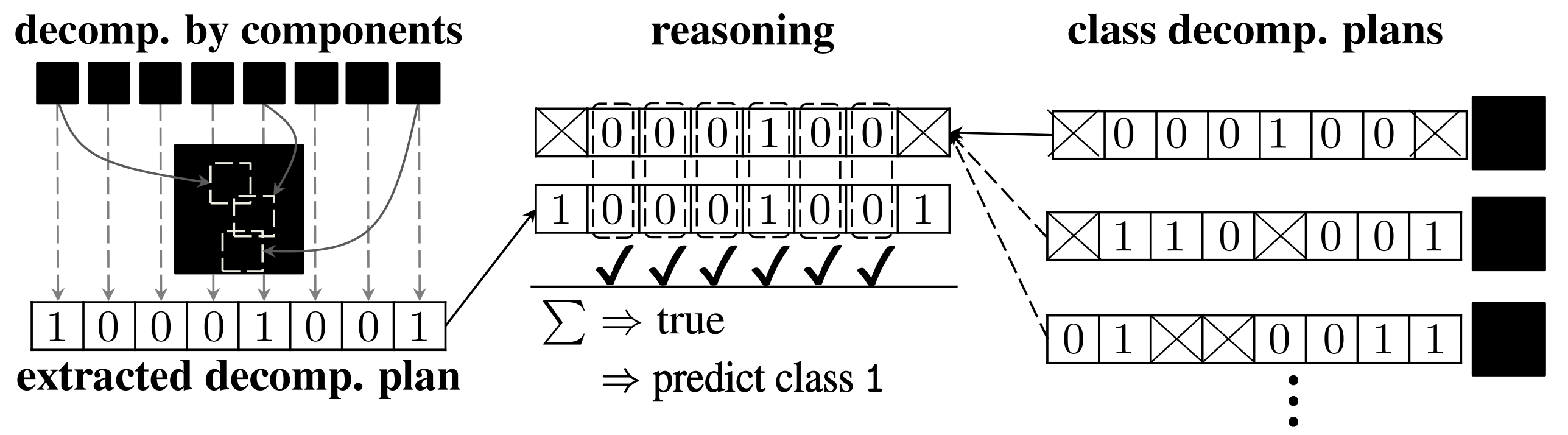
Classification-by-Components: Probabilistic Modeling of Reasoning over a Set of Components

Sascha Saralajew[1], Lars Holdijk[1], Maike Rees[1], Ebubekir Asan[1], Thomas Villmann[2]

[1]Dr. Ing. h.c. F. Porsche AG, [2]University of Applied Sciences Mittweida

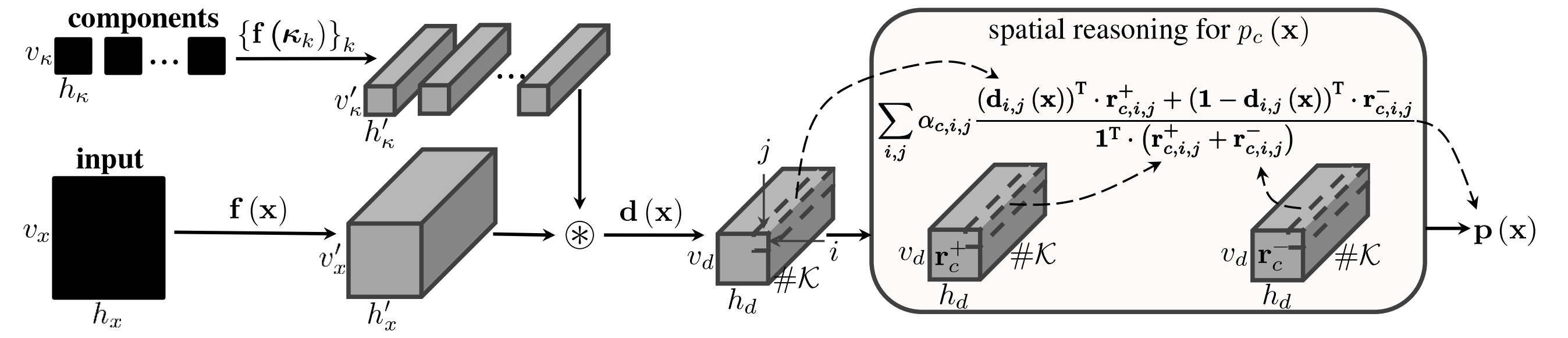
Introduction: Interpretability by Design

