Contextures: Representations from Contexts

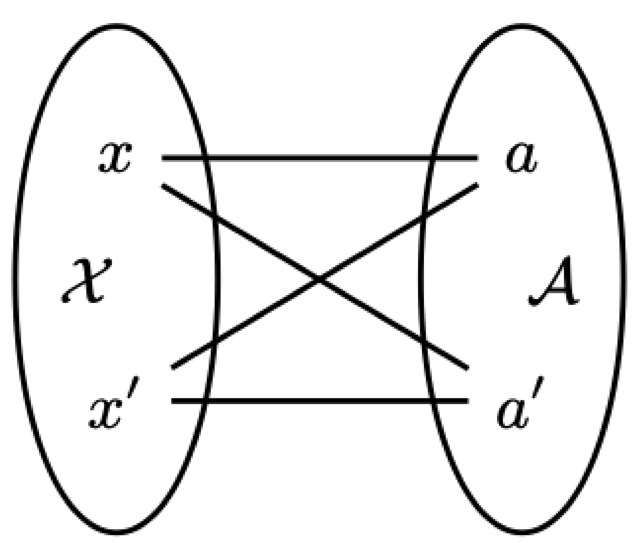
Runtian Zhai, Kai Yang, Burak Varıcı, Che-Ping Tsai, Zico Kolter, Pradeep Ravikumar

Contexture theory: Representations are learned from the association between input X and context variable A

What representations do modern models learn?

- Transferability to downstream tasks completely different from pretraining?
- Representation similarity: Why different models learn ***similar*** representations?
- Scaling law: Are bigger models always better?

Method	Input X	Context A
Supervised	Sample	Label of X
Contrastive	Image	Crop of X
LLMs (GPT)	Text	First k tokens
Vision-language	Image	Text caption



Learn encoder $\Phi : \mathcal{X} \to \mathbb{R}^d$

Intuition: models learn low-order spectral approximation of an implicit kernel induced by input-context pair

- Joint distribution: $P^+(X, A)$, marginals: P_X, P_A
- L^2 space: $f \in L^2(P_X) \implies E_{P_X}[f(X)^2] < \infty$
- **Expectation Operator** $T_{P^+} : L^2(P_A) \to L^2(P_X)$

$$(T_{P^+} g)(x) = \mathbb{E}\left[g(A) \mid x\right]$$

- SVD of T_{P^+}: $\begin{cases} \text{sing. values: } 1 = s_0 \geq s_1 \geq \cdots \geq 0 \\ \text{sing. func. } (\mu_i) \in L^2(P_X),\ (\nu_i) \in L^2(P_A) \end{cases}$
- $P^+(x, a) = \sum_{i \geq 0} s_i \mu_i(x) \nu_i(a) P_X(x) P_A(a)$

Result 1: What representations do we really learn?

Foundation models recover the space spanned by the top-d singular functions of T_{P^+}:

- Supervised learning
- Contrastive / noncontrastive learning
- Masked autoencoders
- Node representation learning on graphs

Informal Theorem:
Optimizer Φ of these objectives over $L^2(P_X)$ span the same subspace as the top-d singular functions of T_{P^+}

$$\mathrm{span}(\phi_1, \ldots, \phi_d) = \mathrm{span}(\mu_1, \ldots, \mu_d)$$

Result 2: When do these representations work?

The representation recovering the top-d eigenspace is optimal over the class of all compatible tasks

Informal: A task is compatible if A helps learn a predictor for it

Compatibility: $\max_{g \in L^2(P_A)} \dfrac{\langle f, T_{P^+} g \rangle_{P_X}}{\|f\|_{P_X} \|g\|_{P_A}}$

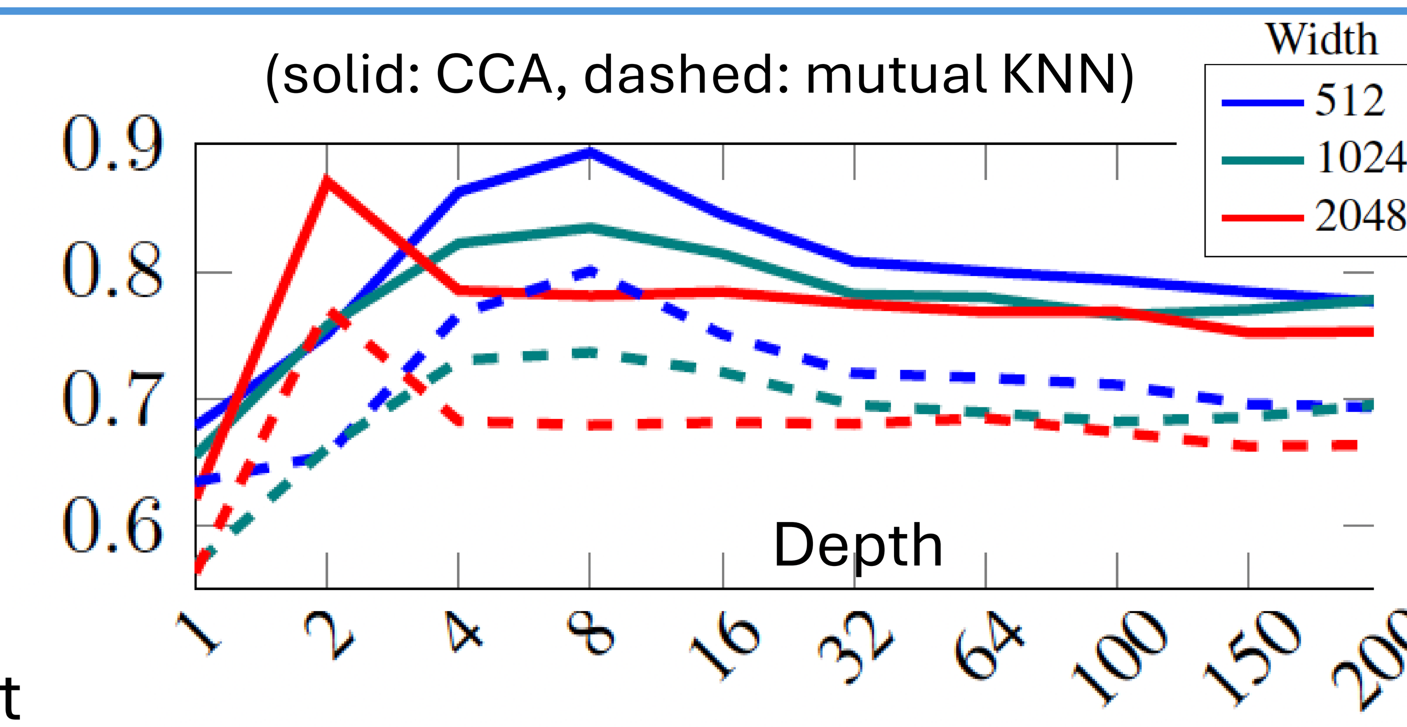
Result 3: Empirical evidence and implications

