A Meta-learning Approach for User-defined Spoken Term Classification with Varying Classes and Examples

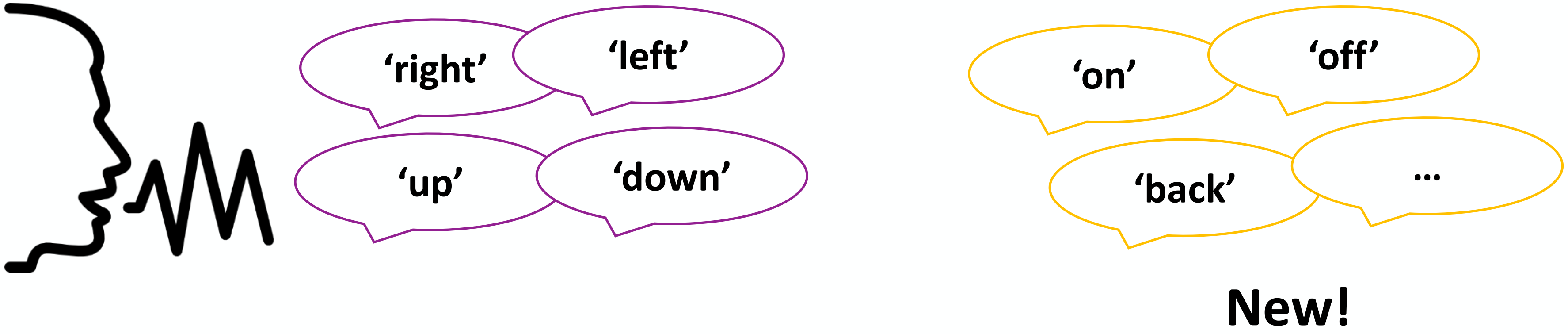
Yangbin Chen[1], Tom Ko[2], Jianping Wang[3]

1. The Chinese University of Hong Kong
2. Southern University of Science and Technology
3. City University of Hong Kong

30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!

User-defined command recognition

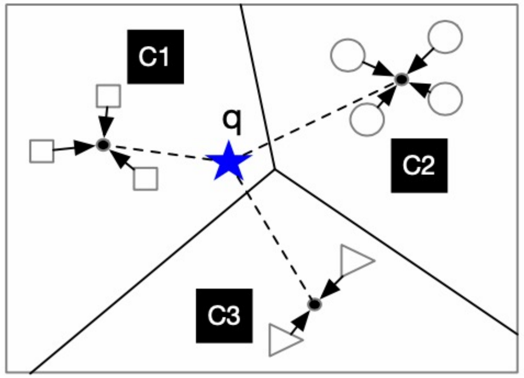
- Allow users to enroll new commands (spoken terms) by recording only a few audio examples in a voice-based human-device interaction system.
- In practice, the number of both newly added commands and pre-recorded audio examples for each command should not be limited.



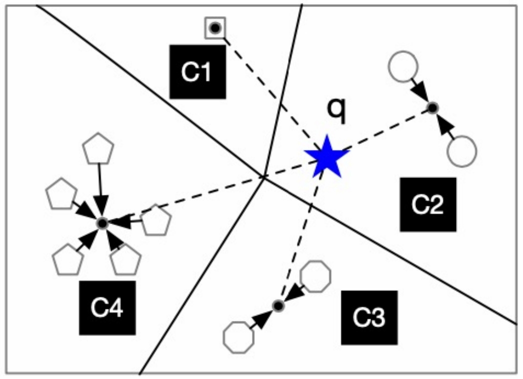
Prototypical networks for few-shot classification

- Learn with limited labelled data of new classes by using knowledge from previous classes.
- Often defined as N-way, K-shot
- In our work, N and K are flexible.

- Sample various few-shot classification tasks and train a backbone model using episodic training.
- It tends to minimize the within-class distance and maximize the between-class distance.



