AI4SEC

NEURAL INFORMATION PROCESSING SYSTEMS

FACE: Faithful Automatic Concept Extraction

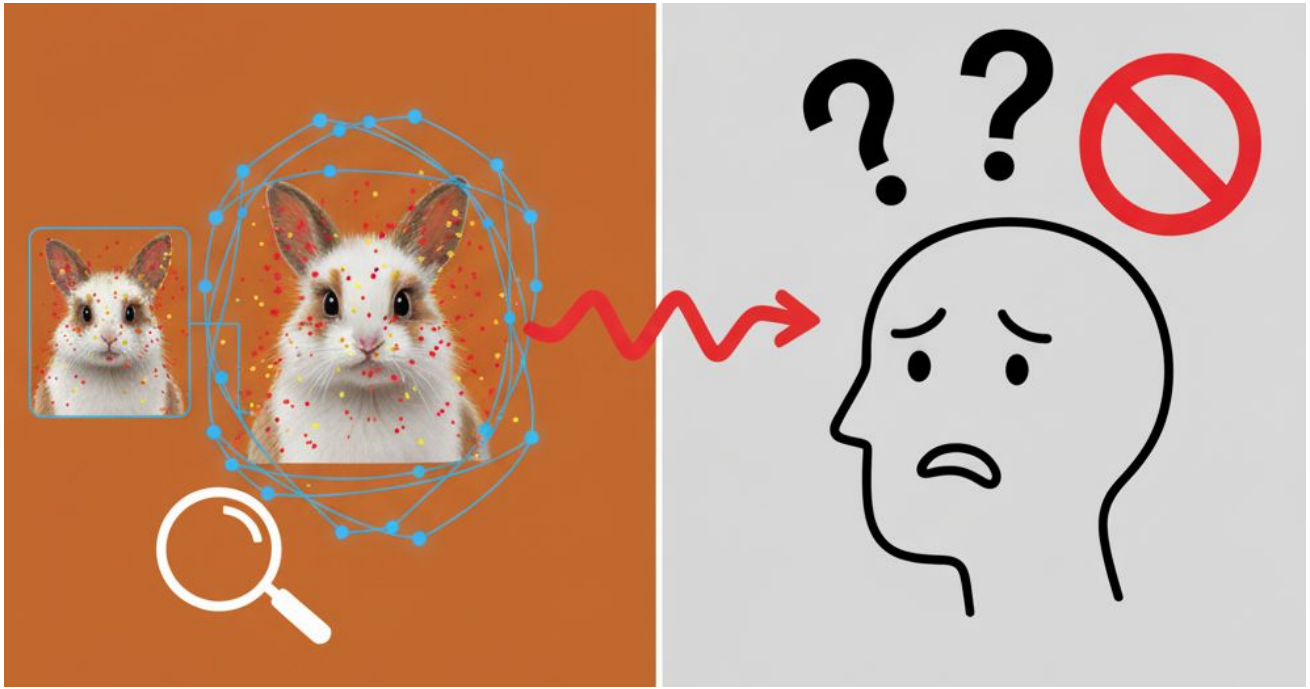
Dipkamal Bhusal[1], Michael Clifford[2], Sara Rampazzi[3], Nidhi Rastogi[1]

[1]Rochester Institute of Technology, [2]Toyota InfoTech Labs (Toyota Motor North America), [3]University of Florida

TOYOTA UF UNIVERSITY of FLORIDA RIT Rochester Institute of Technology

Supported by Toyota InfoTech Labs through Unrestricted Research Funds.

