Identifying functionally important features with end-to-end sparse dictionary learning

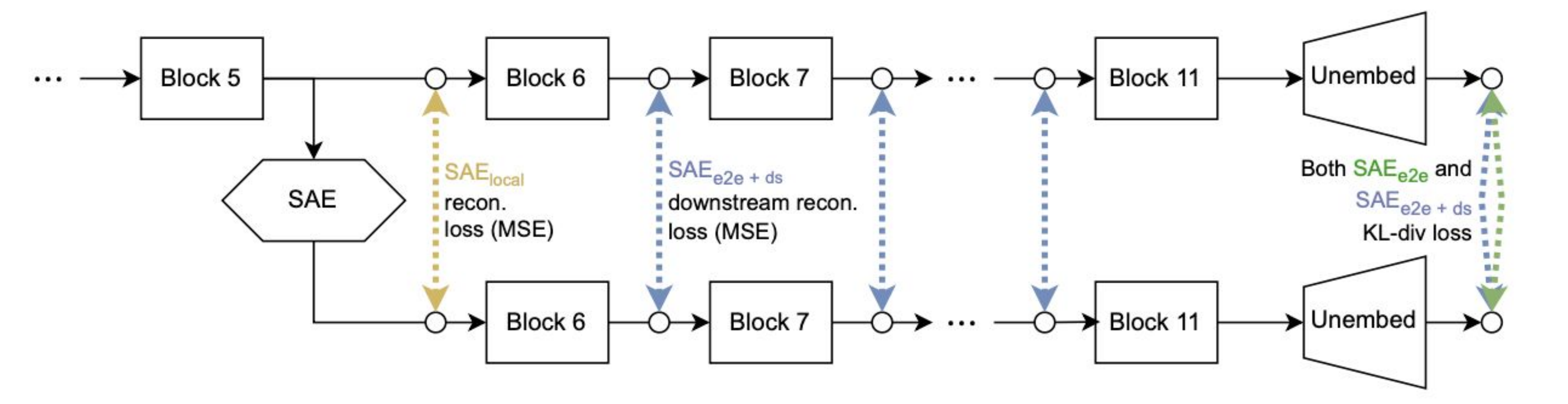
Dan Braun, Jordan Taylor, Nix Goldowsky-Dill,
Lee Sharkey

