Representer Point Selection for Explaining Deep Neural Networks

Chih-Kuan Yeh, Joon Sik Kim,
Ian E.H. Yen, Pradeep Ravikumar
CMU

NIPS-2018

what's about

for a given test point prediction, select **representer** points in the training set which can explain the predictions

Representer point

Φ: model
Θ: parameter of model
x_t: testing data
L: loss
f_i: feature of training data
f_t: feature of testing data

Prediction of test data

$$\Phi(\mathbf{x}_t, \mathbf{\Theta}^*) = \sum_i^n k(\mathbf{x}_t, \mathbf{x}_i, \alpha_i),$$

Decomposition by theorem

$$\alpha_i = \frac{1}{-2\lambda n} \frac{\partial L(\mathbf{x}_i, \mathbf{y}_i, \mathbf{\Theta})}{\partial \Phi(\mathbf{x}_i, \mathbf{\Theta})}$$

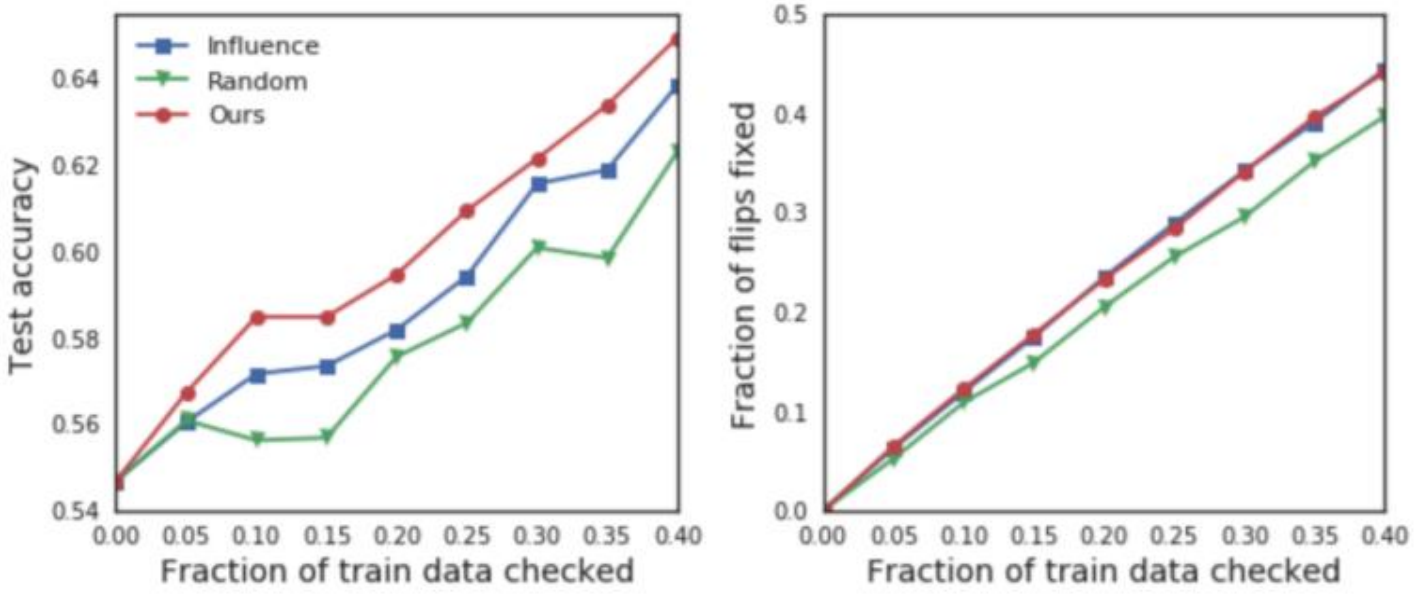
α_i: importance of i-th data on the parameter

$$k(\mathbf{x}_t, \mathbf{x}_i, \alpha_i) = \alpha_i \mathbf{f}_i^T \mathbf{f}_t,$$

