Selective Amnesia using Contrastive Subnet Erasure for Class Level Unlearning in Vision Models

Scientific Achievement

- Developed **Contrastive Subnet Erasure (CSE)**, a training-free method that selectively removes a target concept from pretrained vision models while preserving other learned knowledge.
- Identified compact neural subnets responsible for specific classes using contrastive eigen analysis and selectively attenuated them to achieve concept-level forgetting.
- Introduced a cross-dataset evaluation protocol to verify true semantic unlearning rather than removal of specific training examples.

Significance and Impact

- Enables machine unlearning in large pretrained models without retraining.
- Supports privacy-compliant AI by removing sensitive data or identities from models.
- Applicable to foundation models across science, industry, and security applications.

Technical Approach

- Identified target-concept channels using contrastive covariance and eigen analysis.
- Selected and attenuated a compact subnet while preserving non-target representations.
- Applied a training-free model edit folded into existing layers with no inference overhead.

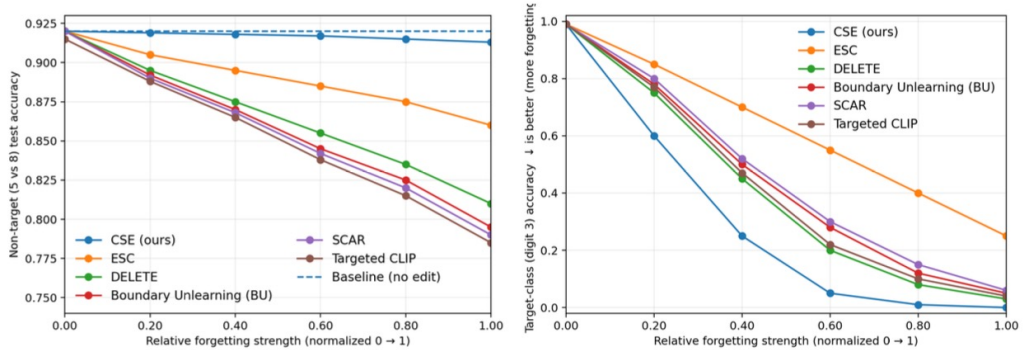

Figure 1. MNIST–EfficientNet toy with a shared, normalized forgetting strength s ∈[0, 1] (0 = no edit). CSE drives the target accuracy to near-zero while preserving non-target utility, whereas others require stronger edits or fail to fully forget within the same range.

PI(s)/Facility Lead(s): K. Kim (PI), O. Kotevska (ORNL's PI)
ASCR Program: AI for Science
ASCR PM: Xujing Davis
Publication(s) for this work: Pramanik, V., Maliha, M., Jha, S., Velasquez, A., Kotevska, O., & Jha, S. K. (2026). Selective Amnesia using Contrastive Subnet Erasure for Class Level Unlearning in Vision Models. In IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR).