	0.408	0.285	0.327	0.173	0.376	0.482	0.262	0.608	0.710	0.555	0.321	0.383	0.182
PICA (Huang et al., 2020)	0.591	0.696	0.512	0.310	0.337	0.171	0.611	0.713	0.531	0.802	0.870	0.761	0.352	0.352	0.201
CC (Li et al., 2021a)	0.705	0.790	0.637	0.431	0.429	0.266	**0.764**	**0.850**	**0.726**	0.859	0.893	0.822	**0.445**	**0.429**	**0.274**
CoHiClust	**0.779**	**0.839**	**0.731**	**0.467**	**0.437**	**0.299**	0.584	0.613	0.474	**0.907**	**0.953**	**0.899**	0.411	0.355	0.232

Comparison with hierarchical models

Method	MNIST			F-MNIST		
	DP	NMI	ACC	DP	NMI	ACC
DeepECT	0.82	0.83	0.85	0.47	0.60	0.52
DeepECT + Aug	0.94	0.93	0.95	0.44	0.59	0.50
IDEC (agglomerative complete*)	0.40	0.86	0.85	0.35	0.58	0.53
AE + k-means (bisecting*)	0.53	0.70	0.77	0.38	0.52	0.48
CoHiClust	**0.97**	**0.97**	**0.99**	**0.52**	**0.62**	**0.65**

Results - Ablation Study

Ablation Study - Backbone

Table 2: The importance of architecture choice.

Method	CoHiClust			CC [24]		
Backbone	NMI	ACC	ARI	NMI	ACC	ARI
ResNet18	0.711	0.768	0.642	0.650	0.736	0.569
ResNet34	0.730	0.788	0.667	**0.705**	**0.790**	**0.637**
ResNet50	**0.767**	**0.840**	**0.720**	0.663	0.747	0.585

Results - Ablation Study

Ablation Study - Impact of losses

Table 3: Ablation study of CoHiClust loss function performed on CIFAR-10.

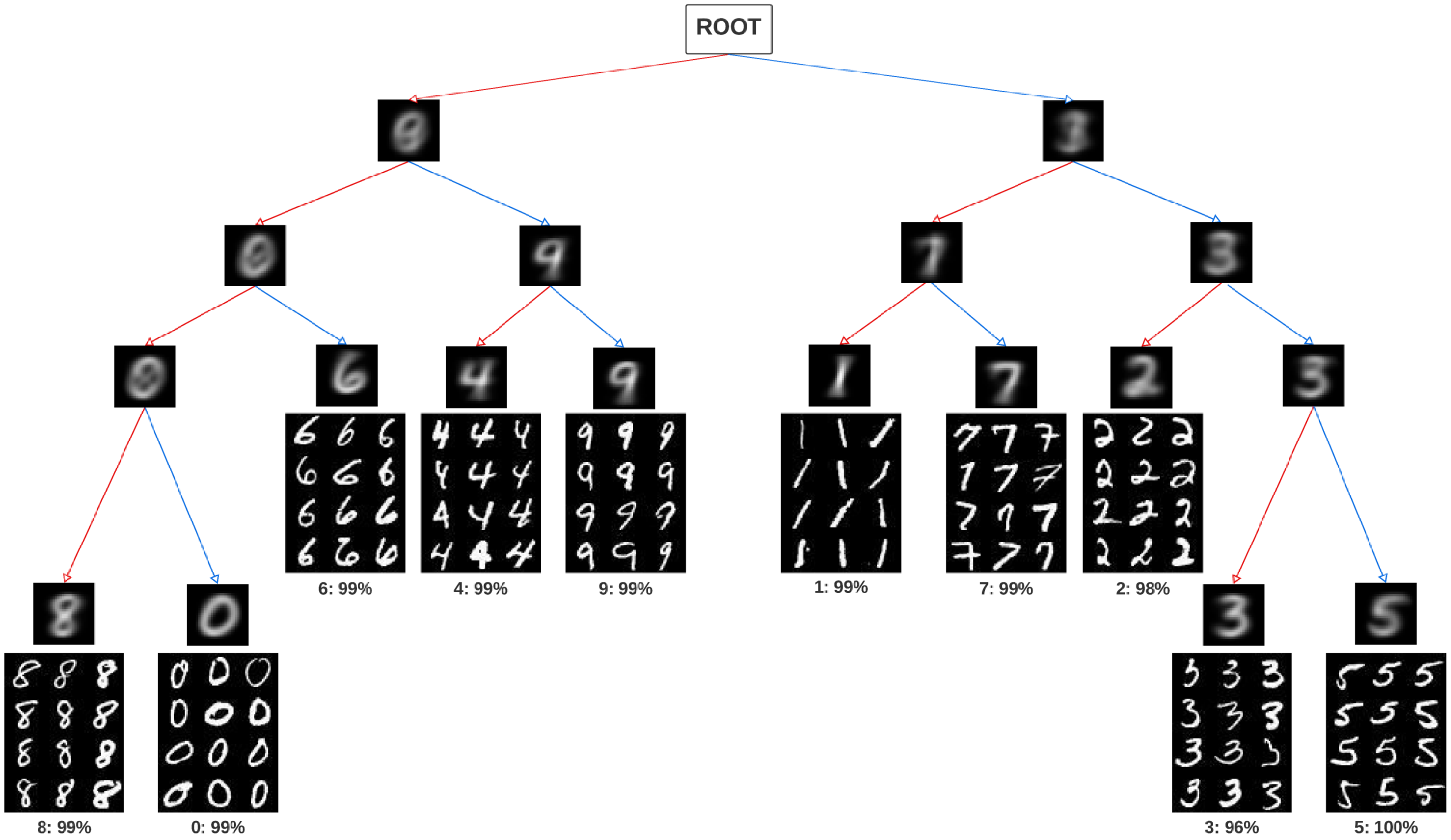
	NMI	ACC	ARI
CoHiLoss	0.567	0.569	0.457
CoHiLoss + R1	0.629	0.726	0.549
CoHiLoss + R1 + R2	**0.767**	**0.84**	**0.72**
CoHiClust w/o pre-training	0.59	0.657	0.50

