B5-10 MLPの重みを反映したSparse Autoencoderの初期化手法の提案

菊谷 幹　北田 俊輔　原 聡

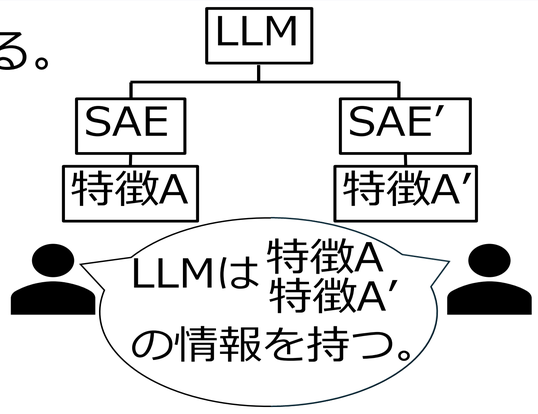
国立大学法人 電気通信大学 The University of Electro-Communications

背景と貢献

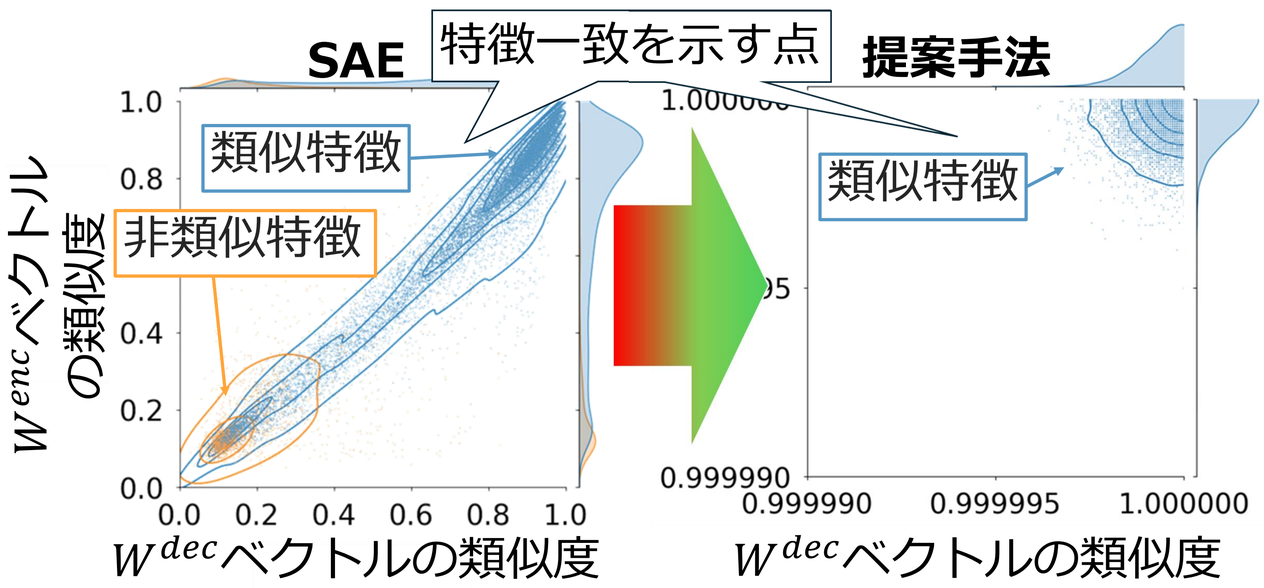
LLM内部の解釈にSAEが使われる。

