Human-Driven FOL Explanations of Deep Learning

Gabriele Ciravegna,

SAILAB, University of Siena, Italy



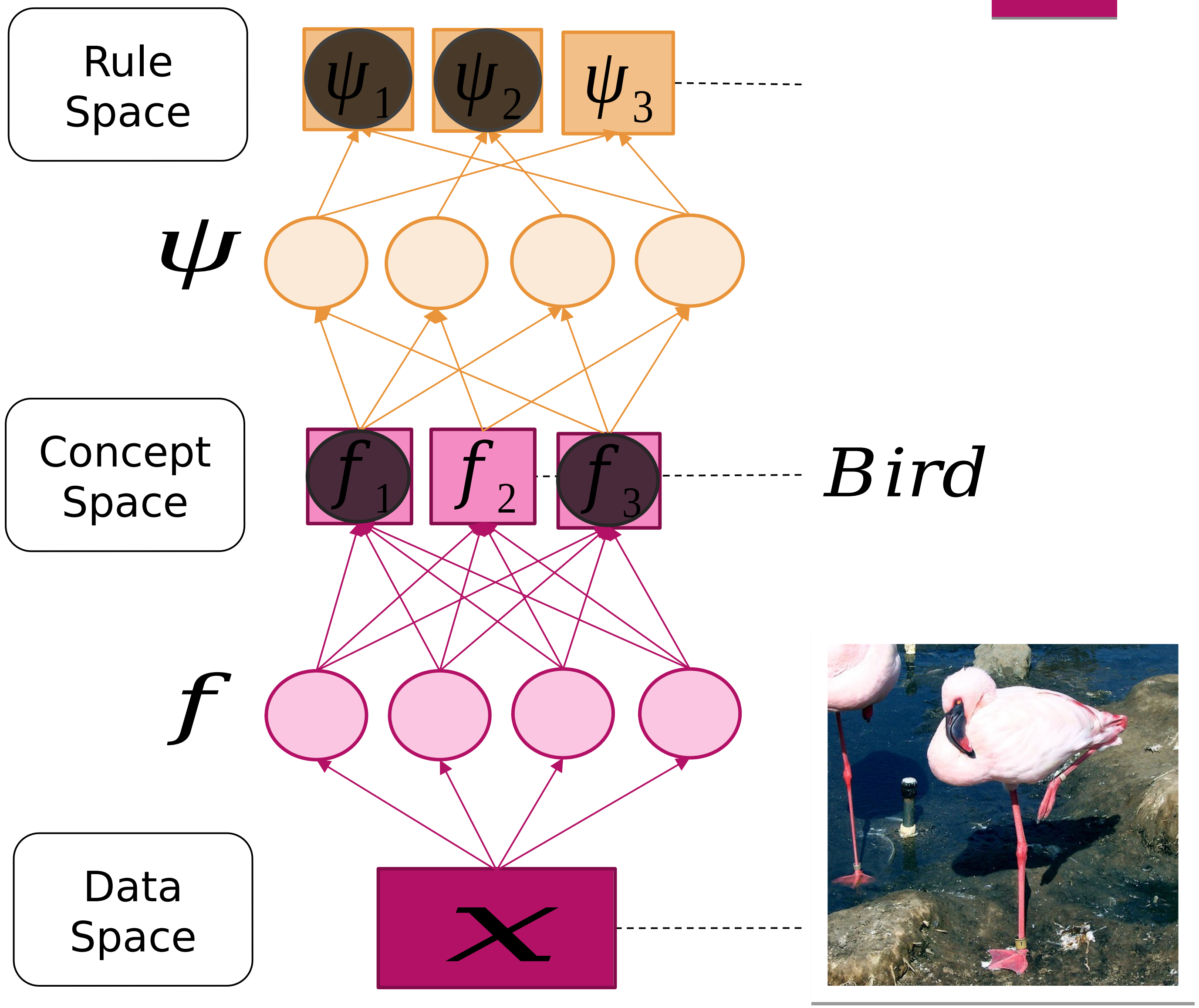
WHY?



