Contrastive Relevance Propagation for Interpreting Predictions by a Single-Shot Object Detector

Hideomi Tsunakawa[1], Yoshitaka Kameya[1],
Hanju Lee[2], Yosuke Shinya[2], and Naoki Mitsumoto[2]

[1]Department of Information Engineering, Meijo University
[2]DENSO CORPORATION

Outline

- Background
- Proposed method: CRP
- Experiments

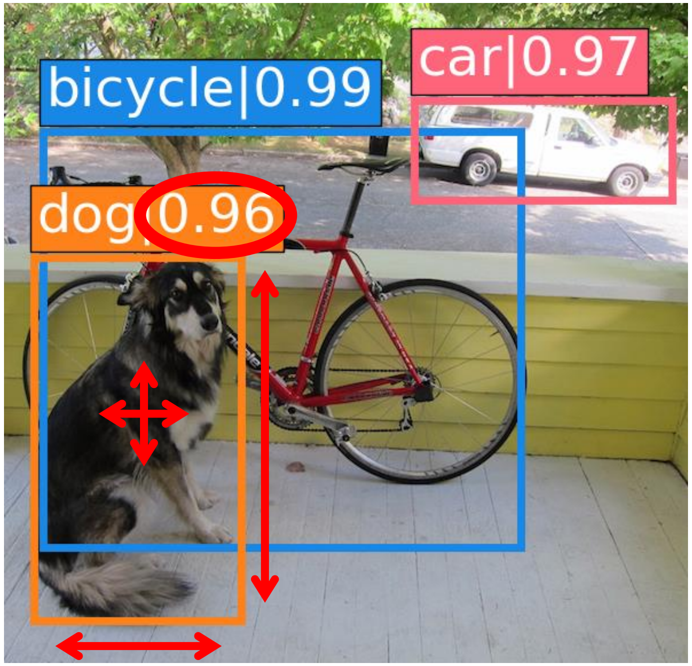
Outline

- Background
- Proposed method: CRP
- Experiments

Background: SSD (1)

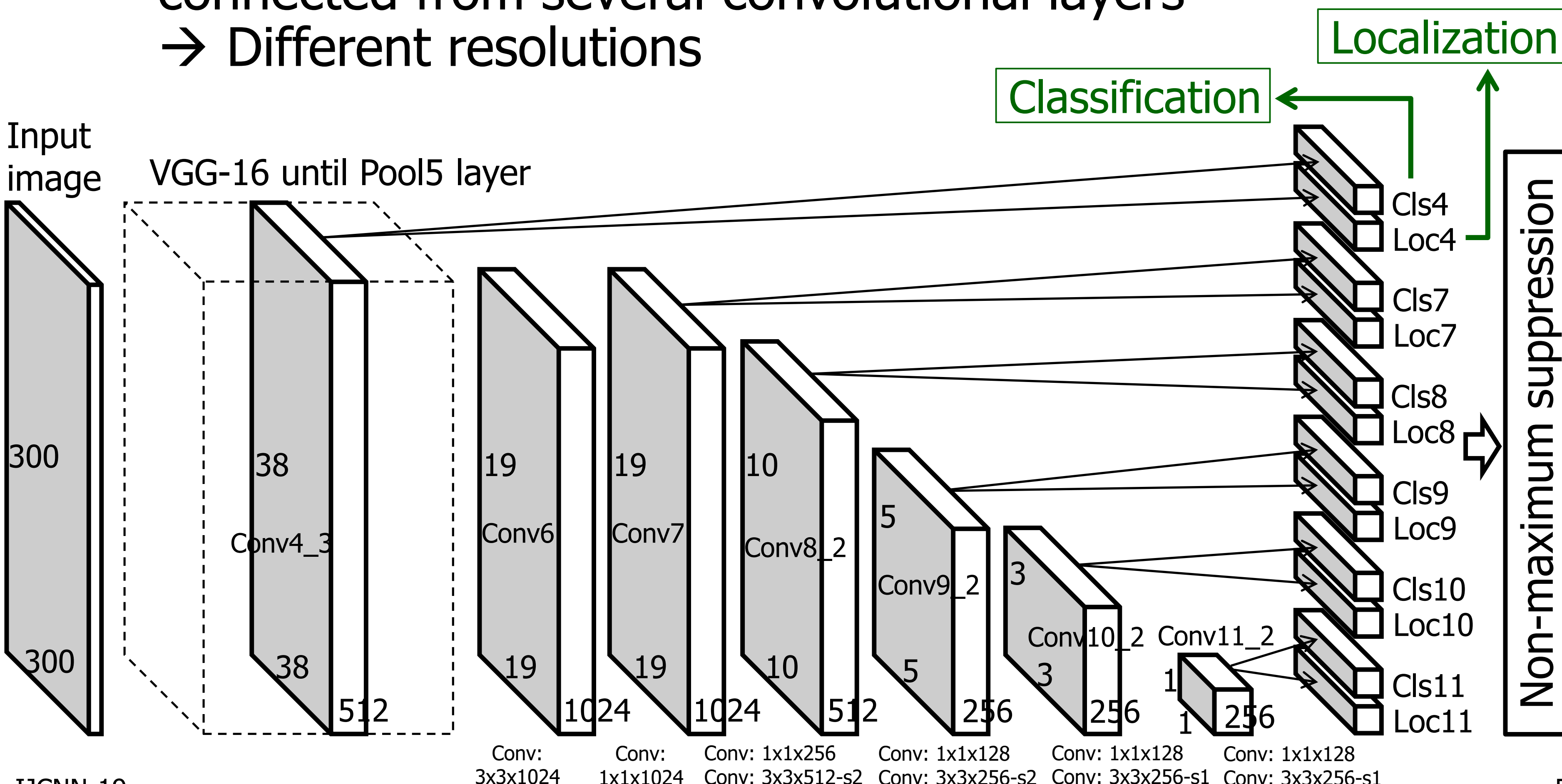
- Object detection is a well-known task in computer vision
- SSD (Single-Shot MultiBox Detector) [Liu+ ECCV-16]:
  - Known for its high speed and accuracy
  - Outputs:
    - Confidences for classes — Classification
    - Location offsets — Localization
      (center on x-axis, center on y-axis, width, height)

Input: 

Output: 

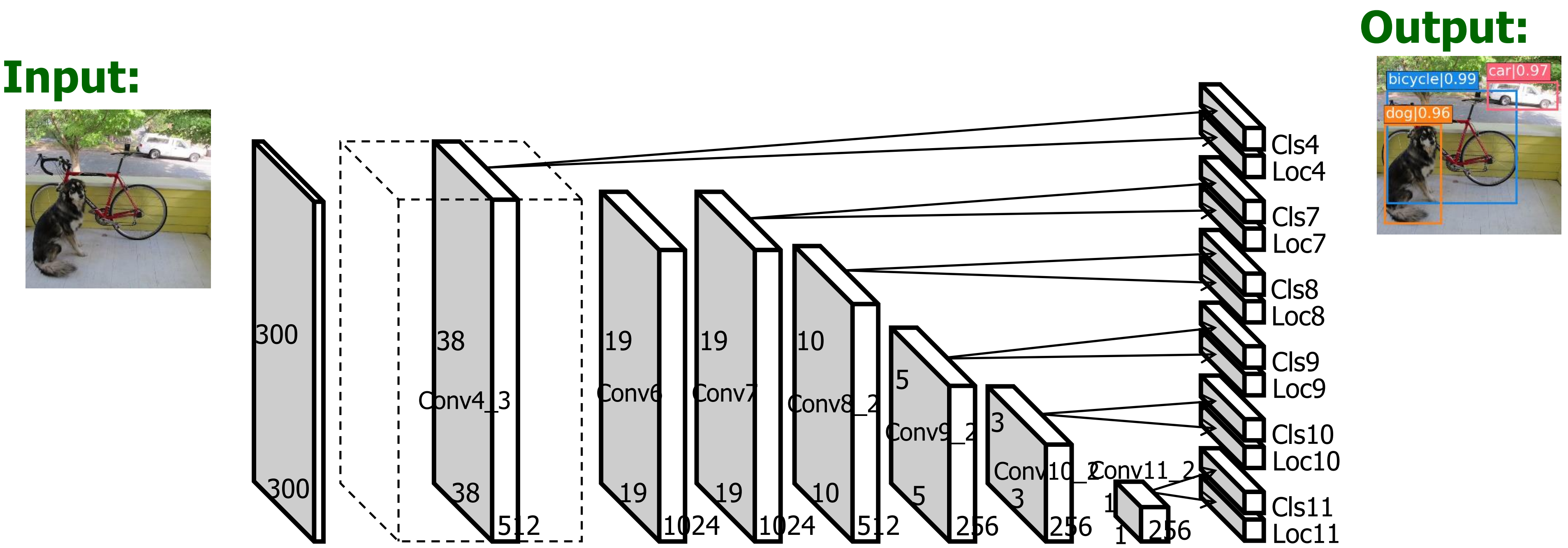
Background: SSD (2)

- SSD:
  - Based on a (large) single convolutional network
  - Layers for classification and layers for localization are connected from several convolutional layers
    → Different resolutions



