A Framework to Learn with Interpretation

Jayneel Parekh*
(Advisors: Prof. Florence d'Alché-Buc* and Assist. Prof. Pavlo Mozharovskyi*)

*LTCI, Télécom Paris, IP Paris

October 16, 2021

NeurIPS 2021

Introduction

- Interpretability is the ability to provide human-understandable insights on the decision process.

Introduction

- Interpretability is the ability to provide human-understandable insights on the decision process.

Two primary problem settings regarding interpretability in literature:

1. Post-hoc approaches
2. Interpretability by design

Introduction

- Interpretability is the ability to provide human-understandable insights on the decision process.

Two primary problem settings regarding interpretability in literature:

1. Post-hoc approaches
2. Interpretability by design

We propose a novel framework FLINT – jointly learns a predictor and its associated interpreter. Primarily to learn interpretable models by design.

Key aspects of FLINT

- A special case applicable for post-hoc interpretations.
- *Means of interpretation*: raw features, simplified representation, prototypes, logical rules, high-level features/concepts.
- *Scope of interpretation*: Local AND Global.

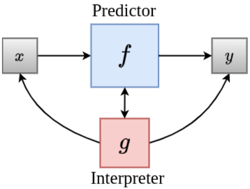
Supervised Learning with Interpretation (SLI)

- Generic task SLI: Considers prediction and interpretation as separate tasks with dedicated models f and g.

Supervised Learning with Interpretation (SLI)

- Generic task SLI: Considers prediction and interpretation as separate tasks with dedicated models f and g.
- Optimization problem:

$$\arg \min_{f \in \mathcal{F}, g \in \mathcal{G}_f} \mathcal{L}_{pred}(f, \mathcal{S}) + \mathcal{L}_{int}(f, g, \mathcal{S})$$



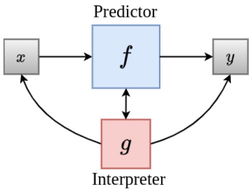
- $\mathcal{F}$ – Space of predictive models.
  $\mathcal{G}_f$ – Family of interpreter models dependent on f.

Supervised Learning with Interpretation (SLI)

- Generic task SLI: Considers prediction and interpretation as separate tasks with dedicated models f and g.

- Optimization problem:

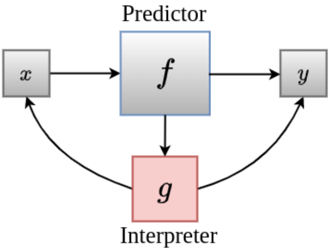
$$\arg \min_{f \in \mathcal{F}, g \in \mathcal{G}_f} \mathcal{L}_{pred}(f, \mathcal{S}) + \mathcal{L}_{int}(f, g, \mathcal{S})$$



- $\mathcal{F}$ – Space of predictive models.
  $\mathcal{G}_f$ – Family of interpreter models dependent on f.

- Our goal is to address SLI when $\mathcal{F}$ instantiated with deep neural networks and task is multi-class classification.

Specializing SLI: Post-hoc interpretation



- A special case with $f = \hat{f}$ is fixed and we only learn g.
- Optimization problem:

$$\arg\min_{g \in \mathcal{G}_{\hat{f}}} \mathcal{L}_{int}(\hat{f}, g, \mathcal{S}),$$

(No gradients are backpropagated to f.)

Design of FLINT

FLINT: Framework to Learn INTerpretable networks

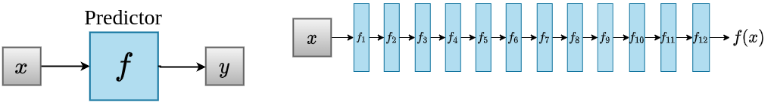


Figure: System Overview

Design of FLINT

FLINT: Framework to Learn INTerpretable networks

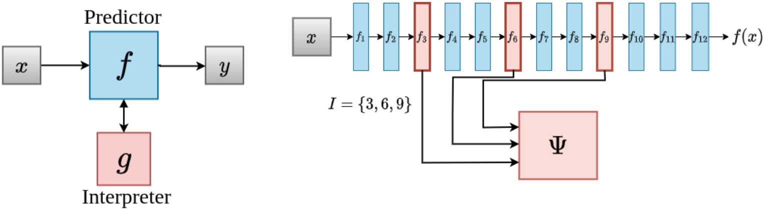


Figure: System Overview

- **Interpreter** $g(x) = h \circ \Psi \circ f_{\mathcal{I}}(x) = h \circ \Phi(x) := \text{softmax}(W^T \Phi(x))$. Computes composition of attribute functions $\Phi(x)$ and interpretable function h characterized by weight matrix W.

