RAPID LEARNING OR FEATURE REUSE? TOWARDS UNDERSTANDING THE EFFECTIVENESS OF MAML

Anonymous authors
Paper under double-blind review

ABSTRACT

An important research direction in machine learning has centered around developing meta-learning algorithms to tackle few-shot learning. An especially successful algorithm has been Model Agnostic Meta-Learning (MAML), a method that consists of two optimization loops, with the outer loop finding a *meta-initialization*, from which the inner loop can efficiently learn new tasks. Despite MAML's popularity, a fundamental open question remains – is the effectiveness of MAML due to the meta-initialization being primed for *rapid learning* (large, efficient changes in the representations) or due to *feature reuse*, with the meta initialization already containing high quality features? We investigate this question, via ablation studies and analysis of the latent representations, finding that feature reuse is the dominant factor. This leads to the ANIL (Almost No Inner Loop) algorithm, a simplification of MAML where we *remove* the inner loop for all but the (task-specific) head of the underlying neural network. ANIL matches MAML's performance on benchmark few-shot image classification and RL and offers computational improvements over MAML. We further study the precise contributions of the head and body of the network, showing that performance on the test tasks is entirely determined by the quality of the learned features, and we can remove even the head of the network (the NIL algorithm). We conclude with a discussion of the rapid learning vs feature reuse question for meta-learning algorithms more broadly.

1 INTRODUCTION

A central problem in machine learning is *few-shot learning*, where new tasks must be learned with a very limited number of labelled datapoints. A significant body of work has looked at tackling this challenge using meta-learning approaches (16; 37; 32; 6; 30; 28; 24). Broadly speaking, these approaches define a family of tasks, some of which are used for training and others solely for evaluation. A proposed meta-learning algorithm then looks at learning properties that generalize across the different training tasks, and result in fast and efficient learning of the evaluation tasks.

One highly successful meta-learning algorithm has been *Model Agnostic Meta-Learning (MAML)* (6). At a high level, the MAML algorithm is comprised of two optimization loops. The outer loop (in the spirit of meta-learning) aims to find an effective *meta-initialization*, from which the inner loop can perform efficient *adaptation* – optimize parameters to solve new tasks with very few labelled examples. This algorithm, with deep neural networks as the underlying model, has been highly influential, with significant follow on work, such as first order variants (24), probabilistic extensions (8), augmentation with generative modelling (29), and many others (15; 7; 12; 35).

Despite the popularity of MAML, and the numerous followups and extensions, there remains a fundamental open question on the basic algorithm. Does the meta-initialization learned by the outer loop result in *rapid learning* on unseen test tasks (efficient but significant changes in the representations) or is the success primarily due to *feature reuse* (with the meta-initialization already providing high quality representations)? In this paper, we explore this question and its many surprising consequences. Our main contributions are:

- We perform layer freezing experiments and latent representational analysis of MAML, finding that feature reuse is the predominant reason for efficient learning.
- Based on these results, we propose the *ANIL (Almost No Inner Loop)* algorithm, a significant simplification to MAML that removes the inner loop updates for all but the head (final layer)

of a neural network during training *and* inference. ANIL performs identically to MAML on standard benchmark few-shot classification and RL tasks and offers computational benefits over MAML.

- We study the effect of the head of the network, finding that once training is complete, the head can be removed, and the representations can be used without adaptation to perform unseen tasks, which we call the *No Inner Loop (NIL)* algorithm.
- We study different training regimes, e.g. multiclass classification, multitask learning, etc, and find that the task specificity of MAML/ANIL at training facilitate the learning of better features. We also find that multitask training, a popular baseline with no task specificity, performs worse than random features.
- We discuss rapid learning and feature reuse in the context of other meta-learning approaches.

2 Related Work

MAML (6) is a highly popular meta-learning algorithm for few-shot learning, achieving competitive performance on several benchmark few-shot learning problems (16; 37; 32; 30; 28; 24). It is part of the family of optimization-based meta-learning algorithms, with other members of this family presenting variations around how to learn the weights of the task-specific classifier. For example (19; 10; 4; 18; 39), first learn functions to embed the support set and target examples of a few-shot learning task, before using the test support set to learn task specific weights to use on the embedded target examples. (14) also proceeds similarly, using a Bayesian approach.

Of these optimization-based meta-learning algorithms, MAML has been especially influential, inspiring numerous direct extensions in recent literature (1; 8; 12; 29). Most of these extensions critically rely on the core structure of the MAML algorithm, incorporating an outer loop (for meta-training), and an inner loop (for task-specific adaptation), and there is little prior work analyzing why this central part of the MAML algorithm is practically successful. In this work, we focus on this foundational question, examining how and why MAML leads to effective few-shot learning. To do this, we utilize analytical tools such as Canonical Correlation Analysis (CCA) (26; 23) and Centered Kernel Alignment (CKA) (17) to study the neural network representations learned with the MAML algorithm, which also demonstrates MAML's ability to learn effective features for few-shot learning.

Insights from this analysis lead to a simplification that almost completely removes the inner optimization loop (the ANIL algorithm) with no reduction in performance. Other work has looked at having outer/inner loop specific parameters (40), but does this in a more complex fashion, partitioning parameters within each layer, and for specific layers, contrasting with the simple head/body separation in ANIL. Our work is complementary to methods extending MAML, and our simplification and insights could be applied to such extensions also.

3 MAML, Rapid Learning, and Feature Reuse

Our goal is to understand whether the MAML algorithm efficiently solves new tasks due to *rapid learning* or *feature reuse*. In rapid learning, large representational and parameter changes occur during adaptation to each new task as a result of favorable weight conditioning from the meta-initialization. In feature reuse, the meta-initialization already contains highly useful features that can mostly be reused as is for new tasks, so little task-specific adaptation occurs. Figure 1 shows a schematic of these two hypotheses.

We start off by overviewing the details of the MAML algorithm, and then we study the rapid learning vs feature reuse questions via layer freezing experiments and analyzing latent representations of models trained with MAML. The results strongly support feature reuse as the predominant factor behind MAML's success. In Section 4, we explore the consequences of this, providing a significant simplification of MAML – the ANIL algorithm, and in Section 6, we outline the connections to meta-learning more broadly.

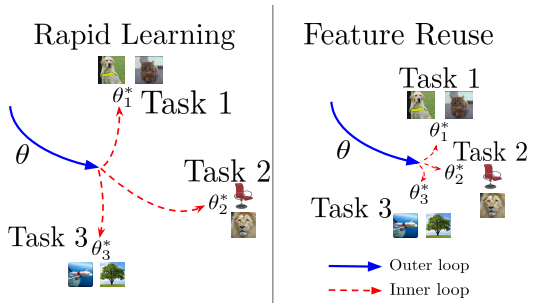

Figure 1: **Rapid learning and feature reuse paradigms.** In Rapid Learning, outer loop training leads to a parameter setting that is well-conditioned for fast learning, and inner loop updates result in significant task specialization. In Feature Reuse, the outer loop leads to parameter values corresponding to reusable features, from which the parameters do not move significantly in the inner loop. Images from (13; 9; 36; 2; 22; 34).

3.1 Overview of MAML

The MAML algorithm finds an *initialization* for a neural network so that new tasks can be learnt with very few examples (k examples from each class for k-shot learning) via two optimization loops:

- **Outer Loop:** Updates the initialization of the neural network parameters (often called the *meta-initialization*) to a setting that enables fast adaptation to new tasks.
- **Inner Loop:** Performs *adaptation*: takes the outer loop initialization, and, *separately for each task*, performs a few gradient updates over the k labelled examples (the *support set*) provided for adaptation.

More formally, we first define our base model to be neural network with meta-initialization parameters θ; let this be represented by f_θ. We have have a distribution $\mathcal{D}$ over tasks, and draw a batch $\{T_1, ..., T_B\}$ of B tasks from $\mathcal{D}$. For each task T_b, we have a *support set* of examples $\mathcal{S}_{T_b}$, which are used for inner loop updates, and a *target set* of examples $\mathcal{Z}_{T_b}$, which are used for outer loop updates. Let $\theta_i^{(b)}$ signify θ after i gradient updates for task T_b, and let $\theta_0^{(b)} = \theta$. In the inner loop, during each update, we compute

$$\theta_m^{(b)} = \theta_{m-1}^{(b)} - \alpha \nabla_{\theta_{m-1}^{(b)}} \mathcal{L}_{S_{T_b}}(f_{\theta_{m-1}^{(b)}(\theta)}) \tag{1}$$

for m fixed across all tasks, where $\mathcal{L}_{S_{T_b}}(f_{\theta_{m-1}^{(b)}(\theta)})$ is the loss on the support set of T_b after $m-1$ inner loop updates.

