Contextual Squeeze-and-Excitation for Efficient Few-Shot Image Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent years have seen a growth in user-centric applications that require effective knowledge transfer across tasks in the low-data regime. An example is personalization, where a pretrained system is adapted by learning on small amounts of labeled data belonging to a specific user. This setting requires high accuracy under low computational complexity, therefore the Pareto frontier of accuracy vs. adaptation cost plays a crucial role. In this paper we push this Pareto frontier in the few-shot image classification setting with two key contributions: (i) a new adaptive block called Contextual Squeeze-and-Excitation (CaSE) that adjusts a pretrained neural network on a new task to significantly improve performance with a single forward pass of the user data (context), and (ii) a hybrid training protocol based on Coordinate-Descent called UpperCaSE that exploits meta-trained CaSE blocks and fine-tuning routines for efficient adaptation. UpperCaSE achieves a new state-of-the-art accuracy relative to meta-learners on the 26 datasets of VTAB+MD and on a challenging real-world personalization benchmark (ORBIT), narrowing the gap with leading fine-tuning methods with the benefit of orders of magnitude lower adaptation cost.

1 Introduction

In recent years, the growth of industrial applications based on recommendation systems (Bennett et al., 2007), speech recognition (Xiong et al., 2018), and personalization (Massiceti et al., 2021) has sparked an interest in machine learning techniques that are able to adapt a model on small amounts of data belonging to a specific user. A key factor in many of these applications is the Pareto frontier of accuracy vs. computational complexity (cost to adapt). For example, in a real-time classification task on a phone, a pretrained model must be personalized by exploiting small amounts of data on the user's device (context). In these applications the goal is twofold: maximize the classification accuracy on unseen data (target) while avoiding any latency and excessive use of computational resources.

Methods developed to face these challenges in the few-shot classification setting, can be grouped in two categories: meta-learning and fine-tuning. Meta-learning is based on the idea of learning-how-to-learn by improving the algorithm itself (Schmidhuber, 1987; Hospedales et al., 2020). Meta-learners are trained across multiple tasks to ingest a labeled context set, adapt the model, and predict the class membership of an unlabeled target point. Fine-tuning methods adjust the parameters of a pretrained neural network on the task at hand by iterative gradient-updates (Chen et al., 2019; Triantafillou et al., 2019; Tian et al., 2020; Kolesnikov et al., 2020; Dumoulin et al., 2021).

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We can gain an insight on the differences between those two paradigms by comparing them in terms of accuracy and adaptation cost. Figure 1 illustrates this comparison by showing on the vertical axis the average classification accuracy on the 18 datasets of the Visual Task Adaptation Benchmark (VTAB, Dumoulin et al. 2021), and on the horizontal axis the adaptation cost measured as the number of multiply–accumulate operations (MACs) required to adapt on a single task (see Appendix C.1 for details). Overall, fine-tuners achieve a higher classification accuracy than meta-learners but are more expensive to adapt. The comparison between two state-of-the-art methods for both categories, Big Transfer (BiT, Kolesnikov et al. 2020) and LITE (Bronskill et al., 2021), shows a substantial performance gap of 14% in favor of the fine-tuner but at a much higher adaptation cost, with BiT requiring 526×10^{12} MACs and LITE only 0.2×10^{12} MACs.

It is crucial to find solutions that retain the best of both worlds: the accuracy of fine-tuners and low adaptation cost of meta-learners. The main bottleneck that hampers the adaptation of fine-tuners is the need for multiple gradient adjustments over the entire set of network parameters. Restricting those adjustments to the last linear layer (head) significantly speeds up fine-tuning, but it harms performance (e.g. see experiments in Section 5.1). Finding a way to rapidly adapt the feature extractor (body) is therefore the main obstacle to bypass. In this paper we propose a hybrid solution to this issue, exploiting meta-learned adapters for rapidly adjusting the body and a fine-tuning routine for optimizing the head.

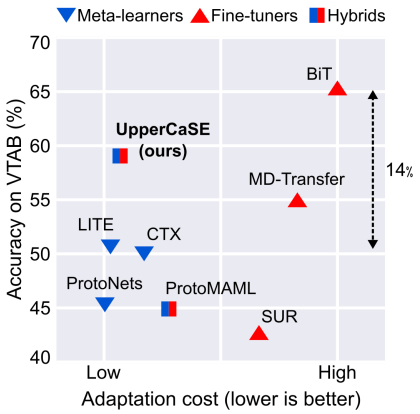

Figure 1: Accuracy and adaptation cost on VTAB for meta-learners (blue), fine-tuners (red), and hybrids (blue-red). Black dotted-line is the previous Pareto front across categories. UpperCaSE narrows the gap with the leading fine-tuning method and represents the best trade-off in terms of accuracy/adaptation-cost.

At the core of our approach is a novel extension of the popular Squeeze-and-Excitation block proposed by Hu et al. (2018) to the meta-learning setting that we call **C**ontextu**a**l **S**queeze-and-**E**xcitation (**CaSE**). We exploit CaSE as building block of a hybrid training protocol called **UpperCaSE** which is based on the idea of adjusting the body of the network in a single forward pass over the context, and reserving the use of expensive fine-tuning routines for the linear head. Figure 1 shows how UpperCaSE substantially improves the performance in the low-cost regime, outperforming meta-learners, fine-tuners such as MD-Transfer (Triantafillou et al., 2019), and reducing the gap with the current state of the art (BiT). When adaptation cost is critical, UpperCaSE is the best method currently available since it can provide substantial computation savings and compelling classification performance. Our *contributions* can be summarized as follows:

1. We introduce a new adapter called **C**ontextu**a**l **S**queeze-and-**E**xcitation (**CaSE**), based on the popular Squeeze-and-Excitation model proposed by Hu et al. (2018), that outperforms other adaptation mechanisms (e.g. the FiLM adapters used in Bronskill et al. 2021) in terms of parameter efficiency (a 75% reduction in the number of adaptation parameters) and classification accuracy (a 1.5% improvement on MetaDataset and VTAB).
2. We use CaSE adaptive blocks in **UpperCaSE**, a novel hybrid training protocol based on Coordinate-Descent (CD) that combines meta-learning and fine-tuning and enables fast adaptation of the entire model without backpropagation through the body.
3. Using UpperCaSE we report an improved classification accuracy compared to the SOTA meta-learner (Bronskill et al., 2021) on the 8 datasets of MDv2 ($+2.5\%$ on average) and the 18 datasets of VTAB ($+6.8\%$ on average), narrowing the gap with BiT (Kolesnikov et al., 2020) with the benefit of orders of magnitude lower adaptation cost.
4. We showcase the potential of UpperCaSE in a real-world personalization task on the ORBIT dataset (Massiceti et al., 2021), where it compares favorably with the leading methods in the challenging cross-domain setting (training on MDv2, testing on ORBIT).

2 Contextual Squeeze-and-Excitation (CaSE)

Problem formulation In this paragraph we introduce the few-shot learning notation, as this will be used to describe the functioning of a CaSE adaptive block. Let us define a collection of meta-training tasks as $\mathcal{D} = \{\tau_1, \ldots, \tau_D\}$ where $\tau_i = (\mathcal{C}_i, \mathcal{T}_i)$ represents a generic task composed of a context set $\mathcal{C}_i = \{(\mathbf{x}, y)_1, \ldots, (\mathbf{x}, y)_M\}$ and a target set $\mathcal{T}_i = \{(\mathbf{x}, y)_1, \ldots, (\mathbf{x}, y)_D\}$ of input-output pairs. Following common practice we use the term *shot* to identify the number of samples per class (e.g. 5-shot is 5 samples per class) and the term *way* to identify the number of classes (e.g. 10-way is 10 classes per task). Given an evaluation task $\tau_* = \{\mathcal{C}_*, \mathbf{x}_*\}$ the goal is to predict the true label y_* of the unlabeled target point $\mathbf{x}_*$ conditioned on the context set $\mathcal{C}_*$.

In fine-tuning methods, we are given a neural network $f_{\boldsymbol{\theta}}(\cdot)$, with parameters $\boldsymbol{\theta}$ estimated via standard supervised-learning on a large labeled dataset (e.g. ImageNet). Given a test task τ_* adaptation consists of minimizing the loss $\mathcal{L}(\cdot)$ via gradient updates to find the task-specific parameters $\boldsymbol{\theta}_{\tau_*} \leftarrow G(\epsilon, \mathcal{L}, \tau_*, f_{\boldsymbol{\theta}})$, where ϵ is a learning rate, and $G(\cdot)$ is a functional representing an iterative routine that returns the adapted parameters $\boldsymbol{\theta}_{\tau_*}$ (used for prediction). This procedure is particularly effective because it can exploit efficient mini-batching, parallelization, and large pretrained models.

In meta-learning methods training and evaluation are performed episodically (Vinyals et al., 2016), with training tasks sampled from a meta-train dataset and evaluation tasks sampled from an unseen meta-test dataset. The distinction in tasks is exploited to define a hierarchy. The parameters are divided in two groups: $\boldsymbol{\phi}$ task-common parameters shared across all tasks (top of the hierarchy), and $\boldsymbol{\psi}_\tau$ task-specific parameters estimated on the task at hand as part of an adaptive mechanism (bottom of the hierarchy). The way $\boldsymbol{\phi}$ and $\boldsymbol{\psi}_\tau$ come into play is method dependent; they can be estimated via gradient updates (e.g. MAML, Finn et al. 2017), learned metrics (e.g. ProtoNets, Snell et al. 2017), or Bayesian methods (Gordon et al., 2018; Patacchiola et al., 2020; Sendera et al., 2021).

Standard Squeeze-Excite (SE) We briefly introduce standard SE (Hu et al., 2018), as we are going to build on top of this work. SE is an adaptive layer used in the supervised learning setting to perform instance based channel-wise feature adaptation, which is trained following a supervised protocol together with the parameters of the neural network backbone. Given a convolutional neural network, consider a subset of L layers and associate to each one of them a Multi-Layer Perceptron (MLP), here represented as a function $g_{\boldsymbol{\phi}}(\cdot)$. The number of hidden units in the MLP is defined by the number of inputs divided by a reduction factor. Given a mini-batch of B input images, each convolution produces an output of size $B \times C \times H \times W$ where C is the number of channels, H the height, and W the width of the resulting tensor. For simplicity we split this tensor into sub-tensors that are grouped into a set $\{\mathbf{H}_1, \ldots, \mathbf{H}_B\}$ with $\mathbf{H}_i \in \mathbb{R}^{C \times H \times W}$. To avoid clutter, we suppress the layer indexing when possible. SE perform a spatial pooling that produces a tensor of shape $B \times C \times 1 \times 1$; this can be interpreted as a set of vectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_B\}$ with $\mathbf{h}_i \in \mathbb{R}^C$. For each layer l, the set is passed to the associated MLP that will generate an individual scale vector $\boldsymbol{\gamma}_i \in \mathbb{R}^C$, where

$$\boldsymbol{\gamma}_1^{(l)} = g_{\boldsymbol{\phi}}^{(l)}\left(\mathbf{h}_1^{(l)}\right) \cdots \boldsymbol{\gamma}_B^{(l)} = g_{\boldsymbol{\phi}}^{(l)}\left(\mathbf{h}_B^{(l)}\right). \tag{1}$$

An elementwise product is then performed between the scale vector and the original tensor

$$\hat{\mathbf{H}}_1^{(l)} = \mathbf{H}_1^{(l)} * \boldsymbol{\gamma}_1^{(l)} \cdots \hat{\mathbf{H}}_B^{(l)} = \mathbf{H}_B^{(l)} * \boldsymbol{\gamma}_B^{(l)}, \tag{2}$$

with the aim of modulating the activation along the channel dimension. This operation can be interpreted as a soft attention mechanism, with the MLP conditionally deciding which channel must be attended to. A graphical representation of SE is provided in Figure 2 (left).