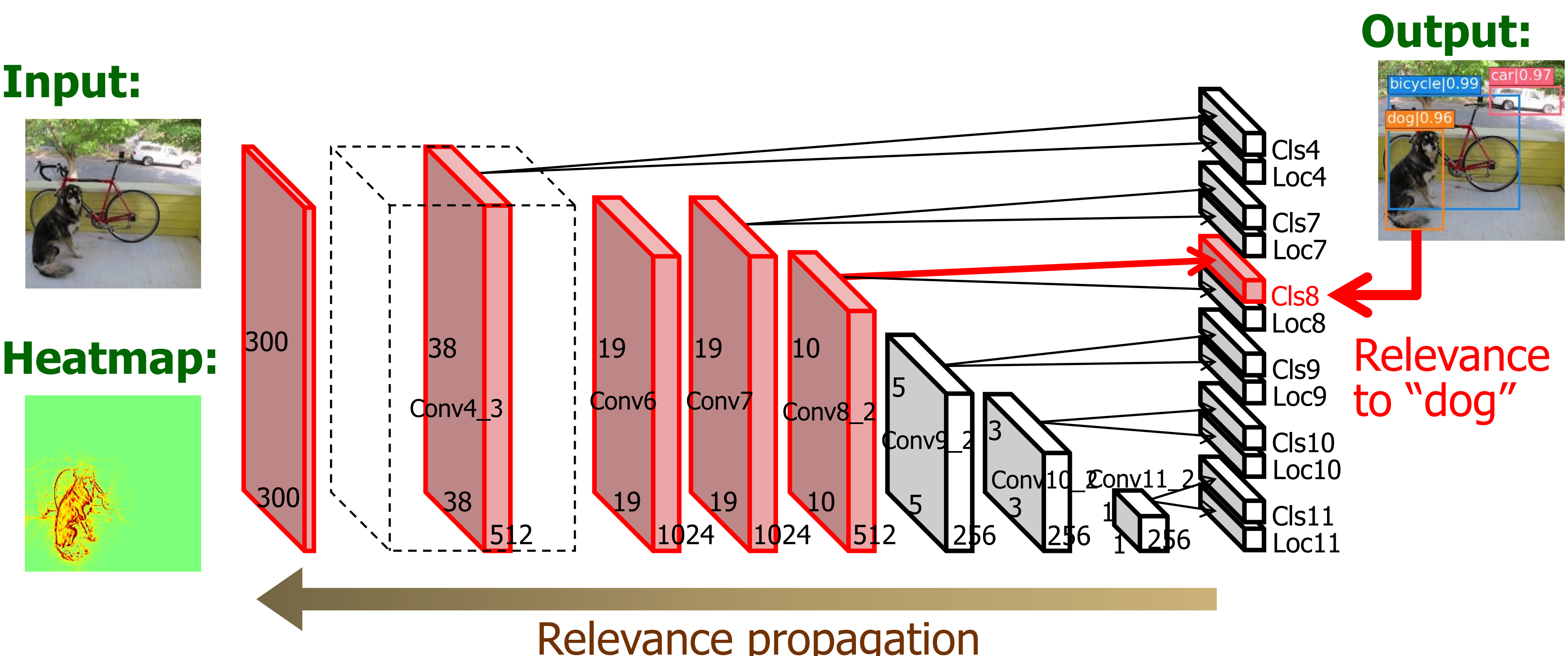
Background: LRP (1)

- LRP (Layer-wise Relevance Propagation) [Bach+ 15]:
  - Often used for interpreting predictions of DNNs



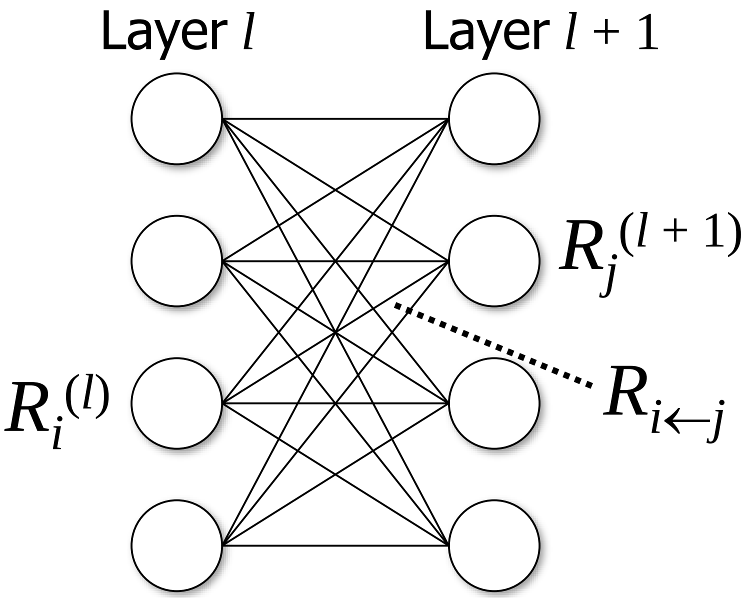
Background: LRP (1)

- LRP (Layer-wise Relevance Propagation) [Bach+ 15]:
  - Often used for interpreting predictions of DNNs
  - Propagates relevance backward from the output to the input features
  - Creates a heatmap using relevance at the input features



Background: LRP (2)

- LRP is equipped with several propagation rules:
  - Common:
    $R_j^{(l+1)}$: distributed to lower units
    $R_i^{(l)} := \sum_j R_{i \leftarrow j}$
    $R_{i \leftarrow j}$: passed through connection



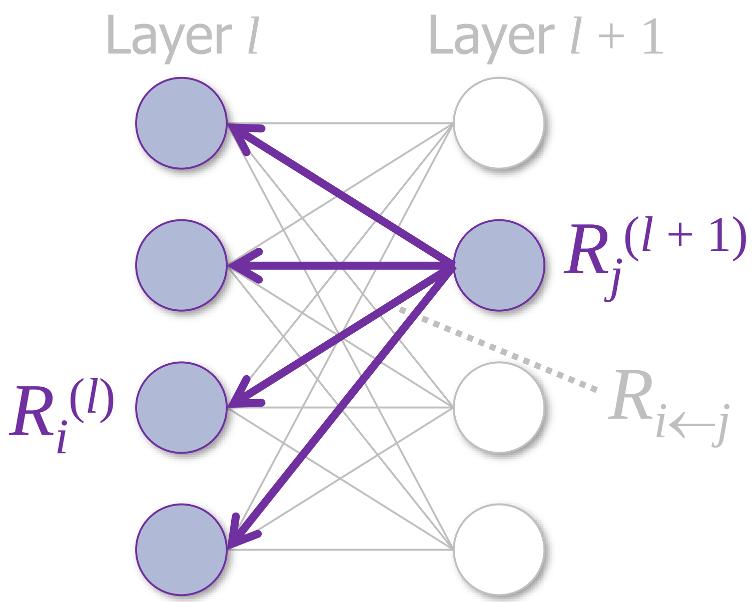
Background: LRP (2)

- LRP is equipped with several propagation rules:
  - Common:

    $R_j^{(l+1)}$: distributed to lower units

    $R_i^{(l)} := \sum_j R_{i \leftarrow j}$

    $R_{i \leftarrow j}$: passed through connection



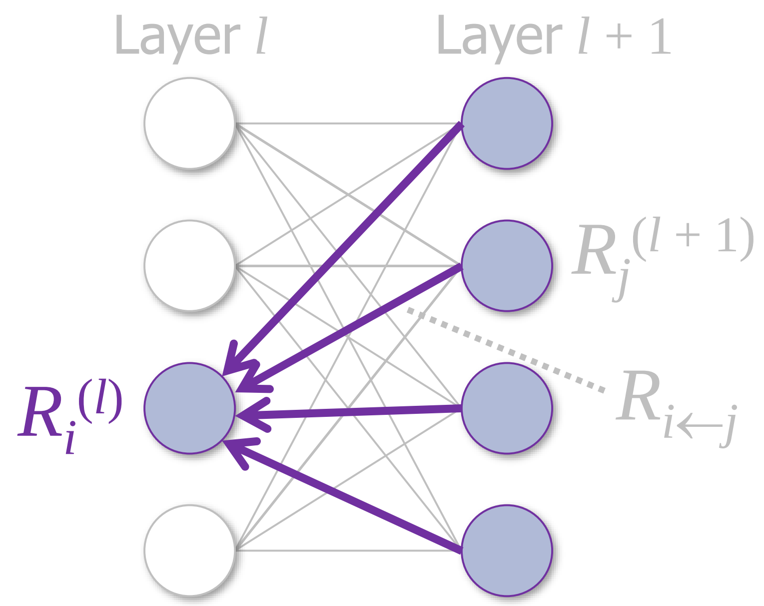
Background: LRP (2)

- LRP is equipped with several propagation rules:
  - Common:

    $R_j^{(l+1)}$: distributed to lower units

    $R_i^{(l)} := \sum_j R_{i \leftarrow j}$

    $R_{i \leftarrow j}$: passed through connection



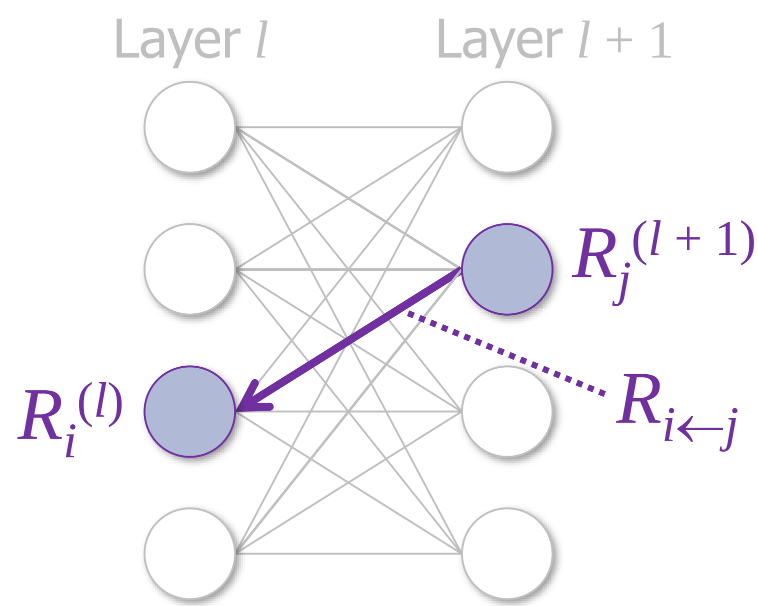
Background: LRP (2)

- LRP is equipped with several propagation rules:
  - Common:

    $R_j^{(l+1)}$: distributed to lower units

    $R_i^{(l)} := \sum_j R_{i\leftarrow j}$

    $R_{i\leftarrow j}$: passed through connection



  - Simple LRP:

$$R_{i\leftarrow j} = \frac{w_{ij}x_i}{\sum_{i'} w_{i'j}x_{i'}} R_j$$

  - ε -LRP:

$$R_{i\leftarrow j} = \frac{w_{ij}x_i}{\sum_{i'} w_{i'j}x_{i'} + \varepsilon \cdot \mathrm{sign}\left(\sum_{i'} w_{i'j}x_{i'}\right)} R_j$$

  - $\alpha\beta$ -LRP:

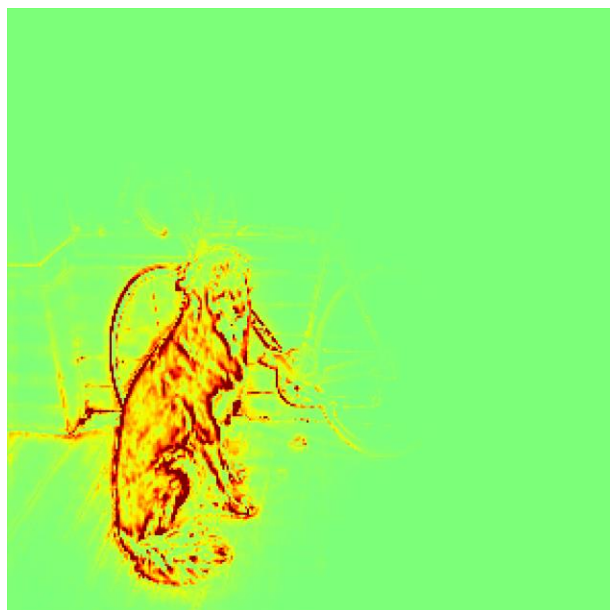
$$R_{i\leftarrow j} = \left(\alpha \frac{w_{ij}^{+}x_i}{\sum_{i'} w_{i'j}^{+}x_{i'}} + \beta \frac{w_{ij}^{-}x_i}{\sum_{i'} w_{i'j}^{-}x_{i'}}\right) R_j$$

$$w_{ij}^{+} \stackrel{\mathrm{def}}{=} \max\{w_{ij}, 0\}$$
$$w_{ij}^{-} \stackrel{\mathrm{def}}{=} \min\{w_{ij}, 0\}$$

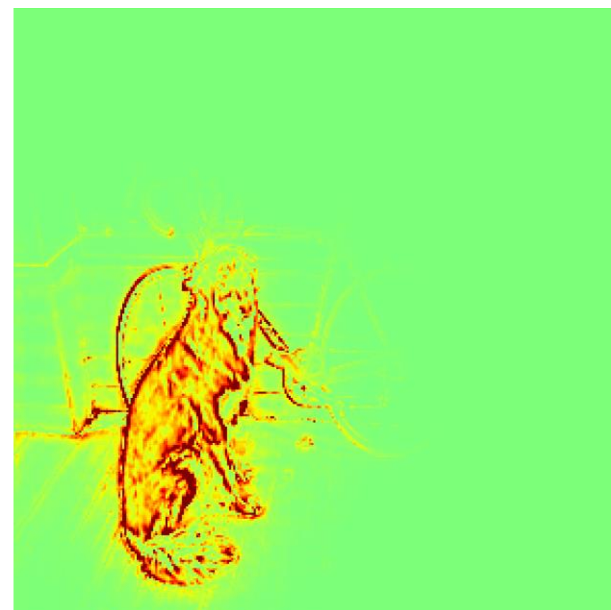
Background: Indistinguishable Heatmaps (1)

- Heatmaps are almost **invariant** even when the target class has been changed
- Heatmaps obtained with $\alpha\beta$ -LRP ($\alpha = 1$, $\beta = 0$):

Target class: “dog”
(actually predicted)



Target class: “cat”
(“what-if” analysis)



Background: Indistinguishable Heatmaps (2)

- Relevance propagated in each layer:

Layer	Relevance for 'dog'				Relevance for 'cat'			
	Max.	95%-tile	Median	Min.	Max.	95%-tile	Median	Min.
Cls8	1.82E-02	0	0	0	2.61E-02	0	0	0
Conv8_2	3.32E-03	3.03E-05	0	0	2.89E-03	3.00E-05	0	0
Conv8_1	3.23E-03	5.54E-06	0	0	3.19E-03	5.41E-06	0	0
Conv7	6.70E-03	0	0	0	7.17E-03	0	0	0
Conv6	2.61E-03	1.22E-05	0	0	2.78E-03	1.16E-05	0	0
Pool5	1.67E-02	0	0	0	1.61E-02	0	0	0
Conv5_3	3.33E-03	9.27E-06	0	0	3.32E-03	8.93E-06	0	0
Conv5_2	4.32E-03	1.00E-05	0	0	4.13E-03	9.66E-06	0	0
Conv5_1	3.05E-03	2.03E-05	0	0	2.92E-03	1.99E-05	0	0
Pool4	3.05E-03	0	0	0	2.92E-03	0	0	0
Conv4_3	9.78E-04	2.89E-06	0	0	9.61E-04	2.82E-06	0	0
Conv4_2	6.41E-04	3.46E-06	0	0	6.35E-04	3.38E-06	0	0
Conv4_1	9.04E-04	1.19E-05	0	0	8.87E-04	1.17E-05	0	0
Pool3	9.04E-04	3.47E-08	0	0	8.87E-04	3.11E-08	0	0
Conv3_3	3.63E-04	2.93E-06	0	0	3.80E-04	2.90E-06	0	0
Conv3_2	1.93E-04	3.27E-06	0	0	2.02E-04	3.25E-06	0	0
Conv3_1	3.71E-04	7.21E-06	0	0	3.89E-04	7.17E-06	0	0
Pool2	3.71E-04	2.76E-07	0	0	3.89E-04	2.63E-07	0	0
Conv2_2	1.41E-04	1.73E-06	0	0	1.38E-04	1.72E-06	0	0
Conv2_1	1.90E-04	3.54E-06	2.04E-11	0	1.99E-04	3.52E-06	1.79E-11	0
Pool1	1.90E-04	2.06E-07	0	0	1.99E-04	2.00E-07	0	0
Conv1_2	1.13E-04	6.88E-07	0	0	1.19E-04	6.85E-07	0	0
Conv1_1	3.60E-04	2.20E-05	2.37E-08	0	3.79E-04	2.21E-05	2.09E-08	0
Input	3.60E-04	2.20E-05	2.37E-08	0	3.79E-04	2.21E-05	2.09E-08	0

Relevance decreases exponentially

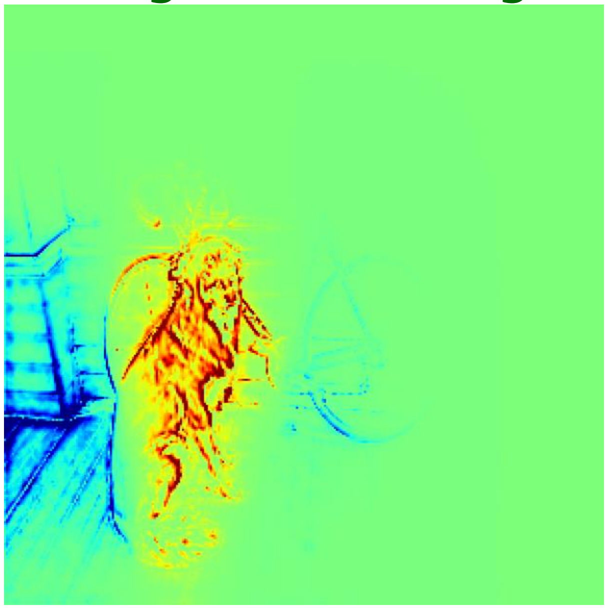
Background: Indistinguishable Heatmaps (3)

