A Channel Coding Benchmark for Meta-Learning

Rui Li[1]

Work with Ondrej Bohdal[2], Hyeji Kim[3], Da Li[1,2], Nicholas D. Lane[1,4], and Timothy Hospedales [1,2]

[1] Samsung AI Center, Cambridge, UK, [2] University of Edinburgh, UK
[3] University of Texas at Austin, US, [4] University of Cambridge, UK

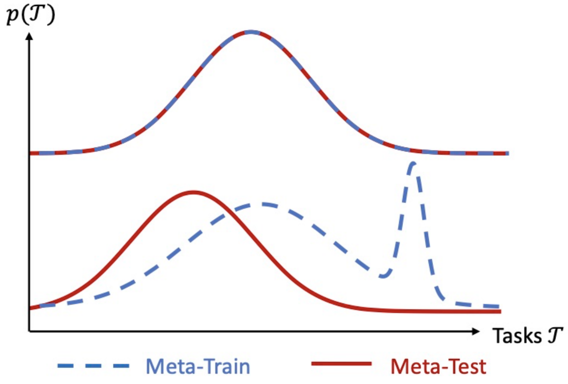
AAAI Workshop on Meta-Learning
9 February 2021

Challenges in Meta-Learning

SOTA Meta-learner often suffer in realistic settings[1][2], when:

- Task distribution is broad and multi-modal
- There is distribution shift between the meta-training and meta-testing tasks

Studying these issues with existing benchmarks lack of quantitative **measure** and ability to **control** of task complexity and distribution shifts



[1] Triantafillou et. al. “Meta-dataset: A dataset of datasets for learning to learn from few examples”. In *ICLR*, 2020.
[2] Yu et.al. “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning”. In *CORL*, 2019.

Which of the following datasets is more **complex**?

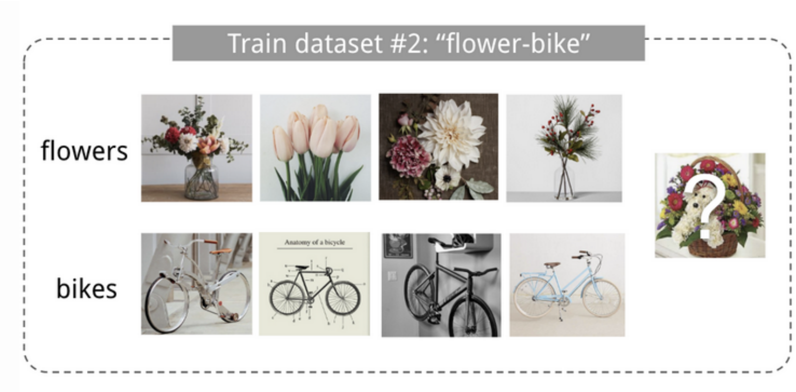
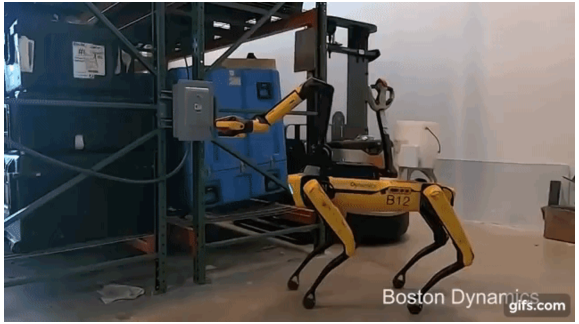


Image Source: https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html and Pinterest

Which of the following transition is associated with **a greater distribution shift**?