Motivation



Feature attribution: Pixel-level scores

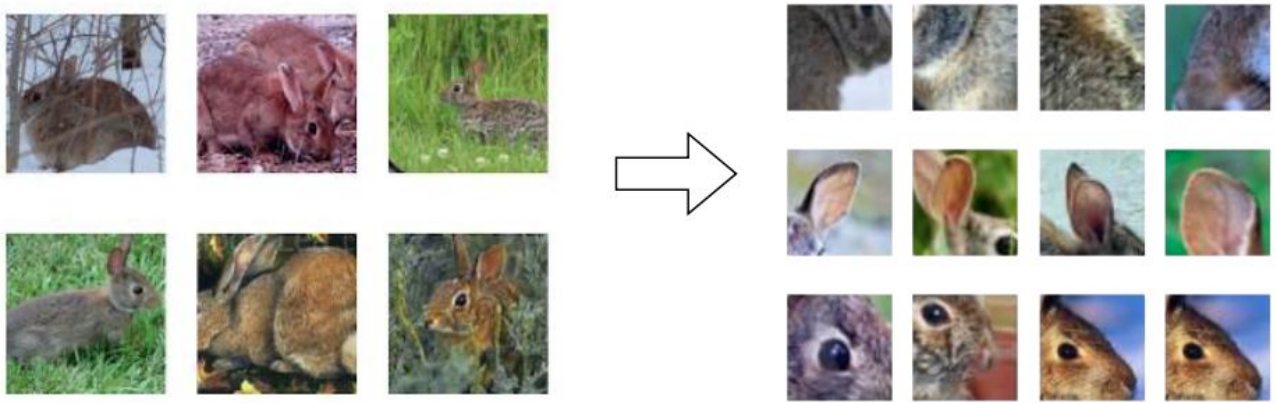
Semantic interpretability gap

AI4SEC

RIT Rochester Institute of Technology

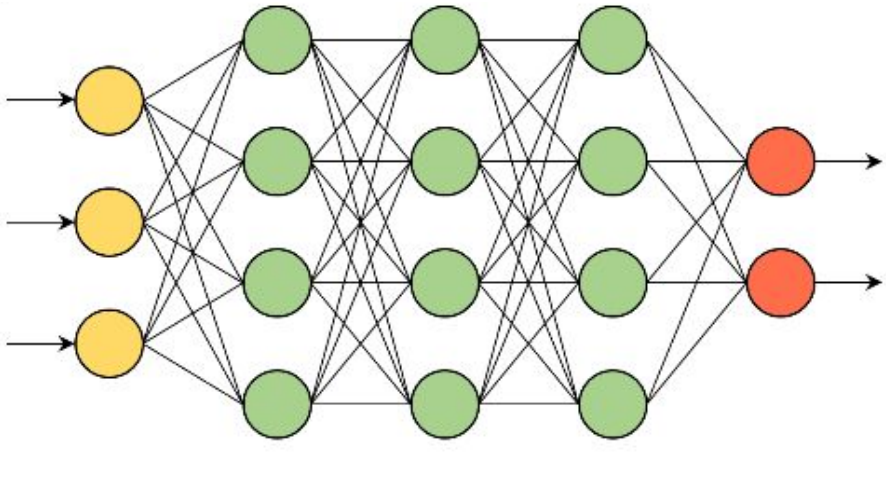
Motivation

Concept based explanation methods *explain* model predictions in terms of high-level, human-interpretable concepts.



Problem formulation: Automatic Concept Extraction Methods

$$f : \mathcal{X} \to \mathcal{Y}$$



$$g : \mathcal{X} \to \mathcal{G} \qquad h : \mathcal{G} \to \mathcal{Y}$$

$$f(\mathbf{x}) = (h \circ g)(\mathbf{x})$$

AI4SEC

RIT Rochester Institute of Technology

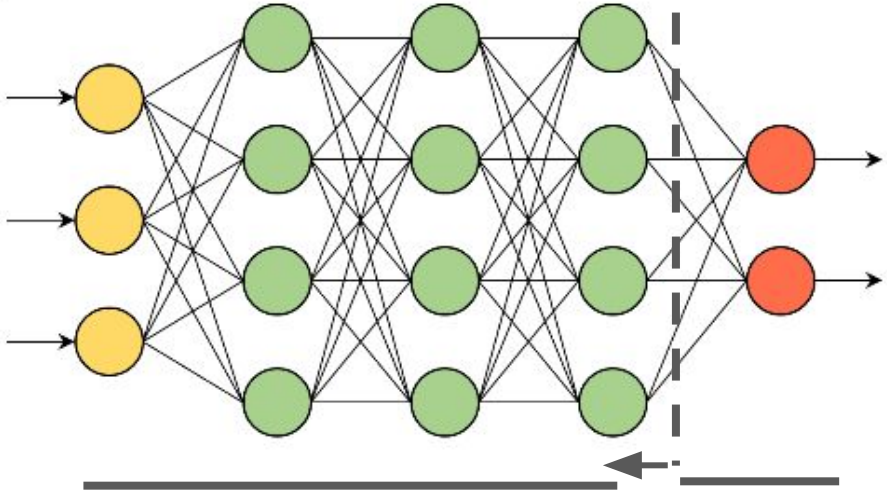
Problem formulation: Automatic Concept Extraction Methods

X

Crop/
Segment

X’

A = g(X’)



$g : \mathcal{X} \to \mathcal{G}$

$h : \mathcal{G} \to \mathcal{Y}$

$$f(\mathbf{x}) = (h \circ g)(\mathbf{x})$$

AI4SEC

RIT Rochester Institute of Technology

Problem formulation: Non-negative matrix factorization

Compute a set of concept activation vector **W** from **A**:

$$(\mathbf{U}, \mathbf{W}) = \underset{\mathbf{U}\geq 0, \mathbf{W}\geq 0}{\arg\min} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^{\mathsf{T}}\|_F^2,$$

W: Concept Activation Vectors
U: Transformed points on W
UW^T: Reconstructed activation vector (A')

AI4SEC

RIT Rochester Institute of Technology

Reconstruction alone is not faithful

A first-order Taylor expansion of the fixed classifier head (h) around reconstructed activations (**A'**) obtained with NMF:

$$h(\mathbf{A}') - h(\mathbf{A}) \approx \nabla h(\mathbf{A}') \cdot (\mathbf{A} - \mathbf{A}')$$

Reconstruction alone is not faithful

A first-order Taylor expansion of the fixed classifier head (h) around reconstructed activations (**A'**) obtained with NMF:

$$h(\mathbf{A}') - h(\mathbf{A}) \approx \nabla h(\mathbf{A}') \cdot (\mathbf{A} - \mathbf{A}')$$

Difference in classifier head prediction on reconstructed activation (A') and original activation (A),

Reconstruction alone is not faithful

A first-order Taylor expansion of the fixed classifier head (h) around reconstructed activations (**A'**) obtained with NMF:

$$h(\mathbf{A}') - h(\mathbf{A}) \approx \nabla h(\mathbf{A}') \cdot (\mathbf{A} - \mathbf{A}')$$

depends on both Jacobian and reconstruction error (A'-A)

Reconstruction alone is not faithful

A first-order Taylor expansion of the fixed classifier head (h) around reconstructed activations (**A'**) obtained with NMF:

$$h(\mathbf{A}') - h(\mathbf{A}) \approx \nabla h(\mathbf{A}') \cdot (\mathbf{A} - \mathbf{A}')$$

Model prediction on A and A' can differ significantly even when A' is close to A due to the Jacobian.

Reconstruction alone is not faithful

A first-order Taylor expansion of the fixed classifier head (h) around reconstructed activations (**A'**) obtained with NMF:

$$h(\mathbf{A}') - h(\mathbf{A}) \approx \nabla h(\mathbf{A}') \cdot (\mathbf{A} - \mathbf{A}')$$

There is no bound on the difference between model prediction on original and reconstructed activation.

Our solution: FACE (Faithful Automatic Concept Extraction)

A faithfulness-aware variant of NMF that explicitly aligns the reconstructed activations with the model's predictive behavior.

Our solution: FACE (Faithful Automatic Concept Extraction)

A faithfulness-aware variant of NMF that explicitly aligns the reconstructed activations with the model's predictive behavior.

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2$$

Original NMF Formulation

Our solution: FACE (Faithful Automatic Concept Extraction)

A faithfulness-aware variant of NMF that explicitly aligns the reconstructed activations with the model's predictive behavior.

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 \; + \lambda \cdot \mathbf{KL}(h(\mathbf{A})\|h(\mathbf{U}\mathbf{W}^\top))$$

AI4SEC

RIT Rochester Institute of Technology

Our solution: FACE (Faithful Automatic Concept Extraction)

A faithfulness-aware variant of NMF that explicitly aligns the reconstructed activations with the model's predictive behavior.

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 + \lambda \cdot \mathbf{KL}(h(\mathbf{A})\|h(\mathbf{U}\mathbf{W}^\top))$$

This ensures that model prediction remains consistent before and after matrix factorization.

AI4SEC

RIT Rochester Institute of Technology

Our solution: FACE (Faithful Automatic Concept Extraction)

A faithfulness-aware variant of NMF that explicitly aligns the reconstructed activations with the model's predictive behavior.

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 + \lambda \cdot \mathbf{KL}(h(\mathbf{A})\|h(\mathbf{U}\mathbf{W}^\top))$$

- Cannot apply multiplicative update rules.
- Use projected gradient descent.
- Convergence is guaranteed under mild conditions*.