初期値による特徴の不安定性が存在する[Paulo+, ICLR'26]。

そのため、**SAEを用いた解釈が不正確になる**懸念がある。

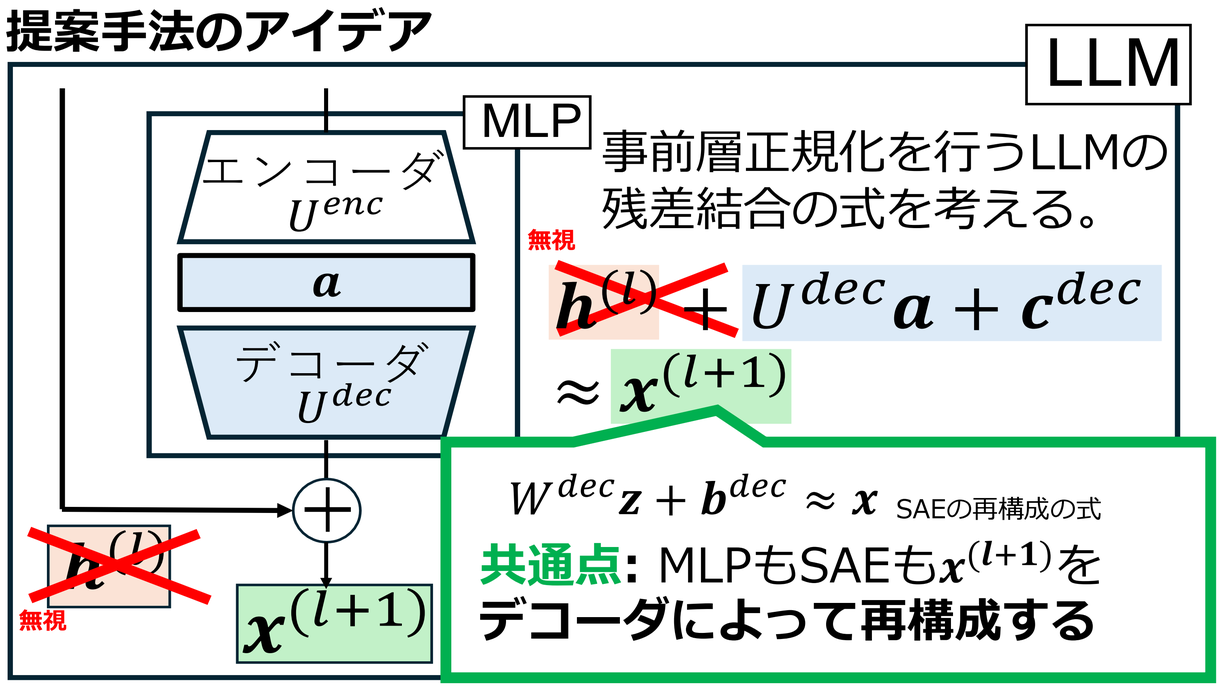


本研究のアイディア: 初期値として MLP のデコーダ
性能を維持しながら安定化。

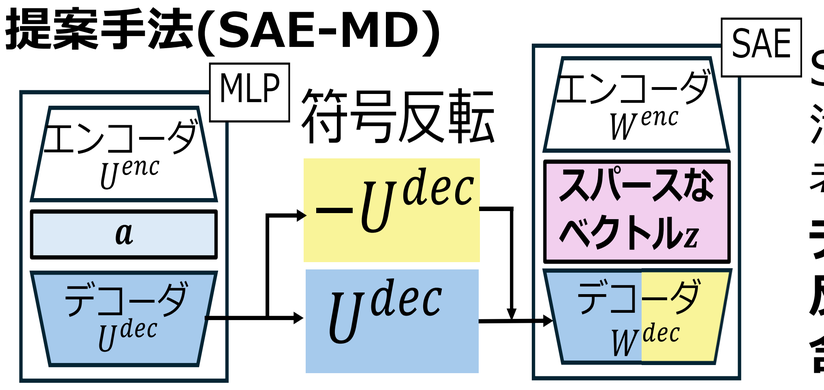


提案手法

提案手法のアイデア



提案手法(SAE-MD)



