MAML is a Noisy Contrastive Learner in Classification

Chia-Hsiang Kao

Wei-Chen Chiu

Pin-Yu Chen

National Yang Ming Chiao Tung University

MIT-IBM Watson AI Lab

Take Home Message

Under a mild assumption, we show that **MAML (model-agnostic meta-learning) is a noisy supervised contrastive learning algorithm** in a few-shot classification paradigm.

Why is MAML effective in learning general-purpose representations?

- Because MAML implicitly exploits contrastive learning.

What is the role of support and query data in MAML?

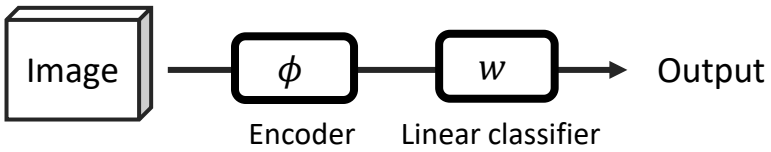
- In first-order MAML, the features of support data act as the prototypes, guiding the update of the features of query data.

What is the role of inner loops and outer loops in MAML?

- In the inner loop, the features of support data are memorized by the linear classifier. Therefore, in the outer loop, the SoftMax output of the query data contains the inner products between the support features and the query feature.

A Motivating Example

Model structure



Condition

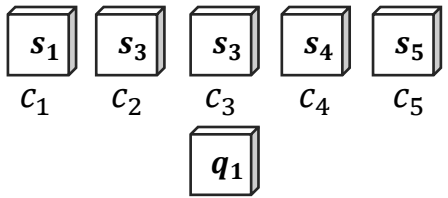
Linear classifier is zeroed ($w = 0$)
at the start of an outer loop

Algorithm

MAML, with
- One inner loop update.
- Inner loop loss function: mean square error.
- Outer loop loss function: mean square error.

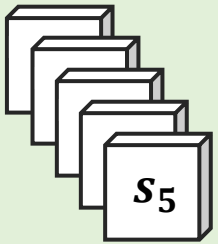
A few-shot learning setting

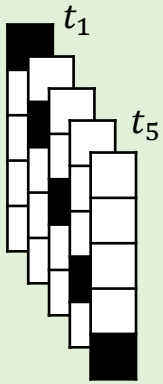
- 5-way: Each task contains 5 classes of images.
- 1-shot: Only one image per class in the support data.



Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

Model:



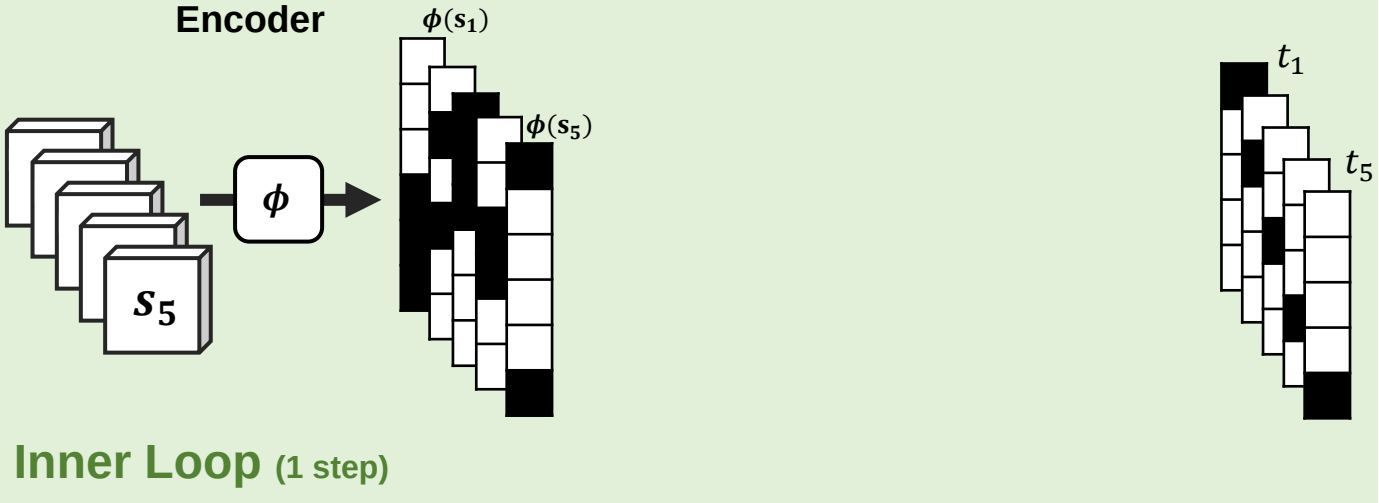


Inner Loop (1 step)

4

Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

Model:

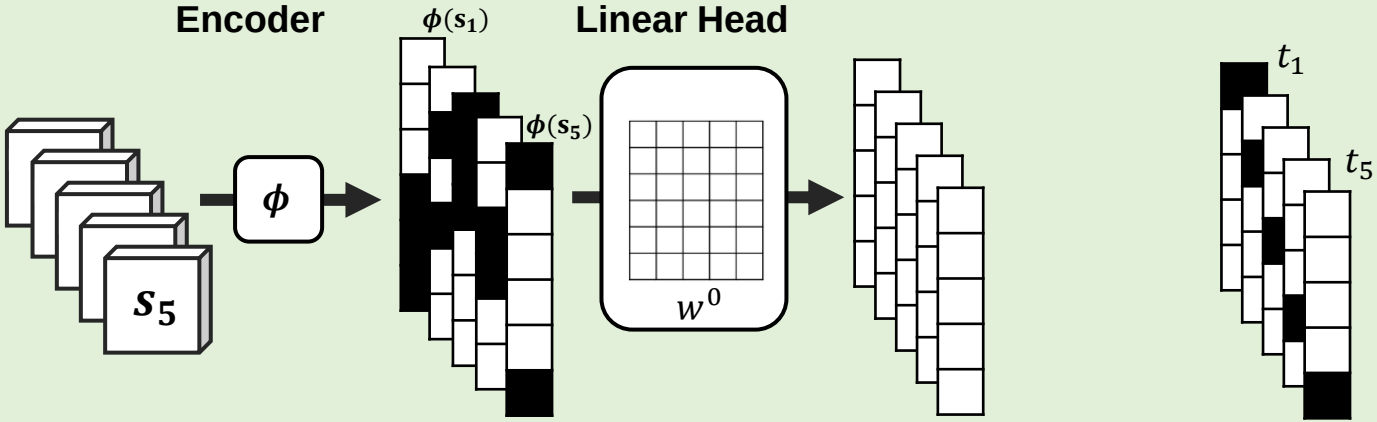


Inner Loop (1 step)

5

Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

Model:

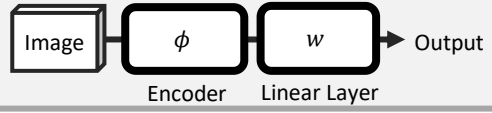


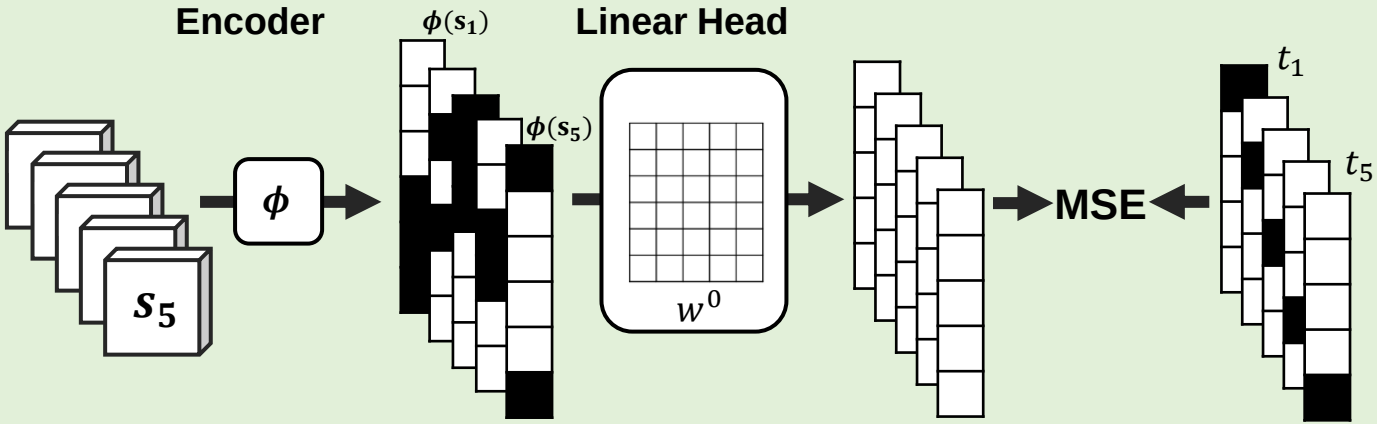
Inner Loop (1 step)

6

Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

Model:





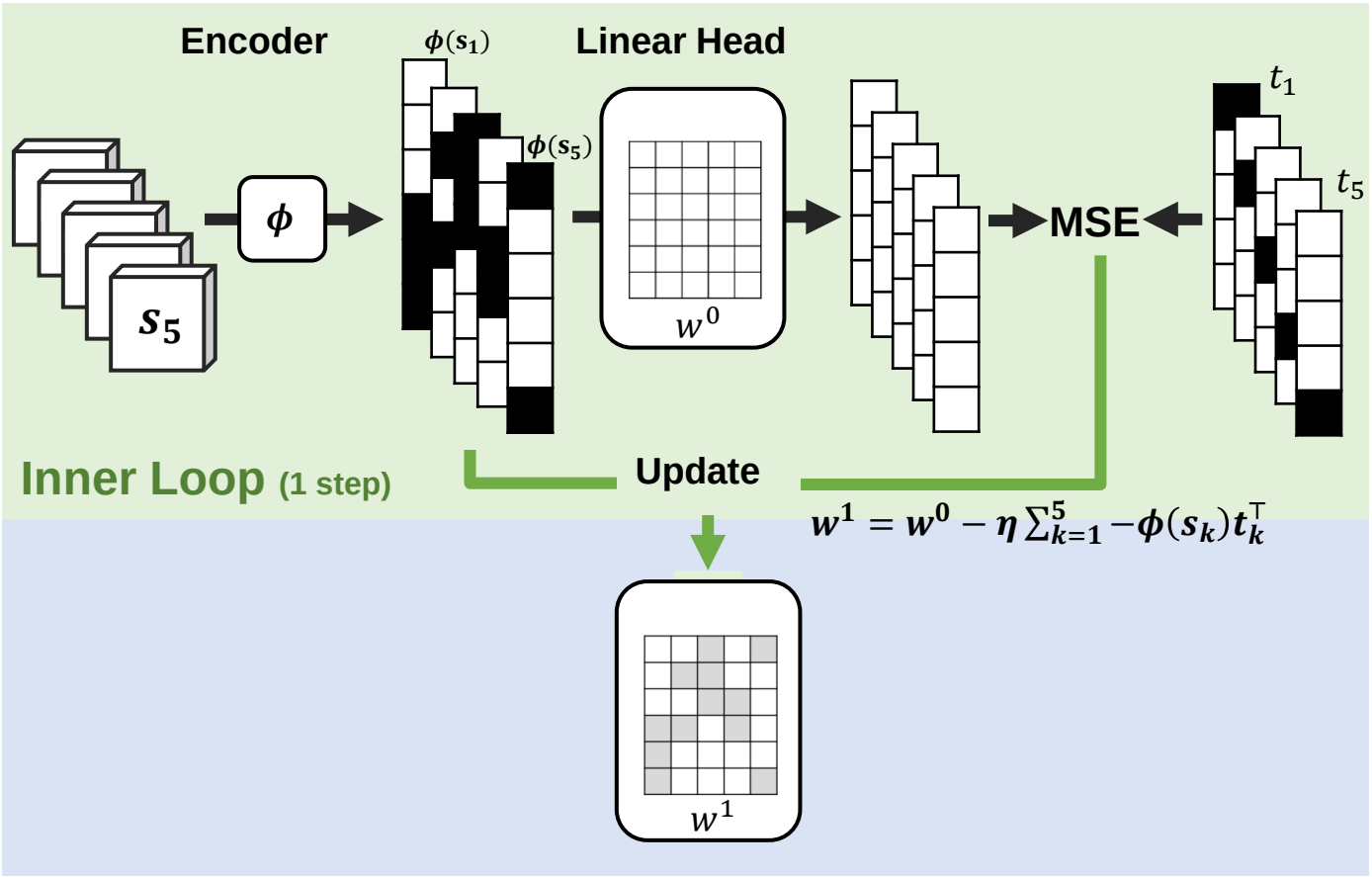
Inner Loop (1 step)