- Recent works that seem to support our observation:
  - [Adebayo+ NeurIPS-18]:
    - Uses Inception v3 (a large network)
    - If relevance = gradient $\times$ input, the input part dominates
      → Heatmaps will be invariant
      (since the input is of course fixed)
  - [Ancona+ ICLR-18]:
    - Several methods tend to return similar heatmaps (theoretically or empirically):
      - Gradient $\times$ input
      - DeepLIFT (Rescale)
      - Integrated Gradients
      - Simple LRP

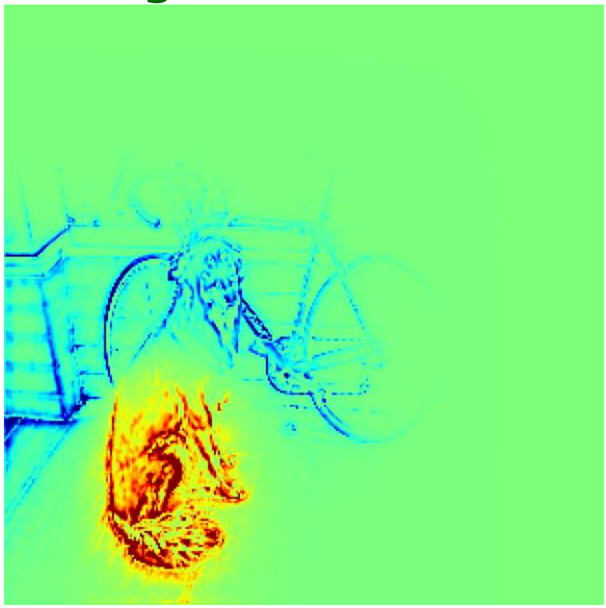
Background: Our Motivation

- We introduce **contrastive relevance** that highlights the more important part to the target class

Target class: “dog”



Target class: “cat”



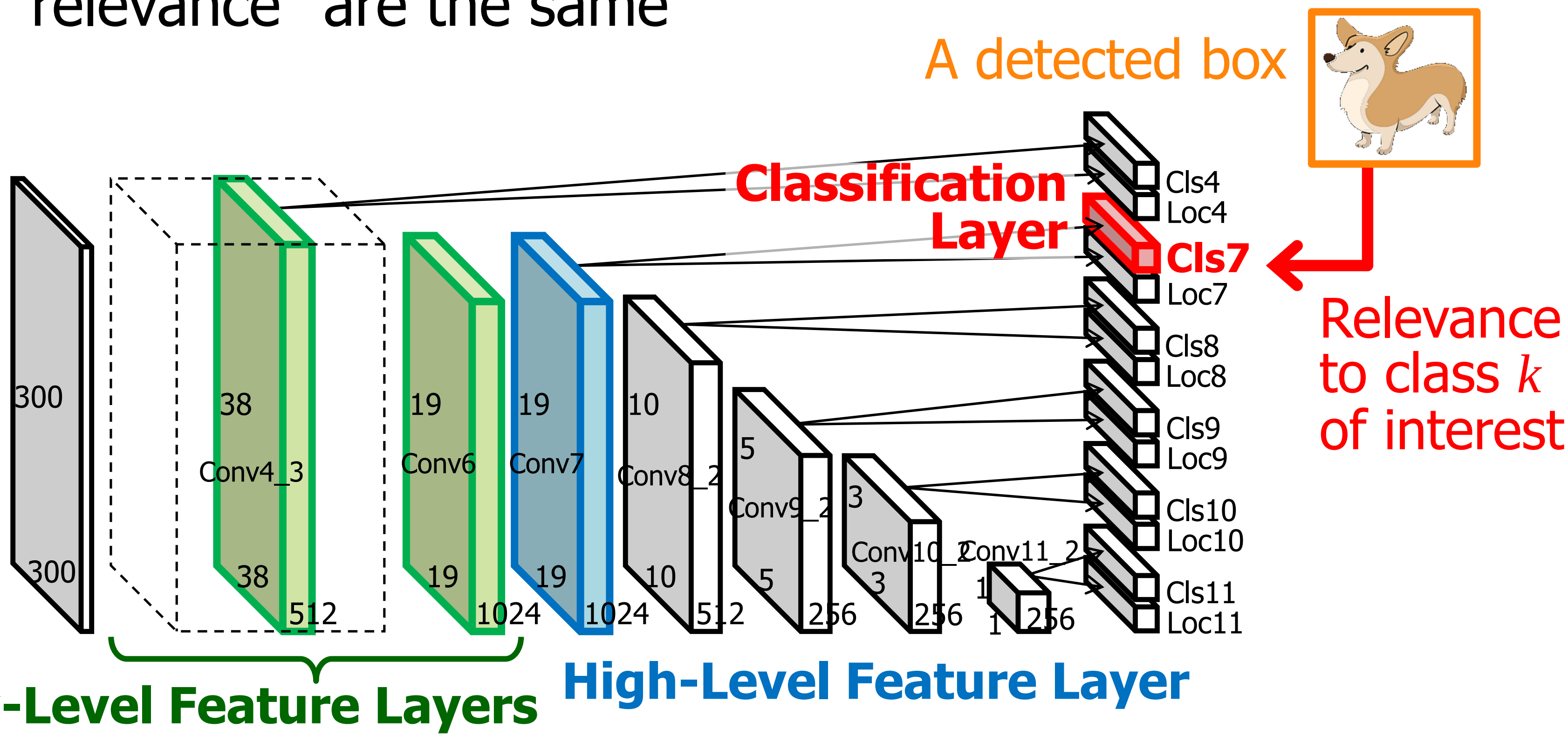
- We design the meaning of relevance to be **consistent** in two heterogeneous tasks in SSD:
  - Classification
  - Localization (Regression)

Outline

✓ Background
- Proposed method: CRP
- Experiments

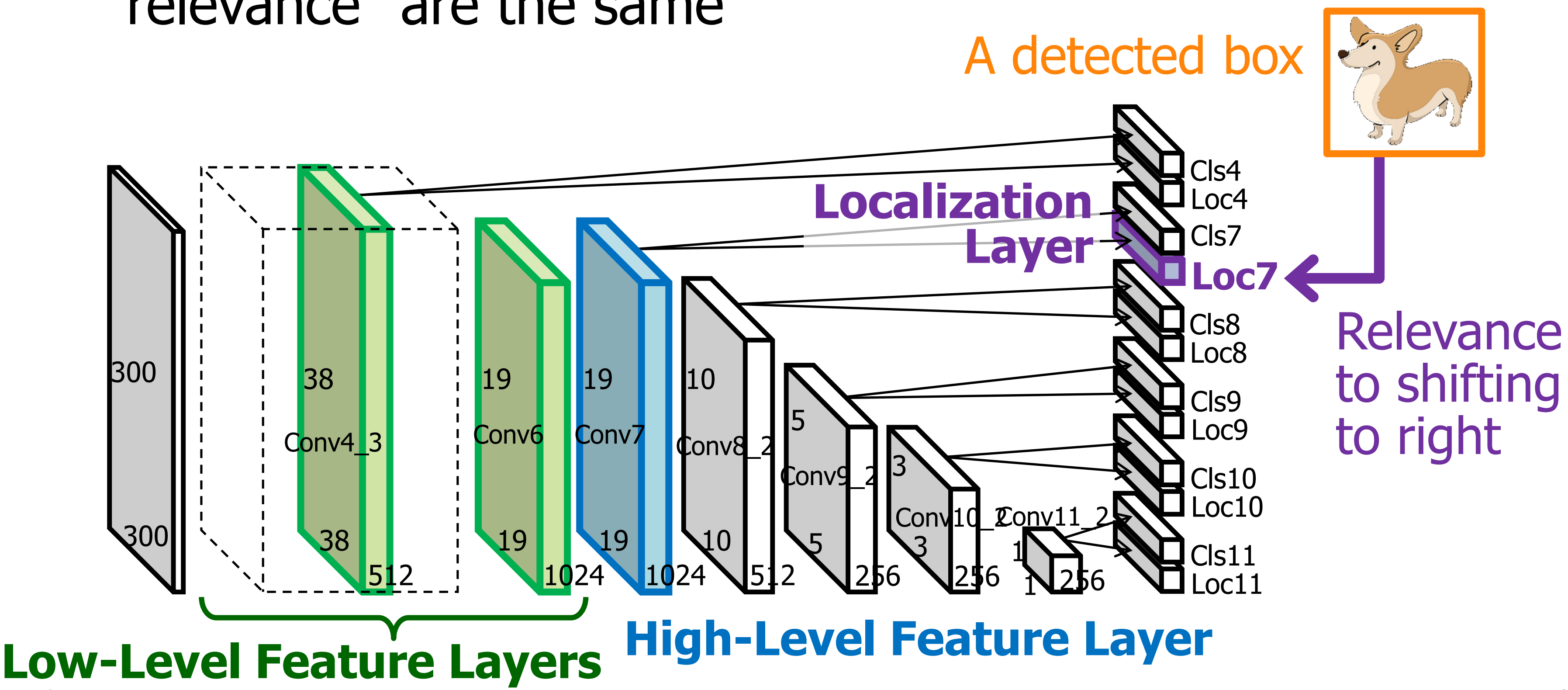
Contrastive Relevance Propagation (CRP)