Feature similarity

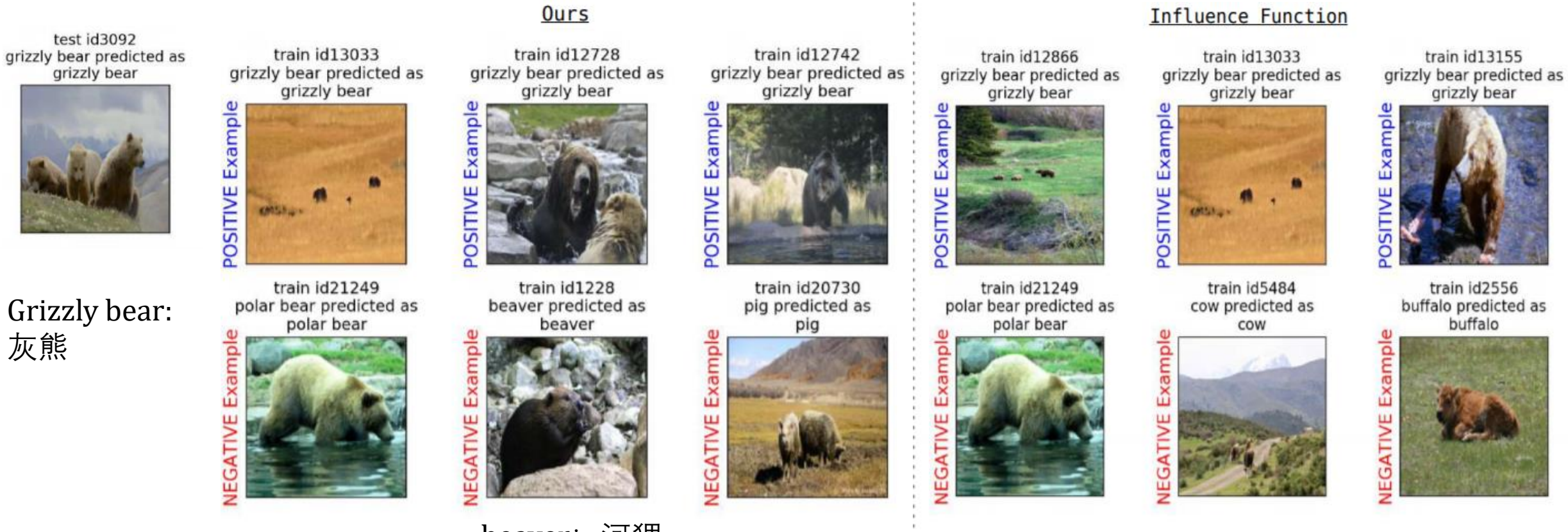
$k(x_t, x_i, \alpha_i)$: representer value for x_i given x_t, k_j represents x_i's impact on class j

Dataset Debugging



40 percent of the data has the label flipped.
find mislabeled data, retrain.

