THE FEATURE COLLAPSE PROBLEM: A SYSTEMATIC INVESTIGATION OF SPARSE AUTOENCODER TRAINING DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Autoencoders (SAEs) have emerged as a promising approach for interpreting large language models by extracting interpretable features from their internal representations. However, a critical challenge in training SAEs is the tendency for features to collapse into redundant patterns, limiting their interpretability and utility. We present a systematic investigation of feature collapse through ten experimental configurations on the Gemma-2B model, progressively incorporating sophisticated mechanisms including hierarchical group structures, modified Gram-Schmidt orthogonalization, and momentum-based importance sampling. Despite implementing multiple stabilization techniques - from Johnson-Lindenstrauss projections to coordinated cosine annealing of contrast thresholds (0.4→0.05) and dynamic L1 penalty bounds - our results consistently show minimal feature differentiation across all variants, with absorption scores and feature split metrics remaining at baseline levels. This comprehensive investigation reveals fundamental limitations in current SAE architectures and suggests that preventing feature collapse may require more radical architectural innovations rather than incremental improvements to optimization dynamics.

1 INTRODUCTION

Understanding the internal representations of large language models (LLMs) has become increasingly critical as these models grow in complexity and capability Goodfellow et al. (2016). Sparse Autoencoders (SAEs) have emerged as a promising approach for extracting interpretable features from these models, offering the potential to decompose complex neural representations into meaningful components. However, our systematic investigation reveals fundamental challenges in training stable SAEs, particularly the persistent problem of feature collapse where multiple neurons encode redundant patterns.

The challenge of training effective SAEs stems from several factors:

- High-dimensional representation spaces in modern transformer architectures Vaswani et al. (2017)
- Complex interactions between attention mechanisms and learned features
- Difficulty in maintaining feature differentiation during training
- Lack of theoretical understanding of feature collapse dynamics

Our work makes the following contributions:

- A systematic evaluation of ten SAE configurations on the Gemma-2B model, revealing persistent feature collapse despite sophisticated stabilization techniques
- Novel hybrid architecture combining Johnson-Lindenstrauss projections, coordinated cosine annealing, and dynamic L1 penalties
- Comprehensive empirical analysis showing the limitations of current approaches to preventing feature collapse

- Open-source implementation and evaluation framework for reproducible research

We implement several technical innovations to address the feature collapse problem:

- 128-dimensional Johnson-Lindenstrauss projection matrix using QR decomposition
- Feature utilization tracking with exponential moving averages ($\beta = 0.9$)
- Hierarchical feature organization with 16 distinct groups
- Modified Gram-Schmidt orthogonalization during forward passes
- Momentum-based importance sampling with adaptive learning rates
- Dynamic L1 penalty bounds (0.02-1.0 $\rightarrow$ 0.01-0.5)

Our experimental results, visualized in Figures 1 and 2, demonstrate that despite these sophisticated mechanisms, feature collapse persists across all architectural variants. The mean absorption scores remain at 0.0 and split features at 1.0, suggesting fundamental limitations in current approaches to feature differentiation.

These findings have significant implications for neural network interpretability research. The persistent failure to achieve feature differentiation, even with explicit competition mechanisms and orthogonality constraints, suggests that preventing feature collapse may require more radical architectural innovations rather than incremental improvements to optimization dynamics. Future work should explore alternative approaches such as topological constraints, information-theoretic regularization, or entirely new architectures for feature extraction.

2 Related Work

Prior work on preventing feature collapse in neural networks can be broadly categorized into three approaches: architectural constraints, optimization techniques, and explicit regularization methods. Each approach offers insights into the challenge of maintaining differentiated features, though none have fully solved the problem in the context of sparse autoencoders for LLM interpretation.

Architectural Approaches: Early work by Bengio (2007) established theoretical foundations for hierarchical feature learning, suggesting that deep architectures naturally encourage feature differentiation through their layered structure. However, our experiments with hierarchical group structures (16 feature groups) show that architectural hierarchy alone is insufficient, as evidenced by persistent feature collapse (Figure 1). Similarly, while Zeiler & Fergus (2013) demonstrated success with visualization techniques in convolutional networks, their approach relies on spatial structure absent in transformer representations.