NEURAL NETWORKS HAVE BECOME INCREDIBLY POWERFUL

Network explaining network

- Provides example-based explanation
- Extract new knowledge

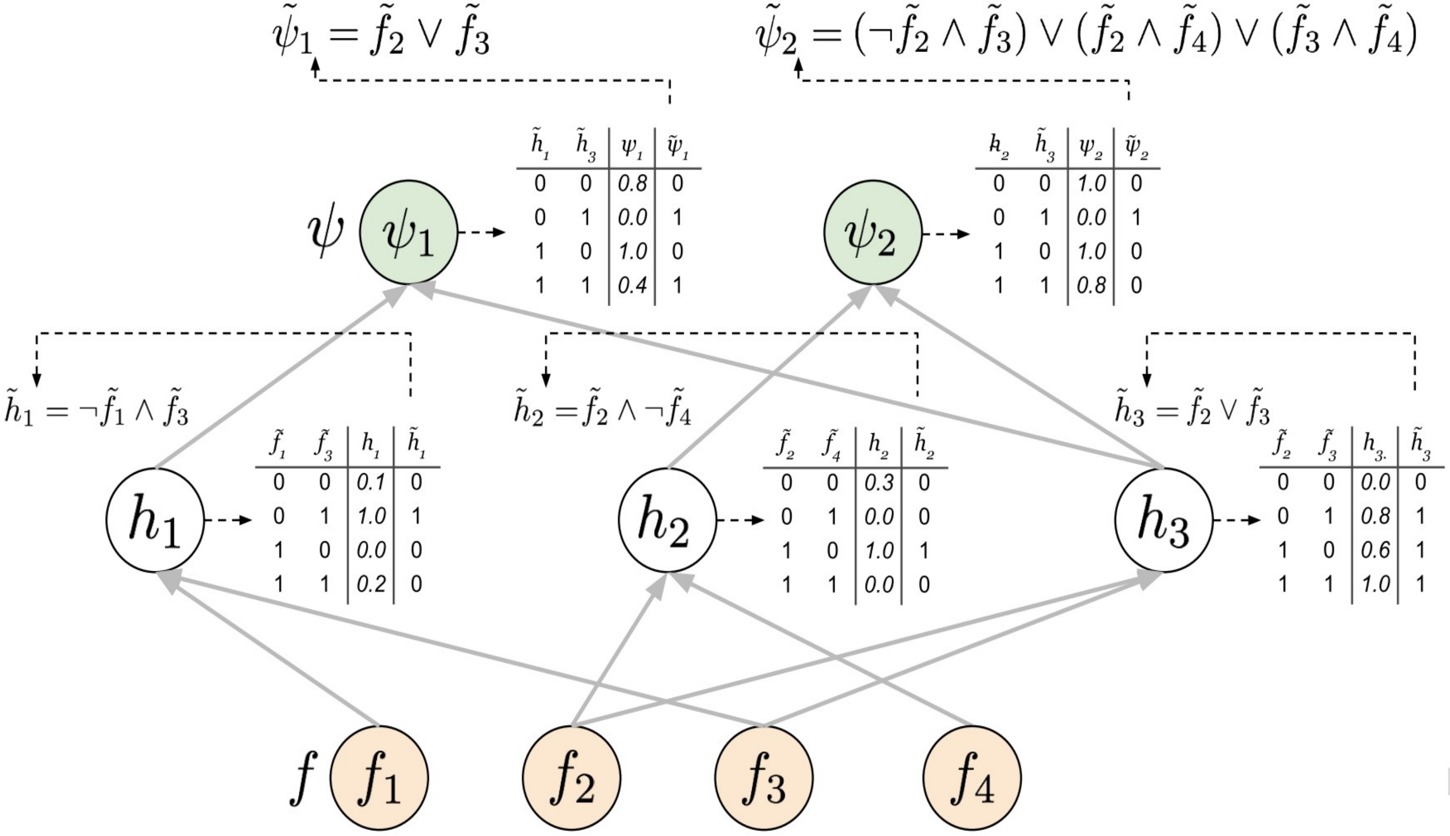


Network Pruning

- Strong L1 regularization
- Progressive pruning of the lowest weights

- **Low Fan-In Neurons**

Explainable Network



$\tilde{h}_1$	$\tilde{h}_3$	ψ_1	$\tilde{\psi}_1$
0	0	0.8	0
0	1	0.0	1
1	0	1.0	0
1	1	0.4	1