7

Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

Model: Image → ϕ (Encoder) → w (Linear Layer) → Output

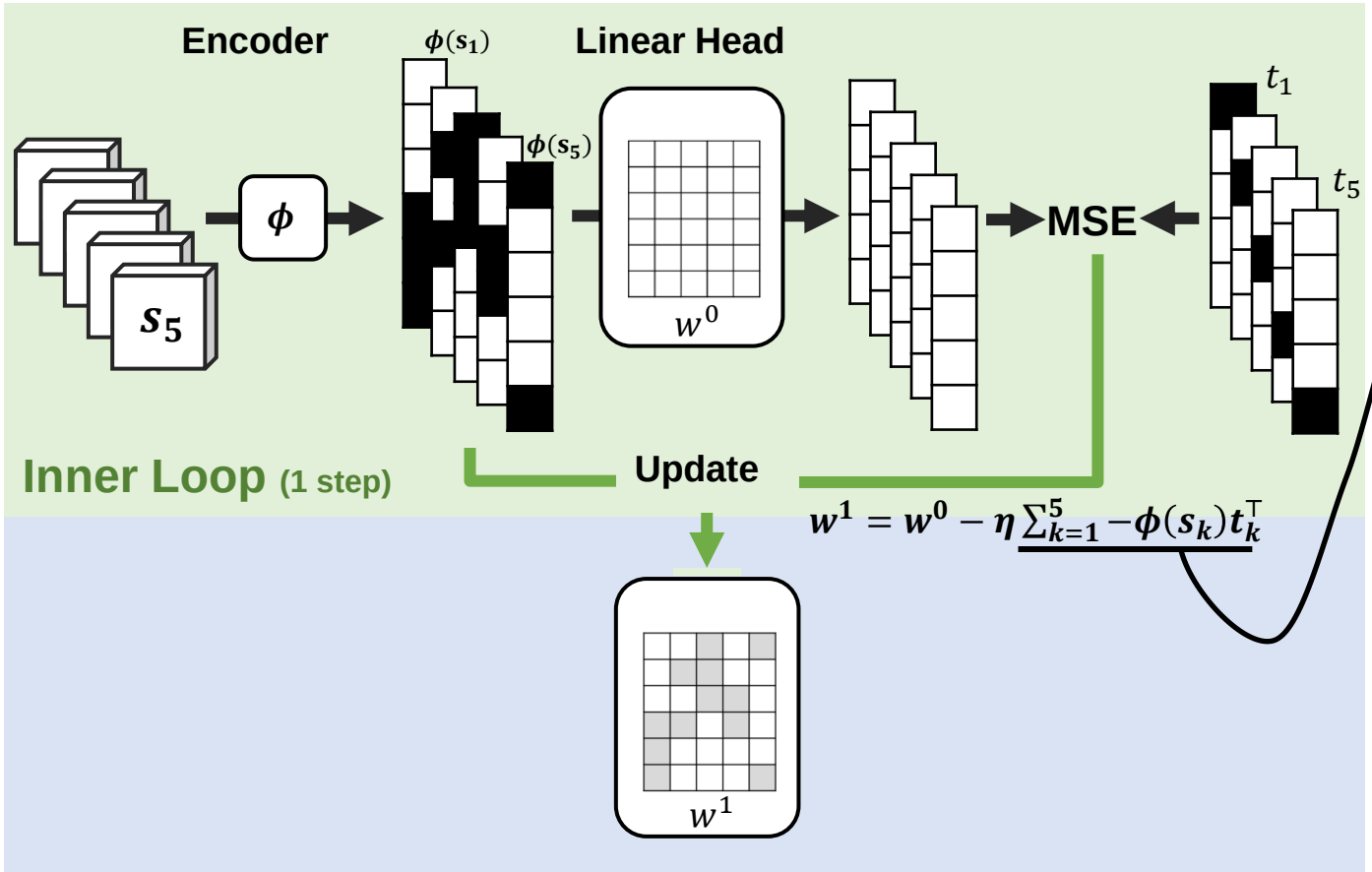
0 1

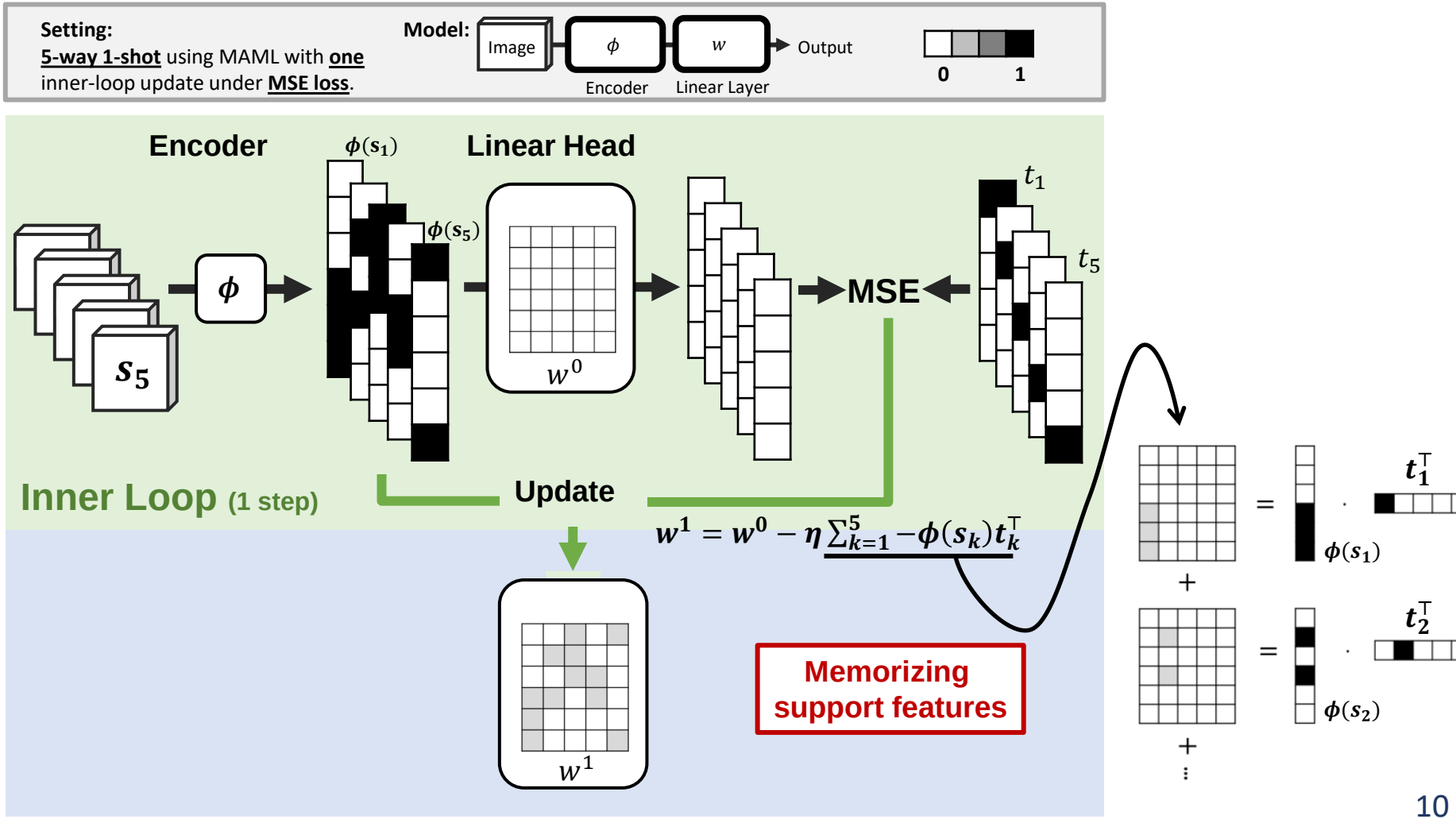


Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

Model: Image → ϕ (Encoder) → w (Linear Layer) → Output

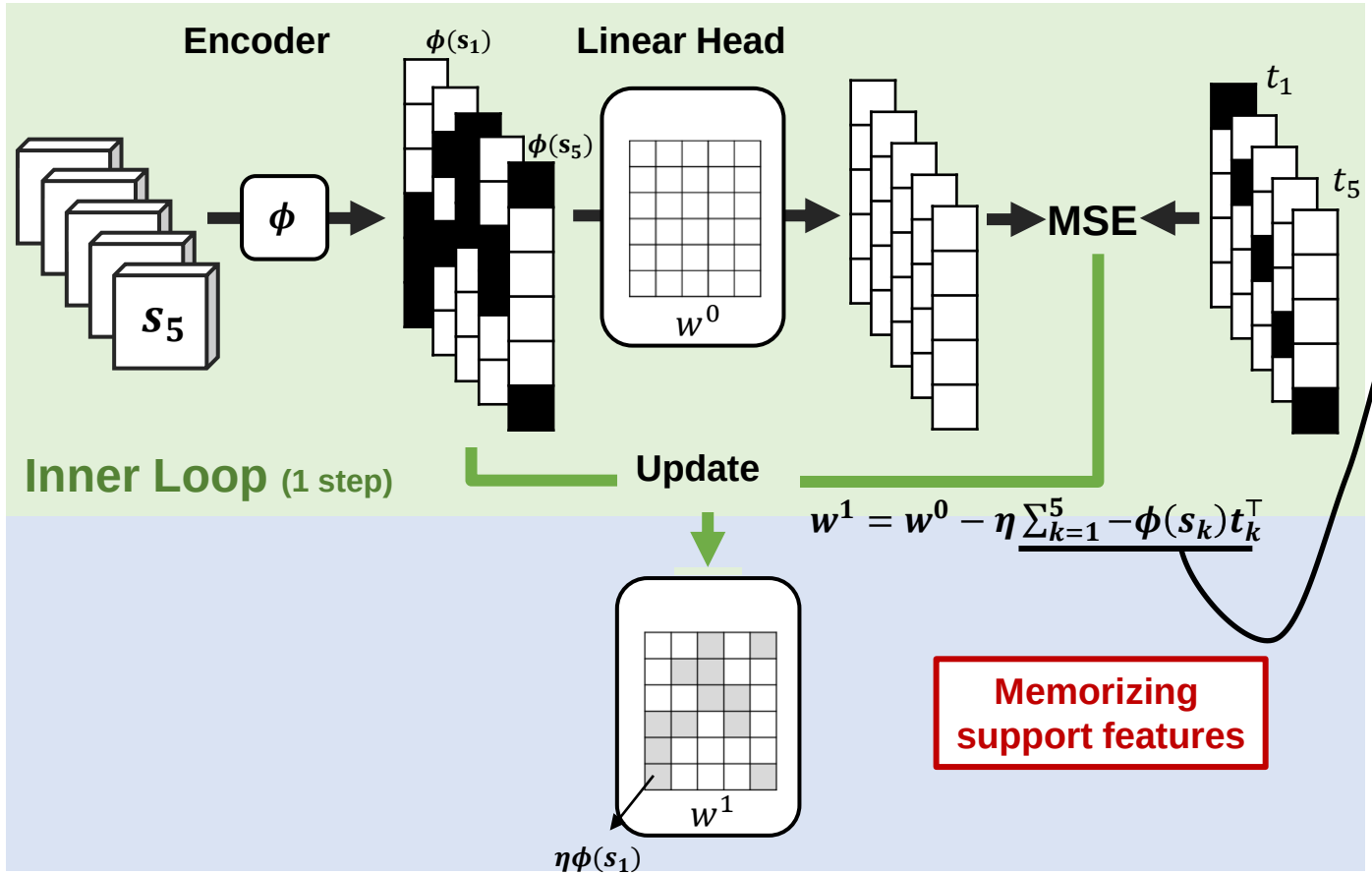
0 1




Setting:
5-way 1-shot using MAML with one inner-loop update under MSE loss.
Model:
Image
ϕ
w
Output
Encoder
Linear Layer
0
1
Encoder
$\phi(s_1)$
Linear Head
$\phi(s_5)$
t_1
t_5
s_5
ϕ
w^0
MSE
Inner Loop (1 step)
Update
$w^1 = w^0 - \eta \sum_{k=1}^{5} -\phi(s_k) t_k^\top$
w^1
Memorizing support features
$t_1^\top$
$\phi(s_1)$
$t_2^\top$
$\phi(s_2)$

Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

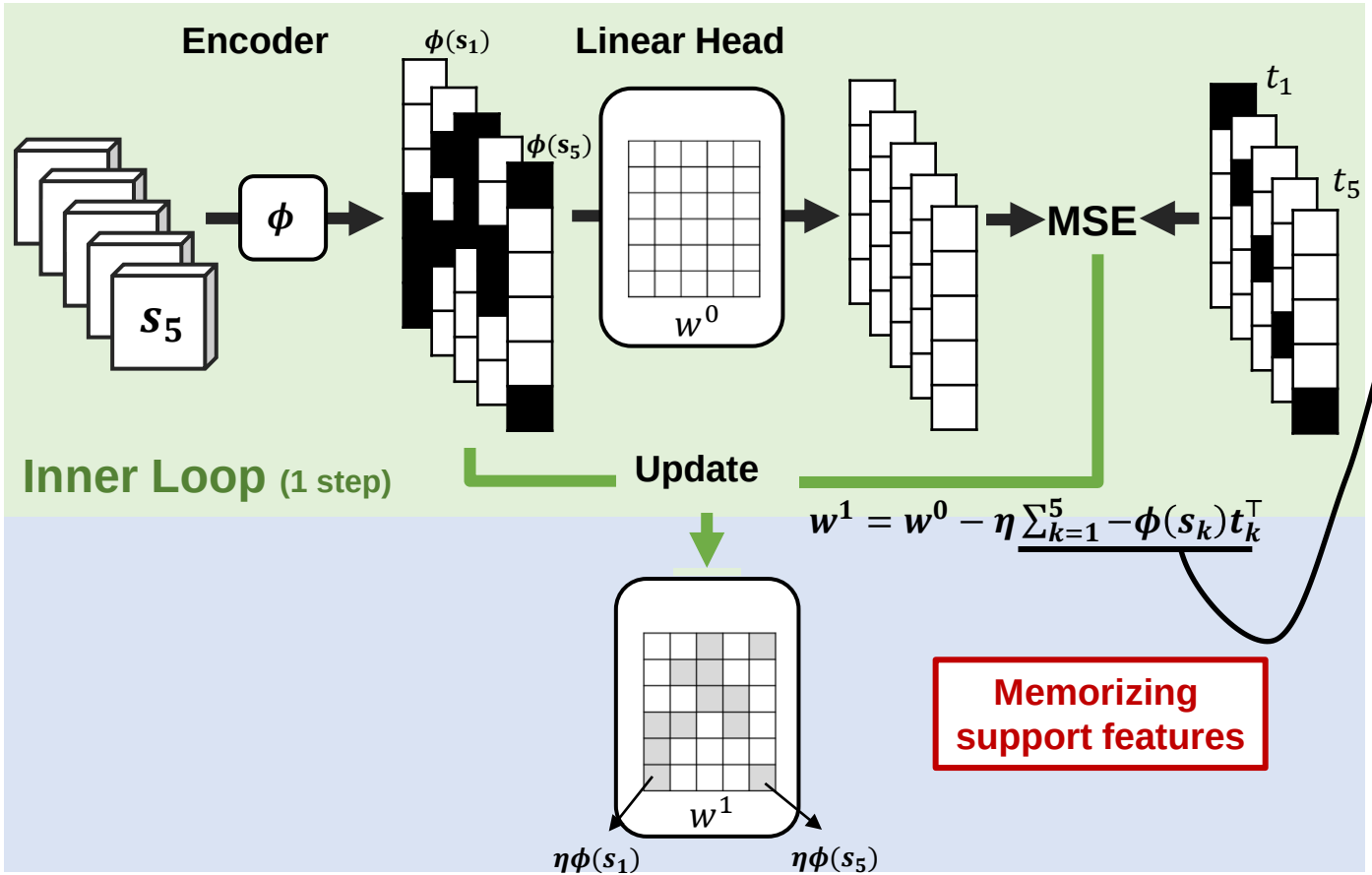
Model: Image → ϕ (Encoder) → w (Linear Layer) → Output

