Kronecker Factorization Improves Efficiency and Interpretability of Sparse Autoencoders

Vadim Kurochkin, Yaroslav Aksenov, Daniil Laptev, Daniil Gavrilov, Nikita Balagansky

T-Tech

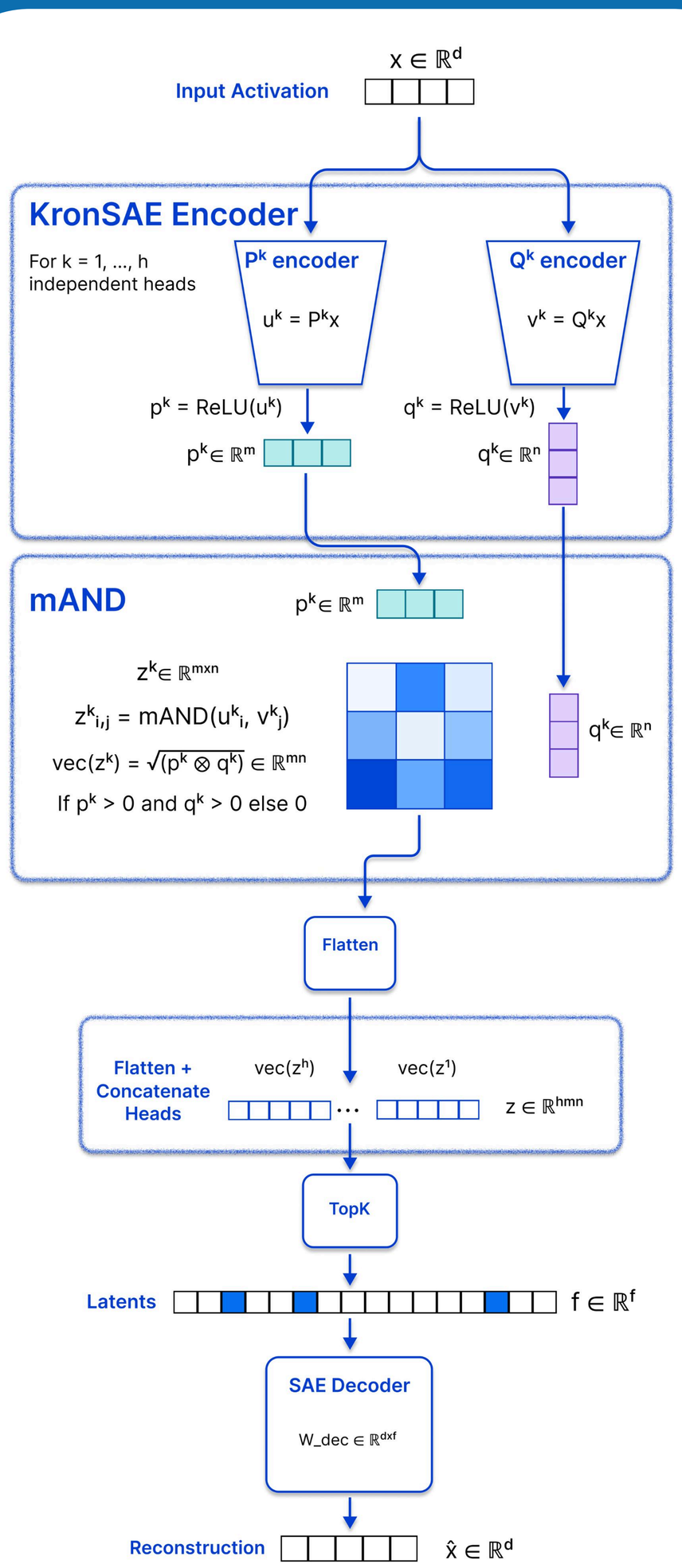

Figure 1. KronSAE splits each activation into head-wise pre-latents, composes them with mAND so post-latents activate only on paired co-activations, then applies TopK and the standard SAE decoder.

