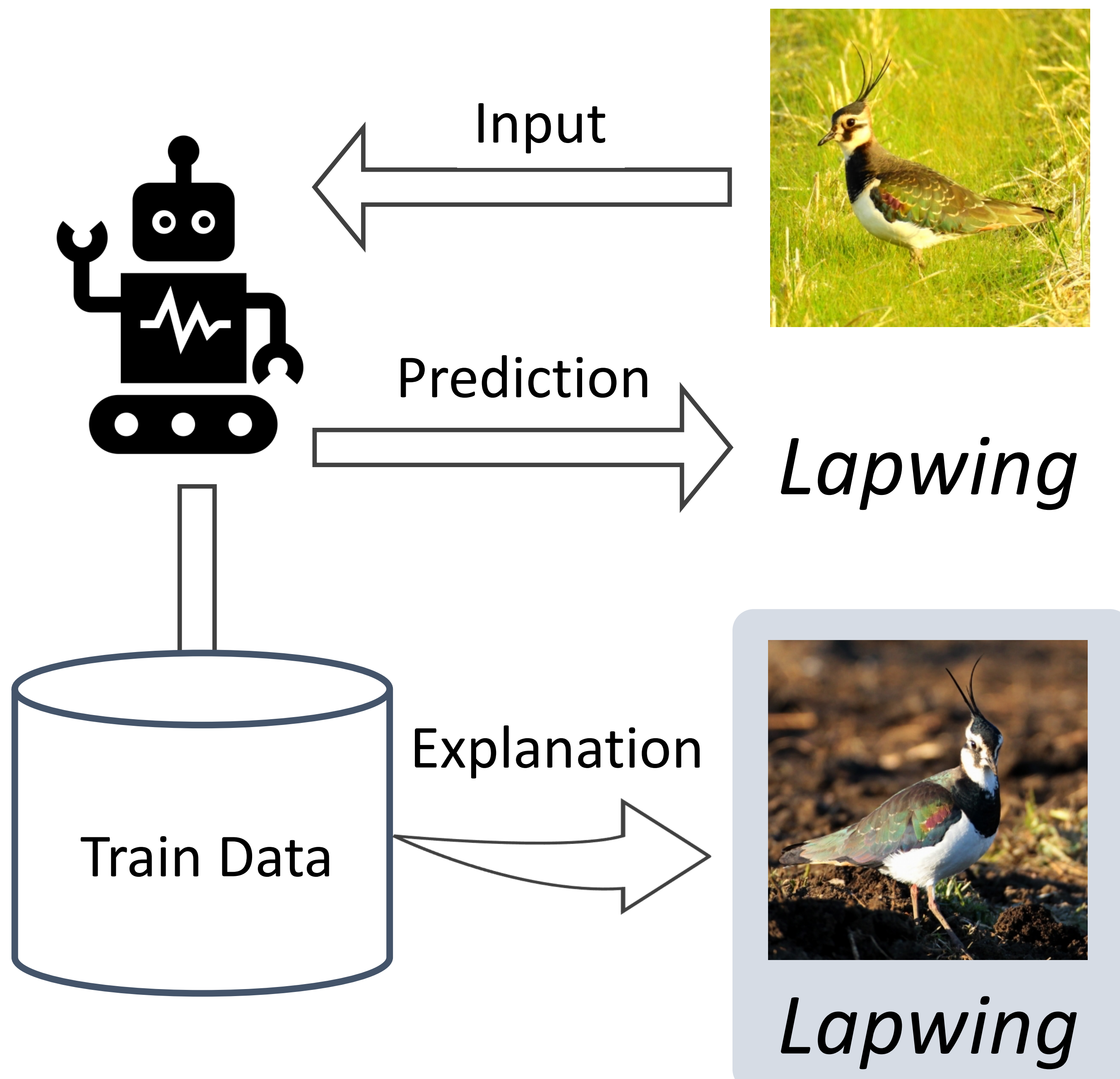
# Evaluation of Similarity-based Explanations