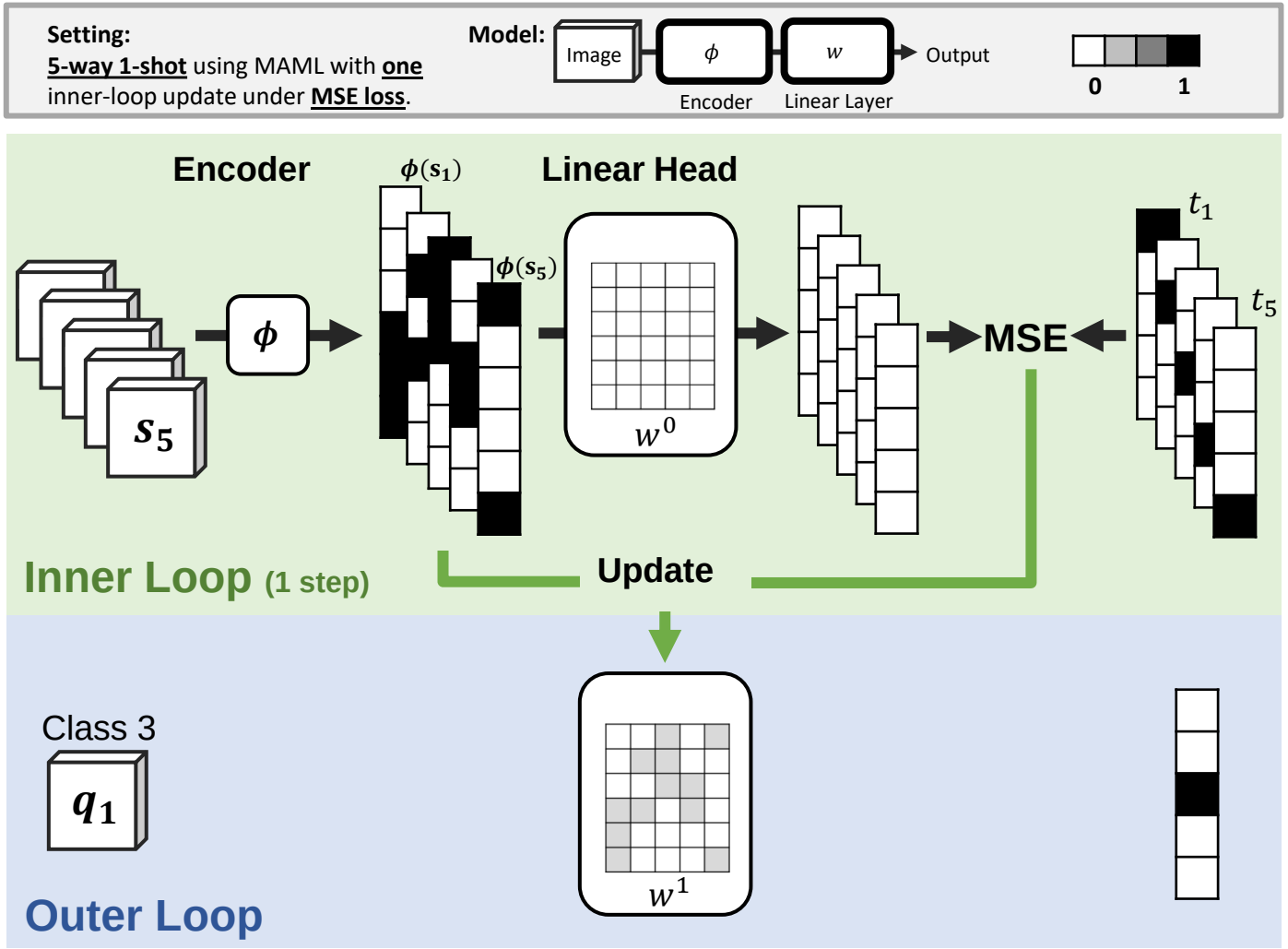


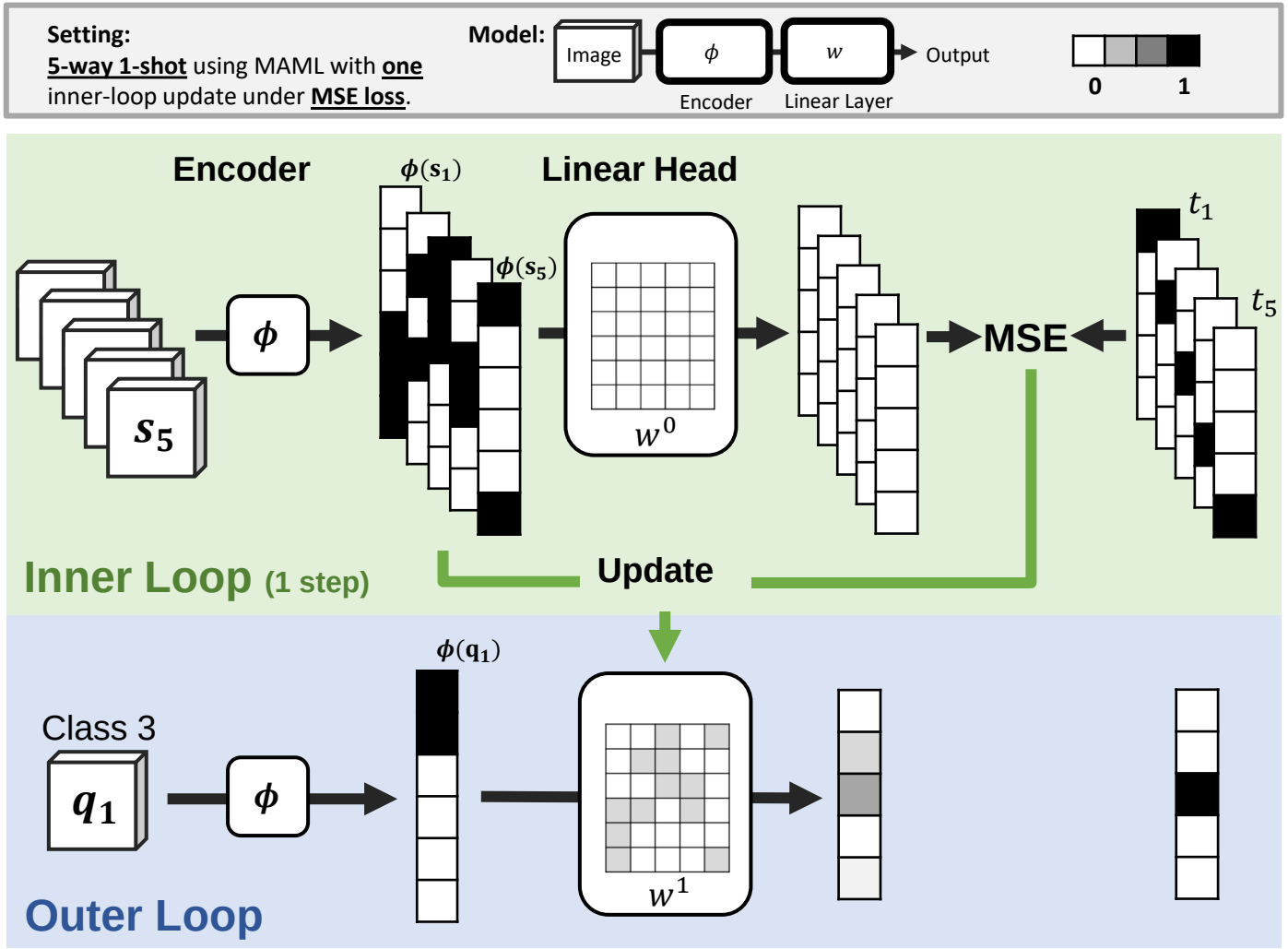
Setting:
5-way 1-shot using MAML with **one** inner-loop update under **MSE loss**.