$\tilde{h}_2$	$\tilde{h}_3$	ψ_2	$\tilde{\psi}_2$
0	0	1.0	0
0	1	0.0	1
1	0	1.0	0
1	1	0.8	0

$\tilde{f}_1$	$\tilde{f}_3$	h_1	$\tilde{h}_1$
0	0	0.1	0
0	1	1.0	1
1	0	0.0	0
1	1	0.2	0

$\tilde{f}_2$	$\tilde{f}_4$	h_2	$\tilde{h}_2$
0	0	0.3	0
0	1	0.0	0
1	0	1.0	1
1	1	0.0	0

$\tilde{f}_2$	$\tilde{f}_3$	h_3	$\tilde{h}_3$
0	0	0.0	0
0	1	0.8	1
1	0	0.6	1
1	1	1.0	1

How does it works?

$$f^\star, \psi^\star = \arg\min_{f,\psi} \{U(f) + D(\psi, f)\}$$

$$U(f) = \left\{ \sum_{j=1}^{c} \sum_{x_k \in X_{\phi_j}} \hat{\phi}_j(f(x_k)) + \gamma_f \|f\| \right\}$$

$$\hat{\phi}_j(f(x)) = \|f(x) - y(x)\|^2$$

$$D(\psi, f) = \{-\hat{I}_{Y,\Psi}(\psi, f, X) + \gamma_\psi \|\psi\|\}$$

$$I_{Y,\Psi}(\psi, f) = H_\Psi(\psi, f) - H_{\Psi|Y}(\psi, f)$$

Regularization Effects

Baseline

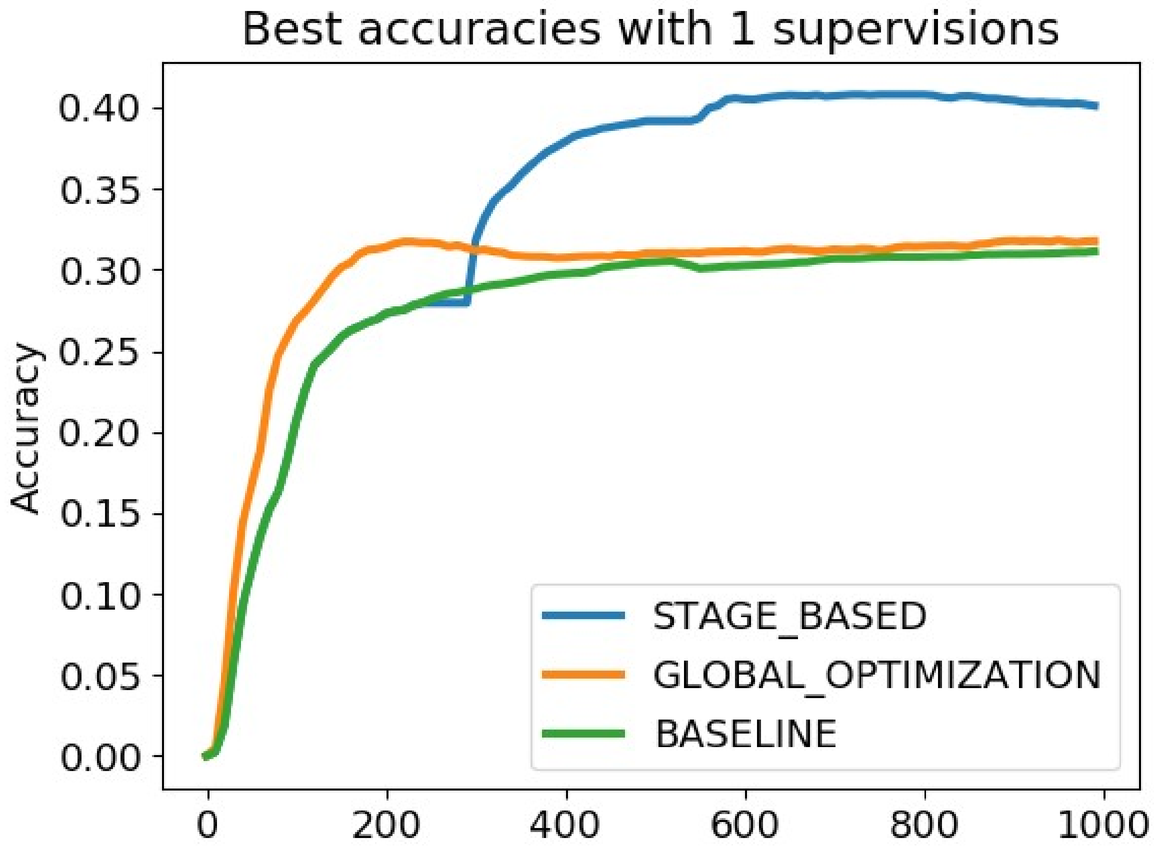
$$U\left(F\right)$$

Global Optimization

$$U\left(F\right)+D\left(f,\psi\right)$$

Stage-based Optimization

Performance Improvements - MNIST



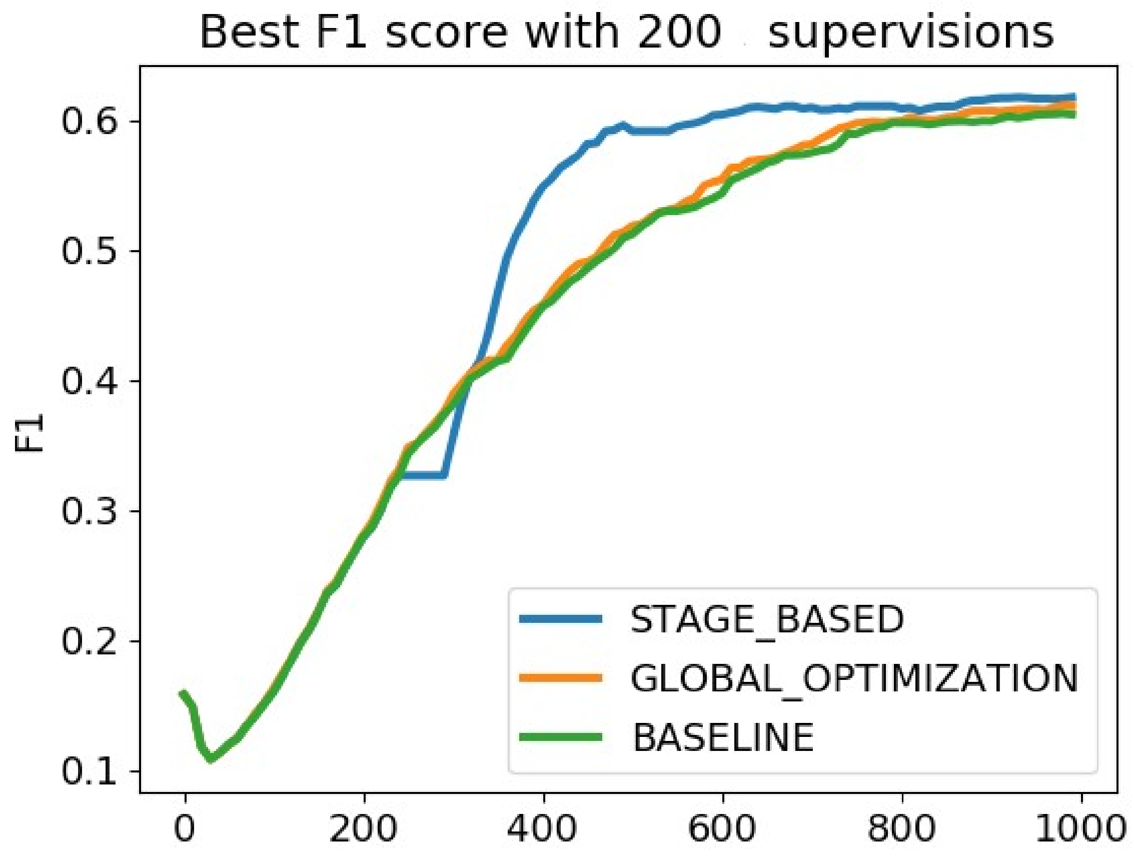
Overall Results - MNIST

# Labeled Examples	Baseline	Stage-based Optimization	Global Optimization
1	31.7 ± 2.1 %	**46.5 ± 1.4 %**	38.7 ± 3.7%
10	68.6 ± 0.4 %	71.5 ± 0.9%	**71.6 ± 1.7%**
