南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Contrastive Label Disambiguation for Partial Label Learning

Submit to ICLR2022

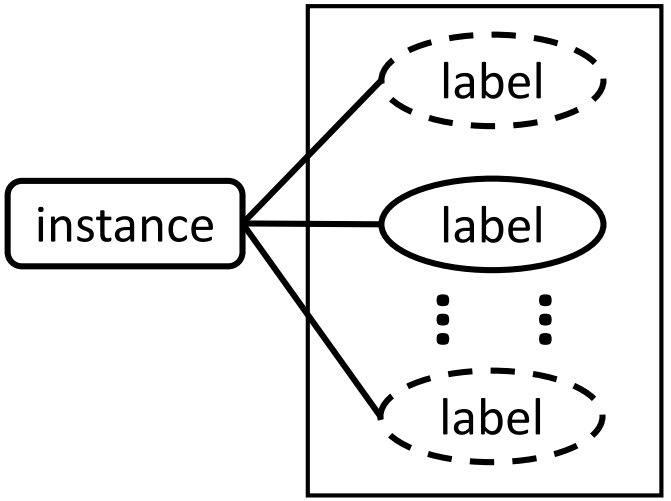
Average score: 8 (Top1)

Partial Label Learning

- **Partial label learning** vs. ordinary supervised learning

Ordinary supervised learning
(only one ground-truth label)

candidate label set



Partial label learning
(a candidate label set, among which only one label is valid)

Partial Label Learning: Applications

- Multi-modal large-scale image annotation

易边再战卢尼打到克劳德面部被吹一级恶意犯规，艾顿篮下得分，普尔命中三分，保罗连得4分，格林连得4分，艾顿篮下打进，约翰逊命中三分，布里奇斯抢断快攻暴扣，艾顿两罚中一，太阳领先9分。随后维金斯打成2+1，普尔罚球得分，佩顿抛投命中，普尔命中三分，佩顿扣篮得分，勇士将分差缩小至1分。紧接着沙梅特命中三分，佩顿连得4分，佩恩命中三分，库里回敬三分球，麦基补扣得手，三节结束太阳80-78领先。

Candidate set
(accurate one in black)

库里 克劳德 保罗 佩顿

Partial Label Learning: Applications

- Crowdsourcing labeling

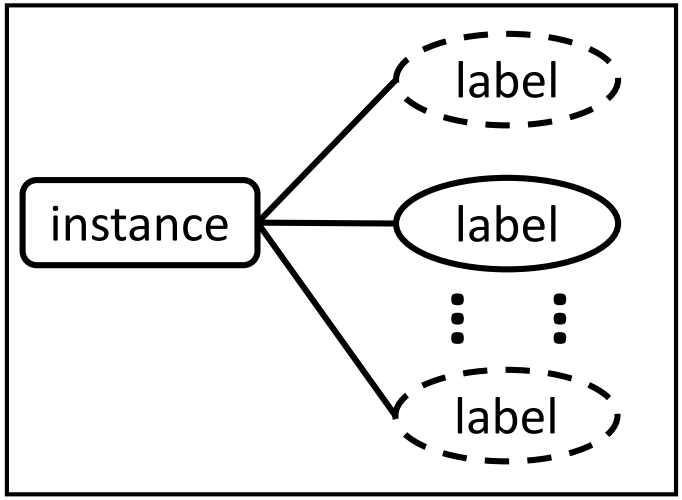
Annotator 1: Horse
Annotator 2: Donkey
Annotator 3: Mule

Candidate set
(accurate one in black)

Horse Donkey Mule

Partial Label Learning Framework

- Learning from partial labels



✓ Each instance is associated with **multiple candidate labels**

✓ Only one of the candidate label is the **unknown ground-truth label**

- Learning a classifier from the candidate set Y_i

$$\mathcal{L}_{\text{cls}}(f; \boldsymbol{x}_i, Y_i) = \sum_{j=1}^{C} -s_{i,j} \log(f^j(\boldsymbol{x}_i)) \quad \text{s.t.} \quad \sum\nolimits_{j \in Y_i} s_{i,j} = 1 \text{ and } s_{i,j} = 0, \forall j \notin Y_i,$$

- **Solution: disambiguation**:
  - Recover the ground-truth confidence s from Y_i

