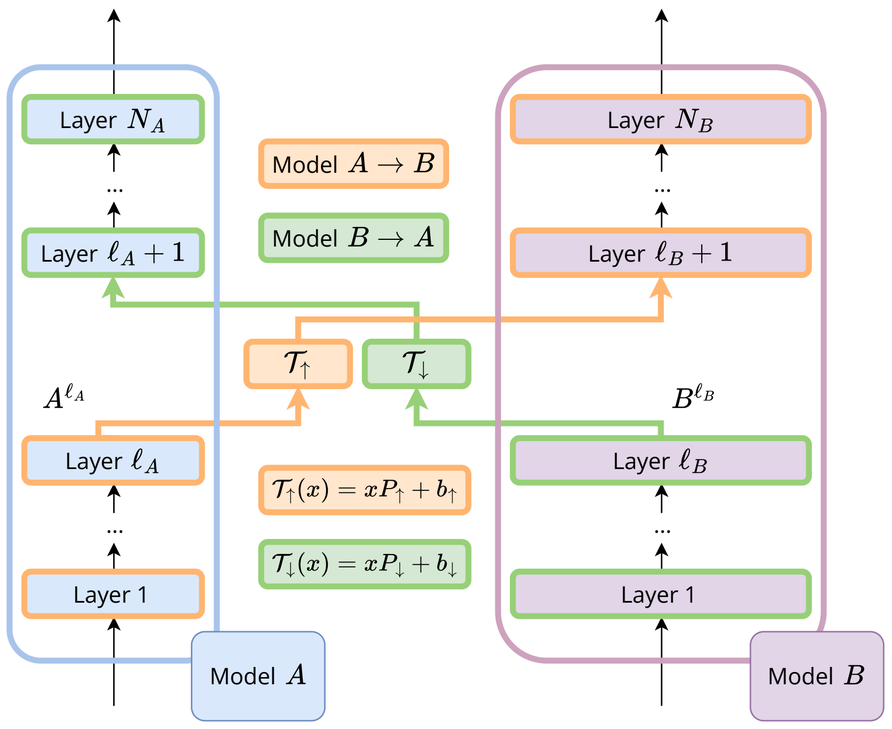
Transferring Features Across Language Models with Model Stitching

Alan Chen[1], Jack Merullo[1,2], Alessandro Stolfo[3], Ellie Pavlick[1]

[1]Brown University [2]Goodfire [3]ETH Zurich

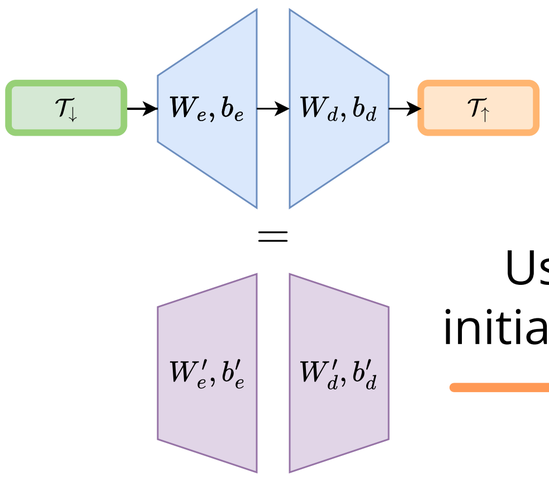
Feature universality can be leveraged to cheaply transfer linear features between models.

(a) Training a Stitch



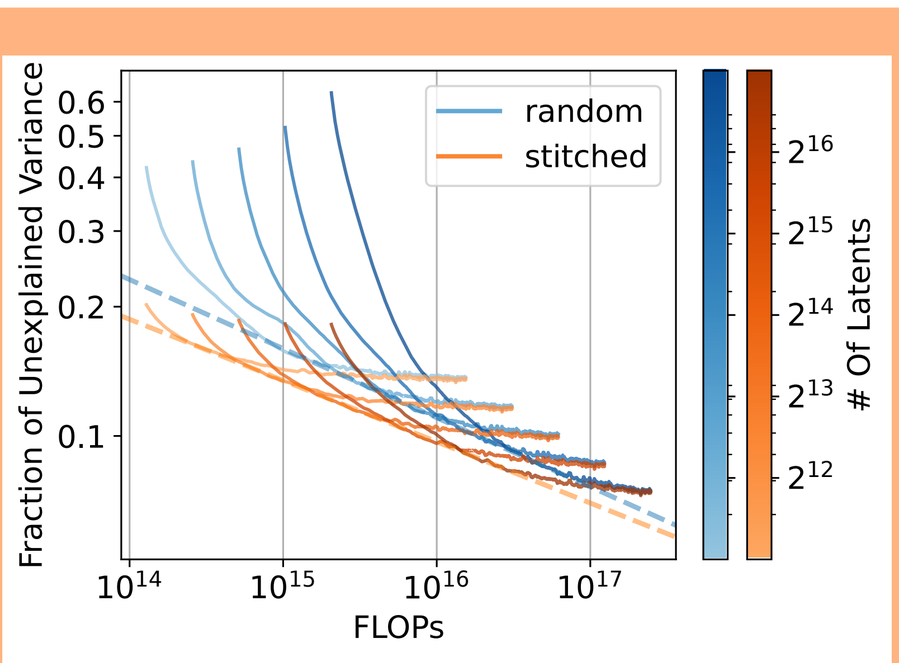
We train two affine transformations together using an MSE-based penalty that includes reconstruction and inversion terms. The transformation mostly preserves downstream next token prediction loss, capturing weak universality.

