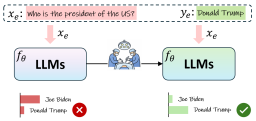
Energy-Regularized Sequential Model Editing on Hyperspheres

Qingyuan Liu[1]*, Jia-Chen Gu[2]*, Yunzhi Yao[3], Hong Wang[4], Nanyun Peng[2]
[1]Columbia [2]UCLA [3]ZJU [4]USTC

Paper Code

Background & Motivation

Model Editing: Localized parameter updates modify target knowledge while **avoiding interference** with unrelated knowledge



Hyperspherical Uniformity

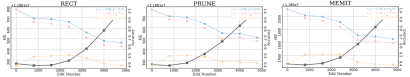
Hyperspherical Energy (HE) is used to quantify the uniformity of neuron weights on hyperspheres :

$$HE[X] = \sum_{i \neq j} \left(\|\hat{x}_i - \hat{x}_j\|^2 + \epsilon\right)^{-s} = \sum_{i \neq j} \left(2(1 - \cos\theta_{ij}) + \epsilon\right)^{-s}$$

Lower Energy

Higher Orthogonality

Correlation between Uniformity and Editing



Energy increases

Weights become less uniform

Statistically significant negative correlation

Maintaining uniformity is critical in stabilizing editing

SPHERE

Sparse **P**rojection for **H**yperspherical **E**nergy **R**egularized **E**diting

Project editing knowledge onto a **sparse** space which is **complementary to principal** hyperspherical directions

Enable new know incorporation

Mitigate ▲'s interfere with ●

● **original weight**
▲ **update weight**

Principal Space Estimation

$$\boldsymbol{v} = \arg\max_{\|\hat{\boldsymbol{v}}\|=1} \left(\frac{1}{n}\|\boldsymbol{W}\hat{\boldsymbol{v}}\|^2\right)$$

Maximize the variance of all neurons in W when projected onto v

Top-r: $\boldsymbol{U} = [\boldsymbol{v}_{d-r+1}, \ldots, \boldsymbol{v}_d] \in \mathbb{R}^{d \times r}$

Sparse Space Definition

$$\boldsymbol{P}_\perp = I - \alpha \boldsymbol{U}\boldsymbol{U}^\top \in \mathbb{R}^{d \times d}$$

Orthogonal

Suppression Strength: how much contribution of U is removed

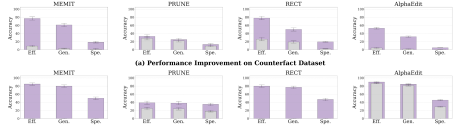
Sparse Space Projection

$$\hat{\boldsymbol{W}} = \boldsymbol{W} + \Delta \boldsymbol{W}\boldsymbol{P}_\perp$$

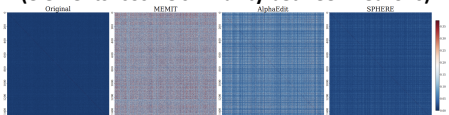
could be any editing method

Results & Analysis

Baseline editing methods can be improved with plug-and-play SPHERE by 43.27% (Eff.) , 36.20% (Gen.), and 20.96% (Spe.) on avg.



SPHERE's angular distribution remains stable (elements: cosine similarity between neurons)



SPHERE-edited models can better preserve the general abilities