Meta-train task: Switch Manipulation

Meta-Train-Test Shift
#1

Meta-Train-Test Shift
#2

Meta-Test task #1: Picking-up and Putting-down an Object

Meta-Test task #2: Opening a Door

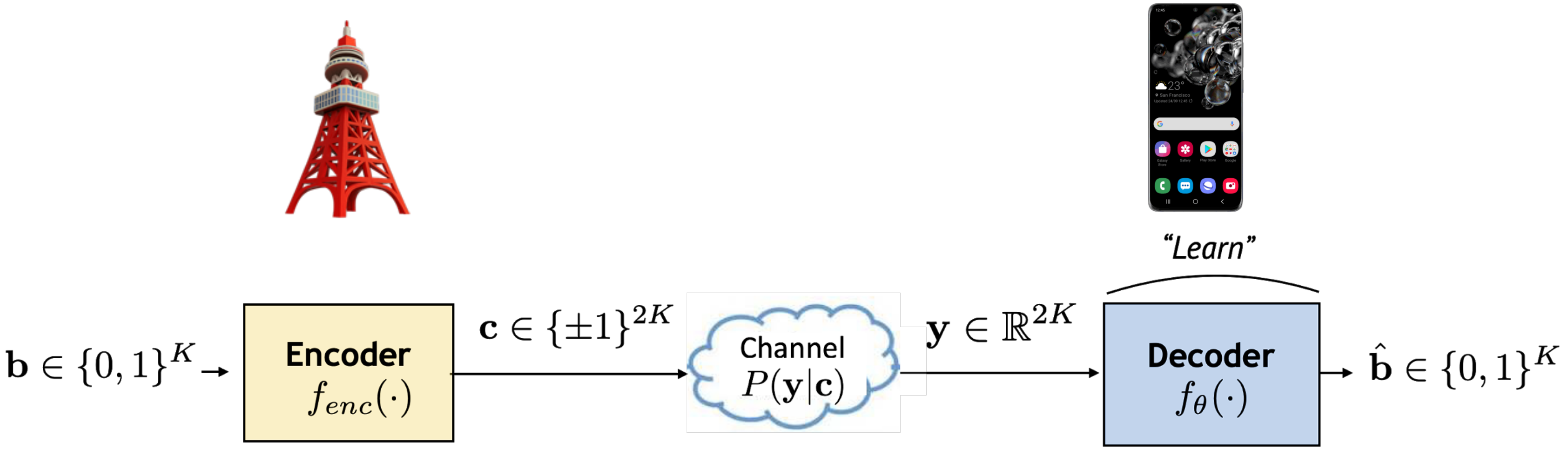
Video Source: Boston Dynamics

Our Contributions

- ➢ Propose a channel coding powered meta-learning benchmark.
- ➢ And use such benchmark to investigate:
  - Q1: *How vulnerable are existing meta-learners to under-fitting when trained on complex task distributions?*
  - Q2: *How robust are existing meta-learners to task-distribution shift between meta-train and meta-test?*
  - Q3: *Are channel coding meta-learners able to rely on the feature re-use shortcut, or must they learn to adapt?*

What is Channel Coding?

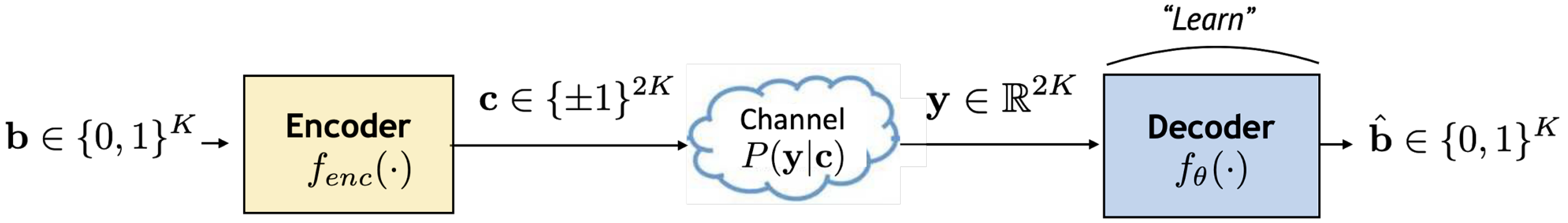
Neural Decoder [3] able to obtain superior performance on complex & realistic channels



[3] Kim, Hyeji, et al. "Communication algorithms via deep learning." NeurIPS, 2018

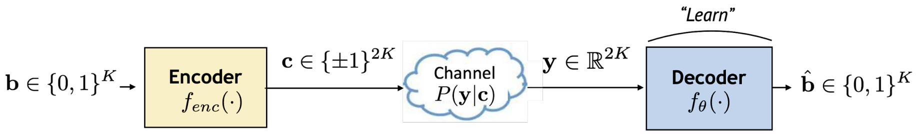
Why Channel Coding as a Benchmark

- A fundamental problem in communications
- Task distributions naturally arise, and fast adaptation to new tasks is practically valuable
- Controllability of task distributions (via controlling e.g. channel noise distributions)
- Information theoretic measures obtainable