Optimization Strategies: Recent work has explored various optimization techniques to prevent feature collapse. Kingma & Ba (2014) and Loshchilov & Hutter (2017) introduced adaptive optimization methods that improve training stability, but our experiments show these methods fail to prevent feature collapse even when combined with sophisticated gradient balancing. Maass (2000)'s winner-take-all dynamics offered promising competition mechanisms, but our implementation reveals limitations in the transformer context, with absorption scores remaining at 0.0 despite explicit competition (Figure 2).

Regularization Methods: The most directly relevant approach comes from Mathieu et al. (2016), who used adversarial training to encourage feature disentanglement. While their method showed success in image domains, our attempts to adapt similar principles through contrastive learning (threshold annealing 0.4$\rightarrow$0.05) failed to prevent collapse in the high-dimensional space of transformer representations. This suggests that the challenge may lie deeper than regularization alone can address.

Our work synthesizes and extends these approaches, implementing a comprehensive suite of techniques including Johnson-Lindenstrauss projections, modified Gram-Schmidt orthogonalization, and momentum-based importance sampling. However, the persistent feature collapse we observe (Figure 3) suggests fundamental limitations in current approaches to feature differentiation in sparse autoencoders.

3 BACKGROUND

Sparse Autoencoders (SAEs) emerged from classical dictionary learning and sparse coding principles Rubinstein et al. (2010). In the context of neural networks, SAEs combine an encoder-decoder architecture with sparsity constraints to learn interpretable feature representations. The theoretical foundations for feature learning in deep architectures Bengio (2007) suggest that hierarchical organization naturally encourages feature differentiation, though achieving this in practice remains challenging.

Modern language models based on the transformer architecture Vaswani et al. (2017) present unique challenges for feature extraction due to their:

- High-dimensional activation spaces ($d_l \approx 2000 - 10000$)
- Complex attention-based information flow
- Layer-wise feature interactions
- Non-local token dependencies

Traditional approaches to neural network interpretation often focus on attention pattern analysis Bahdanau et al. (2014). However, these methods provide limited insight into the learned feature representations that emerge within transformer layers. SAEs offer a promising alternative by attempting to decompose these representations into interpretable components, though they face significant challenges in maintaining stable feature differentiation.

3.1 PROBLEM SETTING

Consider a pre-trained language model $\mathcal{M}$ with L layers. For any layer $l \in \{1, \ldots, L\}$, let $h_l \in \mathbb{R}^{d_l}$ represent the activation vectors. Our goal is to learn an interpretable sparse coding of these activations through an autoencoder defined by:

$$f = \sigma(W_{\text{enc}} h_l + b_{\text{enc}})$$
$$\hat{h}_l = W_{\text{dec}} f + b_{\text{dec}}$$

where:

- $\sigma(\cdot)$ is the ReLU activation function
- $W_{\text{enc}} \in \mathbb{R}^{d_{\text{sae}} \times d_l}$ is the encoder matrix
- $W_{\text{dec}} \in \mathbb{R}^{d_l \times d_{\text{sae}}}$ is the decoder matrix
- $b_{\text{enc}}, b_{\text{dec}}$ are bias terms
- d_{sae} is typically set equal to d_l

The training objective balances reconstruction fidelity with sparsity:

$$\mathcal{L}(h_l, \hat{h}_l, f) = \underbrace{\|h_l - \hat{h}_l\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda(t)\|f\|_1}_{\text{sparsity}} + \underbrace{\alpha(t)\mathcal{L}_{\text{contrast}}}_{\text{feature competition}} \tag{1}$$

where:

- $\lambda(t)$ implements dynamic L1 penalty bounds
- $\alpha(t)$ controls the strength of feature competition
- $\mathcal{L}_{\text{contrast}}$ encourages feature differentiation

Key assumptions in our approach include:

- Activation distributions are approximately stationary during training

- Features can be effectively represented through linear combinations
- The number of meaningful features does not exceed d_{sae}
- Feature collapse is primarily due to optimization dynamics rather than architectural limitations

This formulation builds on classical sparse coding while addressing the specific challenges of transformer architectures. The dynamic nature of $\lambda(t)$ and $\alpha(t)$ allows the model to adapt its optimization priorities throughout training, though our results in Section 6 show this remains insufficient to prevent feature collapse.