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

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

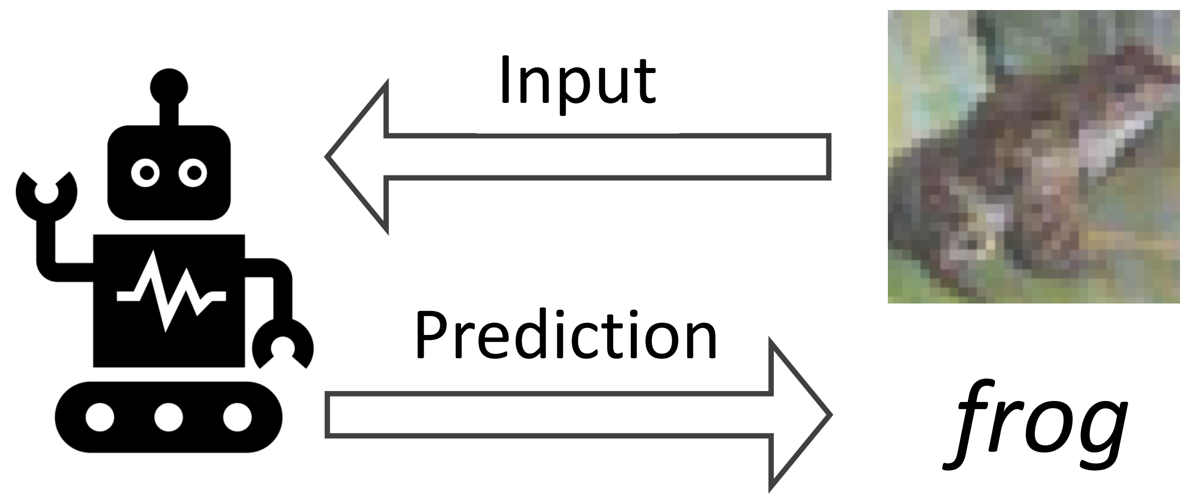
# Background: Similarity-based Explanation

- Explanation by “presenting similar examples” [Charpiat+, 2019; Barshan+, 2020]



