CS 330 Autumn 2021/2022 Homework 2: Prototypical Networks and Model-Agnostic Meta-Learning

Kevin Lee (kelelee@stanford.edu)

Due on Monday October 18, 11:59 PM PST

1 Prototypical Networks (Protonets)

1.2

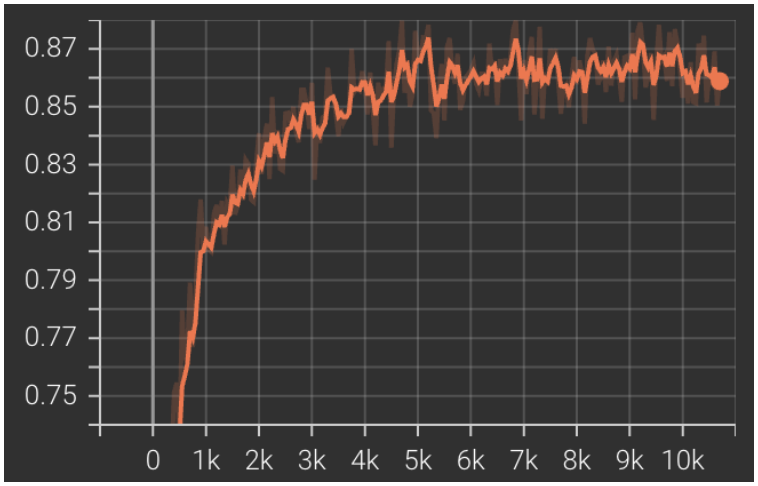

Figure 1: Validation query accuracy over the course of 5-way 5-shot Omniglot training on Protonet

1.3

(a) What do you notice about the train support and val support accuracy? What does this suggest about where the protonet places support examples of the same class in feature space?

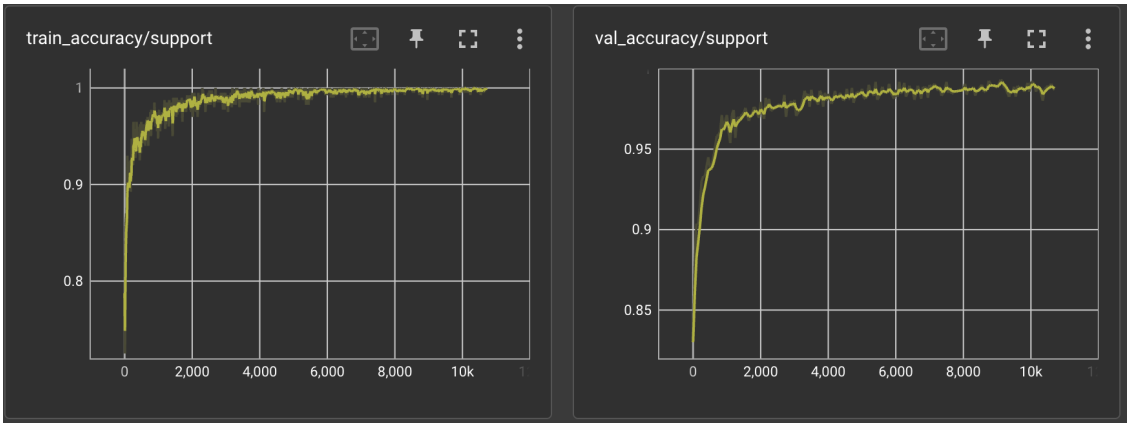

Figure 2: Train support accuracy (left) and validation support accuracy (right) over the course of 5-way 5-shot Omniglot training on Protonet

Answer:
Both train and val support accuracies are similarly very high. This suggests that protonet placed the support examples fairly optimally in feature space.

(b) Compare train query and val query. Is the model generalizing to new tasks? If not, is it overfitting or underfitting?