T1

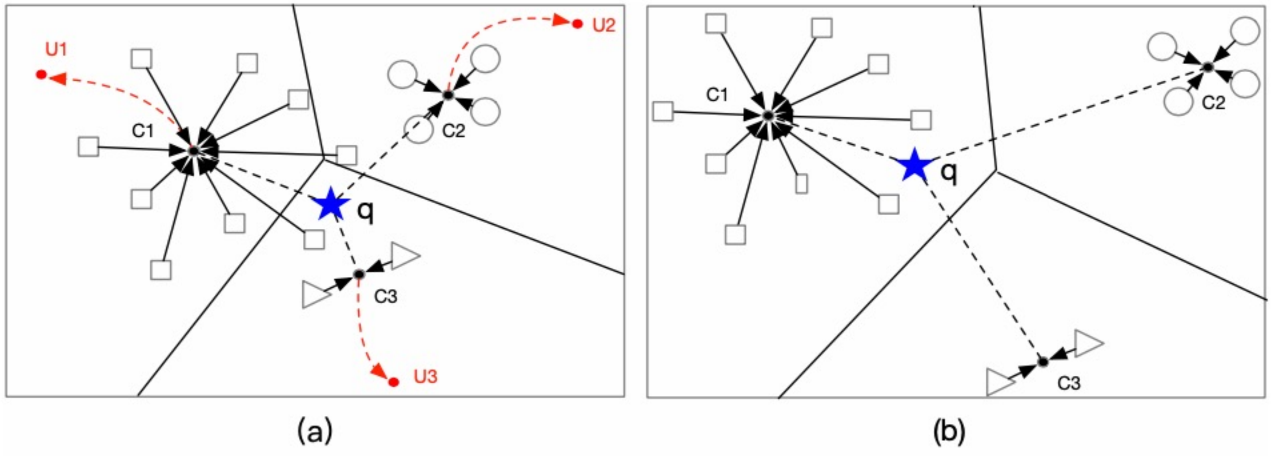


T2

⋮

Improved strategies towards varying classes and examples

- After investigating the effect of N and K in the training phase, we use a significant N and a varying K for training.
- We add a Max-Mahalanobis Center (MMC) loss-based regularizer to force the prototypical representations of different classes to move far apart from each other in the embedding space.



$$L_{\tau}^{reg} = \frac{1}{2} log \frac{\sum_i K_{\tau,i} ||\boldsymbol{c}_i - \boldsymbol{u}_i||_2^2}{\sum_i K_{\tau,i}}$$

$$L_{\tau}^{total} = L_{\tau} + \lambda L_{\tau}^{reg}$$

Experimental results

Table 2: *Accuracy with 95% confidence intervals of experiments on N+2-way, 5-shot classification tasks.*

Training	Testing 5+2-way	Testing 10+2-way	Testing N_τ+2-way
Superv.L.	27.52 ± 0.27	24.83 ± 0.38	-
Transf.L.	62.67 ± 0.38	54.43 ± 0.47	-
MAML	67.57 ± 0.91	63.22 ± 0.71	-
1+2-way	62.73 ± 0.12	52.32 ± 0.05	63.14 ± 0.21
2+2-way	74.33 ± 0.10	65.21 ± 0.05	54.42 ± 0.25
3+2-way	75.32 ± 0.10	66.38 ± 0.04	75.09 ± 0.16
5+2-way	76.38 ± 0.10	67.84 ± 0.04	76.47 ± 0.16
10+2-way	76.30 ± 0.09	67.92 ± 0.04	76.39 ± 0.15
15+2-way	76.28 ± 0.09	67.55 ± 0.04	76.23 ± 0.16
20+2-way	76.86 ± 0.09	$\mathbf{68.44 \pm 0.04}$	76.78 ± 0.15
N_τ**+2-way**	$\mathbf{76.90 \pm 0.09}$	68.13 ± 0.04	$\mathbf{76.82 \pm 0.16}$

