Addressing Catastrophic Forgetting in Few-Shot Problems

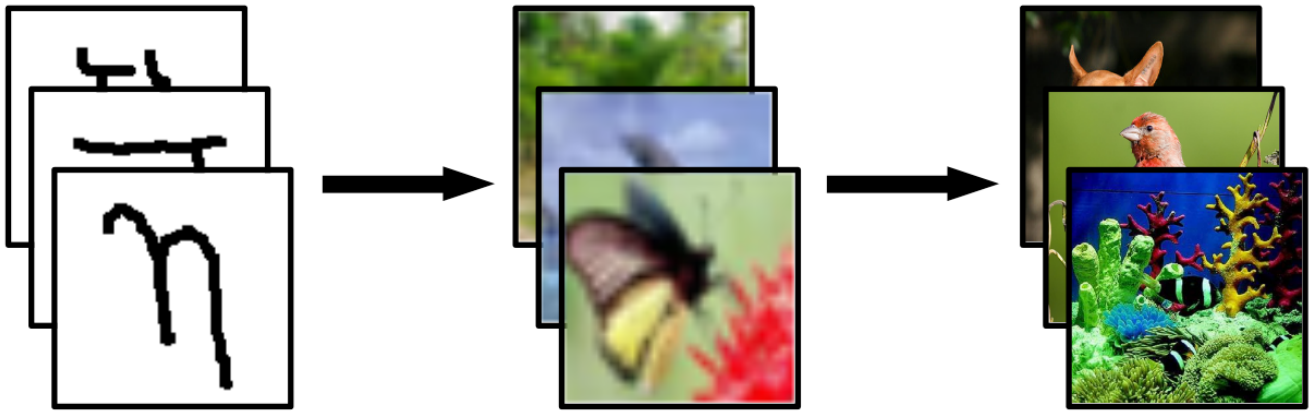
Pauching Yap Hippolyt Ritter David Barber

Centre for Artificial Intelligence
University College London

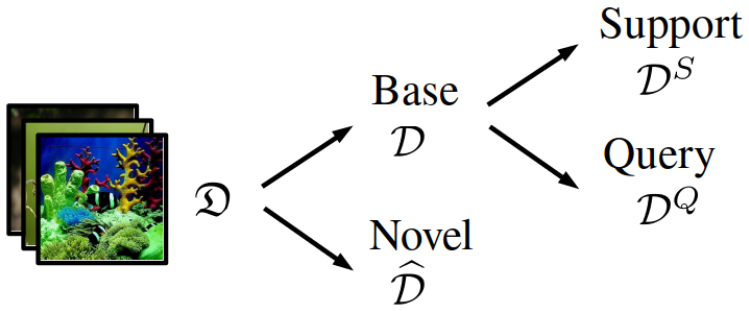
ICML 2021

Motivation

- Meta-learning is a promising solution for few-shot classification.
- Most of them assume for stationary task distribution.
- Can we meta-learn:
  - many datasets with evident distributional shift?
  - datasets arriving sequentially for meta-training?



Background



- Meta-learning algorithm (MAML) [Finn et al., 2017]:

$$\arg\min_{\theta} \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(\underbrace{SGD_k(\mathcal{L}(\theta, \mathcal{D}^{m,S}))}_{=:\tilde{\theta}^m,\ \text{inner loop task-adapted parameters}}, \mathcal{D}^{m,Q}) \tag{1}$$

- Bayesian online learning (BOL) [Opper, 1998]:

$$p(\theta|\mathfrak{D}_{1:t+1}) \propto p(\mathfrak{D}_{t+1}|\theta)p(\theta|\mathfrak{D}_{1:t}) \tag{2}$$

Update posterior using old approximate posterior
Project the updated posterior into the same parametric family

Bayesian Online Meta-Learning Framework

Base sets $\mathcal{D}_1, \ldots \mathcal{D}_t, \ldots$ arrive sequentially for meta-training.

Overview

The posterior of meta-parameters:

$$
\begin{aligned}
& p(\theta|\mathcal{D}_{1:t+1}) \\
& \quad \propto p(\mathcal{D}^S_{t+1}, \mathcal{D}^Q_{t+1}|\theta)\, p(\theta|\mathcal{D}_{1:t}) \quad \textit{BOL} && (3) \\
& \quad = p(\mathcal{D}^Q_{t+1}|\theta, \mathcal{D}^S_{t+1})\, p(\mathcal{D}^S_{t+1}|\theta)\, p(\theta|\mathcal{D}_{1:t}) && (4) \\
& \quad = \left\{ \int p(\mathcal{D}^Q_{t+1}|\tilde{\theta})\, p(\tilde{\theta}|\theta, \mathcal{D}^S_{t+1})\, d\tilde{\theta} \right\} p(\mathcal{D}^S_{t+1}|\theta)\, p(\theta|\mathcal{D}_{1:t}) \quad \textsc{Boml} && (5)
\end{aligned}
$$