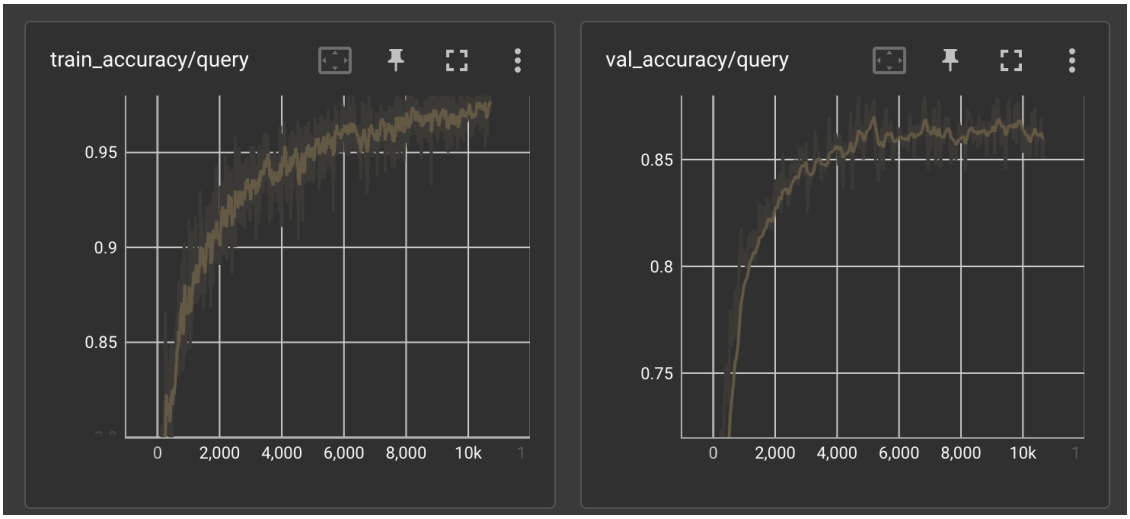

Figure 3: Train query accuracy (left) and validation query accuracy (right) over the course of 5-way 5-shot Omniglot training on Protonet

Answer:
The model is generalizing moderately well to new tasks, but since the validation query accuracy plateaued after 5,000 steps, it has at least achieved its best fit and is therefore not underfitting. Since there is no sign of decrease in validation query accuracy at later steps, there is also no signs of overfitting on the training tasks.

1.4

K	Mean Accuracy	95% confidence interval
1	0.489	0.008
5	0.692	0.008

Table 1: Test query accuracy for 5-way 1-shot vs 5-way 5-shot Omniglot on protonet

How did you choose which checkpoint to use for testing for each model?

Answer:
I selected the checkpoints at step 5000 for both $K = 1$ and $K = 5$, respectively, because that is approximately when the validation query accuracies have converged for each.

What do you notice about the test performance? If there is a difference, what could explain this difference?

Answer:
The test accuracy for $K = 1$ is lower than for $K = 5$ because the model has less samples to calculate more precise *prototypes*. The test accuracies are significantly lower than the validation accuracies in both $K = 1$ and $K = 5$ cases because there is bias in selecting a checkpoint based on the models' performances on the validation set, in addition to the fact that there are more test tasks than validation tasks.

2 Model-Agnostic Meta-Learning (MAML)

2.2

Submit a plot of the val post-adaptation query accuracy over the course of training.