Presenter: Dan Braun
21 October 2024

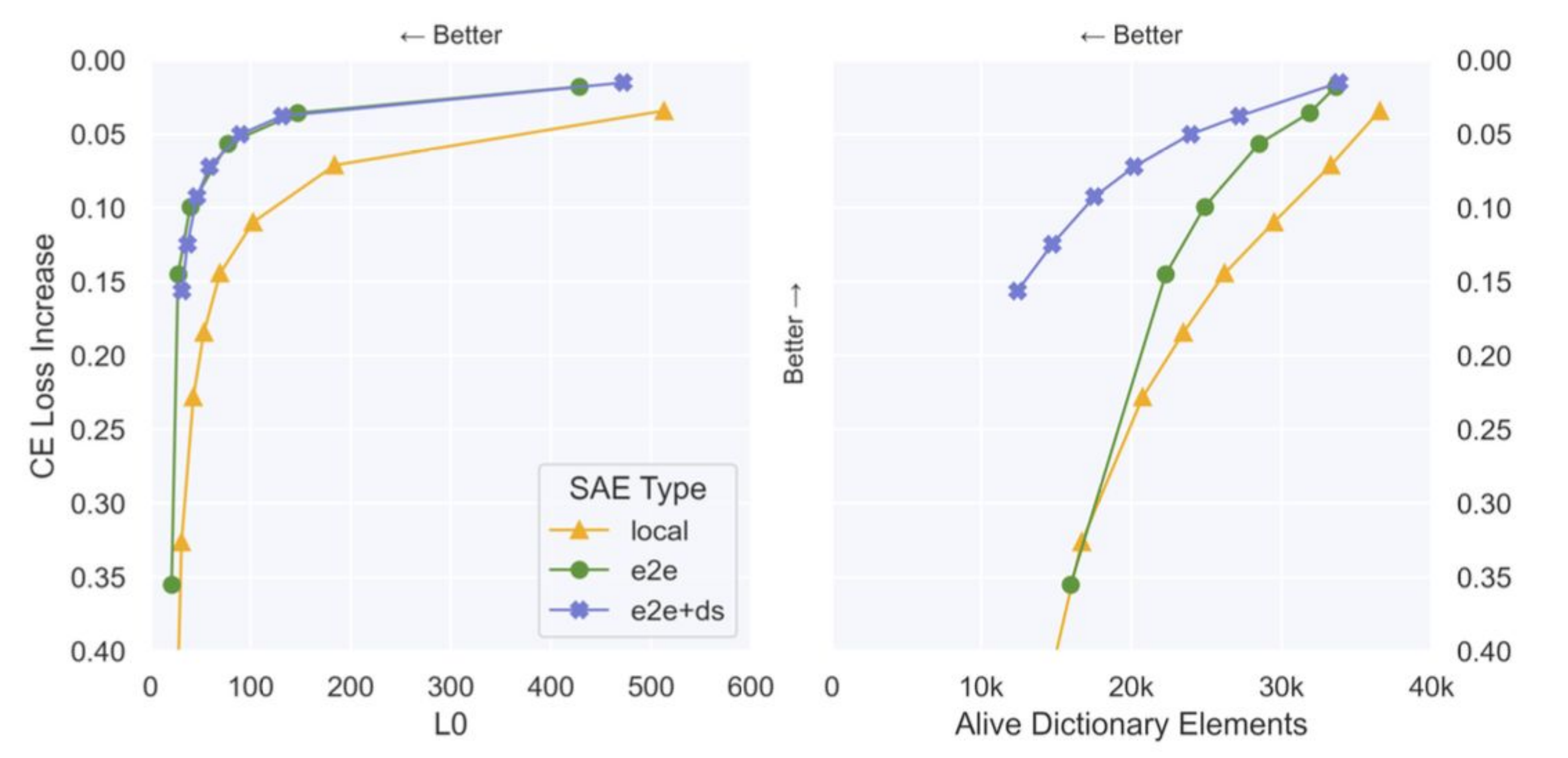
Problems with standard “local” SAEs

- Local SAEs are trained only to minimize the reconstruction loss at a layer.
- This does not prioritize learning features based on explaining network performance.
- Therefore, they may learn less important or irrelevant features.

Three settings: local, e2e, e2e+downstream



Results: Pareto Improvement

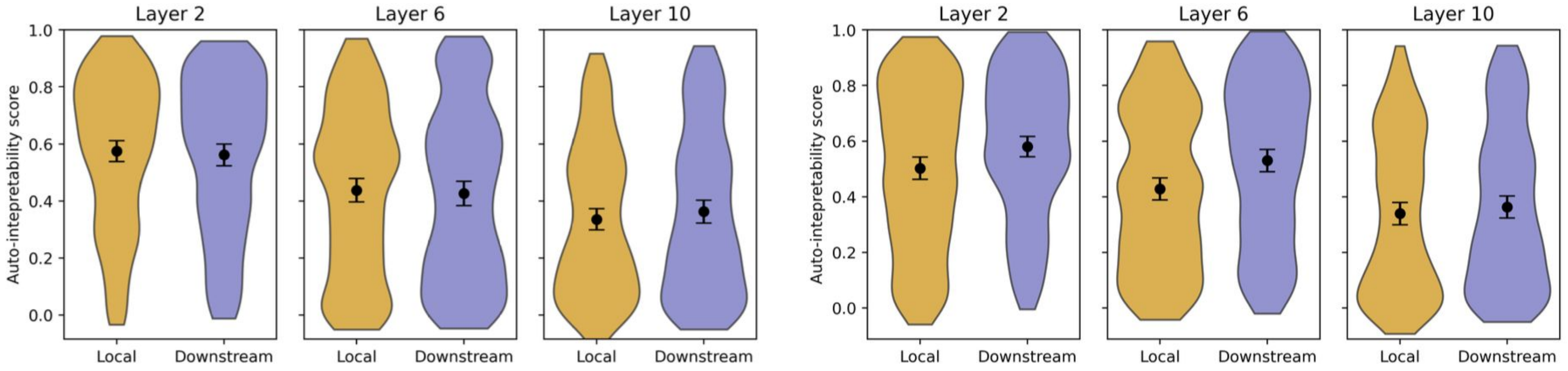


Fewer than half the active features per datapoint (L0) to explain same performance.