Existing Disambiguating Strategies

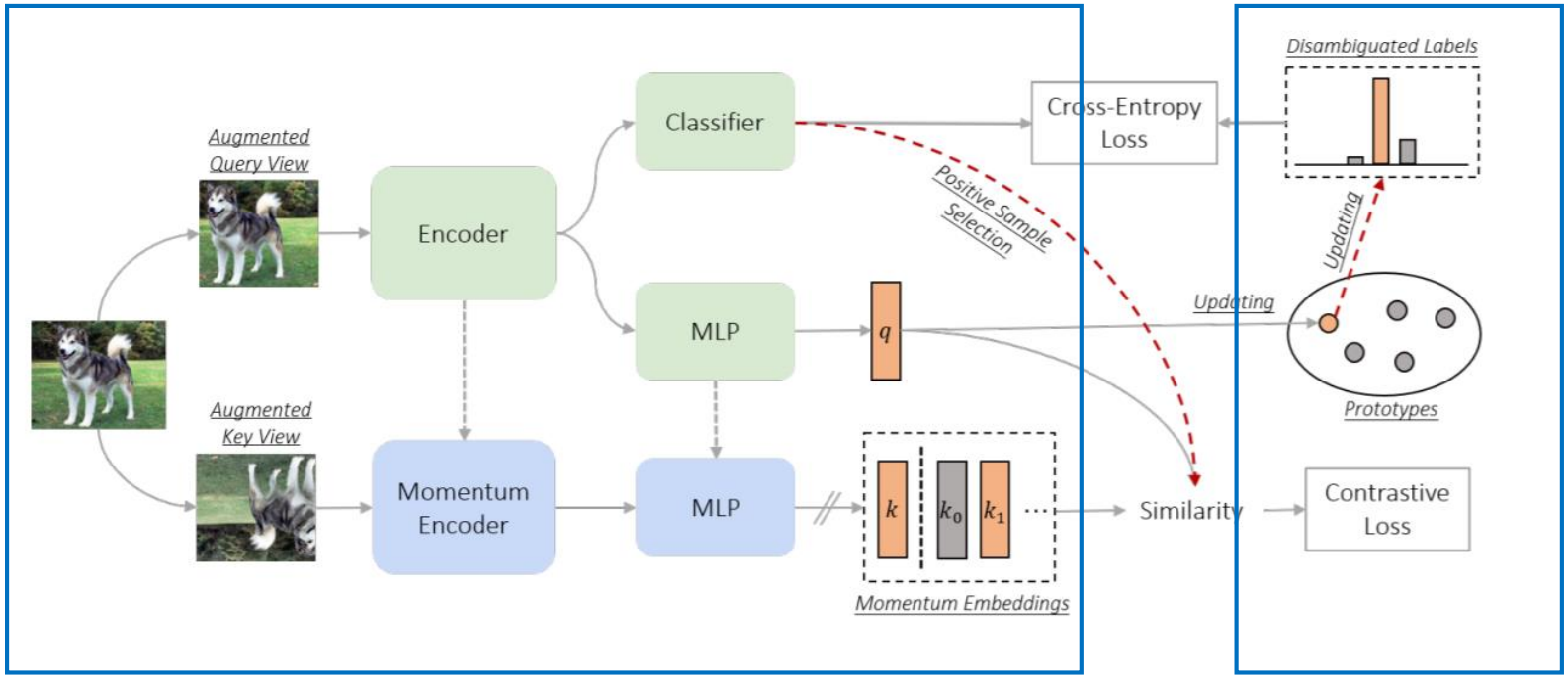
- Regularization term
  - Entropy regularizer: $\mathcal{L}_d(\hat{\mathbf{Y}}) = -\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}_i^{\top}\log\hat{\mathbf{y}}_i$ [AAAI'20]
  - Smooth assumption: $\mu\sum_{i,j}^{m} s_{ij}\left\|\frac{\mathbf{p}_i}{\sqrt{d_{ii}}}-\frac{\mathbf{p}_j}{\sqrt{d_{jj}}}\right\|_2^2$ [AAAI'18]
  - Self training: $\mathrm{w}_{ij}=\begin{cases} g_j(\boldsymbol{x}_i)\Big/\sum_{k\in s_i} g_k(\boldsymbol{x}_i) & \text{if } j\in s_i, \\ 0 & \text{otherwise.}\end{cases}$ [ICML'20]
- Noisy label identification