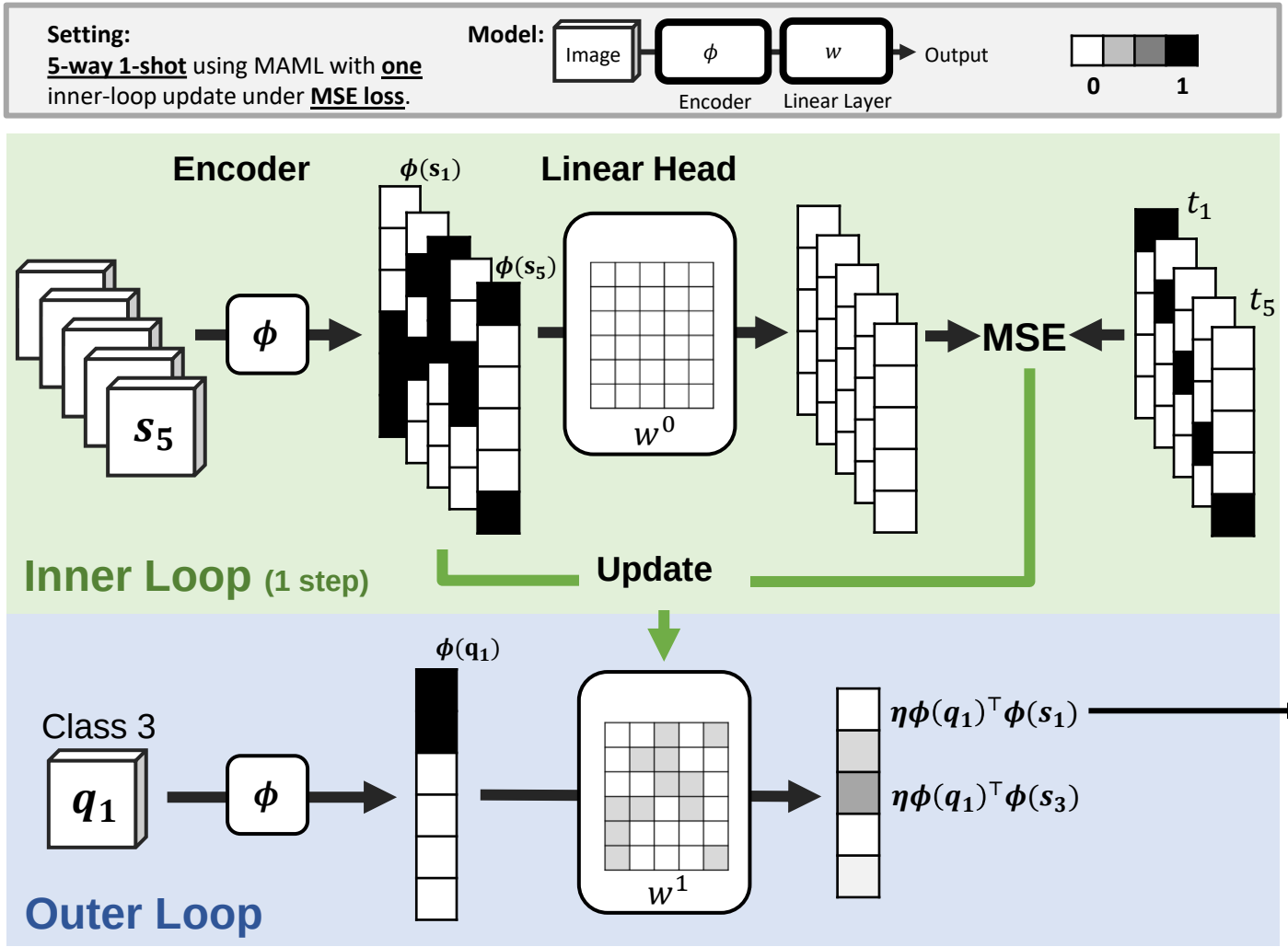
Model: Image → ϕ (Encoder) → w (Linear Layer) → Output