- Classification-By-Components networks (CBCs) are interpretable through their architecture
- Inspired by Recognition-by-Components theory [1]: ‘Humans recognize complex objects by decomposing them into components’
- Classification output is computed with a probability tree (no softmax)



1: detection / positive reasoning
0: no detection / negative reasoning
x: indefinite reasoning

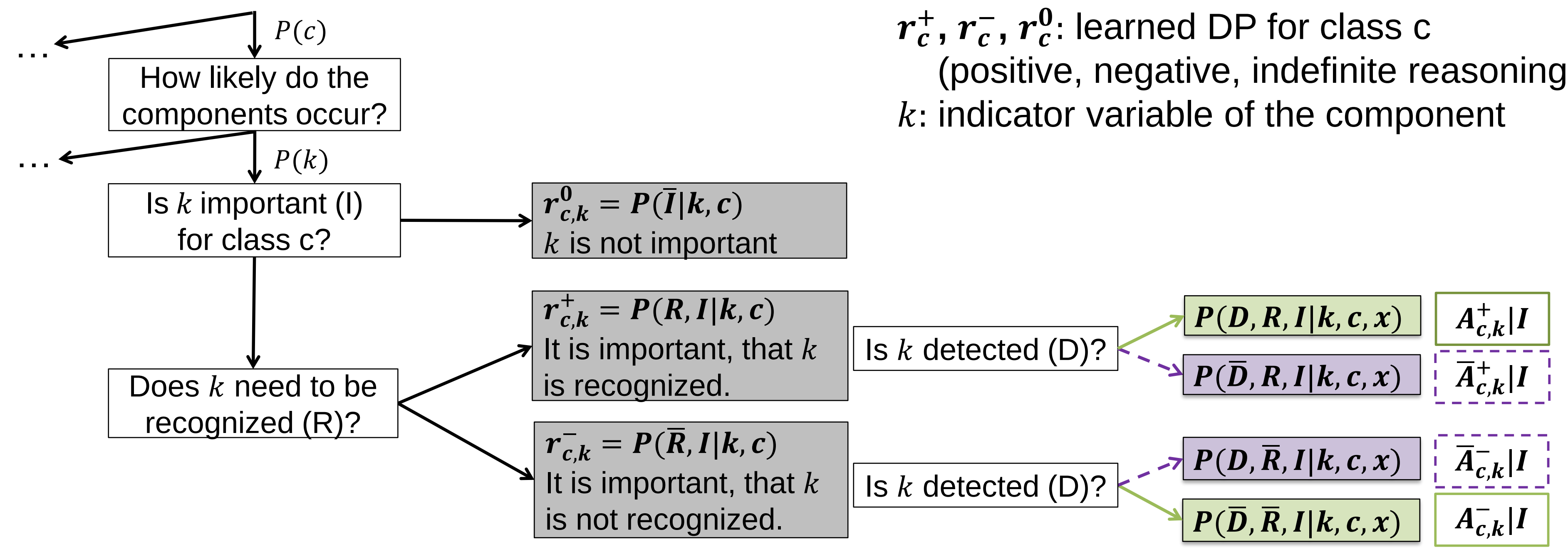
The Model: Components and Decomposition Plans


components
v_κ
h_κ
$\{\mathbf{f}(\boldsymbol{\kappa}_k)\}_k$
v'_κ
h'_κ
input
v_x
h_x
$\mathbf{f}(\mathbf{x})$
v'_x
h'_x
$\circledast$
$\mathbf{d}(\mathbf{x})$
v_d
h_d
j
i
$\#\mathcal{K}$
spatial reasoning for $p_c(\mathbf{x})$
$\sum_{i,j} \alpha_{c,i,j} \frac{(\mathbf{d}_{i,j}(\mathbf{x}))^\mathrm{T} \cdot \mathbf{r}^+_{c,i,j} + (\mathbf{1}-\mathbf{d}_{i,j}(\mathbf{x}))^\mathrm{T} \cdot \mathbf{r}^-_{c,i,j}}{\mathbf{1}^\mathrm{T} \cdot (\mathbf{r}^+_{c,i,j} + \mathbf{r}^-_{c,i,j})}$
$\mathbf{r}^+_c$
$\mathbf{r}^-_c$
$\mathbf{p}(\mathbf{x})$