Excitatory(positive) and Inhibitory(negative) Examples

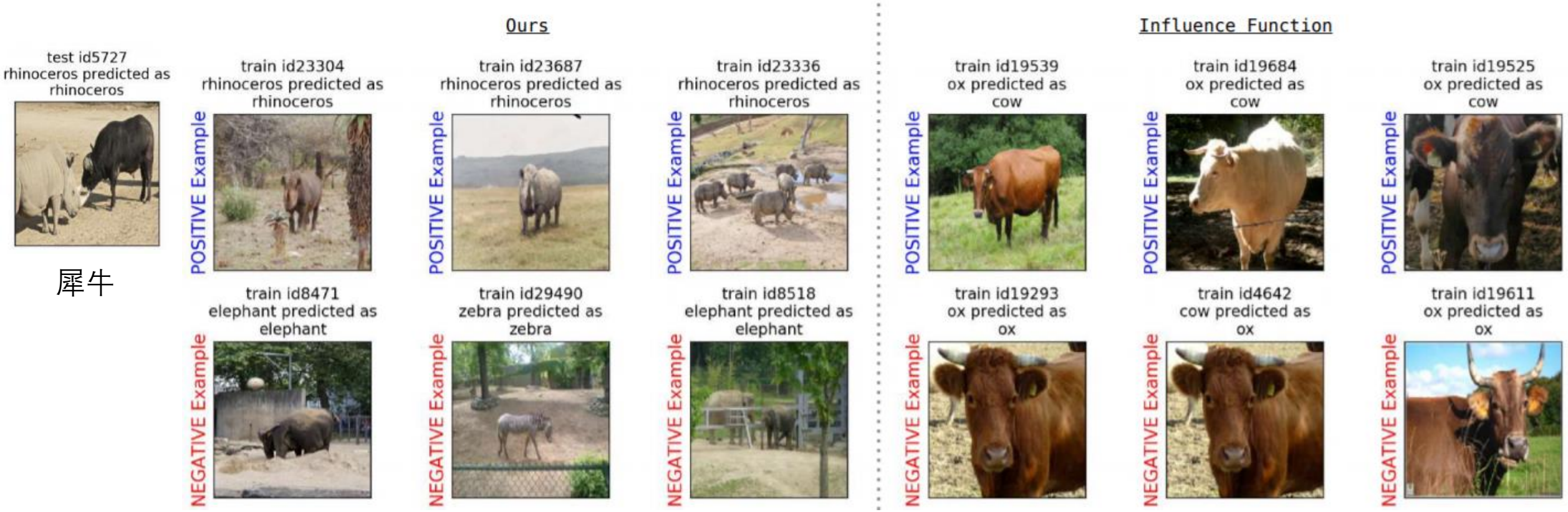


Grizzly bear: 灰熊

beaver：河狸

Postive: similar and same label
Negative: similar but different label

Excitatory and Inhibitory Examples



犀牛

Ox: 牛

Misclassified Examples

test id7
predicted as deer
true label is antelope

train id29372
predicted as zebra
true label is zebra

train id688
predicted as deer
true label is antelope

train id8090
predicted as elephant
true label is elephant

train id29208
predicted as zebra
true label is zebra

羚羊被标记为鹿

All of these pictures have deer,
but labeld different class

Example selection for dictionary learning

Tomoki Tsuchida ,Garrison W. Cottrell
UCSD

ICLR-2015

Dictionary Learning

Given the input dataset $X = [x_1 \dots x_K]$, $x_i \in R^d$, we wish to find a dictionary $D \in R^{d*n}$: $D = [d_1, d_2 \dots d_n]$, and a representation $R = [r_1, r_2 \dots r_k]$, $r_i \in R^n$ such that both $\|x_i - Dr_i\|_F^2$ is minimized and the representations r_i are sparse enough.

This can be formulated as the following optimization problem:

$$\underset{\mathbf{D}\in\mathcal{C},r_i\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{K} \|x_i - \mathbf{D}r_i\|_2^2 + \lambda\|r_i\|_0$$

We want to use less data to learn dictionary

Score function

Err: $g_j\left(r^{(i)}, x^{(i)}\right) = \left\| \widehat{D} r^{(i)} - x^{(i)} \right\|_1$

Grad: $g_j\left(r^{(i)}, x^{(i)}\right) = \left\| \widehat{D} r^{(i)} - x^{(i)} \right\|_1 * \mathrm{r_j^{(i)}}$

SNR: $g_j\left(r^{(i)}, x^{(i)}\right) = \frac{\|x^i\|_2^2}{\|\widehat{D} r^i - x^i\|_2^2} * r_j^{(i)}$, 信噪比