Design of FLINT

FLINT: Framework to Learn INTerpretable networks

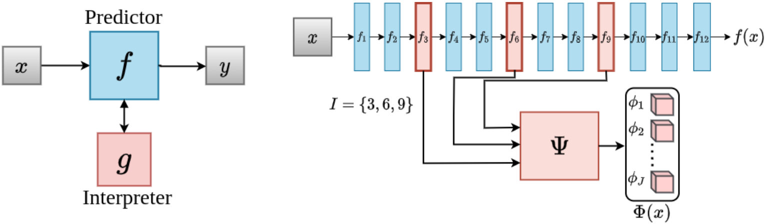


Figure: System Overview

- **Interpreter** $g(x) = h \circ \Psi \circ f_{\mathcal{I}}(x) = h \circ \Phi(x) := \text{softmax}(W^T \Phi(x))$. Computes composition of attribute functions $\Phi(x)$ and interpretable function h characterized by weight matrix W.

Design of FLINT

FLINT: Framework to Learn INTerpretable networks

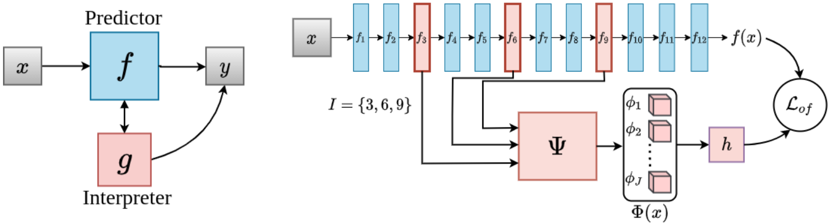


Figure: System Overview

- **Interpreter** $g(x) = h \circ \Psi \circ f_{\mathcal{I}}(x) = h \circ \Phi(x) := \text{softmax}(W^T \Phi(x))$. Computes composition of attribute functions $\Phi(x)$ and interpretable function h characterized by weight matrix W.
- **Attribute dictionary**: functions $\phi_j : \mathcal{X} \to \mathbb{R}^+, j = 1, \ldots J$. $\phi_j(x)$ is activation of some high level attribute, i.e. a "concept" over $\mathcal{X}$.

Design of FLINT

FLINT: Framework to Learn INTerpretable networks

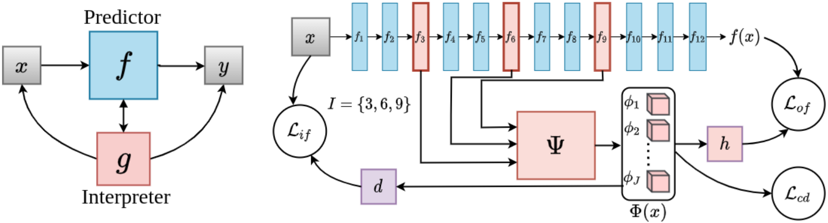


Figure: System Overview

- **Interpreter** $g(x) = h \circ \Psi \circ f_{\mathcal{I}}(x) = h \circ \Phi(x) := \text{softmax}(W^T \Phi(x))$. Computes composition of attribute functions $\Phi(x)$ and interpretable function h characterized by weight matrix W.
- **Attribute dictionary**: functions $\phi_j : \mathcal{X} \to \mathbb{R}^+, j = 1, \ldots J$. $\phi_j(x)$ is activation of some high level attribute, i.e. a "concept" over $\mathcal{X}$.