100	84.5 ± 0.4 %	**86.7 ± 0.3 %**	86.0 ± 0.1%

Learned Rules

(a)	$\text{Odd} \dot{\vee} \text{Even}$ $\text{Nine} \wedge \neg\text{Even}$	$\text{Eight} \dot{\vee} \text{Zero}$ $\text{Six} \wedge \neg\text{Odd}$
(b)	$\text{Odd} \wedge \neg\text{Even} \wedge [(\text{One} \dot{\vee} \text{Five} \dot{\vee} \text{Seven} \dot{\vee} \text{Nine}) \vee$ $(\neg\text{One} \wedge \neg\text{Five} \wedge \neg\text{Seven} \wedge \neg\text{Nine})]$ $\text{Even} \wedge \neg\text{Odd} \wedge \neg\text{One} \wedge \neg\text{Five} \wedge \neg\text{Nine} \wedge (\text{Eight} \vee \neg\text{Eight})$	

Performance Improvements - PascalPart



Overall Results - PascalPart

# Labeled Examples	Baseline	Stage-based Optimization	Global Optimization
10	$54.0 \pm 0.3\%$	**$56.8 \pm 0.2\%$**	$55.8 \pm 0.7\%$
50	$60.5 \pm 0.2\%$	**$62.0 \pm 0.4\%$**	$61.3 \pm 0.5\%$
Whole dataset	$62.6 \pm 0.3\%$	**$63.8 \pm 0.2\%$**	$63.7 \pm 0.4\%$

Learned Rules

(a)	$\mathrm{Hand} \wedge \mathrm{Foot} \wedge \neg \mathrm{Beak}$, $\mathrm{Handlebar} \wedge \neg \mathrm{Boat}$,	$\neg \mathrm{Foot} \wedge \mathrm{Dog}$, $\neg \mathrm{AirplaneBody} \wedge \mathrm{Cow}$