See the paper for proof.

AI4SEC

RIT Rochester Institute of Technology

Reconstruction with KL-loss is faithful

Let **p** and **q** be softmax/predictions before and after reconstruction. Then, FACE minimizes

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 + \lambda \cdot \mathbf{KL}(p\|q)$$

Reconstruction with KL-loss is faithful

Let **p** and **q** be softmax/predictions before and after reconstruction. Then, FACE minimizes

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 + \lambda \cdot \mathbf{KL}(p\|q)$$

Minimizing KL divergence between p and q

Reconstruction with KL-loss is faithful

Let **p** and **q** be softmax/predictions before and after reconstruction. Then, FACE minimizes

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 + \lambda \cdot \mathbf{KL}(p\|q)$$

Bounds the difference between two distribution. With Pinsker's inequality,

Reconstruction with KL-loss is faithful

Let **p** and **q** be softmax/predictions before and after reconstruction. Then, FACE minimizes

$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 + \lambda \cdot \mathbf{KL}(p\|q)$$

Bounds the difference between two distribution. With Pinsker's inequality,

$$\|p - q\|_1 \leq \sqrt{2 \cdot \mathbf{KL}(\cdot\|\cdot)} \leq \sqrt{2\varepsilon}$$

Reconstruction with KL-loss is faithful

Let **p** and **q** be softmax/predictions before and after reconstruction. Then, FACE minimizes

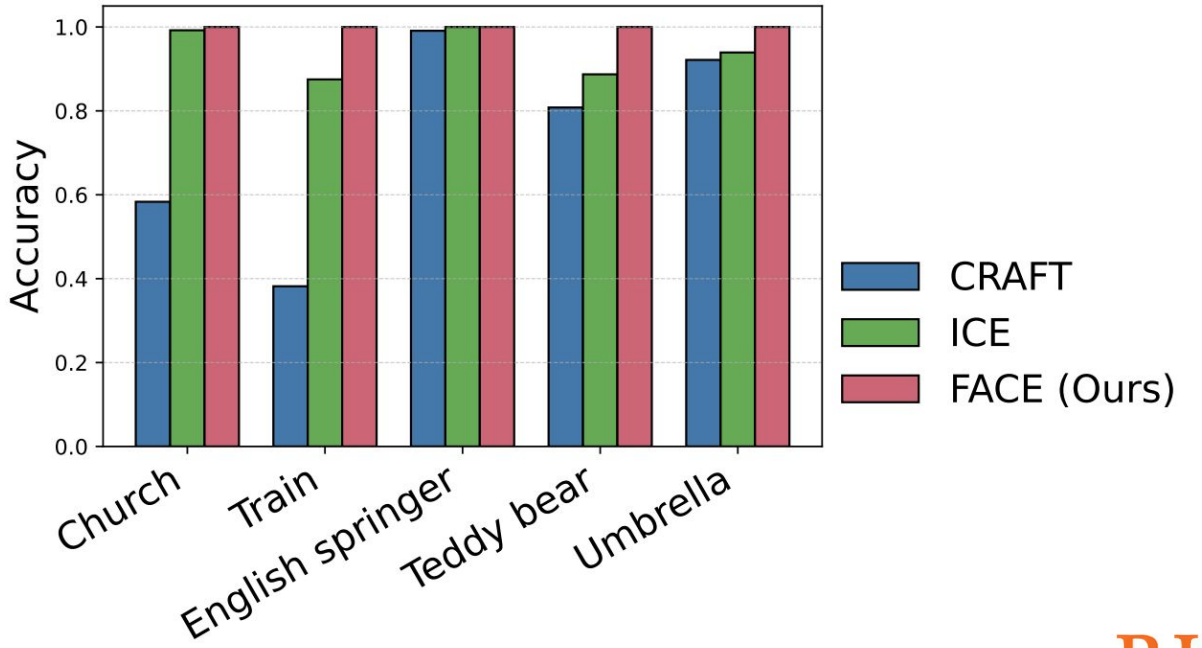
$$\min_{\mathbf{U}\geq 0,\, \mathbf{W}\geq 0} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{W}^\top\|_F^2 + \lambda \cdot \mathbf{KL}(p\|q)$$

Bounds the difference between two distribution. With Pinsker's inequality,

$$\|p - q\|_1 \leq \sqrt{2 \cdot \mathbf{KL}(\cdot\|\cdot)} \leq \sqrt{2\varepsilon}$$

Minimizing reconstruction alone does not offer any such guarantee.