ML ALIGNMENT & THEORY SCHOLARS

APOLLO RESEARCH

Results: Auto-Interpretability



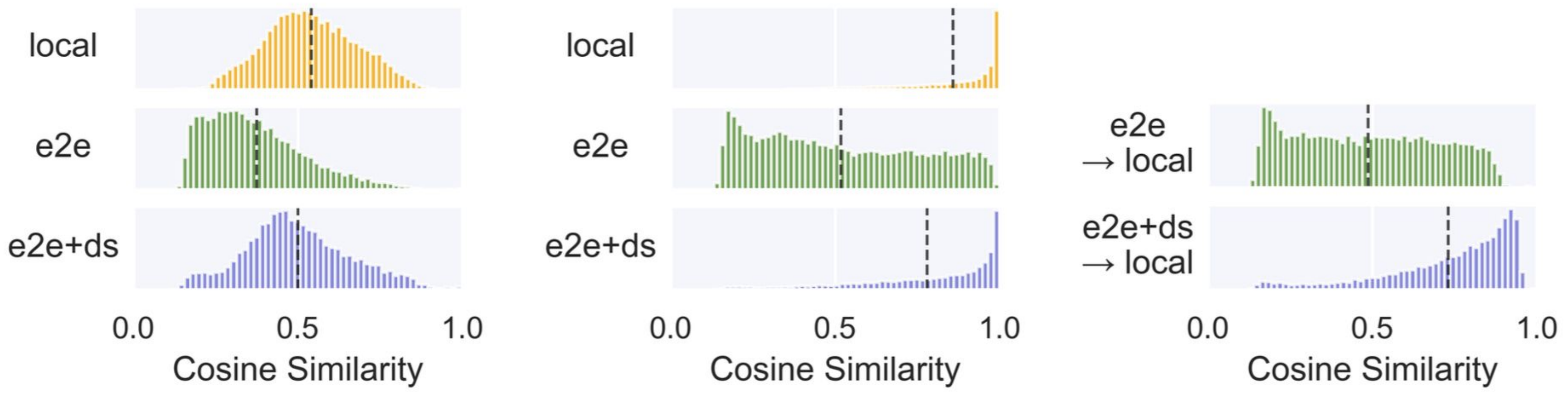
(a) Similar L_0

(b) Similar CE Loss increase

Figure 9: Comparison of auto-interpretability scores between $\text{SAE}_{\text{e2e+ds}}$ and $\text{SAE}_{\text{local}}$. We choose two pairs at every layer, one with similar L_0 (see Table 3) and the other with similar CE loss increase (see Table 2). Error bars are a bootstraped 95% confidence interval for the true mean auto-interpretability scores. Measured on approximately $200(\pm 2)$ randomly selected features per dictionary.

e2e+downstream SAEs are approximately interpretable as local SAEs

Results: Dictionary Geometry



(a) Within-SAE cosine similarity (b) Cross-seed cosine similarity (c) Cross-SAE-type cosine similarity

Figure 3: Geometric comparisons for our set of GPT2-small layer 6 SAEs with similar CE loss increases (Table 1). For each dictionary element, we find the max cosine similarity between itself and all other dictionary elements. In 3a we compare to others directions in the same SAE, in 3b to directions in an SAE of the same type trained with a different random seed, in 3c to directions in the $\mathrm{SAE}_{\mathrm{local}}$ with similar CE loss increase.

Downside

- e2e SAEs are 2.5x slower to train (from scratch) on GPT2-small.

Scan for paper

SEE OUR POSTER!