The PiCO Framework

- Contrastive label disambiguation

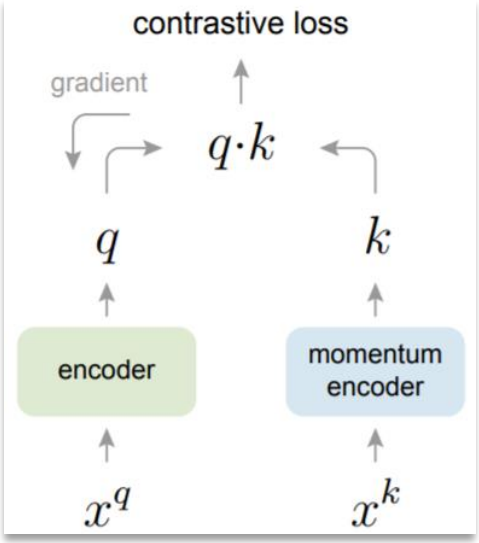
Contrastive representation learning

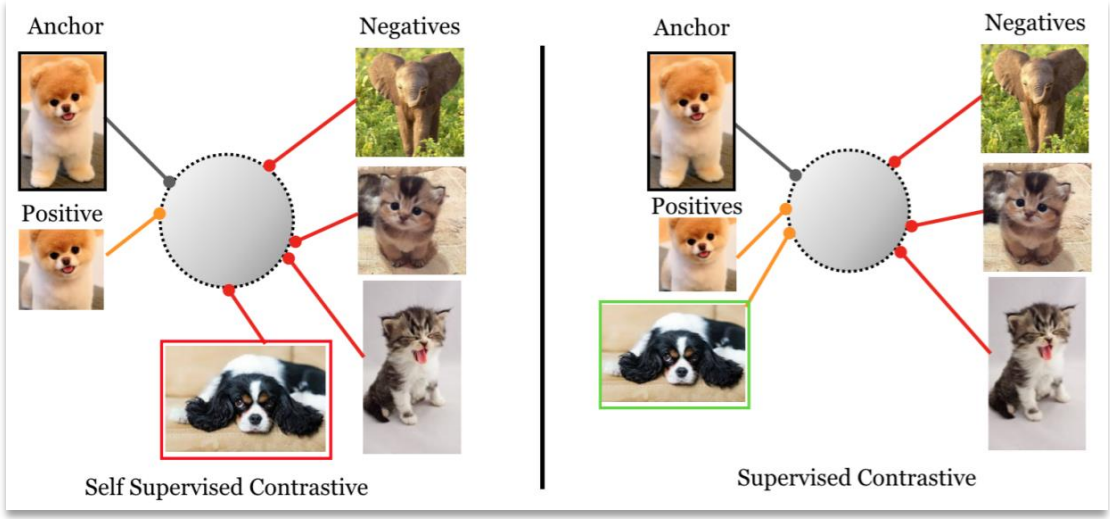
Prototype-based label disambiguation



The PiCO Framework (cont.)

- Supervised contrastive learning





Self-Supervised contrastive loss:

$$-\sum_{i \in I} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_{j(i)} / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)}$$

Supervised contrastive loss:

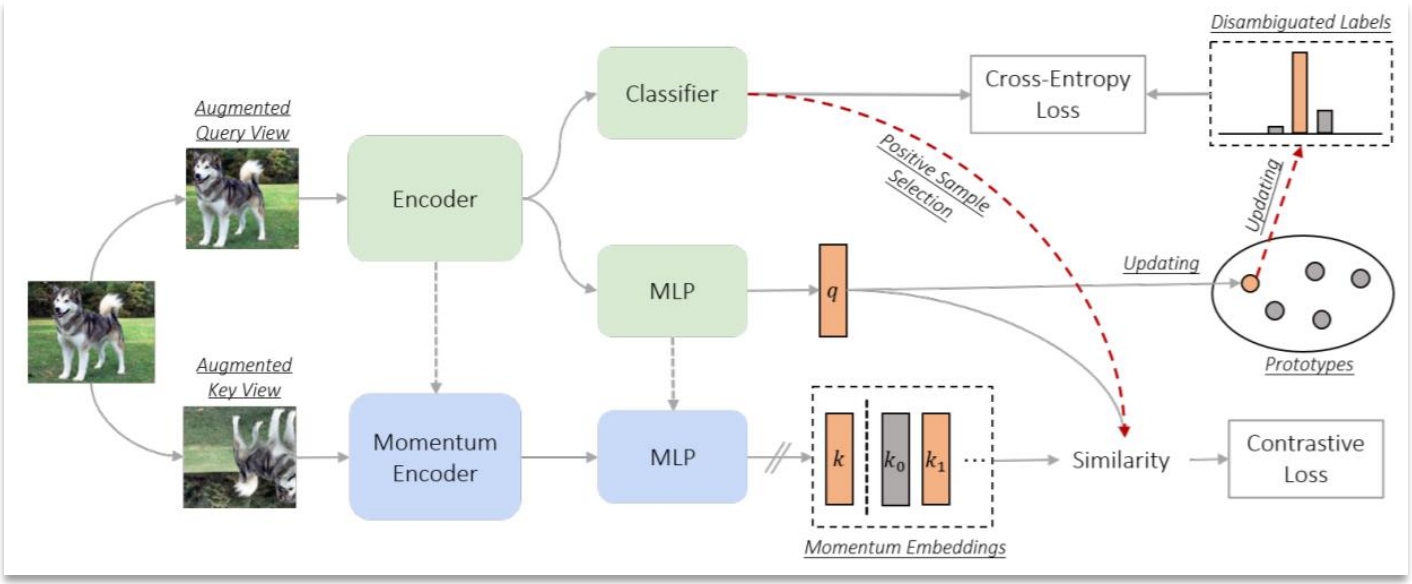
$$\sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)}$$

The PiCO Framework (cont.)

- Contrastive representation learning
  - Positive set selection: $P(\boldsymbol{x}) = \{\boldsymbol{k}' | \boldsymbol{k}' \in A(\boldsymbol{x}), \tilde{y}' = \tilde{y}\}$
  - The contrastive loss:

$$\mathcal{L}_{\text{cont}}(g; \boldsymbol{x}, \tau, A) = -\frac{1}{|P(\boldsymbol{x})|} \sum_{\boldsymbol{k}_+ \in P(\boldsymbol{x})} \log \frac{\exp(\boldsymbol{q}^\top \boldsymbol{k}_+ / \tau)}{\sum_{\boldsymbol{k}' \in A(\boldsymbol{x})} \exp(\boldsymbol{q}^\top \boldsymbol{k}' / \tau)}$$

  - The overall loss: $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{cont}}$ $\mathcal{L}_{\text{cls}}(f; \boldsymbol{x}_i, Y_i) = \sum_{j=1}^{C} -s_{i,j} \log(f^j(\boldsymbol{x}_i))$



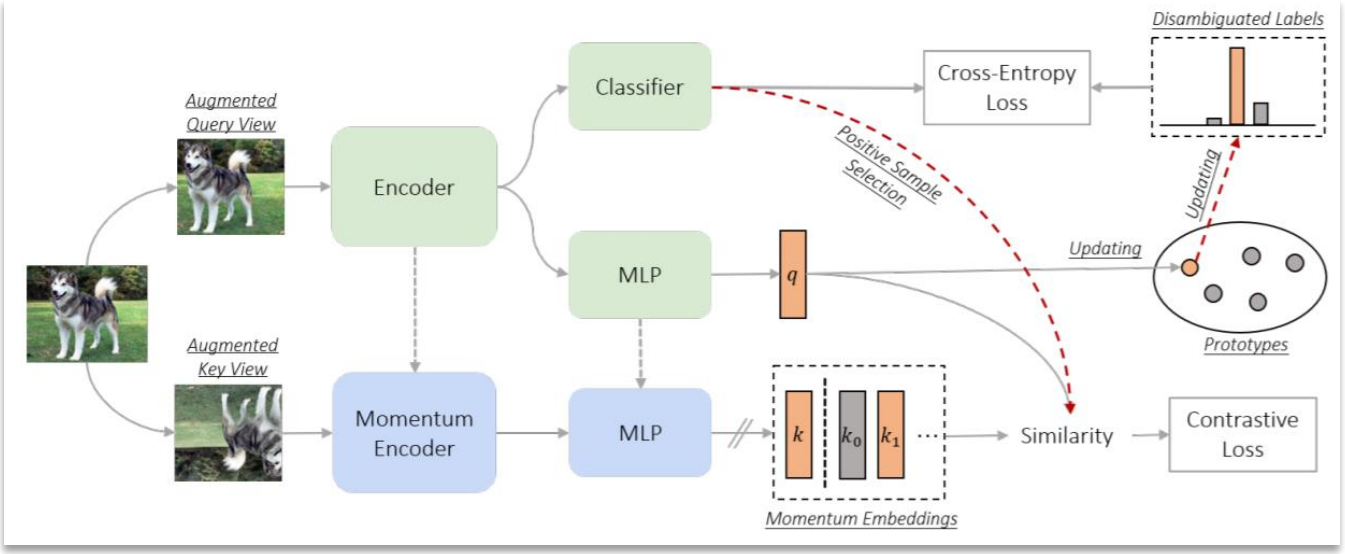
The PiCO Framework (cont.)

