Evaluation of Similarity-based Explanations

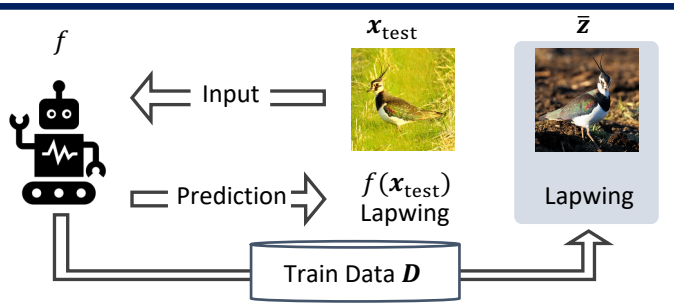
Kazuaki Hanawa[1,2], Sho Yokoi[2,1], Satoshi Hara[3], Kentaro Inui[2,1]

[1]RIKEN AIP, [2]Tohoku University, [3]Osaka University

Background

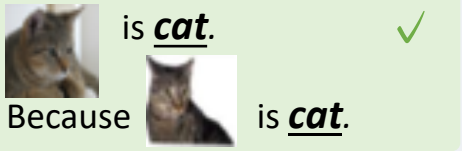
- **Similarity-based explanation**
  - Present a similar instance $\bar{\boldsymbol{z}} = \mathrm{argmax}_{\boldsymbol{z}_i \in \boldsymbol{D}} R(\boldsymbol{z}_{\mathrm{test}}, \boldsymbol{z}_i)$
- **Relevance metric** $R(\boldsymbol{z}_{\mathrm{test}}, \boldsymbol{z}_i) \in \mathbb{R}$
  - Quantify the relationship between two instances

$\boldsymbol{D}$: Training data
f: Classification model
$\boldsymbol{z}_{\mathrm{test}} = (\boldsymbol{x}_{\mathrm{test}}, f(\boldsymbol{x}_{\mathrm{test}}))$
$\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i) \in D$

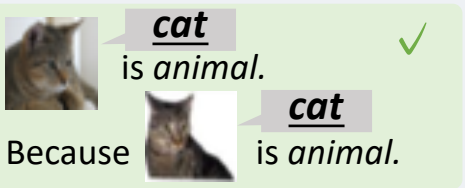


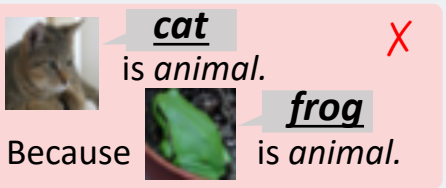
Evaluation Criteria

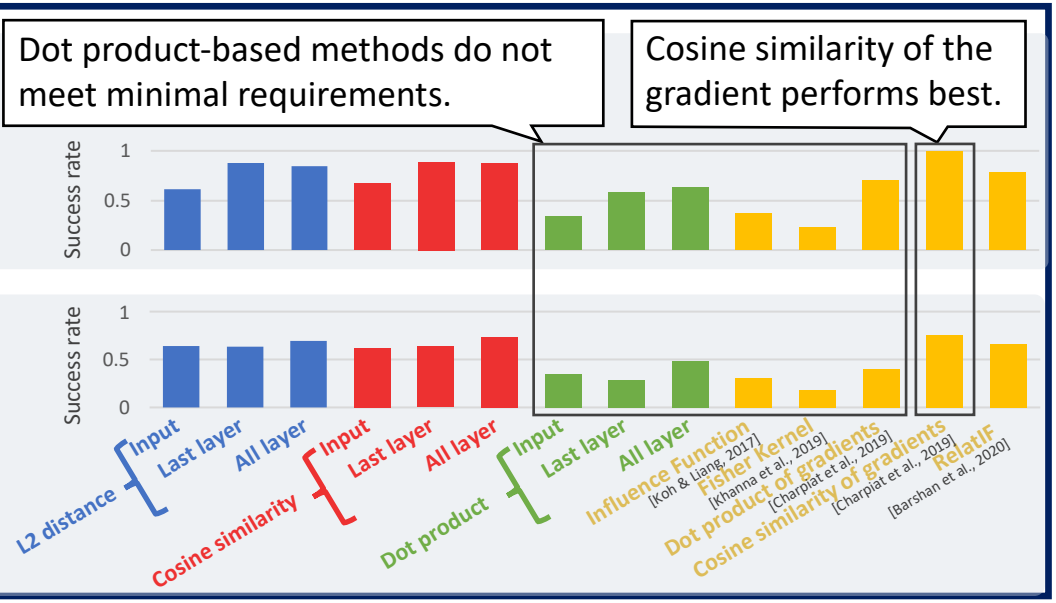
Identical class test



Identical subclass test







Analysis of the Failure

- Dot product-based methods do not work well
- Some instances are judged as similar to various test instances because of **the large norm**

Example of **Dot product of gradients** $\langle \boldsymbol{g}_{\mathrm{test}}, \boldsymbol{g}_i \rangle$

$\boldsymbol{g}_{\mathrm{test}}$: Gradient of the test instance
$\boldsymbol{g}_i$: Gradient of the i-th training instance