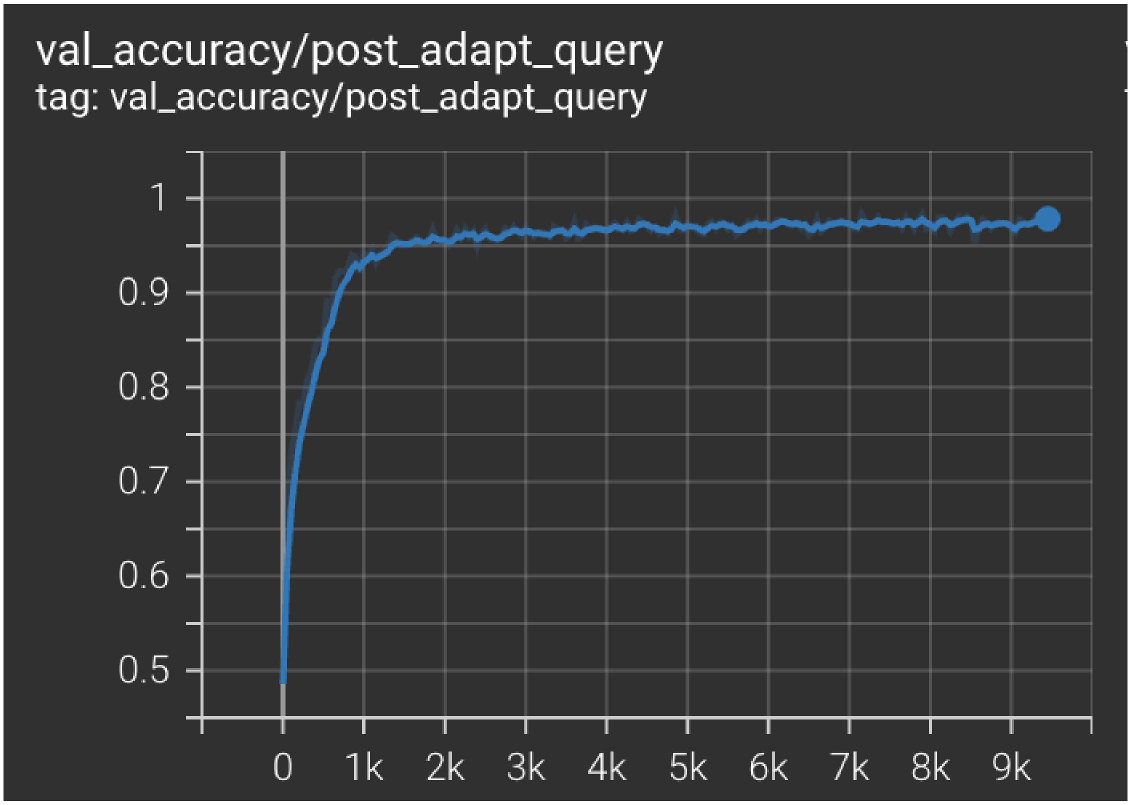

Figure 4: Validation post-adaptation query accuracy over the course of 5-way 1-shot Omniglot training on MAML

2.3

6 accuracy metrics are logged. Examine these in detail to reason about what MAML is doing. **Submit** responses to the following questions:

(a) What do you notice about the `train_pre_adapt_support` and `val_pre_adapt_support` accuracies? Why does this make sense given the task sampling process?

Answer:
Both `train_pre_adapt_support` and `val_pre_adapt_support` remained noisy throughout the first 10,000 steps of training and eventually converges to a stable mean value of $\frac{1}{N} = 0.2$. This makes sense because the model learns a "middle-ground" weight, θ to be used as a starting point for all tasks which, assuming tasks are distinct enough, would be far from optimal for any given task, at which point the model would do no better than guessing the labels for each query sample.

(b) What can you infer about the model from comparing the `train_pre_adapt_support` and `train_post_adapt_support` accuracies? And the corresponding val accuracies?

Answer:
Since `num_inner_steps` are kept constant, we see that the model progressively learns to achieve better train support accuracies after adaptation. This indicates that the learnt initial weights θ is getting closer and closer to all training tasks on average. The same observation holds true when we compare `val_pre_adapt_support` and `val_post_adapt_support`, indicating that the model did not overfit and can adapt similarly on unseen tasks.

(c) What about by comparing the `train_post_adapt_support` and `train_post_adapt_query` accuracies? And the corresponding val accuracies?

Answer:
Both `train_post_adapt_support` and `train_post_adapt_query` plots show a rise in accuracy followed by a plateau close to perfect accuracy within very few steps, and the same is observed for `val_post_adapt_support` and `val_post_adapt_query`. This indicates that the model is capable of adapting to the tasks quickly and accurately.

2.4

Try MAML with a fixed inner learning rate of 0.04. **Submit** a plot of the validation post-adaptation query accuracy over the course of training with for the two inner learning rates $(0.04, 0.4)$.

