An Investigation of Few-shot Learning in Spoken Term Classification

Yangbin Chen[1], Tom Ko[2], Lifeng Shang[3], Xiao Chen[3], Xin Jiang[3], Qing Li[4]

1. City University of Hong Kong
2. Southern University of Science and Technology
3. Huawei Noah's Ark Lab
4. The Hong Kong Polytechnic University

INTERSPEECH 2020

香港城市大學
City University of Hong Kong

HUAWEI

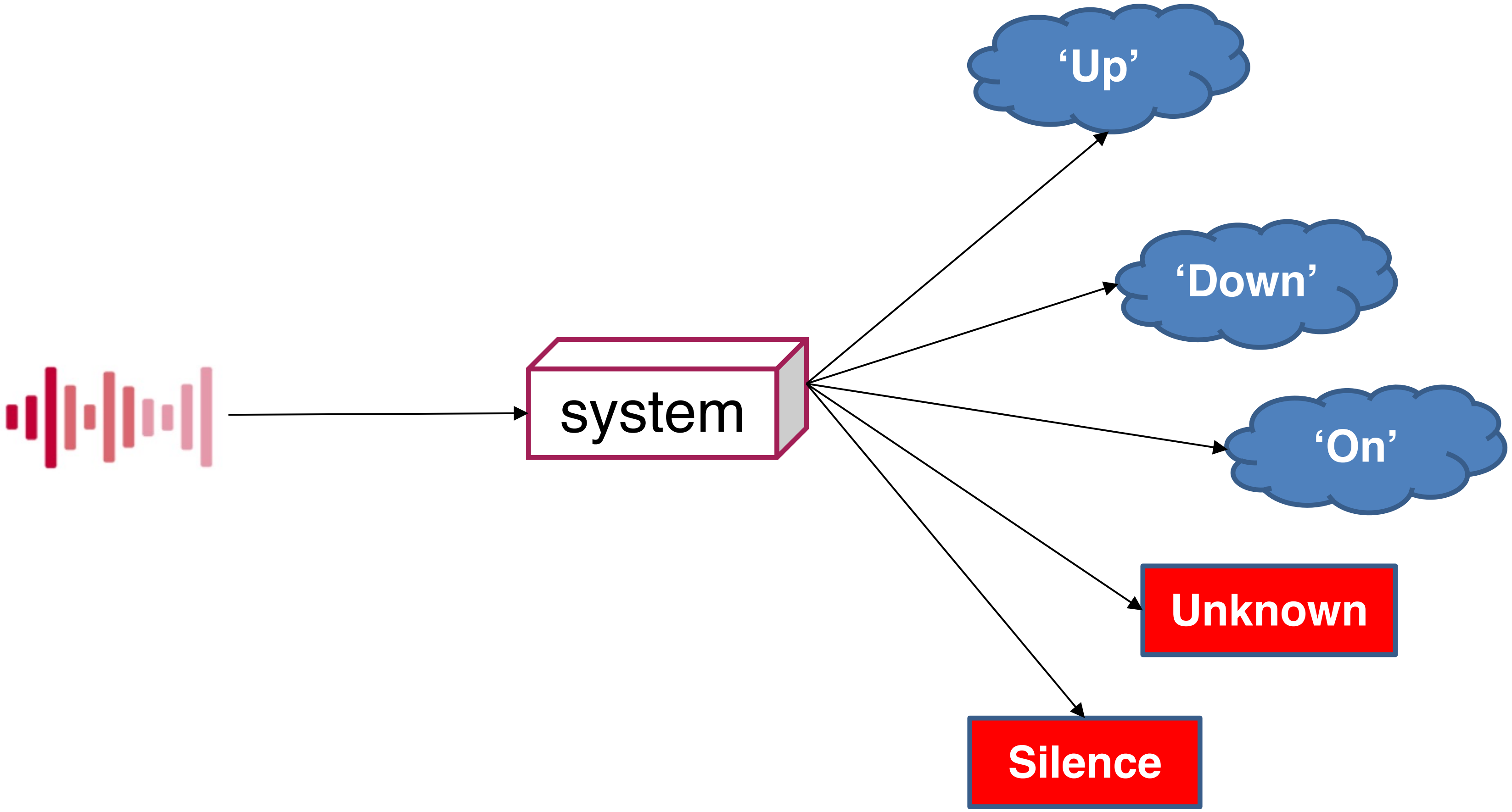
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Motivation