Update step: using Eq. (5) with old approximate posterior $q(\theta|\phi_t)$
Projection step:

1. Laplace approximation (Gaussian q with structured precision)
2. Variational inference (Gaussian mean-field q)

BOML Implementation

$$p(\theta|\mathcal{D}_{1:t+1}) \propto \left\{ \int p(\mathcal{D}^Q_{t+1}|\tilde{\theta}) \underbrace{p(\tilde{\theta}|\theta, \mathcal{D}^S_{t+1})}_{\substack{SGD \text{ inner loop} \\ \text{as in Eq. (1)}}} d\tilde{\theta} \right\} p(\mathcal{D}^S_{t+1}|\theta)\, q(\theta|\phi_t)$$

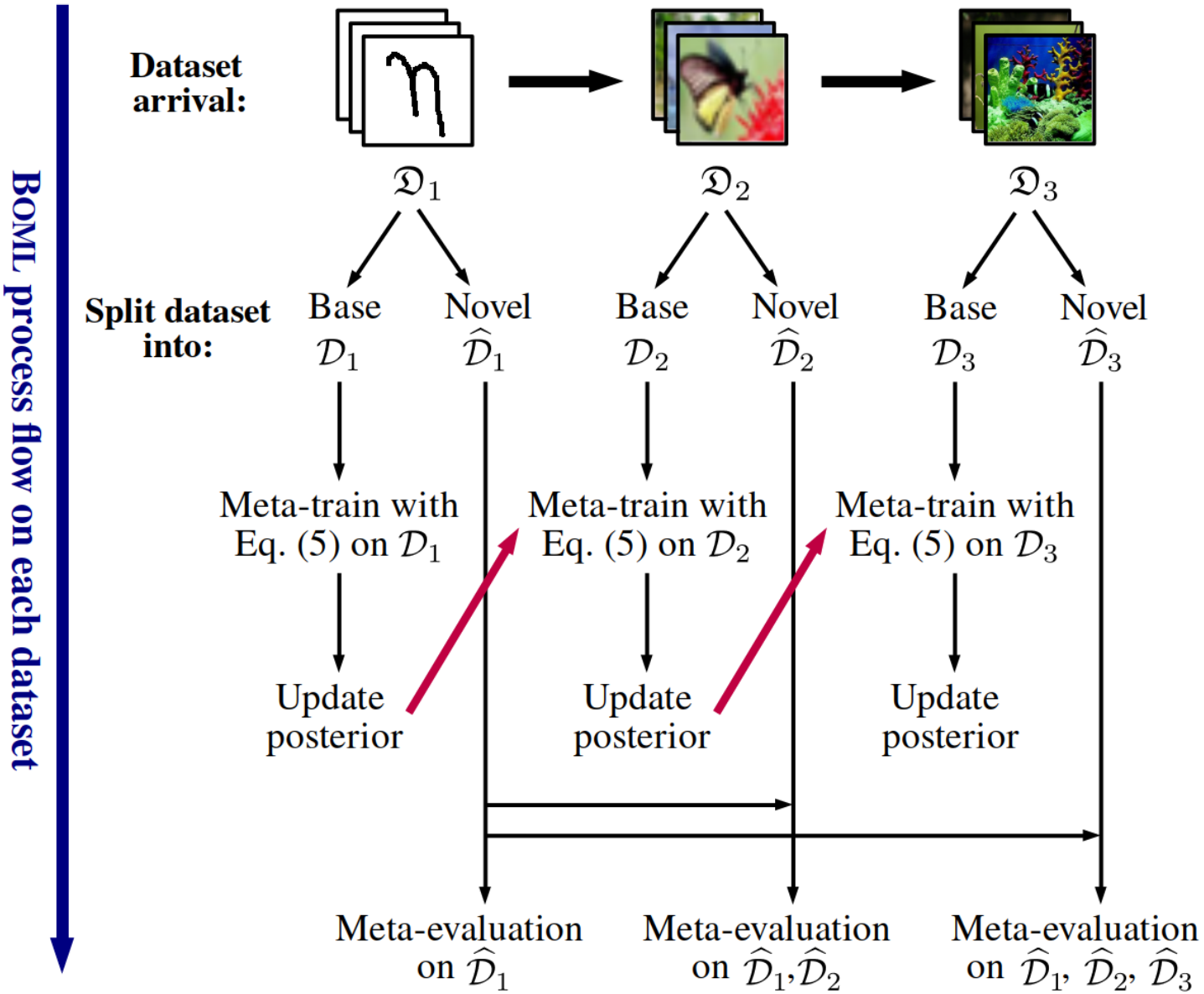
Laplace approximation [MacKay, 1992]