Channel Coding Benchmark for Meta-Learning

- 4 Families (modes) of common **channel models**: Additive White Gaussian Noise (AWGN), Bursty, Memory, and Multipath interference channels, and corresponding decoding tasks.
- A **task distribution** corresponds to a channel class and is **parameterized** by continuous channel parameters ω, e.g., SNR value.
- Implementation: Based on and extended Learn2Learn [4] framework.



[4] Arnold, Sebastien M. R., et al. "learn2learn: A Library for Meta-Learning Research."

Diversity Score and Train-Test Task-Shift Measures

➢ **Definition 1: The Diversity Score D($\mathscr{T}$)** of a task distribution *p($\mathscr{T}$)* is defined as mutual information between the channel parameter *ω* and the received signal *y*:

$$D(\mathcal{T}) = \mathbb{E}_{\mathbf{c}}[I(\omega; \mathbf{y}|\mathbf{c})],$$

where *ω* denotes the channel parameter (latent variable) for the task distribution, i.e.

$$p(\mathbf{y}|\mathbf{c}) = \int_{\omega} p(\mathbf{y}|\mathbf{c}, \omega) p_{\omega}(\omega).$$

➢ **Definition 2: Train-Test Task-Shift *S(p_a(𝒯),p_b(𝒯))***

Distance between a test distribution $\mathscr{T}_a$ and a train distribution $\mathscr{T}_b$ using Kullback–Leibler divergence (KLD):

$$S(p_{\mathbf{a}}(\mathcal{T}), p_b(\mathcal{T})) := \mathbb{E}_{\mathbf{c}}[D_{KL}(p_a(\mathbf{y}_a|\mathbf{c})||p_b(\mathbf{y}_b|\mathbf{c}))] + \mathbb{E}_{\mathbf{c}}[D_{KL}(p_b(\mathbf{y}_b|\mathbf{c})||p_a(\mathbf{y}_a|\mathbf{c}))],$$

In which $p_a(y_a|c)$ and $p_b(y_b|c)$ denote the channels associated with $\mathscr{T}_a$ and $\mathscr{T}_b$, respectively.

Experiment Setup

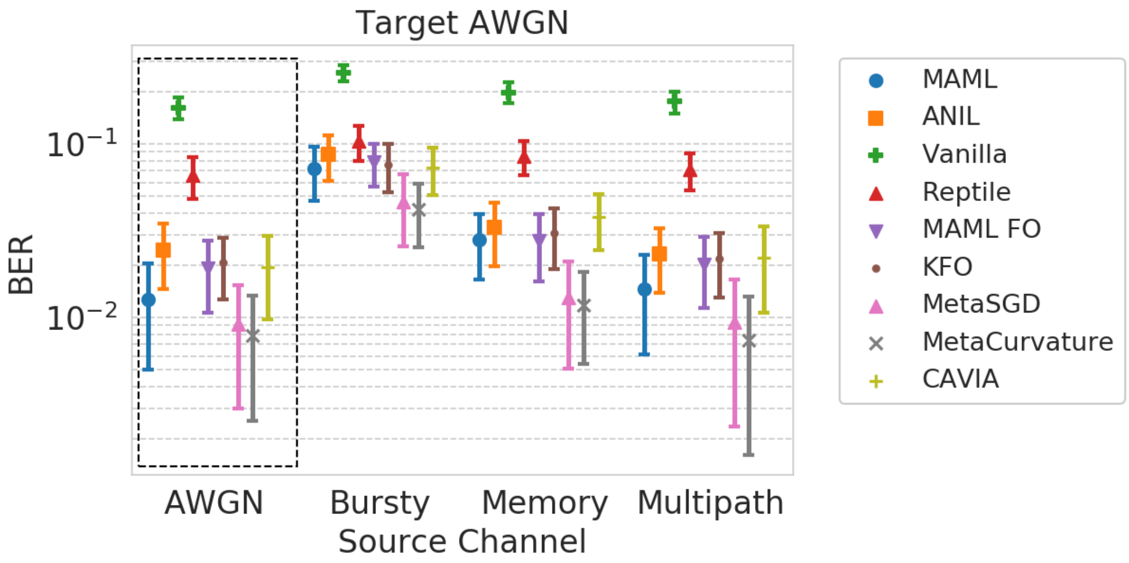
- 8 Meta-learners: MAML, MAML FO, Reptile, ANIL, KFO, CAVIA, MetaSGD, and MetaCurvature
- Non-meta-learner: empirical risk minimisation (ERM) baseline “Vanilla”
- 4 Channel families, each sample 200 noise setups