We then define the meta loss as

$$\mathcal{L}_{meta}(\theta) = \sum_{b=1}^{B} \mathcal{L}_{\mathcal{Z}_{T_b}}(f_{\theta_m^{(b)}(\theta)})$$

where $\mathcal{L}_{\mathcal{Z}_{T_b}}(f_{\theta_m^{(b)}(\theta)})$ is the loss on the target set of T_b after m inner loop updates, making clear the dependence of $f_{\theta_m^{(b)}}$ on θ. The outer optimization loop then updates θ as

$$\theta = \theta - \eta \nabla_\theta \mathcal{L}_{meta}(\theta)$$

At test time, we draw unseen tasks $\{T_1^{(test)}, ..., T_n^{(test)}\}$ from the task distribution, and evaluate the loss and accuracy on $\mathcal{Z}_{T_i^{(test)}}$ after inner loop adaptation using $\mathcal{S}_{T_i^{(test)}}$ (e.g. loss is $\mathcal{L}_{\mathcal{Z}_{T_i^{(test)}}}\left(f_{\theta_m^{(i)}(\theta)}\right)$).

3.2 Rapid Learning or Feature Reuse?

We now turn our attention to the key question: *Is MAML's efficacy predominantly due to rapid learning or feature reuse?* In investigating this question, there is an important distinction between the *head* (final layer) of the network and the earlier layers (the *body* of the network). In each few-shot

Freeze layers	MiniImageNet-5way-1shot	MiniImageNet-5way-5shot
None	44.5 ± 0.8	61.7 ± 0.6
1	44.8 ± 0.7	61.7 ± 0.7
1,2	44.8 ± 0.8	61.4 ± 0.7
1,2,3	44.7 ± 0.8	60.2 ± 0.7
1,2,3,4	44.7 ± 0.8	60.2 ± 0.7

Table 1: **Freezing successive layers (preventing inner loop adaptation) does not affect accuracy, supporting feature reuse.** To test the amount of feature reuse happening in the inner loop adaptation, we test the accuracy of the model when we freeze (prevent inner loop adaptation) a contiguous block of layers at test time. We find that freezing even all four convolutional layers of the network (all layers except the network head) hardly affects accuracy. This strongly supports the feature reuse hypothesis: layers don't have to change rapidly at adaptation time; they already contain good features from the meta-initialization.

learning task, there is a different alignment between the output neurons and classes. For instance, in task $\mathcal{T}_1$, the (wlog) five output neurons might correspond, in order, to the classes (dog, cat, frog, cupcake, phone), while for a different task, $\mathcal{T}_2$, they might correspond, in order, to (airplane, frog, boat, car, pumpkin). This means that the head must necessarily change for each task to learn the new alignment, and for the rapid learning vs feature reuse question, we are primarily interested in the behavior of the body of the network. We return to this in more detail in Section 5, where we present an algorithm (NIL) that does not use a head at test time.

To study rapid learning vs feature reuse in the network body, we perform two sets of experiments: (1) We evaluate few-shot learning performance when freezing parameters after MAML training, without test time inner loop adaptation; (2) We use representational similarity tools to directly analyze how much the network features and representations change through the inner loop. We use the MiniImageNet dataset, a popular standard benchmark for few-shot learning, and with the standard convolutional architecture in (6). Results are averaged over three random seeds. Full implementation details are in Appendix B.

3.2.1 Freezing Layer Representations

To study the impact of the inner loop adaptation, we freeze a contiguous subset of layers of the network, during the inner loop at test time (after using the standard MAML algorithm, incorporating both optimization loops, for training). In particular, the frozen layers are not updated at all to the test time task, and must reuse the features learned by the meta-initialization that the outer loop converges to. We compare the few-shot learning accuracy when freezing to the accuracy when allowing inner loop adaptation.

Results are shown in Table 1. We observe that even when freezing all layers in the network body, performance hardly changes. This suggests that the meta-initialization has already learned good enough features that can be reused as is, without needing to perform any rapid learning for each test time task.

3.2.2 Representational Similarity Experiments

We next study how much the latent representations (the latent functions) learned by the neural network change during the inner loop adaptation phase. Following several recent works (26; 31; 23; 21; 27; 11; 3) we measure this by applying Canonical Correlation Analysis (CCA) to the latent representations of the network. CCA provides a way to the compare representations of two (latent) layers L_1, L_2 of a neural network, outputting a similarity score between 0 (not similar at all) and 1 (identical). For full details, see (26; 23). In our analysis, we take L_1 to be a layer before the inner loop adaptation steps, and L_2 after the inner loop adaptation steps. We compute CCA similarity between L_1, L_2, averaging the similarity score across different random seeds of the model and different test time tasks. Full details are in Appendix B.2

The result is shown in Figure 2, left pane. Representations in the body of the network (the convolutional layers) are highly similar, with CCA similarity scores of > 0.9, indicating that the inner loop induces little to no functional change. By contrast, the head of the network, which does change

significantly in the inner loop, has a CCA similarity of less than 0.5. To further validate this, we also compute CKA (Centered Kernel Alignment) (17) (Figure 2 right), another similarity metric for neural network representations, which illustrates the same pattern. These representational analysis results strongly support the feature reuse hypothesis, with further results in Appendix Section B.4, B.3 providing yet more evidence.

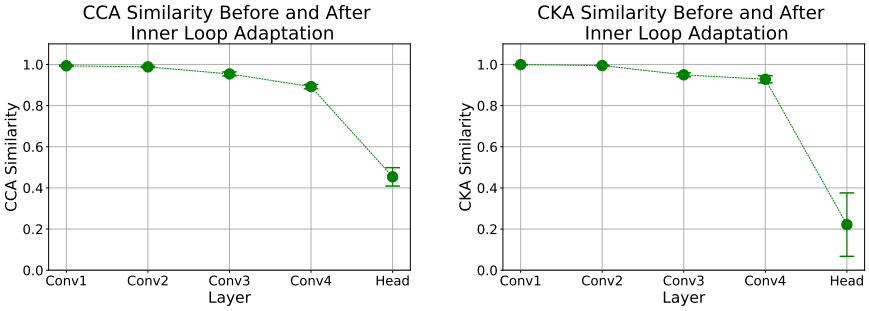

Figure 2: **High CCA/CKA similarity between representations before and after adaptation for all layers except the head.** We compute CCA/CKA similarity between the representation of a layer before the inner loop adaptation and after adaptation. We observe that for all layers except the head, the CCA/CKA similarity is almost 1, indicating perfect similarity. This suggests that these layers do not change much during adaptation, but mostly perform feature reuse. Note that there is a slight dip in similarity in the higher conv layers (e.g. conv3, conv4); this is likely because the slight representational differences in conv1, conv2 have a compounding effect on the representations of conv3, conv4. The head of the network must change significantly during adaptation, and this is reflected in the much lower CCA/CKA similarity.

3.2.3 Feature Reuse Happens Early in Learning

Having observed that the inner loop does not significantly affect the learned representations with a fully trained model, we extend our analysis to see whether the inner loop affects representations and features earlier on in training. We take MAML models at 10000, 20000, and 30000 iterations into training, perform freezing experiments (as in Section 3.2.1) and representational similarity experiments (as in Section 3.2.2).

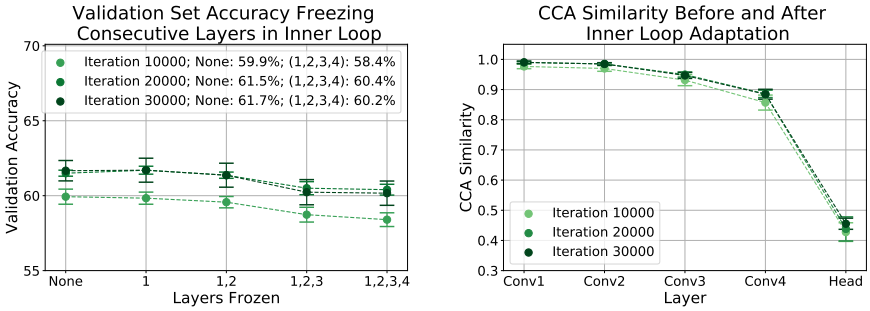

Figure 3: **Inner loop updates have little effect on learned representations from early on in learning.** Left pane: we freeze contiguous blocks of layers (no adaptation at test time), on MiniImageNet-5way-5shot and see almost identical performance. Right pane: representations of all layers except the head are highly similar pre/post adaptation – i.e. features are being reused. This is true from very early (iteration 10000) in training.