- Taylor expand log-posterior around a mode up to 2nd order $\rightarrow$ **Gaussian** log-probability
- Consider a MAP estimate $\arg\max_\theta \log p(\theta|\mathcal{D}_{1:t+1})$
- Corresponds to BOML projection step

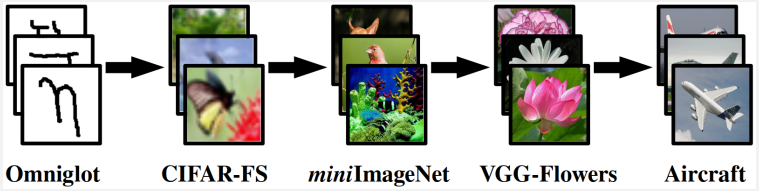
Variational continual learning [Nguyen et al., 2018]

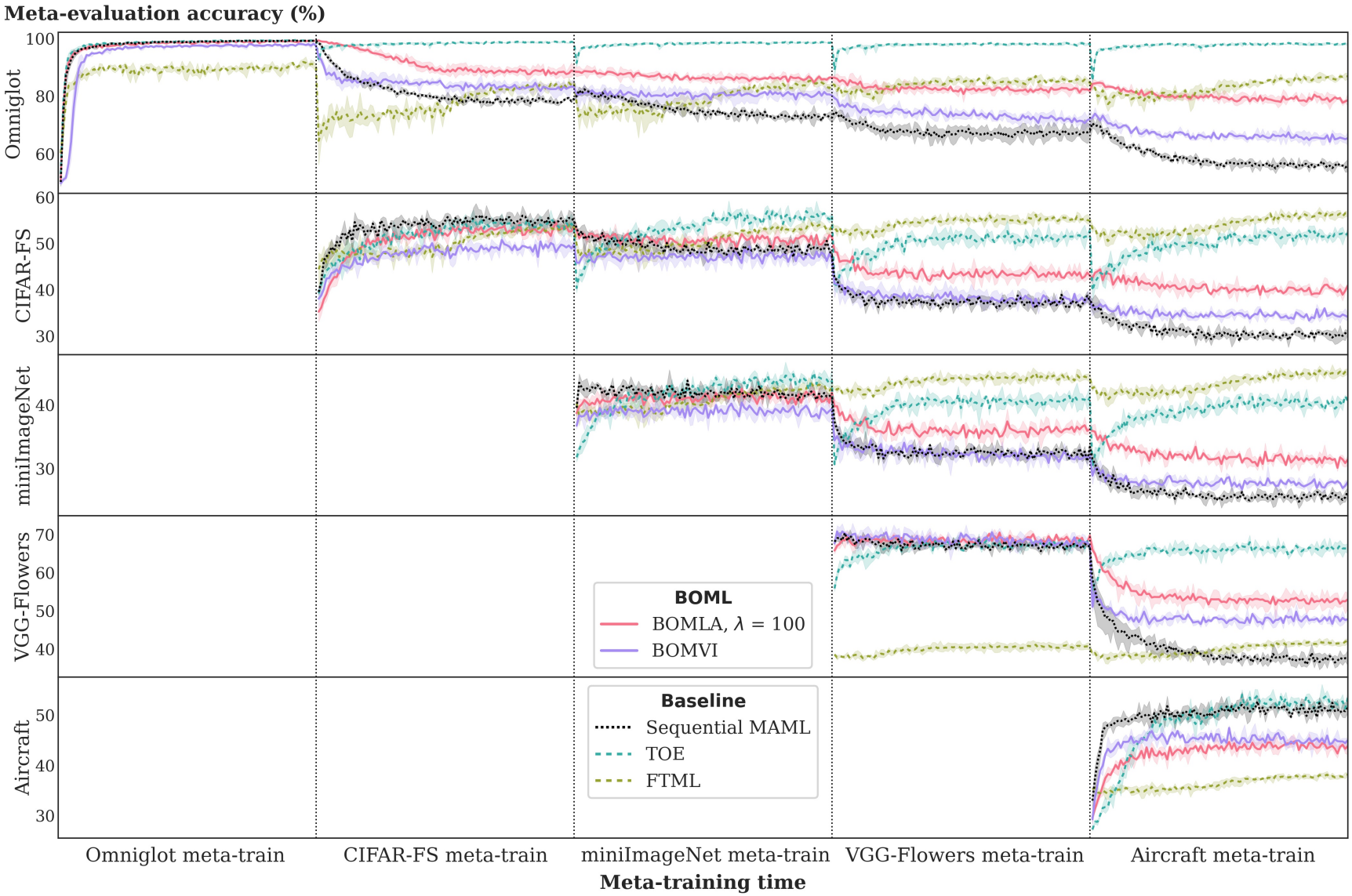
- Use Gaussian **mean-field** $q(\theta|\phi_t) = \prod_{d=1}^D N(\mu_{t,d}, \sigma^2_{t,d})$
- Minimise KL-divergence between parametric form and posterior
- Corresponds to BOML projection step