(b)	$\mathrm{Beak} \wedge \neg \mathrm{Ear} \wedge \neg \mathrm{Nose}$, $\mathrm{TvMonitor} \wedge (\mathrm{Screen} \vee \mathrm{Table})$	$(\mathrm{Cat} \vee \mathrm{Dog}) \wedge (\neg \mathrm{Foot} \vee \mathrm{Paw})$, $\mathrm{Aeroplane} \wedge (\mathrm{Engine} \vee \mathrm{Stern})$

Following steps

Improve rules quality

- From Local to Global rules:
- Predicate explaining rules:

Framework Applications

- Adversarial Defense:
  Learned constraints as Attack detector

From Local to Global Explanations

$$\forall x \in X_j,\ \hat{\psi}_j(f(x)) \quad X \ = \ \bigcup_{j=1}^{m} X_j$$

CNF Conversion

$$\hat{\Psi} = \bigvee_{j=1}^{m} \hat{\psi}_j(f(x)) \equiv \bigwedge_{k=1}^{K} \hat{\psi}'_k(f(x))$$

$$\forall x \in X,\ \hat{\psi}'_k(f(x))$$

From Local to Class-Driven Explanations

Local Rule rewrite

$$\hat{\psi}_i(f(x)) = 1_{\psi_i}(f(x)) \leftrightarrow \psi_i(f(x)), \quad \forall x \in X,$$

Enforce Support

$$1_{\hat{\psi}_i(f(x))} = 1_{\hat{f}_i(x)}$$

Class-Driven Rule

$$\hat{\psi}_i(f(x)) = \hat{f}_i(x) \leftrightarrow \psi_i(f_k(x)), \ \forall x \in X, \ i, k \in (1, ..., n), \ k \neq i$$

$$\hat{\psi}_i(f(x)) = Man(x) \leftrightarrow Head(x) \vee Hand(x) \vee Foot(x), \ \forall x \in X$$

IFF ↔ or IF → ?

