EVALUATION OF SIMILARITY-BASED EXPLANATIONS

Kamil Grudzień
Krystian Sztenderski

Plan

1. What are similarity based explanations
2. What similarity metrics are used
3. Evaluation criterias
4. Results

What are similarity based explanations

What are similarity based explanations

- Instance-based
- Similar instances as evidence to support model prediction
- "I(the model) think this image is cat because similar images I saw in the past were also cat"
- Very simple to understand without machine learning expertise
- Better than SHAP or LIME?

$$D = \{z_{train}^{(i)} = (x_{train}^{(i)}, y_{train}^{(i)})\}_{i=1}^{N}$$

$$z_{test} = (x_{test}, f(x_{test}))$$

$$\bar{z} = argmax_{z_{train} \in D} R(z_{test}, z_{train})$$

What similarity measures are used

What similarity metrics are used

- **L2 Metric:** $R_{\ell_2}(\boldsymbol{z}, \boldsymbol{z}') := -\|\phi(\boldsymbol{z}) - \phi(\boldsymbol{z}')\|^2$
- **Cosine Metric:** $R_{\cos}(\boldsymbol{z}, \boldsymbol{z}') := \cos(\phi(\boldsymbol{z}), \phi(\boldsymbol{z}'))$
- **Dot Metric:** $R_{\mathrm{dot}}(\boldsymbol{z}, \boldsymbol{z}') := \langle \phi(\boldsymbol{z}), \phi(\boldsymbol{z}') \rangle$

ϕ can be any from:

$\phi(\boldsymbol{z}) = \boldsymbol{x}$ $\qquad$ $\phi(\boldsymbol{z}) = \boldsymbol{h}^{\mathrm{last}}$ $\qquad$ $\phi(\boldsymbol{z}) = \boldsymbol{h}^{\mathrm{all}}$

Not only so simple- Gradient-based metrics

- **IF**: $R_{\mathrm{IF}}(\boldsymbol{z}, \boldsymbol{z}') := \langle \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}}, \boldsymbol{H}^{-1} \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}'} \rangle$
- **RIF**: $R_{\mathrm{RIF}}(\boldsymbol{z}, \boldsymbol{z}') := \cos(\boldsymbol{H}^{-\frac{1}{2}} \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}}, \boldsymbol{H}^{-\frac{1}{2}} \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}'})$
- **FK**: $R_{\mathrm{FK}}(\boldsymbol{z}, \boldsymbol{z}') := \langle \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}}, \boldsymbol{I}^{-1} \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}'} \rangle$,
- **GD**: $R_{\mathrm{GD}}(\boldsymbol{z}, \boldsymbol{z}') := \langle \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}}, \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}'} \rangle$
- **GC**: $R_{\mathrm{GC}}(\boldsymbol{z}, \boldsymbol{z}') := \cos(\boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}}, \boldsymbol{g}_{\widehat{\boldsymbol{\theta}}}^{\boldsymbol{z}'})$

Where ***H*** and ***I*** are the Hessian and Fisher information matrices of the loss $\mathcal{L}_{train}$

Evaluation criterias

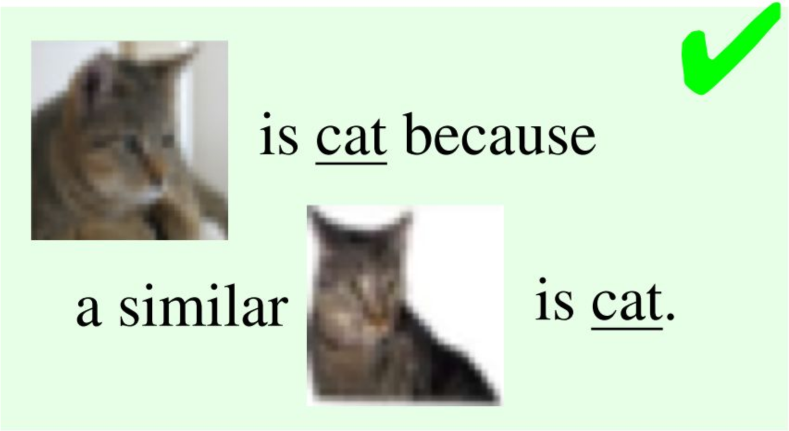
Model Randomization Test

$$R(z_{test}, z_{train}^{(\pi_f(1))}) \geq R(z_{test}, z_{train}^{(\pi_f(2))}) \geq \dots \geq R(z_{test}, z_{train}^{(\pi_f(N))})$$