- Prototype-based label disambiguation
  - Pseudo target updating:

$$\boldsymbol{s} = \phi \boldsymbol{s} + (1-\phi)\boldsymbol{z}, \quad z_c = \begin{cases} 1 & \text{if } c = \arg\max_{j \in Y} \boldsymbol{q}^\top \boldsymbol{\mu}_j \\ 0 & \text{else} \end{cases}$$

  - Prototype updating

$$\boldsymbol{\mu}_c = \text{Normalize}(\gamma \boldsymbol{\mu}_c + (1-\gamma)\boldsymbol{q}), \quad \text{if } c = \arg\max_{j \in Y} f^j(\text{Aug}_q(\boldsymbol{x}))$$



Experiments

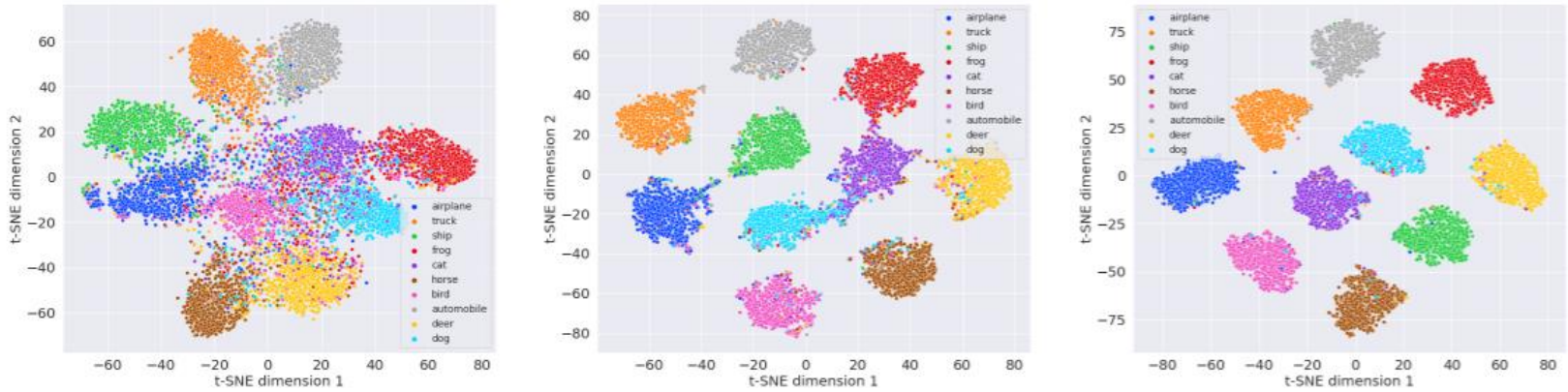
- PiCO achieves SOTA results

Table 1: Accuracy comparisons on benchmark datasets. Bold indicates superior results. Notably PiCO achieves comparable results to the fully supervised learning (less than 1% in accuracy with 1 false candidate).