- In recent years, few-shot learning has drawn a lot of attention in the machine learning community.
- A lot of elegant solutions have been developed.
- It is worth to investigate the feasibility of applying few-shot learning methods to speech tasks.

Spoken Term Classification

- It aims to recognize spoken terms in the voice signal.

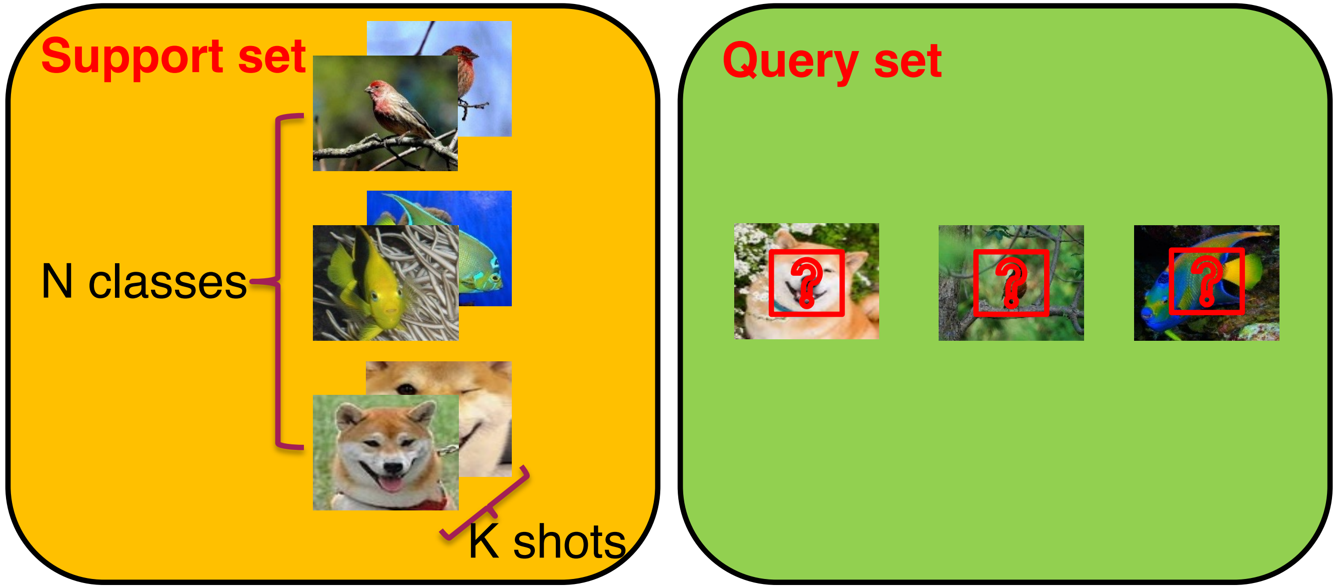


User-defined Spoken Term Classification

- Normally, the spoken term is predefined.
  - Given plenty of training data, conventional supervised learning could have solved the problem nicely.
- What about a user-defined scenario?
  - Users can define new spoken terms by providing a few audio examples.
- We formulate this problem as a few-shot learning problem, specifically, a few-shot classification task.