SUN: $g_j\left(r^{(i)}, x^{(i)}\right) = r_j^{(i)}$

SalMap: $g_j\left(r^{(i)}, x^{(i)}\right) = SaliencyMap(x^{(i)})$

From other papers

Selector function

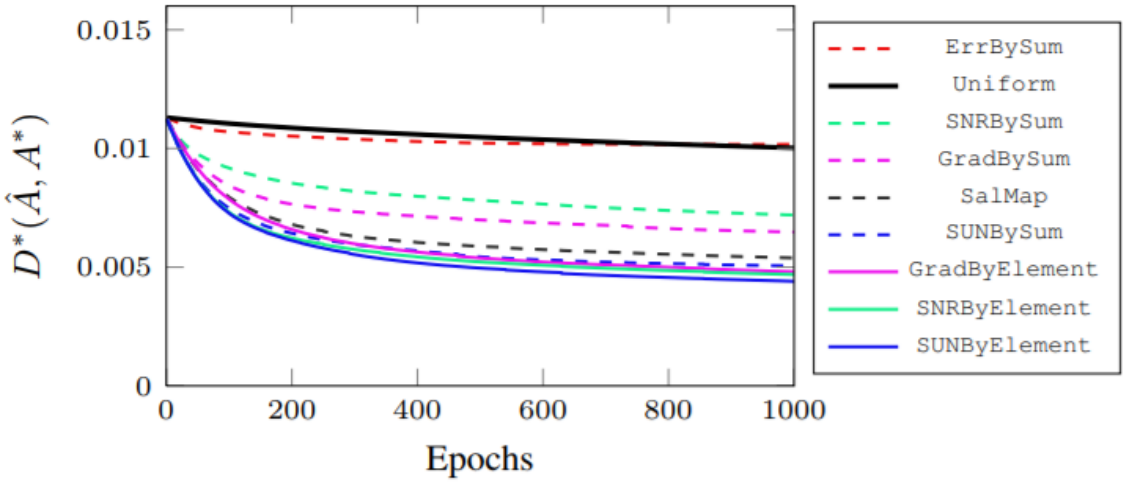
$BySum: f_{sel} = top\ n\ elements\ of \sum_{j=1}^{n} g_j^i$

计算每个element的分值，最后加总，选出总分最高的样本

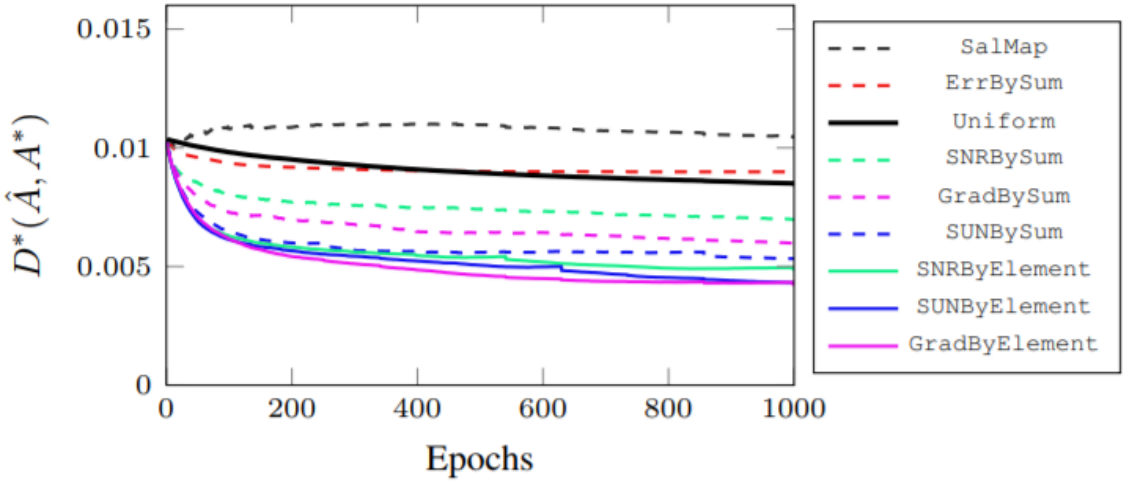
$ByElement: f_{sel} = \{top\ k\ elements\ of\ g_j^i | j \in 1 \ldots n\}$