Results in Figure 3 show the same patterns from early in training, with CCA similarity between activations pre and post inner loop update on MiniImageNet-5way-5shot being very high for the body (just like Figure 2), and similar to Table 1, test accuracy remaining approximately the same when freezing contiguous subsets of layers, even when freezing all layers of the network body. This shows that even early on in training, significant feature reuse is taking place, with the inner loop having minimal effect on learned representations and features. Results for 1shot MiniImageNet are in Appendix B.5, and show very similar trends.

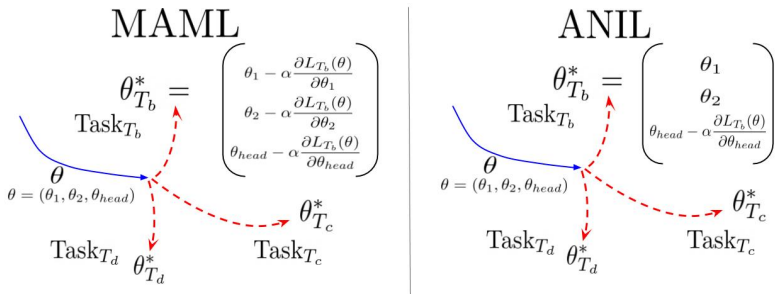

Figure 4: **Schematic of MAML and ANIL algorithms.** The difference between the MAML and ANIL algorithms: in MAML (left), the inner loop (task-specific) gradient updates are applied to all parameters θ, which are initialized with the meta-initialization from the outer loop. In ANIL (right), only the parameters corresponding to the network head θ_{head} are updated by the inner loop, during training **and** testing.

Method	Omniglot-20way-1shot	Omniglot-20way-5shot	MiniImageNet-5way-1shot	MiniImageNet-5way-5shot
MAML	93.7 ± 0.7	96.4 ± 0.1	46.9 ± 0.2	63.1 ± 0.4
ANIL	96.2 ± 0.5	98.0 ± 0.3	46.7 ± 0.4	61.5 ± 0.5

Method	HalfCheetah-Direction	HalfCheetah-Velocity	2D-Navigation
MAML	170.4 ± 21.0	-139.0 ± 18.9	-20.3 ± 3.2
ANIL	363.2 ± 14.8	-120.9 ± 6.3	-20.1 ± 2.3

Table 2: **ANIL matches the performance of MAML on few-shot image classification and RL.** On benchmark few-shot classification tasks MAML and ANIL have comparable accuracy, and also comparable average return (the higher the better) on standard RL tasks (6).

4 THE ANIL (ALMOST NO INNER LOOP) ALGORITHM

In the previous section we saw that for all layers except the head of the neural network, the meta-initialization learned by the outer loop of MAML results in very good features that can be reused as is on new tasks. Inner loop adaptation does not significantly change the representations of these layers, even from early on in training. This suggests a natural simplification of the MAML algorithm: the *ANIL (Almost No Inner Loop) algorithm*.

In ANIL, during training **and** testing, we *remove* the inner loop updates for the network body, and apply inner loop adaptation *only to the head*. The head requires the inner loop to allow it to align to the different classes in each task. In Section 5.1 we consider another variant, the *NIL (No Inner Loop) algorithm*, that removes the head entirely at test time, and uses learned features and cosine similarity to perform effective classification, thus avoiding inner loop updates altogether.

For the ANIL algorithm, mathematically, let $\theta = (\theta_1, ..., \theta_l)$ be the (meta-initialization) parameters for the l layers of the network. Following the notation of Section 3.1, let $\theta_m^{(b)}$ be the parameters after m inner gradient updates for task $\mathcal{T}_b$. In ANIL, we have that:

$$\theta_m^{(b)} = \left(\theta_1, \ldots, (\theta_l)_{m-1}^{(b)} - \alpha \nabla_{(\theta_l)_{m-1}^{(b)}} \mathcal{L}_{S_b}(f_{\theta_{m-1}^{(b)}})\right)$$

i.e. only the final layer gets the inner loop updates. As before, we then define the meta-loss, and compute the outer loop gradient update. The intuition for ANIL arises from Figure 3, where we observe that inner loop updates have little effect on the network body even early in training, suggesting the possibility of removing them entirely. Note that this is distinct to the freezing experiments, where we only removed the inner loop at inference time. Figure 4 presents the difference between MAML and ANIL, and Appendix C.1 considers a simple example of the gradient update in ANIL, showing how the ANIL update differs from MAML.

Computational benefit of ANIL: As ANIL almost has no inner loop, it significantly speeds up both training and inference. We found an average speedup of 1.7x per training iteration over MAML and an average speedup of 4.1x per inference iteration. In Appendix C.5 we provide the full results.

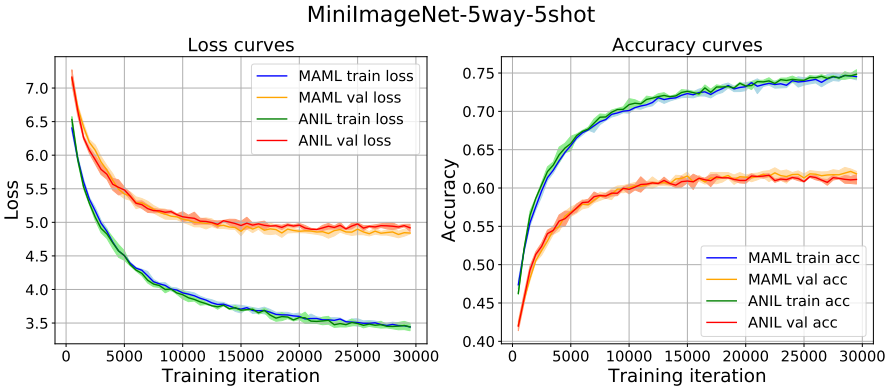

Figure 5: **MAML and ANIL learn very similarly**. Loss and accuracy curves for MAML and ANIL on MiniImageNet-5way-5shot, illustrating how MAML and ANIL behave similarly through the training process.

Model Pair	CCA Similarity	CKA Similarity
MAML-MAML	0.51	0.83
ANIL-ANIL	0.51	0.86
ANIL-MAML	0.50	0.83

Table 3: **MAML and ANIL models learn comparable representations.** Comparing CCA/CKA similarity scores of the of MAML-ANIL representations (averaged over network body), and MAML-MAML and ANIL-ANIL similarity scores (across different random seeds) shows algorithmic differences between MAML/ANIL does not result in vastly different types of features learned.

Results of ANIL on Standard Benchmarks: We evaluate ANIL on few-shot image classification and RL benchmarks, using the same model architectures as the original MAML authors, for both supervised learning and RL. Further implementation details are in Appendix C.4. The results in Table 2 (mean and standard deviation of performance over three random initializations) show that ANIL matches the performance of MAML on both few-shot classification (accuracy) and RL (average return, the higher the better), demonstrating that the inner loop adaptation of the body is unnecessary for learning good features.

MAML and ANIL Models Show Similar Behavior: MAML and ANIL perform equally well on few-shot learning benchmarks, illustrating that removing the inner loop during training does not hinder performance. To study the behavior of MAML and ANIL models further, we plot learning curves for both algorithms on MiniImageNet-5way-5shot, Figure 5. We see that loss and accuracy for both algorithms look very similar throughout training. We also look at CCA and CKA scores of the representations learned by both algorithms, Table 3. We observe that MAML-ANIL representations have the same average similarity scores as MAML-MAML and ANIL-ANIL representations, suggesting both algorithms learn comparable features (removing the inner loop doesn't change the kinds of features learned.) Further learning curves and representational similarity results are presented in Appendices C.2 and C.3.

5 Contributions of the Network Head and Body

So far, we have seen that MAML predominantly relies on feature reuse, with the network body (all layers except the last layer) already containing good features at meta-initialization. We also observe that such features can be learned even without inner loop adaptation during training (ANIL algorithm). The head, however, requires inner loop adaptation to enable task specificity.