Few-Shot Classification

- Few-Shot Learning (FSL) Problem is a machine learning problem that learns with limited labeled data of target tasks by incorporating external source data, which has a different distribution from target data.
- Few-Shot Classification is a few-shot learning task, which is defined as N-way, K-shot, where
  - N is the number of classes in the target task
  - K is the number of examples per class

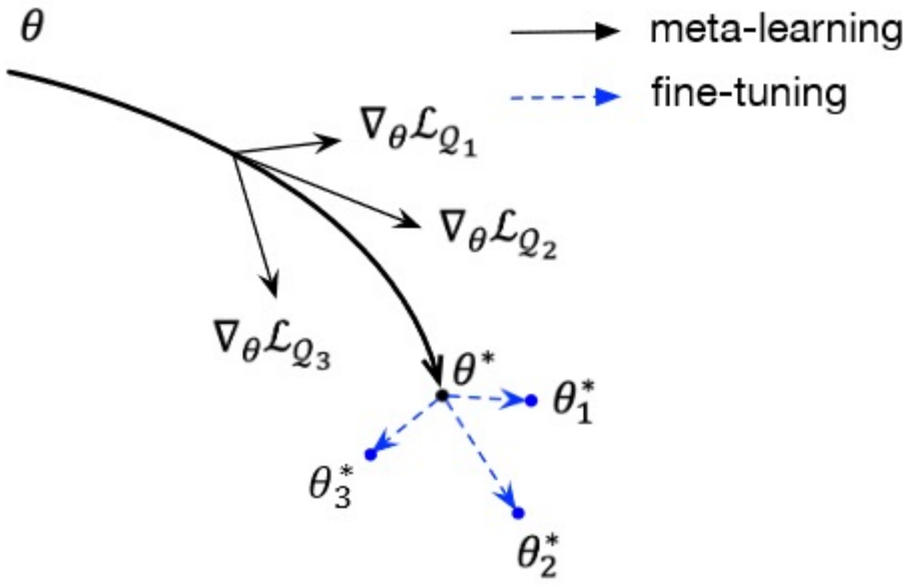


Meta-Learning

- Most popular solutions of few-shot learning problems right now use meta-learning.
- Also known as ‘learning to learn’, aims to make a quick adaptation to new tasks with only a few examples.
- Many elegant solutions are proposed:
  - Matching Network
  - Prototypical Network
  - Model-Agnostic Meta-Learning

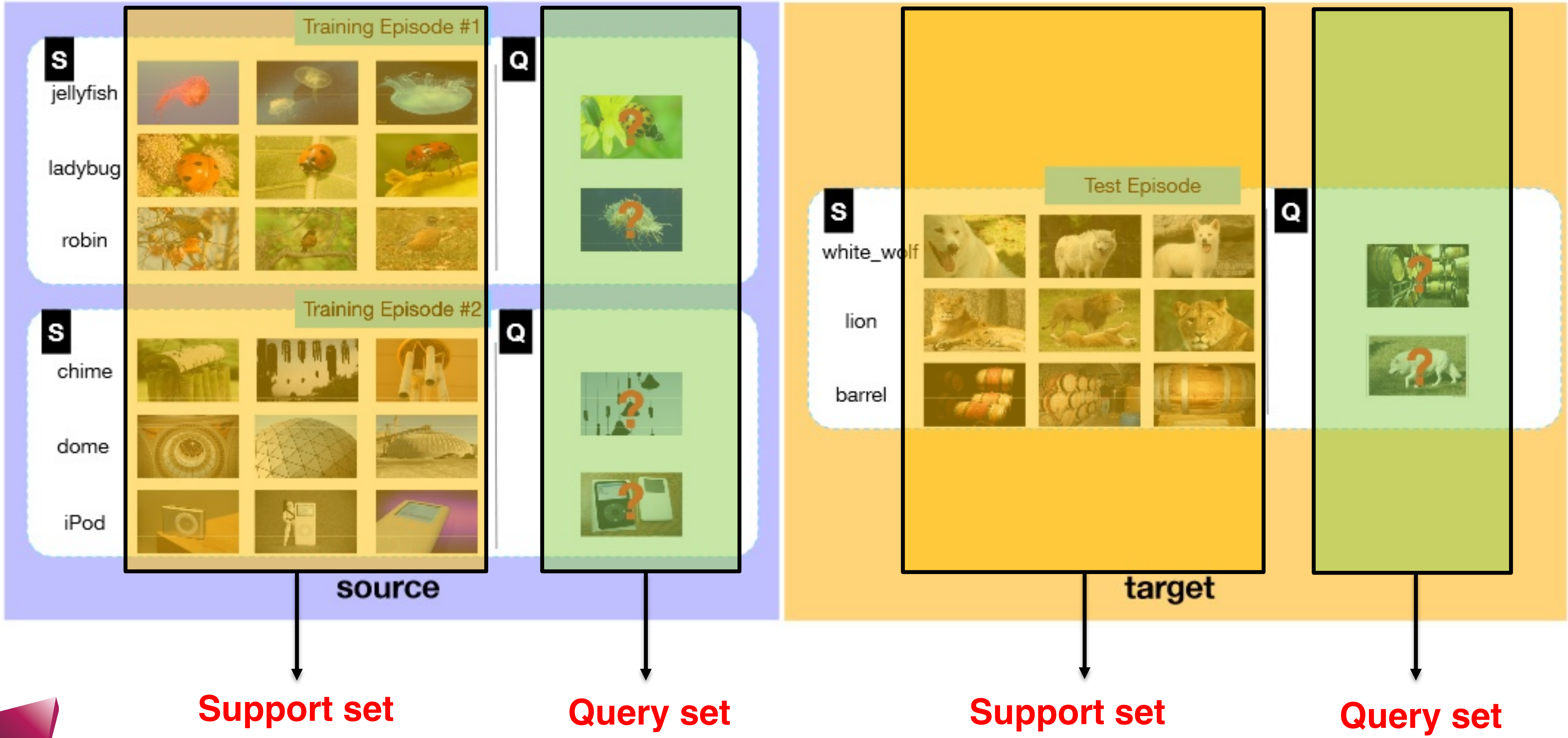
Model-Agnostic Meta-Learning (MAML)

