Modular GEM

The new proposed rethinking of GEM consists of two main aspects: (1) the modular partitioning of the units of each of the network's layers, and (2) the discrepancy estimation of each task's representation projected in each group. The first aspect concerns the creation of the groups g_1^d, ..., $g_{K_d}^d$ for each layer $d \in \{2, \dots, L-1\}$, so the total number of groups is $K = \sum_2^{L-1} K_d$. And the second aspect leads to the computation of the discrepancy $(r_i^d)_k = D\left(P_t\left(y|G_i^d(x;\theta)\right) :: P_k\left(y|G_i^d(x;\theta)\right)\right)$ between task T_t and each previous task T_k $(k < t)$ given the group g_i^d.

GEM tries to solve the quadratic program problem:

$$min_v \ g^T G^T v + \frac{1}{2} v^T G G^T v \ \ s.t. \ \ v \geq 0 \qquad (1)$$

where $G = -(g_1, \dots, g_{t-1}) \in \mathbb{R}^{t-1 \times p}$ represents the gradients associated with previous tasks. $g \in \mathbb{R}^p$ stands for the gradient of the current task t. The goal of GEM is to find v^* to project the proposed gradient g to the closest gradient $\tilde{g} = G^T v^* + g$ (in squared ℓ_2 norm) satisfying the constraint.

Now, because of the modular partition, the previous gradients $G = -\left(g_1^1, \dots, g_1^{K_{L-1}}, \dots, g_{t-1}^1, \dots, g_{t-1}^{K_{L-1}}\right) \in \mathbb{R}^{(K \times (t-1)) \times p}$ has also been split into groups. When we only consider the group partition without discrepancy involved and want to reconstruct the overall previous gradients $G \in \mathbb{R}^{t-1 \times p}$ from the partitioned previous gradients $G \in \mathbb{R}^{(K \times (t-1)) \times p}$, we only need to sum up the individual group gradients which belong to the same task. But how can we do the reconstruction with discrepancy involved?

The measure of relatedness between task T_t and T_k $(k < t)$ can be inversely proportional to the symmetric discrepancy $(r_i^d)_k$. It means, if the current task T_t doesn't get updated, the more discrepancy is, the more catastrophic forgetting the corresponding groups will suffer. Hence, those less related groups should play a more important role for gradient reconstruction.

First, we have to merge the partitioned gradients $G \in \mathbb{R}^{(K \times (t-1)) \times p}$ into $G = -(G_1, \dots, G_{t-1})$, where $G_i = (g_i^1, \dots, g_i^{K_{L-1}}) \in \mathbb{R}^{K \times p}$. Second, the discrepancy $(r_i^d)_k$ should be merged to $r = (r_1, \dots, r_{t-1})$, where $r_i = (r_i^1, \dots, r_i^K) \in \mathbb{R}^K$. For reconstruction, the mean of discrepancy within the same task should also be moved to 1, so $\widetilde{r_i} = \left(\widetilde{r_i^1}, \dots, \widetilde{r_i^K}\right)$. Modular GEM aims at:

$$min_v \ g^T \tilde{G}^T v + \frac{1}{2} v^T \tilde{G} \tilde{G}^T v \ \ s.t. \ \ v \geq 0 \qquad (2)$$

where the reconstructed overall previous gradients $\tilde{G}_i = \widetilde{r_i} G_i$ and $\tilde{G} = \left(\tilde{G}_1, \dots, \tilde{G}_{t-1}\right) \in \mathbb{R}^{t-1 \times p}$.