In this section, we explore the contributions of the network head and body. We first ask: *How important is the head at test time, when good features have already been learned?* Motivating this question is that these features needed no adaptation at inference time, so perhaps they are themselves

Method	Omniglot-20way-1shot	Omniglot-20way-5shot	MiniImageNet-5way-1shot	MiniImageNet-5way-5shot
MAML	93.7 ± 0.7	96.4 ± 0.1	46.9 ± 0.2	63.1 ± 0.4
ANIL	96.2 ± 0.5	98.0 ± 0.3	46.7 ± 0.4	61.5 ± 0.5
NIL	96.7 ± 0.3	98.0 ± 0.04	48.0 ± 0.7	62.2 ± 0.5

Table 4: **NIL algorithm performs as well as MAML and ANIL on few-shot image classification.** Performance of MAML, ANIL, and NIL on few-shot image classification benchmarks. We see that with no test-time inner loop, and just learned features, NIL performs comparably to MAML and ANIL, indicating the strength of the learned features, and the relative lack of importance of the head at test time.

sufficient to perform classification, with no head. In Section 5.1, we find that test time performance is entirely determined by the quality of these representations, and we can use similarity of the frozen meta-initialization representations to perform unseen tasks, *removing the head entirely*. We call this the NIL (No Inner Loop) algorithm.

Given this result, we next study how useful the head is at training (in ensuring the network body learns good features). We look at multiple different training regimes (some without the head) for the network body, and evaluate the quality of the representations. We find that MAML/ANIL result in the best representations, demonstrating the importance of the head during training for feature learning.

5.1 The Head at Test Time and the NIL (No Inner Loop) Algorithm

Here, we study how important the head (and task specific alignment) are, when good features have *already* been learned (through training) by the meta-initialization. At test time, we find that the representations can be used directly, with *no* adaptation, which leads to the *No Inner Loop (NIL) algorithm*:

1. Train a few-shot learning model with ANIL/MAML algorithm as standard. We use ANIL training.
2. At test time, remove the head of the trained model. For each task, first pass the k labelled examples (support set) through the body of the network, to get their penultimate layer representations. Then, for a test example, compute cosine similarities between its penultimate layer representation and those of the support set, using these similarities to weight the support set labels, as in (37).

The results for the NIL algorithm, following ANIL training, on few-shot classification benchmarks are given in Table 4. Despite having no network head and no task specific adaptation, NIL performs comparably to MAML and ANIL. This demonstrates that the features learned by the network body when training with MAML/ANIL (and reused at test time) are the critical component in tackling these benchmarks.

5.2 Training Regimes for the Network Body

The NIL algorithm and results of Section 5.1, lead to the question of how important task alignment and the head are during training to ensure good features. Here, we study this question by examining the quality of features arising from different training regimes for the body. We look at (i) MAML and ANIL training; (ii) *multiclass classification*, where all of the training data and classes (from which training tasks are drawn) are used to perform standard classification; (iii) *multitask training*, a standard baseline, where no inner loop or task specific head is used, but the network is trained on all the tasks at the same time; (iv) *random features*, where the network is not trained at all, and features are frozen after random initialization; (v) NIL at training time, where there is no head and cosine distance on the representations is used to get the label.

After training, we apply the NIL algorithm to evaluate test performance, and quality of features learned at training. The results are shown in Table 5. MAML and ANIL training performs best. Multitask training, which has no task specific head, performs the worst, even worse than random features (adding evidence for the need for task specificity at training to facilitate feature learning.)

Method	MiniImageNet-5way-1shot	MiniImageNet-5way-5shot
MAML training-NIL head	48.4 ± 0.3	61.5 ± 0.8
ANIL training-NIL head	48.0 ± 0.7	62.2 ± 0.5
Multiclass training-NIL head	39.7 ± 0.3	54.4 ± 0.5
Multitask training-NIL head	26.5 ± 1.1	34.2 ± 3.5
Random features-NIL head	32.9 ± 0.6	43.2 ± 0.5
NIL training-NIL head	38.3 ± 0.6	43.0 ± 0.2

Table 5: **MAML/ANIL training leads to superior features learned, supporting importance of head at training.** Training with MAML/ANIL leads to superior performance over other methods which do not have task specific heads, supporting the importance of the head at training.

Using NIL during training performs worse than MAML/ANIL. These results demonstrate that the head is important at training to learn good features in the network body.

In Appendix D.1, we study test time performance variations from using a MAML/ANIL head instead of NIL, finding (as suggested by Section 5.1) very little performance difference. Additional results on similarity between the representations of different training regimes is given in Appendix D.2.

6 Feature Reuse in Other meta-learning Algorithms

Up till now, we have closely examined the MAML algorithm, and have demonstrated empirically that the algorithm's success is primarily due to feature reuse, rather than rapid learning. We now discuss rapid learning vs feature reuse more broadly in meta-learning. By combining our results with an analysis of evidence reported in prior work, we find support for many meta-learning algorithms succeeding via feature reuse, identifying a common theme characterizing the operating regime of much of current meta-learning.

6.1 Optimization and Model Based meta-learning

MAML falls within the broader class of *optimization based* meta-learning algorithms, which at inference time, directly optimize model parameters for a new task using the support set. MAML has inspired many other optimization-based algorithms, which utilize the same two-loop structure (19; 29; 8). Our analysis so far has thus yielded insights into the feature reuse vs rapid learning question for this class of algorithms. Another broad class of meta-learning consists of *model based* algorithms, which also have notions of rapid learning and feature reuse.

In the model-based setting, the meta-learning model's parameters are *not* directly optimized for the specific task on the support set. Instead, the model typically conditions its output on some representation of the task definition. One way to achieve this conditioning is to *jointly encode* the entire support set in the model's latent representation (37; 33), enabling it to adapt to the characteristics of each task. This constitutes rapid learning for model based meta-learning algorithms.

An alternative to joint encoding would be to encode each member of the support set independently, and apply a cosine similarity rule (as in (37)) to classify an unlabelled example. This mode of operation is purely feature reuse – we do not use information defining the task to directly influence the decision function.

If joint encoding gave significant test-time improvement over non-joint encoding, then this would suggest that rapid learning of the test-time task is taking place, as task specific information is being utilized to influence the model's decision function. However, on analyzing results in prior literature, this improvement appears to be minimal. Indeed, in e.g. Matching Networks (37), using joint encoding one reaches 44.2% accuracy on MiniImageNet-5way-1shot, whereas with independent encoding one obtains 41.2%: a small difference. More refined models suggest the gap is even smaller. For instance, in (5), many methods for one shot learning were re-implemented and studied, and baselines without joint encoding achieved 48.24% accuracy in MiniImageNet-5way-1shot, whilst other models using joint encoding such as Relation Net (33) achieves very similar accuracy of 49.31%

(they also report MAML, at 46.47%). As a result, we believe that the dominant mode of "feature reuse" rather than "rapid learning" is what has currently dominated both MAML-styled optimization based meta-learning *and* model based meta-learning.

7 CONCLUSION

In this paper, we studied a fundamental question on whether the highly successful MAML algorithm relies on rapid learning or feature reuse. Through a series of experiments, we found that *feature reuse* is the dominant component in MAML's efficacy. This insight led to the ANIL (Almost No Inner Loop) algorithm, a simplification of MAML that has identical performance on standard image classification and reinforcement learning benchmarks, and provides computational benefits. We further study the importance of the head (final layer) of a neural network trained with MAML, discovering that the body (lower layers) of a network is sufficient for few-shot classification at test time, allowing us to remove the network head for testing (NIL) and still match performance. We connected our results to the broader literature in meta-learning, identifying feature reuse to be a common mode of operation for other meta-learning algorithms also. Based off of our conclusions, future work could look at developing and analyzing new meta-learning algorithms that perform more rapid learning, which may expand the datasets and problems amenable to these techniques.

REFERENCES

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018.

[2] PMP Art. Overview figure: Lion image. https://www.pinterest.com/pin/350436414739113168/. Accessed: 2019-09-09.

[3] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*, 2018.

[4] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[5] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[7] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *arXiv preprint arXiv:1710.11622*, 2017.

[8] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, pages 9516–9527, 2018.

[9] Nathan Glover. Overview figure: Cat image. https://thesecatsdonotexist.com/. Accessed: 2019-09-09.