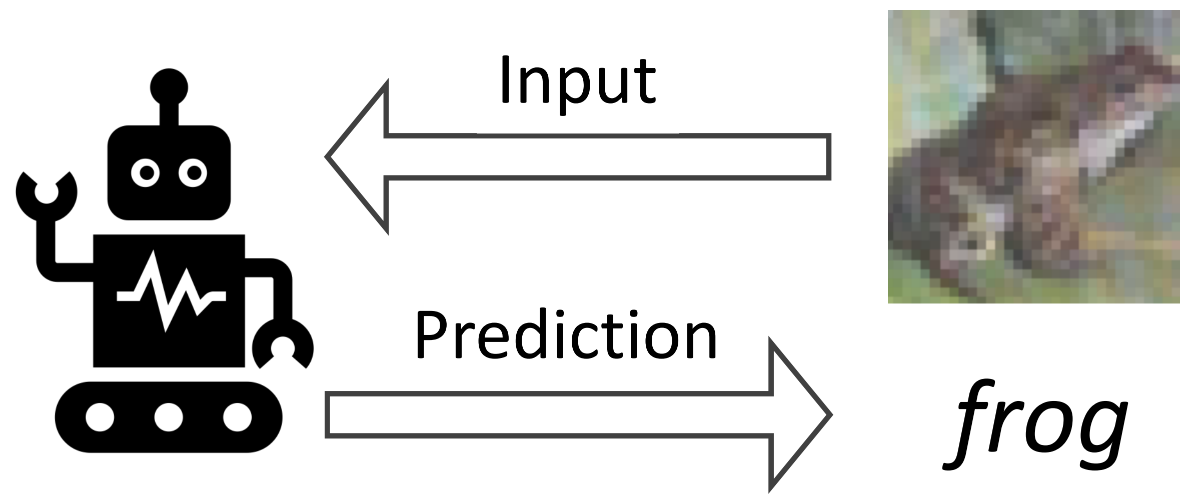

Present a similar training instance
as the reason for the prediction

# Can existing methods provide reasonable explanations?

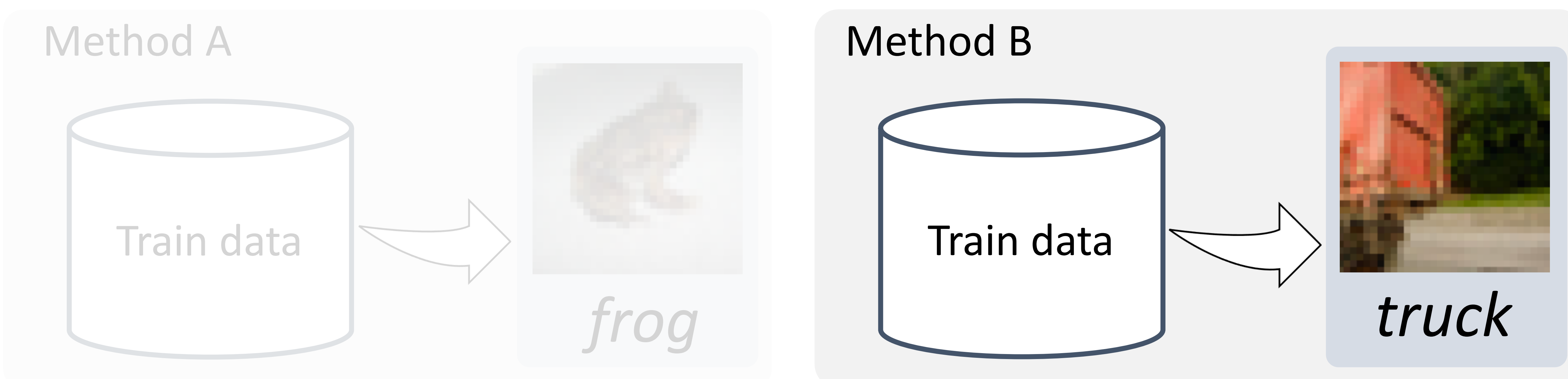



The reason for "predicting this image to be a *frog*" is …

# Can existing methods provide reasonable explanations?




The reason for "predicting this image to be a *frog*" is …




- The instance obtained by Method B (*truck*) will not be convincing.

# Contributions: Investigating appropriate explanation methods

- Evaluate the similarity-based explanation with three tests from two perspectives
- Explanations need to be plausible and faithful [Jacovi & Goldberg, 2020].

# Contributions: Investigating appropriate explanation methods

- Evaluate the similarity-based explanation with three tests from two perspectives
- Explanations need to be plausible and faithful [Jacovi & Goldberg, 2020].
  - Perspective 1: **Plausibility** [Lei+, 2016; Lage+, 2019; Strout+, 2019]
    - Explanation must be convincing to humans.
    - Test 1: **Identical class test**
    - Test 2: **Identical subclass test**

# Contributions: Investigating appropriate explanation methods

- Evaluate the similarity-based explanation with three tests from two perspectives
- Explanations need to be plausible and faithful [Jacovi & Goldberg, 2020].
  - Perspective 1: **Plausibility** [Lei+, 2016; Lage+, 2019; Strout+, 2019]
    - Explanation must be convincing to humans.
    - Test 1: **Identical class test**
    - Test 2: **Identical subclass test**
  - Perspective 2: **Faithfulness** [Adebayo+, 2018; Lakkaraju+, 2019; Jacovi & Goldberg, 2020]
    - Explanation must reflect the underlying inference process.
    - Test 3: **Randomization test**

# Contributions: Investigating appropriate explanation methods

- Evaluate the similarity-based explanation with three tests from two perspectives
- Explanations need to be plausible and faithful [Jacovi & Goldberg, 2020].
  - Perspective 1: **Plausibility** [Lei+, 2016; Lage+, 2019; Strout+, 2019]
    - Explanation must be convincing to humans.
    - Test 1: **Identical class test**
    - Test 2: **Identical subclass test**
  - Perspective 2: **Faithfulness** [Adebayo+, 2018; Lakkaraju+, 2019; Jacovi & Goldberg, 2020]
    - Explanation must reflect the underlying inference process.
    - Test 3: **Randomization test**

# Identical Class Test

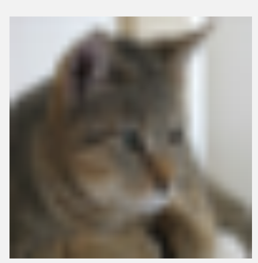
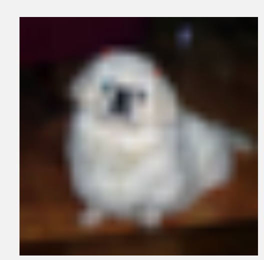
- Check if **the predicted class** and **the presented class** are the same
- Evaluate the plausibility of the explanation

