Meta-Learning Representations for Continual Learning

Overview

- Artificial Neural Networks trained online on a correlated stream of data suffer from catastrophic interference.
- We propose learning a representation that is robust to catastrophic interference.
- To learn the representation, we propose MRCL, a second order meta-learning objective that directly minimizes interference and maximizes forward transfer.

Motivation

We hypothesize that in the observation space, gradient w.r.t one task interference with other tasks, and by learning a representation in which solution manifolds of different tasks are either parallel or orthogonal, we can mitigate catastrophic interference as shown in Figure 1. The Effect of the representation on continual learning, for a problem where targets are generated from three different distributions $p1(Y |x), p2(Y |x)$ and $p3(Y |x)$. The representation results in different solution manifolds for the three distributions; we depict two different possibilities here. We show the learning trajectory when training incrementally from data generates first by $p1$, then $p2$ and $p3$. On the left, the online updates interfere, jumping between distant points on the manifolds. On the right, the online updates either generalize appropriately—for parallel manifolds—or avoid interference because manifolds are orthogonal.

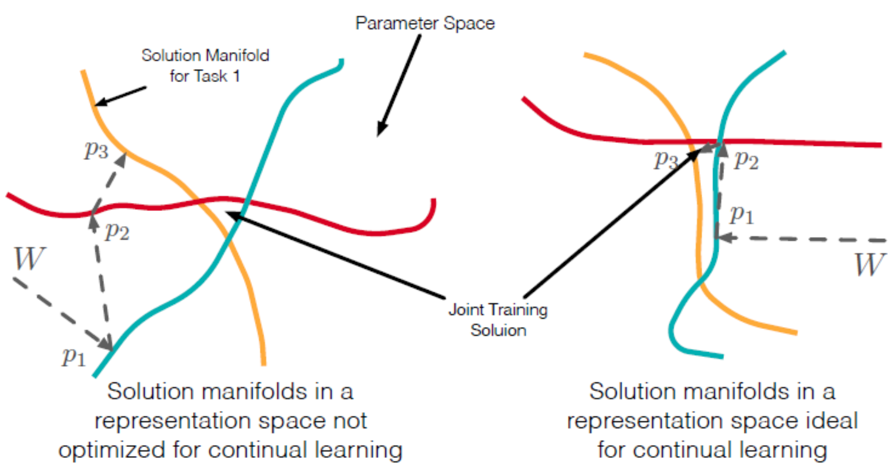

Figure 1 : We want to find a representation such that solution manifold of different tasks are parallel or orthogonal.

More concretely, we propose transforming the input into R, a large d dimensional vector, using a deep Representation Learning Network (RLN) such that it is possible to learn without interference from R. An example architecture of our method is shown in Figure 2. An example of our proposed architecture for learning representations for continual learning. During the inner gradient steps for computing the meta-objective, we only update the parameters in the prediction learning network (PLN). We then update both the representation learning network (RLN) and the prediction learning network (PLN) by taking a gradient step with respect to our meta-objective. The online updates for continual learning also only modify the PLN. Both RLN and PLN can be arbitrary models.

To train RLN, we treat the parameters in the RLN as meta-parameters which are updated to minimize interference when learning on correlated sequences of data using a Task Learning Network (TLN).

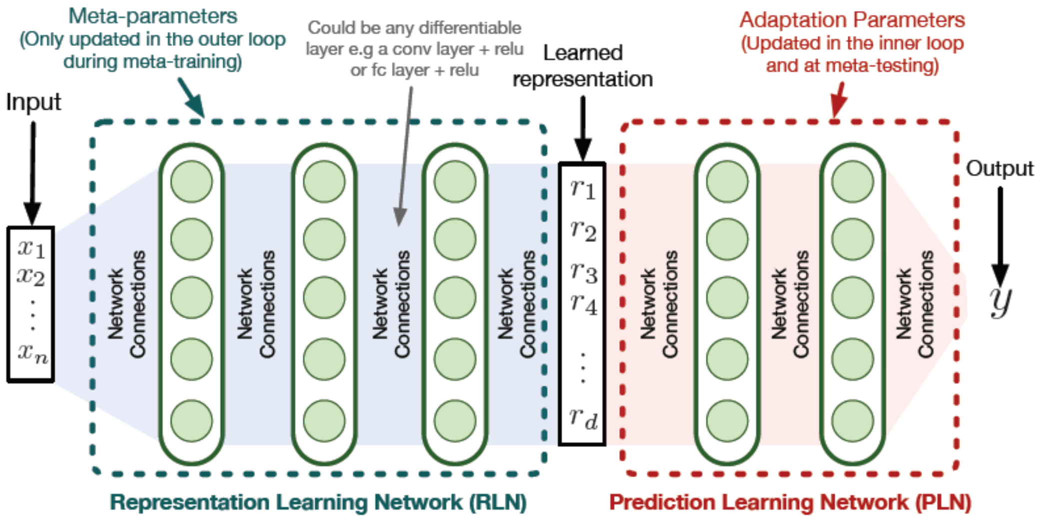

Figure 2 : One possible realization of an architecture used by MRCL. RLN learns to transform inputs into a space that allows learning without forgetting.

Proposed model

To apply neural network to the CLP problem, we propose meta-learning a function $\phi_\theta(X)$ – a deep Representation Learning Network (RLN) parametrized by θ – from $X \rightarrow R^d$. We then learn another function g_W from $R^d \rightarrow Y$, called a Prediction Learning Network (PLN). By composing the two functions we get $f_{W, \theta}(X)= g_W(\phi_\theta(X))$, which constitute our model for the CLP tasks as shown in Figure 1. We treat θ as meta-parameters that are learned by minimizing a meta-objective and then later fixed at meta-test time. After learning θ, we learn g_W from $R^d \rightarrow Y$ for a CLP problem from a single trajectory $\mathcal{S}$ using fully online SGD updates in a single pass. A similar idea has been proposed by Bengio et al. (2019) for learning causal structures.

$$\min_{W,\theta} \sum_{\mathcal{T}_i\sim p(\mathcal{T})} \mathrm{OML}(W,\theta) \overset{\text{def}}{=} \sum_{\mathcal{T}_i\sim p(\mathcal{T})} \sum_{\mathcal{S}_k^j\sim p(\mathcal{S}_k|\mathcal{T}_i)} \left[\mathcal{L}_{CLP_i}\left(U(W,\theta,\mathcal{S}_k^j)\right)\right] \quad (2)$$

where S_k^j represents an update function where W_{t+k} is the weight vector after k steps of stochastic gradient descent. The goal of the OML objective is to learn representations suitable for online continual learnings.