Components: generic features, either full (input size) or patch (< input size)
Reasoning: one Decomposition Plan (DP) per class.
At inference: Match extracted DP with learned DPs.
Training: With contrastive loss and SGD

$$p_c(\boldsymbol{x}) = \frac{(\boldsymbol{d}(\boldsymbol{x}))^T * \boldsymbol{r}_c^+ + (\boldsymbol{1}-\boldsymbol{d}(\boldsymbol{x}))^T * \boldsymbol{r}_c^-}{\boldsymbol{1}^T * (\boldsymbol{1}-\boldsymbol{r}_c^0)}$$

$\boldsymbol{x}$: input
$\boldsymbol{d}(\boldsymbol{x})$: component detection possibility vector (DP extracted from the input)
$p_c(\boldsymbol{x})$: class hypothesis probability for class c
$\boldsymbol{r}_c^+, \boldsymbol{r}_c^-, \boldsymbol{r}_c^0$: learned DP for class c (positive, negative, indefinite reasoning)
k: indicator variable of the component



How likely do the components occur?

Is k important (I) for class c?

$\boldsymbol{r}_{c,k}^0 = \boldsymbol{P}(\bar{\boldsymbol{I}}|\boldsymbol{k},\boldsymbol{c})$
k is not important

Does k need to be recognized (R)?

$\boldsymbol{r}_{c,k}^+ = \boldsymbol{P}(\boldsymbol{R},\boldsymbol{I}|\boldsymbol{k},\boldsymbol{c})$
It is important, that k is recognized.

$\boldsymbol{r}_{c,k}^- = \boldsymbol{P}(\bar{\boldsymbol{R}},\boldsymbol{I}|\boldsymbol{k},\boldsymbol{c})$
It is important, that k is not recognized.

Is k detected (D)?

$\boldsymbol{P}(\boldsymbol{D},\boldsymbol{R},\boldsymbol{I}|\boldsymbol{k},\boldsymbol{c},\boldsymbol{x})$ — $\boldsymbol{A}_{c,k}^+|\boldsymbol{I}$

$\boldsymbol{P}(\bar{\boldsymbol{D}},\boldsymbol{R},\boldsymbol{I}|\boldsymbol{k},\boldsymbol{c},\boldsymbol{x})$ — $\bar{\boldsymbol{A}}_{c,k}^+|\boldsymbol{I}$

Is k detected (D)?

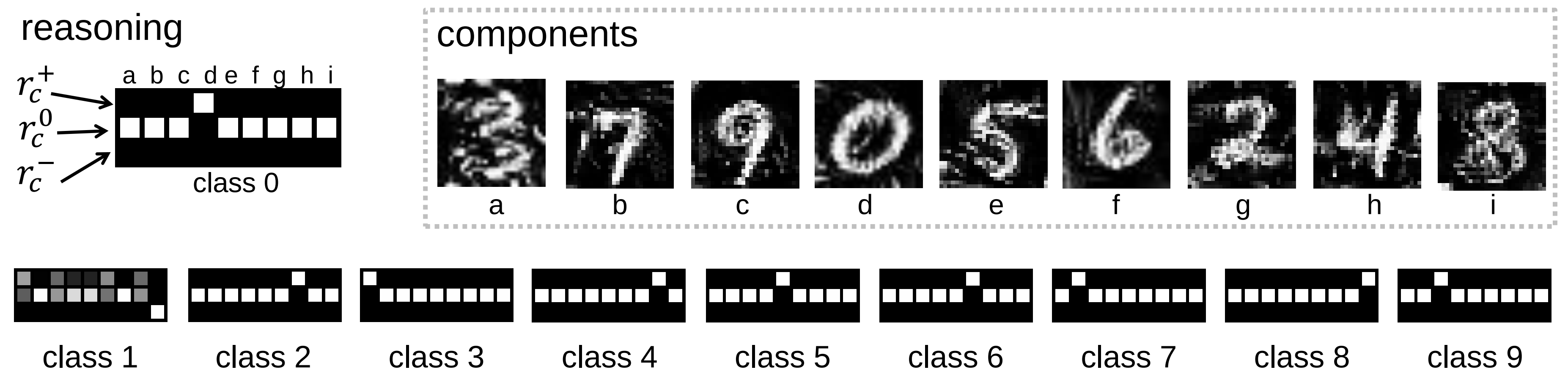
$\boldsymbol{P}(\boldsymbol{D},\bar{\boldsymbol{R}},\boldsymbol{I}|\boldsymbol{k},\boldsymbol{c},\boldsymbol{x})$ — $\bar{\boldsymbol{A}}_{c,k}^-|\boldsymbol{I}$

