PACEAttention: Principled and Adaptive Feature Compression-Expansion Grounded in the Geometry of MCR^2

Xiaojie Yu[1] Haibo Zhang[2,*] Jeremiah D. Deng[1,*] Lizhi Peng[3,4]
[1]University of Otago, [2]University of New South Wales, [3]Quancheng Laboratory, [4]University of Jinan
[1] Contact: haibo.zhang@unsw.edu.au, jeremiah.deng@otago.ac.nz

ICML International Conference On Machine Learning

WeChat

1. Background

The maximal coding rate reduction (MCR^2) objective provides a principled framework for learning low-dimensional subspace representations Z by promoting intra-class compactness (i.e., feature compression) and inter-class separation (i.e., feature expansion). White-box deep models can be derived by unrolling optimization steps of this objective into layer-wise architectures. However, they do not fully preserve the design principles implied by the MCR^2 gradient, as the approximation of the gradient may deviate from the intended compression objective.
Our goal aims to optimize the following MCR^2-like objective, where $U_{[K]}$ are K trainable subspace matrices.

$$\Delta R(Z, \mathbf{\Pi}) = R(Z) - R_c(Z|U_{[K]}) = \underbrace{\frac{1}{2}\log\det(I + \frac{d}{n\epsilon^2}ZZ^T)}_{\text{Expansion term}} - \underbrace{\frac{1}{2}\sum_{k=1}^{K}\log\det(I + \frac{p}{n\epsilon^2}(U_k^T Z)^T(U_k^T Z))}_{\text{Compression term}}$$

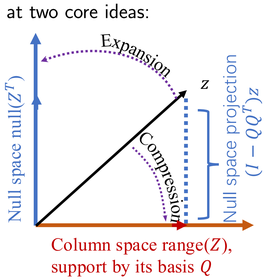
2. Key Contributions

- **Principled feature updates derived from the geometry of MCR^2 gradient.**
- **Using randomization to efficiently approximate subspace structures.**
- Linear complexity.
- Learnable α, β balance expansion and compression per layer.
- Heads learn fine-grained object-region memberships.

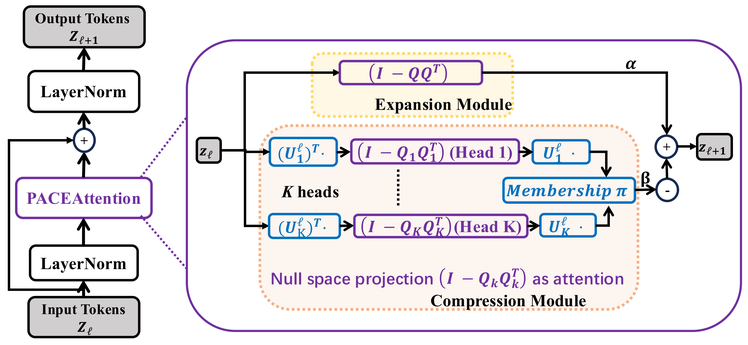
3. Geometric Intuition

We exploit the geometric interpretation of the gradient of $\log\det(I + ZZ^T)$, a term included in MCR^2-like objective. To increase or decrease the value of $\log\det(I+ZZ^T)$, we can enlarge or reduce the dimensionality of column space range(Z). Therefore, let Q be the orthonormal basis of range(Z). We arrive at two core ideas:

- For **feature expansion**, adding the null space projection $(\mathbf{I} - \mathbf{QQ}^T)\mathbf{z}$ moves a token $\mathbf{z}$ toward the null space, hence increasing the column space dimensionality and promoting separability.
- For **feature compression**, subtracting null space projection moves the token $\mathbf{z}$ toward the column space, hence decreasing the column space dimensionality.



4. PACEAttention Layer



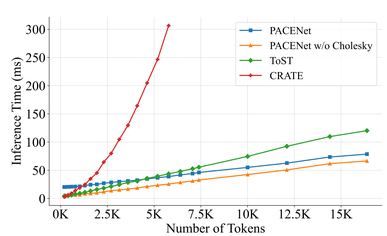
$$\mathbf{z}_{\ell+1} = \mathbf{z}_\ell + \underbrace{\alpha(\mathbf{I} - \mathbf{QQ}^T)\mathbf{z}_\ell}_{\text{Expansion}} - \underbrace{\beta\sum_{k=1}^{K}\pi_k\mathbf{U}_k(\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^T)\mathbf{U}_k^T\mathbf{z}_\ell}_{\text{Compression}}$$

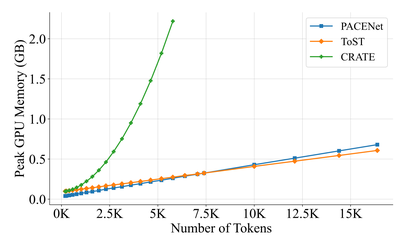
Expansion: Increase global coding rate $R(\mathbf{Z})$ by expanding toward null-space directions.
Compression: Reduce $R_c(\mathbf{Z} \mid \mathbf{U}_{[K]})$ by subtracting group-wise null-space components.
Self-adaptive dynamics: α controls expansion; β controls compression.

5. Efficient Subspace Approximation based on Randomization

$$\mathbf{Y} = \mathbf{Z}\Omega, \qquad \mathbf{Q} = \mathrm{orth}(\mathbf{Y})$$

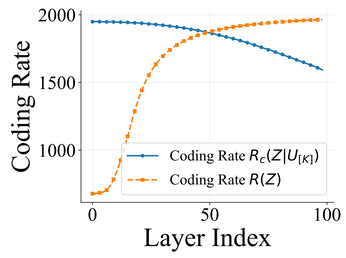
Random matrix Ω captures the column space.
Cholesky decomposition extract orthonormal basis efficiently.

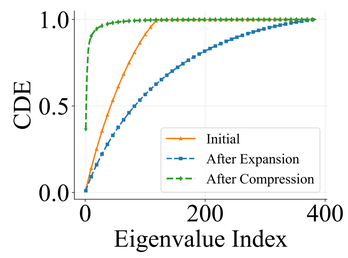




Scaling. Linear complexity in the number of tokens.

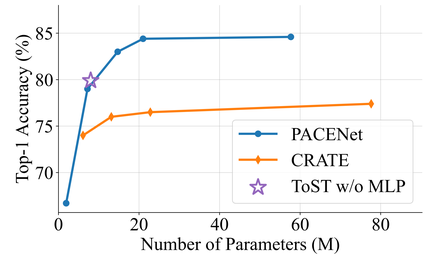
6. Verification on Toy Data



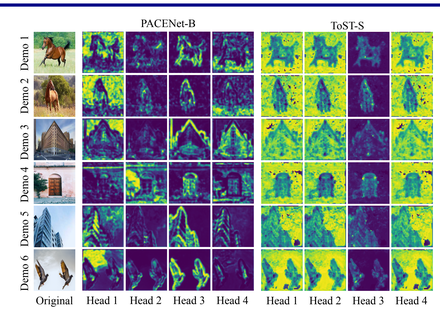


- Expansion increases $R(\mathbf{Z})$ layer by layer.
- Compression decreases $R_c(\mathbf{Z} \mid \mathbf{U}_{[K]})$.
- Eigen-spectrum confirms effective rank expansion/compression.

7. Top-1 on ImageNet-Real



8. Interpretability



Different heads attend to distinct object parts, enabling fine-grained structural modeling.