- To train a model which can adapt to any new task using only a few labeled examples
- The model is trained on various tasks (meta-tasks) and it treats the entire task as a training example
- The model is forced to face different tasks so that it can get used to adapting to new tasks

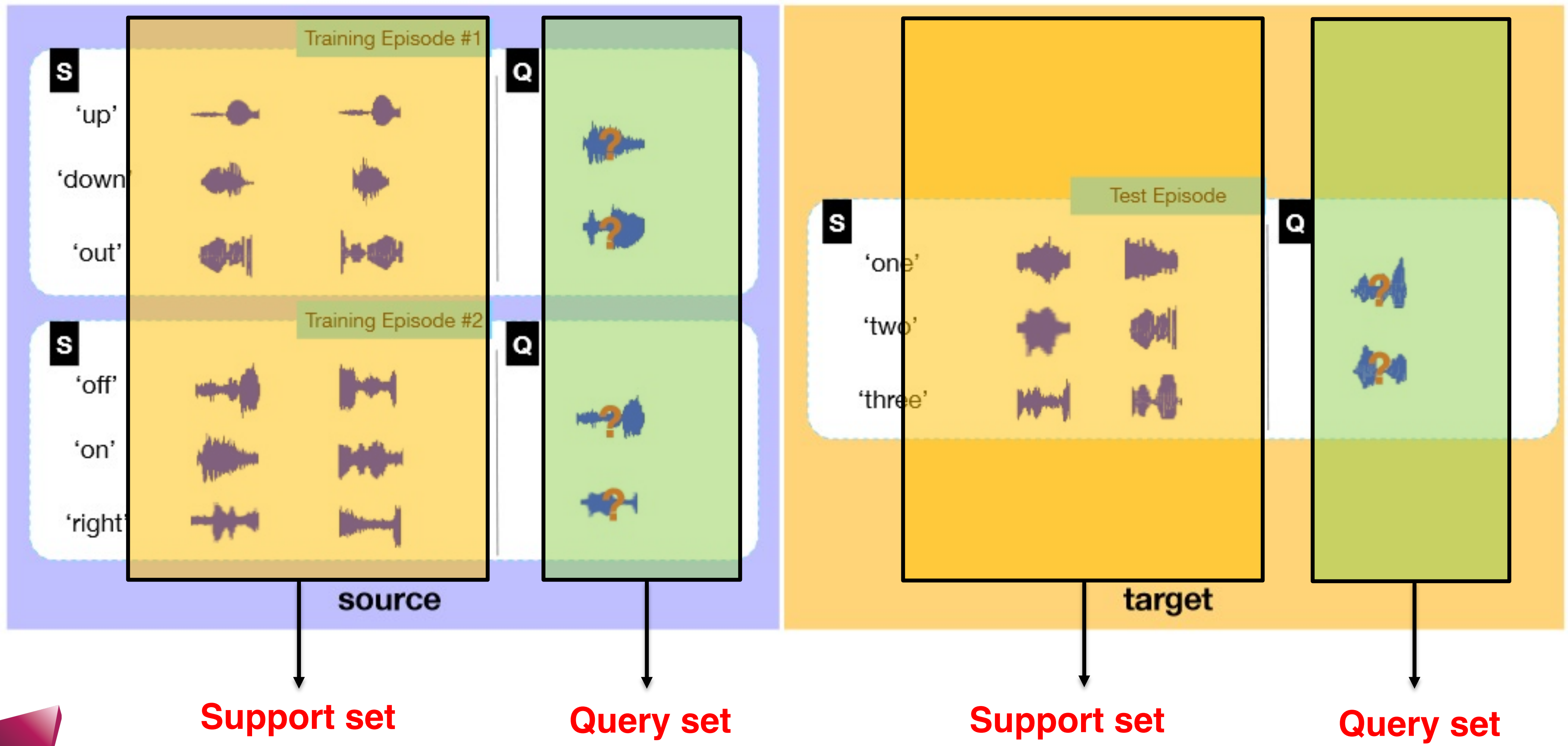


Chelsea Finn, Pieter Abbeel, Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks,"in Proceedings of the 34th ICML-Volume 70. JMLR. org, 2017, pp. 1126–1135.

MAML on Image Tasks



MAML on Speech Tasks



MAML – The Meta-learning Stage

- Given an initial model f_θ and a meta-task $\mathcal{T}_i$, a loss is computed with the support set:

$$\mathcal{L}_{S_i}(f_\theta) = -\sum_{(\boldsymbol{x}_j, \boldsymbol{y}_j) \in S_i} \boldsymbol{y}_j log f_\theta(\boldsymbol{x}_j) \qquad (1)$$

inner loop

- Then a gradient update is done:

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{S_i}(f_\theta) \qquad (2)$$

- Then another loss is computed with the query set:

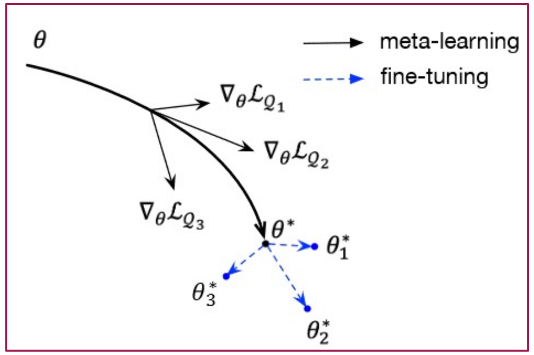
$$\mathcal{L}_{Q_i}\left(f_{\theta_i'}\right) = -\sum_{(\boldsymbol{x}_u', \boldsymbol{y}_u') \in Q_i} \boldsymbol{y}_u' log f_{\theta'}(\boldsymbol{x}_u') \qquad (3)$$

outer loop

- A gradient is computed on equation (3) with respect to θ, the model is updated:

$$\theta^* \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{Q_i}(f_{\theta_i'}) \;/\; \theta^* \leftarrow \theta - \beta \nabla_\theta \textstyle\sum_i \mathcal{L}_{Q_i}(f_{\theta_i'}) \qquad (4)$$