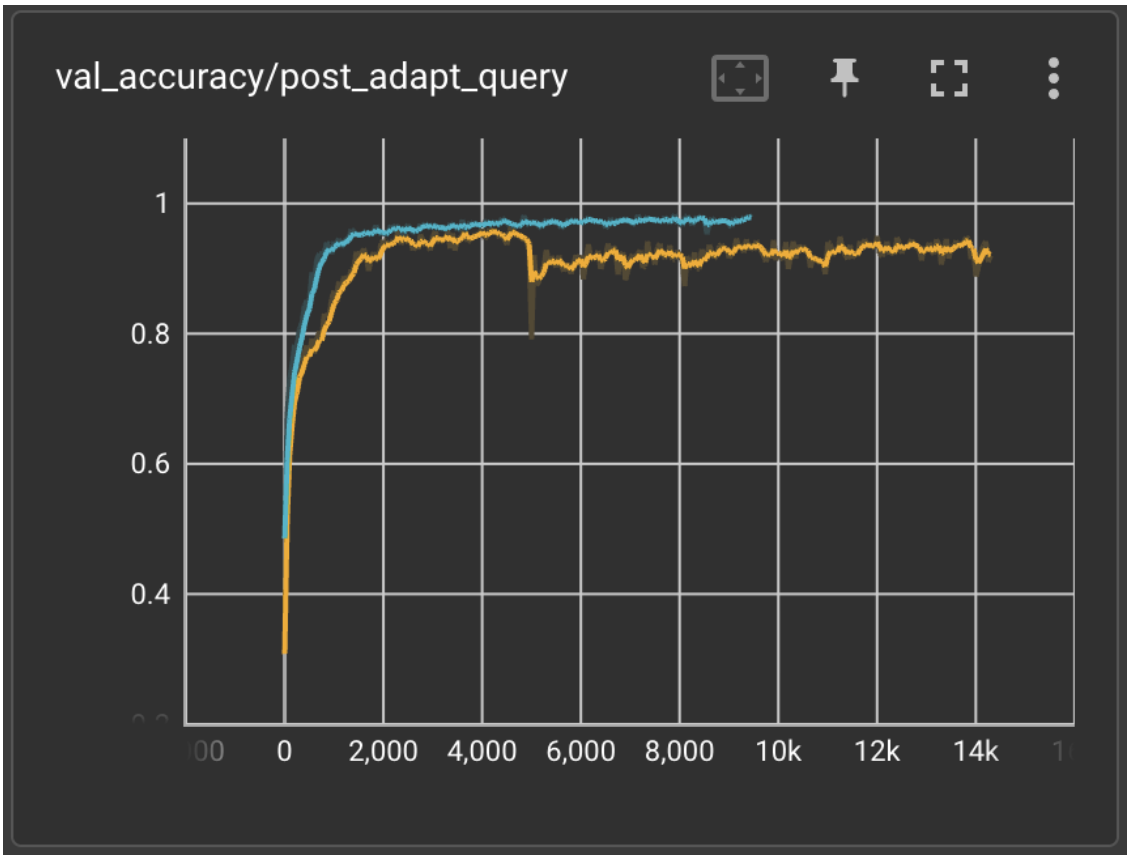

Figure 5: Validation post-adaptation query accuracy over the course of 5-way 1-shot Omniglot training on MAML, with inner learning rate of 0.4 (blue) and 0.04 (yellow).

Submit a response to the following question: Why would these different values affect training?

Answer:
A lower learning rate results in slower convergence.

2.5

Try MAML with learning the inner learning rates. Initialize the inner learning rates with 0.4. **Submit** a plot of the validation post-adaptation query accuracy over the course of training for learning and not learning the inner learning rates, initialized at 0.4.