- R - relevance metric
- π_f - permutation of indices with decreasing relevance
- π_f and π_{f_rand} must have a small rank correlation
- Checks faithfulness

Identical Class Test

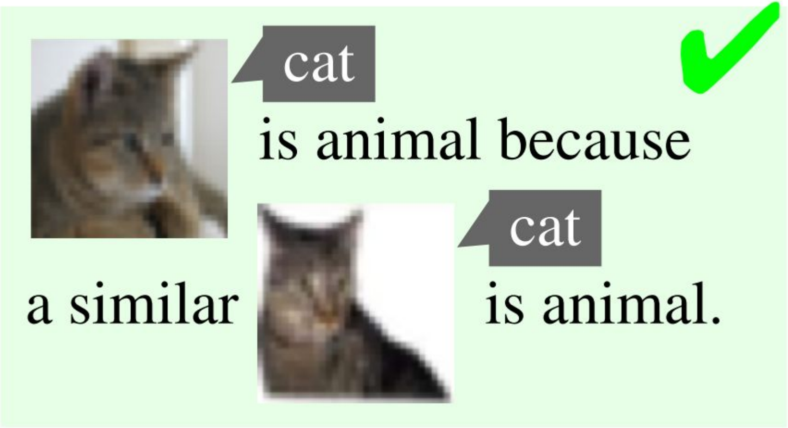
$$\underset{\boldsymbol{z}=(\boldsymbol{x},y)\in\mathcal{D}}{\arg\max}\ R\big(\boldsymbol{z}_{\text{test}}, \boldsymbol{z}\big) = (\bar{\boldsymbol{x}}, \bar{y}) \implies \bar{y} = \widehat{y}_{\text{test}}$$



Identical Subclass Test

s(z) is a subclass for class ***y*** $s(z) \subset y$

$$argmax_{z_{train} \in D} R(z_{test}, z_{train}) = \bar{z} \Rightarrow s(\bar{z}) = s(z_{test})$$



