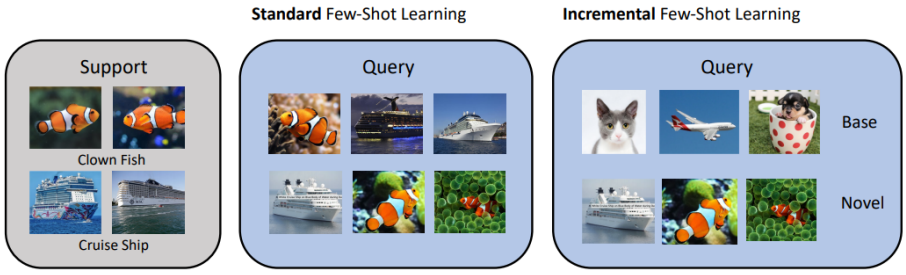
Incremental Few-Shot Learning with Attention Attractor Networks

Chen Zhao

NeurIPS 2019 (`https://arxiv.org/pdf/1810.07218.pdf`)

The availability of large scale datasets with detailed annotation, such as ImageNet, played a significant role in the recent success of deep learning. The need for such a large dataset is however a limitation, since its collection requires intensive human labor. This is also strikingly different from human learning, where new concepts can be learned from very few examples. One line of work that attempts to bridge this gap is few-shot learning, where a model learns to output a classifier given only a few labeled examples of the unseen classes. While this is a promising line of work, its practical usability is a concern, because few-shot models only focus on learning novel classes, ignoring the fact that many common classes are readily available in large datasets.



Incremental few-shot learning combines incremental learning where we want to add new classes without catastrophic forgetting, with few-shot learning when the new classes, unlike the base classes, only have a small amount of examples. This work presents a novel method for incremental few-shot learning where during meta-learning we optimize a regularizer that reduces catastrophic forgetting from the incremental few-shot learning.

In Figure 1

- **Pretraining stage**: Learn the base class weights W_α and the feature extractor CNN backbone.

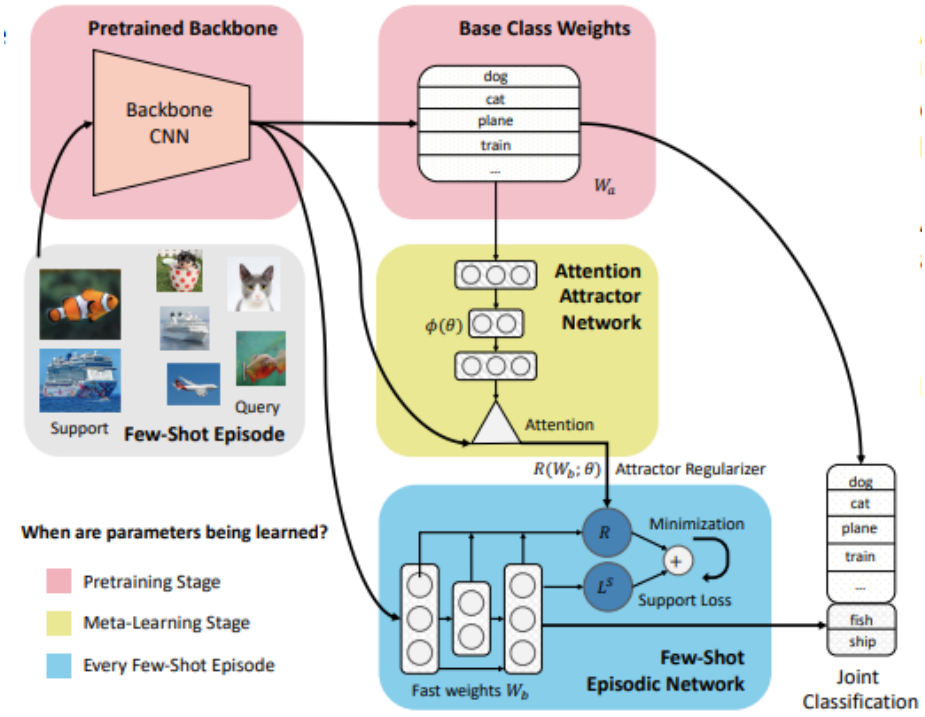


Figure 1: The proposed attention attractor network for incremental few-shot learning. During pretraining we learn the base class weights Wa and the feature extractor CNN backbone. In the meta-learning stage, a few-shot episode is presented. The support set only contains novel classes, whereas the query set contains both base and novel classes. We learn an episodic classifier network through an iterative solver, to minimize cross entropy plus an additional regularization term predicted by the attention attractor network by attending to the base classes. The attention attractor network is meta-learned to minimize the expected query loss. During testing an episodic classifier is learned in the same way.

- **Few-shot episode**: The support set only contains novel classes, whereas the query set contains both base and novel classes.
- **Meta-learning stage**: The attractor network is meta-learned to minimize the expected query loss. Gradient accumulated using the RBP algorithm.
- **Attention attractor network**: Use cosine similarity α of base classes to generate regularizer. Each base class has attractor $U_k = \phi(W_{\alpha,k};\theta)$. At-

tractor is then computed by attention $u_{k'} = \sum_k \alpha_{k',k} U_k + U_0$. Regularizer $R(W_b) = \sum_{k'} (W_{b,k'} - u_{k'})^T \gamma (W_{b,k'} - u_{k'})$.

- **Episodic stage**: An episodic classifier network learned through an iterative solver, to minimize cross entropy plus the attractor regularizer.

Joint Prediction of Old and New

- Learn a fast weight model on the novel classes, minimizing regularized cross-entropy objective

$$\begin{aligned} W_b^* &= \arg\min L^S(W_b) \\ &= \mathbb{E}_{x,y\sim S} y \log \hat{y}(W_b, x) + R(W_b) \end{aligned}$$

- A linear classifier or an MLP for the fast model
- Concatenate the logits with pretrained classification together

Recurrent Backprop

- Backprop through time (BPTT) expensive, need to store intermediate states of the inner loop.
- Truncated BPTT (T-BPTT) which optimizes the initial several iterations (e.g. MAML)
- Recurrent backprop (RBP) allows us to backpropagate through fixed-point iterations efficiently.
- The difference between two gradient descent steps ϕ can be written as

$$\phi(W_b^{(t)}) = W_b^{(t)} - F(W_b^{(t)}) = -\alpha \nabla L^S(W_b^{(t)})$$

  At optimum, we have $\frac{\partial W_b^*}{\partial \theta_E} = (I - J_{F,W_b^*}^T)^{-1} \frac{\partial F}{\partial \theta_E}$.

- Neumann series: $(I - J^T)^{-1} v = \sum_{n=0}^{\infty} (J^T)^n v$. Use standard backpropagation to compute $J^T v$.
- Same time complexity for backward pass
- No need to store intermediate states.

Summary

- Few-shot learning may not be very practical
- Incremental few-shot learning: novel classes on top of base classes, with a few examples, with only novel data.

- Training on novel classes without seeing old classes can be problematic.
- We propose to add learned regularizers conditioned on the episode.
- The proposed attention attractor produce regularizers by attending to base classes