- **IFF ↔ rules:**

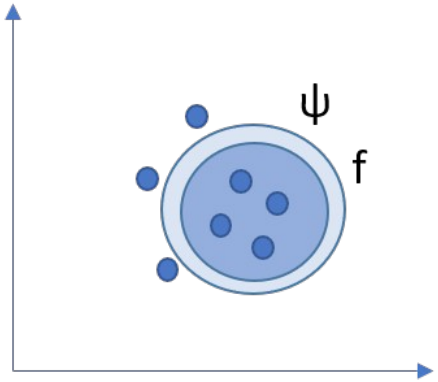
$$1_{\hat{\psi}_i(f(x))} = 1_{\hat{f}_i(x)}$$

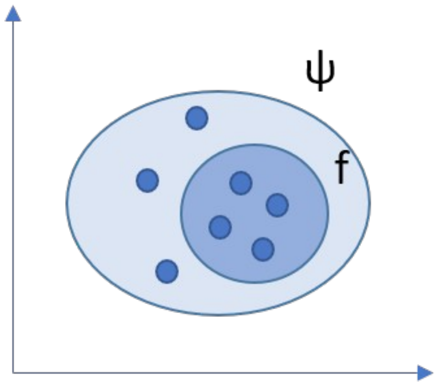
$$L_{\leftrightarrow}(\psi, f, X) = \sum_{i \in S, x \in X} |f_i(x) - \psi_{h(i)}(f(x))|$$