Example of CIFAR-10

Test instance | Training instance

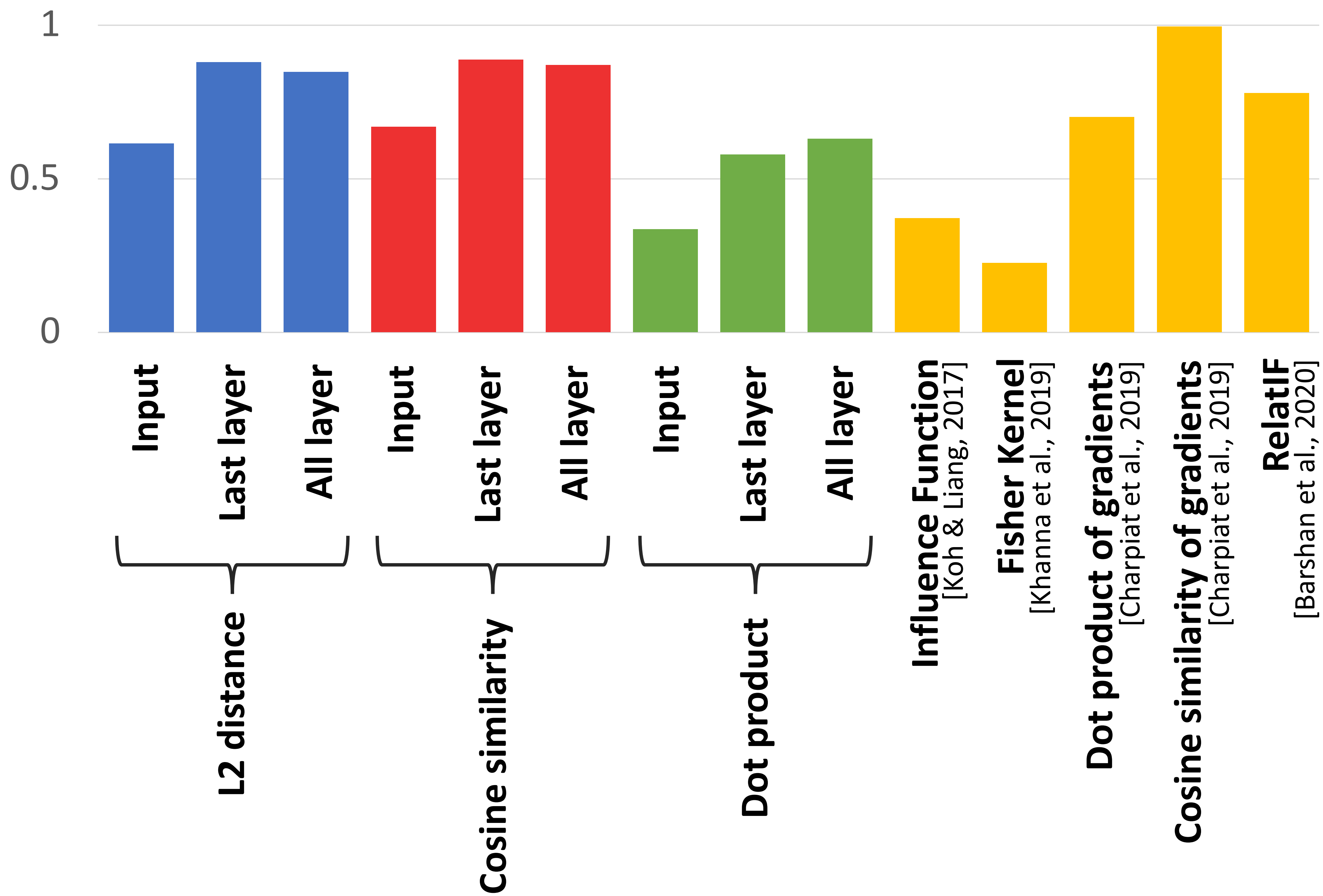
 is ***cat***. Because  is ***cat***.

Test instance | Training instance

 is ***cat***. Because  is ***dog***.

