Contextures: Mechanism of Representation Learning

Runtian Zhai, CMU CS PhD Dissertation

Carnegie Mellon University

Why are foundation models so good on downstream tasks so different from pretraining task?

When will the scaling law end?

The Contexture Theory

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

Method	Input X	Context Var A
Supervised	Sample	Label of X
KNN	Sample	Neighbor of X
Contrastive	Image	Cropped image
LLMs (GPT)	Text	First k tokens
Diffusion	Image	X plus noise
Vision-lang	Image	Text caption

Joint dist. $P^+(X, A)$, marginals P_X, P_A

L^2 function spaces $L^2(P_X), L^2(P_A)$

Expectation operator

$$T_{P^+}: L^2(P_A) \to L^2(P_X)$$

$$(T_{P^+}g)(x) = \mathbb{E}_{P^+}[g(A)|x]$$

SVD of T_{P^+} (analogous to PCA)

- Singular values $1 = s_0 \geq s_1 \geq \cdots \geq 0$
- Left singular func. $\mu_0, \mu_1, \cdots \in L^2(P_X)$
- $\mu_0 \equiv 1, (\mu_i)_{i \geq 0}$ forms basis of $L^2(P_X)$

Result 1: Pretrained models transfer to tasks compatible with the context

- Compatible: The context helps learn a predictor of target $f^* \in L^2(P_X)$
- Metric: $\rho(f^*, P^+) = \max_{g \neq 0} \frac{\langle f^*, T_{P^+}g \rangle}{\|f^*\|_{P_X} \|g\|_{P_A}}$
- Comp. set: $\mathcal{F}_c = \{f^* : \rho(f^*, P^+) \geq c\}$
- Predictor: $\hat{f}(x) = \boldsymbol{W}\Phi(x) + \boldsymbol{b}$

Result 2: The optimal encoder Φ on $\mathcal{F}_c$ learns the span of top-d singular func.

$$span(\phi_1, \cdots, \phi_d) = span(\mu_1, \cdots, \mu_d)$$

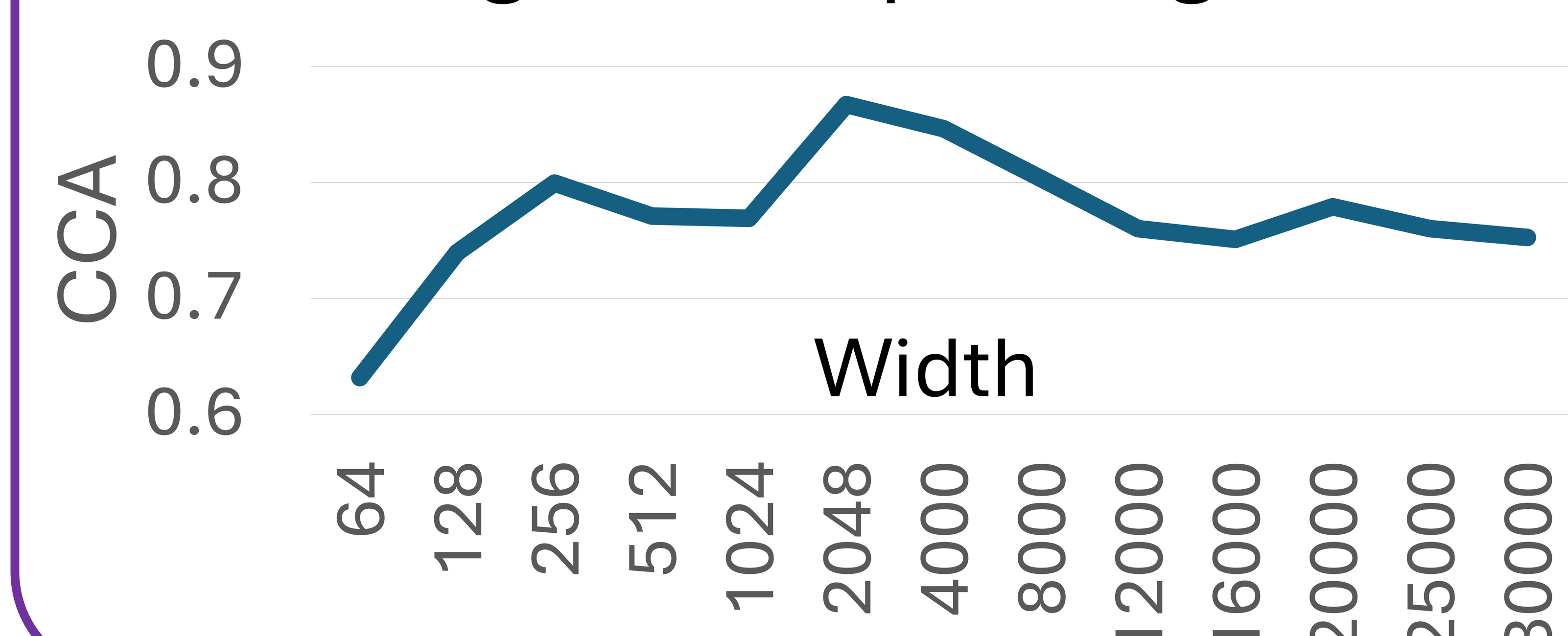
$\mu_0 \equiv 1$ is implicitly included in $\boldsymbol{b}$

Theorem: This Φ achieves the lowest worst-case approximation error on $\mathcal{F}_c$

Analogy: PCA

- Spaces of X and A are finite
- $f^* \in \mathbb{R}^{|X|}$, $T_{P^+} = \boldsymbol{T}$ is a matrix
- Goal: Learn embedding $E \in \mathbb{R}^{|X| \times d}$
- **PCA:** Top-d singular vectors of $\boldsymbol{T}$ minimize the prediction error of f^*

Result 3: Representations of deep models align with top-d singular func.



