Paper ID 840

SHENZHEN INSTITUTE
of ARTIFICIAL INTELLIGENCE AND ROBOTICS *for* SOCIETY
深圳市人工智能与机器人研究院

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

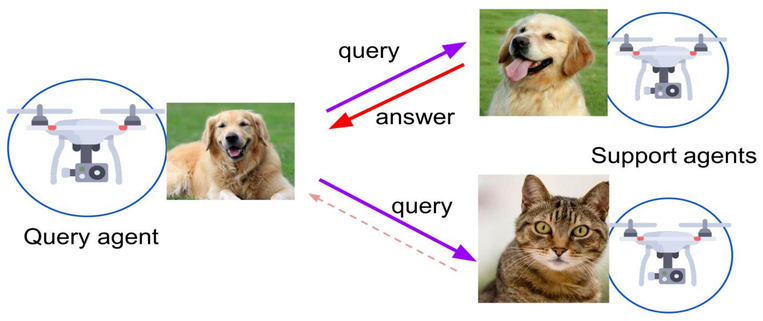
ACM multimedia
Chengdu, China OCT 20-24 2021

Few-Shot Multi-Agent Perception

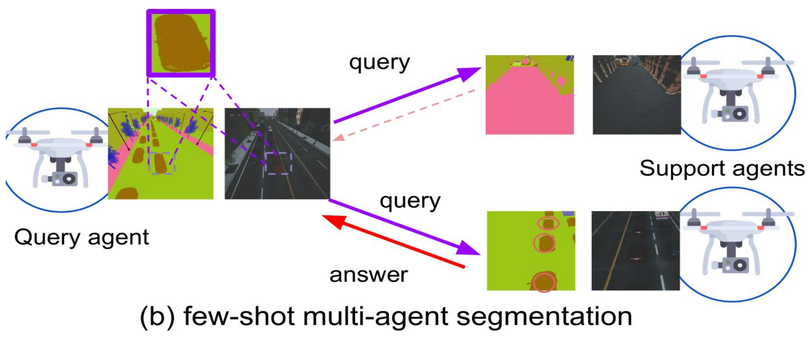
Chenyou Fan[†], Junjie Hu[†], Jianwei Huang[‡,†,*]
{fanchenyou, hujunjie, jianweihuang}@cuhk.edu.cn
[†]Shenzhen Institute of Artificial Intelligence and Robotics for Society, China
[‡]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

Motivations and Scenarios:

- In many real-world visual perception tasks, multiple agents are observing the environment from different perspectives at the same time. We study how to collaborate many agents to make effective and efficient perceptions in data-scarce scenarios.



(a) few-shot multi-agent classification



(b) few-shot multi-agent segmentation

- For instance, in realistic air-ground collaboration scenarios, a fleet of UAVs and UGVs can be operating in the same or different scenes to track moving objects in the air or ground; sensor networks can share viewpoints to capture objects (e.g., wild animals) or natural phenomena (e.g., wild fires). Yet the events could happen very rarely, causing the few-shot support data.

- We propose a multi-agent communication and metric learning framework to coordinate many agents to query the labels for their new observations from a set of support agents who can share their own observations and labels for partial categories.

Our Approach:

- Definition of a ***C-way K-shot N-agent learning task:***
  1. There is a total number of C data classes.
  2. N support agents (s-agent) are collaboratively perceiving the environment; each s-agent observes K data instances of one or several classes of data; K is small, e.g., 1 or 5.
  3. Each query agent (q-agent) sends query data (w/o labels) to s-agents to query the true labels by searching for correct support data.

- For a given query image, q-agent extract 3-D query feature maps $\boldsymbol{q_u}$; for given support images, s-agent extracts key features $\boldsymbol{k_v}$.

- To reduce communication costs, $\boldsymbol{q_u}$ is set to small dimensions (e.g., 32), as it needs broadcast to all s-agents; while $\boldsymbol{k_v}$ is set to large dimensions (e.g., 1024) for keeping rich information.

Reference:
[1] Yen-Cheng Liu et al. When2com: Multi-Agent Perception via Communication Graph Grouping. In CVPR 2020.
[2] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. In IEEE Transactions on Speech and Audio Processing 2002
[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In NeurIPS 2013.

- We design our FS-MAP architecture:
  1. shared backbone networks to produce 3D-feature maps and key/query features of different dimensions.
  2. a RegOT module to measure distance between query and support data.

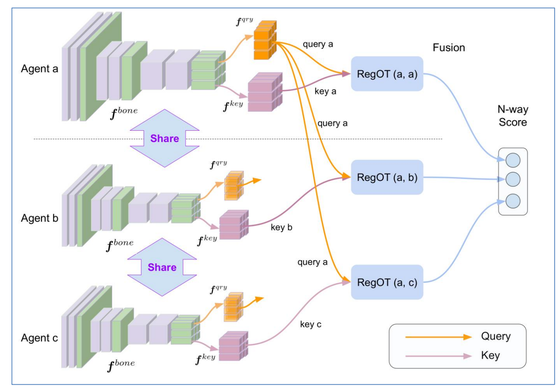


Fig 1: Model overview.

