Ridge Regression, Hubness, and Zero-Shot Learning

Yutaro Shigeto · Ikumi Suzuki · Kazuo Hara · Masashi Shimbo · Yuji Matsumoto

Zero-shot learning — special type of multi-class classification

Zero-shot learning (ZSL) is a type of classification task in which labels in the training and test sets are disjoint

Standard classification

$Y_{\text{train}} = \{\text{gorilla, lion, tiger}\}$
$Y_{\text{test}} = \{\text{gorilla, lion, tiger}\}$

$Y_{\text{train}} = Y_{\text{test}}$

Zero-shot setting

$Y_{\text{train}} = \{\text{gorilla, lion, tiger}\}$
$Y_{\text{test}} = \{\text{chimpanzee, leopard}\}$

$Y_{\text{train}} \cap Y_{\text{test}} = \emptyset$

Applications: Image labeling, bilingual lexicon extraction, and many other cross-domain matching tasks

Regression-based approach to ZSL

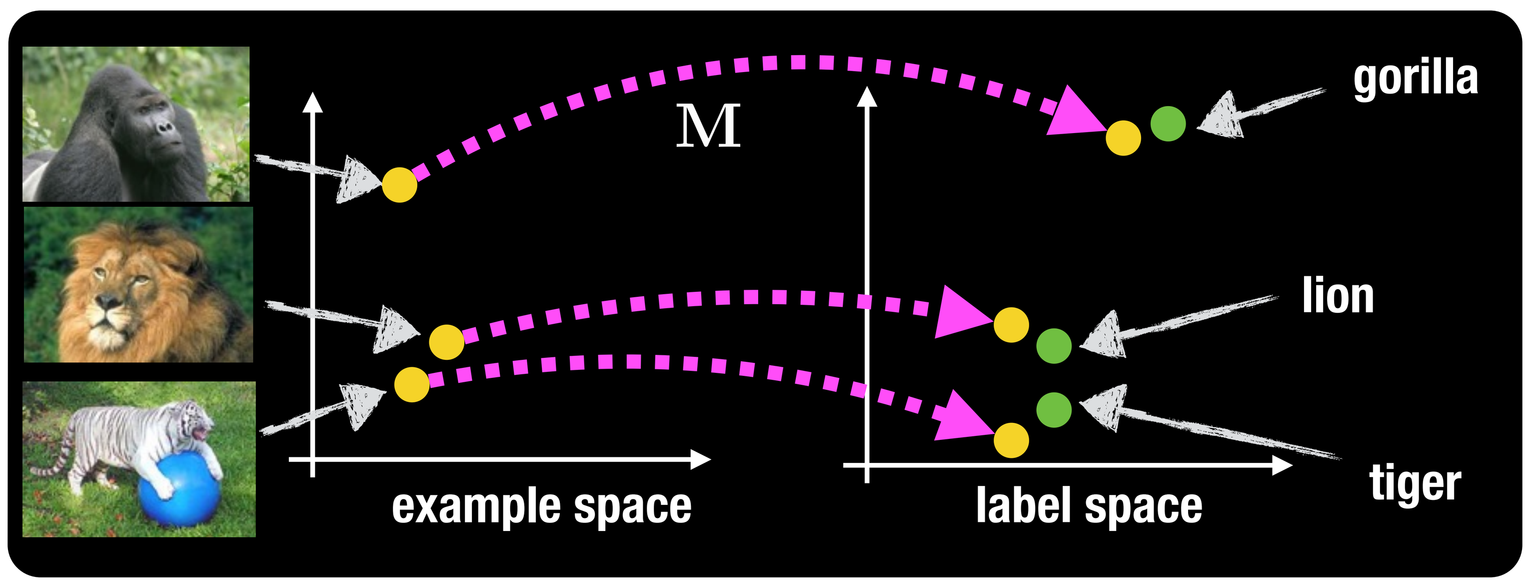
1. Embed labels as vectors in some "label space" Y