Result 4: Scaling up the model size produces diminishing returns. Further progress requires better contexts

- Size ↑: Func. class of $\Phi \to L^2(P_X)$
- Model → Top-d singular functions
- If close enough, scaling has no use

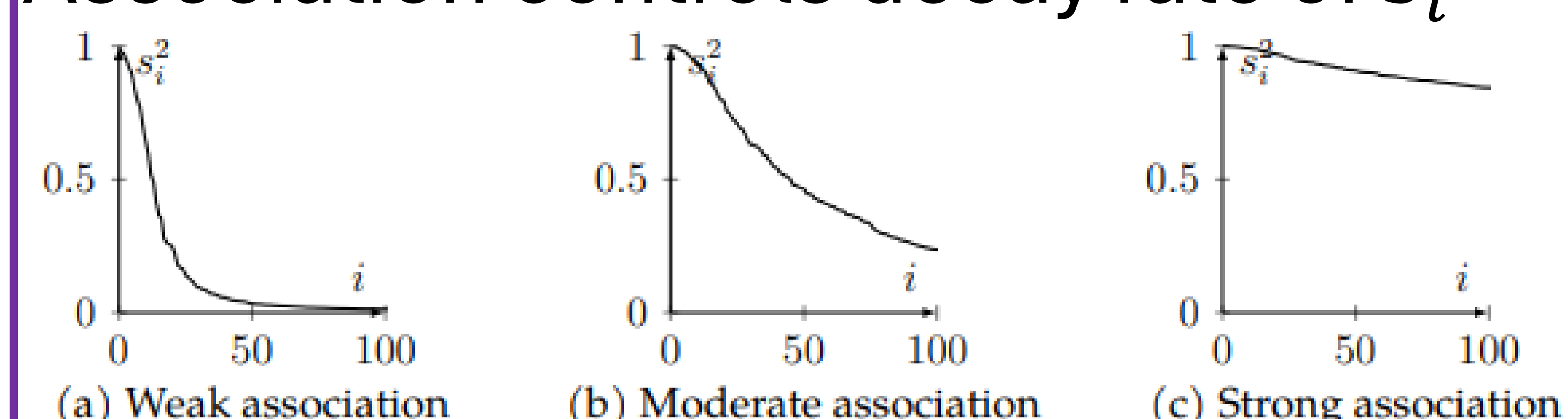
Scaling law is not all we need.

Scientific understanding is a must

Result 5: A good context should have moderate association between X and A

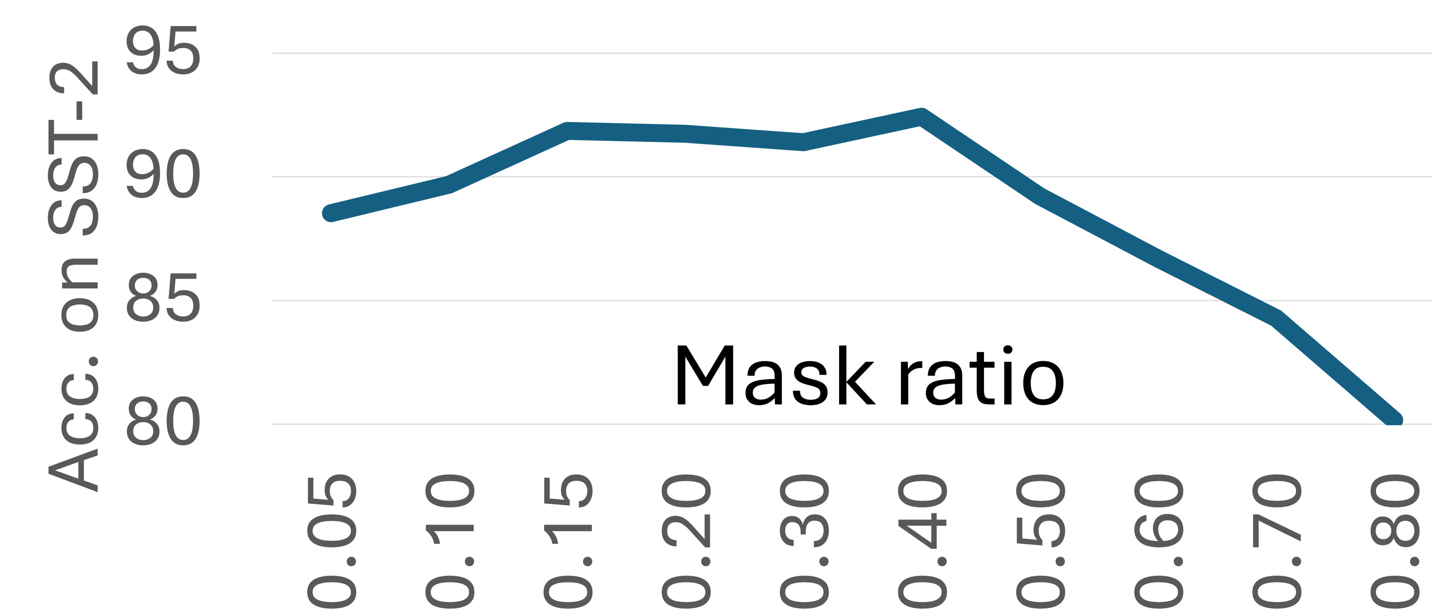
- Extremely weak: A is random noise
- Extremely strong: $A = X$

Association controls decay rate of s_i



- Too weak: Few tasks are compatible
- Too strong: High sample complexity

BERT is best with moderate mask ratio



Result 6: Mixing multiple contexts can produce better contexts by balancing their associations