Losses for Interpretability

- Complete interpretability loss term:

$$\mathcal{L}_{int}(f, \Phi, h, d, \mathcal{S}) = \beta\mathcal{L}_{of}(f, \Phi, h, \mathcal{S}) + \delta\mathcal{L}_{cd}(\Phi, \mathcal{S}) + \gamma\mathcal{L}_{if}(\Phi, h, d, \mathcal{S})$$

Losses for Interpretability

- Complete interpretability loss term:

$$\mathcal{L}_{int}(f, \Phi, h, d, \mathcal{S}) = \beta\mathcal{L}_{of}(f, \Phi, h, \mathcal{S}) + \delta\mathcal{L}_{cd}(\Phi, \mathcal{S}) + \gamma\mathcal{L}_{if}(\Phi, h, d, \mathcal{S})$$

- Composed of three individual terms:
  - *Fidelity to output* term $\mathcal{L}_{of}$: Generalized cross-entropy between $g(x)$ and $f(x)$. Their outputs should match.

Losses for Interpretability

- Complete interpretability loss term:

$$\mathcal{L}_{int}(f, \Phi, h, d, \mathcal{S}) = \beta \mathcal{L}_{of}(f, \Phi, h, \mathcal{S}) + \delta \mathcal{L}_{cd}(\Phi, \mathcal{S}) + \gamma \mathcal{L}_{if}(\Phi, h, d, \mathcal{S})$$

- Composed of three individual terms:
  - *Fidelity to output* term $\mathcal{L}_{of}$: Generalized cross-entropy between $g(x)$ and $f(x)$. Their outputs should match.
  - *Conciseness and Diversity* term $\mathcal{L}_{cd}$: For single sample, small # of ϕ_j's should activate (Conciseness).

Losses for Interpretability

- Complete interpretability loss term:

$$\mathcal{L}_{int}(f, \Phi, h, d, \mathcal{S}) = \beta \mathcal{L}_{of}(f, \Phi, h, \mathcal{S}) + \delta \mathcal{L}_{cd}(\Phi, \mathcal{S}) + \gamma \mathcal{L}_{if}(\Phi, h, d, \mathcal{S})$$

- Composed of three individual terms:
  - *Fidelity to output* term $\mathcal{L}_{of}$: Generalized cross-entropy between $g(x)$ and $f(x)$. Their outputs should match.
  - *Conciseness and Diversity* term $\mathcal{L}_{cd}$: For single sample, small # of ϕ_j's should activate (Conciseness). Across many samples, multiple attributes should be used (Diversity). Entropy based loss (Jain et al).

Losses for Interpretability

- Complete interpretability loss term:

$$\mathcal{L}_{int}(f, \Phi, h, d, \mathcal{S}) = \beta \mathcal{L}_{of}(f, \Phi, h, \mathcal{S}) + \delta \mathcal{L}_{cd}(\Phi, \mathcal{S}) + \gamma \mathcal{L}_{if}(\Phi, h, d, \mathcal{S})$$

- Composed of three individual terms:
  - *Fidelity to output* term $\mathcal{L}_{of}$: Generalized cross-entropy between $g(x)$ and $f(x)$. Their outputs should match.
  - *Conciseness and Diversity* term $\mathcal{L}_{cd}$: For single sample, small # of ϕ_j's should activate (Conciseness). Across many samples, multiple attributes should be used (Diversity). Entropy based loss (Jain et al).
  - *Fidelity to input* term $\mathcal{L}_{if}$. To promote encoding high-level patterns relevant to input. Use of autoencoder via decoder d (Melis & Jaakkola).

Losses for Interpretability

- Complete interpretability loss term:

$$\mathcal{L}_{int}(f, \Phi, h, d, \mathcal{S}) = \beta \mathcal{L}_{of}(f, \Phi, h, \mathcal{S}) + \delta \mathcal{L}_{cd}(\Phi, \mathcal{S}) + \gamma \mathcal{L}_{if}(\Phi, h, d, \mathcal{S})$$

- Composed of three individual terms:
  - *Fidelity to output* term $\mathcal{L}_{of}$: Generalized cross-entropy between $g(x)$ and $f(x)$. Their outputs should match.
  - *Conciseness and Diversity* term $\mathcal{L}_{cd}$: For single sample, small # of ϕ_j's should activate (Conciseness). Across many samples, multiple attributes should be used (Diversity). Entropy based loss (Jain et al).
  - *Fidelity to input* term $\mathcal{L}_{if}$. To promote encoding high-level patterns relevant to input. Use of autoencoder via decoder d (Melis & Jaakkola).
- $\mathcal{L}_{pred}(f, S)$ is the standard cross-entropy loss.

