Matryoshka Neural Collapse

1 Motivation

We aim to leverage the **Neural Collapse** (NC) phenomenon—particularly the NC1 property (within-class variability collapse)—to facilitate knowledge distillation. Our teacher model is **BGE-M3**, which empirically exhibits strong NC1 behavior. The student is a $\mathbf{BERT_{base}}$ model, whose hidden dimensionality is 768.

We construct a set of nested sub-representations (Matryoshka subnets) from BERT by truncating the final hidden layer to dimensions 768, 512, 256, and 128. Each of these subnets should learn to match the teacher's representation. Our key motivation is that, due to neural collapse, the teacher's class representations are centralized and simple. Thus, each student subnet, even the smaller ones, only needs to learn this simplified structure rather than modeling the full complexity of deep representations.

2 Methodology

Let:

- $\boldsymbol{Z}_T \in \mathbb{R}^{n \times D}$ be the teacher embeddings for a batch of n samples in D-dim space.
- $\boldsymbol{Z}_S^{(d_1)} \in \mathbb{R}^{n \times d_1}$ be the output from a student subnet with dimension $d_1 \in \{768, 512, 256, 128\}$.

For each student subnet, we introduce a **semi-orthogonal projection matrix** $\boldsymbol{P}^{(d_1)} \in \mathbb{R}^{d_1 \times D}$, constrained by:

$$\boldsymbol{P}^{(d_1)\top} \boldsymbol{P}^{(d_1)} = \boldsymbol{I}_{d_1},$$

but not necessarily $\boldsymbol{P}^{(d_1)} \boldsymbol{P}^{(d_1)\top} = \boldsymbol{I}_D$. This makes $\boldsymbol{P}^{(d_1)}$ a *semi-orthogonal matrix*.

Why Orthogonality? Using orthogonal (or semi-orthogonal) projection helps preserve the geometric structure of the student representation during transformation to the teacher space. In particular:

- Norms and angles are preserved, preventing distortion of feature relationships.
- It acts as a structured linear transformation without introducing bias or scale.
- It simplifies optimization and avoids overfitting to noisy alignment signals.

Projection and Alignment

The projected student embeddings in the teacher space are:

$$\hat{\boldsymbol{Z}}_S^{(d_1)} = \boldsymbol{Z}_S^{(d_1)} \boldsymbol{P}^{(d_1)}.$$

The alignment loss (e.g., MSE) is:

$$\mathcal{L}_{\text{align}}^{(d_1)} = \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{\boldsymbol{z}}_{S,i}^{(d_1)} - \boldsymbol{z}_{T,i} \right\|_2^2.$$

The total alignment loss over all subnet dimensions is:

$$\mathcal{L}_{\text{align}} = \sum_{d_1 \in \{768,512,256,128\}} \mathcal{L}_{\text{align}}^{(d_1)}.$$

Optional: Classification Loss

We can also add classification loss for each subnet:

$$\mathcal{L}_{\text{cls}}^{(d_1)} = \frac{1}{n} \sum_{i=1}^{n} \text{CE}(f_{\text{cls}}^{(d_1)}(\hat{\boldsymbol{z}}_{S,i}^{(d_1)}), y_i),$$

where $f_{\text{cls}}^{(d_1)}$ is the classifier for subnet d_1, and y_i is the true label.

The overall objective is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \sum_{d_1} \mathcal{L}_{\text{cls}}^{(d_1)},$$

where λ_1 and λ_2 control the balance between distillation and classification.

3 Conclusion

This method introduces a structured multi-scale distillation strategy leveraging the simplicity induced by Neural Collapse. Semi-orthogonal projections ensure faithful alignment from student subnets to the teacher's feature space. By training multiple Matryoshka subnets simultaneously, the model enables efficient and robust distillation at various compression levels.