Contextual Squeeze-Excite (CaSE) Standard SE is an instance-based mechanism that is suited for i.i.d. data in the supervised setting. In a meta-learning setting we can exploit the distinction in tasks to define a new version of SE for task-based channel-wise feature adaptation. For a task $\tau = (\mathcal{C}, \mathcal{T})$, consider the N images from the context set $\mathcal{C}$, and the tensors produced by each convolution in the layers of interest $\{\mathbf{H}_1, \ldots, \mathbf{H}_N\}$ with $\mathbf{H}_i \in \mathbb{R}^{C \times H \times W}$. As in a standard SE, we first apply a *spatial* pooling to each tensor $\mathbf{H}_i$ which produces N vectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_N\}$ of shape $\mathbf{h}_i \in \mathbb{R}^C$. Then a *context* pooling is performed; this corresponds to an empirical mean over $\{\mathbf{h}_1, \ldots, \mathbf{h}_N\}$ (see

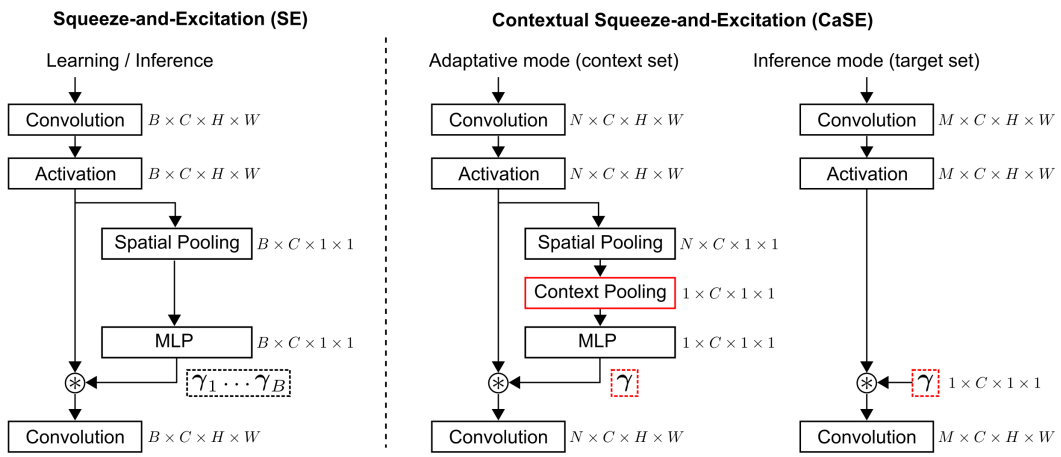

Figure 2: Comparison between the standard Squeeze-Excite (left) and the proposed Contextual Squeeze-Excite (right). Red frames highlight the two key differences between SE and CaSE: context pooling and scale transfer from context to target. B = mini-batch size, C = channels, H = height, W = width, N = context-set size, M = target-set size, $*$ elementwise multiplication.

Appendix A for more details about context pooling). The pooled representation is passed to the associated MLP to produce a single scale-vector for that layer

$$\boldsymbol{\gamma}^{(l)} = g_{\boldsymbol{\phi}}^{(l)}\left(\bar{\mathbf{h}}^{(l)}\right) \quad \text{with} \quad \bar{\mathbf{h}}^{(l)} = \frac{1}{N}\left(\mathbf{h}_1^{(l)} + \cdots + \mathbf{h}_N^{(l)}\right), \tag{3}$$

which is then multiplied elementwise by the original tensors

$$\hat{\mathbf{H}}_1^{(l)} = \mathbf{H}_1^{(l)} * \boldsymbol{\gamma}^{(l)} \;\cdots\; \hat{\mathbf{H}}_N^{(l)} = \mathbf{H}_N^{(l)} * \boldsymbol{\gamma}^{(l)}. \tag{4}$$

The scale vector is estimated in adaptive mode and transferred to the target points $\mathcal{T}$ in inference mode (no forward pass on the MLPs), as shown in the rightmost part of Figure 2. In synthesis, the three major differences between SE and CaSE are: (i) CaSE uses a contextual pooling with the aim of generating an adaptive vector per-task instead of per-instance as in SE; (ii) CaSE distinguishes between an adaptive mode and an inference mode that transfers the scale from context to target, while SE does not make such a distinction; and (iii) CaSE parameters are estimated via episodic meta-training while SE parameters via standard supervised-training. In Section 5.1 we show that those differences are fundamental to achieve superior performance in the few-shot setting. A representation of a CaSE block is reported in Figure 2 (right), additional technical details are provided in Appendix A.

3 UpperCaSE: system description and optimization protocol

We exploit CaSE blocks as part of UpperCaSE, a new hybrid training protocol based on Coordinate-Descent (CD). We call this protocol *hybrid* because it combines a meta-training procedure to optimize the CaSE parameters (body) with a fine-tuning routine to estimate the task-specific parameters (head).