Generating Interpretations

How do we get local and global interpretability from our trained model?

Generating Interpretations

How do we get local and global interpretability from our trained model?

1. **Local relevance** of an attribute j for sample x ($r_{j,x}$): Obtained via activation $\phi_j(x)$ and weight for that attribute $w_{j,\hat{y}}$.

$$r_{j,x} = \frac{\alpha_{j,\hat{y},x}}{\max_i |\alpha_{i,\hat{y},x}|}, \alpha_{j,\hat{y},x} = \phi_j(x).w_{j,\hat{y}}$$

Generating Interpretations

How do we get local and global interpretability from our trained model?

1. **Local relevance** of an attribute j for sample x ($r_{j,x}$): Obtained via activation $\phi_j(x)$ and weight for that attribute $w_{j,\hat{y}}$.

$$r_{j,x} = \frac{\alpha_{j,\hat{y},x}}{\max_i |\alpha_{i,\hat{y},x}|}, \alpha_{j,\hat{y},x} = \phi_j(x).w_{j,\hat{y}}$$

2. **Global relevance**: Average out $r_{j,x}$ for samples with same predicted class to get relationship of class-attribute relationships $r_{j,c}$.

$$r_{j,c} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} r_{j,x}, \mathcal{S}_c = \{x \in \mathcal{S} | \hat{y} = c\}$$

Generating Interpretations

How do we get local and global interpretability from our trained model?

1. **Local relevance** of an attribute j for sample x ($r_{j,x}$): Obtained via activation $\phi_j(x)$ and weight for that attribute $w_{j,\hat{y}}$.

$$r_{j,x} = \frac{\alpha_{j,\hat{y},x}}{\max_i |\alpha_{i,\hat{y},x}|}, \alpha_{j,\hat{y},x} = \phi_j(x).w_{j,\hat{y}}$$

2. **Global relevance**: Average out $r_{j,x}$ for samples with same predicted class to get relationship of class-attribute relationships $r_{j,c}$.

$$r_{j,c} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} r_{j,x}, \mathcal{S}_c = \{x \in \mathcal{S} | \hat{y} = c\}$$

3. Understanding concept encoded by an attribute.

Generating Interpretations

How do we get local and global interpretability from our trained model?

1. **Local relevance** of an attribute j for sample x ($r_{j,x}$): Obtained via activation $\phi_j(x)$ and weight for that attribute $w_{j,\hat{y}}$.

$$r_{j,x} = \frac{\alpha_{j,\hat{y},x}}{\max_i |\alpha_{i,\hat{y},x}|}, \alpha_{j,\hat{y},x} = \phi_j(x).w_{j,\hat{y}}$$

2. **Global relevance**: Average out $r_{j,x}$ for samples with same predicted class to get relationship of class-attribute relationships $r_{j,c}$.

$$r_{j,c} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} r_{j,x}, \mathcal{S}_c = \{x \in \mathcal{S} | \hat{y} = c\}$$

3. Understanding concept encoded by an attribute.

$$1 + 3 \longrightarrow \text{local interpretability}$$

$$2 + 3 \longrightarrow \text{global interpretability}$$

Generating Interpretations

Last piece: How do we understand concept encoded by an attribute ϕ_j?

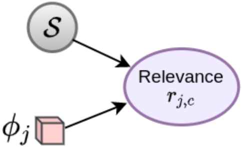


Figure: Flow to understand encoded concept by attribute ϕ_j

- Compute global relevance $r_{j,c}$ (for each class c).

Generating Interpretations

Last piece: How do we understand concept encoded by an attribute ϕ_j?

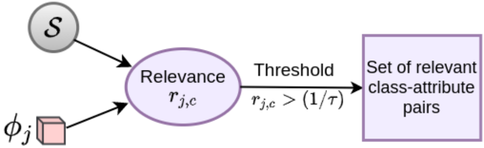


Figure: Flow to understand encoded concept by attribute ϕ_j

- Compute global relevance $r_{j,c}$ (for each class c).
- Select relevant class-attribute pairs by thresholding $r_{j,c}$.