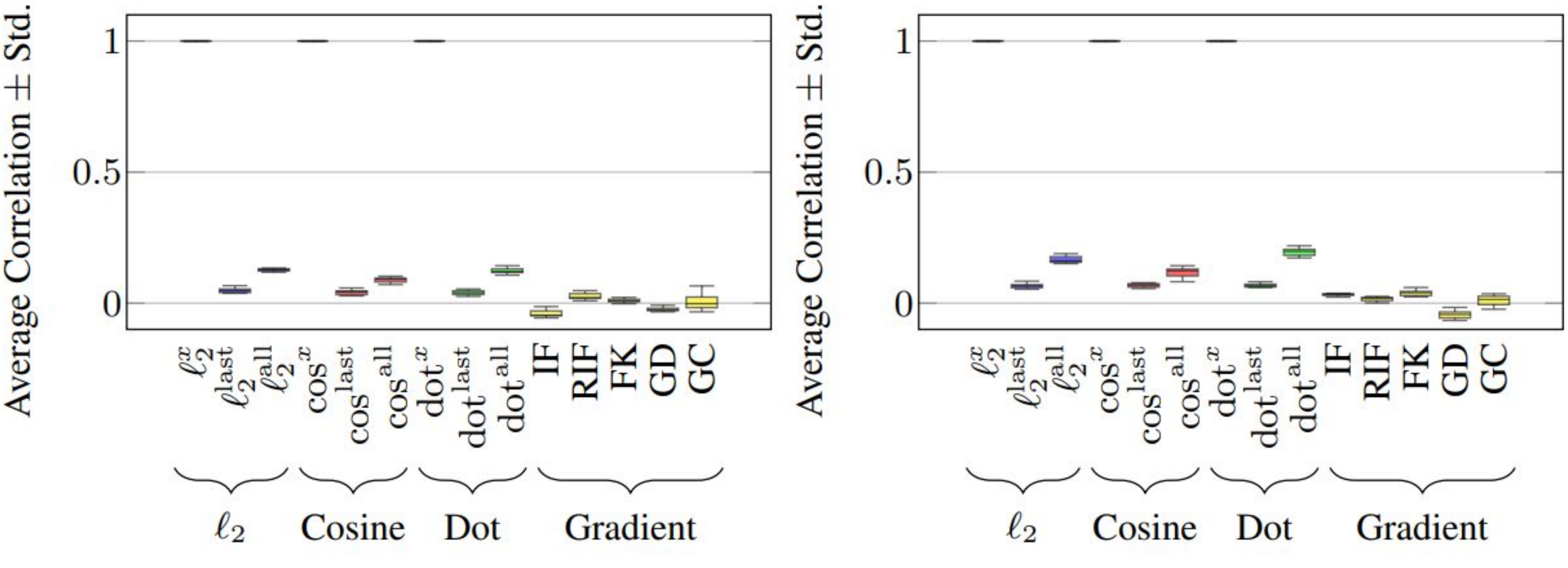
Dataset split

Class A

Class B

Results

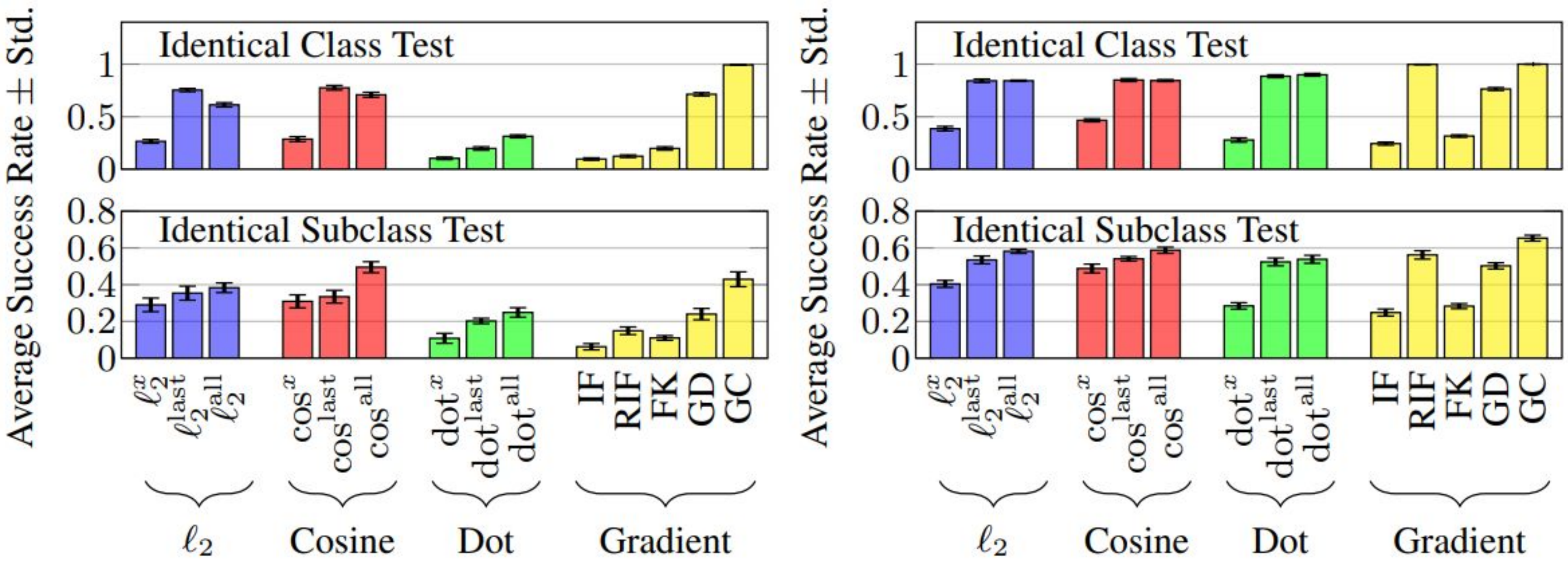
Model randomization test



(a) CIFAR10 with CNN

(b) AGNews with Bi-LSTM

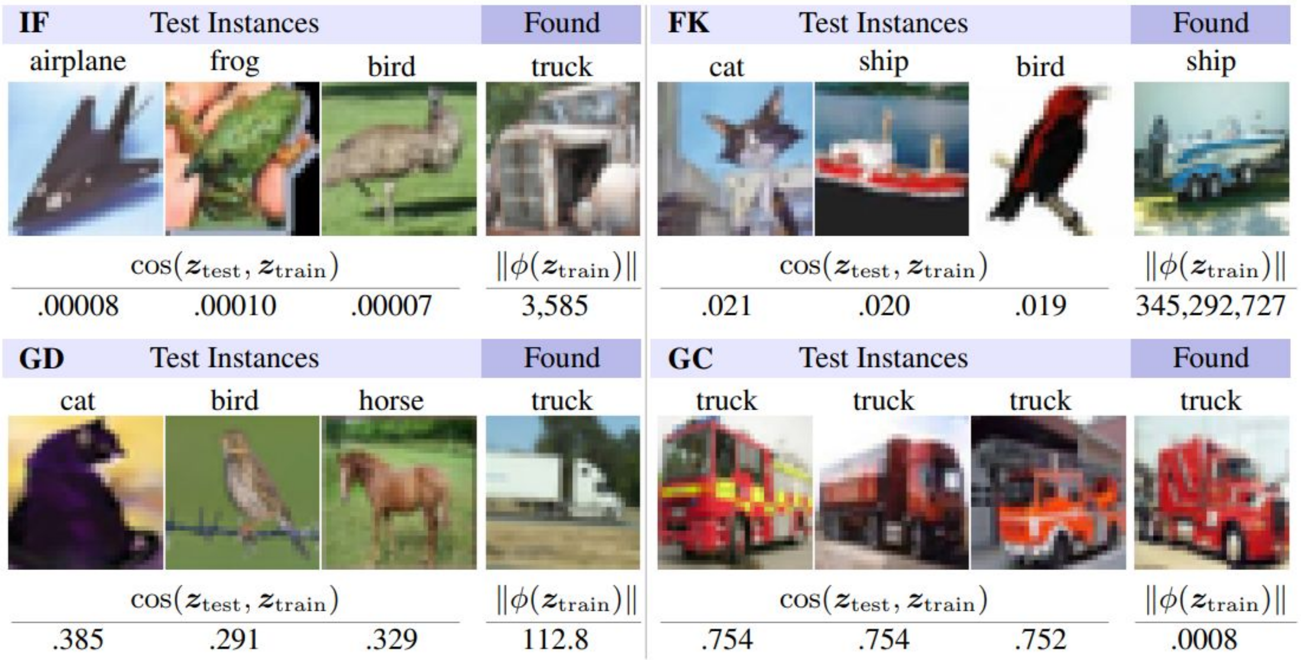
Identical class and subclass test



(a) CIFAR10 with CNN

(b) AGNews with Bi-LSTM

Frequently selected training instances

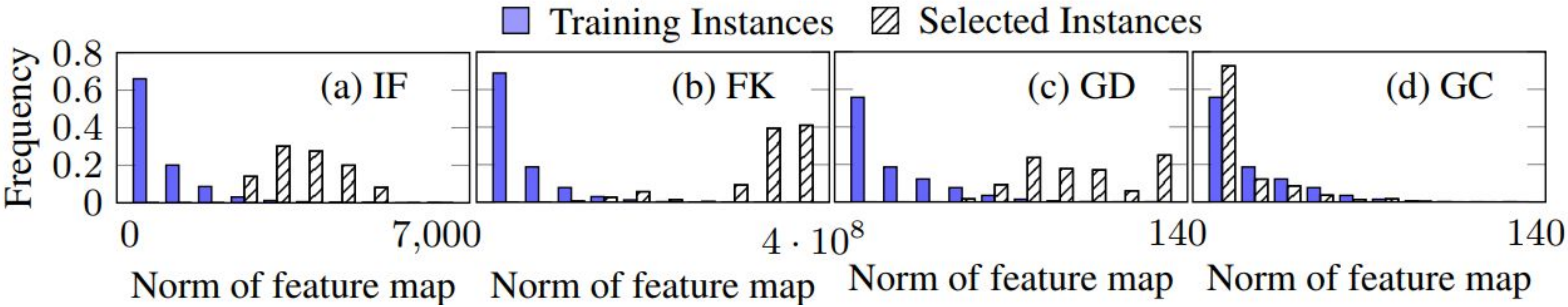


Failure of dot metrics & gradient based metrics

$$\langle \phi(\boldsymbol{z}_{\text{test}}), \phi(\boldsymbol{z}_{\text{train}}^{(i)}) \rangle < \langle \phi(\boldsymbol{z}_{\text{test}}), \phi(\boldsymbol{z}_{\text{train}}^{(j)}) \rangle$$

$$\|\phi(\boldsymbol{z}_{\text{train}}^{(i)})\| < \|\phi(\boldsymbol{z}_{\text{train}}^{(j)})\| \cos(\phi(\boldsymbol{z}_{\text{test}}), \phi(\boldsymbol{z}_{\text{train}}^{(j)}))$$

Frequently selected training instances



Takeaways

- 3 minimal requirement tests:

  Model Randomization Test Identical Class Test Identical Subclass Test

- $\ell_2^{\text{last}}, \cos^{\text{last}}$ and **gradient-based** metrics **passed** the Model Randomization Test
- **Dot metrics** as well as **IF, FK and GD** **failed** the Identical Class and Subclass Test
- **GC** performed **the best** in most of the tests – **recommended** method

Thanks for your attention