Results [Q1]: Impact of Training Distribution Diversity

Uni-modal/ within family (AWGN): Focused (SNR -0.5 ~ 0.5); Expanded (SNR -5 ~ 5)
Multi-modal/mixed: AWGN + Bursty + Memory + Multi-path
=> moderate degradation as diversity increases

BER: Bit-error-rate
(lower the better)

Results [Q2]: Impact of **Train-Test Distribution Shift**

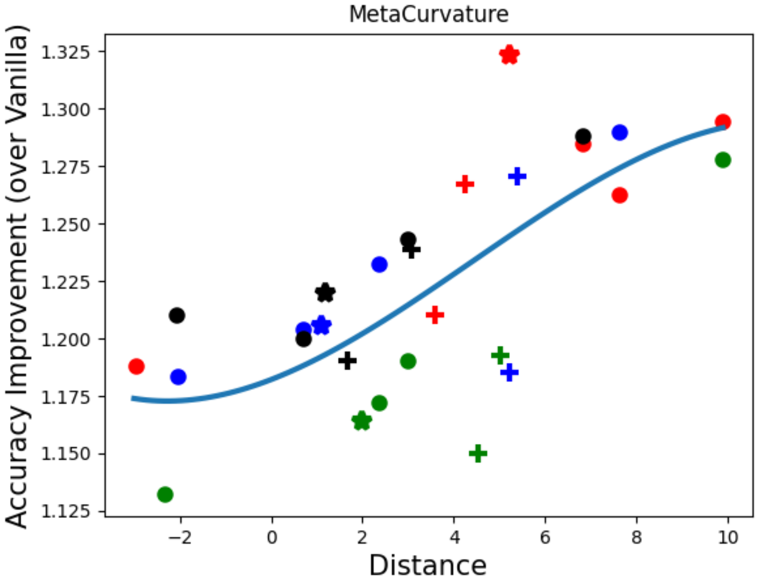


Results [Q2]: Distance Score vs Accuracy Gain over Vanilla