Generating Interpretations

Last piece: How do we understand concept encoded by an attribute ϕ_j?

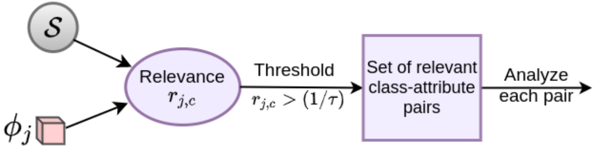


Figure: Flow to understand encoded concept by attribute ϕ_j

- Compute global relevance $r_{j,c}$ (for each class c).
- Select relevant class-attribute pairs by thresholding $r_{j,c}$.
- Analyze each pair by repeating this:

Generating Interpretations

Last piece: How do we understand concept encoded by an attribute ϕ_j?

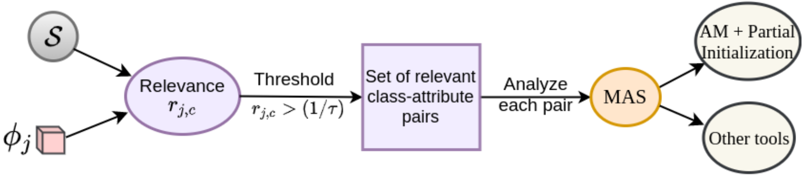


Figure: Flow to understand encoded concept by attribute ϕ_j

- Compute global relevance $r_{j,c}$ (for each class c).
- Select relevant class-attribute pairs by thresholding $r_{j,c}$.
- Analyze each pair by repeating this:
  - Select samples of class c maximally activating ϕ_j (MAS).
  - Use Activation Maximization w/ Partial Initialization (AM+PI) as tool – *optimizes* weakly initialized input to maximally activate ϕ_j.

Generating Interpretations

Last piece: How do we understand concept encoded by an attribute ϕ_j?

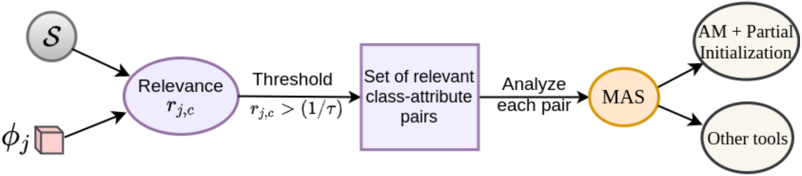


Figure: Flow to understand encoded concept by attribute ϕ_j

- Compute global relevance $r_{j,c}$ (for each class c).
- Select relevant class-attribute pairs by thresholding $r_{j,c}$.
- Analyze each pair by repeating this:
  - Select samples of class c maximally activating ϕ_j (MAS).
  - Use Activation Maximization w/ Partial Initialization (AM+PI) as tool – *optimizes* weakly initialized input to maximally activate ϕ_j.
- Can use AM+PI to analyze any sample for local interpretations.

Experimental Validation

- **Datasets & Networks**:
  - MNIST, FashionMNIST – LeNet,
  - CIFAR10, QuickDraw subset (Hand sketch recognition) – ResNet18.
- **Quantitative Evaluation Metrics**:
  - *Accuracy*: Two goals (1) Comparison to other interpretable NN architectures, (2) Training f & g jointly does not negatively affect performance.
  - *Fidelity of interpreter*: Fraction of samples where prediction of g is same as f.
  - *Conciseness of interpretations*: Average number of attributes "important" to interpretations.

$$\mathrm{CNS}_{g,x} = |\{j : |r_{j,x}| > 1/\tau\}|$$

Results – Quantitative I

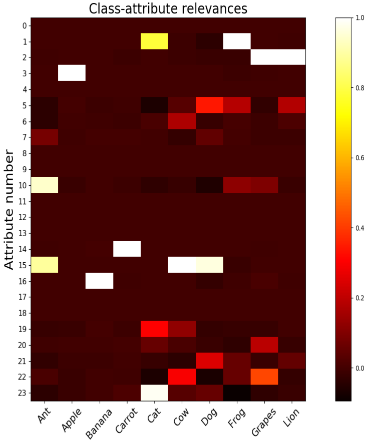
	BASE-f	SENN	PrototypeDNN	FLINT-f	FLINT-g
MNIST	98.9±0.1	98.4±0.1	**99.2**	98.9±0.2	98.3±0.2
FashionMNIST	90.4±0.1	84.2±0.3	90.0	**90.5±0.2**	86.8±0.4
CIFAR10	84.7±0.3	77.8±0.7	–	**84.5±0.2**	84.0±0.4
QuickDraw	85.3±0.2	85.5±0.4	–	**85.7±0.3**	85.4±0.1