(b) Transferring SAEs

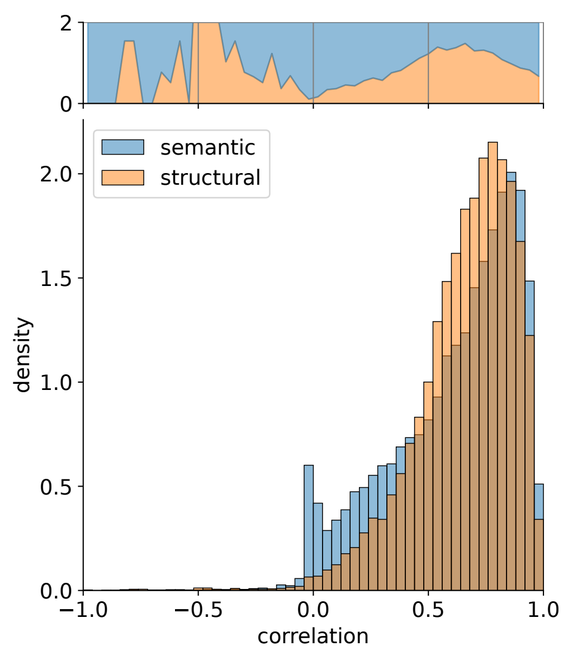


Use as initialization

We derive a method to transfer entire sets of linear features learned by SAEs. We first stitch a latent B->A, apply the original SAE, then stitch A->B. Because everything is affine, we can equivalently construct an SAE on B that represents this computation.

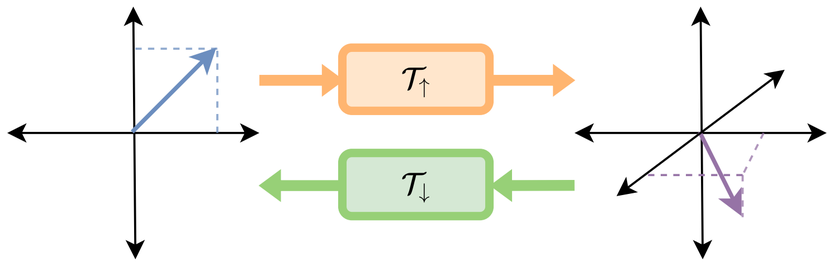


We can use the transferred SAEs as initialization for training runs, saving the compute of "relearning" features that transfer well.



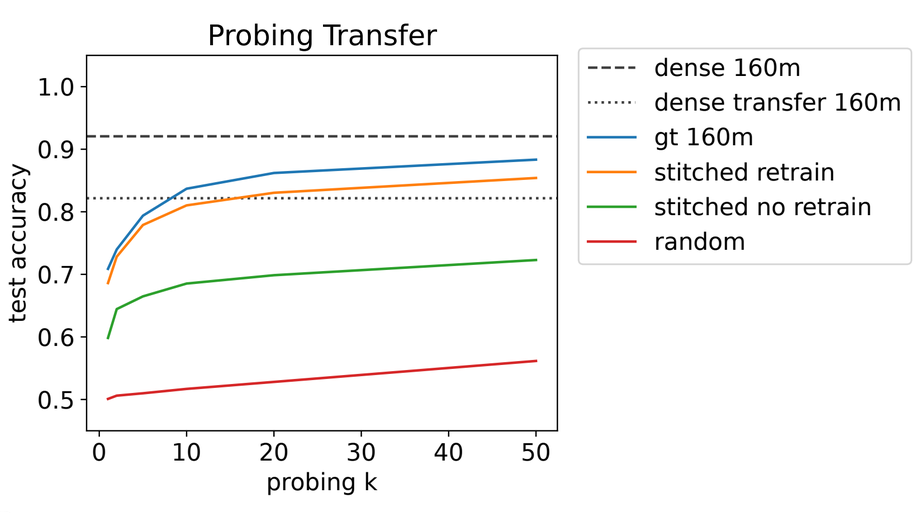
Structural and semantic features transfer differently according to attribution correlation.

(c) Transferring Features

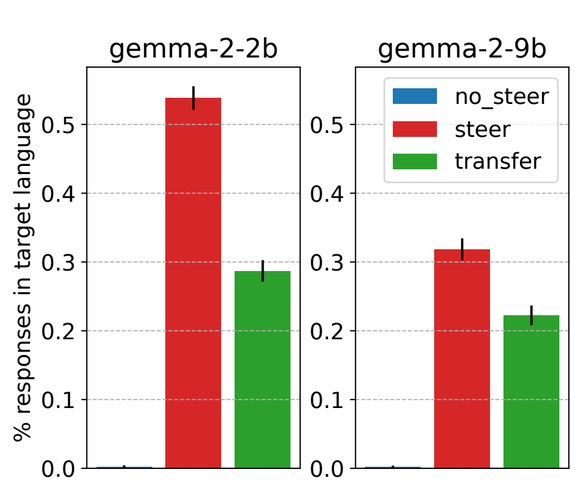


The stitches can easily be used to transfer linear features, allowing us to transfer (i) probes and (ii) steering vectors. We also find functional features that are preserved (e.g. entropy neurons, attention deactivation).

(i) Probing



(ii) Language Steering



Acknowledgements: LUNAR Lab (Brown), CSCI2222 @ Brown, Neel Nanda (DeepMind)