$\boldsymbol{P}(\bar{\boldsymbol{D}},\bar{\boldsymbol{R}},\boldsymbol{I}|\boldsymbol{k},\boldsymbol{c},\boldsymbol{x})$ — $\boldsymbol{A}_{c,k}^-|\boldsymbol{I}$

Results: Interpretable Classification

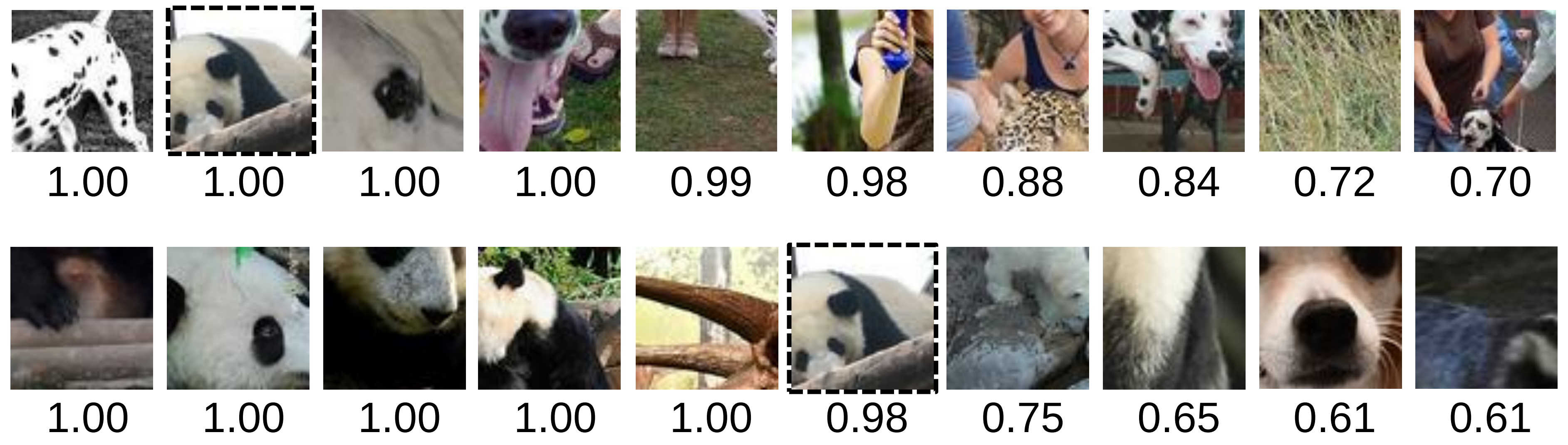
Components are prototypical and interpretable (MNIST):

To classify a 1, the CBC depends on negative (not look like an 8) and positive reasoning. For the other classes, the CBC depends on strong positive reasoning over one and indefinite reasoning over all the other components.



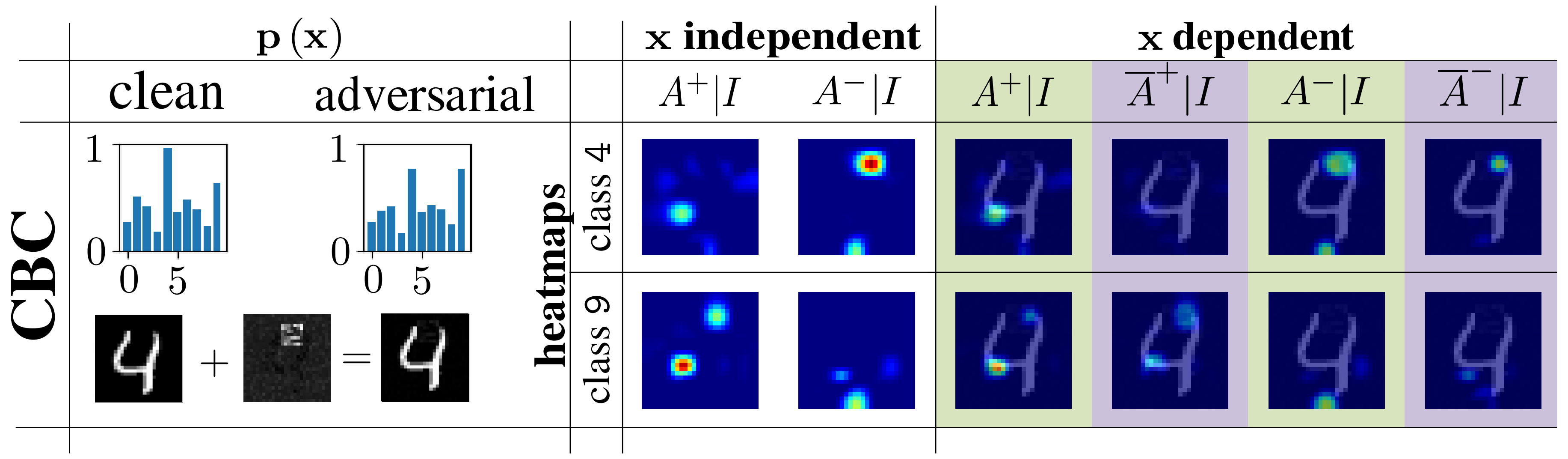
Components are conceptionally meaningful (ImageNet):