- CRP: LRP tailored for SSD
  - Classifies SSD's layers into 4 types
  - Applies semantically appropriate propagation rules to each layer type
  - In both classification and localization, the meanings of "relevance" are the same



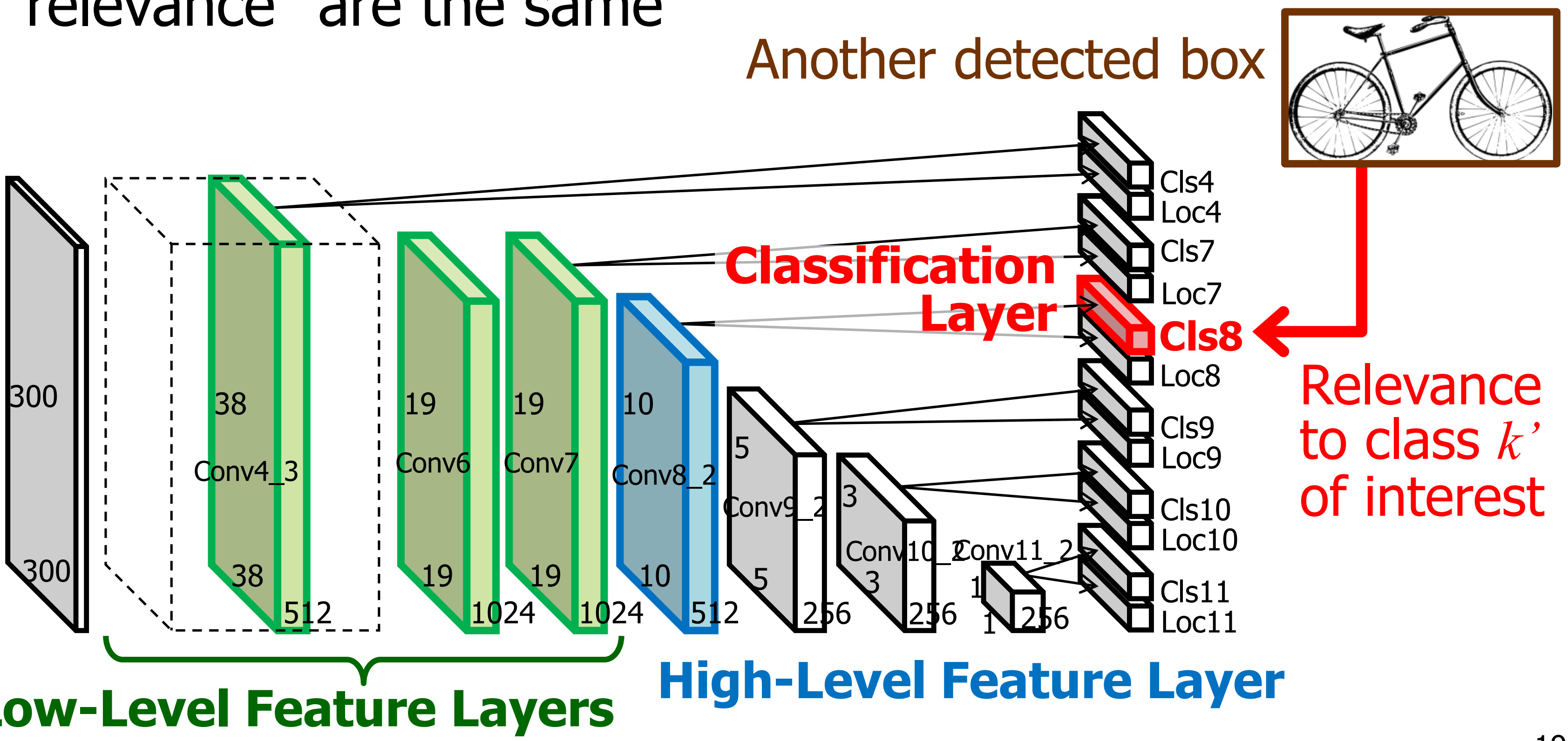
Contrastive Relevance Propagation (CRP)

- CRP: LRP tailored for SSD
  - Classifies SSD’s layers into 4 types
  - Applies semantically appropriate propagation rules to each layer type
  - In both classification and localization, the meanings of “relevance” are the same

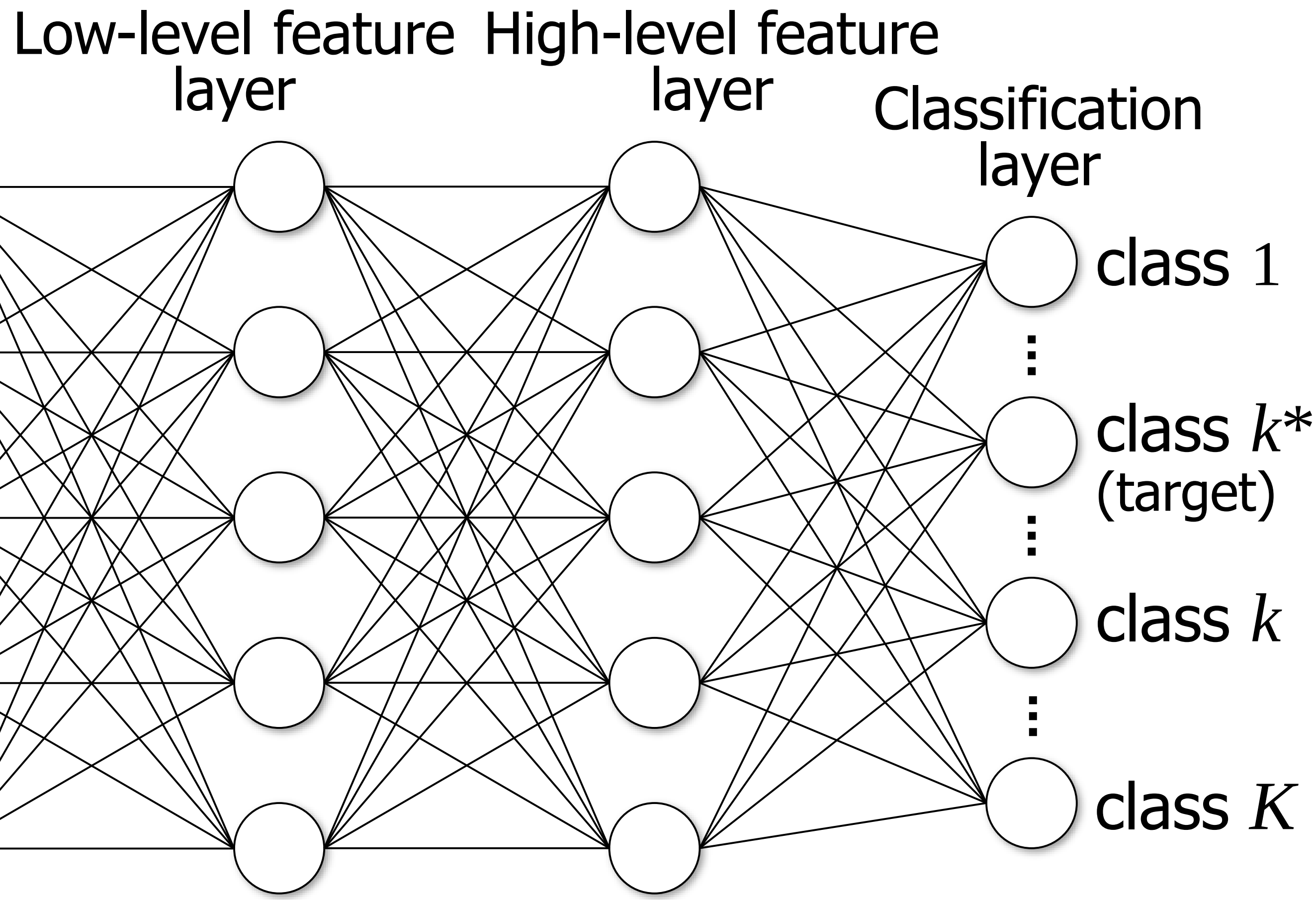


Contrastive Relevance Propagation (CRP)