$$(\mathbf{x}_i, \mathbf{y}_i) \in X \times Y \qquad i = 1, \cdots, N$$

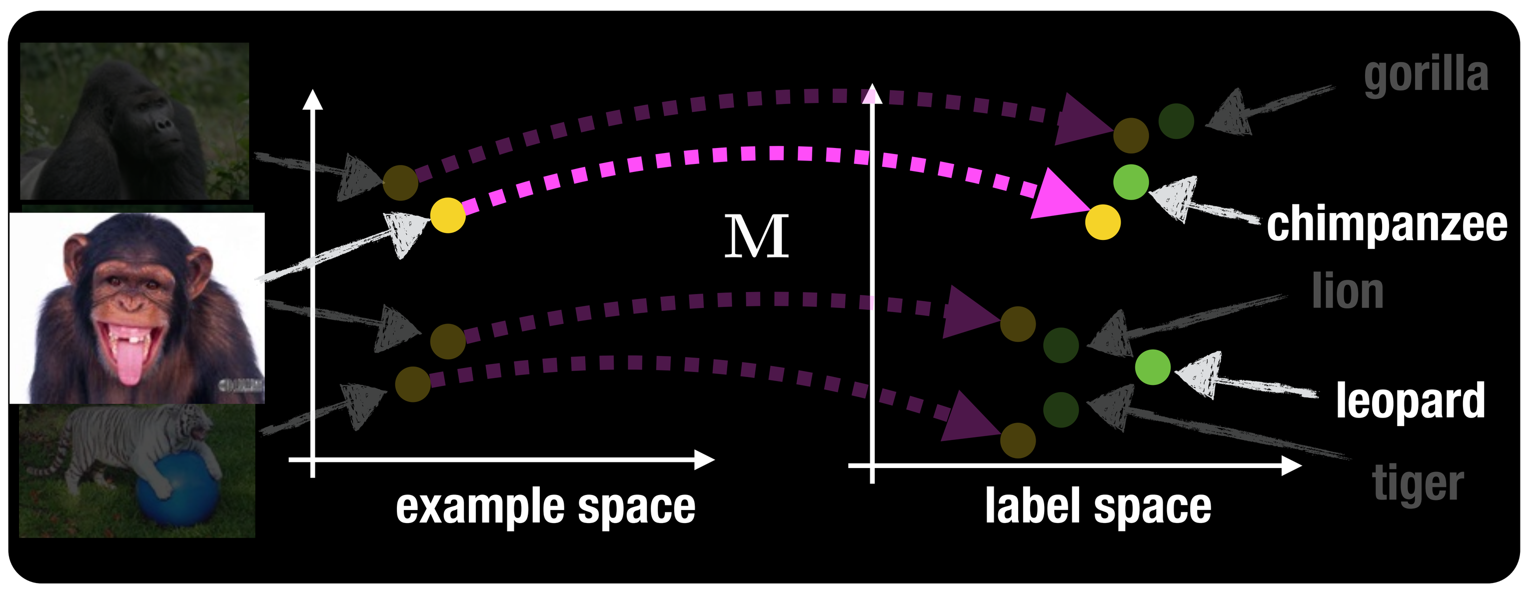
Both examples and labels are vectors

2. **(Training)** Find projection **M**: X →Y such that

$$\min_{\mathbf{M}} \sum_{i=1}^{N} \|\mathbf{M}\mathbf{x}_i - \mathbf{y}_i\|^2 + \lambda\|\mathbf{M}\|_F^2 \qquad \text{(Ridge regression)}$$