# Results of Identical Class Test

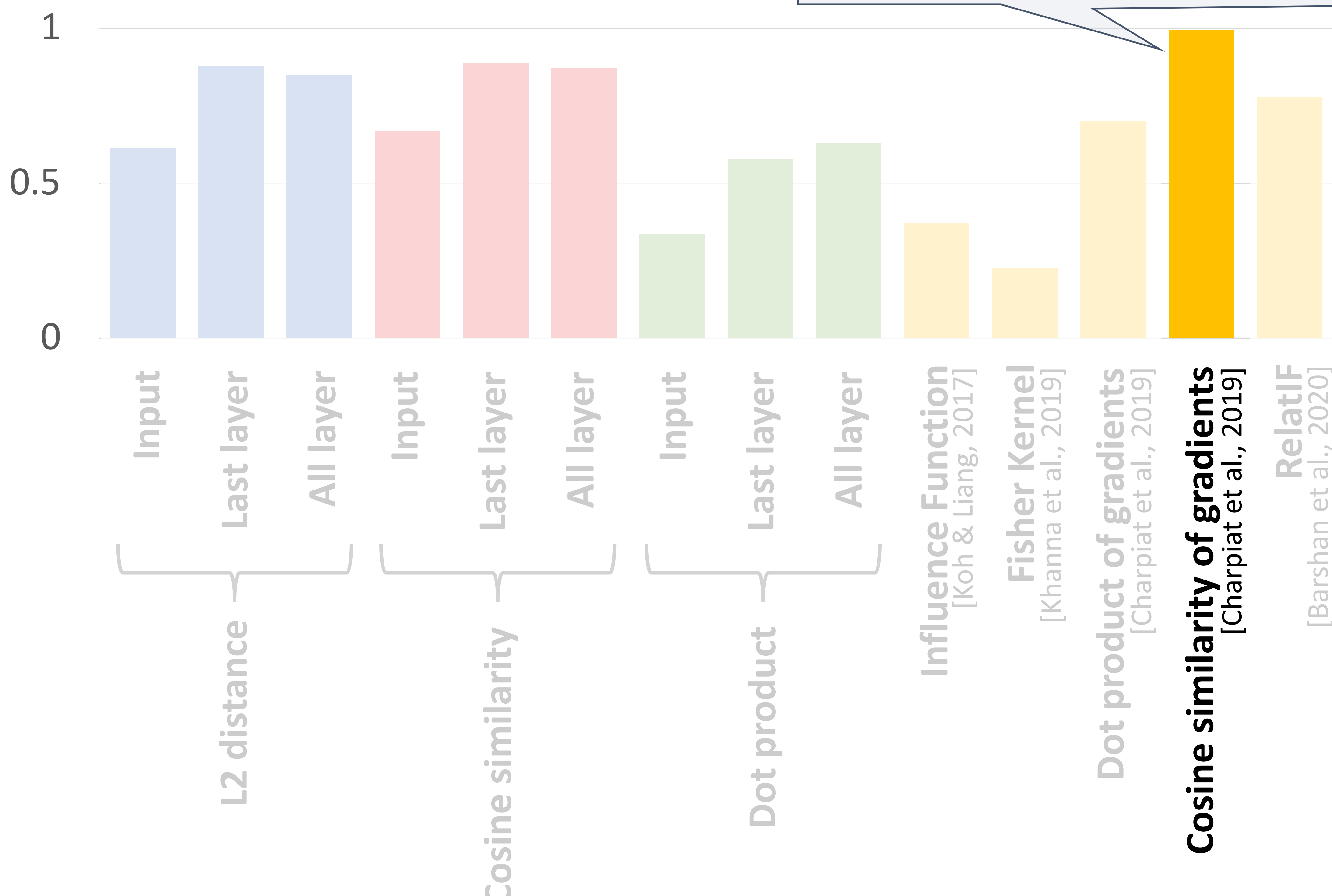
- Measure the percentage of the **most similar instance** in the same class



# Results of Identical Class Test

- Measure the percentage of the **most similar instance** in the same class

Cosine similarity of the gradients performs best.