Dataset	Method	$q = 0.1$	$q = 0.3$	$q = 0.5$
CIFAR-10	PiCO (ours)	**94.39** $\pm$ 0.18%	**94.18** $\pm$ 0.12%	**93.58** $\pm$ 0.06%
	LWS	90.30 $\pm$ 0.60%	88.99 $\pm$ 1.43%	86.16 $\pm$ 0.85%
	PRODEN	90.24 $\pm$ 0.32%	89.38 $\pm$ 0.31%	87.78 $\pm$ 0.07%
	CC	82.30 $\pm$ 0.21%	79.08 $\pm$ 0.07%	74.05 $\pm$ 0.35%
	MSE	79.97 $\pm$ 0.45%	75.64 $\pm$ 0.28%	67.09 $\pm$ 0.66%
	EXP	79.23 $\pm$ 0.10%	75.79 $\pm$ 0.21%	70.34 $\pm$ 1.32%
	Fully Supervised		94.91 $\pm$ 0.07%	
Dataset	Method	$q = 0.01$	$q = 0.05$	$q = 0.1$
CIFAR-100	PiCO (ours)	**73.09** $\pm$ 0.34%	**72.74** $\pm$ 0.30%	**69.91** $\pm$ 0.24%
	LWS	65.78 $\pm$ 0.02%	59.56 $\pm$ 0.33%	53.53 $\pm$ 0.08%
	PRODEN	62.60 $\pm$ 0.02%	60.73 $\pm$ 0.03%	56.80 $\pm$ 0.29%
	CC	49.76 $\pm$ 0.45%	47.62 $\pm$ 0.08%	35.72 $\pm$ 0.47%
	MSE	49.17 $\pm$ 0.05%	46.02 $\pm$ 1.82%	43.81 $\pm$ 0.49%
	EXP	44.45 $\pm$ 1.50%	41.05 $\pm$ 1.40%	29.27 $\pm$ 2.81%
	Fully Supervised		73.56 $\pm$ 0.10%	

Experiments

- PiCO learns more distinguishable representations



(a) Uniform features (b) PRODEN features (c) PiCO features (ours)

Figure 3: T-SNE visualization of the image representation on CIFAR-10 with $q = 0.5$. Different colors represent the corresponding classes.

$$s_j = 1/|Y| \ (j \in Y)$$

Experiments

- Effective of contrastive loss and label disambiguation

Table 2: Ablation study on CIFAR-10 with $q = 0.5$ and CIFAR-100 with $q = 0.05$.

Ablation	$\mathcal{L}_{\text{cont}}$	Label Disambiguation	CIFAR-10 ($q = 0.5$)	CIFAR-100 ($q = 0.05$)
PiCO	✓	Ours	**93.58**	**72.74**
PiCO w/o Disambiguation	✓	Uniform Pseudo Target	84.50	64.11
PiCO w/o $\mathcal{L}_{\text{cont}}$	✗	Uniform Pseudo Target	76.46	56.87
PiCO with $\phi = 0$	✓	Soft Prototype Probs	91.60	71.07
PiCO with $\phi = 0$	✓	One-hot Prototype	91.41	70.10
PiCO	✓	MA Soft Prototype Probs	81.67	63.75

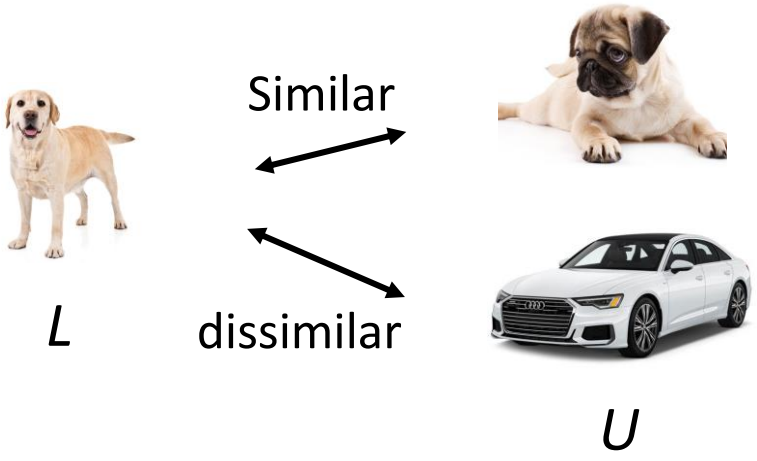
$$\boldsymbol{s} = \phi \boldsymbol{s} + (1-\phi)\boldsymbol{z}, \quad z_c = \begin{cases} 1 & \text{if } c = \arg\max_{j \in Y} \boldsymbol{q}^\top \boldsymbol{\mu}_j \\ 0 & \text{else} \end{cases}$$