Deep nets learn the top-d eigenspace empirically.

Implications for scaling laws

Increasing model size = diminishing returns

- Encoder converges to the top-d eigenspace
- When close enough, further scaling has little effect



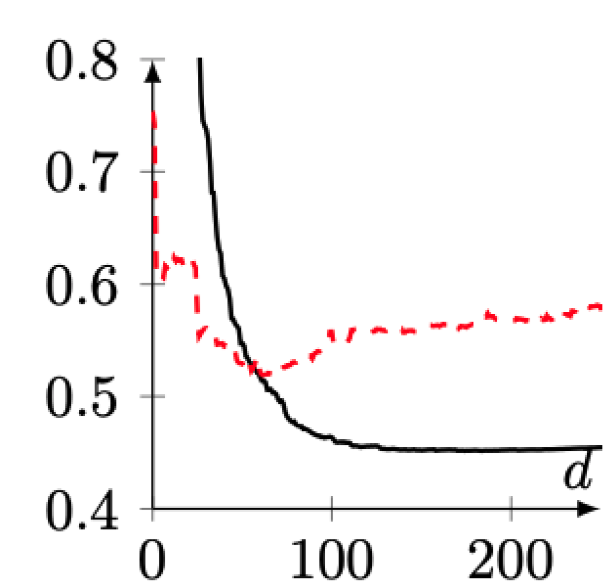
Alignment of representations and true eigenfunctions (abalone dataset)

Result 4: Evaluating contexts

A metric to predict the downstream error

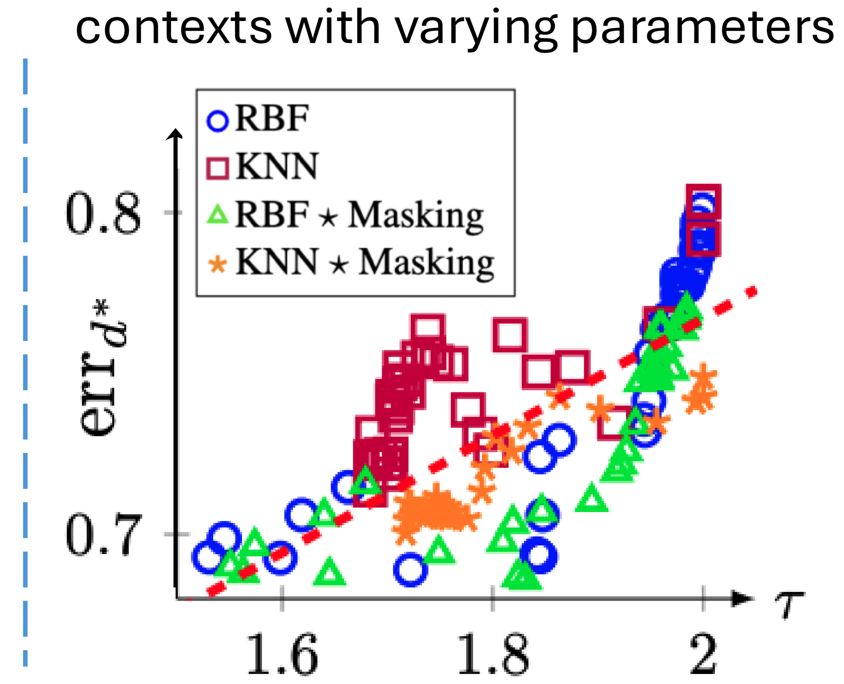
✓ Only depends on the singular values
✓ Strong correlation with error on real datasets
(over 28 datasets, 0.43 mean, 0.58 median Pearson correlation)
✓ Selecting pretraining methods and hyperparams

$$\tau_d = \frac{1}{1 - s_{d+1}^2} + \beta \frac{\sum_{i=1}^d s_i^2}{\sum_{i=1}^\infty s_i^2}, \quad \tau = \min_d \tau_d$$



τ_d vs. error (abalone dataset)

contexts with varying parameters



metric-downstream error correlation (diabetes dataset)