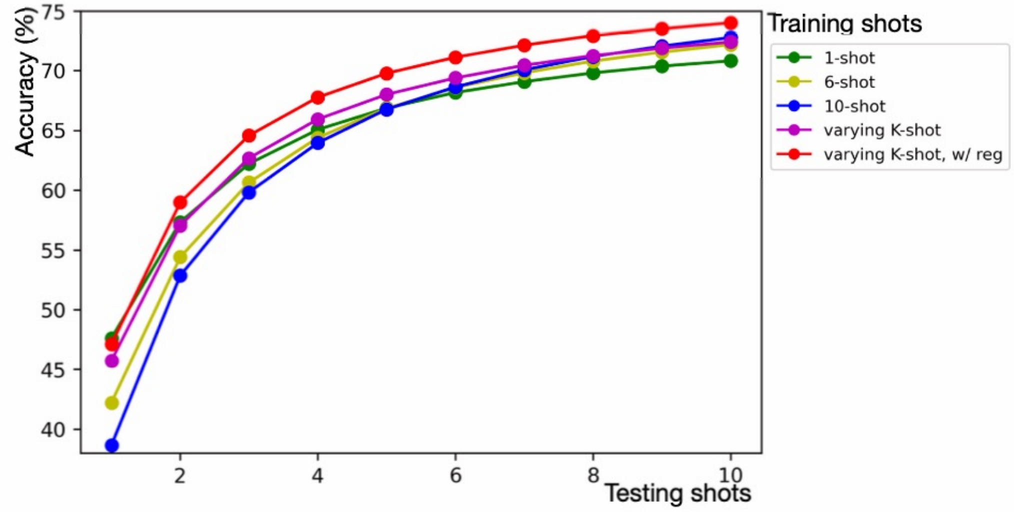


Figure 3: *Experiments on 20+2-way, K-shot tasks for training and 10+2-way, K-shot tasks for testing.*

Table 3: *Accuracy with 95% confidence intervals of experiments on 10+2-way, $K_{\tau,i}$-shot tasks for testing.*

Training	Testing $K_{\tau,i}$-shot
1-shot	66.32 ± 0.05
2-shot	65.94 ± 0.06
4-shot	66.20 ± 0.06
6-shot	64.96 ± 0.07
8-shot	64.11 ± 0.07
10-shot	64.60 ± 0.07
$K_{\tau,i}$**-shot**	$\mathbf{67.29 \pm 0.06}$
$K_{\tau,i}$**-shot (w/reg)**	$\mathbf{68.87 \pm 0.06}$

Experimental results

Table 4: *Accuracy with 95% confidence intervals of experiments on N_τ-way, $K_{\tau,i}$-shot tasks for testing.*

Training	**Testing** N_τ**-way,** $K_{\tau,i}$**-shot**
1+2-way, 1-shot	75.02 ± 0.17
1+2-way, 5-shot	74.92 ± 0.17
5+2-way, 1-shot	75.02 ± 0.17
5+2-way, 5-shot	74.92 ± 0.17
10+2-way, 1-shot	75.23 ± 0.17
10+2-way, 5-shot	74.56 ± 0.17
20+2-way, 1-shot	74.90 ± 0.17
20+2-way, 5-shot	74.95 ± 0.17
20+2-way, 10-shot	72.88 ± 0.17
20+2-way, $K_{\tau,i}$**-shot**	$\mathbf{75.77 \pm 0.16}$
20+2-way, $K_{\tau,i}$**-shot (w/ reg)**	$\mathbf{77.21 \pm 0.16}$

Conclusion

- Prototypical networks learn discriminative representations for few-shot classification tasks.
- When testing in N-way, K-shot tasks with varying N and K, episodic training with a significant N and a varying K improves the final performance.
- The MMC loss strengthens representation learning of prototypical networks by moving the centers of different classes apart from each other.

THANK YOU!