Preliminaries We are given a feature extractor (body) pretrained with supervised learning on a large dataset (e.g. ImageNet), defined as $b_{\boldsymbol{\theta}}(\cdot)$ where $\boldsymbol{\theta}$ are the pretrained parameters. CaSE blocks, parameterized by $\boldsymbol{\phi}$, are added to the model at specific locations to give $b_{\boldsymbol{\theta},\boldsymbol{\phi}}(\cdot)$ (see Appendix A in supp. material for details about this step). We are interested in learning the CaSE parameters $\boldsymbol{\phi}$ keeping constant the pretrained parameters $\boldsymbol{\theta}$ (omitted from here to keep the notation uncluttered). At training time, we are given a series of tasks $\tau = \{\mathcal{C}, \mathcal{T}\} \sim \mathcal{D}$, where $\mathcal{D}$ is the training set. The number of classes (way) is calculated from the context set and used to define a linear classification head $h_{\boldsymbol{\psi}_\tau}(\cdot)$ parameterized by $\boldsymbol{\psi}_\tau$. The complete model is obtained by nesting the two functions as $h_{\boldsymbol{\psi}_\tau}(b_{\boldsymbol{\phi}}(\cdot))$. We indicate a forward pass through the body over the context inputs with the shorthand $b_{\boldsymbol{\phi}}(\mathcal{C}^x) \rightarrow \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, where $\mathbf{z}_n$ is the context embedding for the input $\mathbf{x}_n$. All the context embeddings and the associated labels are stored in $\mathcal{M} = \{(\mathbf{z}_n, y_n)\}_{n=1}^N$.

Optimization challenges We have two sets of learnable parameters, $\boldsymbol{\phi}$ the CaSE parameters, and $\boldsymbol{\psi}_\tau$ the parameters of the linear head for the task τ. While $\boldsymbol{\phi}$ is shared across all tasks (task-common), $\boldsymbol{\psi}_\tau$ must be inferred on the task at hand (task-specific). In both cases, the objective is the minimization of a classification loss $\mathcal{L}$. There are some challenges in optimizing the CaSE parameters in the body, as shown by the decomposition of the full gradient

$$\frac{d\mathcal{L}}{d\boldsymbol{\phi}} = \sum_\tau \left(\frac{\partial \mathcal{L}_\tau}{\partial \boldsymbol{\psi}_\tau} \frac{d\boldsymbol{\psi}_\tau}{d\boldsymbol{\phi}} + \frac{\partial \mathcal{L}_\tau}{\partial \boldsymbol{\phi}} \right). \tag{5}$$

The first term $\partial \mathcal{L}_\tau / \partial \boldsymbol{\psi}_\tau$ (sensitivity of the loss w.r.t. the head) and the direct gradient $\partial \mathcal{L} / \partial \boldsymbol{\phi}$ (sensitivity of the loss w.r.t. the adaptation parameters with a fixed head) can be obtained with auto-differentiation as usual. The second term $d\boldsymbol{\psi}_\tau / d\boldsymbol{\phi}$ (sensitivity of the head w.r.t. the adaptation parameters) is problematic because $\boldsymbol{\psi}_\tau$ is obtained iteratively after a sequence of gradient updates. Backpropagating the gradients to $\boldsymbol{\phi}$ includes a backpropagation through all the gradient steps performed to obtain the task-specific $\boldsymbol{\psi}_\tau$. Previous work has showed that this produces instability, vanishing gradients, and high memory consumption (Antoniou et al., 2018; Rajeswaran et al., 2019).

Meta-training via Coordinate-Descent A potential solution to these issues is the use of implicit gradients (Chen et al., 2020; Rajeswaran et al., 2019; Chen et al., 2022). The main problem with implicit gradients is the computation and inversion of the Hessian matrix as part of Cauchy's implicit function theorem, which is infeasible when the number of parameters in the linear head is large. Here we propose a simpler yet effective solution, which is based on a novel Coordinate-Descent (CD) procedure (Wright, 2015). The idea behind CD is to consider the minimization of a complex multi-variate function as a set of simpler objectives that can be solved one at a time. In our case, we can consider the joined landscape w.r.t. $\boldsymbol{\phi}$ and $\boldsymbol{\psi}_\tau$ as composed of two separate sets of coordinates (block CD, Wright 2015). By minimizing $\boldsymbol{\psi}_\tau$ first, we reach a local minima where $\partial \mathcal{L}_\tau / \partial \boldsymbol{\psi}_\tau \approx 0$. Therefore CD induces a direct optimization objective w.r.t. $\boldsymbol{\phi}$, with Equation (5) reducing to $\partial \mathcal{L}_\tau / \partial \boldsymbol{\phi}$ (no red term). See Appendix B for more details. In practice, at each training iteration we sample a task $\tau = (\mathcal{C}, \mathcal{T}) \sim \mathcal{D}$, perform a forward pass on the body (with CaSE in adaptive mode) to get

$$b_{\boldsymbol{\phi}}(\mathcal{C}^x) \rightarrow \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}. \tag{6}$$

The context embeddings are temporarily stored in a buffer with their associated labels $\mathcal{M} = \{(\mathbf{z}_n, y_n)\}_{n=1}^N$ to avoid expensive calls to $b_{\boldsymbol{\phi}}(\cdot)$. We then set the head parameters to zero, and solve the first minimization problem (inner-loop), obtaining the task-specific parameters $\boldsymbol{\psi}_\tau$ via

$$\boldsymbol{\psi}_\tau \leftarrow G\left(\alpha, \mathcal{M}, \mathcal{L}, h_{\boldsymbol{\psi}_\tau}\right) \tag{7}$$

where α is a learning rate, and $G(\cdot)$ is a functional representing an iterative gradient-descent routine for parameter estimation (e.g. maximum likelihood estimation or maximum a posteriori estimation). Note that the iterative routine in Equation (7) only relies on the head $h_{\boldsymbol{\psi}_\tau}(\cdot)$ and not on the body $b_{\boldsymbol{\phi}}(\cdot)$, which is the primary source of memory savings and the crucial difference with common fine-tuning methods. Moreover, the inner-loop is agnostic to the choice of optimizer, it can handle many gradient steps without complications, exploit parallelization and efficient mini-batching.