Each dot corresponds to an experiment

Blue curve: fitted accuracy gain

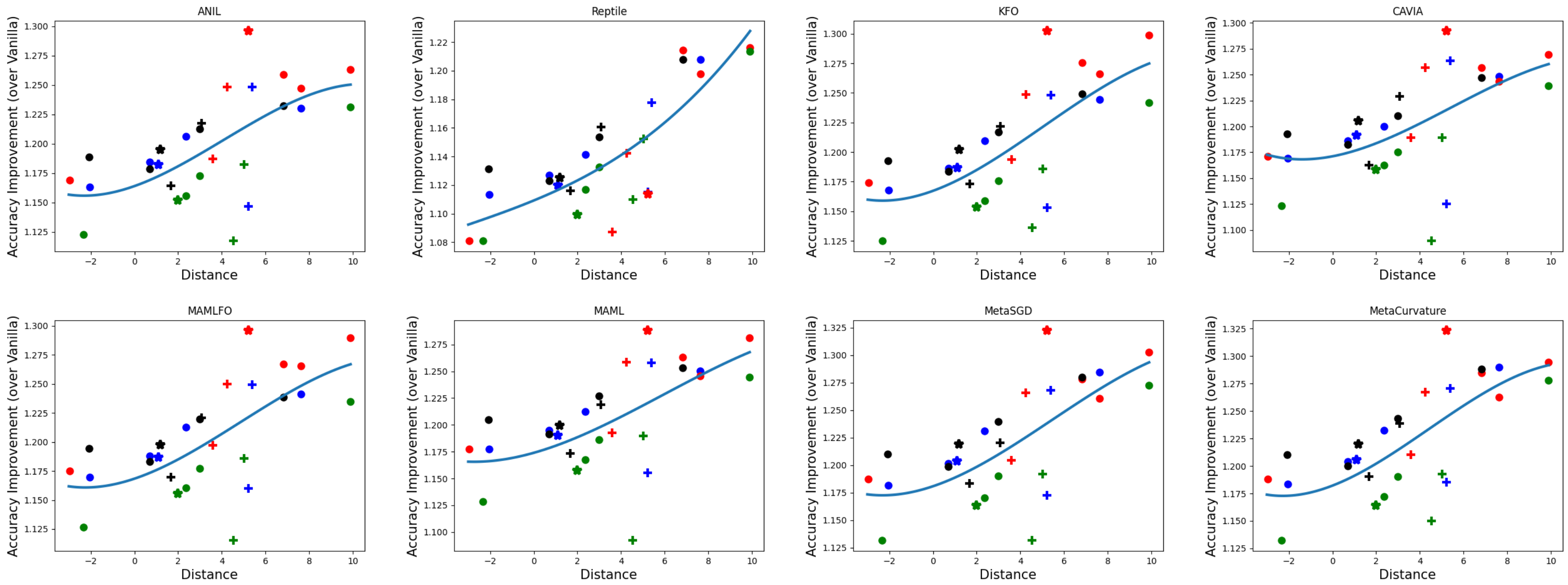
X-axis: Our distance score; Y-axis: Accuracy gain



★ shift from mixed channels
+, shift within family
o, shift across family

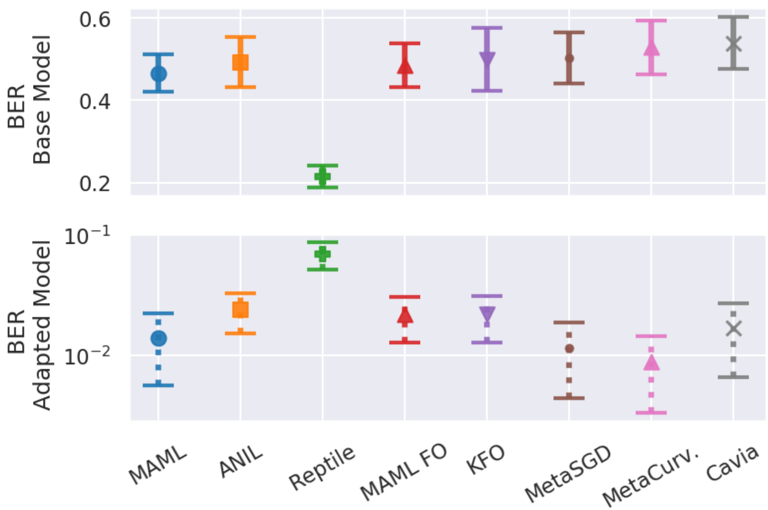
Blue: AWGN
Red: Bursty
Green: Memory
Black: Multipath

Results [Q2]: Distance Score vs Accuracy Gain over Vanilla



Follow-up Studies (if time allows)