1 TL;DR

- Standard SAEs use flat latent dictionaries, KronSAE makes feature structure explicit.
- KronSAE uses a **Kronecker** factorization for encoder plus mAND to build latents.
- KronSAE uses **>40% fewer FLOPs** with minimal EV drop.
- KronSAE **reducing feature absorption** and **increasing AutoInterp scores**.

2 Method

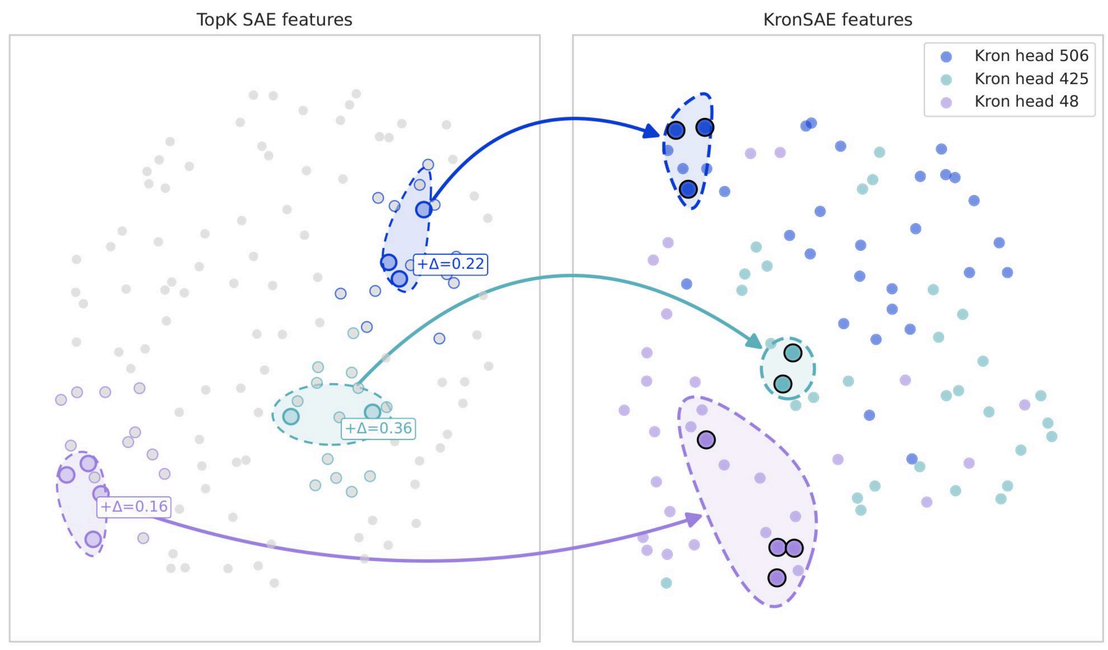

Figure 2. KronSAE turns default SAE features into structured groups: correlated TopK features map to nearby KronSAE features within the same head, suggesting that heads capture feature co-activation structure.