0 1

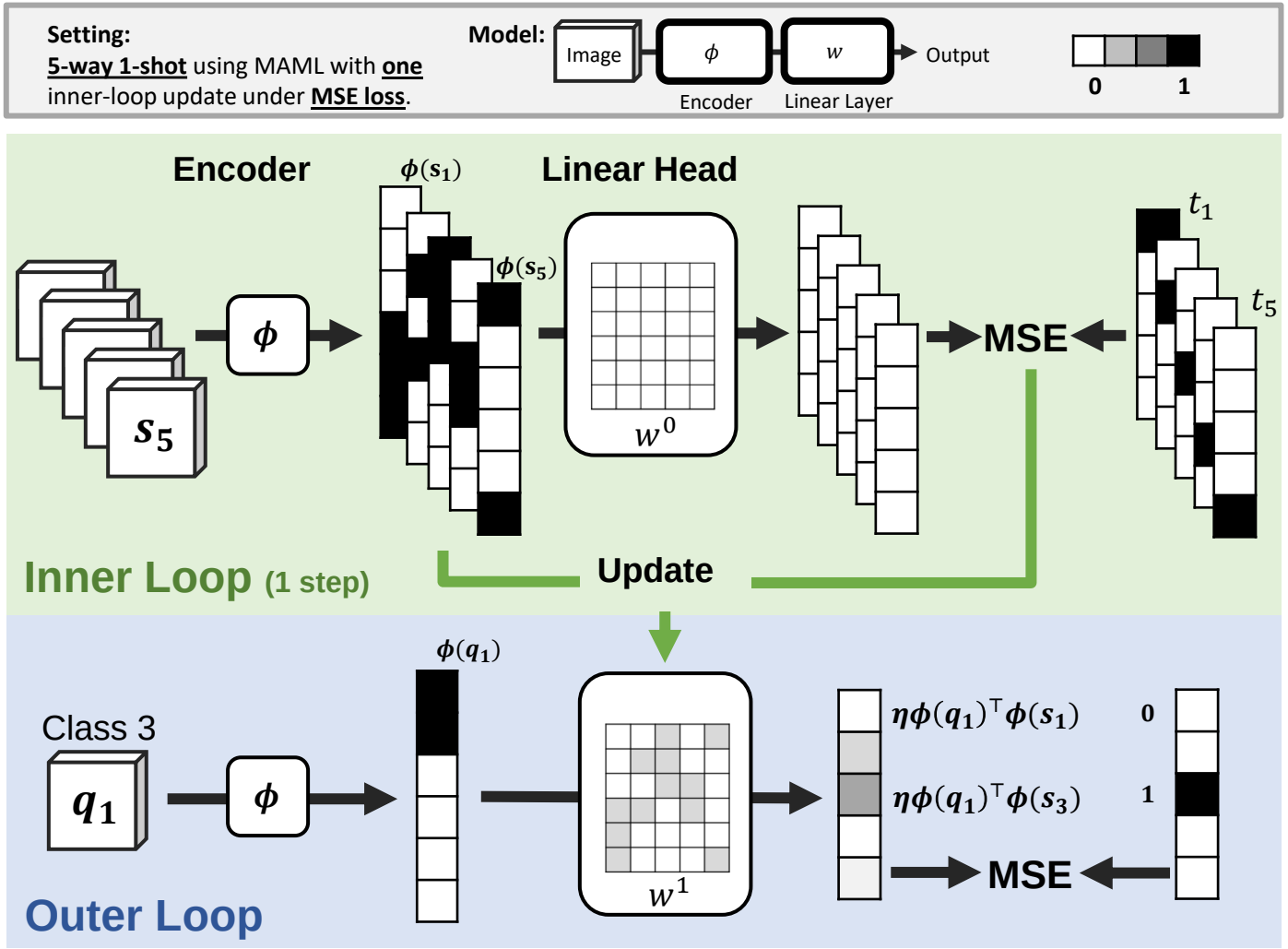


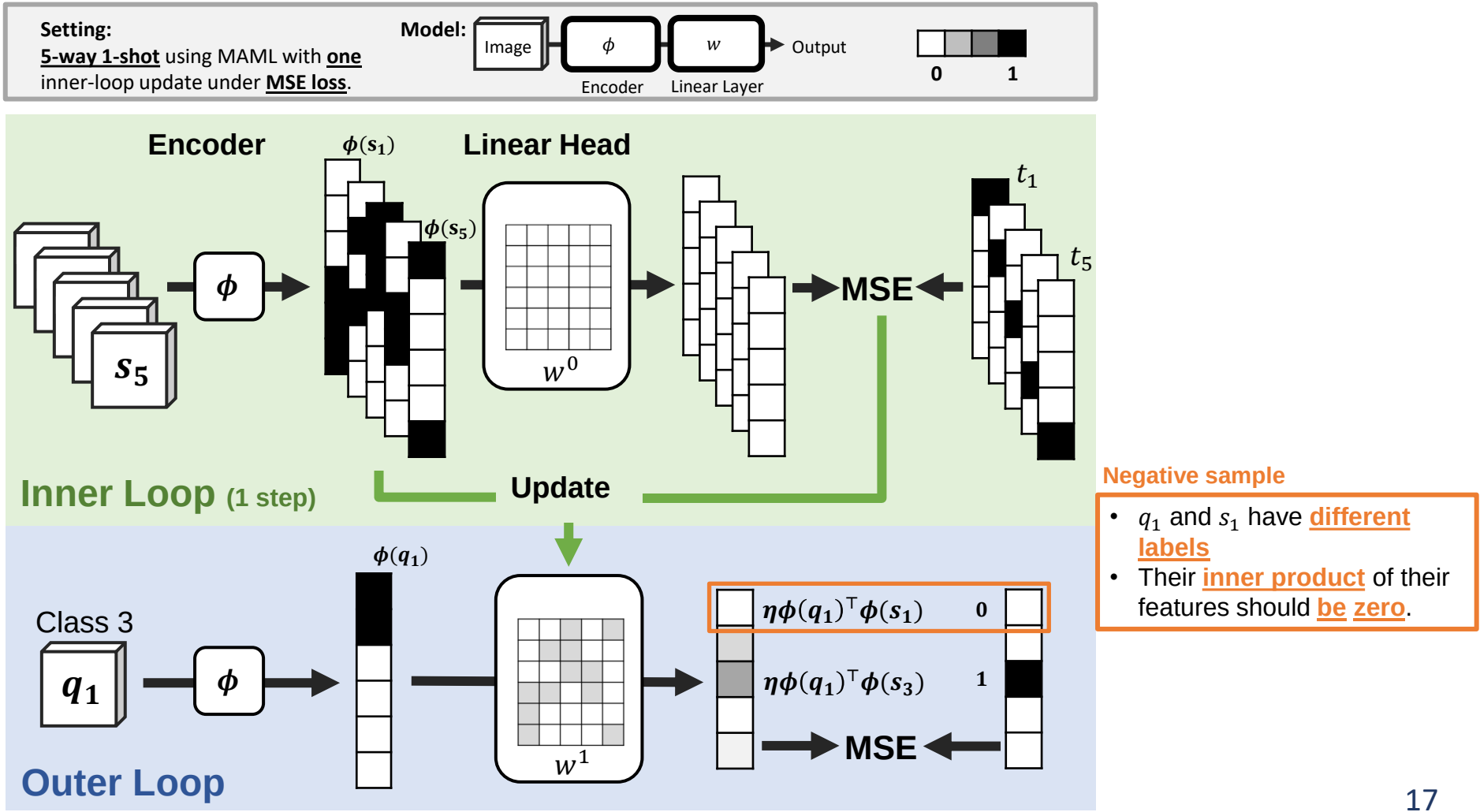

Setting:
5-way 1-shot using MAML with one inner-loop update under MSE loss.
Model:
Image
ϕ
w
Output
Encoder
Linear Layer
0
1
Encoder
$\phi(s_1)$
$\phi(s_5)$
Linear Head
s_5
ϕ
w^0
MSE
t_1
t_5
Inner Loop (1 step)
Update
Class 3
q_1
w^1
Outer Loop


Setting:
5-way 1-shot using MAML with one inner-loop update under MSE loss.
Model:
Image
ϕ
w
Output
Encoder
Linear Layer
0
1
Encoder
$\phi(s_1)$
Linear Head
$\phi(s_5)$
t_1
t_5
s_5
ϕ
w^0
MSE
Inner Loop (1 step)
Update
$\phi(q_1)$
Class 3
q_1
ϕ
w^1
Outer Loop



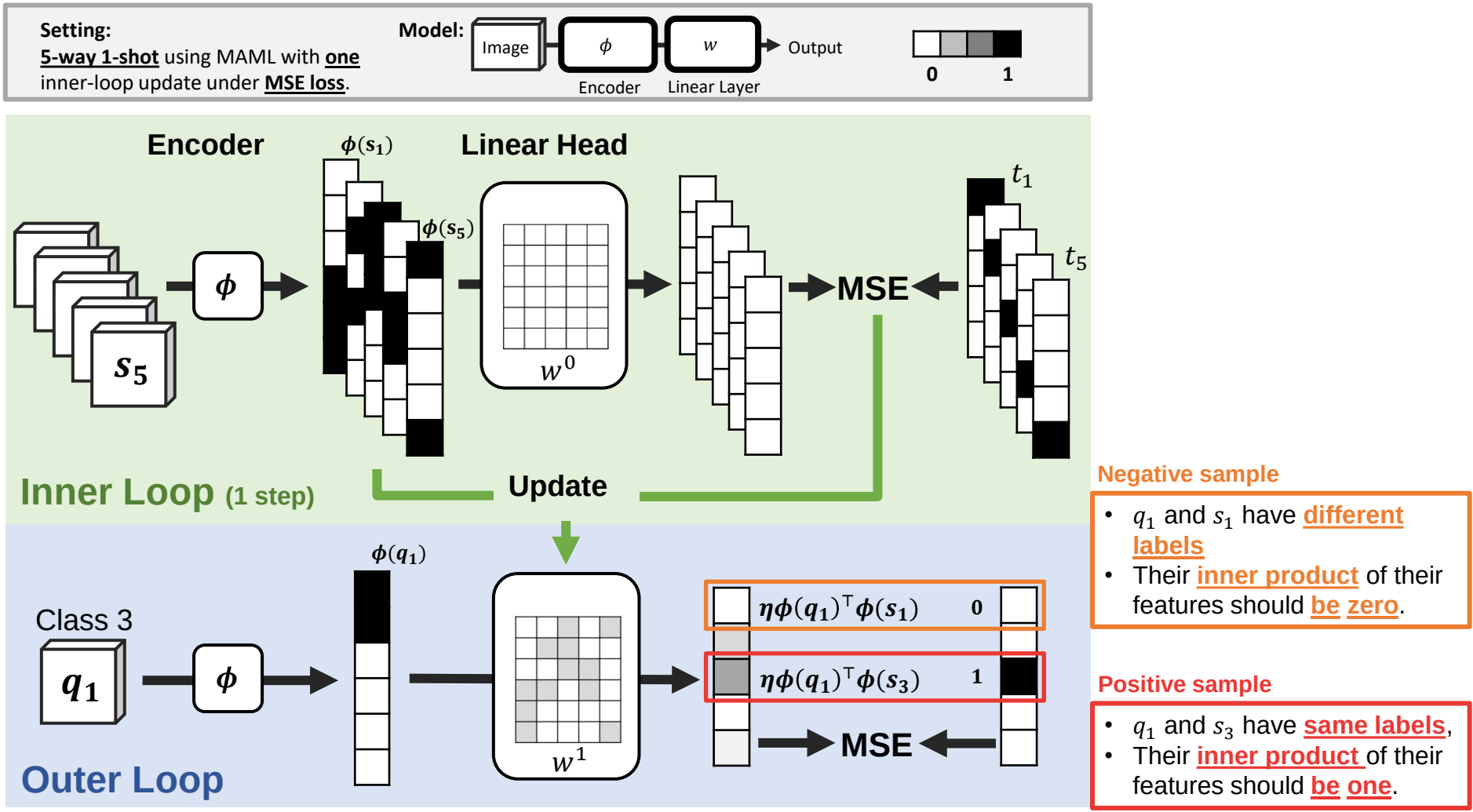
The inner product between support feature s_1 and query feature q_1.


Setting:
5-way 1-shot using MAML with one inner-loop update under MSE loss.
Model:
Image
ϕ
w
Output
Encoder
Linear Layer
0
1
Encoder
$\phi(s_1)$
Linear Head
$\phi(s_5)$
t_1
t_5
s_5
ϕ
w^0
MSE
Inner Loop (1 step)
Update
$\phi(q_1)$
Class 3
q_1
ϕ
w^1
$\eta\phi(q_1)^\top\phi(s_1)$
$\eta\phi(q_1)^\top\phi(s_3)$
0
1
MSE
Outer Loop



Negative sample

- q_1 and s_1 have different labels
- Their inner product of their features should be zero.



Negative sample

- q_1 and s_1 have different labels
- Their inner product of their features should be zero.

Positive sample

- q_1 and s_3 have same labels,
- Their inner product of their features should be one.

Results
Insights from the motivating example

Under a mild assumption, we show that **MAML (model-agnostic meta-learning) is a noisy supervised contrastive learning algorithm** in a few-shot classification paradigm.

Why is MAML effective in learning general-purpose representations?

- Because MAML implicitly exploits contrastive learning.

What is the role of support and query data in MAML?

- In first-order MAML, the features of support data act as the prototypes, guiding the update of the features of query data.

What is the role of inner loops and outer loops in MAML?

- In the inner loop, the features of support data are memorized by the linear classifier. Therefore, in the outer loop, the SoftMax output of the query data contains the inner products between the support features and the query feature.