Failure case of unconstrained NMF: Recovering model accuracy on reconstructed activation vector



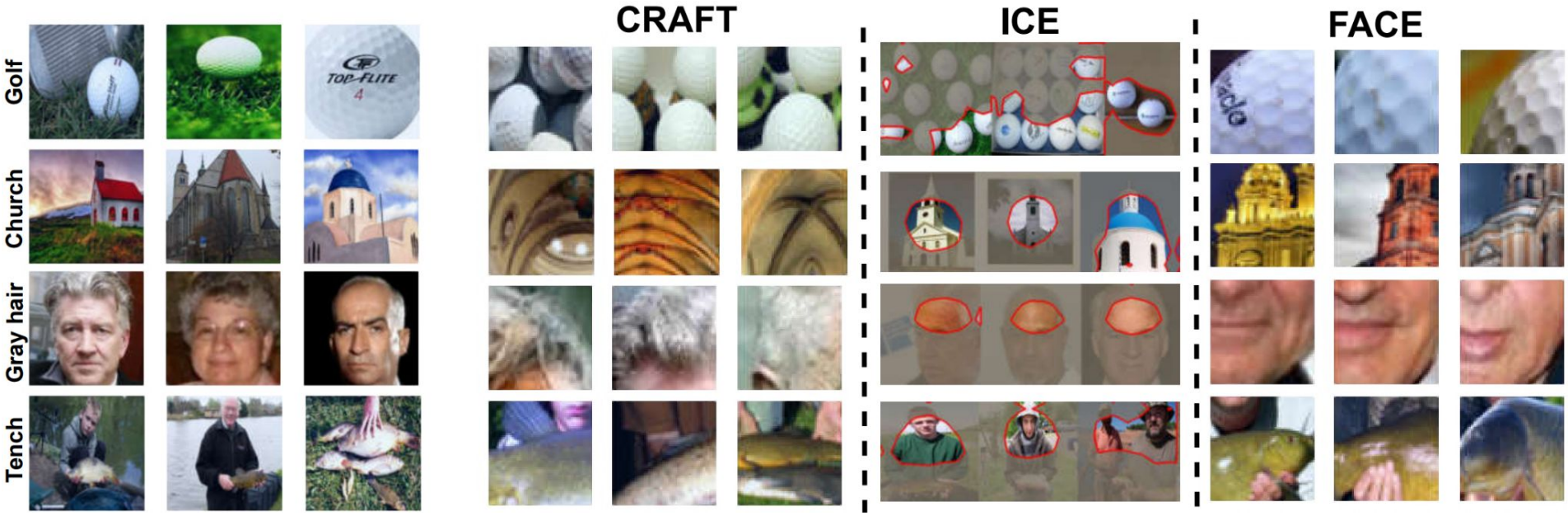
Evaluation

- **Concept Insertion (C-Ins):**
  - How fast the model accuracy increases when the most important concepts are added into a blank representation?
- **Concept Deletion (C-Del):**
  - How fast the model accuracy drops when the most important concepts are removed from the latent representation?
- **Concept sparsity (C-Gini):**
  - How is the distribution of importance scores for each concept?
  - Concept importance are computed using Sobol-concept importance.