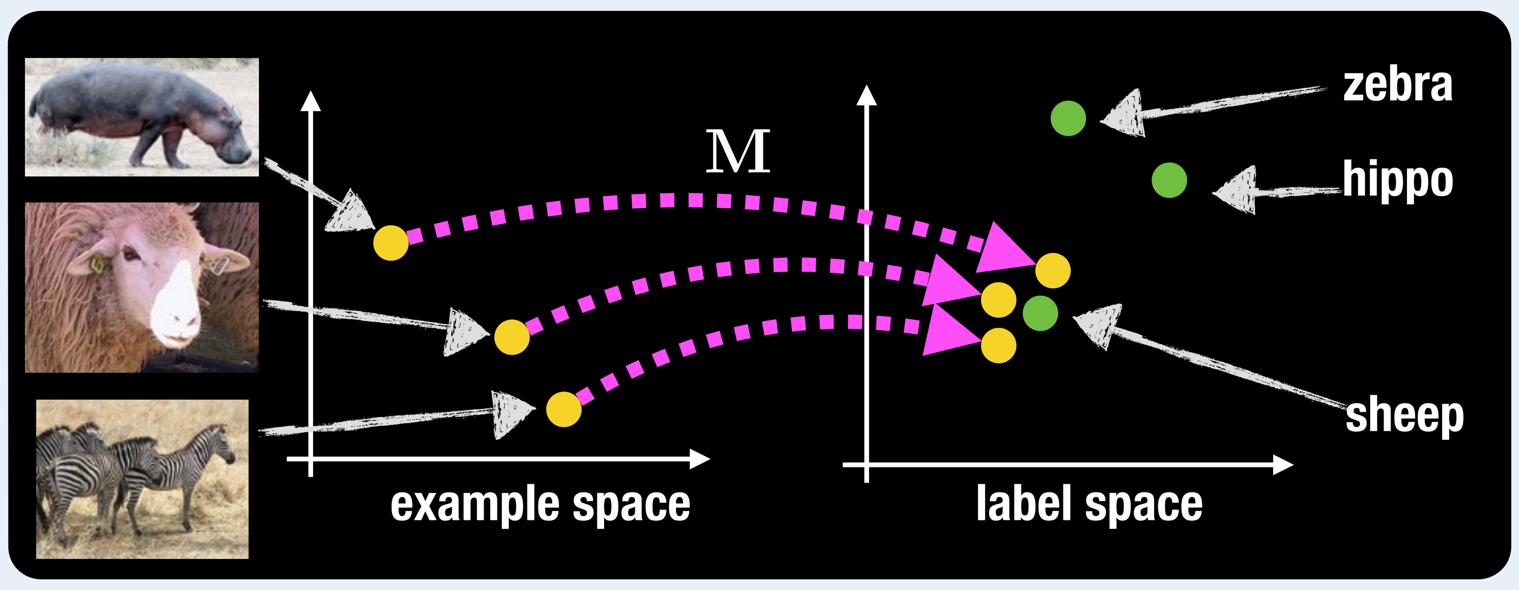


3. **(Prediction)** To predict the class label of a test example $\mathbf{x}$, project it to the label space by **M** and find the nearest label there.



Hubness: problem with the current approach

The learned classifier frequently predicts the same labels regardless of input example = emergence of "**hub**" labels



Empirical evaluation

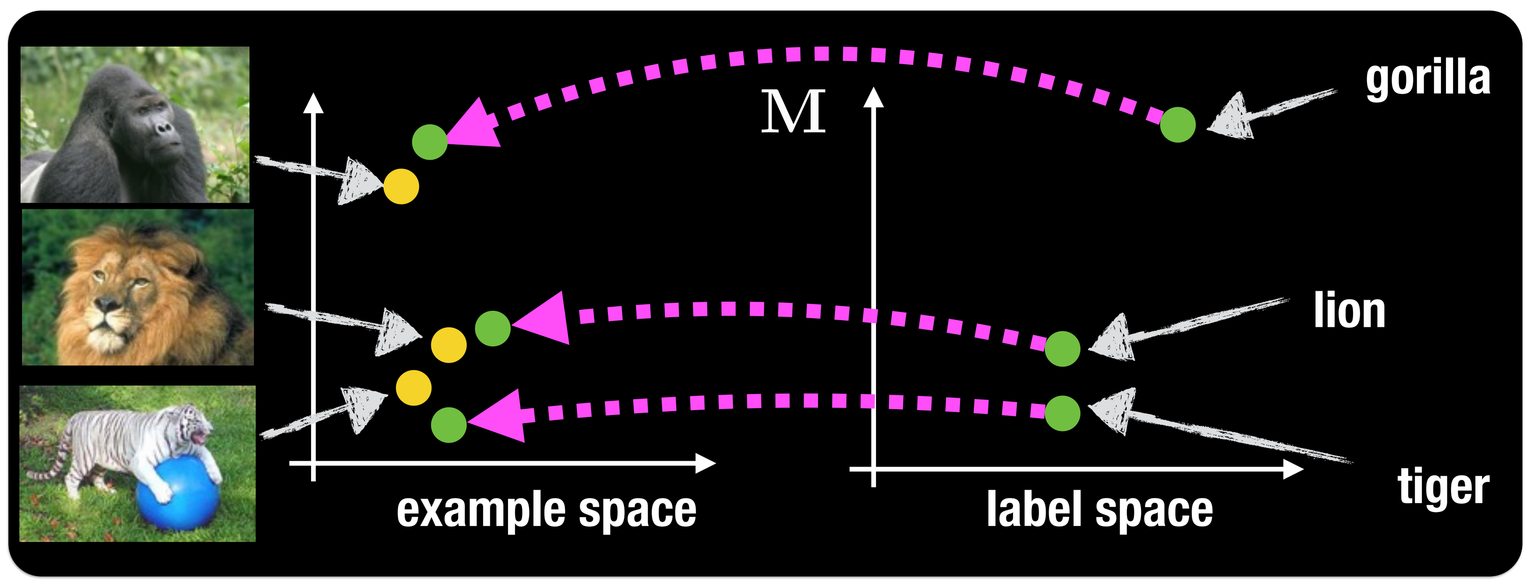
	Image labeling		Bilingual lexicon extraction (fr→en)	
	Accuracy [%]	Hubness	Accuracy [%]	Hubness
Current	22.6	2.61	0.3	67.79
Proposed	41.3	0.08	36.6	2.56

Hubness indicates N_1 skewness

Proposed approach — reverse the mapping direction

Reverse mapping direction (project labels into the example space):

$$\min_{\mathbf{M}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{M}\mathbf{y}_i\|^2 + \lambda\|\mathbf{M}\|_F^2$$



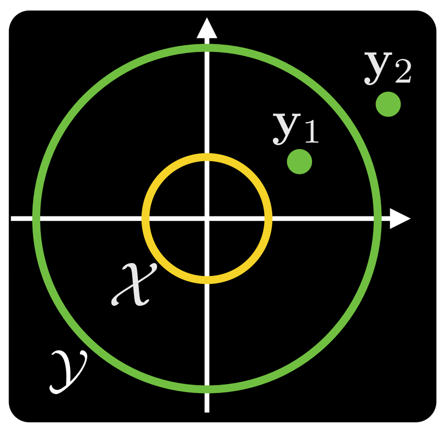
Then, using M, project all test labels into the example space. When a test example is given, find the nearest (projected) label in the example space.