Results

Difference between FOMAML and SOMAML

We also explain the difference between first-order MAML and second-order MAML.

$$L_{FOMAML} = \sum_{k=\{1,2,4,5\}} \eta\phi(q_1)^\top \phi(s_k) + \left(1 - \eta\phi(q_1)^\top \phi(s_3)\right)^2$$

Gradient stopping Gradient stopping

$$\frac{\partial L_{FOMAML}}{\partial \varphi} = \frac{\partial L_{FOMAML}}{\partial \phi(q_1)} \frac{\partial \phi(q_1)}{\partial \varphi}$$

$$\frac{\partial L_{FOMAML}}{\partial \varphi} = \sum_{k=\{1,2,4,5\}} \eta\phi(s_k) - 2\left(1 - \eta\phi(q_1)^\top \phi(s_3)\right)\phi(s_3)$$

→

In FOMAML, the encoder is updated s.t.

1. query feature is moving towards the same-class support features;
2. query feature is moving further to the different-class support features."

$$L_{SOMAML} = \sum_{k=\{1,2,4,5\}} \eta\phi(q_1)^\top \phi(s_k) + \left(1 - \eta\phi(q_1)^\top \phi(s_3)\right)^2$$

$$\frac{\partial L_{SOMAML}}{\partial \varphi} = \frac{\partial L_{FOMAML}}{\partial \phi(q_1)} \frac{\partial \phi(q_1)}{\partial \varphi} + \sum_{k=1}^{5} \frac{\partial L_{FOMAML}}{\partial \phi(s_k)} \frac{\partial \phi(s_k)}{\partial \varphi}$$

→

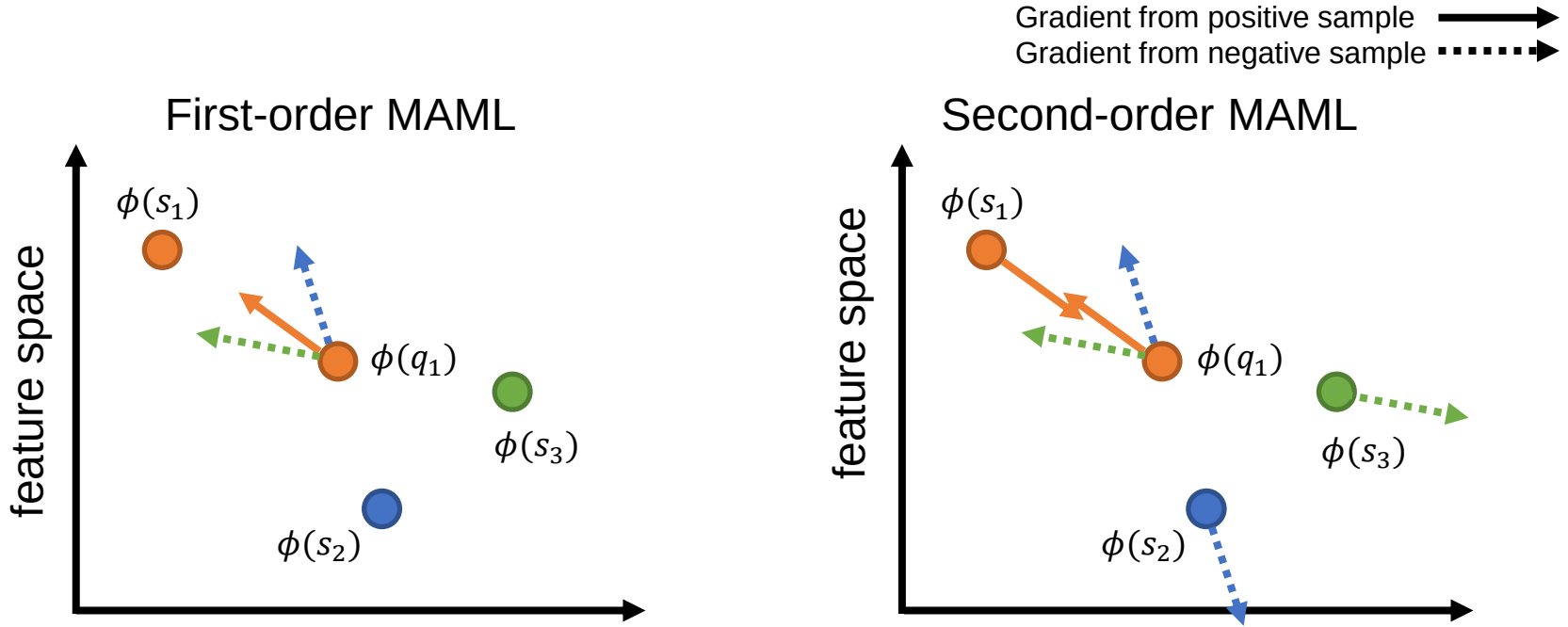
In SOMAML, the encoder is updated s.t.

1. query feature and its same-class support features are closer;
2. query feature and its different-class support features are further

Results

Difference between FOMAML and SOMAML

We illustrate the difference between first-order MAML and second-order MAML.

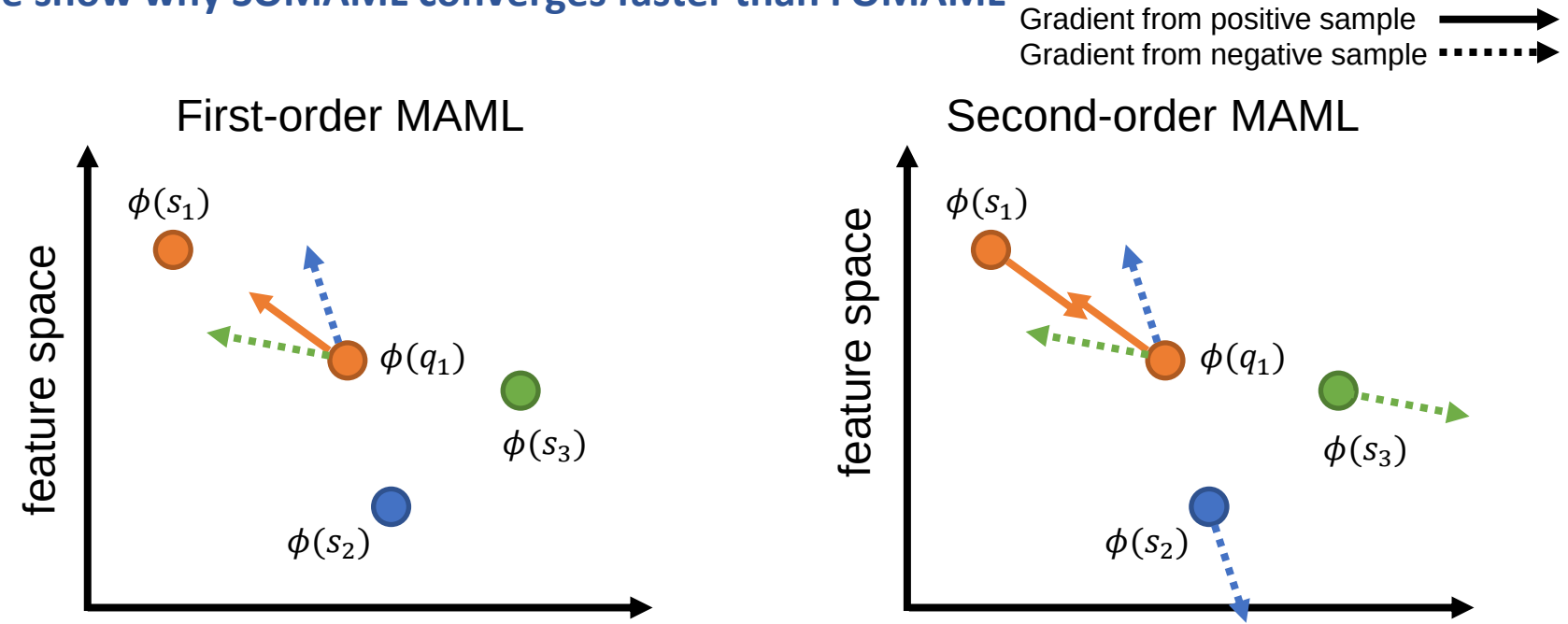


Results

Difference between FOMAML and SOMAML

We illustrate the difference between first-order MAML and second-order MAML.