SAEにおける正の活性化値の制約を考慮し、**MLPのデコーダの符号を反転させたものも含めて利用する。**

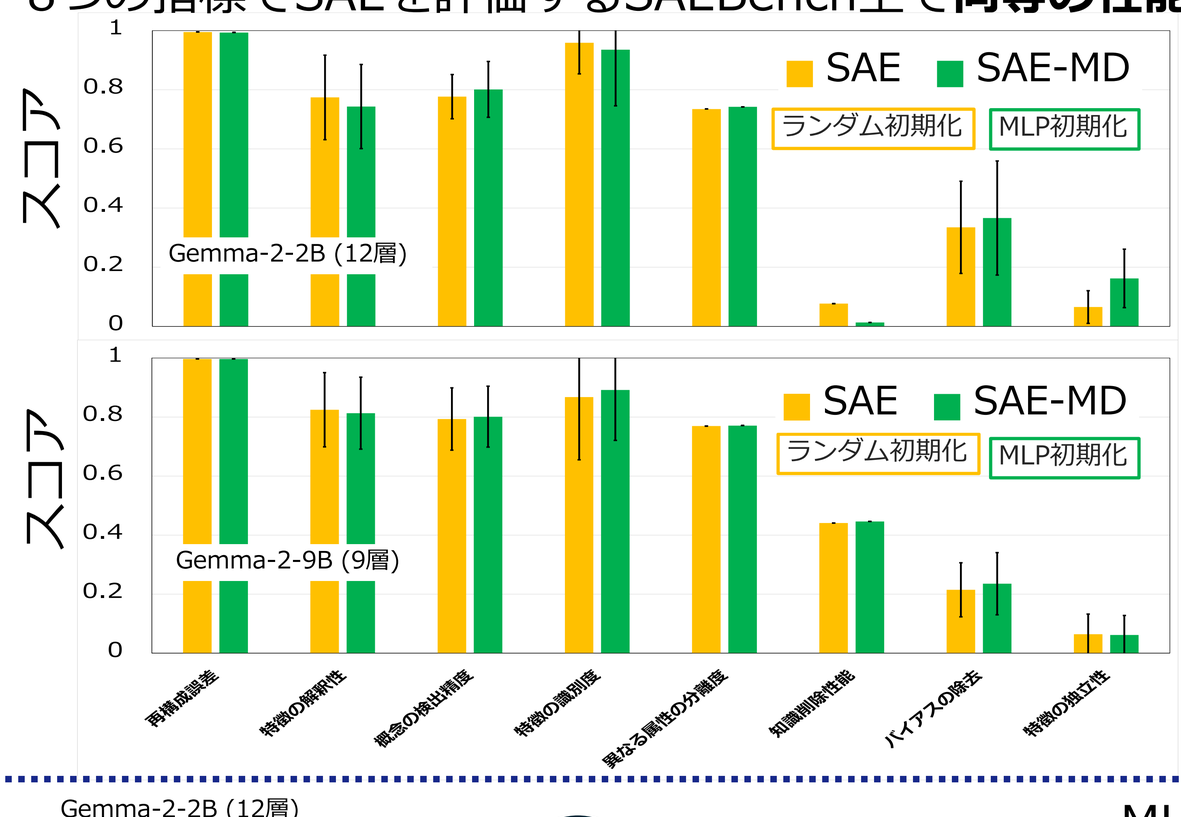
実験手法と結果

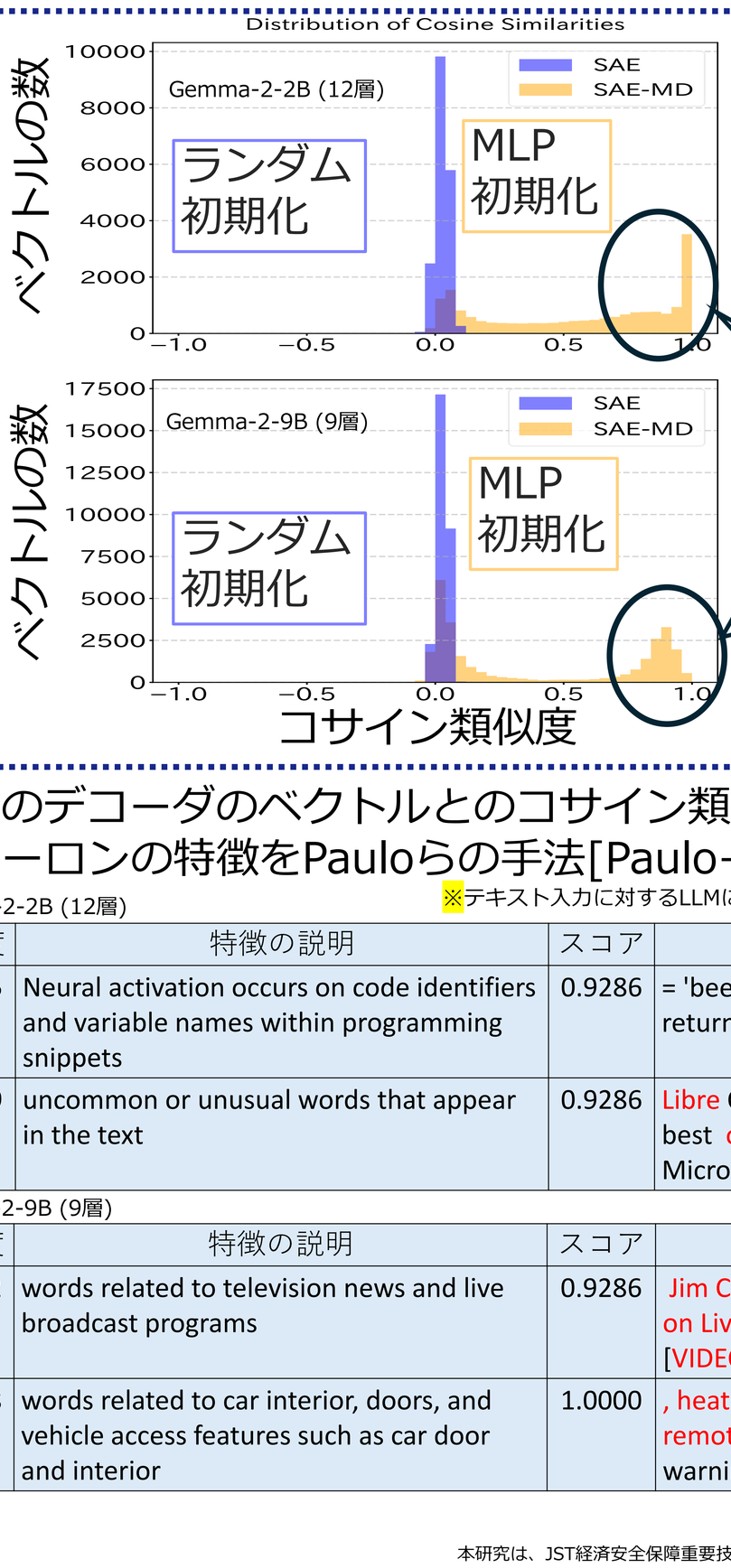
Gemma-2-2B, Gemma-2-9B [Gemma Team, 2024] → SAE, SAE-MD

1. SAE間のW^{enc}, W^{dec}同士を比較する手法[Paulo+, ICLR'26]を用いて特徴の共有度を調査。
2. SAEBench [Adam+, ICLR'25]による性能調査。

データセットはmonology/pile-uncopyrighted[1] (500M)
seed=24,42でSAEの中間層の非零要素数は 20, 40, 80, 160, 320になるように学習。