Why proposed method reduces hubness

Hubness and variance of label objects

Assume

- example distribution: $\mathcal{X}$ = any distribution with zero mean
- label distribution: $\mathcal{Y} = \mathcal{N}(\mathbf{0}, s^2\mathbf{I})$
- two fixed objects: $\mathbf{y}_1$ and $\mathbf{y}_2$ such that
  $$\|\mathbf{y}_2\|^2 - \|\mathbf{y}_1\|^2 = \sqrt{\mathrm{Var}_{\mathcal{Y}}[\|\mathbf{y}\|^2]} > 0$$



Then,

$$\mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_2\|^2] - \mathrm{E}_{\mathcal{X}}[\|\mathbf{x} - \mathbf{y}_1\|^2] = s^2\sqrt{2d} > 0 \quad (*)$$

Implication:

- $\mathbf{y}_1$ is more likely to be closer to $\mathbf{x}$: i.e, more likely to be a hub
- Quantity (*) can be interpreted as the degree of bias in the data which makes objects closest to the origin hubs
- In particular, the smaller the variance s^2, the smaller the bias (*)

For a fixed $\mathcal{X}$, distribution $\mathcal{Y}$ with smaller variance s^2 is preferable in order to reduce hubs

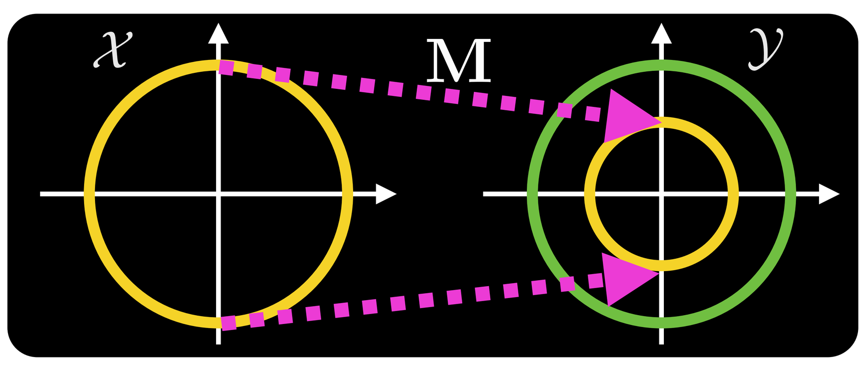
Shrinkage of projected objects

If we optimize

$$\min_{\mathbf{M}} \|\mathbf{M}\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{M}\|_F^2$$

then,

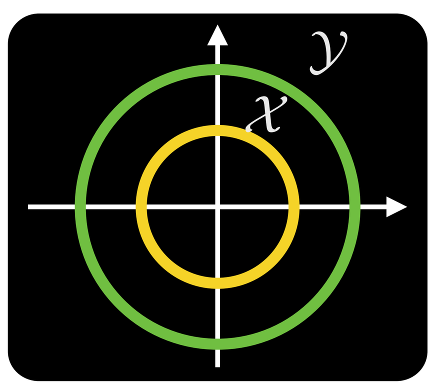
$$\|\mathbf{M}\mathbf{X}\|_2 \le \|\mathbf{Y}\|_2$$



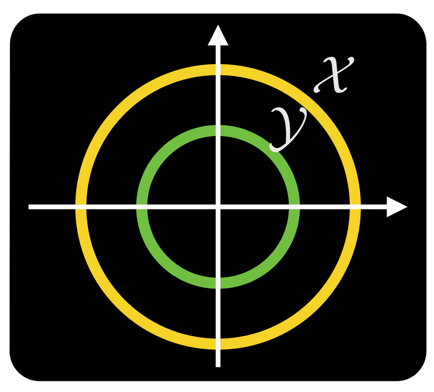
Projected objects tend to lie closer to the origin

Example-label configuration after projection

Current: map $\mathbf{x}$ into space Y



Proposed: map $\mathbf{y}$ into space X



Variance of $\mathcal{Y}$ (relative to $\mathcal{X}$) smaller with the proposed approach

Proposed approach is less biased to produce hubs