[10] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

[11] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

[12] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

[13] GuideDogVerity. Overview figure: Dog image. https://twitter.com/guidedogverity. Accessed: 2019-09-09.

[14] James Harrison, Apoorva Sharma, and Marco Pavone. Meta-learning priors for efficient online bayesian regression. *arXiv preprint arXiv:1807.08912*, 2018.

[15] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. *arXiv preprint arXiv:1810.02334*, 2018.

[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[17] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019.

[18] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.

[19] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. *arXiv preprint arXiv:1801.05558*, 2018.

[20] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John E Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *FE@ NIPS*, pages 196–212, 2015.

[21] Niru Maheswaranathan, Alex H. Willams, Matthew D. Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *arXiv preprint arXiv:1907.08549*, 2019.

[22] Ahmed Malik. Overview figure: Plane image. `https://appadvice.com/app/real-airplane-pilot-flight-simulator-game-for-free/1186146488`. Accessed: 2019-09-09.

[23] Ari S Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *arXiv preprint arXiv:1806.05759*, 2018.

[24] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2, 2018.

[25] Maithra Raghu. Svcca code and tutorials. `https://github.com/google/svcca`.

[26] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017.

[27] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

[28] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[29] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[30] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

[31] Naomi Saphra and Adam Lopez. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*, 2018.

[32] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[33] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[34] Save Sheffield Trees. Overview figure: Tree image. `https://twitter.com/saveshefftrees`. Accessed: 2019-09-09.

[35] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

[36] Vexels. Overview figure: Chair image. `https://www.vexels.com/png-svg/preview/148959/small-office-chair-clipart`. Accessed: 2019-09-09.

[37] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[38] Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John E. Hopcroft. To what extent do different neural networks learn the same representation: A neuron activation subspace match approach. In *NIPS 2018*, 2018.

[39] Fengwei Zhou, Bin Wu, and Zhenguo Li. Deep meta-learning: Learning to learn in the concept space. *arXiv preprint arXiv:1802.03596*, 2018.

[40] Luisa M Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Caml: Fast context adaptation via meta-learning. *arXiv preprint arXiv:1810.03642*, 2018.

A FEW-SHOT IMAGE CLASSIFICATION DATASETS AND EXPERIMENTAL SETUPS

We consider the few-shot learning paradigm for image classification to evaluate MAML and ANIL. We evaluate using two datasets often used for few-shot multiclass classification – the Omniglot dataset and the MiniImageNet dataset.

Omniglot: The Omniglot dataset consists of over 1600 different handwritten character classes from 23 alphabets. The dataset is split on a character-level, so that certain characters are in the training set, and others in the validation set. We consider the 20-way 1-shot and 20-way 5-shot tasks on this dataset, where at test time, we wish our classifier to discriminate between 20 randomly chosen character classes from the held-out set, given only 1 or 5 labelled example(s) from each class from this set of 20 testing classes respectively. The model architecture used is identical to that in the original MAML paper, namely: 4 modules with a 3 x 3 convolutions and 64 filters with a stride of 2, followed by batch normalization, and a ReLU nonlinearity. The Omniglot images are downsampled to 28 x 28, so the dimensionality of the last hidden layer is 64. The last layer is fed into a 20-way softmax. Our models are trained using a batch size of 16, 5 inner loop updates, and an inner learning rate of 0.1.

MiniImageNet: The MiniImagenet dataset was proposed by (28), and consists of 64 training classes, 12 validation classes, and 24 test classes. We consider the 5-way 1-shot and 5-way 5-shot tasks on this dataset, where the test-time task is to classify among 5 different randomly chosen validation classes, given only 1 and 5 labelled examples respectively. The model architecture is again identical to that in the original paper: 4 modules with a 3 x 3 convolutions and 32 filters, followed by batch normalization, ReLU nonlinearity, and 2 x 2 max pooling. Our models are trained using a batch size of 4. 5 inner loop update steps, and an inner learning rate of 0.01 are used. 10 inner gradient steps are used for evaluation at test time.

B ADDITIONAL FREEZING AND REPRESENTATIONAL SIMILARITY DETAILS

In this section, we provide further experimental details and results from freezing and representational similarity experiments.

B.1 EXPERIMENTAL DETAILS

We concentrate on MiniImageNet for our freezing and representational similarity experiments in Section 3.2, as it is more complex than Omniglot.

The model architecture used for our experiments is identical to that in the original paper: 4 modules with a 3 x 3 convolutions and 32 filters, followed by batch normalization, ReLU nonlinearity, and 2 x 2 max pooling. Our models are trained using a batch size of 4, 5 inner loop update steps, and an inner learning rate of 0.01. 10 inner gradient steps are used for evaluation at test time. We train models 3 times with different random seeds. Models were trained for 30000 iterations.

B.2 DETAILS OF REPRESENTATIONAL SIMILARITY

CCA takes in as inputs $L_1 = \{z_1^{(1)}, z_2^{(1)}, ..., z_m^{(1)}\}$ and $L_2 = \{z_1^{(2)}, z_2^{(1)}, ..., z_n^{(2)}\}$, where L_1, L_2 are layers, and $z_i^{(j)}$ is a neuron activation vector: the vector of outputs of neuron i (of layer L_j) over a set of inputs X. It then finds linear combinations of the neurons in L_1 and neurons in L_2 so that the resulting activation vectors are maximally correlated, which is summarized in the canonical correlation coefficient. Iteratively repeating this process gives a similarity score (in $[0, 1]$ with 1 identical and 0 completely different) between the representations of L_1 and L_2.

We apply this to compare corresponding layers of two networks, net1 and net2, where net1 and net2 might differ due to training step, training method (ANIL vs MAML) or the random seed. When comparing convolutional layers, as described in (25), we perform the comparison over channels, flattening out over all of the spatial dimensions, and then taking the mean CCA coefficient. We average over three random repeats.

B.3 SIMILARITY BEFORE AND AFTER INNER LOOP WITH EUCLIDEAN DISTANCE

In addition to assessing representational similarity with CCA/CKA, we also consider the simpler measure of Euclidean distance, capturing how much weights of the network change during the inner loop update (task-specific finetuning). We note that this experiment does not assess functional changes on inner loop updates as well as the CCA experiments do; however, they serve to provide useful intuition.

We plot the per-layer average Euclidean distance between the initialization θ and the finetuned weights $\theta_m^{(b)}$ across different tasks T_b, i.e.

$$\frac{1}{N}\sum_{b=1}^{N} ||(\theta_l) - (\theta_l)_m^{(b)}||$$

across different layers l, for MiniImageNet in Figure 6. We observe that very quickly after the start of training, all layers except for the last layer have small Euclidean distance difference before and after finetuning, suggesting significant feature reuse. (Note that this is despite the fact that these layers have more parameters than the final layer.)

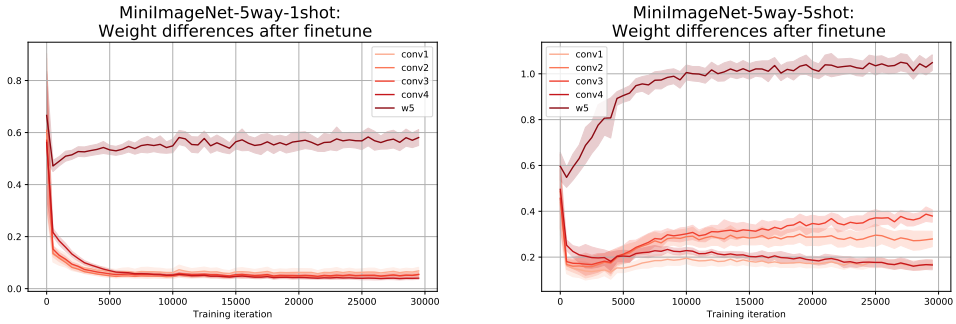

Figure 6: **Euclidean distance before and after finetuning for MiniImageNet**. We compute the average (across tasks) Euclidean distance between the weights before and after inner loop adaptation, separately for different layers. We observe that all layers except for the final layer show very little difference before and after inner loop adaptation, suggesting significant feature reuse.