The components with the highest $r^+_{c,k}$ for *dalmatian* (top) and *giant panda* (bottom) are displayed. Components are shared among classes which shows that CBCs are capable of learning complex, class independent structures.



Reasoning explains why an adversarial attack succeeds (MNIST):

The CBC learns to recognize only as few parts as needed to distinguish between classes. For the 4, it checks that there is no stroke at the bottom ($\boldsymbol{A}^-|\boldsymbol{I}$) but a corner at the left ($\boldsymbol{A}^+|\boldsymbol{I}$). The adversary attacks a 4 to being classified as a 9 by inserting exactly enough noise to make the CBC believe that there is a stroke at the top ($\bar{\boldsymbol{A}}^-|\boldsymbol{I}$).



Accuracies: MNIST: 99.32 (CNN: 99.8), ImageNet: 82.4 (AlexNet: 82.8)

This paper in 30seconds!

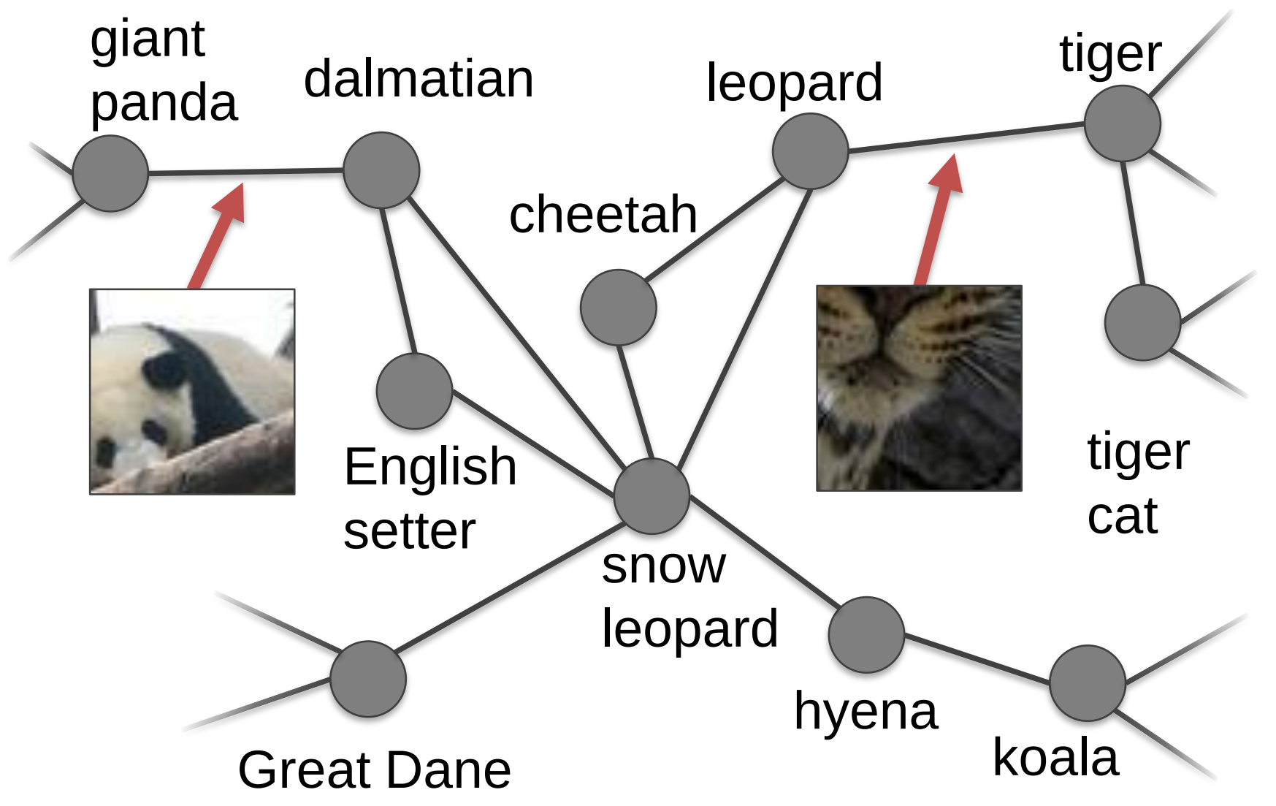
- Classification-by-Components networks are probabilistic and interpretable by design.
- They can be stacked onto almost every existing classification NN with almost no computational overhead at inference.
- The concept is to decompose complex objects into components and then compare the extracted with a learned class decomposition plan.