We then turn our attention to the second coordinate: the task-common parameters of the CaSE blocks in the body. For a single task, the update consists of a single optimization step w.r.t. $\boldsymbol{\phi}$ (outer-loop) given support/target points and the task-specific parameters $\boldsymbol{\psi}_\tau$ identified previously. The final form of the equation depends on the optimizer, for a generic SGD the update is given by

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta \nabla_{\boldsymbol{\phi}} \mathcal{L}\left(\mathcal{C}^y \cup Q^y, h_{\boldsymbol{\psi}_\tau}, b_{\boldsymbol{\phi}}\right), \tag{8}$$

where β is a learning rate. CaSE blocks must be in adaptive mode to allow the backpropagation of the gradients to the MLPs. The process repeats, alternating the minimization along the two sets of coordinates. The pseudo-code for train and test is provided in Appendix B (supp. material).

Inference on unseen tasks After the training phase, we are given an unseen task $\tau_* = (\mathcal{C}_*, \mathbf{x}_*)$ where $\mathbf{x}_*$ is a single target input and y_* the associate true label to estimate. Inference consists of three steps: (i) forward pass on the body for all the context inputs with CaSE set to adaptive mode as in Equation (6) and embeddings/labels stored in $\mathcal{M}$, (ii) estimation of the task-specific parameters $\boldsymbol{\psi}_*$ via iterative updates as in Equation (7), and (iii) inference of the target-point membership via a forward pass over body and head $\hat{y}_* = h_{\boldsymbol{\psi}_*}\left(b_{\boldsymbol{\phi}}(\mathbf{x}_*)\right)$ with CaSE in inference mode.

4 Related work

Meta-learning There has been a large volume of publications related to meta-learning. Here we focus on those methods that are the most related to our work, and refer the reader to a recent survey for additional details (Hospedales et al., 2020). LITE (Bronskill et al., 2021) is a protocol for training meta-learners on large images, that achieved SOTA accuracy on VTAB+MD. LITE is particularly relevant in this work, as its best performing method is based on Simple CNAPs (Bateni et al., 2020) that exploits FiLM for fast body adaptation. We compare against LITE in Section 5.2 showing that UpperCaSE is superior in terms of classification accuracy and parameter efficiency.

Fine-tuning Chen et al. (2019) were the first to expose the potential of simple fine-tuning baselines for transfer learning. MD-Transfer has been proposed in Triantafillou et al. (2019) as an effective fine-tuning baseline for the MetaDataset benchmark. More recently Kolesnikov et al. (2020) have presented Big Transfer (BiT), showing that large models pretrained on ILSVRC-2012 ImageNet and the full ImageNet-21k are very effective at transfer learning. MD-Transfer and BiT differ in terms of classification head, learning schedule, normalization layers, and batching. Fine-tuning only the last linear layer can be effective (Bauer et al., 2017; Tian et al., 2020). We compare against this baseline in Section 5.1, showing that adapting the body via CaSE significantly boosts the performance.

Hybrids Hybrid methods are trained episodically like meta-learners but rely on fine-tuning routines for adaptation. Model Agnostic Meta-Learning (MAML, Finn et al. 2017) finds a set of parameters that is a good starting point for adaptation towards new tasks in a few gradient steps. MAML has been the inspiration for a series of other models such as MAML++ (Antoniou et al., 2018), ProtoMAML (Triantafillou et al., 2019), and Reptile (Nichol et al., 2018).

Dynamic networks CaSE blocks belong to the wider family of dynamic networks, models that can adapt their structure or parameters to different inputs (Han et al., 2021). Adaptive components have been used in a variety of applications, such as neural compression (Veit and Belongie, 2018; Wu et al., 2018), generation of artistic styles (Dumoulin et al., 2016; Huang and Belongie, 2017), or routing (Guo et al., 2019). Residual adapters (Rebuffi et al., 2017, 2018) have been used in transfer learning (non few-shot) but they rely on fine-tuning routines which are significantly slow during adaptation.

Feature-wise Linear Modulation (FiLM) Perez et al. (2018) proposed FiLM as flexible block for relational reasoning. This modulator has been used in meta-learning for fast adaptation (Requeima et al., 2019; Zintgraf et al., 2019; Bronskill et al., 2021). SOTA meta-learners (Bronskill et al., 2021) use FiLM in conjunction with a separate convolutional set-encoder to produce a context embedding. The embedding is forwarded to local MLPs to produce scale and shift vectors that modulate a pretrained model. There are two key differences between CaSE and those FiLM adapters (i) CaSE exploits context pooling for aggregating the representations instead of a separate set-encoder (see Appendix A for details) which is more efficient, and (ii) FiLM uses scale and shift to modulate the activations, CaSE only the scale (half the parameters). We compare FiLM and CaSE in Section 5.1, showing that CaSE is superior in terms of accuracy while using a fraction of the parameters.

5 Experiments

In this section we report on experiments on VTAB+MD (Dumoulin et al., 2021) and ORBIT (Massiceti et al., 2021). VTAB+MD has become the standard evaluation protocol for few-shot approaches, and it includes a large number of datasets (8 test dataset for MD, 18 for VTAB). For a description of ORBIT, see Section 5.2. In all experiments we used the following pretrained (on ImageNet) backbones: EfficientNetB0 from the official Torchvision repository; ResNet50x1-S released with BiT (Kolesnikov et al., 2020). We used three workstations (CPU 6 cores, 110GB of RAM, and a Tesla V100 GPU), the meta-training protocol of Bronskill et al. (2021) ($10K$ training tasks, updates every 16 tasks), the Adam optimizer with a linearly-decayed learning rate in $[10^{-3}, 10^{-5}]$ for both the CaSE and linear-head. The head is updated 500 times using a random mini-batch of size 128. MD test results are averaged over 1200 tasks per-dataset (confidence intervals in appendix). We did not use data augmentation. Code to reproduce the experiments is provided in the supplementary.