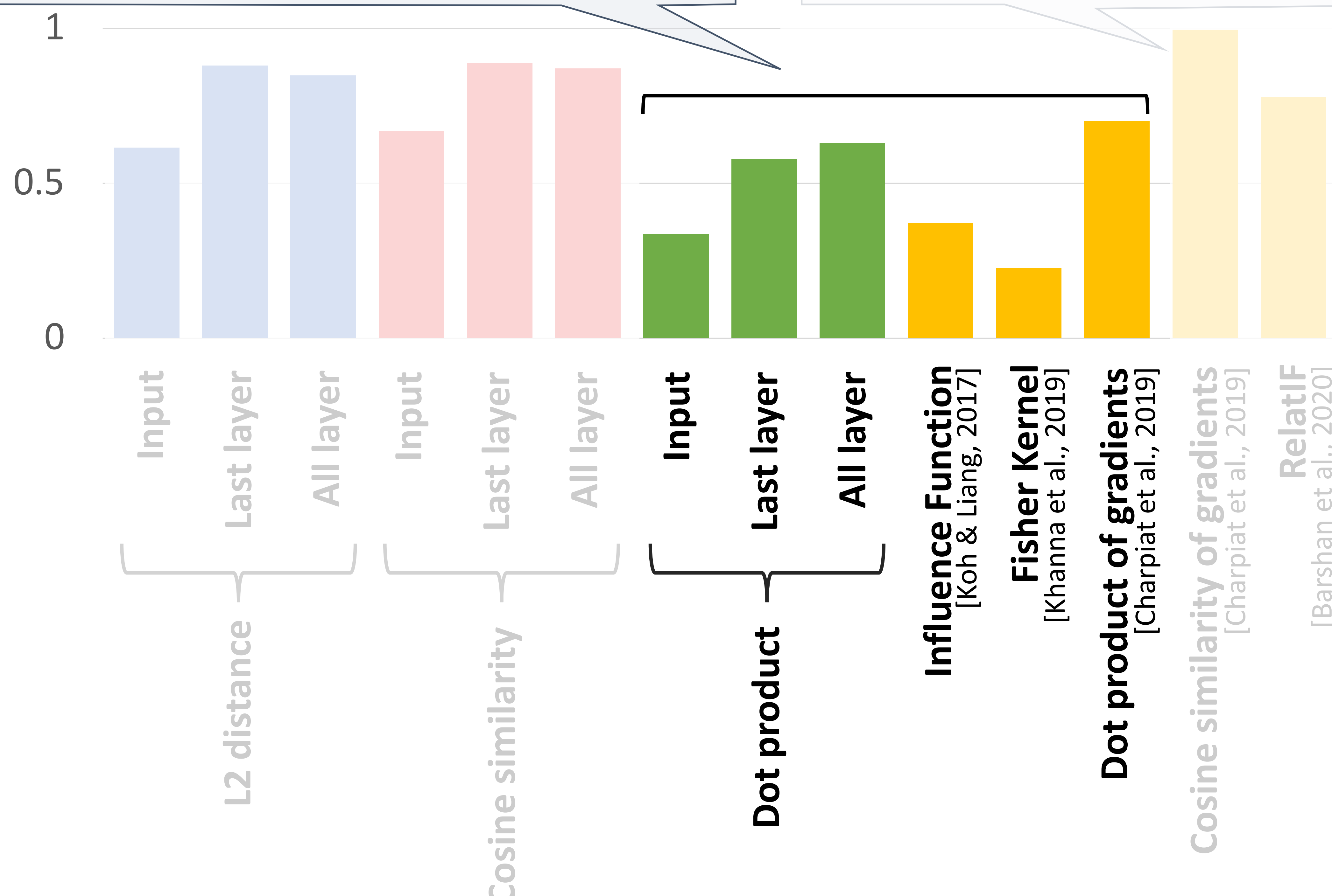


# Results of Identical Class Test

- Measure the percentage of the **most similar instance** in the same class

Dot product-based methods do not meet this minimal (easy) requirement.

Cosine similarity of the gradients performs best



# Why Are Dot Product-based Metrics Not Successful ?

- Some instances are judged as similar to various test instances due to **the large norm**.