Comparison to Agglomerative Clustering

Table 5: Comparison with agglomerative clustering trained on the representation generated by the self-supervised learning model.

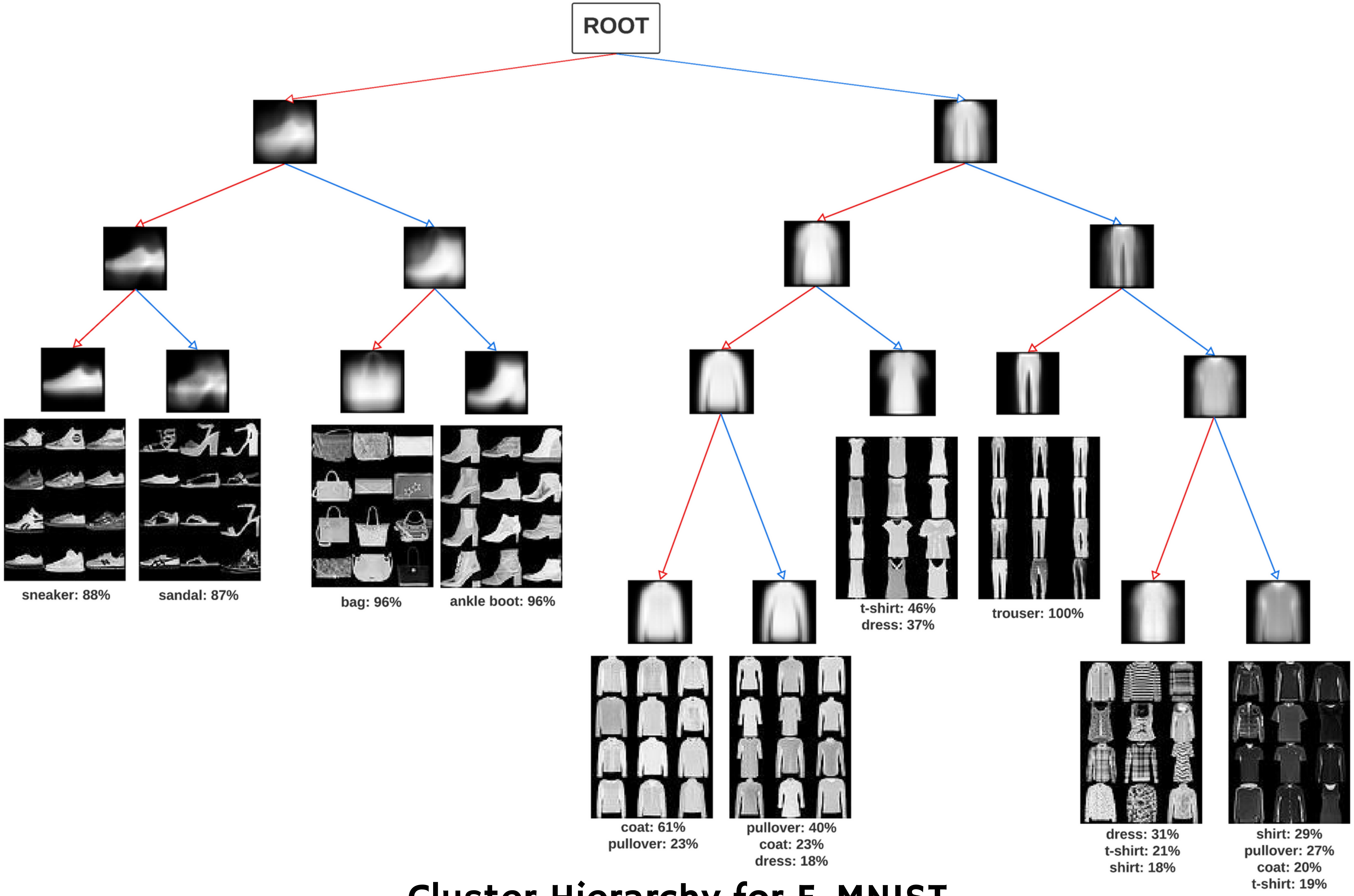
	NMI	ACC	ARI
Agglomerative clustering	0.265	0.363	0.147
CoHiClust	**0.767**	**0.84**	**0.72**