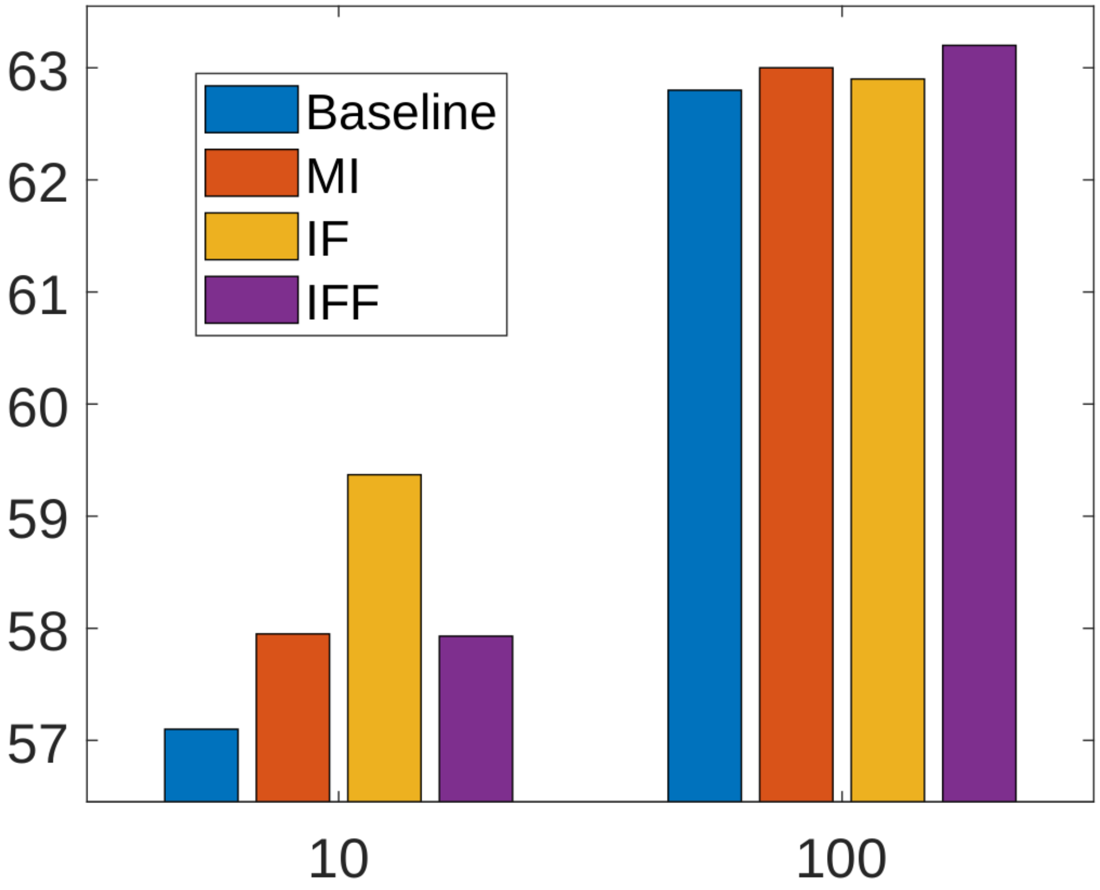
- **IF → rule:**

$$1_{\hat{\psi}_i(f(x))} \supseteq 1_{\hat{f}_i(x)}$$

$$L_{\rightarrow}(\psi, f, X) = \sum_{i \in P, x \in X} \max\{0, f_i(x) - \psi_{h(i)}(f(x))\}$$





