First-Order Approximations for Efficient Meta-Learning

B. King, A. Black, A. J. McLeay

Introduction

Following on from the promising performance of the Model Agnostic Meta Learning (MAML) algorithm [1], Reptile was written as a strong approximator for MAML that retains its performance whilst boasting much improved algorithmic complexity. This project seeks to replicate the experiments of the Reptile paper [2] and extend them into more complex contexts.

We can mathematically justify Reptile (and, as it happens, MAML) by thinking in terms of inner products of gradients after k inner-loop update steps [2]:

\[ \mathbf{g}_{k+1}^{\text{Reptile}} = \frac{1}{n} \left( \mathbf{g}_k - \theta \right) - \sum_{i=1}^{k} (\nabla L_i^{\left( \theta \right)}) + \left( \nabla L_i^{\left( \theta \right)} \right) \]

\[ \approx \sum_{i=1}^{k} \left( \nabla L_i^{\left( \theta \right)} \right) - \frac{n}{k} \left( \nabla L_i^{\left( \theta \right)} \right) \]

\[ \mathbb{E}_i [\mathbf{g}_{k+1}^{\text{Reptile}}] \approx \mathbb{E}_i [\nabla L_i^{\left( \theta \right)}] - \frac{n}{k} \left( \nabla L_i^{\left( \theta \right)} \right) \]

Inner-Loop Optimiser Comparisons

Figure 2: Using different optimisers for the rapid-learning updates on the sinusoid regression experiment from the original paper [1]. This concerns how \(g_0\) is found in [1]. Whilst ADAM improved learning, like [1], including momentum made learning more unstable and Reptile performed best when the momentum coefficient was near zero. Lastly, overfitting on the test task can be seen after about 20 ADAM inner-loop iterations when trained for many meta iterations.

Extensions to Few-Shot Regression

Ground truth

After 0 SGD steps (initialisation)

After 256 inner iter

Before training

1000 meta iter

2000 meta iter

Figure 3: Results from a 2D version of the 1D few-shot regression sine wave example in the Reptile paper. The input space is scaled from 50 to 10K points, and the minibatch size is reduced from 10 to 20. The amplitude is fixed at 2.0 and the phase is varied from 0 to 2\(\pi\). An SGD optimiser is used. After 1000 meta iterations on initialisation is learned that enables the ‘bowl’ shape of the function to be modelled. After 2000 epochs the learned initialisation enables the ‘tails’ of the function to also be modelled.

References


Conclusions

1. On 1D sinusoid regression, an ADAM inner-loop optimiser achieves lowest MSE but with less stability due to momentum.
2. Reptile learns few-shot approximations to 2D sine waves as well as mixtures of 1D function families (Fourier series and sine waves).
3. Reptile performs comparably to MAML [1] on few-shot image classification. Pretraining elevates Reptile performance during early training iterations, but does not produce a sustained advantage. However we did not extensively tune pretraining hyperparameters and further research could develop a more rigorous pretraining protocol.