Table: Accuracy (in %) on different datasets. BASE-f is system trained with just accuracy loss. FLINT-f, FLINT-g denote the predictor and interpreter trained in our framework.

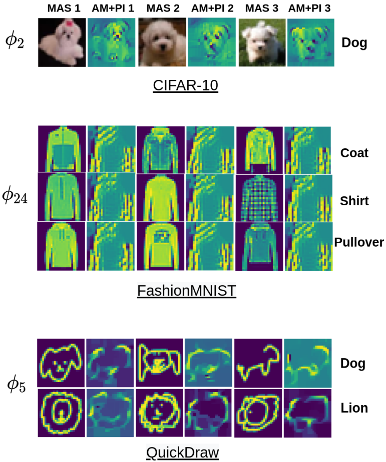
Dataset	LIME	VIBI	FLINT-g
MNIST	95.6±0.4	96.6±0.7	**98.7±0.1**
FashionMNIST	67.3±1.3	88.4±0.3	**91.5±0.1**
CIFAR-10	31.5±0.9	65.5±0.3	**93.2±0.2**
QuickDraw	76.3±0.1	78.6±0.4	**90.8±0.4**

Table: Results for fidelity to FLINT-f (in %)

Global Interpretations I



(a) Global relevances ($r_{j,c}$) for all class-attribute pairs for QuickDraw



(b) Sample class-attribute pairs with high relevance

Local Interpretations

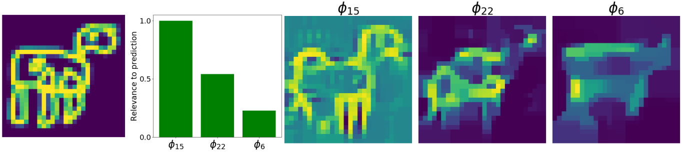


Figure: Local interpretation example. True label 'Cow'

Local Interpretations

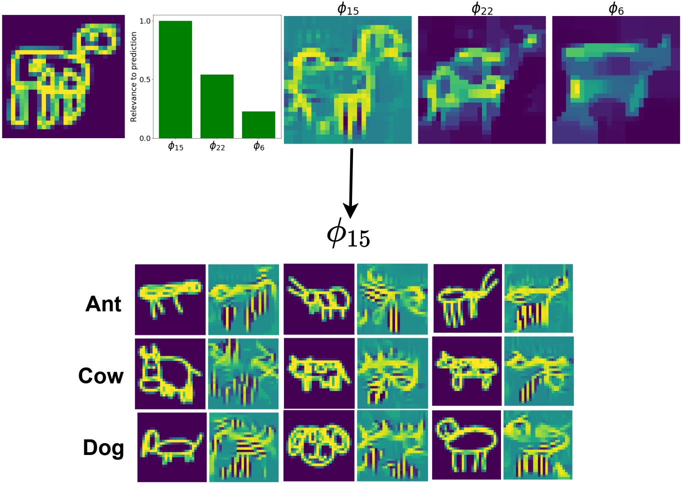


Figure: Local interpretation example. True label 'Cow'

Local Interpretations

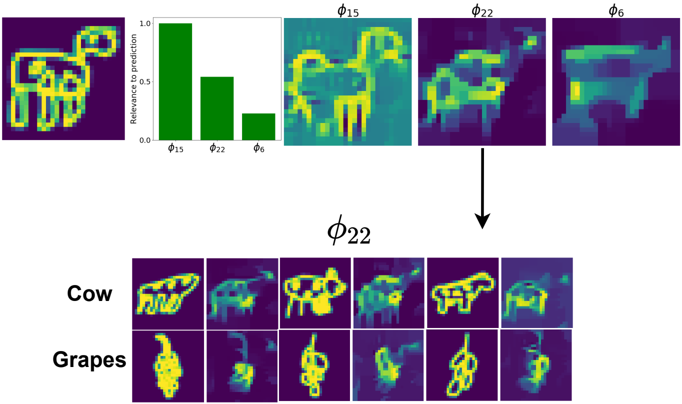


Figure: Local interpretation example. True label 'Cow'

