FLITE: Focusing LITE for Memory-Efficient Meta-Learning

Sarthak Consul, Sharan Ramjee, Julia Xu
{sarthakc, sramjee, juliaxu} [at] stanford.edu

Background and Motivation

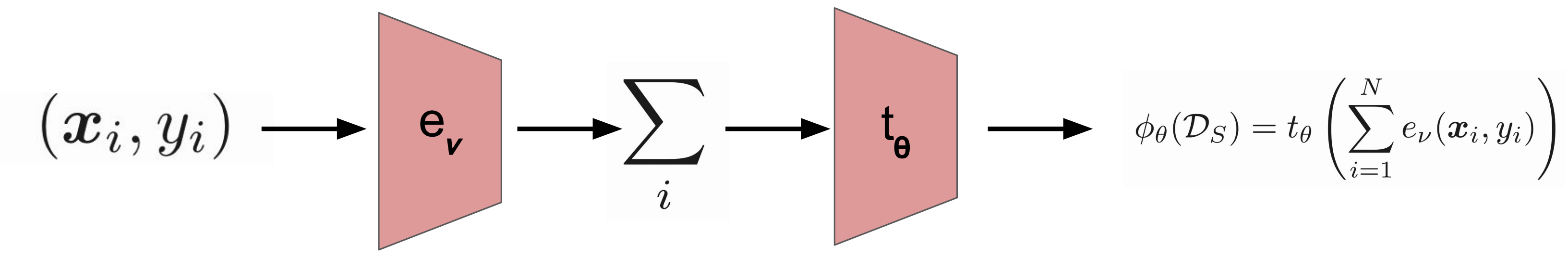
The ORBIT dataset was developed to serve as a benchmark to train object recognizer models to assist people who are visually impaired. ORBIT requires models to deal with cluttered frames where the objects are in cluttered backgrounds, and the dataset clips are collected by visually imparied users. LITE is a meta-training scheme that enables efficient meta-learning on a single GPU for tasks with large images by computing an unbiased estimate of gradients by sampling a random subset of images from the support set to backpropagate on.

However, an unbiased estimate may not always the best choice of the gradient for gradient descent. We propose using support set sampling heuristics to estimate the gradient in a modification we call FLITE. Additionally, we investigate attention and object detection approaches in meta-learning to tackle clutter in the images.

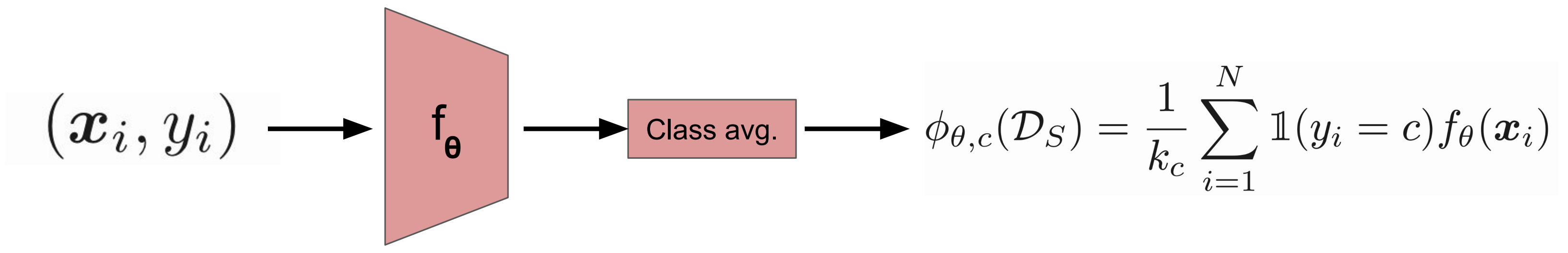
Meta-Learning

$$\underset{\theta}{\operatorname{argmin}} \sum_{\tau=1}^{T} \sum_{m=1}^{M} \mathcal{L}(y_m^*, f(\boldsymbol{x}_m^*, \phi_\theta(\mathcal{D}_S^\tau)))$$