Results - Structure Analysis



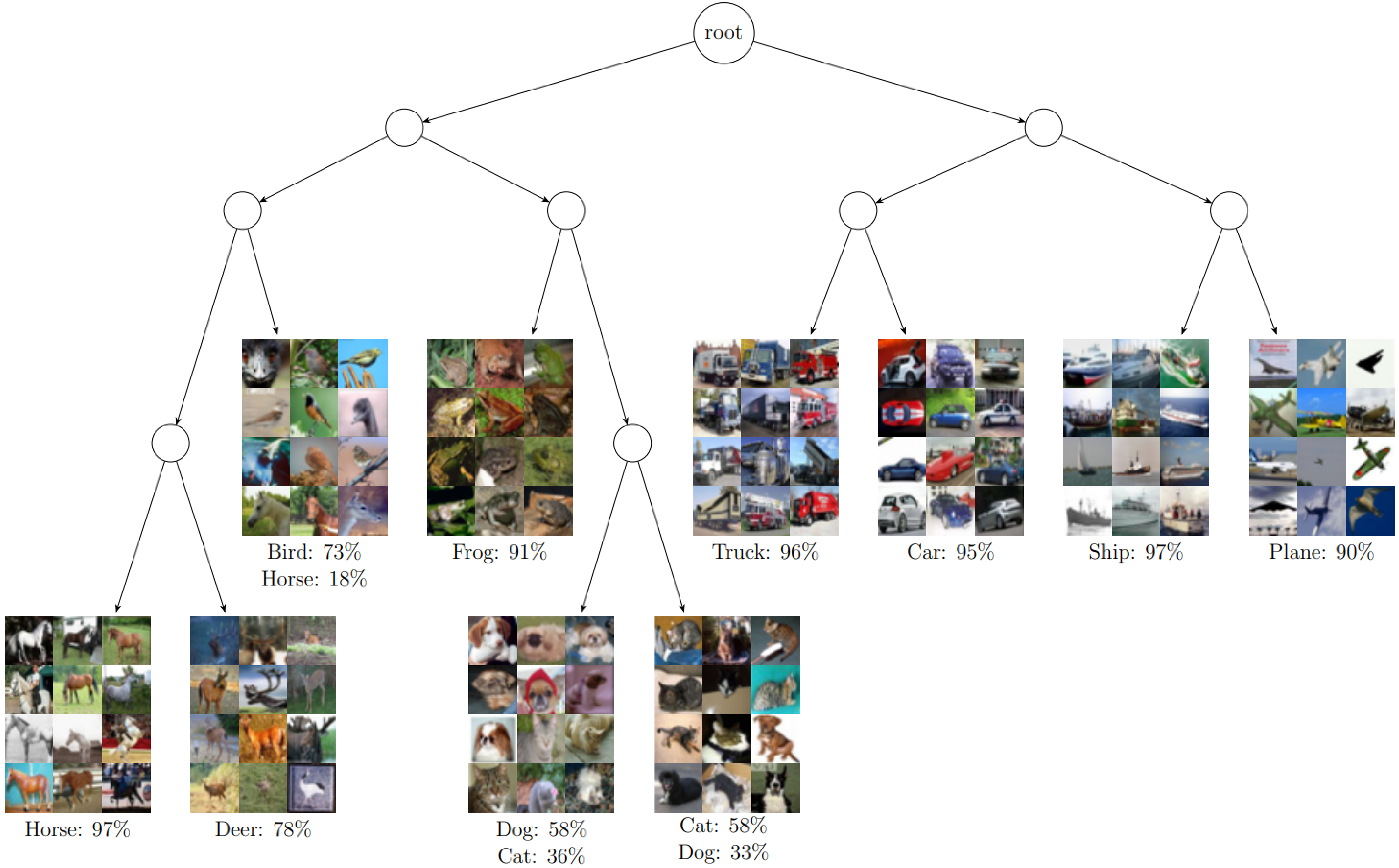
Cluster Hierarchy for MNIST

Results - Structure Analysis



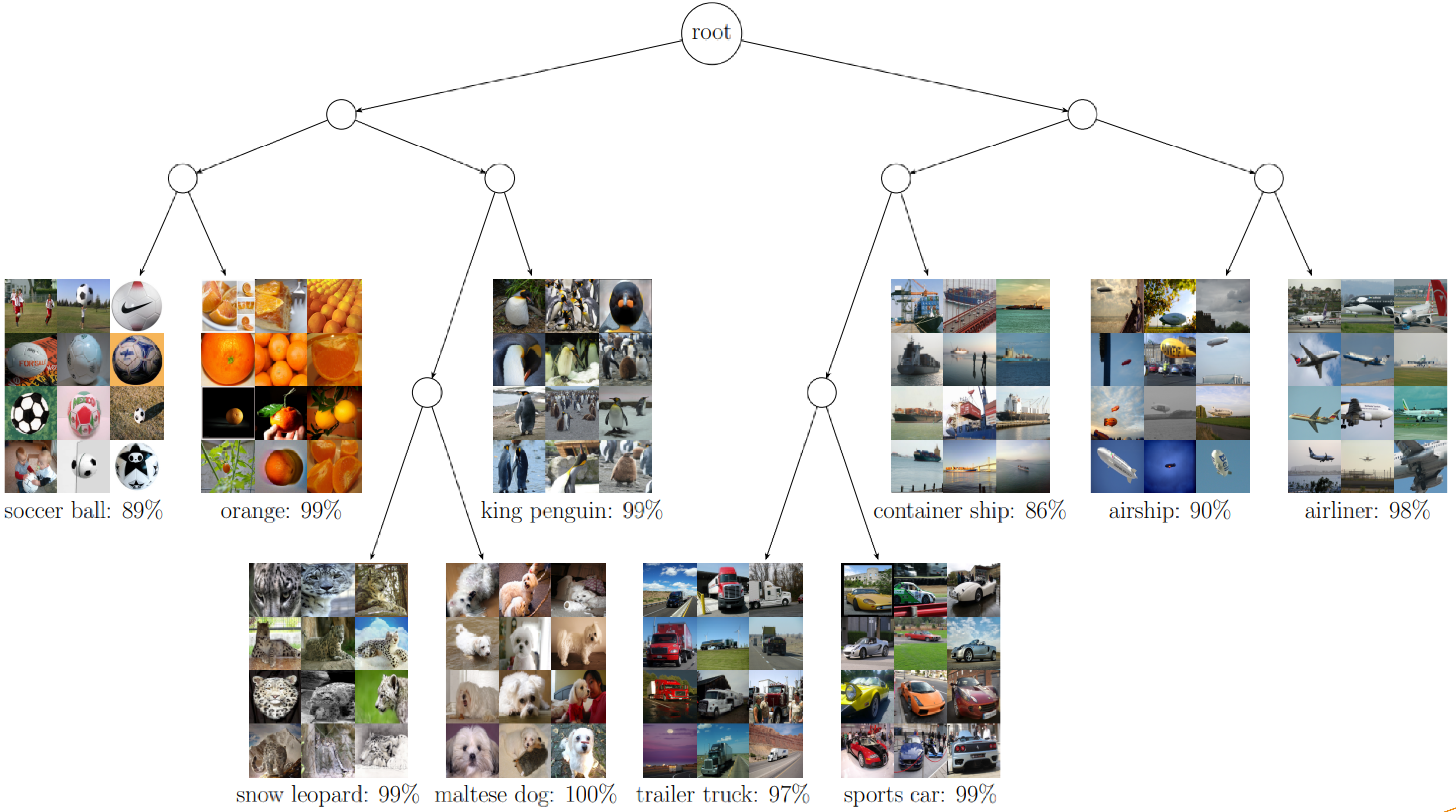
Cluster Hierarchy for F-MNIST

Results - Structure Analysis



Cluster Hierarchy for CIFAR10

Results - Structure Analysis



Cluster Hierarchy for ImageNet10

Conclusions

- Our method provides significantly **more information** about the data than typical flat clustering models.
- Analysis performed on typical clustering benchmarks confirms that the produced partitions are **highly similar to ground-truth classes**.
- Our method generates a reasonable structure of clusters, which is consistent with **human intuition** and **image semantics**.

Future works

- Experiment with datasets that have more complex structures:
  - More classes.
  - More relationships between classes.
- Extend work beyond image datasets:
  - Medicine - Molecular datasets.

Thank you

SCAN TO READ THE PAPER

AMD

group of machine
gmum
learning methods