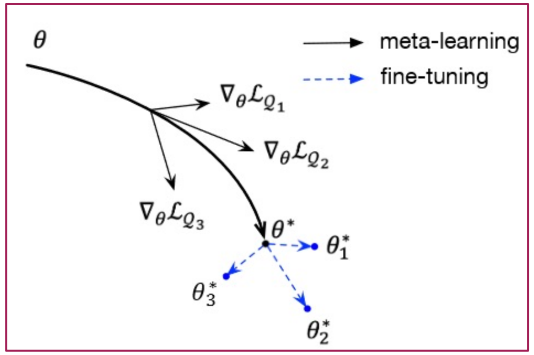
- This is a **second-order gradient optimization**.



MAML – The Fine-tuning Stage

- Before evaluation, the model will be fine-tuned for a few iterations according to the equation (2):

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{S_i}(f_\theta)$$



Extend the Few-Shot Classification Problem

- In most few-shot studies, all the classes are assumed to be new.
- In real-life applications, some of the classes are known.
- We define an N+M-way, K-shot problem where
  - M is the number of **fixed** classes
  - N is the number of **new** classes in the target task
  - K is the number of examples of each **new** class

Our approach – Extended MAML

- We fix the output positions of the fixed classes in the neural network classifier.
- The fixed classes occur in every meta-task in the meta-learning stage.
- The adaptation of fixed classes is not needed in the fine-tuning stage as they have already been learned in the meta-learning stage.

Few-Shot Spoken Term Classification

- 10+2-way, K-shot
- 10 keywords
- 2 fixed class: silence and unknown
- In the meta-learning stage, meta-tasks are randomly formed from a pool of keywords.