B.4 CCA SIMILARITY ACROSS RANDOM SEEDS

The experiment in Section 3.2.2 compared representational similarity of L_1 and L_2 at different points in training (before/after inner loop adaptation) but corresponding to the same random seed. To complete the picture, it is useful to study whether representational similarity across *different* random seeds is also mostly unaffected by the inner loop adaptation. This motivates four natural comparisons: assume layer L_1 is from the first seed, and layer L_2 is from the second seed. Then we can compute the representational similarity between (L_1 pre, L_2 pre), (L_1 pre, L_2 post), (L_1 post, L_2 pre) and (L_1 post, L_2 post), where pre/post signify whether we take the representation before or after adaptation.

Prior work has shown that neural network representations may vary across different random seeds (26; 23; 20; 38), organically resulting in CCA similarity scores much less than 1. So to identify the effect of the inner loop on the representation, we plot the CCA similarities of (i) (L_1 pre, L_2 pre) against (L_1 pre, L_2 post) and (ii) (L_1 pre, L_2 pre) against (L_1 post, L_2 pre) and (iii) (L_1 pre, L_2 pre) against (L_1 post, L_2 post) separately across the different random seeds and different layers. We then compute the line of best fit for each plot. If the line of best fit fits the data and is close to $y = x$, this suggests that the inner loop adaptation doesn't affect the features much – the similarity before adaptation is very close to the similarity after adaptation.

The results are shown in Figure 7. In all of the plots, we see that the line of best fit is almost exactly $y = x$ (even for the pre/pre vs post/post plot, which could conceivably be more different as both seeds change) and a computation of the coefficient of determination R^2 gives $R^2 \approx 1$ for all three plots. Putting this together with Figure 2, we can conclude that the inner loop adaptation step doesn't

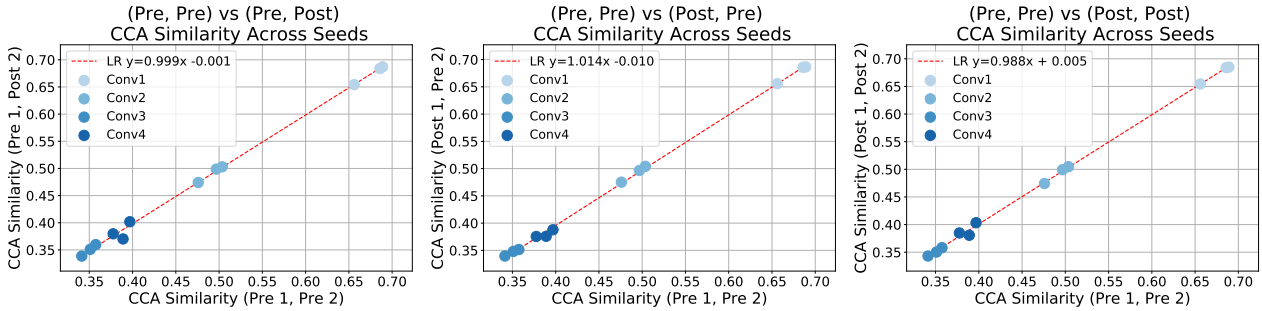

Figure 7: **Computing CCA similarity pre/post adaptation across different random seeds further demonstrates that the inner loop doesn't change representations significantly.** We compute CCA similarity of L_1 from seed 1 and L_2 from seed 2, varying whether we take the representation *pre* (before) adaptation or *post* (after) adaptation. To isolate the effect of adaptation from inherent variation in the network representation across seeds, we plot CCA similarity of of the representations before adaptation against representations after adaptation in three different combinations: (i) (L_1 pre, L_2 pre) against (L_1 pre, L_1 post), (ii) (L_1 pre, L_2 pre) against (L_1 pre, L_1 post) (iii) (L_1 pre, L_2 pre) against (L_1 post, L_2 post). We do this separately across different random seeds and different layers. Then, we compute a line of best fit, finding that in all three plots, it is almost identical to $y = x$, demonstrating that the representation does not change significantly pre/post adaptation. Furthermore a computation of the coefficient of determination R^2 gives $R^2 \approx 1$, illustrating that the data is well explained by this relation. In Figure 8, we perform this comparison with CKA, observing the same high level conclusions.

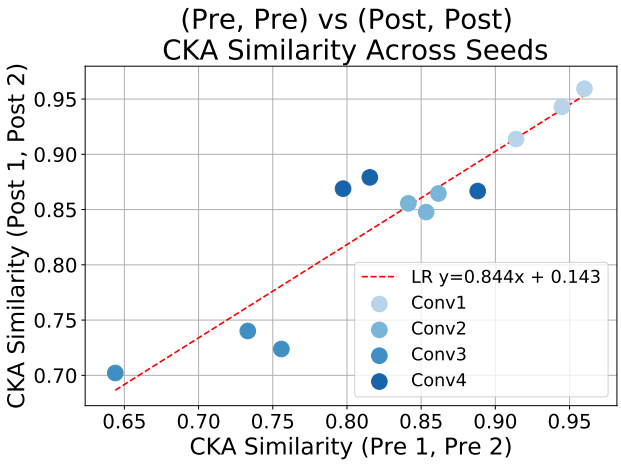

Figure 8: We perform the same comparison as in Figure 7, but with CKA instead. There is more variation in the similarity scores, but we still see a strong correlation between (Pre, Pre) and (Post, Post) comparisons, showing that representations do not change significantly over the inner loop.

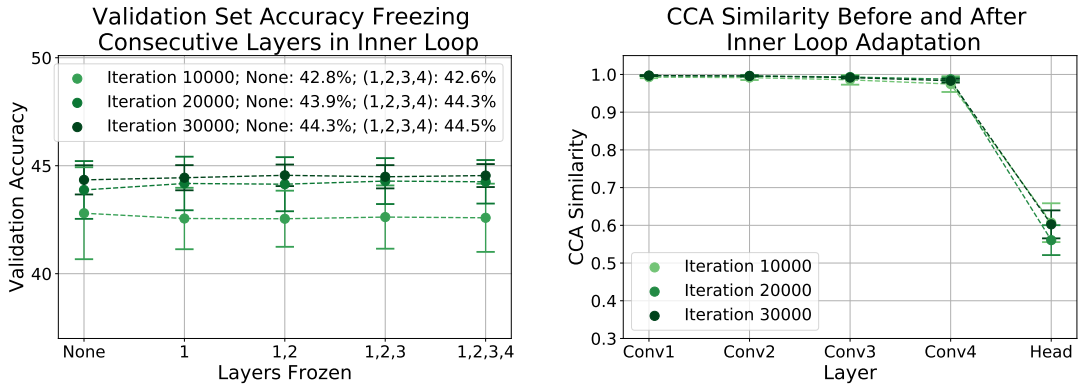

Figure 9: **Inner loop updates have little effect on learned representations from early on in learning.** We consider freezing and representational similarity experiments for MiniImageNet-5way-1shot. We see that early on in training (from as few as 10k iterations in), the inner loop updates have little effect on the learned representations and features, and that removing the inner loop updates for all layers but the head have little-to-no impact on the validation set accuracy.

affect the representation learned by any layer except the head, and that the learned representations and features are mostly reused as is for the different tasks.

B.5 MiniImageNet-5way-1shot Freezing and CCA Over Training

Figure 9 shows that from early on in training, on MiniImageNet-5way-1shot, that the CCA similarity between activations pre and post inner loop update is very high for all layers but the head. We further see that the validation set accuracy suffers almost no decrease if we remove the inner loop updates and freeze all layers but the head. This shows that even early on in training, the inner loop appears to have minimal effect on learned representations and features. This supplements the results seen in Figure 3 on MiniImageNet-5way-5shot.

C ANIL Algorithm: More Details

In this section, we provide more details about the ANIL algorithm, including an example of the ANIL update, implementation details, and further experimental results.

C.1 An Example of the ANIL Update

Consider a simple, two layer linear network with a single hidden unit in each layer: $\hat{y}(x; \boldsymbol{\theta}) = \theta_2(\theta_1 x)$. In this example, θ_2 is the *head*. Consider the 1-shot regression problem, where we have access to examples $\left\{(x_1^{(t)}, y_1^{(t)}), (x_2^{(t)}, y_2^{(t)})\right\}$ for tasks $t = 1, \ldots, T$. Note that $(x_1^{(t)}, y_1^{(t)})$ is the (example, label) pair in the meta-training set (used for inner loop adaptation – support set), and $(x_2^{(t)}, y_2^{(t)})$ is the pair in the meta-validation set (used for the outer loop update – target set).