CNAPs

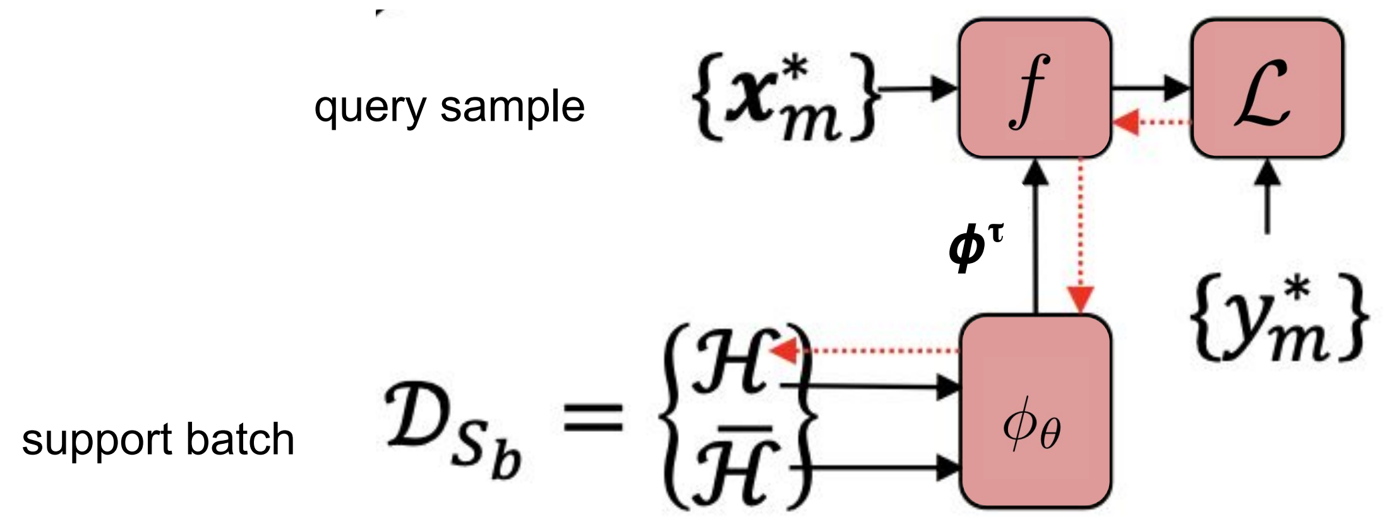

$(\boldsymbol{x}_i, y_i)$
e_ν
$\sum_i$
t_θ
$\phi_\theta(\mathcal{D}_S) = t_\theta\left(\sum_{i=1}^{N} e_\nu(\boldsymbol{x}_i, y_i)\right)$

ProtoNet


$(\boldsymbol{x}_i, y_i)$
f_θ
Class avg.
$\phi_{\theta,c}(\mathcal{D}_S) = \frac{1}{k_c}\sum_{i=1}^{N} \mathbb{1}(y_i = c) f_\theta(\boldsymbol{x}_i)$

LITE

During the inner-loop of meta-learning, LITE training proposes to process the support set by computing the forward pass on the entire support set, but estimating the gradient by only looking at the contribution of a **random subset** of the support set. This allows for a significant reduction in compute required.



Selected References

1. D. Massiceti, L. Zintgraf, J. Bronskill, L. Theodorou, M. T. Harris, E. Cutrell, C. Morrison, K. Hofmann, and S. Stumpf, "ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition," inProceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021
2. J. Bronskill, D. Massiceti, M. Patacchiola, K. Hofmann, S. Nowozin, and R. E. Turner, "Memory Efficient Meta-Learning with Large Images,"arXiv preprint arXiv:2107.01105, 2021

