A Generalization Theory for Zero-Shot Prediction

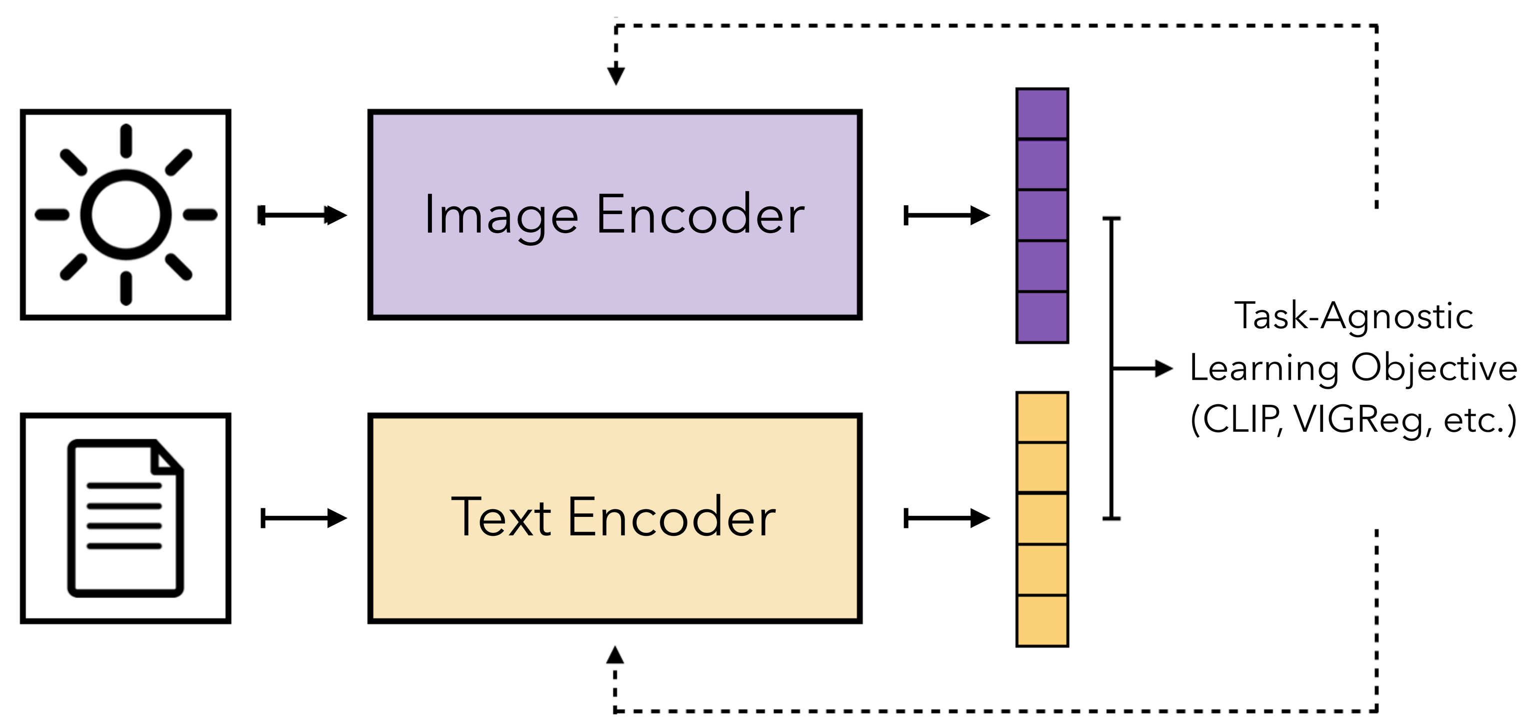
Ronak Mehta and Zaid Harchaoui

Zero-Shot Prediction (ZSP)