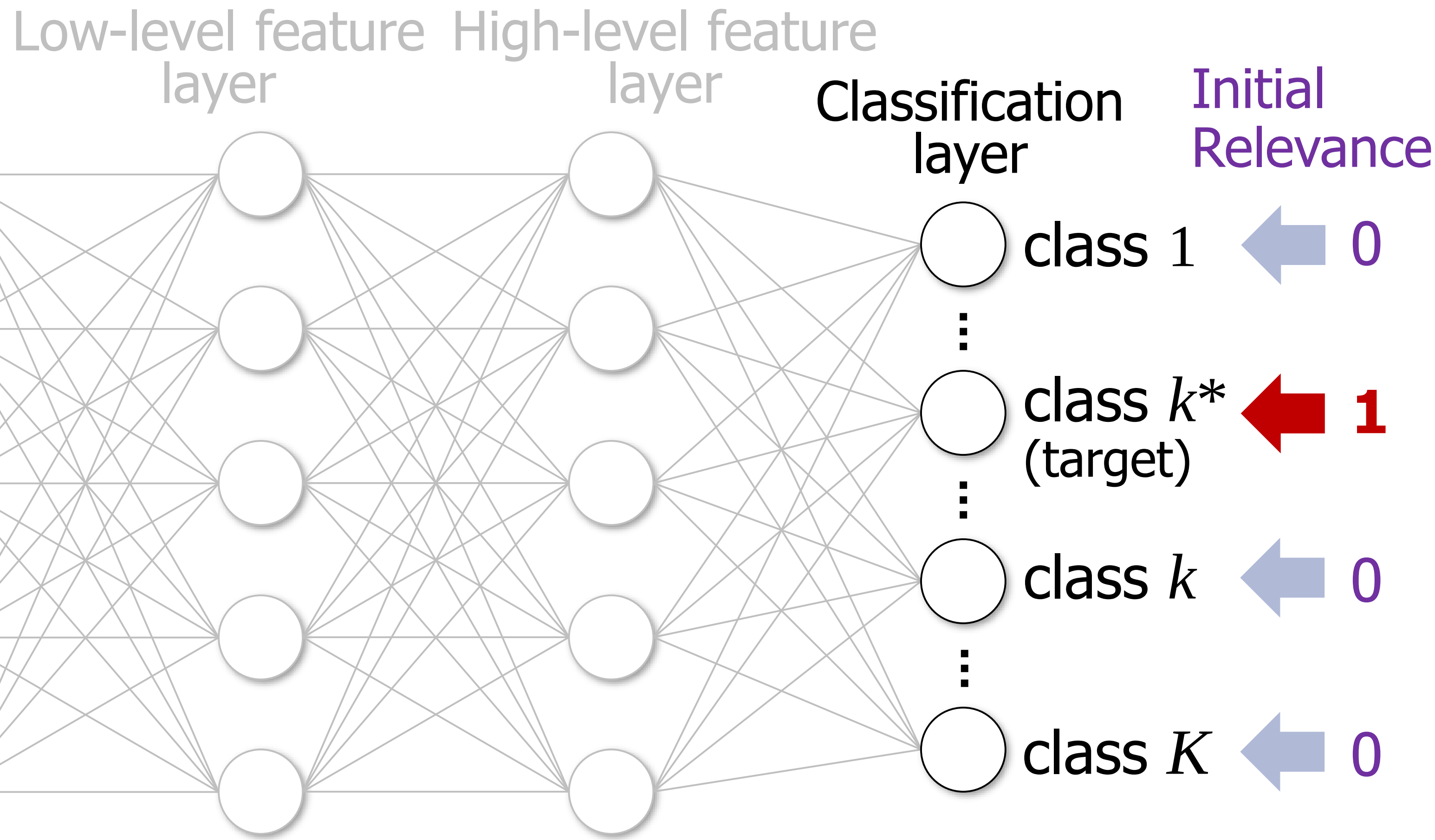
- CRP: LRP tailored for SSD
  - Classifies SSD's layers into 4 types
  - Applies semantically appropriate propagation rules to each layer type
  - In both classification and localization, the meanings of "relevance" are the same



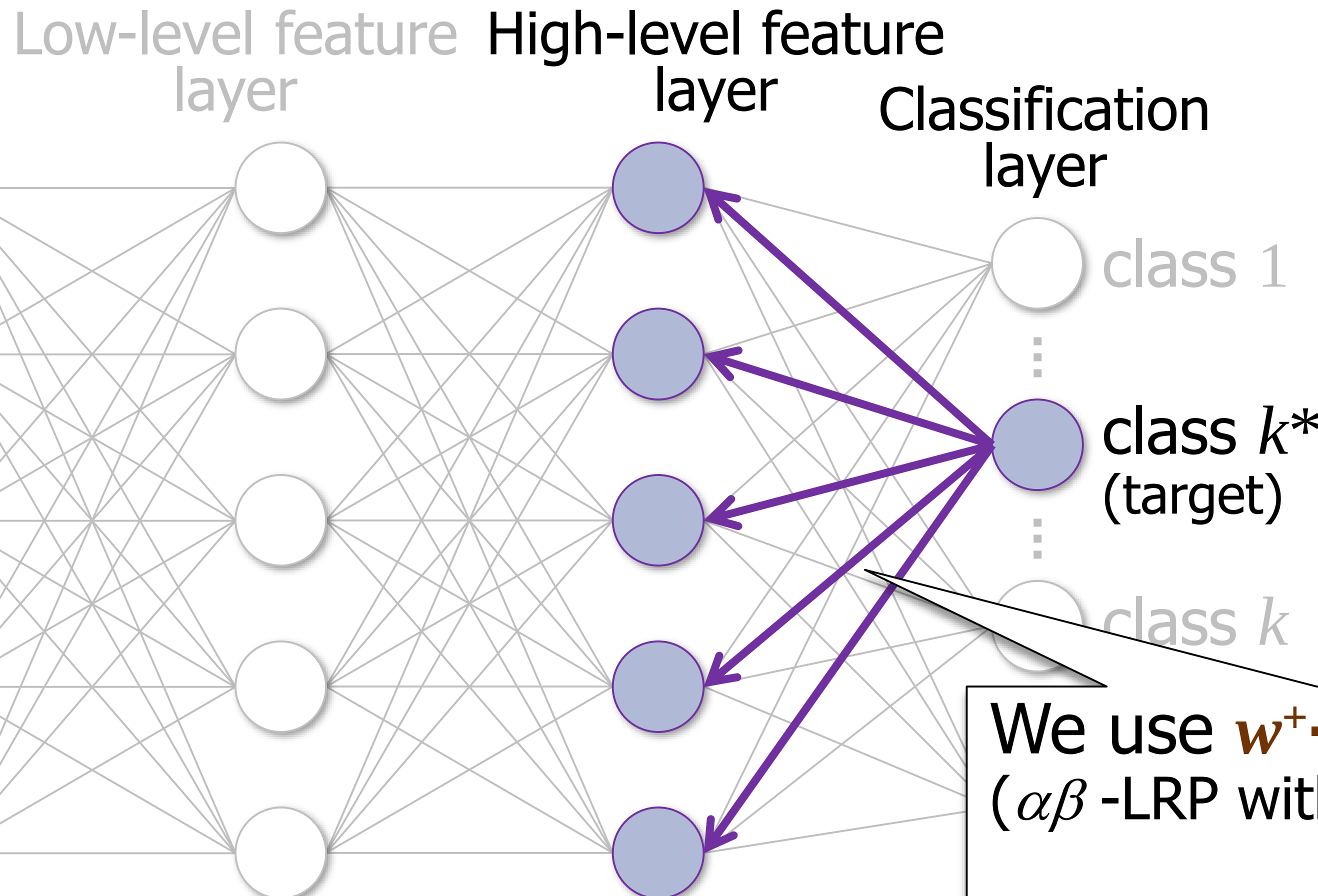
CRP: Propagation Rules in Classification