Methodology

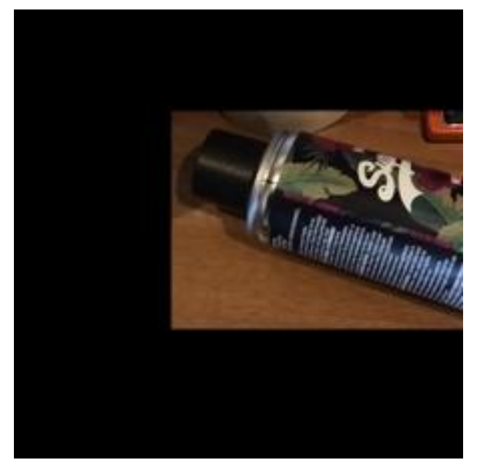
Setup

Given that our goal is to efficiently classify cluttered objects, we **subsampled** the dataset with a sampling rate of 1/10 (i.e. discarded 9/10 frames for every clip) to reduce meta-training time by 10x **without any noticeable loss in performance**. We replaced the meta-training dataset consisting of clean images with cluttered images to better match realistic scenarios where clean data in not available.

Forced Attention

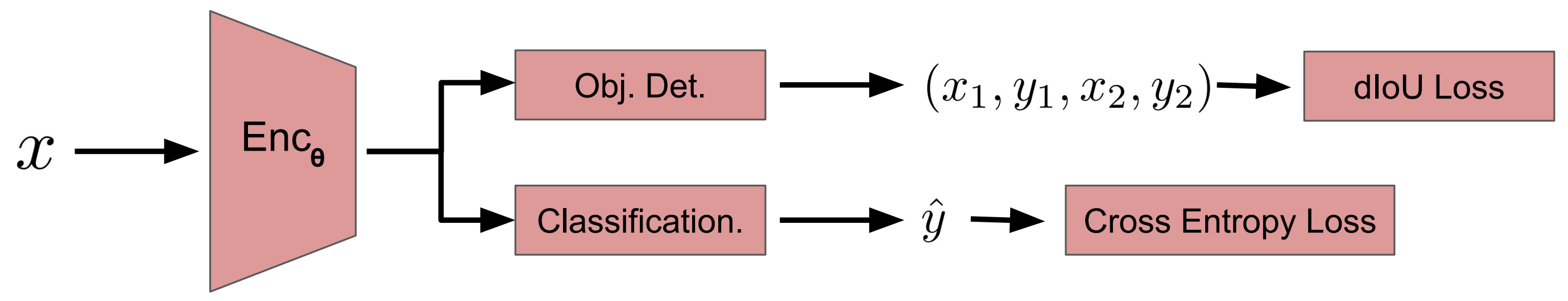
In order to enable the model to focus on the object in cluttered backgrounds, we **blacked out the part of the images** not in the bounding box in order to simulate the effect of using an attention mechanism. We use two approaches: applying the masked images during meta-training on the support set only to generate prototype features and applying the masked images during meta-training on the query set only for classification.

Deodorant
(Before forced attention)


Deodorant
(After forced attention)

Object Detection Head

We hypothesized that using forced attention led to such a drastic drop in performance as a result of using deterministic bounding boxes instead of learning to probabilistically estimate them. Thus, we added another **output head to estimate the bounding box coordinates** and dimensions during meta-training on the query set.



LITE Backpropagation Sampling Heuristics

Blur Heuristic

We applied a blurriness heuristic where the most blurry images (as determined by the variance of the laplacian) were sampled from the support set during backpropagation. This resulted in a small improvement in performance, perhaps as a result of increasing the difficulty of the meta-training tasks.

BBox Heuristic

We applied a bounding box heuristic where the images with the smallest bounding boxes were sampled from the support set during backpropagation. This resulted in noticeable improvements in performance, perhaps due to the same reason earlier stated reason.