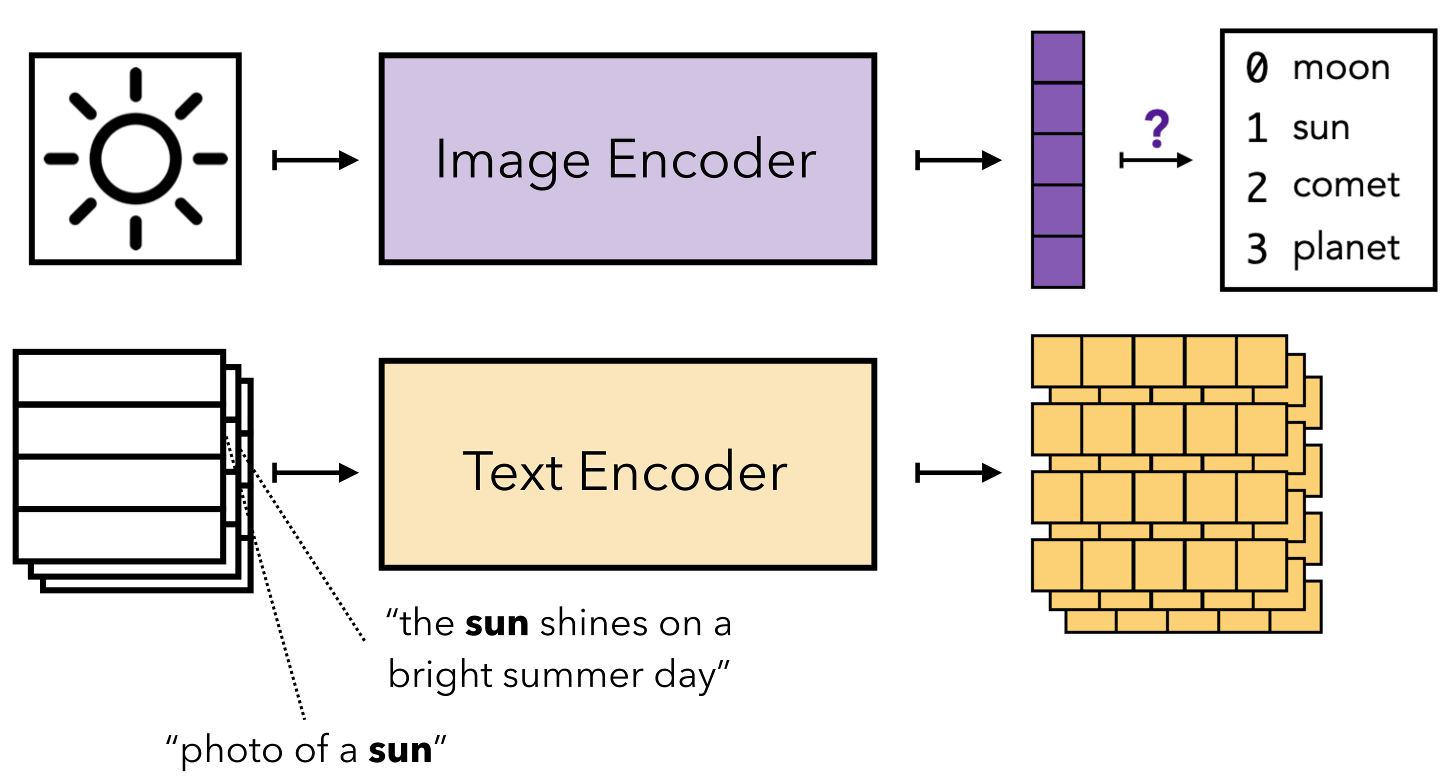
Motivation: Zero-shot prediction is a modern method that reuses foundation models to build classifiers for tasks without seeing *any* directly labeled training data.

Need for theoretical understanding has arisen.

Contrastive Pre-Training



Evaluation



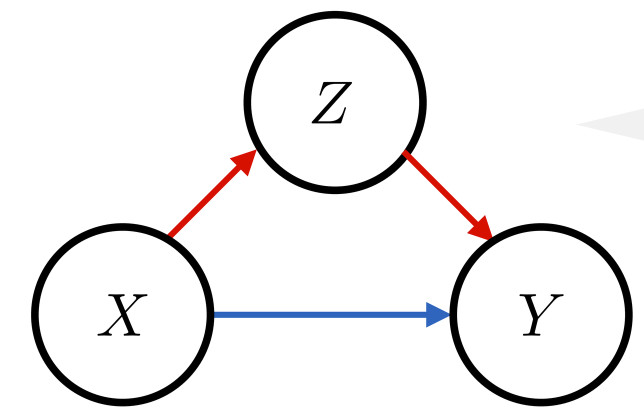
Research Question: How does the downstream performance of ZSP depend on the pre-training distributions, downstream task distribution, and prompting strategy?

Theoretical Framework

Fundamental limits of ZSP rely on the compatibility of three distributions.

$P_{X,Y}$	$Q_{X,Z}$	$\rho_{Y,Z}$
Evaluation	Pre-Training	Prompting

X = image
Y = label
Z = caption



How well can prediction through Z perform for any joint distribution $P_{X,Y,Z}$ that nearly agrees with marginals?