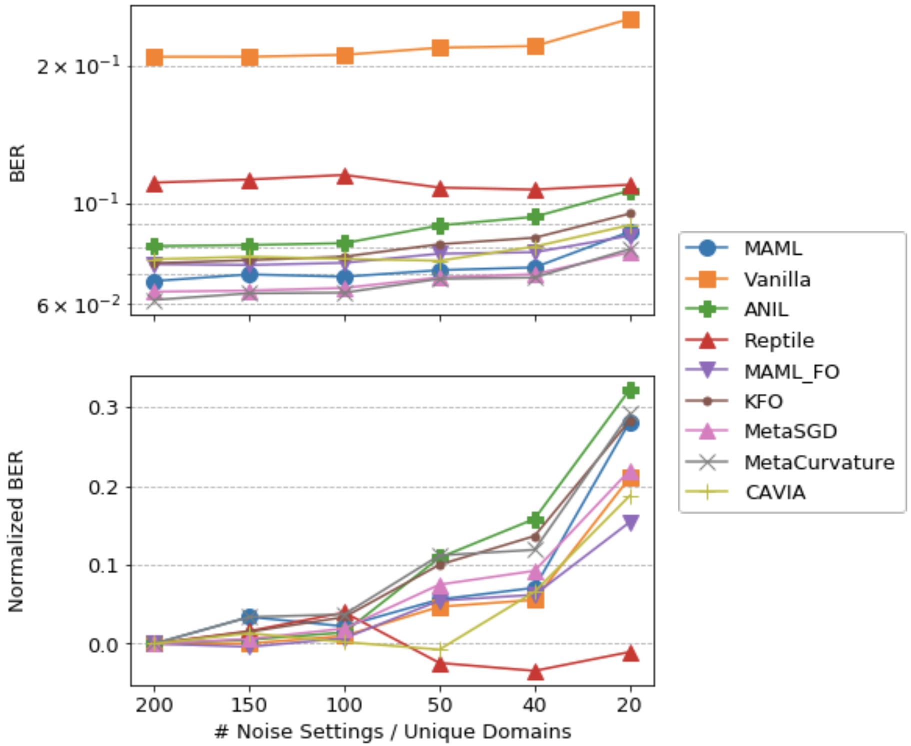
[Q3] *Who is taking the **feature re-use** short-cut?* 🔍



[5] Raghu, Aniruddh, et al. "Rapid learning or feature reuse? towards understanding the effectiveness of maml." *arXiv preprint arXiv:1909.09157* (2019).

Follow-up Studies (if time allows)

Impact of #domains available

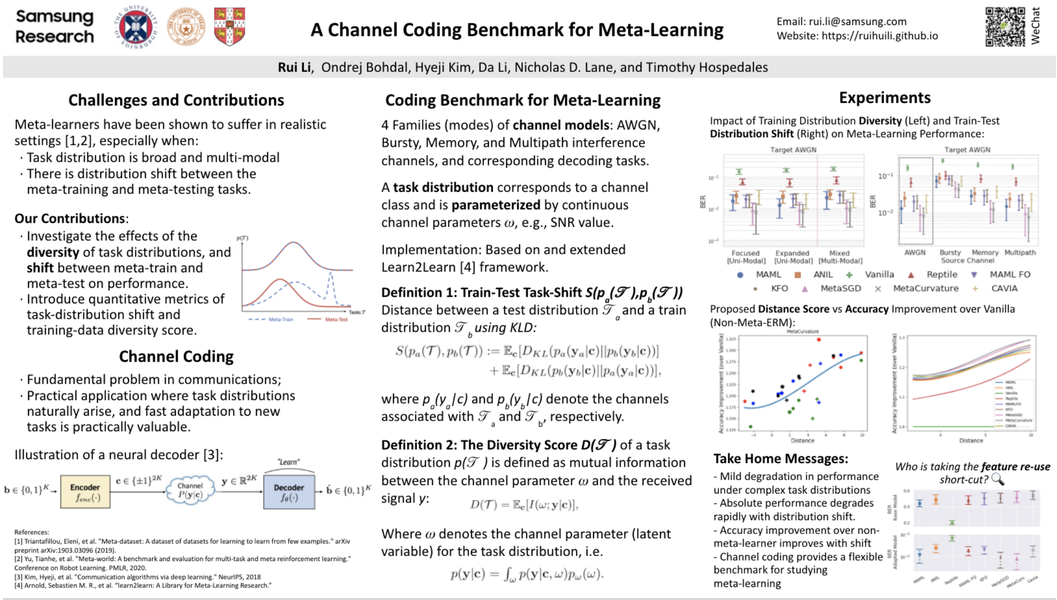


Conclusions

- ➢ Channel coding provides a flexible benchmark for studying meta-learning
- ➢ Mild degradation in performance under complex task distributions (Q1)
- ➢ Absolute performance degrades rapidly with distribution shift. (Q2)
- ➢ Accuracy improvement over non-meta-learner improves with shift (Q2)
- ➢ Less features re-use in channel coding than vision tasks (Q3)

Thank You & Questions!

Poster Session 12:00-13:00
Room 7



Contact:

Email: rui.li@samsung.com
Website: https://ruihuili.github.io