BOML Process Flow

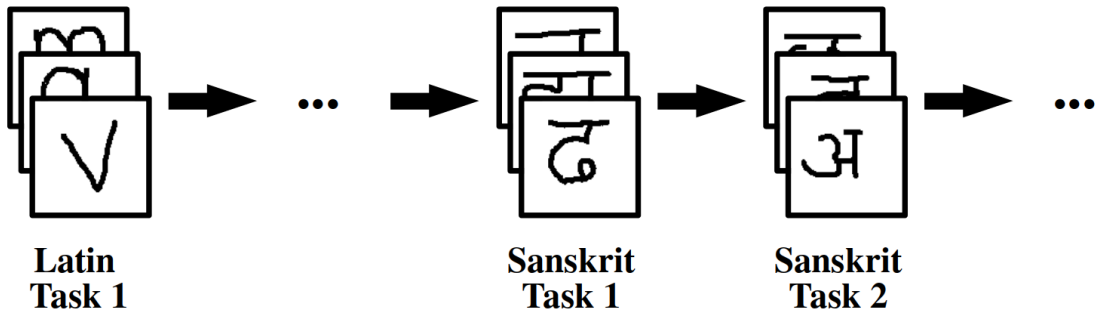


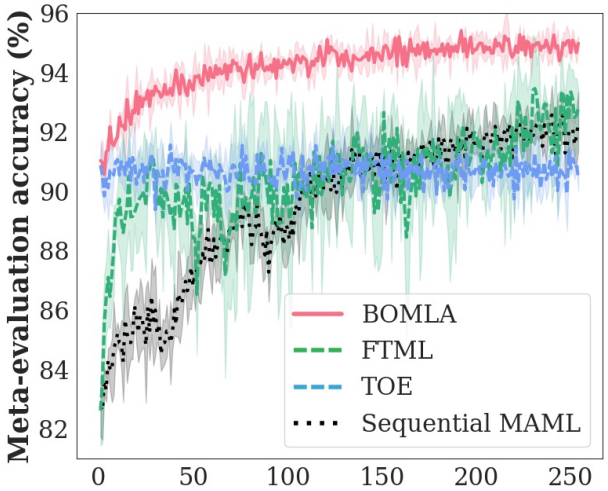
Experiment: Pentathlon

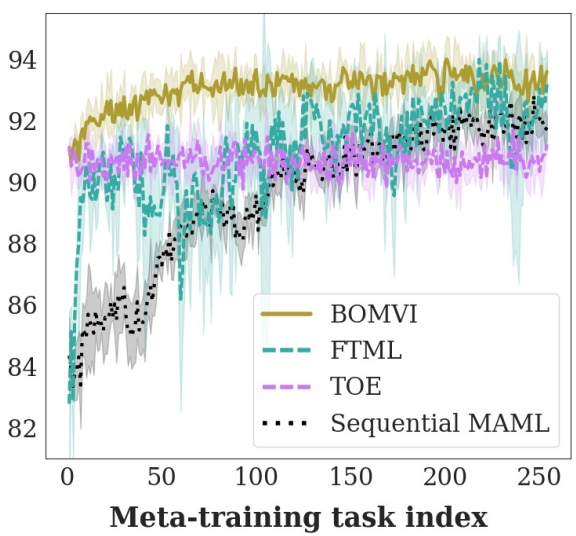


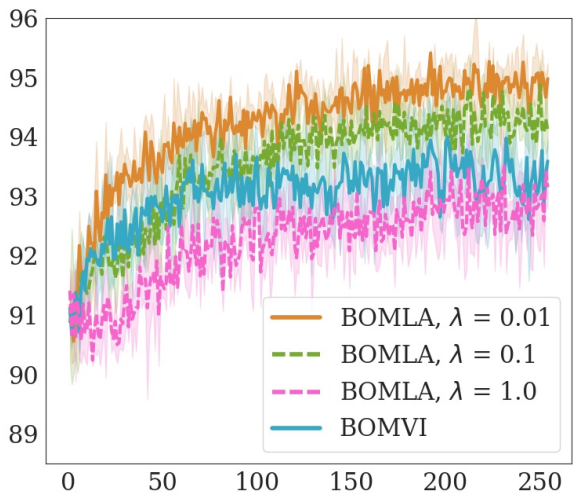


Experiment: Omniglot – Stationary Task Distribution









References

C. Finn, P. Abbeel, and S. Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

D. J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 1992.

C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational Continual Learning. In *International Conference on Learning Representations*, 2018.

M. Opper. *A Bayesian Approach to Online Learning*. Cambridge University Press, 1998.