Assignment 4: Extension

Advanced Topics In Machine Learning (CS6360)

Rahul Vigneswaran

CS23MTECH02002

Paper presented as part of assignment 1

Neural Collapse Terminus: A Unified Solution for Class Incremental Learning and Its Variants

Thanks to Deepika, Sayanta and Arvind for the heated discussions!

We consider the following problem,

$$\min_{\mathbf{M}^{(t)}} \quad \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right),\ 0 \leq t \leq T,$$

$$s.t. \quad \|\mathbf{m}_{k,i}^{(t)}\|^2 \leq 1, \quad \forall 1 \leq k \leq K^{(t)},\ 1 \leq i \leq n_k,$$

where,
$\mathbf{m}_{k,i}^{(t)} \in \mathbb{R}^d$: i-th sample of class k in session t feature
n_k : no. of samples in class k
$K^{(t)}$: no. of classes in session t
$N^{(t)} = \sum_{k=1}^{K^{(t)}} n_k$
$\mathbf{M}^{(t)} \in \mathbb{R}^{d \times N^{(t)}}$: collection of $\mathbf{m}_{k,i}^{(t)}$
$K = \sum_{t=0}^{T} K^{(t)}$
$\hat{\mathbf{W}}_{\mathrm{ETF}} \in \mathbb{R}^{d \times K}$: neural collapse terminus all K

A simplex equiangular tight frame (ETF) refers to a collection of vectors $\{\mathbf{e}_i\}_{i=1}^{K}$ in $\mathbb{R}^d$, $d \geq K-1$, that satisfies:

$$\mathbf{e}_{k_1}^T \mathbf{e}_{k_2} = \frac{K}{K-1}\delta_{k_1,k_2} - \frac{1}{K-1}, \qquad \text{DEF1}$$

$$\forall k_1, k_2 \in [1, K],$$

where $\delta_{k_1,k_2} = 1$ when $k_1 = k_2$, and 0 otherwise. All vectors have the same ℓ_2 norm and any pair of two different vectors has the same inner product of $-\frac{1}{K-1}$, which is the minimum possible cosine similarity for K equiangular vectors in $\mathbb{R}^d$.

Theorem

Let $\hat{\mathbf{M}}^{(t)}$ denotes the global minimizer by optimizing the model incrementally from $t = 0$, and we have $\hat{\mathbf{M}} = [\hat{\mathbf{M}}^{(0)}, \cdots, \hat{\mathbf{M}}^{(T)}] \in \mathbb{R}^{d \times \sum_{t=0}^{T} N^{(t)}}$. No matter if $\mathcal{L}$ is CE or misalignment loss, for any column vector $\hat{\mathbf{m}}_{k,i}$ in $\hat{\mathbf{M}}$ whose class label is k, we have:

$$\|\hat{\mathbf{m}}_{k,i}\| = 1, \ \ \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_{k'} = \frac{K}{K-1}\delta_{k,k'} - \frac{1}{K-1},$$

for all $k, k' \in [1, K]$, $1 \leq i \leq n_k$, where $K = \sum_{t=0}^{T} K^{(t)}$ denotes the total number of classes of the whole label space, $\delta_{k,k'} = 1$ when $k = k'$ and 0 otherwise, and $\hat{\mathbf{w}}_{k'}$ is the class prototype in $\hat{\mathbf{W}}_{\mathrm{ETF}}$ for class k'.

Extension 1: Relaxing implicit weight assumption

$$\min_{\mathbf{M}^{(t)}} \quad \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right),\ 0 \leq t \leq T,$$
$$s.t. \quad \|\mathbf{m}_{k,i}^{(t)}\|^2 \leq 1, \quad \forall 1 \leq k \leq K^{(t)},\ 1 \leq i \leq n_k,$$

$$\min_{\mathbf{M}^{(t)}} \quad \sum_{k=1}^{K^{(t)}} \frac{1}{N_k^{(t)}} \sum_{i=1}^{n_k} \mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right),\ 0 \leq t \leq T,$$
$$s.t. \quad \|\mathbf{m}_{k,i}^{(t)}\|^2 \leq 1, \quad \forall 1 \leq k \leq K^{(t)},\ 1 \leq i \leq n_k,$$

Extension 1

1. Take $\alpha_k = \frac{1}{N_k}$.
2. $\lambda = 0$ doesn't end in contradiction like earlier. Might have to add another clause in the theorem for it. But how λ can be interpreted?
3. The final theorem would be dependent on α_k

Extension 2: Adding $t-1$ dependency

$$\min_{\mathbf{M}^{(t)}} \quad \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right), \; 0 \le t \le T,$$
$$s.t. \quad \|\mathbf{m}_{k,i}^{(t)}\|^2 \le 1, \quad \forall 1 \le k \le K^{(t)}, \; 1 \le i \le n_k,$$

$$\min_{\mathbf{M}^{(t)}} \quad \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \mathcal{L}_{\mathrm{CE}}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right) + L_{\mathrm{Distill}}\left(\mathbf{m}_{k,i}^{(t)}, \mathbf{m}_{k,i}^{(t-1)}\right), \; 0 \le t \le T,$$
$$s.t. \quad \|\mathbf{m}_{k,i}^{(t)}\|^2 \le 1, \quad \forall 1 \le k \le K^{(t)}, \; 1 \le i \le n_k,$$

Extension 3: Mixture of experts

1. Towards Understanding the Mixture-of-Experts Layer in Deep Learning (NeurIPS 22) [Thanks to Piyushi!]
    1. Provides proof of why experts don't converge to the same function.
2. We could prove that the experts indeed converge to independent desired functions based on L_1 and L_2 losses. L_1 and L_2 focuses on a subset of classes for each expert.
3. Should think through it more.