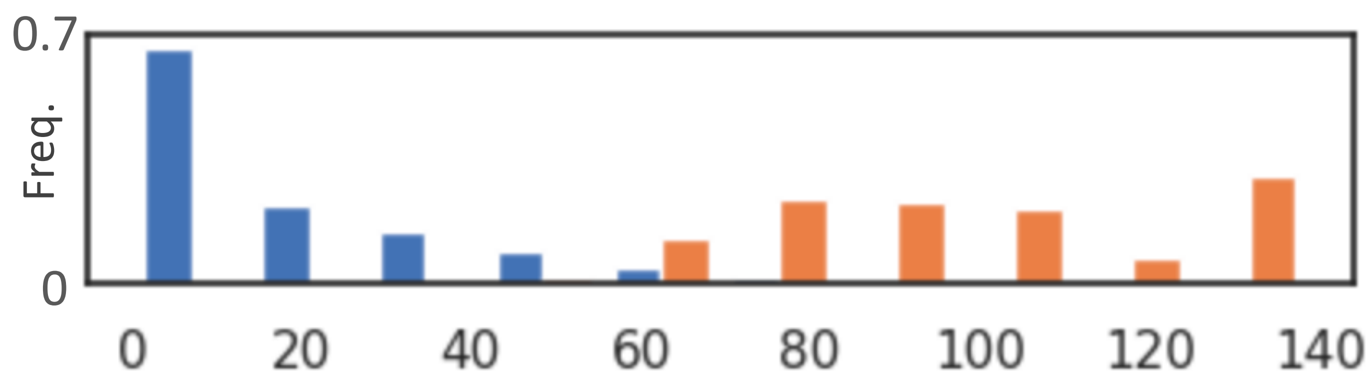
Example of **Dot product of gradients** $\langle \boldsymbol{g}_{\mathbf{test}}, \boldsymbol{g}_{\boldsymbol{i}} \rangle$ [Charpiat et al., 2019]

$\boldsymbol{g}_{\mathbf{test}}$: Gradient of the test instance

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

Norms for the entire training data

Norms for selected training instances

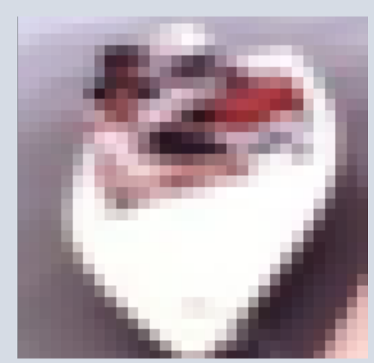



| Test instance | Explanation | Test instance | Explanation | Test instance | Explanation |
|---|---|---|---|---|---|

