A Generalization Theory for Zero-Shot Prediction

Ronak Mehta and Zaid Harchaoui

University of Washington · IFML · NSF · Institute for Foundations of Data Science · IFDS · Washington - Wisconsin - Santa Cruz - Chicago

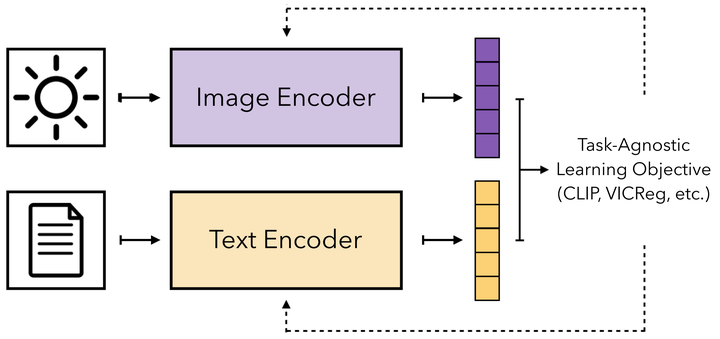
Paper

Zero-Shot Prediction (ZSP)

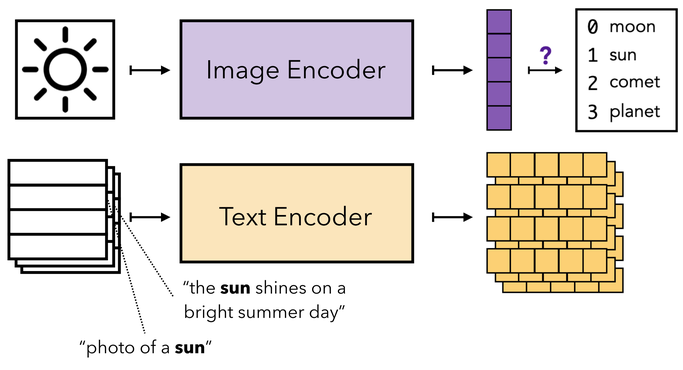
Motivation: Zero-shot prediction is a modern method that reuses foundation models to build classifiers for tasks without seeing *any* directly labeled training data.

Need for theoretical understanding has arisen.

Contrastive Pre-Training



Evaluation



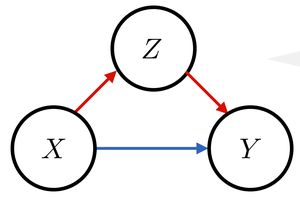
Research Question: How does the downstream performance of ZSP depend on the pre-training distributions, downstream task distribution, and prompting strategy?

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 $P_{X,Y,Z}$ satisfying $P_{Z|X} = Q_{Z|X}$ (i.e., agrees with caption distribution)?

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.

$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le 2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right] + 2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]$$

information-theoretic error · 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.

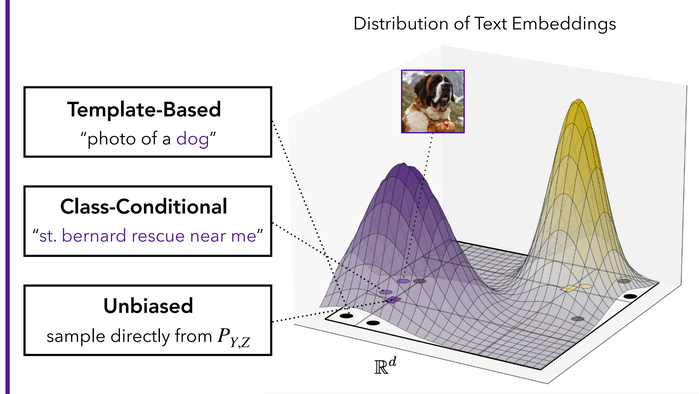
Interpretation: Bias of prompt distribution.

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.

Prompting Strategies

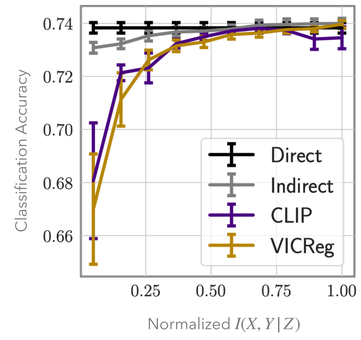


Experiments

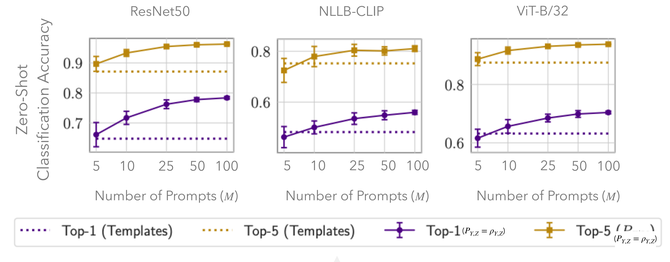
Synthetic Data: Controllable Residual Dependence and Prompt Bias

$$\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{X|\boldsymbol{y}} \\ \boldsymbol{\mu}_{Z|\boldsymbol{y}} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{XX|\boldsymbol{y}} & \mathbf{C}_{XZ|\boldsymbol{y}} \\ \mathbf{C}_{ZX|\boldsymbol{y}} & \mathbf{C}_{ZZ|\boldsymbol{y}} \end{bmatrix}\right)$$

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



Real Data: Language-Image Pre-Training and Zero-Shot Image Classification



When $P_{Y,Z}$ is observable, unbiased prompting strategy outperforms template-based.