Conclusion: Interpretability at Scale

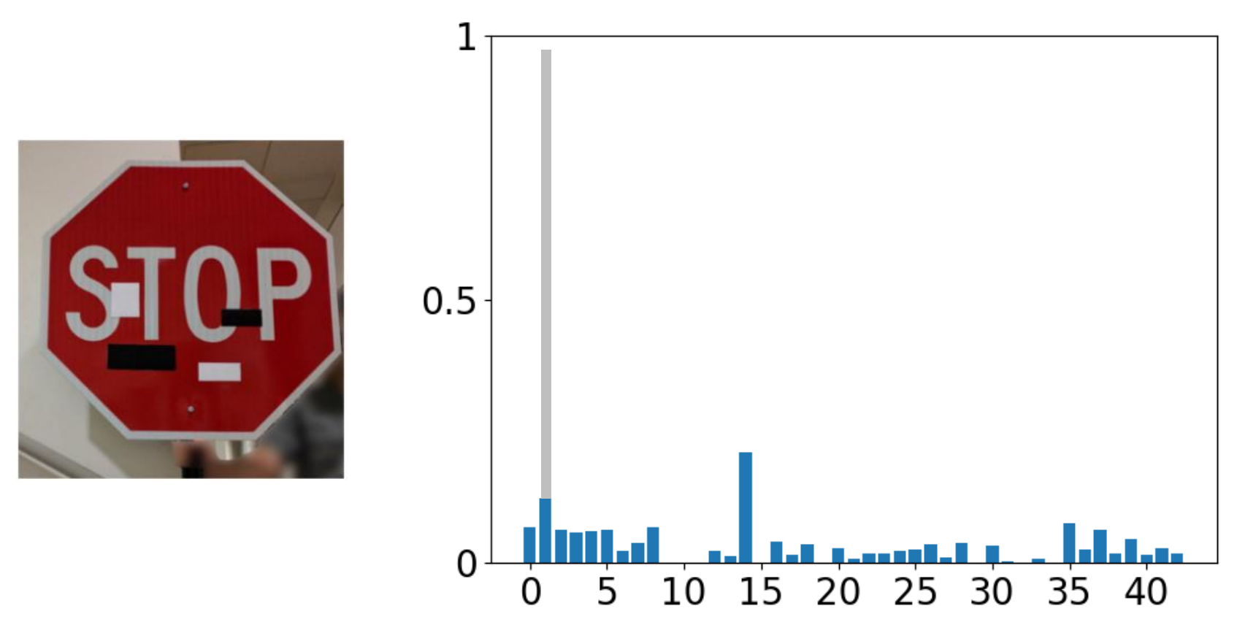
- Probabilistic framework for the last and penultimate layer of a NN
- Intuitive visualizations by back projection, heatmaps of DP for agreement ($\boldsymbol{A}$) and disagreement ($\bar{\boldsymbol{A}}$)
- Achieves satisfying accuracies for different datasets with small or large number of classes.

Current Work: Exciting Improvements in the Making

Investigation of components, that are shared by classes, can reveal a hierarchical structure between classes.



The class hypothesis possibility vector of a CBC might be used to reject adversarial examples.



CBC: *stop* with 21%

CNN: *speed limit 30* with 99%

Inspirations and Theory

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[2] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206, 2019.