针对每个element计算分值，选出该element分值最高的几个样本

Distance from true dictionary



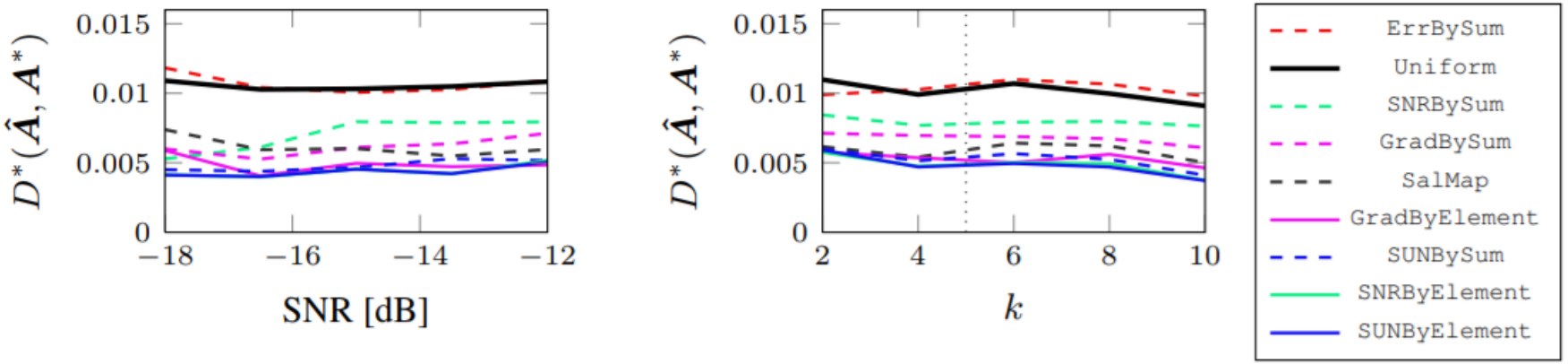
(a) Gabor dictionary

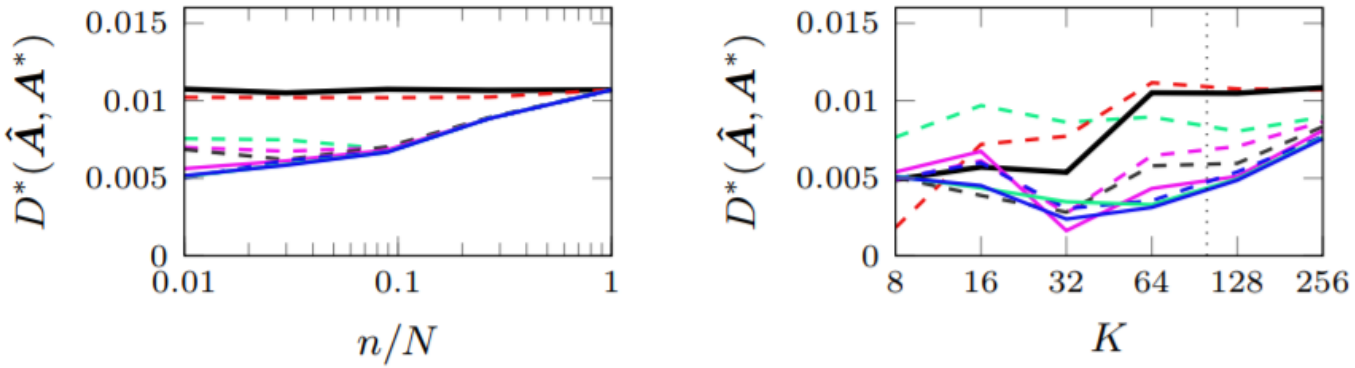


(b) Alphanumeric dictionary

robustness

SNR: signal to noise ratio $(10\log_{10}(2\lambda^2/\sigma_\epsilon^2)$
k: number of nonzero elements
n/N: ratio of selected examples to the original training set
K: number of dictionary elements





Algorithm is robust