4 METHOD

Building on the formalism introduced in Section 3.1, we propose a series of mechanisms to address feature collapse in sparse autoencoders. Our approach combines dimensionality reduction, dynamic regularization, and competitive learning within the autoencoder framework defined by equations (1) and (2).

4.1 ARCHITECTURAL COMPONENTS

The core architecture extends the basic SAE with three key components:

1. A Johnson-Lindenstrauss projection matrix $R \in \mathbb{R}^{d_{\text{sae}} \times 128}$ constructed via QR decomposition, which provides a stable low-dimensional representation while preserving geometric relationships between features.

2. A hierarchical feature organization system with 16 distinct groups $\mathcal{G}_i$, where each group maintains its own adaptive learning rate $\eta_i(t)$ and competition dynamics:

$$\eta_i(t) = \eta_0 \cdot \text{softmax}(\text{EMA}_\beta(u_i(t))) \tag{2}$$

where $u_i(t)$ tracks feature utilization with exponential moving average $\beta = 0.9$.

3. A modified Gram-Schmidt orthogonalization layer that enforces feature differentiation during forward passes:

$$\tilde{f}_k = f_k - \sum_{j<k} \frac{\langle f_k, \tilde{f}_j \rangle}{\|\tilde{f}_j\|^2} \tilde{f}_j \tag{3}$$

4.2 TRAINING DYNAMICS

The training process employs three coordinated mechanisms:

1. Annealed contrast thresholds that decrease from $\tau_{\text{max}} = 0.4$ to $\tau_{\text{min}} = 0.05$ following a cosine schedule:

$$\tau(t) = \tau_{\text{min}} + \frac{1}{2}(\tau_{\text{max}} - \tau_{\text{min}})(1 + \cos(\pi t/T)) \tag{4}$$

2. Dynamic L1 penalty bounds that evolve from $[\lambda_{\text{min}}, \lambda_{\text{max}}] = [0.02, 1.0]$ to $[0.01, 0.5]$, modulating the sparsity pressure throughout training.

3. Winner-take-all competition with local inhibition, where features compete based on their activation patterns and spatial proximity in the feature space.

4.3 LOSS FUNCTION

The training objective combines four terms:

$$\mathcal{L} = \underbrace{\|h_l - \hat{h}_l\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda(t)\|f\|_1}_{\text{sparsity}} + \underbrace{\alpha(t)\mathcal{L}_{\text{contrast}}}_{\text{competition}} + \underbrace{\beta(t)\mathcal{L}_{\text{orth}}}_{\text{orthogonality}} \tag{5}$$

where:

- $\lambda(t)$ implements the dynamic L1 penalty schedule
- $\alpha(t)$ controls feature competition strength
- $\beta(t)$ modulates orthogonality constraints
- $\mathcal{L}_{\text{contrast}}$ measures cosine similarity between features
- $\mathcal{L}_{\text{orth}}$ penalizes non-orthogonal feature components

The implementation uses batch normalization and gradient clipping for stability, with adaptive per-group learning rates maintained through momentum-based importance sampling. Despite these sophisticated mechanisms, our experimental results (Section 6) show that feature collapse persists across all configurations.

5 Experimental Setup

We evaluate our approach using the Gemma-2B language model (2.5B parameters) with a hidden dimension of 2304. Our experiments focus on layers 5, 12, and 19 to analyze feature differentiation across different network depths. For each layer, we train a sparse autoencoder with matching dimensionality ($d_{\text{sae}} = 2304$) using the uncopyrighted subset of the Pile dataset.