CRP: Propagation Rules in Classification



CRP: Propagation Rules in Classification



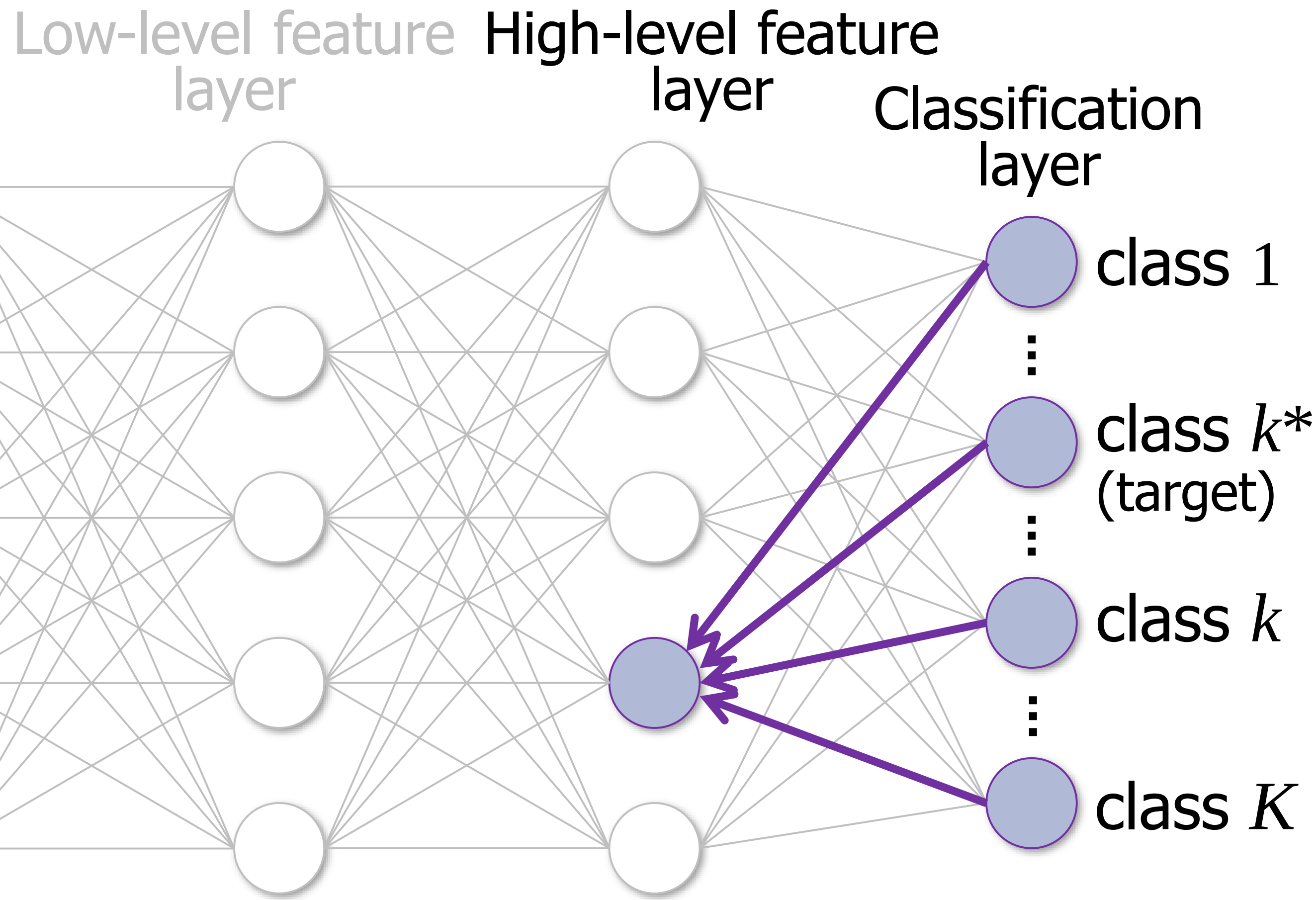
We use ***w*⁺-rule**

($\alpha\beta$ -LRP with $\alpha = 1$, $\beta = 0$)

$$R_{i \leftarrow j} = \frac{w_{ij}^{+} x_i}{\sum_{i'} w_{i'j}^{+} x_{i'}} R_j$$

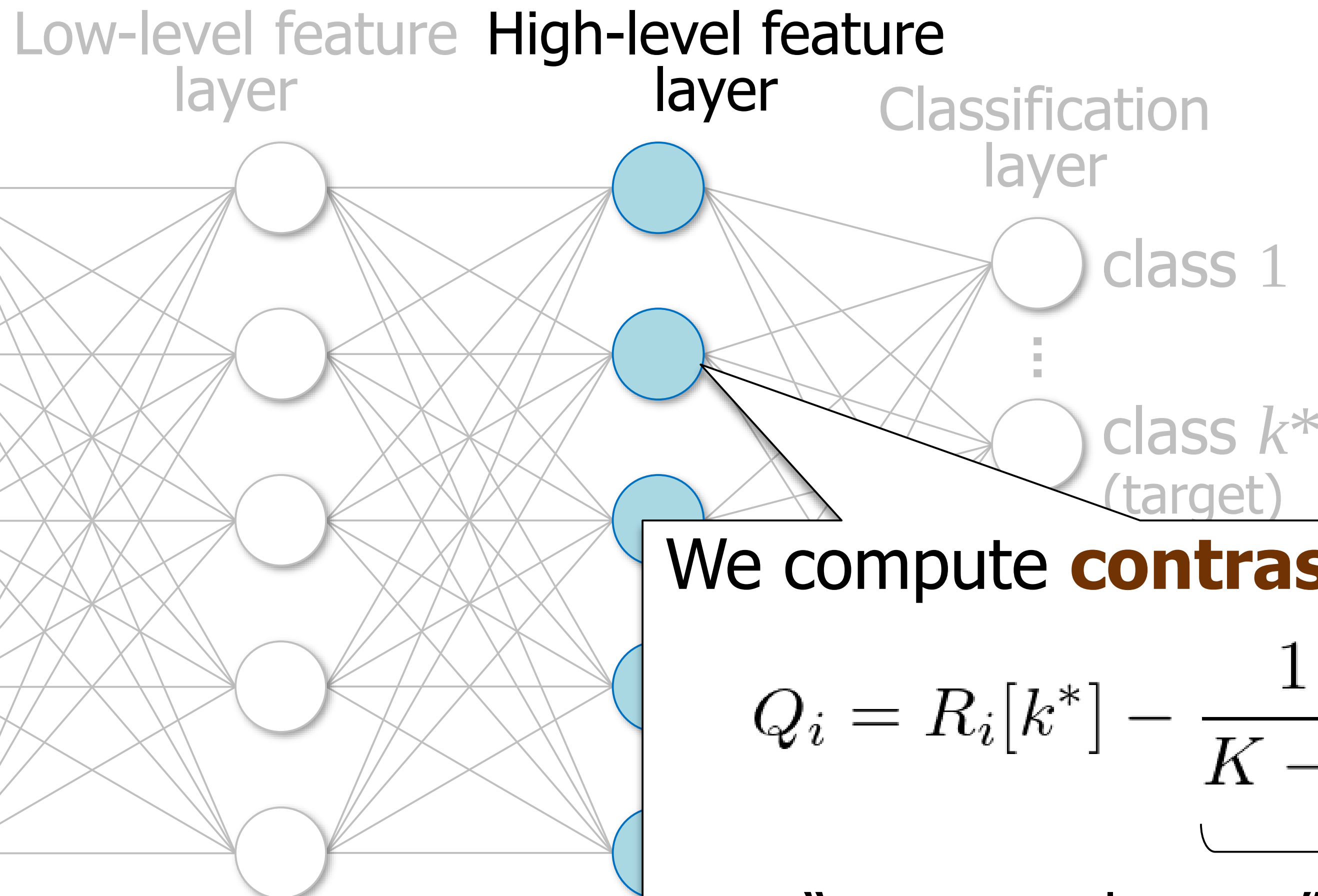
to find units that **positively** contribute to class k*

CRP: Propagation Rules in Classification



At this moment, we can compute a **class-specific** relevance $R_i[k^*]$ for the target class k^* by summing up the passed relevance

CRP: Propagation Rules in Classification

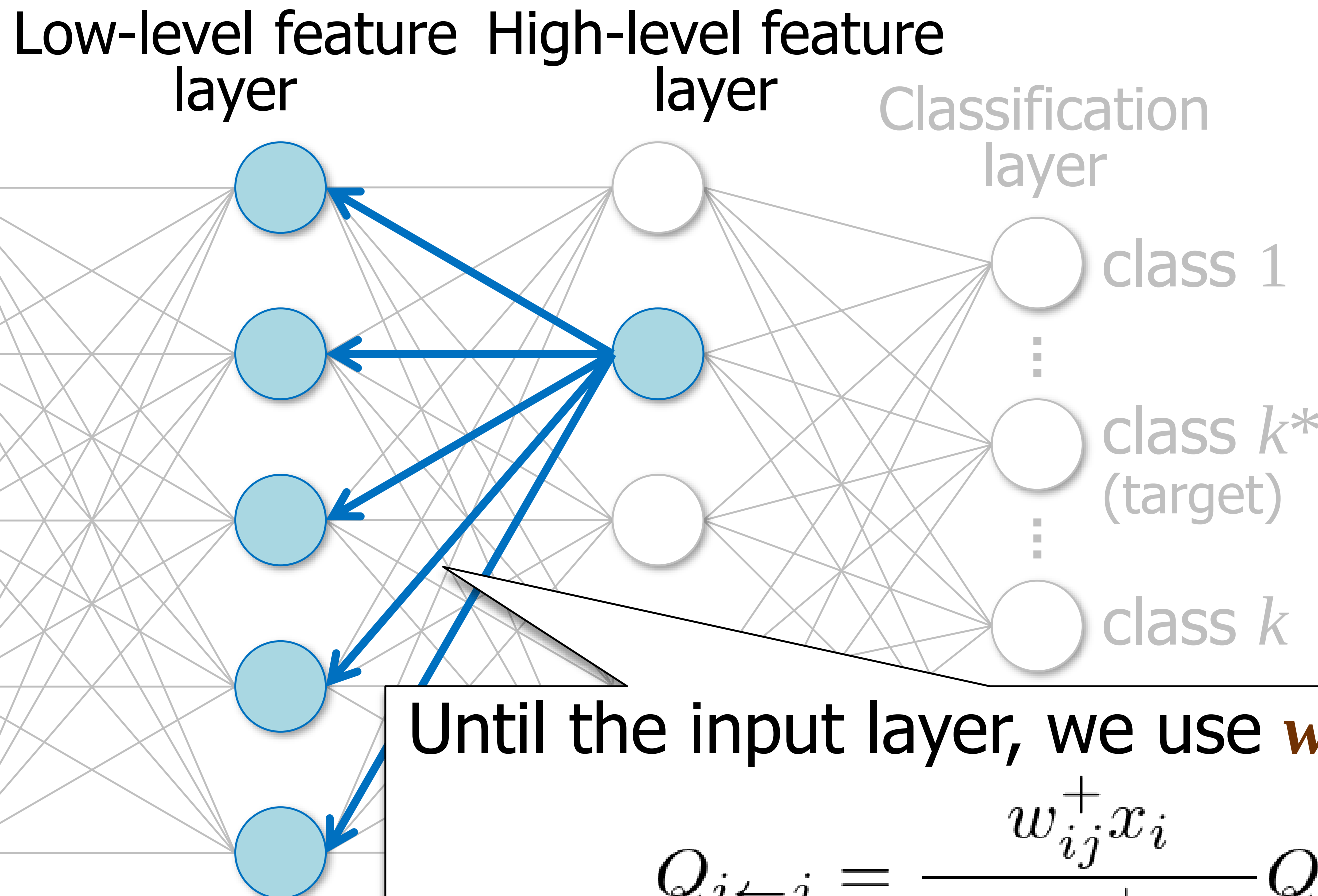


We compute **contrastive relevance**

$$Q_i = R_i[k^*] - \underbrace{\frac{1}{K-1} \sum_{k:k \neq k^*} R_i[k]}_{\text{"average relevance" over other classes}}$$

to find units that make a **significantly positive** or a **significantly negative** contribution to the target class k^*