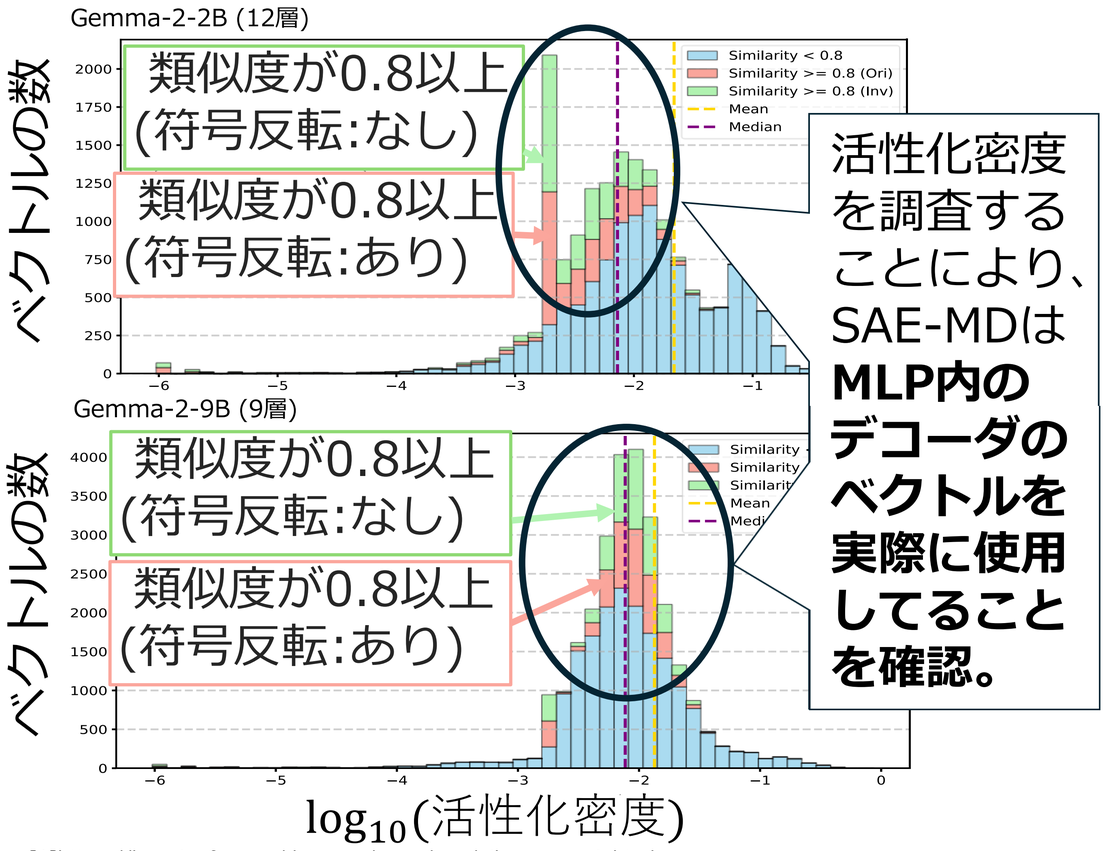
8つの指標でSAEを評価するSAEBench上で**同等の性能**。





SAE-MD, SAEのデコーダで初期値のベクトルと学習後のベクトルのコサイン類似度分布について調査した。

SAE-MDは**MLP内のデコーダのベクトルを保持しながら学習を行える。**



活性化密度を調査することにより、SAE-MDは**MLP内のデコーダのベクトルを実際に使用してることを確認。**

MLPのデコーダのベクトルとのコサイン類似度とスコア※が高いニューロンの特徴をPauloらの手法[Paulo+, ICML'25]で調査。

※テキスト入力に対するLLMによるSAE-MDのニューロン発火の予測正解率

Gemma-2-2B (12層)

類似度	特徴の説明	スコア	活性化テキスト
0.9895	Neural activation occurs on code identifiers and variable names within programming snippets	0.9286	= 'beep';↵↵ function opts (basedir) {↵ return {↵ basedir: path
0.9759	uncommon or unusual words that appear in the text	0.9286	Libre Office Draw module is one of the best open source alternatives to Microsoft Visio. With the help of

Gemma-2-9B (9層)

類似度	特徴の説明	スコア	活性化テキスト
0.9542	words related to television news and live broadcast programs	0.9286	Jim Cantore Defends Himself From Attack on Live TV with a Swift Knee to the Nuts [VIDEO]
0.9368	words related to car interior, doors, and vehicle access features such as car door and interior	1.0000	, heated rear seats and steering wheel, remote engine start and speed warnings.<eos>Richard Broun↵↵Richard

[1]https://huggingface.co/datasets/monology/pile-uncopyrighted

本研究は、JST経済安全保障重要技術育成プログラム JPMJKP24C3 の支援を受けたものです。