In the few-shot learning setting, we firstly draw a set of N tasks and labelled examples from our meta-training set: $\left\{(x_1^{(1)}, y_1^{(1)}), \ldots, (x_1^{(N)}, y_1^{(N)})\right\}$. Assume for simplicity that we only apply one gradient step in the inner loop. The inner loop updates for each task are thus defined as follows:

$$\theta_1^{(t)} \leftarrow \theta_1 - \frac{\partial L(\hat{y}(x_1^{(t)}; \boldsymbol{\theta}), y_1^{(t)})}{\partial \theta_1} \tag{1}$$

$$\theta_2^{(t)} \leftarrow \theta_2 - \frac{\partial L(\hat{y}(x_1^{(t)}; \boldsymbol{\theta}), y_1^{(t)})}{\partial \theta_2} \tag{2}$$

where $L(\cdot, \cdot)$ is the loss function, (e.g. mean squared error) and $\theta_i^{(t)}$ refers to a parameter after inner loop update for task t.

The task-adapted parameters for MAML and ANIL are as follows. Note how only the head parameters change per-task in ANIL:

$$\boldsymbol{\theta}^{(t)}_{\text{MAML}} = \left[\theta_1^{(t)}, \theta_2^{(t)}\right] \tag{3}$$

$$\boldsymbol{\theta}^{(t)}_{\text{ANIL}} = \left[\theta_1, \theta_2^{(t)}\right] \tag{4}$$

In the outer loop update, we then perform the following operations using the data from the meta-validation set:

$$\theta_1 \leftarrow \theta_1 - \sum_{t=1}^{N} \frac{\partial L(\hat{y}(x_2^{(t)}; \boldsymbol{\theta}^{(t)}), y_2^{(t)})}{\partial \theta_1} \tag{5}$$

$$\theta_2 \leftarrow \theta_2 - \sum_{t=1}^{N} \frac{\partial L(\hat{y}(x_2^{(t)}; \boldsymbol{\theta}^{(t)}), y_2^{(t)})}{\partial \theta_2} \tag{6}$$

Considering the update for θ_1 in more detail for our simple, two layer, linear network (the case for θ_2 is analogous), we have the following update for MAML:

$$\theta_1 \leftarrow \theta_1 - \sum_{t=1}^{N} \frac{\partial L(\hat{y}(x_2^{(t)}; \boldsymbol{\theta}^{(t)}_{\text{MAML}}), y_2^{(t)})}{\partial \theta_1} \tag{7}$$

$$\hat{y}(x_2^{(t)}; \boldsymbol{\theta}^{(t)}_{\text{MAML}}) = \left(\left[\theta_2 - \frac{\partial L(\hat{y}(x_1^{(t)}; \boldsymbol{\theta}), y_1^{(t)})}{\partial \theta_2} \right] \cdot \left[\theta_1 - \frac{\partial L(\hat{y}(x_1^{(t)}; \boldsymbol{\theta}), y_1^{(t)})}{\partial \theta_1} \right] \cdot x_2 \right) \tag{8}$$

For ANIL, on the other hand, the update will be:

$$\theta_1 \leftarrow \theta_1 - \sum_{t=1}^{N} \frac{\partial L(\hat{y}(x_2^{(t)}; \boldsymbol{\theta}^{(t)}_{\text{ANIL}}), y_2^{(t)})}{\partial \theta_1} \tag{9}$$

$$\hat{y}(x_2^{(t)}; \boldsymbol{\theta}^{(t)}_{\text{ANIL}}) = \left(\left[\theta_2 - \frac{\partial L(\hat{y}(x_1^{(t)}; \boldsymbol{\theta}), y_1^{(t)})}{\partial \theta_2} \right] \cdot \theta_1 \cdot x_2 \right) \tag{10}$$

Note the lack of inner loop update for θ_1, and how we do not remove second order terms in ANIL (unlike in first-order MAML); second order terms still persist through the derivative of the inner loop update for the head parameters.

C.2 ANIL Learns Almost Identically to MAML

We implement ANIL on MiniImageNet and Omniglot, and generate learning curves for both algorithms in Figure 10. We find that learning proceeds almost identically for ANIL and MAML, showing that removing the inner loop has little effect on the learning dynamics.

C.3 ANIL and MAML Learn Similar Representations

We compute CCA similarities across representations in a MAML seed and an ANIL seed, and then plot these against the same MAML seed representation compared to a different MAML seed (and similarly for ANIL). We find a strong correlation between these similarities (Figure 11), which suggests that MAML and ANIL are learning similar representations, despite their algorithmic differences. (ANIL and MAML are about as similar to each other as two ANILs are to each other, or two MAMLs are to each other.)

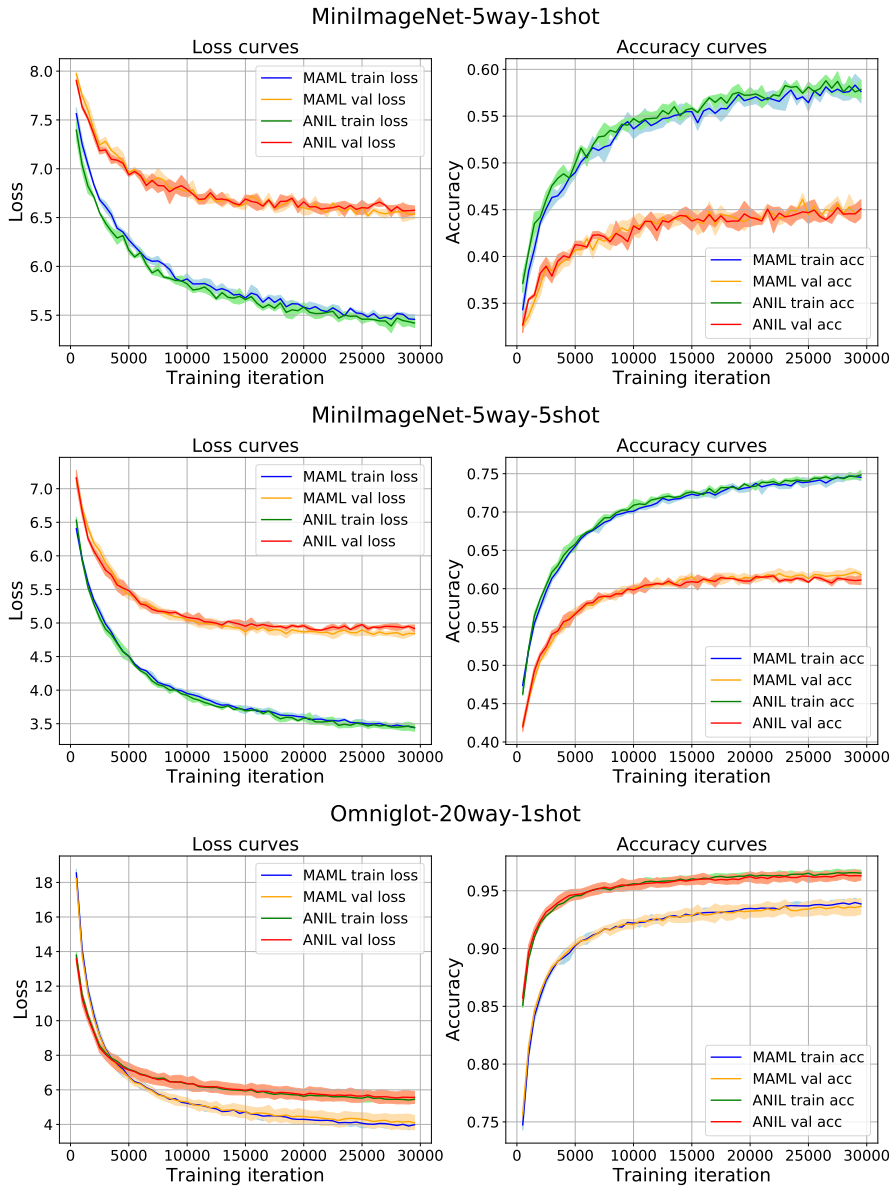

Figure 10: **ANIL and MAML on MiniImageNet and Omniglot**. Loss and accuracy curves for ANIL and MAML on (i) MiniImageNet-5way-1shot (ii) MiniImageNet-5way-5shot (iii) Omniglot-20way-1shot. These illustrate how both algorithms learn very similarly over training.