3 Improved Interpretability

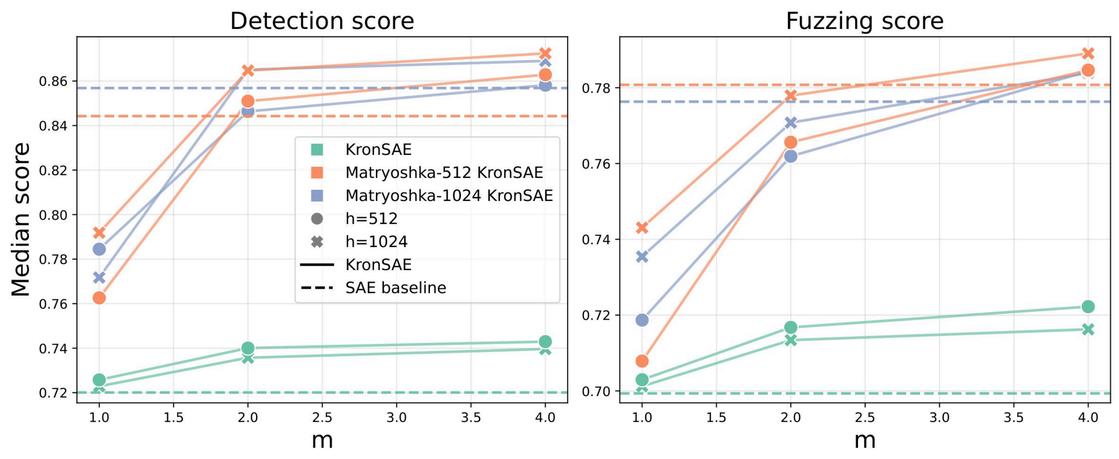

Figure 3. For all SAEs - KronSAE have improvement in scores, with inreasing m. With m ≥ 2 all modifications match or exceed baselines.

4 Same features - same heads

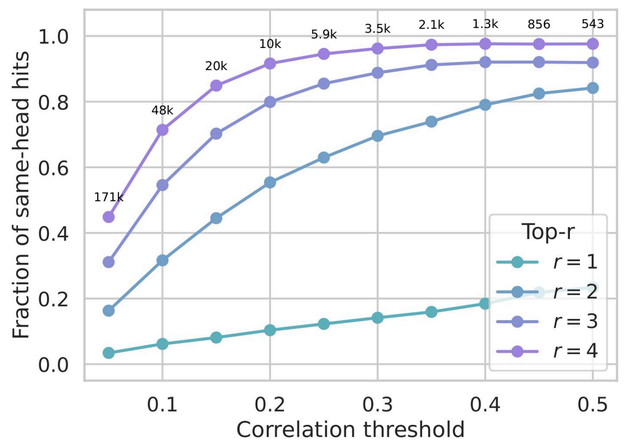

Figure 4. KronSAE heads successfully group correlated TopK features: highly correlated feature pairs are much more likely to map into the same head than expected by chance.

5 Reduced FLOPs

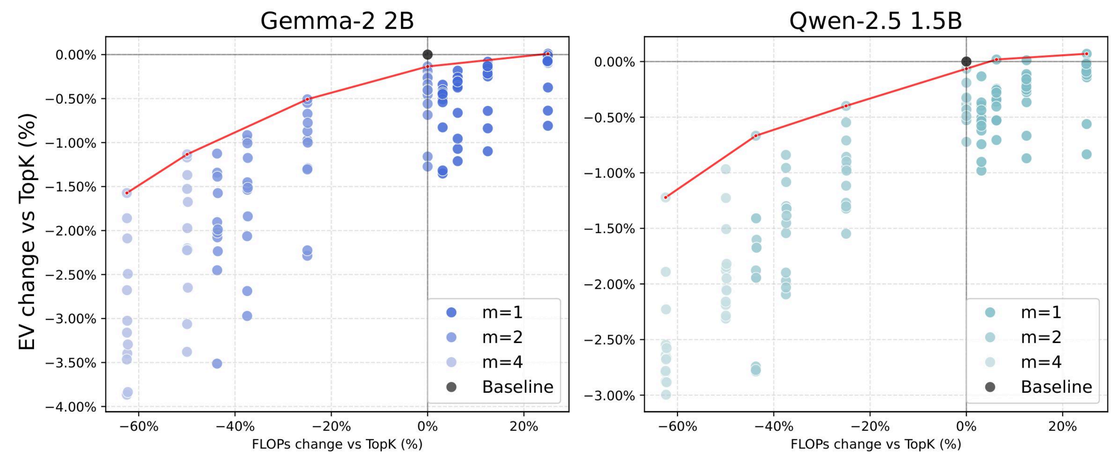

Figure 5. KronSAE offer better trade-off in term of reconstruction quality: while we have less than 1% percent drop, we get more than 40% speedup