Direct Predictor
$$f_\star(\boldsymbol{x}) = \mathbb{E}_{P_{X,Y}}[Y|X=\boldsymbol{x}]$$

Indirect Predictor (Population Version of ZSP)
$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}[Y|Z]\,|X=\boldsymbol{x}\right]$$

Main Results

Error decomposition for ZSP procedures.

$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \leq$$
$$\underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

Theorem.
$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right] \lesssim I(X,Y|Z) + \mathrm{err}(P_{Y,Z},\rho_{Y,Z})$$

Interpretation: Residual dependence between image/label not explained by text.

Interpretation: Bias of prompt distribution.

Theorem.
$$\mathbb{E}_{X\sim P_X}[(\bar{f}(X)-\hat{f}(X))^2] \lesssim C_N(Q_{X,Z}) + C_M(\rho_{Y,Z})$$

Interpretation: Complexity of learning foundation model (e.g., CLIP) from N pre-training examples.

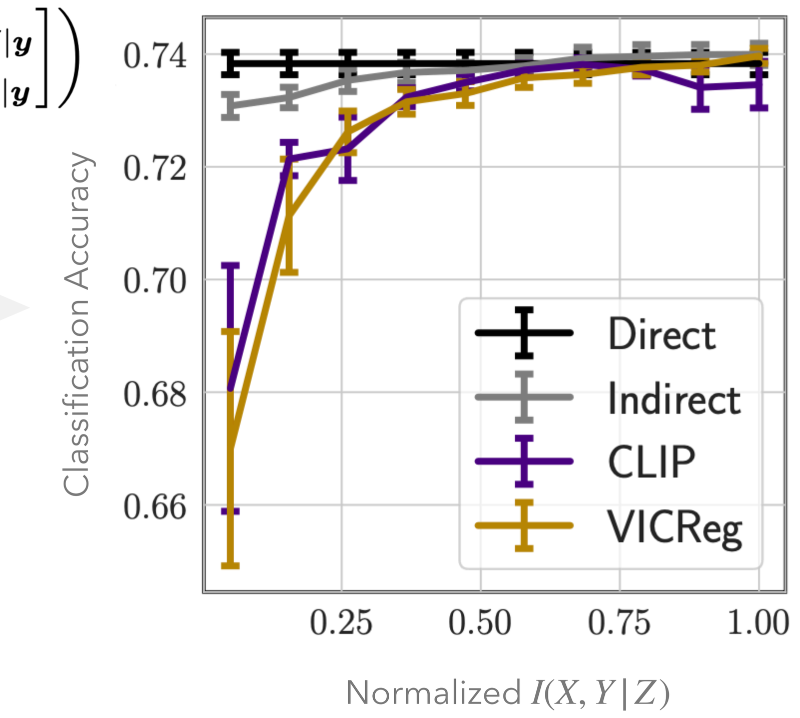
Interpretation: Complexity of approximating prompt distribution with M prompts.

Experiments

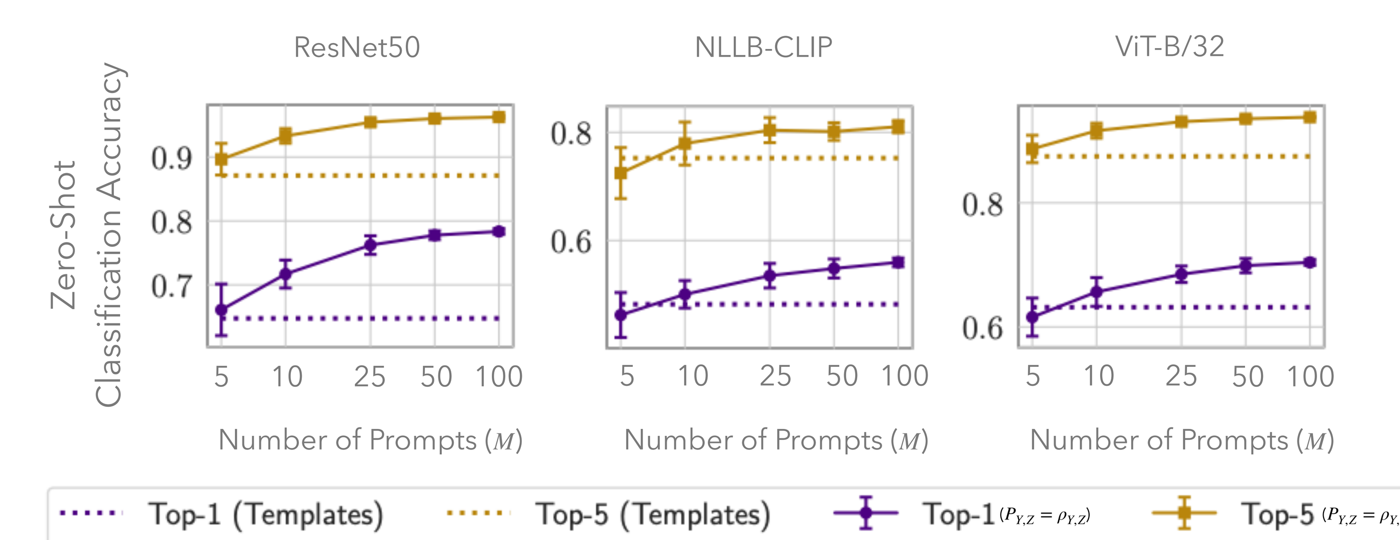
Synthetic Data Example: Controllable Residual Dependence and Prompt Bias

