Learning to Learn via Self-Critique

Antreas Antoniou
University of Edinburgh
{a.antoniou}@sms.ed.ac.uk

Amos Storkey
University of Edinburgh
{a.storkey}@ed.ac.uk

1 Supplementary Material: High-End Backbone details

The motivations behind each of the design choices can be found below.

1. Using DenseNet as the backbone, which decreases probability of gradient degradation problems and by allowing feature-reuse across all blocksm improves parameter/training efficiency. MAML is highly vulnerable to gradient degradation issues, and thus ensuring that our backbone decreases probability of such issues is of vital importance.
2. Using a shallow, yet wide backbone: Previous works Qiao et al. (2018); Rusu et al. (2018) have demonstrated that using features from the 20th layer of a pretrained ResNet produces superior generalization performance. The authors made the case that using features from shallower parts of the network decreases the probability that the features are too class-specific, and thus allow for better generalization on previously unseen classes. In both Qiao et al. (2018); Rusu et al. (2018) the authors did not train their meta-learning system end-to-end, and instead trained the feature backbone and the meta-learning components separately. However, in preliminary experiments we found that ResNet and DenseNet backbones tend to overfit very heavily, and in pursuit of a high-generalization end-to-end trainable meta-learning system, we experimented with explicit reduction of the effective input region of the layers in a backbone. Doing so, ensures that features learned will be local. We found that keeping the effective input region of the deepest layer to approximately 15x15/20x20 produced the best results for both Mini-ImageNet and CUB. Furthermore, we found that widening the network produced additional generalization improvements. We theorize that this is because of a higher probability for a randomly initialized feature to lie in just the right subspace to produce a highly generalizable feature once optimized.
3. Using bottleneck blocks, preceded by squeeze-excite-styleHu et al. (2018) convolutional attention: We empirically found that this improves generalization performance.
4. Inner-Loop optimize only the last squeeze excite linear layer, as well as the last convolutional layer and the final linear layer, whilst sharing the rest of the backbone across the inner loop steps. This design choice was hinted by the learned per-layer, per-step learning rates learned by MAML++ on the low-end baseline. More specifically, the learned learning rates where close to zero, for all layers, in all steps, except the last convolutional and last linear layers. Thus, we attempted to train a MAML++ instance where only those two layers where optimized in the inner loop, while the rest of the layers where shared across steps. In doing so, we found that doing so makes no difference to generalization performance, whilst increasing the training and inference speeds by at least 15 times.

References

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Qiao, S., Liu, C., Shen, W., and Yuille, A. L. (2018). Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. (2018). Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*.