- Region-wise similarity score is calculated by general dot-product as in Eq(1), to measure the importance of each feature region.

$$\hat{s}_{u,i} = \max\left(\boldsymbol{q}_{u,i}^T W_g \frac{\sum_{j=1}^{HW} \boldsymbol{k}_{v,j}}{HW}, \eta\right), \quad s_{u,i} = \frac{\hat{s}_{u,i}}{\sum_{i=1}^{HW} \hat{s}_{u,i}}$$
$$\hat{d}_{v,j} = \max\left(\left(\frac{\sum_{i=1}^{HW} \boldsymbol{q}_{u,i}}{HW}\right)^T W_g \boldsymbol{k}_{v,j}, \eta\right), \quad d_{v,j} = \frac{\hat{d}_{v,j}}{\sum_{j=1}^{HW} \hat{d}_{v,j}}$$
$$\boldsymbol{s}_u = \{s_{u,i}, i \in HW\}, \ \boldsymbol{d}_v = \{d_{v,j}, j \in HW\}$$

Eq(1): regionwise weights

- Define the distance of two feature maps as the minimum cost of transporting the region weights of query data to the support data.

- Solve the regularized optimal transportation task regOT(u,v) as shown in Eq(2) in end-to-end way with Sinkhorn algorithm [3] in $O(n^2 \log n)$ time complexity.

$$\text{regOT}(u,v) = \min_{\boldsymbol{P} \in \mathcal{U}_{\boldsymbol{s},\boldsymbol{d}}} \langle \boldsymbol{P}, \boldsymbol{C}_{u,v} \rangle - \frac{1}{\lambda} H(\boldsymbol{P})$$
$$\mathcal{U}_{\boldsymbol{s},\boldsymbol{d}} := \{\boldsymbol{P} \in \mathcal{R}_+^{n \times n} : \boldsymbol{P}\mathbf{1} = \boldsymbol{s}_u, \ \boldsymbol{P}^\top \mathbf{1} = \boldsymbol{d}_v\}$$

Eq(2): RegOT objective

Experimental results:

- **Datasets.**
  1. FS-AirSim dataset built upon AirSim-MAP [1] to simulate flying multiple drones in the AirSim “CityEnviron" environment.
  2. FS-AirFace dataset. We collect from 16 persons with patrolling robots and DJI drones of manually labeled 354 and 307 human faces from air and ground perspectives.
  3. GTZAN [2] is a widely used music genre dataset with sound-tracks of 10 genres such as blues, classical, pop, etc.

- **Results for 1-/5-shot segmentation and classification.**

1. For scene semantic segmentation task with FS-AirSim.

Method	3-Way 1-Shot		3-Way 5-Shot	
	Acc	IoU	Acc	IoU
When2Com+MAML [8, 22]	0.593	0.203	0.733	0.310
When2Com+MTL [22, 37]	0.652	0.259	0.735	0.321
TarMAC+MTL [7, 37]	0.660	0.310	0.752	0.328
TarMAC+PANet [43, 44]	0.661	0.292	0.762	0.335
MPNet [19]	0.705	0.287	0.770	0.346
MAP-OT (ours)	0.692	0.261	0.764	0.318
MAP-RegOT (ours)	**0.727**	**0.334**	**0.783**	**0.366**

Table 3: Segmentation results on FS-AirSim dataset.

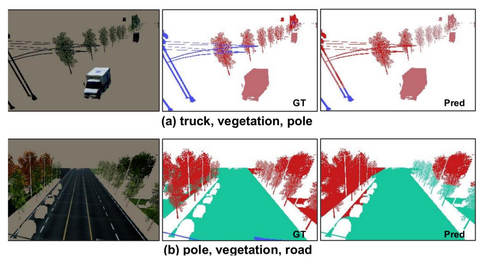


Figure 4: Sample images of FS-AirSim containing truck, vegetation, pole and road, with ground truth masks (mid) and predicted masks (right) with MAP-RegOT.

2. For face recognition tasks with FS-AirFace.

Method	5-Way 1-Shot		5-Way 5-Shot	
	Acc	mAP	Acc	mAP
When2Com+MTL [22, 37]	0.283	0.301	0.309	0.322
TarMAC+MTL [7, 37]	0.310	0.312	0.315	0.345
TarMAC+ProtoNet [35, 43]	0.596	0.642	0.602	0.643
TarMAC+RelationNet [7, 38]	0.564	0.665	0.627	0.687
MAP-OT (ours)	0.636	0.690	0.670	0.737
MAP-RegOT (ours)	**0.671**	**0.740**	**0.693**	**0.751**

Table 5: Face recognition results on FS-AirFace.

Summary/Conclusion

- We tackle the challenge of collaborating distributed agents for learning few-shot tasks. Our method balances the performance and cross-agent communication costs by designing asymmetric query and support feature maps.

- We propose a RegOT-based distance metric which robustly measures the relevance of structured query and support data with invariance to translation and viewpoints.

- Our approach significantly outperforms state-of-the-art methods by 10%-15% on segmentation and classifications tasks upon multimedia data including images and sounds.