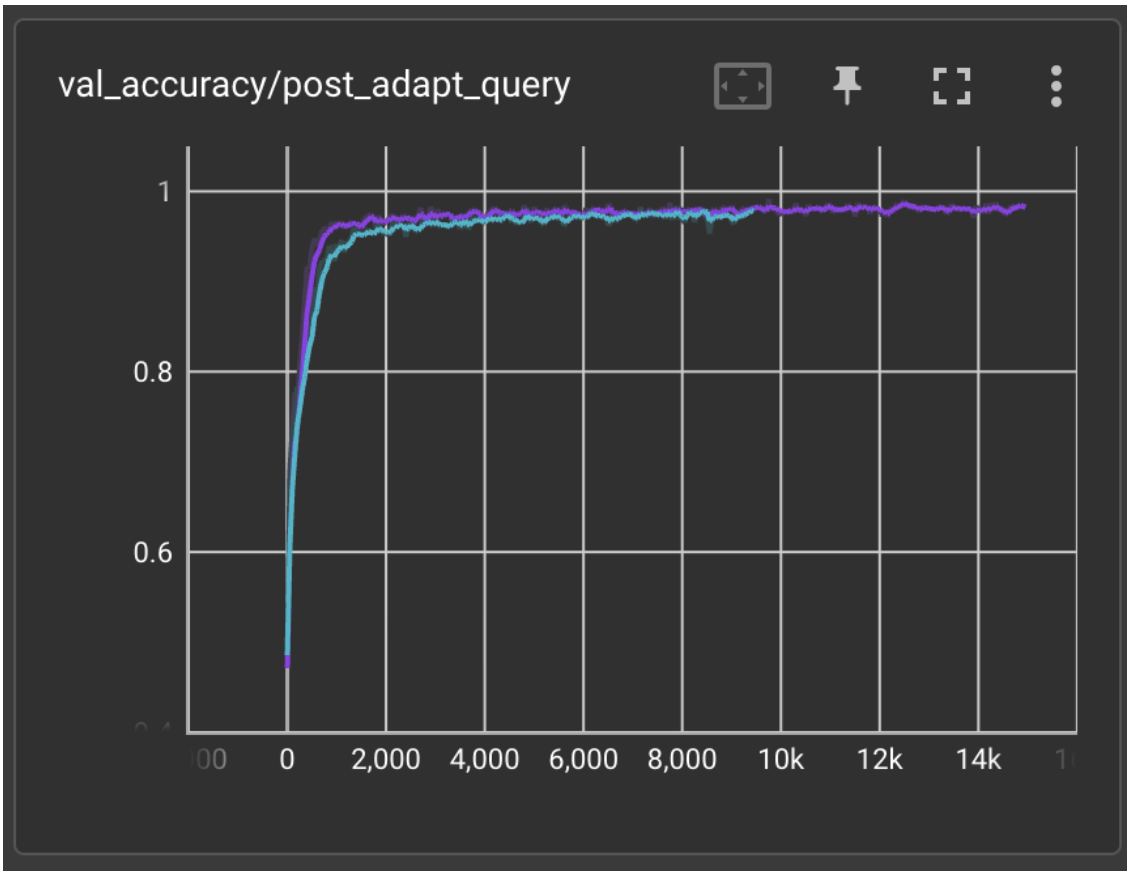

Figure 6: Validation post-adaptation query accuracy over the course of 5-way 1-shot Omniglot training on MAML, with fixed inner learning rate (blue) and learned inner learning rate (purple), both initialized at 0.4.

Submit a response to the following question: What is the effect of learning the inner learning rates?

Answer:
Allowing the inner learning rates to be learned allows the model to pick up on ideal learning rates for each layer.

3 More Support Data at Test Time

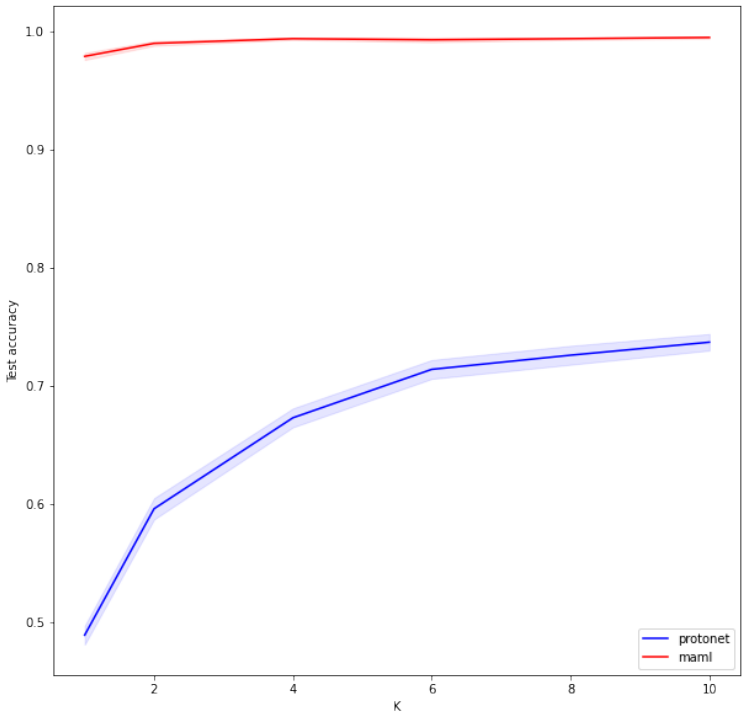

Figure 7: Test accuracies with 95% confidence interval represented as shaded regions, for $K \in \{1, 2, 4, 6, 8, 10\}$. Protonet plot is for 5-way 1-shot Omniglot, and MAML plot is for 5-way 1-shot with learned inner learning rates initialized at 0.4

How well is each model able to use additional data in a task without being explicitly trained to do so?

Answer:
The test accuracy for the protonet model trained on 1-shot increases with K, indicating that it is able to leverage more samples effectively. On the other hand, since MAML is already able to achieve close to perfect accuracy with $K = 2$, it is hard to determine whether MAML was able to use the additional samples effectively since the increase in test accuracy as K increases is marginal.