$$\begin{bmatrix}X\\Z\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_{X|y}\\ \mu_{Z|y}\end{bmatrix}, \begin{bmatrix}C_{XX|y} & C_{XZ|y}\\ C_{ZX|y} & C_{ZZ|y}\end{bmatrix}\right)$$

Methods such as CLIP and VICReg have similar performance trends to the indirect predictor.

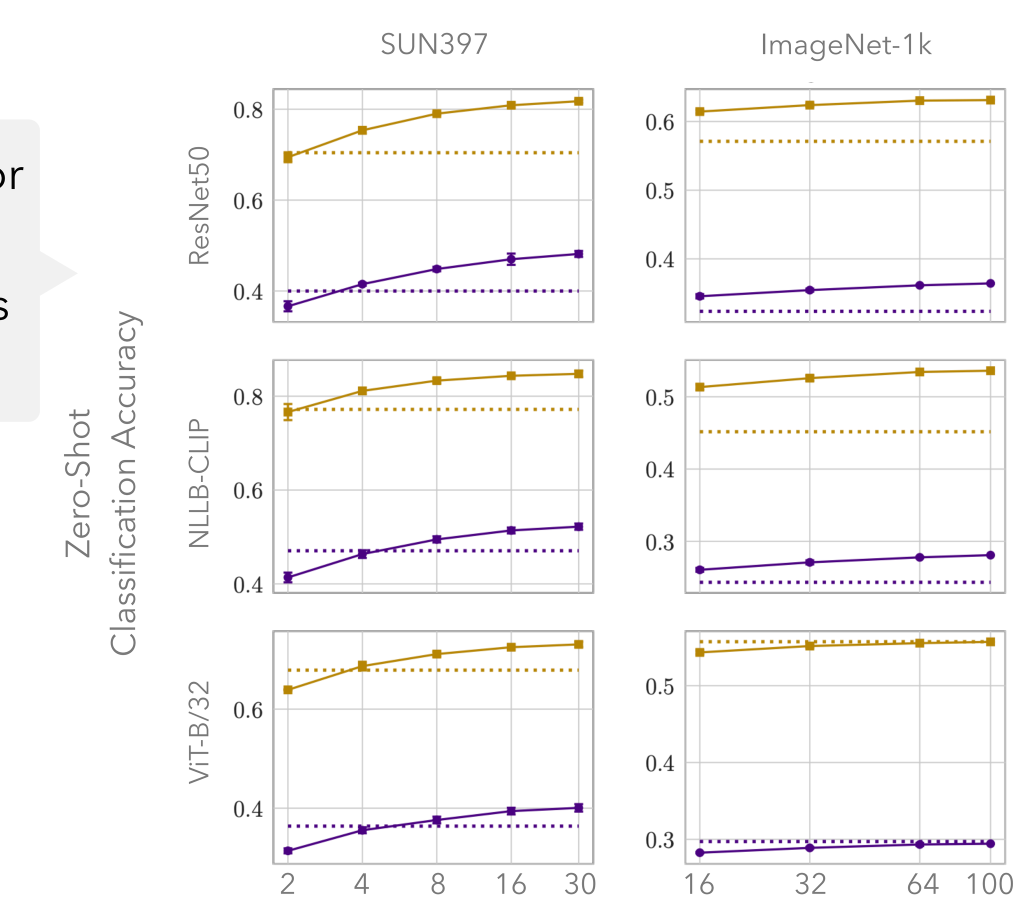


Real Data Examples: Language-Image Pre-Training and Image Classification



When $P_{Y,Z}$ is observed, using held-out prompt examples outperforms templates.

Saturation points for M is often much higher than what is used in practice.



Paper