IIT Hyderabad **Rahul Vigneswaran K, CS23MTECH02002**
Advanced Topics in Machine Learning (CS6360) **Assignment #5**
Instructor: Dr. Vineeth N Balasubramanian **March 27, 2024**

Primary Paper. *Neural Collapse Terminus: A Unified Solution for Class Incremental Learning and Its Variants. Extension of ICLR 23 work and a possible TPAMI submission.*

Secondary Paper. *Learning Imbalanced Datasets with Label-Distribution-Aware Margin. Published in NeurIPS 2019.*

1 Recap of the proposed extension in Assignment 4

The primary paper solves the following problem to prove their theorem,

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

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

where,
$\mathbf{m}_{k,i}^{(t)} \in \mathbb{R}^d$: Feature variable that belongs to i-th sample of class k in session t
n_k : Number of samples in class k
$K^{(t)}$: Number of classes in session t
$N^{(t)}$: Number of samples in session t
$\mathbf{M}^{(t)} \in \mathbb{R}^{d \times N^{(t)}}$: Collection of $\mathbf{m}_{k,i}^{(t)}$
$\hat{\mathbf{W}}_{\mathrm{ETF}} \in \mathbb{R}^{d \times K}$: Neural collapse terminus for the whole label space

But they make an implicit equal weight assumption in their problem definition above. So I wanted to relax that assumption and re-define the problem as follows,

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

$$\alpha_k^{(t)} = \frac{1}{N_k^{(t)}}$$

In the above-relaxed problem, the weight of each class is defined as the inverse of the number of samples in that respective class.

2 Road block with the proposed extension

Since the $\alpha_k^{(t)}$ is a function of k, while following the flow of the proof, there comes a point where it becomes difficult to further proceed with the derivation. It would make things simpler if we somehow take α_k out of the outermost summation or make it in such a way that we can separate the summation of α_k and the summation of the rest of the terms dependent on k. Currently, the form is something like below,

$$\sum_{k=1}^{K^{(t)}} \alpha_k^{(t)} A(k)$$

and having something like the below would make things simpler,

$$\sum_{k=1}^{K^{(t)}} \left(\alpha_k^{(t)} + A(k)\right) = \sum_{k=1}^{K^{(t)}} \alpha_k^{(t)} + \sum_{k=1}^{K^{(t)}} A(k)$$

3 Relevant Contributions from the secondary Paper

1. It proposes a novel Label-Distribution-Aware Margin (LDAM) loss function that encourages larger margins for minority classes during training. This is motivated by theoretical generalisation bounds that suggest minority classes should have larger margins to improve their generalisation performance.

2. It derives an optimal trade-off between the per-class margins for the binary classification case, showing that the margin for each class should scale as $n^{(-1/4)}$ where n is the number of examples in that class. The LDAM loss approximates this optimal margin scaling for multi-class problems

4 How the secondary paper can help with the current roadblock?

The proposed extension relaxes the implicit equal weighting assumption but in doing so we introduce a k dependant multiplier which yields an undesirable form.

The secondary paper can help us look at the same assumption and relaxation from a different angle. We can look at it as an implicit equal margin assumption. So we can relax it by adding a class-specific margin. In doing so, instead of a multiplication term, we would have k dependant addition term. This could (I am not sure yet, I am still in the process of deriving it) potentially provide us with a much more desirable form like the one below,

$$\sum_{k=1}^{K^{(t)}} \left(\alpha_k^{(t)} + A(k)\right) = \sum_{k=1}^{K^{(t)}} \alpha_k^{(t)} + \sum_{k=1}^{K^{(t)}} A(k)$$

5 Conclusion

By leveraging the margin idea from the secondary paper, we can potentially overcome the earlier mentioned roadblock and simplify the derivation process while still catering to the generalised case of imbalanced classes.