C.4 ANIL Implementation Details

Supervised Learning Implementation: We used the TensorFlow MAML implementation open-sourced by the original authors (6). We used the same model architectures as in the original MAML paper for our experiments, and train models 3 times with different random seeds. All models were trained for 30000 iterations, with a batch size of 4, 5 inner loop update steps, and an inner learning rate of 0.01. 10 inner gradient steps were used for evaluation at test time.

Reinforcement Learning Implementation: We used the open source PyTorch implementation of MAML for RL [1], due to challenges encountered when running the open-sourced TensorFlow implementation from the original authors. We note that the results for MAML in these RL domains do

[1] https://github.com/tristandeleu/pytorch-maml-rl

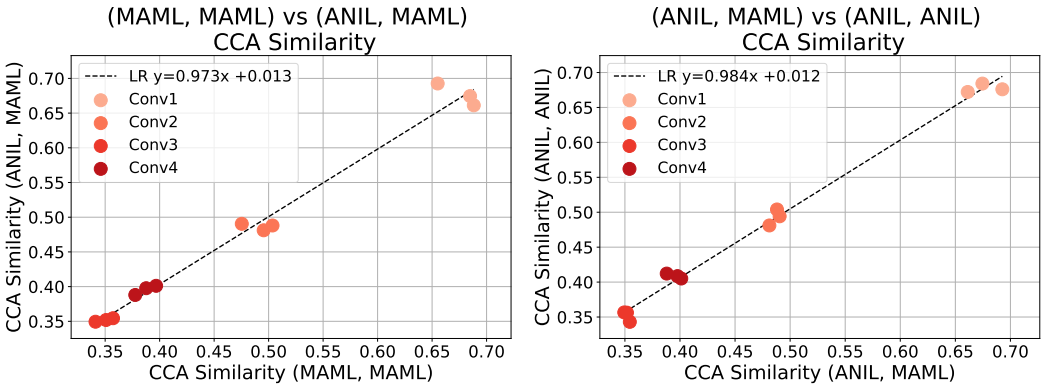

Figure 11: **Computing CCA similarity across different seeds of MAML and ANIL networks suggests these representations are similar.** We plot the CCA similarity between an ANIL seed and a MAML seed, plotted against (i) the MAML seed compared to a different MAML seed (ii) the ANIL seed compared to a different ANIL seed. We observe a strong correlation of similarity scores in both (i) and (ii). This tells us that (i) two MAML representations vary about as much as MAML and ANIL representations (ii) two ANIL representations vary about as much as MAML and ANIL representations. In particular, this suggests that MAML and ANIL learn similar features, despite having significant algorithmic differences.

	Training: 5way-1shot			Training: 5way-5shot		
	Mean (s)	Median (s)	Speedup	Mean (s)	Median (s)	Speedup
MAML	0.15	0.13	1	0.68	0.67	1
First Order MAML	0.089	0.083	1.69	0.40	0.39	1.7
ANIL	0.084	0.072	1.79	0.37	0.36	1.84

	Inference: 5way-1shot			Inference: 5way-5shot		
	Mean (s)	Median (s)	Speedup	Mean (s)	Median (s)	Speedup
MAML	0.083	0.078	1	0.37	0.36	1
ANIL	0.020	0.017	4.15	0.076	0.071	4.87

Table 6: **ANIL offers significant computational speedup over MAML, during both training and inference.** Table comparing execution times and speedups of MAML, First Order MAML, and ANIL during training (above) and inference (below) on MiniImageNet domains. Speedup is calculated relative to MAML's execution time. We see that ANIL offers noticeable speedup over MAML, as a result of removing the inner loop almost completely. This permits faster training and inference.

not exactly match those in the original paper; this may be due to large variance in results, depending on the random initialization. We used the same model architecture as the original paper (two layer MLP with 100 hidden units in each layer), a batch size of 40, 1 inner loop update step with an inner learning rate of 0.1 and 20 trajectories for inner loop adaptation. We trained three MAML and ANIL models with different random initialization, and quote the mean and standard deviation of the results. As in the original MAML paper, for RL experiments, we select the best performing model over 500 iterations of training and evaluate this model at test time on a new set of tasks.

C.5 ANIL is Computationally Simpler Than MAML

Table 6 shows results from a comparison of the computation time for MAML, First Order MAML, and ANIL, during training and inference, with the TensorFlow implementation described previously, on both MiniImageNet domains. These results are average time for executing forward and backward passes during training (above) and a forward pass during inference (bottom), for a task batch size of 1, and a target set size of 1. Results are averaged over 2000 such batches. Speedup is calculated relative to MAML's execution time. Each batches' images were loaded into memory before running the TensorFlow computation graph, to ensure that data loading time was not captured in the timing. Experiments were run on a single NVIDIA Titan-Xp GPU.

Method	MiniImageNet-5way-1shot	MiniImageNet-5way-5shot
MAML training-MAML head	46.9 ± 0.2	63.1 ± 0.4
MAML training-NIL head	48.4 ± 0.3	61.5 ± 0.8
ANIL training-ANIL head	46.7 ± 0.4	61.5 ± 0.5
ANIL training-NIL head	48.0 ± 0.7	62.2 ± 0.5
Multiclass pretrain-MAML head	38.4 ± 0.8	54.6 ± 0.4
Multiclass pretrain-NIL head	39.7 ± 0.3	54.4 ± 0.5
Multitask pretrain-MAML head	26.5 ± 0.8	32.8 ± 0.6
Multitask pretrain-NIL head	26.5 ± 1.1	34.2 ± 3.5
Random features-MAML head	32.1 ± 0.5	43.1 ± 0.3
Random features-NIL head	32.9 ± 0.6	43.2 ± 0.5

Table 7: **Test time performance is dominated by features learned, with no difference between NIL/MAML heads.** We see identical performances of MAML/NIL heads at test time, indicating that MAML/ANIL training leads to better learned features.

During training, we see that ANIL is as fast as First Order MAML (which does not compute second order terms during training), and about 1.7x as fast as MAML. This leads to a significant overall training speedup, especially when coupled with the fact that the rate of learning for these ANIL and MAML is very similar; see learning curves in Appendix C.2. Note that unlike First Order MAML, ANIL also performs very comparably to MAML on benchmark tasks (on some tasks, First Order MAML performs worse (6)). During inference, ANIL achieves over a 4x speedup over MAML (and thus also 4x over First Order MAML, which is identical to MAML at inference time). Both training and inference speedups illustrate the significant computational benefit of ANIL over MAML.

D Further Results on the Network Head and Body

D.1 Training Regimes for the Network Body

We add to the results of Section 5.2 in the main text by seeing if training a head and applying that to the representations at test time (instead of the NIL algorithm) gives in any change in the results. As might be predicted by Section 5.1, we find no change the results.

More specifically, we do the following:

- We train MAML/ANIL networks as standard, and do standard test time adaptation.
- For multiclass training, we first (pre)train with multiclass classification, then throw away the head and freeze the body. We initialize a new e.g. 5-class head, and train that (on top of the frozen multiclass pretrained features) with MAML. At test time we perform standard adaptation.
- The same process is applied to multitask training.
- A similar process is applied to random features, except the network is initialized and then frozen.

The results of this, along with the results from Table 5 in the main text is shown in Table 7. We observe very little performance difference between using a MAML/ANIL head and a NIL head for each training regime. Specifically, task performance is purely determined by the quality of the features and representations learned during training, with task-specific alignment at test time being (i) unnecessary (ii) unable to influence the final performance of the model (e.g. multitask training performance is equally with a MAML head as it is with a NIL-head.)

D.2 Representational Analysis of Different Training Regimes

Here we include results on using CCA and CKA on the representations learned by the different training methods. Specifically, we studied how similar representations of different training methods

were to MAML training, finding a direct correlation with performance – training schemes learning representations most similar to MAML also performed the best. We computed similarity scores by averaging the scores over the first three conv layers in the body of the network.

Feature pair	CCA Similarity	CKA Similarity
(MAML, MAML)	0.51	0.83
(Multiclass pretrain, MAML)	0.48	0.79
(Random features, MAML)	0.40	0.72
(Multitask pretrain, MAML)	0.28	0.65

Table 8: **MAML training most closely resembles multiclass pretraining, as illustrated by CCA and CKA similarities**. On analyzing the CCA and CKA similarities between different baseline models and MAML (comparing across different tasks and seeds), we see that multiclass pretraining results in features most similar to MAML training. Multitask pretraining differs quite significantly from MAML-learned features, potentially due to the alignment problem.