Local Interpretations

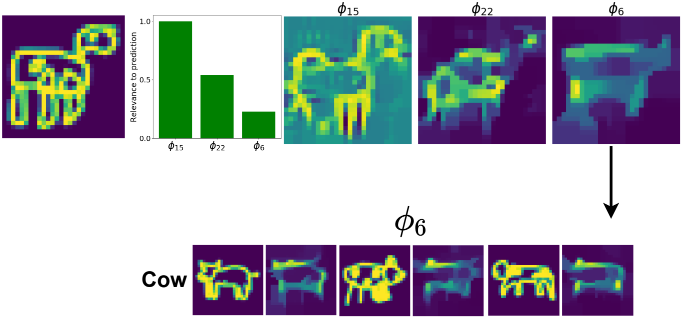


Figure: Local interpretation example. True label 'Cow'

Further Evaluation & Experiments

- **Subjective Evaluation**: Survey with 20 participants to evaluate meaningfulness of interpretations.

Further Evaluation & Experiments

- **Subjective Evaluation**: Survey with 20 participants to evaluate meaningfulness of interpretations. Visualization + textual description of an attribute. Asked to indicate agreement/disagreement

Further Evaluation & Experiments

- **Subjective Evaluation**: Survey with 20 participants to evaluate meaningfulness of interpretations. Visualization + textual description of an attribute. Asked to indicate agreement/disagreement

- **Post-hoc Experiments**: Interpreting the BASE-f model (trained only for accuracy).

Further Evaluation & Experiments

- **Subjective Evaluation**: Survey with 20 participants to evaluate meaningfulness of interpretations. Visualization + textual description of an attribute. Asked to indicate agreement/disagreement
- **Post-hoc Experiments**: Interpreting the BASE-f model (trained only for accuracy).
- Additional results on more complex datasets CIFAR100, CUB-200.

Further Evaluation & Experiments

- **Subjective Evaluation**: Survey with 20 participants to evaluate meaningfulness of interpretations. Visualization + textual description of an attribute. Asked to indicate agreement/disagreement
- **Post-hoc Experiments**: Interpreting the BASE-f model (trained only for accuracy).
- Additional results on more complex datasets CIFAR100, CUB-200.
- **Shuffling experiment**: Extreme test by shuffling attribute activations and observing drop in accuracy

Further Evaluation & Experiments

- **Subjective Evaluation**: Survey with 20 participants to evaluate meaningfulness of interpretations. Visualization + textual description of an attribute. Asked to indicate agreement/disagreement
- **Post-hoc Experiments**: Interpreting the BASE-f model (trained only for accuracy).
- Additional results on more complex datasets CIFAR100, CUB-200.
- **Shuffling experiment**: Extreme test by shuffling attribute activations and observing drop in accuracy
- Multiple ablation studies, more visualizations in supplementary

Conclusion & Future Work

- We have proposed a framework covering interpretable systems by design as well as generating post-hoc interpretations, which provides local and global interpretations in terms of high level attributes.

Conclusion & Future Work

- We have proposed a framework covering interpretable systems by design as well as generating post-hoc interpretations, which provides local and global interpretations in terms of high level attributes.
- To guarantee complete faithfulness, FLINT-g can always be used as the final prediction model.

Conclusion & Future Work

- We have proposed a framework covering interpretable systems by design as well as generating post-hoc interpretations, which provides local and global interpretations in terms of high level attributes.
- To guarantee complete faithfulness, FLINT-g can always be used as the final prediction model.
- Compression and interpretability through g.
- Application to other types of tasks, other input modalities. Search for different representations of attributes/concepts, adapt constraints according to task.

The End

THANK YOU!

For complete details please check out our paper + supplementary