Appendices

A Further Experimental Details

We used the same hyperparameters as [1] and terminated training after the same number of examples specified in [1]. We used a temperature $c$ of 1, which was recommended in [2]. We initialized $T$ to be an identity matrix and all $\zeta$ to zero.

Compared to MAML [1], training a convolutional MT-net takes roughly 0.4 times longer (omniglot 40k steps took 7h 19m for MT-net and 5h 14m for MAML). This gap is fairly small because $1 \times 1$ convolutions require little compute compared to regular convolutions. This gap is larger (roughly 1.1 times) for fully connected MT-nets. We additionally observed that MT-nets take less training steps to converge compared to MAML.

We provide our official implementation of MT-nets at https://github.com/yoonholee/MT-net.

References


B Proofs for Section 4

B.1 MT-nets Learn a Subspace

Proposition 1. Fix $x$ and $A$. Let $U$ be a $d$-dimensional subspace of $\mathbb{R}^n$ ($d \leq n$). There exist configurations of $T$, $W$, and $\zeta$ such that the span of $y^{\text{new}} - y$ is $U$ while satisfying $A = TW$.

Proof. We show by construction that Proposition 1 is true.

Suppose that $\{v_1, v_2, \ldots, v_n\}$ is a basis of $\mathbb{R}^n$ such that $\{v_1, v_2, \ldots, v_d\}$ is a basis of $U$. Let $T$ be the $n \times n$ matrix $[v_1, v_2, \ldots, v_n]$. $T$ is invertible since it consists of linearly independent columns. Let $W = T^{-1}A$ and let $\zeta_1, \zeta_2, \ldots, \zeta_d \to \infty$ and $\zeta_{d+1}, \ldots, \zeta_n \to -\infty$. The resulting mask $M$ that $\zeta$ generates is a matrix with only ones in the first $d$ rows and zeroes elsewhere.

$$y^{\text{new}} - y = T(W^{\text{new}} - W)x$$
$$= T(M \odot \nabla_W \mathcal{L}_T)x$$

(1)

Since all but the first $d$ rows of $M$ are 0, $(M \odot \nabla_W \mathcal{L}_T)x$ is an $n$-dimensional vector in which nonzero elements can only appear in the first $d$ dimensions. Therefore, the vector $T(M \odot \nabla_W \mathcal{L}_T)x$ is a linear combination of $\{v_1, v_2, \ldots, v_d\}$. Thus the span of $y^{\text{new}} - y$ is $U$. \qed
B.2 MT-nets Learn a Metric in their Subspace

**Proposition 2.** Fix $x$, $A$, and a loss function $L_T$. Let $U$ be a $d$-dimensional subspace of $\mathbb{R}^n$, and $g(\cdot, \cdot)$ a metric tensor on $U$. There exist configurations of $T$, $W$, and $\zeta$ such that the vector $y^{\text{new}} - y$ is in the steepest direction of descent on $L_T$ with respect to the metric $du$.

**Proof.** We show Proposition 2 is true by construction as well.

We begin by constructing a representation for the arbitrary metric tensor $g(\cdot, \cdot)$. Let $\{v_1, v_2, \ldots, v_n\}$ be a basis of $\mathbb{R}^n$ such that $\{v_1, v_2, \ldots, v_d\}$ is a basis of $U$. Vectors $u_1, u_2 \in U$ can be expressed as $u_1 = \sum_{i=0}^d c_{1i}v_i$ and $u_2 = \sum_{i=0}^d c_{2i}v_i$. We can express any metric tensor $g(\cdot, \cdot)$ using such coefficients $c$:

$$g(u_1, u_2) = \begin{bmatrix} c_{11} & \ldots & c_{1d} \end{bmatrix} \begin{bmatrix} g_{11} & \ldots & g_{1d} \\ \vdots & \ddots & \vdots \\ g_{d1} & \ldots & g_{dd} \end{bmatrix} \begin{bmatrix} c_{21} \\ \vdots \\ c_{2d} \end{bmatrix},$$

where $G$ is a positive definite matrix. Because of this, there exists an invertible $d \times d$ matrix $H$ such that $G = H^\top H$. Note that $g(u_1, u_2) = (Hc_1)^\top(Hc_2)$: the metric $g(\cdot, \cdot)$ is equal to the inner product after multiplying $H$ to given vectors $c$.

Using $H$, we can alternatively parameterize vectors in $U$ as

$$u_1 = \begin{bmatrix} v_1 & \ldots & v_d \end{bmatrix} c_1 \quad \text{(3)}$$

$$= VH^{-1}(Hc_1). \quad \text{(4)}$$

Here, we are using $Hc_1$ as a $d$-dimensional parameterization and the columns of the $n \times d$ matrix $VH^{-1}$ as an alternative basis of $U$.

Let $v_1^H, \ldots, v_d^H$ be the columns of $VH^{-1}$, and set $T = [v_1^H, \ldots, v_d^H, v_{d+1}, \ldots, v_n]$. Since $H$ is invertible, $\{v_1^H, \ldots, v_d^H\}$ is a basis of $U$ and thus $T$ is an invertible matrix. As in Proposition 1, set $W = T^{-1}A$, $\zeta_1, \zeta_2, \ldots, \zeta_d \rightarrow \infty$, and $\zeta_{d+1}, \ldots, \zeta_n \rightarrow -\infty$. Note that this configuration of $\zeta$ generates a mask $M$ that projects gradients onto the first $d$ rows, which will later be multiplied by the vectors $\{v_1^H, \ldots, v_d^H\}$.

We can express $y$ as $y = Vc_y = VH^{-1}(Hc_y)$, where $c_y$ is again a $d$-dimensional vector. Note that $VH^{-1}$ is constant in the network and change in $W$ only affects $Hc_y$. Since $\nabla_W L_T = (\nabla_{Wx} L_T)x^\top$, the task-specific update is in the direction of steepest descent of $L_T$ in the space of $Hc_y$ (with the Euclidean metric). This is exactly the direction of steepest descent of $L_T$ in $U$ with respect to the metric $g(\cdot, \cdot)$.

$\square$
C Additional Experiments

Figure 1: Additional qualitative results from the polynomial regression task