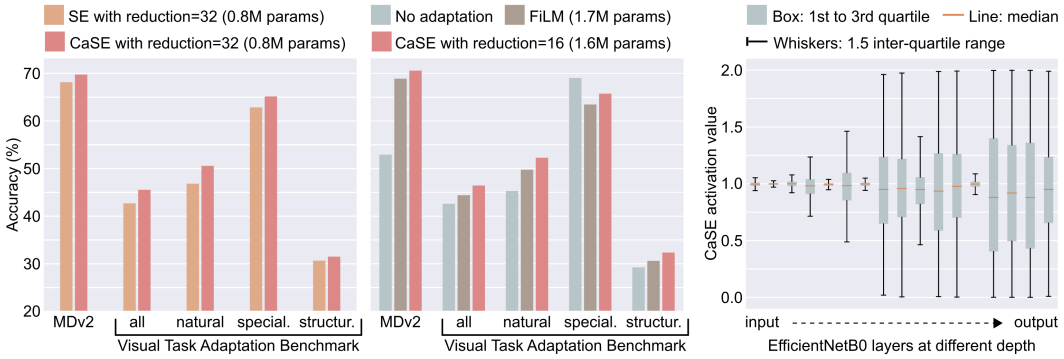

Figure 3: **Left**: CaSE vs Squeeze-and-Excitation (SE) (both methods use EfficientNetB0, 84×84 inputs, Mahalanobis-distance head). CaSE outperforms SE in all conditions. **Center**: CaSE vs the FiLM adapters of Bronskill et al. (2021) and a baseline with no body adaptation (all methods use EfficientNetB0, 84×84 inputs, Mahalanobis-distance head). CaSE outperforms FiLM in all conditions. **Right**: boxplot of CaSE activations at different depth of an EfficientNetB0 for 800 tasks sampled from the MDv2 test set (224×224 inputs, UpperCaSE with linear head). The modulation of CaSE is minimal at early stages for general-purpose filters and increases at deeper stages.

5.1 Analysis of CaSE blocks

In this sub-section we report empirical results related to CaSE blocks in three directions: **1)** we compared standard SE (Hu et al., 2018) and CaSE on MDv2 and VTAB, confirming that a) adaptation helps over not adapting, b) contextual adaptation (CaSE) outperforms instance based adaptation (SE); **2)** we compare CaSE against a SOTA adapter based on FiLM (Bronskill et al., 2021), showing that CaSE is significantly more efficient using 75% fewer parameters while boosting the classification accuracy on average by $+1.5\%$ on VTAB and MD2; and **3)** we provide an insight on the effectiveness of CaSE blocks with a series of qualitative analysis.

Comparing SE vs. CaSE We compare standard SE and the proposed CaSE on VTAB and MD-v2. For a fair comparison we keep constant all factors of variation (backbone, training schedule, hyperparameters, etc.) and use the same reduction of 32 (0.8M adaptive parameters). In order to compare the results with the other experiments in this section, we use a Mahalanobis-distance head as in Bronskill et al. (2021), reporting results with a linear head in the appendix. We summarize the results in Figure 3 (left) and add a full breakdown over all the 26 datasets in Appendix C.2 (supp. material). CaSE outperforms SE in all conditions, confirming that a contextual adaptation mechanism is fundamental to transfer knowledge effectively across tasks.

Comparing adaptation mechanisms We perform a comparison on VTAB+MD of CaSE against FiLM adapters (Bronskill et al., 2021), and a baseline that uses a pretrained model but no adaptation of the body. Methods are compared in identical conditions, using a Mahalanobis-distance head, an EfficientNetB0 backbone, and same training schedule. We show a summary of the results in Figure 3 (center) and provide a full breakdown in the supplementary material. CaSE is able to outperform FiLM in all conditions. In Appendix C.3 (supp. material) we report the results for CaSE with reduction 64 (0.4M parameters) showing that it is able to outperform FiLM adapters (1.7M parameters) using a fraction of the parameters. The comparison with the baseline with no adaptation, shows that in all but one condition (VTAB specialized) adaptation is beneficial. This is likely due to the strong domain shift introduced by some of the specialized datasets.

Role of CaSE blocks To examine the role of CaSE blocks we analyze the aggregated activations at different stages of the body for 800 tasks sampled from the MDv2 test set using an EfficientNetB0 trained with UpperCaSE on 224×224 images. In Figure 3 (right) we report the agregated distribution as boxplots, and in Appendix C.4 (supp. material) we provide a per-dataset breakdown. Overall the median is close to 1.0 (identity) which is the expected behavior as on average we aim at exploiting the underlying pretrained model. The variance is small at early stages, indicating that CaSE has

learned to take advantage of general-purpose filters that are useful across all tasks. In deeper layers the variance increases, showing a task-specific modulation effect. In Appendix C.4 (supp. material) we also include a plot with per-channel activations for all datasets at different depths, showing that the modulation is similar across datasets at early stages and it diverges later on.

5.2 Performance evaluation of UpperCaSE

In this sub-section we analyze the performance of UpperCaSE in two settings: **1)** comparison on the VTAB+MD benchmark against SOTA fine-tuners and meta-learners, where we show that UpperCaSE is able to outperform all the meta-learners, narrowing the gap with Big Transfer (BiT) on VTAB; **2)** we show an application of UpperCaSE in a real-world personalization task on the challenging ORBIT dataset (Massiceti et al., 2021) for the cross-domain condition MDv2→ORBIT, where we achieve the best average-score in most metrics, although these improvements are within the error bars.

Table 1: **UpperCaSE outperforms fine-tuners on MDv2 and narrows the gap on VTAB with the leading method (BiT) with a much lower adaptation cost**. Average accuracy on the 26 datasets of VTAB+MD. RN=ResNet, EN=EfficientNet. Img: image size. Param.: total parameters (no adapters) in millions. Cost: MACs to adapt on a task (10-shot, 100-way), in Teras. Best results in bold.