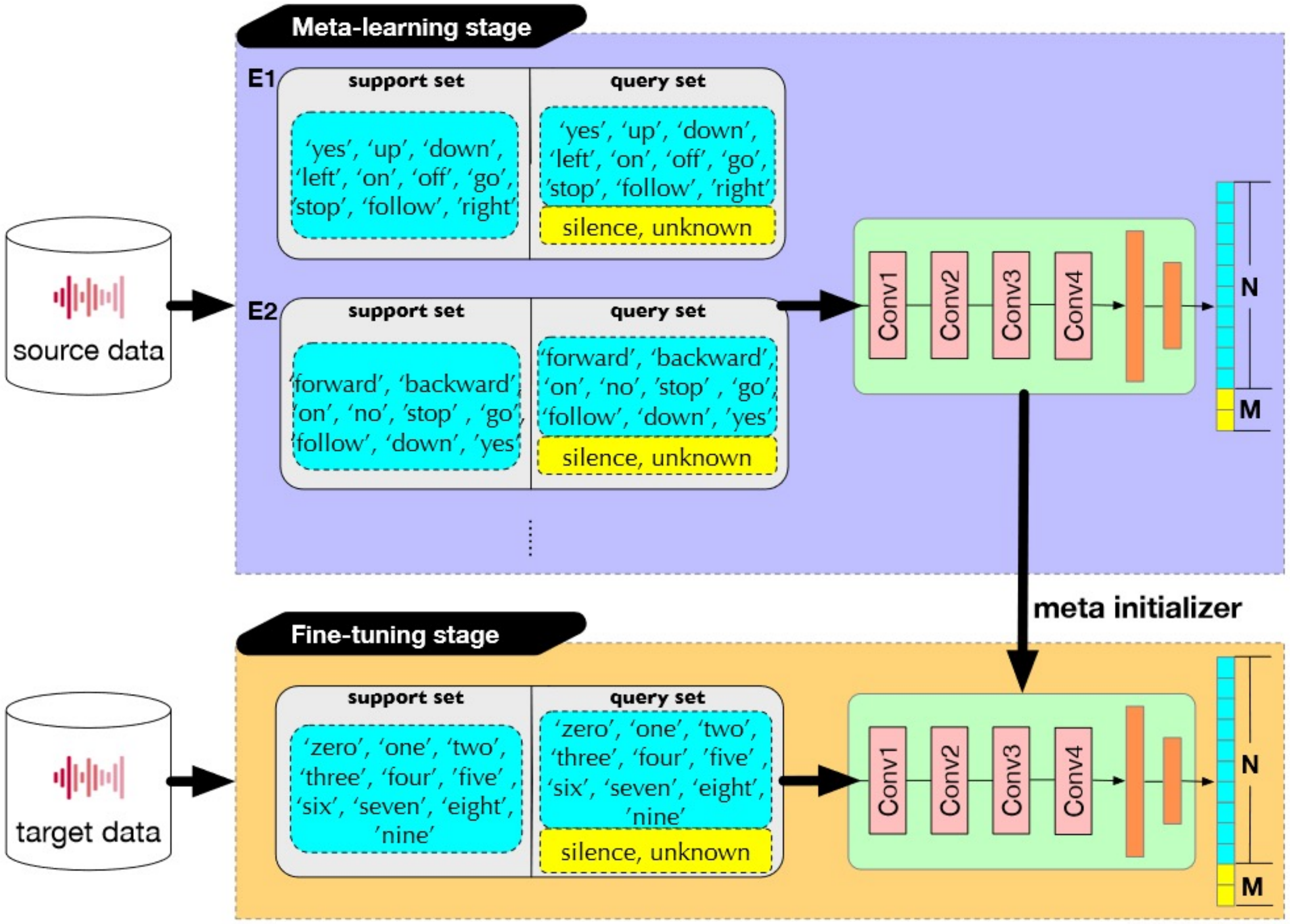


Fig. 1. Framework of our extended-MAML approach for few-shot spoken term classification.

The Algorithm

Algorithm 1 extended-MAML approach for few-shot spoken term classification

Require: $p(\mathcal{T})$: distribution over tasks
Require: $\mathcal{X}$: training keywords set
Require: $\mathcal{S}_{il}$: silence class set, $\mathcal{U}_{nk}$: unknown class set
Require: $\mathcal{S}_i$: support set, $\mathcal{Q}_i$: query set
Require: α, β: learning rates

1: Randomly initialize base model parameters $\boldsymbol{\theta}$
2: **while** not done **do**
3:     Sample a batch of meta-tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4:     **for all** $\mathcal{T}_i$ **do**
5:         Sample a support set $\mathcal{S}_i$ from $\mathcal{X}$
6:         Compute the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}_i}(f_{\boldsymbol{\theta}})$ using $\mathcal{S}_i$ and $\mathcal{L}_{\mathcal{S}_i}(f_{\boldsymbol{\theta}})$
7:         Update base model parameters with gradient descent: $\boldsymbol{\theta}'_i = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}_i}(f_{\boldsymbol{\theta}})$ ▷ step 6 and step 7 can be repeated for several times