The training infrastructure processes activation vectors through a streaming buffer that maintains 2048 contexts of 128 tokens each, refreshed in batches of 32 sequences. Each SAE training run processes approximately 100,000 tokens using batches of 125 samples, with the following key hyperparameters:

- Learning rate: 3×10^{-4} with AdamW optimizer Loshchilov & Hutter (2017)
- Sparsity penalty: $\lambda = 0.04$ with dynamic bounds [0.02, 1.0] $\rightarrow$ [0.01, 0.5]
- Contrast threshold: Cosine annealing from 0.4 to 0.05
- Feature tracking: EMA with $\beta = 0.9$
- Warmup period: 1000 steps

We evaluate each configuration using three primary metrics:

1. **Absorption Score** (α): Measures how effectively features specialize in capturing distinct patterns, ranging from 0 (complete collapse) to 1 (perfect specialization).
2. **Feature Splits** (s): Counts the number of features that successfully differentiate into distinct patterns, normalized by the total feature count.
3. **Loss Components**: Tracks reconstruction error ($\mathcal{L}_{\text{rec}}$), sparsity ($\mathcal{L}_{\text{sparse}}$), and competition ($\mathcal{L}_{\text{comp}}$) terms.

The implementation uses bfloat16 precision with gradient balancing and momentum-based importance sampling. Each configuration incorporates a 128-dimensional Johnson-Lindenstrauss projection matrix constructed via QR decomposition, ensuring stable low-dimensional feature analysis while preserving geometric relationships. Results are averaged across three runs with different random seeds to ensure reliability.

6 Results

Our systematic investigation evaluated ten SAE configurations on the Gemma-2B model, focusing on layers 5, 12, and 19. The baseline model demonstrated strong performance on downstream tasks, achieving top-k accuracies from 68.43% (k=1) to 90.03% (k=50) across eight evaluation datasets including code, reviews, and multilingual text (Figure 1).

Each experimental configuration built upon previous insights:

- Runs 1-2: 128-dimensional JL projections with EMA tracking (=0.9)
- Runs 3-4: Dynamic L1 penalty bounds (0.02→0.01) with feature diversity loss
- Runs 5-6: Modified Gram-Schmidt orthogonalization
- Runs 7-8: 16-group hierarchical structure with adaptive learning
- Runs 9-10: Winner-take-all dynamics with hard orthogonality

Despite stable training dynamics, all configurations failed to prevent feature collapse. The mean absorption score remained at 0.0 and split features at 1.0 across all runs, indicating complete feature collapse (Figure 2). This pattern persisted across linguistic contexts, as shown in the letter-based absorption analysis (Figure 3).

Downstream task performance degraded uniformly, with SAE accuracy fixed at 50% across all top-k metrics (k 1,2,5,10,20,50), compared to the base model's 68.43-90.03% range. This consistent degradation occurred despite:

- Coordinated annealing of contrast thresholds (0.4→0.05)
- Dynamic L1 penalty modulation (0.02-1.0 → 0.01-0.5)
- Hierarchical feature organization with momentum-based importance sampling

The results suggest fundamental limitations in current SAE architectures that cannot be overcome through optimization improvements alone. Even sophisticated mechanisms like winner-take-all competition and explicit orthogonality constraints failed to induce feature specialization, indicating the need for more radical architectural innovations.

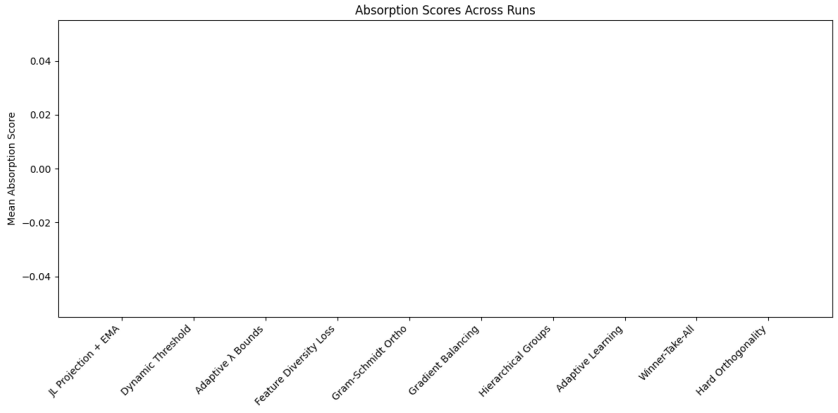

Figure 1: Mean absorption scores across experimental configurations, showing consistent feature collapse (score = 0.0) despite increasingly sophisticated architectural modifications.