					Cost ↓	**MDv2 ↑**	**VTAB ↑**			
Method	**Protocol**	**Net**	**Img**	**Param.**	**MACs**	**all**	**all**	**natur.**	**spec.**	**struc.**
MD-Transfer	fine-tuning	RN18	126	11.2	118.6	63.4	55.6	52.4	72.9	49.3
SUR	fine-tuning	RN50	224	164.6	28.8	71.3	43.7	50.9	66.2	27.2
Big Transfer	fine-tuning	RN50	224	23.5	526.3	73.3	**65.4**	**69.4**	**81.0**	**54.5**
UpperCaSE	hybrid	RN50	224	23.5	0.8	74.9	56.6	66.3	80.1	37.6
UpperCaSE	hybrid	ENB0	224	4.0	**0.2**	**76.1**	58.4	69.1	80.3	39.4

Table 2: **UpperCaSE outperforms all meta-learning/hybrid methods and uses the lowest number of parameters per adaptive blocks**. Average accuracy on the 26 datasets of VTAB+MD. RN=ResNet, EN=EfficientNet. Img: image size. Param.: total parameters (excluding adapters). Adapt.: total adaptive parameters in millions. Best results in bold.

					Adapt. ↓	**MDv2 ↑**	**VTAB ↑**			
Method	**Protocol**	**Net**	**Img**	**Param.**	**count**	**all**	**all**	**natur.**	**spec.**	**struc.**
ProtoMAML	hybrid	RN18	126	11.2	n/a	64.2	45.0	45.7	70.7	31.5
CTX	meta-learning	RN34	224	21.3	n/a	71.6	50.5	61.1	67.3	34.0
ProtoNet	meta-learning	ENB0	224	4.0	n/a	72.7	46.1	60.9	64.2	25.9
LITE	meta-learning	ENB0	224	4.0	1.7	73.8	51.4	65.2	71.9	30.8
UpperCaSE	hybrid	RN50	224	23.5	0.8	74.9	56.6	66.3	80.1	37.6
UpperCaSE	hybrid	ENB0	224	4.0	**0.4**	**76.1**	**58.4**	**69.1**	**80.3**	**39.4**

Comparison on VTAB+MD We compare UpperCaSE against fine-tuners, meta-learners, and hybrids on the 18 datasets of VTAB and the 8 datasets of MetaDataset-v2 (MDv2) and report the results in Table 1 and Table 2. UpperCaSE outperforms all methods (including BiT) on MDv2 with an accuracy of 74.9% (ResNet50) and 76.1% (EfficientNetB0). On VTAB, UpperCaSE outperforms most methods, narrowing the gap with BiT. A closer look at the differences in performance on VTAB between UpperCaSE and BiT (see Table 1) shows that the gap is narrower on the natural and specialized splits ($+3.1\%$ and $+0.9\%$) but larger on structured ($+16.9\%$). The breakdown by dataset reported in Appendix C.6 (supp. material) shows that the major performance drops are on tasks that require localization and counting (e.g. dSprites, SmallNORB). Similar issues are encountered by methods such as LITE (Bronskill et al., 2021) which are based on FiLM adapters, suggesting that those tasks may introduce a strong domain shift w.r.t. the meta-training set that is difficult to compensate without fine-tuning the body. It is not clear whether transfer learning is beneficial on these datasets in the first place. The results in terms of adaptation cost (see Table 1) over a synthetic

Table 3: **ORBIT: UpperCaSE obtains the best average-score in most metrics, being within error bars with leading methods.** Average accuracy and 95% confidence interval for frames, videos, and frames to recognition (FTR). Cost: average MACs over all tasks (Teras). Results and setup from Massiceti et al. (2021): meta-train on MetaDataset and test on ORBIT, image-size 84×84, ResNet18 backbone, 85 test tasks (17 test users, 5 tasks per user). Best results (within error bars) in bold.

	Cost	**Clean Video Evaluation (CLE-VE)**			**Clutter Video Evaluation (CLU-VE)**		
Method	**MACs↓**	**frame acc.↑**	**FTR↓**	**video acc.↑**	**frame acc.↑**	**FTR↓**	**video acc.↑**
ProtoNet	**3.2**	**59.0±2.2**	**11.5±1.8**	**69.2±3.0**	**47.0±1.8**	**20.4±1.7**	**52.8±2.5**
CNAPs	3.5	51.9±2.5	20.8±2.3	60.8±3.2	41.6±1.9	30.7±2.1	43.0±2.5
MAML	95.3	42.5±2.7	37.3±3.0	47.0±3.2	24.3±1.8	62.3±2.3	25.7±2.2
FineTuner	317.7	**61.0±2.2**	**11.5±1.8**	**72.6±2.9**	**48.4±1.9**	**19.1±1.7**	**54.1±2.5**
UpperCaSE	3.5	**63.0±2.2**	**8.8±1.6**	**74.4±2.8**	**48.1±1.8**	**18.2±1.7**	**54.5±2.5**

task (10-shot, 100 way) show that UpperCaSE is orders of magnitude more efficient (0.2×10^{12} MACs) than all fine-tuners, with BiT being the most expensive method overall (526.3×10^{12} MACs). The comparison against meta-learners in terms of number of adaptive parameters (see Table 2) shows that UpperCaSE requires a fraction of the parameters (0.4 vs 1.7 millions for an EfficientNetB0) compared to LITE (Bronskill et al., 2021) which is based on FiLM adapters.