8:         Sample a query set $\mathcal{Q}_i$ from the union $\{\mathcal{X}, \mathcal{S}_{il}, \mathcal{U}_{nk}\}$ ▷ selected keywords from $\mathcal{X}$ in $\mathcal{Q}_i$ and $\mathcal{S}_i$ within $\mathcal{T}_i$ are the same
9:         Compute the loss $\mathcal{L}_{\mathcal{Q}_i}(f_{\boldsymbol{\theta}'_i})$ using $\mathcal{Q}_i$ and the updated model $f_{\boldsymbol{\theta}'}$
10:     **end for**
11:     Update parameters $\boldsymbol{\theta}$ using each $\mathcal{Q}_i$ and $\mathcal{L}_{\mathcal{Q}_i}(f_{\boldsymbol{\theta}'})$: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}} \sum_i \mathcal{L}_{\mathcal{Q}_i}(f_{\boldsymbol{\theta}'_i})$
12: **end while**

Experimental Setup

- Google Speech Commands dataset (v0.02)
- 105,829 1-second audio clips of 35 keywords
- We formulate two 10+2-way, K-shot tasks using the same setup as the “Audio Recognition” tutorial in the official Tensorflow package
  - ten keywords, silence, and unknown
  - Digits classification, which uses digits zero to nine as ten keywords
  - Commands classification, which contains ten keywords as: “yes”, “no”, “up”, “down”, “left”, “right”, “on”, “off”, “stop”, or “go”