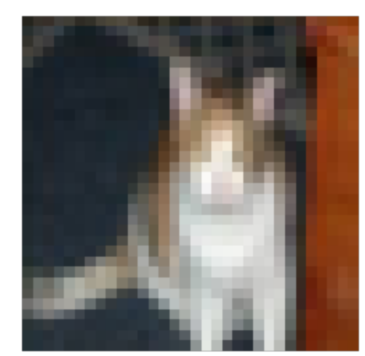
*frog*

*ship*

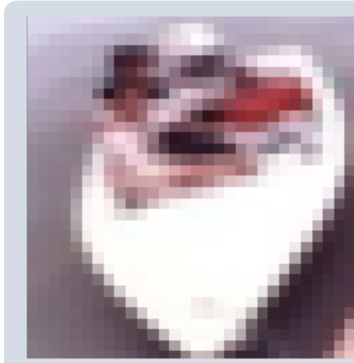
*horse*



*ship*



*cat*



*ship*

# Why Are Dot Product-based Metrics Not Successful ?

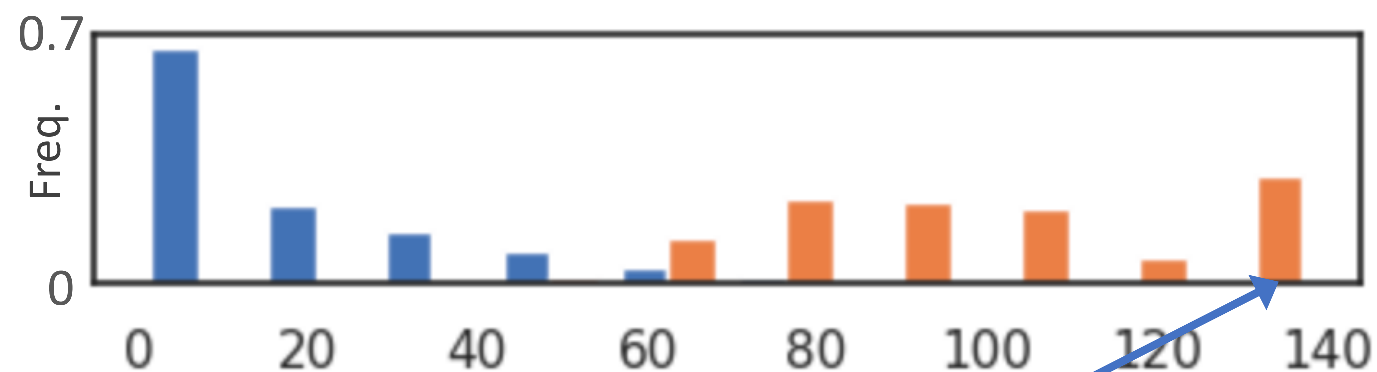
- Some instances are judged as similar to various test instances due to **the large norm**.

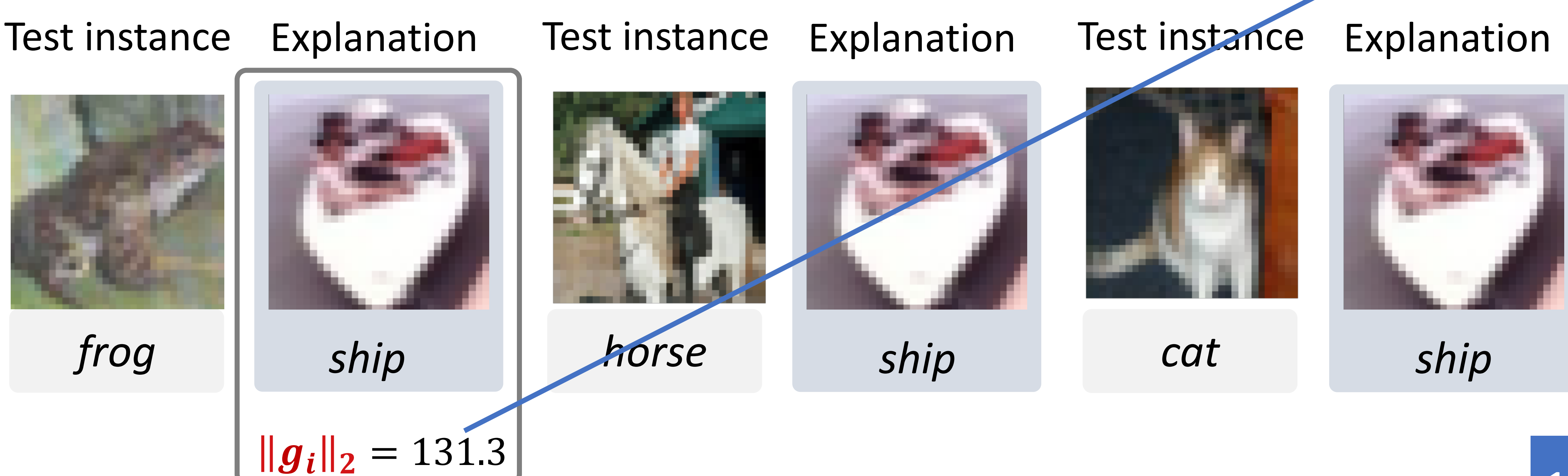
Example of **Dot product of gradients** $\langle g_{\text{test}}, g_i \rangle$ [Charpiat et al., 2019]

$g_{\text{test}}$: Gradient of the test instance
$g_i$: Gradient of the *i*-th training instance

Norms for the entire training data

Norms for selected training instances



# Summary

- Evaluated the appropriateness of the **similarity-based explanation**
  - Perspective 1: **Plausibility** [Lei+, 2016; Lage+, 2019; Strout+, 2019]
    - Test 1: **Identical class test**
    - Test 2: **Identical  subclass test**
  - Perspective 2: **Faithfulness** [Adebayo+, 2018; Lakkaraju+, 2019; Jacovi & Goldberg, 2020]
    - Test 3: **Randomization test**
- The results of the evaluation are as follows:
  - **Cosine similarity of the gradients** performs best.
  - **Dot product-based methods** do not meet minimal requirements.
- Expect that our work will help select/design better explanation methods