Comparison on ORBIT We compare UpperCaSE to other methods on ORBIT (Massiceti et al., 2021), a real-world dataset for teachable object recognizers. ORBIT consists of 3822 videos of 486 objects recorded by 77 blind/low-vision people on their mobile phones. The dataset is challenging because objects are poorly framed, occluded, blurred, and in a wide variation of backgrounds and lighting. The dataset includes two sets of target videos, one for clean video evaluation (CLE-VE) with well-centered objects, and another for clutter video evaluation (CLU-VE) with objects in complex, cluttered environments. We consider a hard transfer-learning condition where classifiers are meta-trained on MetaDataset and tested on ORBIT. Results are reported in Table 3. UpperCaSE outperforms all other methods (on average) on most metrics, being within error bars with the two leading methods. Comparing UpperCaSE with FineTuner, the gap in favor of UpperCaSE is marginal on CLU-VE but substantial on CLE-VE (frame accuracy $+2\%$, video accuracy $+1.8\%$, and FTR -2.7). Comparison in terms of adaptation cost (average MACs over all tasks) shows that UpperCaSE is orders of magnitude more efficient than FineTuner and close to the leading method (ProtoNet).

6 Conclusions

We have introduced a new adaptive block called CaSE, which is based on the popular Squeeze-and-Excitation (SE) block proposed by Hu et al. (2018). CaSE is effective at modulating a pretrained model in the few-shot setting, outperforming other adaptation mechanisms. Exploiting CaSE we have designed UpperCaSE, a hybrid method based on a Coordinate-Descent training protocol, that combines the performance of fine-tuners with the low adaptation cost of meta-learners. UpperCaSE achieves SOTA accuracy w.r.t. meta-learners on the 26 datasets of VTAB+MD and it compares favorably with leading methods in the ORBIT personalization benchmark. There are two *limitations* that are worth mentioning: (i) UpperCaSE requires iterative gradient updates that are hardware-dependent and may be slow/unavailable in some portable devices; (ii) breakdown VTAB results per-dataset shows that the method falls short on structured datasets. This indicates that fine-tuning the body may be necessary for high accuracy when the shift w.r.t. the meta-training set is large.

Societal impact Applications based on CaSE and UpperCaSE could be deployed in few-shot classification settings that can have a positive impact such as: medical diagnosis, recommendation systems, object detection, etc. The efficiency of our method can reduce energy consumption and benefit the environment. Certain applications require careful consideration to avoid biases that can harm specific groups of people (e.g. surveillance, legal decision-making).

References

Antoniou, A., Edwards, H., and Storkey, A. (2018). How to train your maml. *arXiv preprint arXiv:1810.09502*.

Bateni, P., Goyal, R., Masrani, V., Wood, F., and Sigal, L. (2020). Improved few-shot visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Bauer, M., Rojas-Carulla, M., Świątkowski, J. B., Schölkopf, B., and Turner, R. E. (2017). Discriminative k-shot learning using probabilistic models. *arXiv preprint arXiv:1706.00326*.

Bennett, J., Lanning, S., et al. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*.

Bronskill, J., Massiceti, D., Patacchiola, M., Hofmann, K., Nowozin, S., and Turner, R. (2021). Memory efficient meta-learning with large images. *Advances in Neural Information Processing Systems*.

Chen, W., Tripp, A., and Hernández-Lobato, J. M. (2022). Meta-learning feature representations for adaptive gaussian processes via implicit differentiation. *arXiv preprint arXiv:2205.02708*.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. (2019). A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*.

Chen, Y., Friesen, A. L., Behbahani, F., Doucet, A., Budden, D., Hoffman, M., and de Freitas, N. (2020). Modular meta-learning with shrinkage. *Advances in Neural Information Processing Systems*.

Dumoulin, V., Houlsby, N., Evci, U., Zhai, X., Goroshin, R., Gelly, S., and Larochelle, H. (2021). Comparing transfer and meta learning approaches on a unified few-shot classification benchmark. *arXiv preprint arXiv:2104.02638*.

Dumoulin, V., Shlens, J., and Kudlur, M. (2016). A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., and Turner, R. E. (2018). Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*.

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., and Feris, R. (2019). Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Han, Y., Huang, G., Song, S., Yang, L., Wang, H., and Wang, Y. (2021). Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2020). Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In *European conference on computer vision*.

Massiceti, D., Zintgraf, L., Bronskill, J., Theodorou, L., Harris, M. T., Cutrell, E., Morrison, C., Hofmann, K., and Stumpf, S. (2021). Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Nichol, A., Achiam, J., and Schulman, J. (2018). On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Patacchiola, M., Turner, J., Crowley, E. J., O'Boyle, M., and Storkey, A. J. (2020). Bayesian meta-learning for the few-shot setting via deep kernels. *Advances in Neural Information Processing Systems*.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. (2019). Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. (2018). Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Requeima, J., Gordon, J., Bronskill, J., Nowozin, S., and Turner, R. E. (2019). Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in Neural Information Processing Systems*.

Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München.

Sendera, M., Tabor, J., Nowak, A., Bedychaj, A., Patacchiola, M., Trzcinski, T., Spurek, P., and Zieba, M. (2021). Non-gaussian gaussian processes for few-shot regression. *Advances in Neural Information Processing Systems*.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. (2020). Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*.

Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.-A., et al. (2019). Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*.

Veit, A. and Belongie, S. (2018). Convolutional networks with adaptive inference graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151.

Wu, Z., Nagarajan, T., Kumar, A., Rennie, S., Davis, L. S., Grauman, K., and Feris, R. (2018). Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., and Stolcke, A. (2018). The microsoft 2017 conversational speech recognition system. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*.

Zintgraf, L., Shiarli, K., Kurin, V., Hofmann, K., and Whiteson, S. (2019). Fast context adaptation via meta-learning. In *International Conference on Machine Learning*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See dedicated paragraph in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Error bars have been reported for ORBIT and for the breakdonwn of the results on MetaDataset (MD) in the supplementary material. The aggregated results on MD in the paper do not report error bars as they are the average across many datasets. On VTAB error bars are not provided as the benchmark is only used for evaluation (this is in line with the original protocol).
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See setup description at beginning of Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]