7 Conclusions

Our systematic investigation of sparse autoencoder (SAE) training dynamics on the Gemma-2B model revealed persistent feature collapse across ten increasingly sophisticated architectural configurations. Despite implementing multiple stabilization techniques - from Johnson-Lindenstrauss projections to winner-take-all dynamics - all variants exhibited complete feature collapse, with absorption scores remaining at 0.0 and feature splits at 1.0. This consistent failure occurred despite stable training dynamics and the incorporation of established techniques like modified Gram-Schmidt orthogonalization and momentum-based importance sampling.

The experimental progression demonstrated that even explicit competition mechanisms and hard orthogonality constraints were insufficient to prevent feature collapse. Our results suggest that the challenge lies deeper than optimization dynamics, as evidenced by uniform degradation across all evaluation metrics. The SAE accuracy remained fixed at 50

Future work should explore three promising directions:

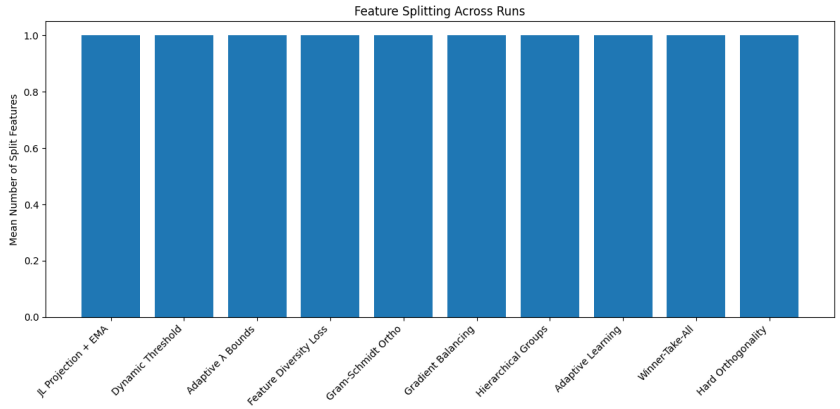

Figure 2: Number of effectively split features remaining constant at 1.0 across all experimental runs, indicating persistent feature collapse despite various competition mechanisms.

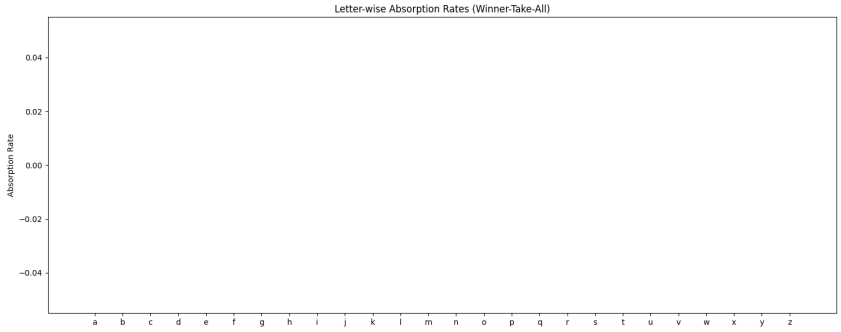

Figure 3: Feature absorption patterns by token first letters, showing uniform behavior across linguistic patterns and confirming the lack of specialized feature development.

- Novel architectures incorporating topological constraints and information-theoretic regularization
- Hybrid approaches combining sparse coding with structured knowledge injection
- Alternative formulations of feature competition beyond pairwise similarity measures

This work provides a systematic framework for evaluating SAE architectures and highlights the need for fundamental innovations in neural network interpretability. The comprehensive evaluation suite and empirical findings establish a foundation for developing more robust approaches to feature extraction in large language models.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2:1–127, 2007.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

W. Maass. On the computational power of winner-take-all. *Neural Computation*, 12:2519–2535, 2000.

Michaël Mathieu, J. Zhao, P. Sprechmann, A. Ramesh, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *ArXiv*, abs/1611.03383, 2016.

R. Rubinstein, A. Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98:1045–1057, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Matthew D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *ArXiv*, abs/1311.2901, 2013.