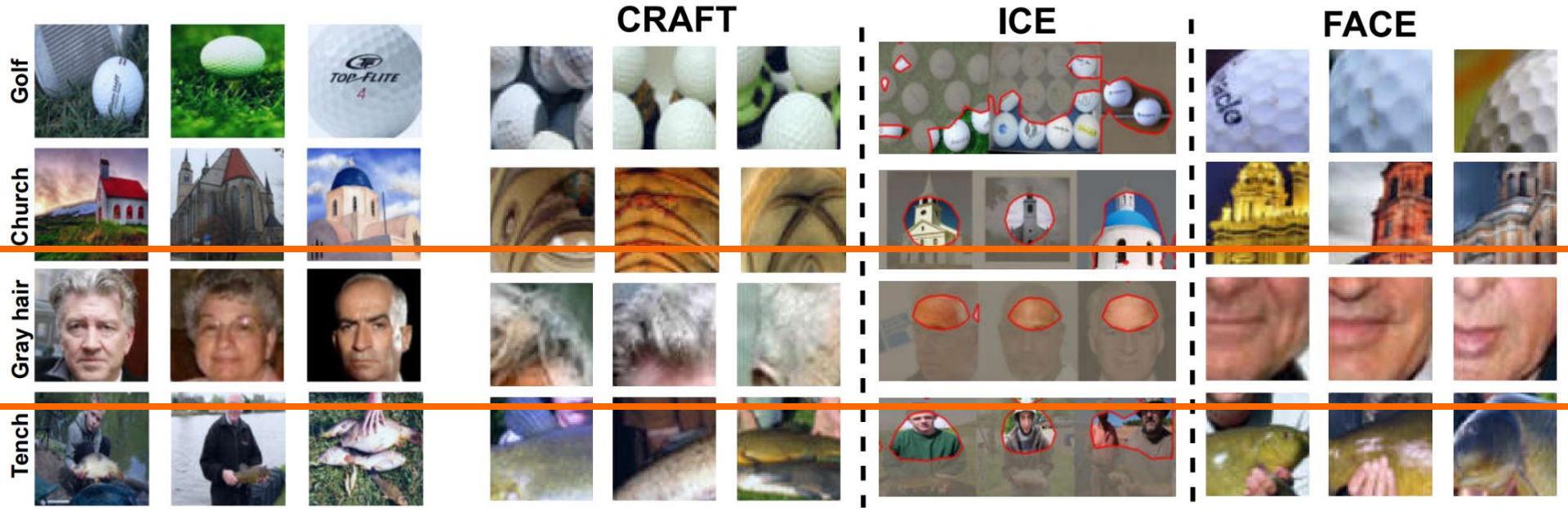
Evaluation

		ResNet-34			MobileNetV2		
		C-Ins ↑	C-Del ↑	C-Gini ↑	C-Ins ↑	C-Del ↑	C-Gini ↑
ImageNet	ICE	0.908 ± 0.034	0.484 ± 0.063	0.537 ± 0.071	0.916 ± 0.020	0.346 ± 0.049	0.605 ± 0.149
	CRAFT	0.932 ± 0.001	0.752 ± 0.031	0.835 ± 0.031	0.886 ± 0.001	0.646 ± 0.024	0.805 ± 0.041
	FACE (Ours)	**0.969 ± 0.010**	**0.891 ± 0.011**	**0.895 ± 0.001**	**0.974 ± 0.003**	**0.882 ± 0.012**	**0.947 ± 0.001**
COCO	ICE	0.883 ± 0.029	0.632 ± 0.020	0.623 ± 0.086	0.906 ± 0.007	0.485 ± 0.051	0.622 ± 0.064
	CRAFT	0.861 ± 0.029	0.691 ± 0.029	0.874 ± 0.035	0.764 ± 0.036	0.571 ± 0.026	0.874 ± 0.047
	FACE (Ours)	**0.971 ± 0.013**	**0.894 ± 0.010**	**0.947 ± 0.000**	**0.974 ± 0.002**	**0.905 ± 0.012**	**0.949 ± 0.000**
CelebA	ICE	0.910 ± 0.008	0.365 ± 0.016	0.662 ± 0.087	0.858 ± 0.007	0.385 ± 0.050	0.728 ± 0.032
	CRAFT	0.953 ± 0.067	0.604 ± 0.036	0.901 ± 0.026	0.960 ± 0.116	0.592 ± 0.070	0.911 ± 0.028
	FACE (Ours)	**0.971 ± 0.012**	**0.635 ± 0.014**	**0.928 ± 0.000**	**0.978 ± 0.001**	**0.649 ± 0.011**	**0.932 ± 0.001**

Qualitative comparison



Qualitative comparison



AI4SEC

RIT Rochester Institute of Technology

Limitations and future work

1. **Limitations:**
   a. Lack of human-centered evaluation.
   b. Suitable only to CNN based architecture.
2. **Future work:**
   a. Extend FACE to Vision Transformers.
   b. Utilize FACE to detect and fix spurious features learned by models.

Conclusion

1. We demonstrate that reconstruction of activation vectors without inductive bias does not guarantee faithful explanations in concept-based methods.
2. We propose FACE, a faithfulness-aware variant of automatic concept extraction.
3. FACE ensures that model predictions remain consistent before and after concept extraction and guarantees faithfulness of output explanations.
4. Empirical evaluation of ImageNet, COCO and CelebA on ResNet and MobileNet show that FACE outperforms existing methods on faithfulness and sparsity.