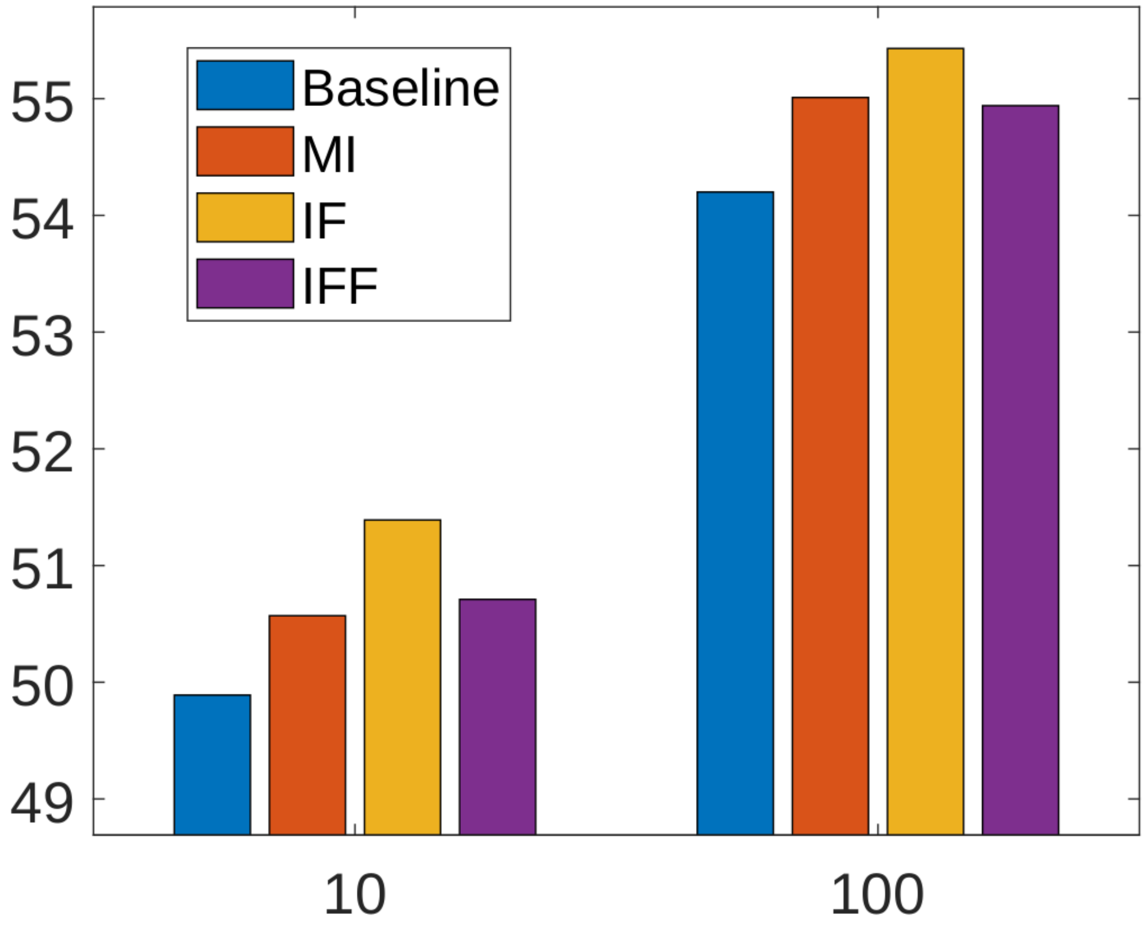


PascalPart Performances

Rules PascalPart

LOCAL	$\forall x \in X_{\psi_i},\ Beak \vee Bird$ $\forall x \in X_{\psi_j},\ Headlight \vee Plate$ $\forall x \in X_{\psi_k},\ Cat \vee Horse$
GLOBAL	$\forall x,\ AeroplaneBody \vee Beak \vee Bird \vee Table \vee Plant$ $\vee Car \vee Headlight \vee Motorbike \vee Muzzle \vee Train$ $\vee Chainwheel \vee \neg Aeroplane$ $\forall x, \neg Horse \vee AeroplaneBody \vee Beak \vee Bird \vee Train$ $\vee Car \vee Chainwheel \vee Headlight \vee Muzzle \vee Table$ $\vee Motorbike \vee Plant$
CLASS-DRIVEN $IF \rightarrow$	$\forall x,\ Car \rightarrow Backside \vee Mirror \vee (Window \wedge \neg Coach)$ $\forall x,\ Bicycle \rightarrow Saddle \vee Handlebar$ $\forall x,\ Train \rightarrow Coach \vee TrainHead$
CLASS-DRIVEN $IFF \leftrightarrow$	$\forall x,\ Horse \leftrightarrow (Hoof \wedge Ear) \vee (Hoof \wedge Neck)$ $\forall x,\ Bird \leftrightarrow Beak \wedge \neg Horn$ $\forall x,\ Bicycle \leftrightarrow (Chainwheel \wedge \neg Cow \wedge Handlebar)$ $\vee(Chainwheel \wedge \neg Cow \wedge Saddle)$

CelebA Performances



Rules CelebA

Local	$\forall x \in X_{\psi_i},\ Bangs \land \neg Bald$ $\forall x \in X_{\psi_j},\ StraightHair \land\ BushyEyebrows$ $\forall x \in X_{\psi_k},\ Female \land Attractive$
Global	$\forall x,\ Bangs \lor BlondHair \lor Blurry \lor Goatee$ $\lor StraightHair \lor WearHat$ $\lor \neg Attractive \lor \neg Female \lor \neg Male$ $\forall x,\ Blurry \lor Goatee \lor WearHat$ $\lor \neg Attractive \lor \neg BlackHair \lor \neg BlondHair$ $\lor \neg Female \lor \neg StraightHair$
Class-driven $IF \rightarrow$	$\forall x,\ Attractive \rightarrow PaleSkin \lor RosyCheeks$ $\lor(\neg Blurry \land \neg Chubby)$ $\forall x,\ Beard \rightarrow Goatee \lor Sideburns$ $\forall x,\ Old \rightarrow GrayHair \lor \neg Attractive$
Class-driven $IFF \leftrightarrow$	$\forall x,\ Bald \leftrightarrow \neg BlackHair \land \neg BrownHair$ $\land \neg StraightHair \land \neg WavyHair$ $\forall x,\ NotBald \leftrightarrow Bangs \lor BrownHair \lor WavyHair$ $\forall x,\ Male \leftrightarrow \neg WearLipstick \land \neg WearNecklace$

SAILab

Siena Artificial Intelligence Lab

Thank you for the attention