Model Setup

- 40 dimensional MFCCs
- CNN based model which contains 4 convolutional blocks
- Each block comprises a 3 x 3 convolutions and 64 filters

Baselines

- Two baselines:
  - Conventional supervised learning approach
  - Original MAML (which treats the 10+2 way problem as a 12-way problem)

Results on Digits Classification

Table 1. Accuracy with 95% confidence intervals on **digits classification**

Methods	1-shot	5-shot	10-shot
Superv. L.	18.14 ± 0.44	24.83 ± 0.38	28.07 ± 0.34
MAML-ori	44.60 ± 0.98	60.88 ± 0.58	65.18 ± 0.62
MAML-ext	$\mathbf{47.42 \pm 0.96}$	$\mathbf{63.22 \pm 0.71}$	$\mathbf{69.48 \pm 0.47}$

Results on Commands Classification

Table 2. Accuracy with 95% confidence intervals on **commands classification**

Methods	1-shot	5-shot	10-shot
Superv. L.	17.03 ± 0.48	22.42 ± 0.33	25.6 ± 0.26
MAML-ori	33.35 ± 0.80	50.31 ± 0.50	57.34 ± 0.41
MAML-ext	$\mathbf{39.54 \pm 0.62}$	$\mathbf{52.20 \pm 0.51}$	$\mathbf{59.36 \pm 0.39}$

Observations

- The overall accuracy in digit classification is better than in command classification.
  - This implies that, in a user-defined scenario, the system performance will be affected by the keywords users pick.
- MAML based approaches perform much better than conventional supervised learning in a few-shot situation.
- Our proposed approach outperforms the original MAML.
  - We attribute the improvement to the use of prior information of the fixed classes.

User-defined vs. Predefined

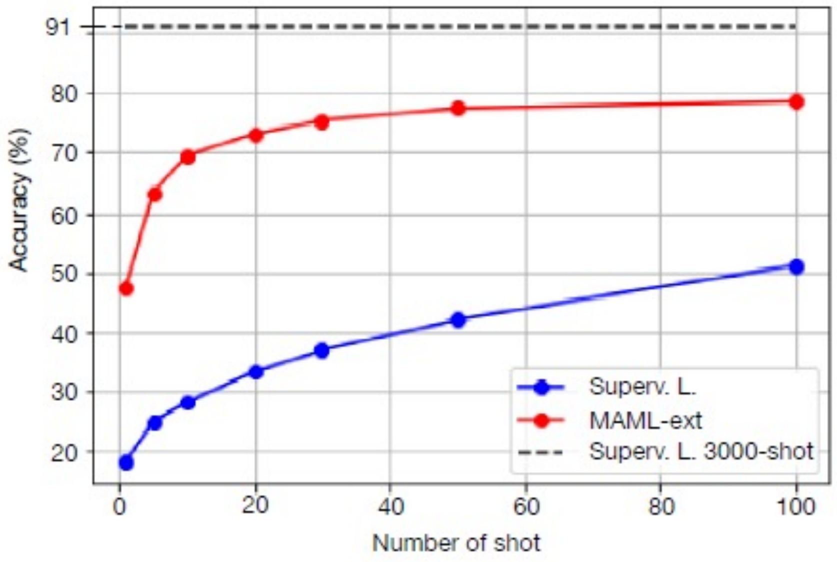


Fig. 2. Accuracy with changing shot on **digits classification**.

Conclusion

- In this piece of work, we formulate a user-defined scenario of spoken term classification as a few-shot learning problem.
- We define a N+M-way K-shot problem which we believe is a more realistic problem.
- We solve the problem by extending the original MAML.

Future Work

- There is a performance gap between a user-defined system and a predefined system.
- Narrow the gap with data augmentation techniques.
- Explore other meta learning methods.