$$\boldsymbol{s} = \boldsymbol{z}$$

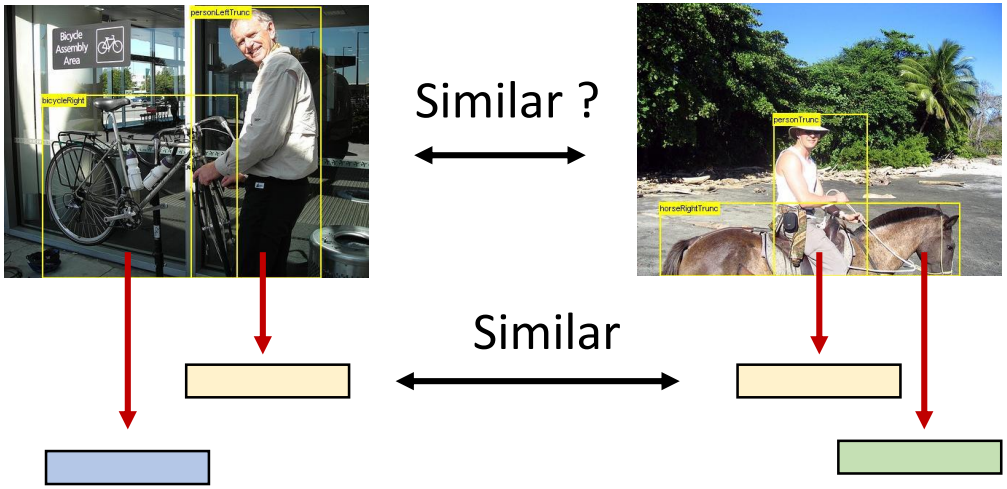
$$s_i = \frac{\exp(\boldsymbol{q}^\top \boldsymbol{\mu}_i / \tau)}{\sum_{j \in Y} \exp(\boldsymbol{q}^\top \boldsymbol{\mu}_j / \tau)}$$

Discussion

- Smooth assumption
  - Single-label

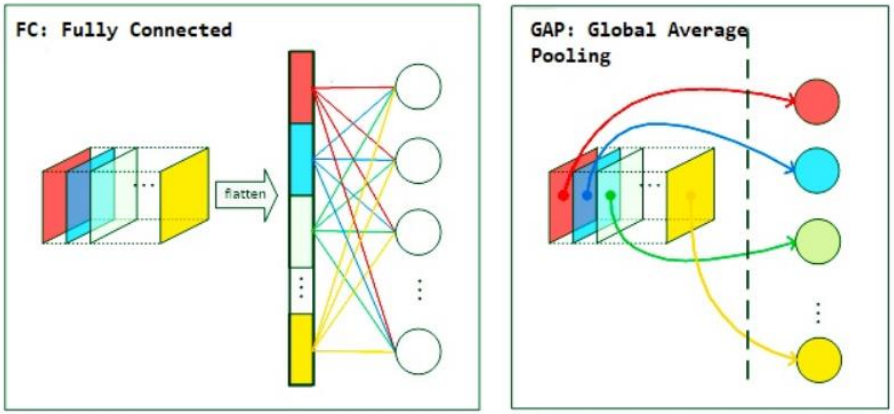


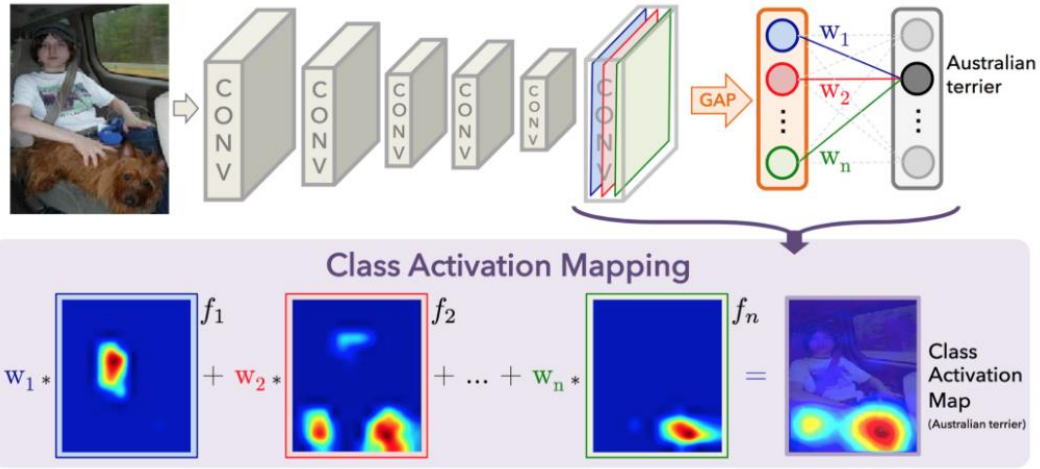
  - Multi-Label



Discussion

- Global Average Pooling (GAP)





THANKS