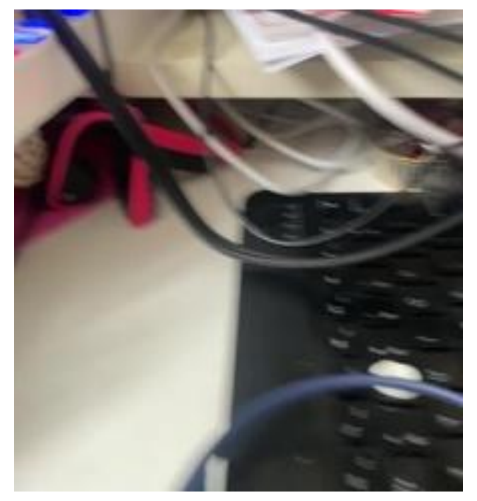
Experimental Results

The table below outlines the average frame and video accuracies over the 95% confidence intervals along with the corresponding standard deviations.

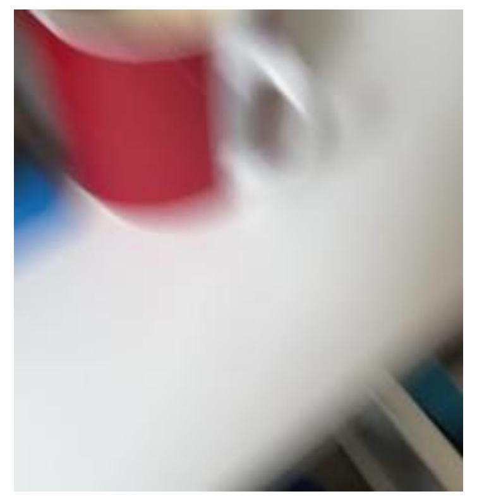
Framework	Method	Support Set	Frame Accuracy	Video Accuracy
CNAPS	Non-Subsampled*	Clean	66.3 (1.80)	72.9 (2.30)
CNAPS	None	Clean	63.92 (1.86)	70.33 (2.31)
CNAPS	Least Blur	Clean	63.93 (1.86)	70.20 (2.31)
CNAPS	Most Blur	Clean	**63.96 (1.86)**	**70.67 (2.30)**
CNAPS	None	Clutter	74.63 (2.29)	77.40 (2.59)
CNAPS	Largest BBox	Clutter	75.16 (2.26)	78.50 (2.55)
CNAPS	Smallest BBox	Clutter	**75.23 (2.26)**	**78.70 (2.54)**
ProtoNet	None	Clutter	78.14 (2.15)	83.10 (2.32)
ProtoNet	Attention (Support)	Clutter	74.20 (2.34)	78.40 (2.55)
ProtoNet	Attention (Query)	Clutter	77.83 (2.14)	82.30 (2.37)
ProtoNet	Object Detection	Clutter	78.03 (2.20)	82.50 (2.36)

*Baseline result from LITE paper

Discussion and Analysis


Headphones
(Poor positioning)

Ball
(Occlusion)


Lip Balm
(Blurry, not in frame)

Keys
(Extreme Clutter)

- The above images show examples of common classifier failure cases due to issues with poor positioning, occlusion, blurriness, and extreme clutter.
- The subsampled dataset provides the best speed-performance trade-off with 10x improvement in speed but minimal decreases in accuracy.
- The **most blurry** backpropagation subsampling heuristic led to **marginal improvements** in performance over the baseline.
- The **smallest bounding box** backpropagation sampling heuristic led to the **greatest increase** in performance over the baseline.
- Both these improvements in performance can be attributed to an increase in the difficulty of the meta-training tasks.
- Masking the support set for prototype generation failed to improve accuracies.
- Performing object detection led to slightly worse performance compared to the baseline due to an overly simplistic/under-trained object detection model.

Future Work

- Fine-tune the bounding box backpropagation subsampling heuristic for the optimal point between using the largest and smallest bounding boxes.
- Use more complex pretrained object detection networks.
- Meta-train with the subsampling heuristics on multi-step frameworks (eg. MAML), unlike the single-step framework of CNAPs and ProtoNet.