We show why SOMAML converges faster than FOMAML

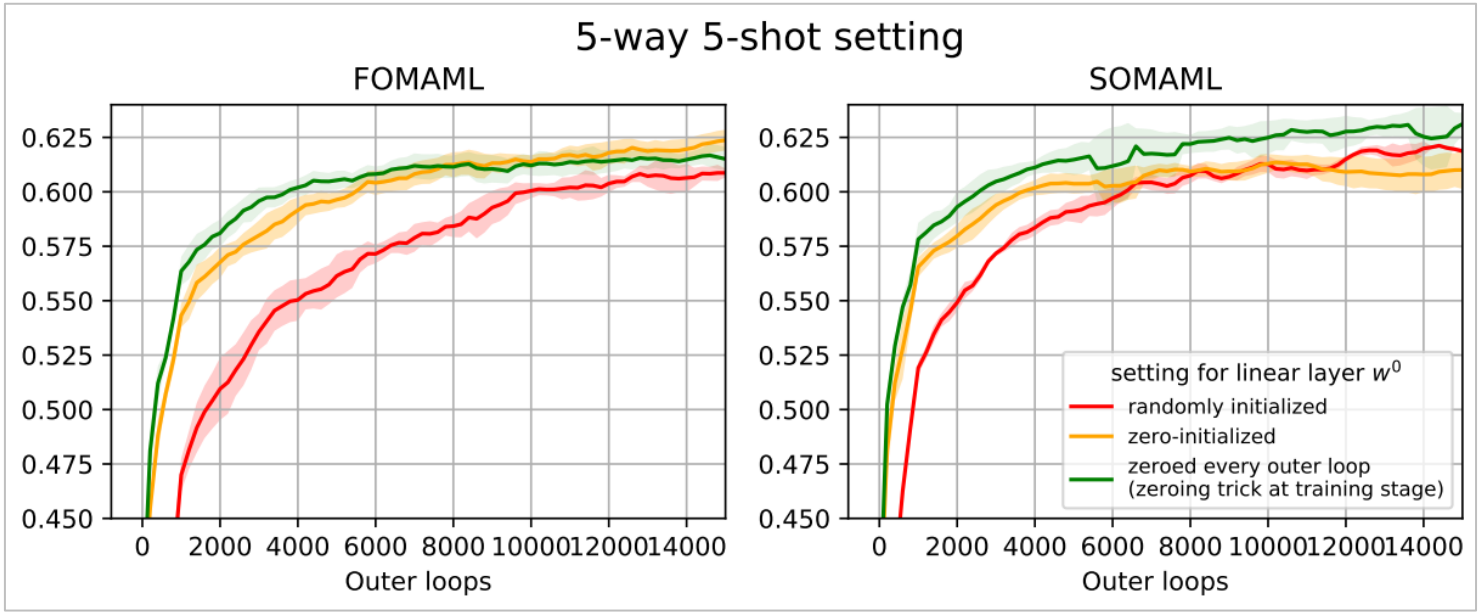


Results
Rethinking vanilla MAML

- It does not zero its linear classifier at the start of each outer loop.

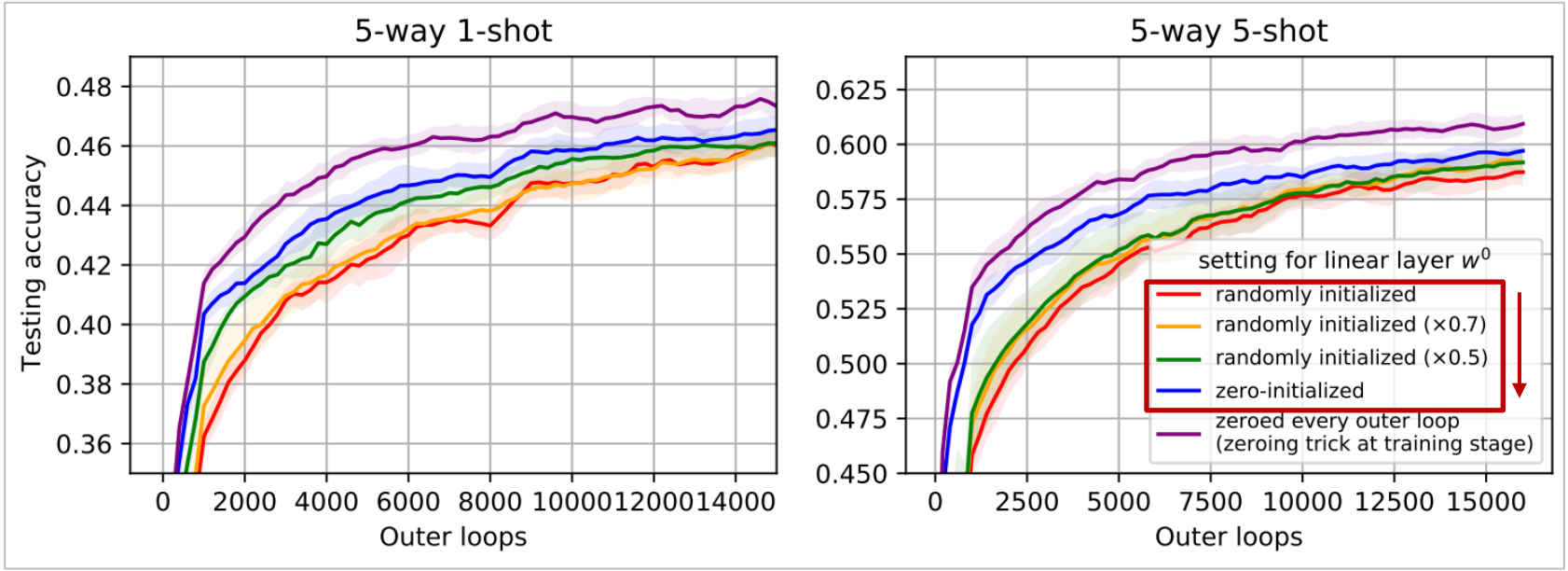
Results
Rethinking vanilla MAML

- It does not zero its linear classifier at the start of each outer loop.
  - Thus, the estimation of supervised contrastiveness in MAML is affected.



Results
Rethinking vanilla MAML

- It does not zero its linear classifier at the start of each outer loop.
  - Thus, the estimation of supervised contrastiveness in MAML is affected.
  - Down-scaling the weight of the classifier also helps mitigate the interferences.

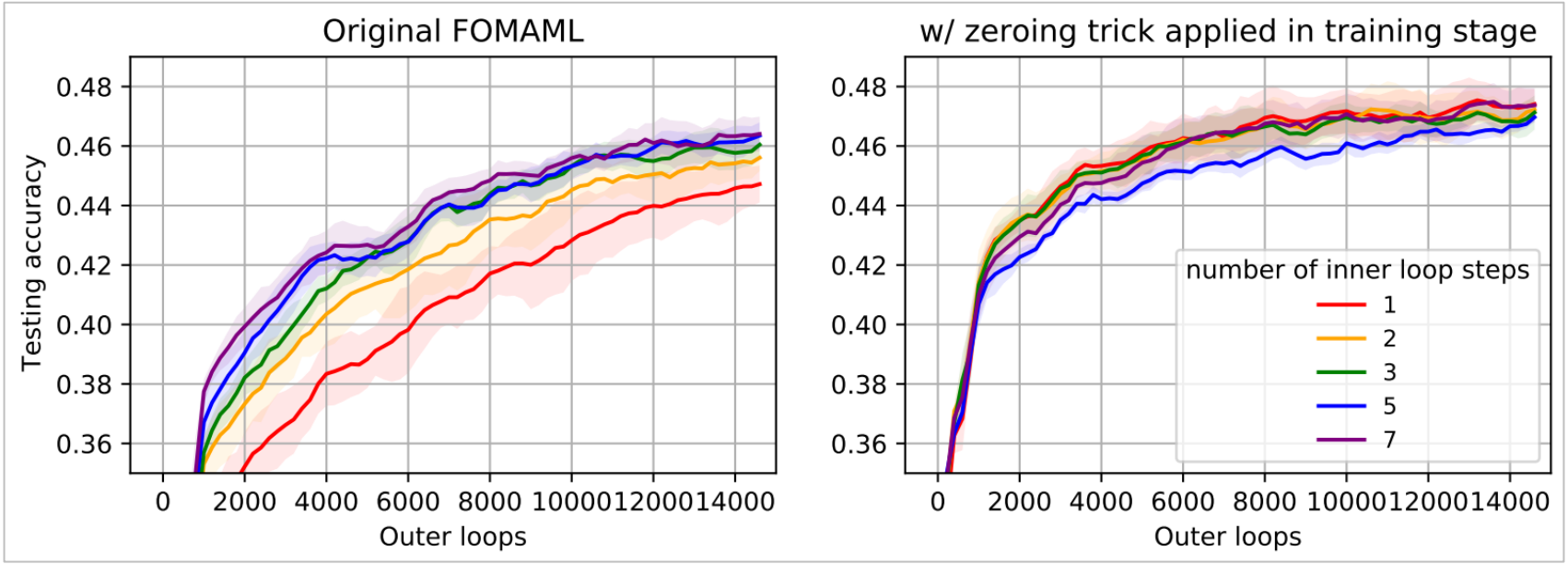


Results
Rethinking vanilla MAML

- Increasing the number of inner loop updates (N_{step}) yield better results, because largerN_{step}helps mitigate the interference.

Results
Rethinking vanilla MAML

- Increasing the number of inner loop updates (N_{step}) yield better results, because largerN_{step}helps mitigate the interference.
  - Thus, with the zeroing trick, increasing N_{step} has no effect on the performance.



Take Home Message

Under a mild assumption, we show that **MAML (model-agnostic meta-learning) is a noisy supervised contrastive learning algorithm** in a few-shot classification paradigm.

Why is MAML effective in learning general-purpose representations?

- Because MAML implicitly exploits contrastive learning.

What is the role of support and query data in MAML?

- In first-order MAML, the features of support data act as the prototypes, guiding the update of the features of query data.

What is the role of inner loops and outer loops in MAML?

- In the inner loop, the features of support data are memorized by the linear classifier. Therefore, in the outer loop, the SoftMax output of the query data contains the inner products between the support features and the query feature.