CRP: Propagation Rules in Classification

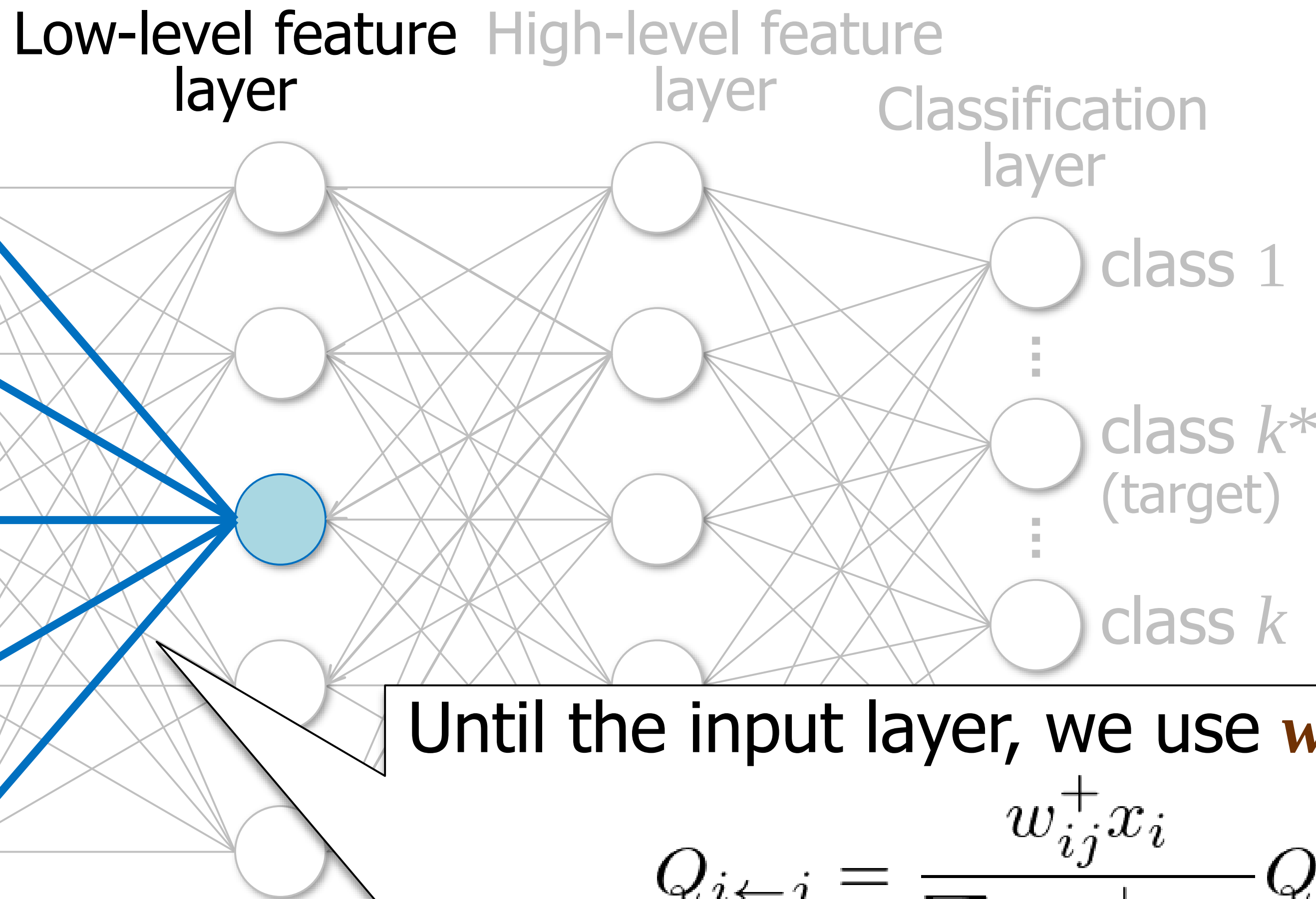


Until the input layer, we use ***w*⁺-rule**

$$Q_{i \leftarrow j} = \frac{w_{ij}^{+} x_i}{\sum_{i'} w_{i'j}^{+} x_{i'}} Q_j$$

to distribute the positivity or the negativity of contrastive relevance
(activations x_i are non-negative due to ReLU)

CRP: Propagation Rules in Classification

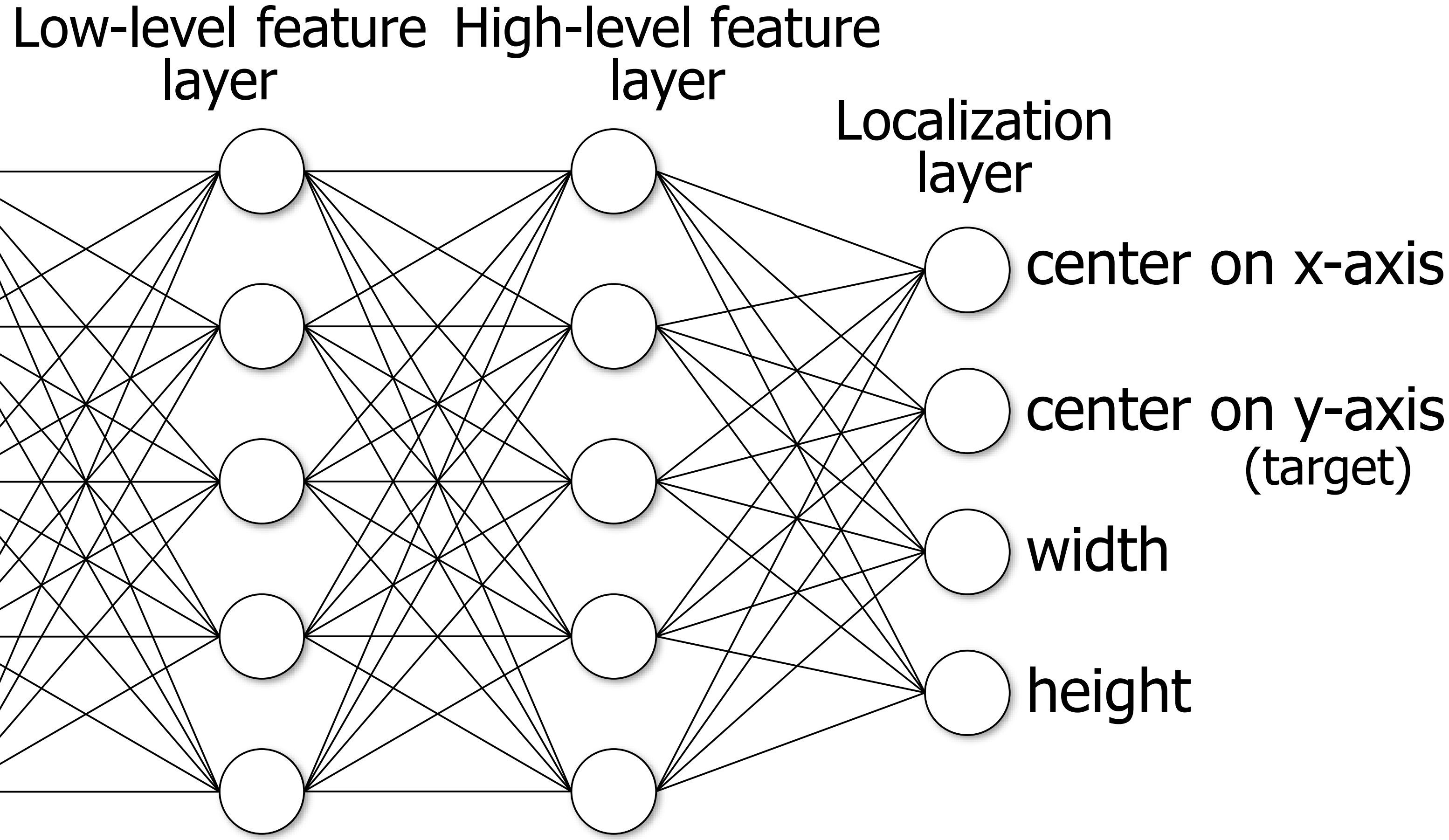


Until the input layer, we use ***w^+*-rule**

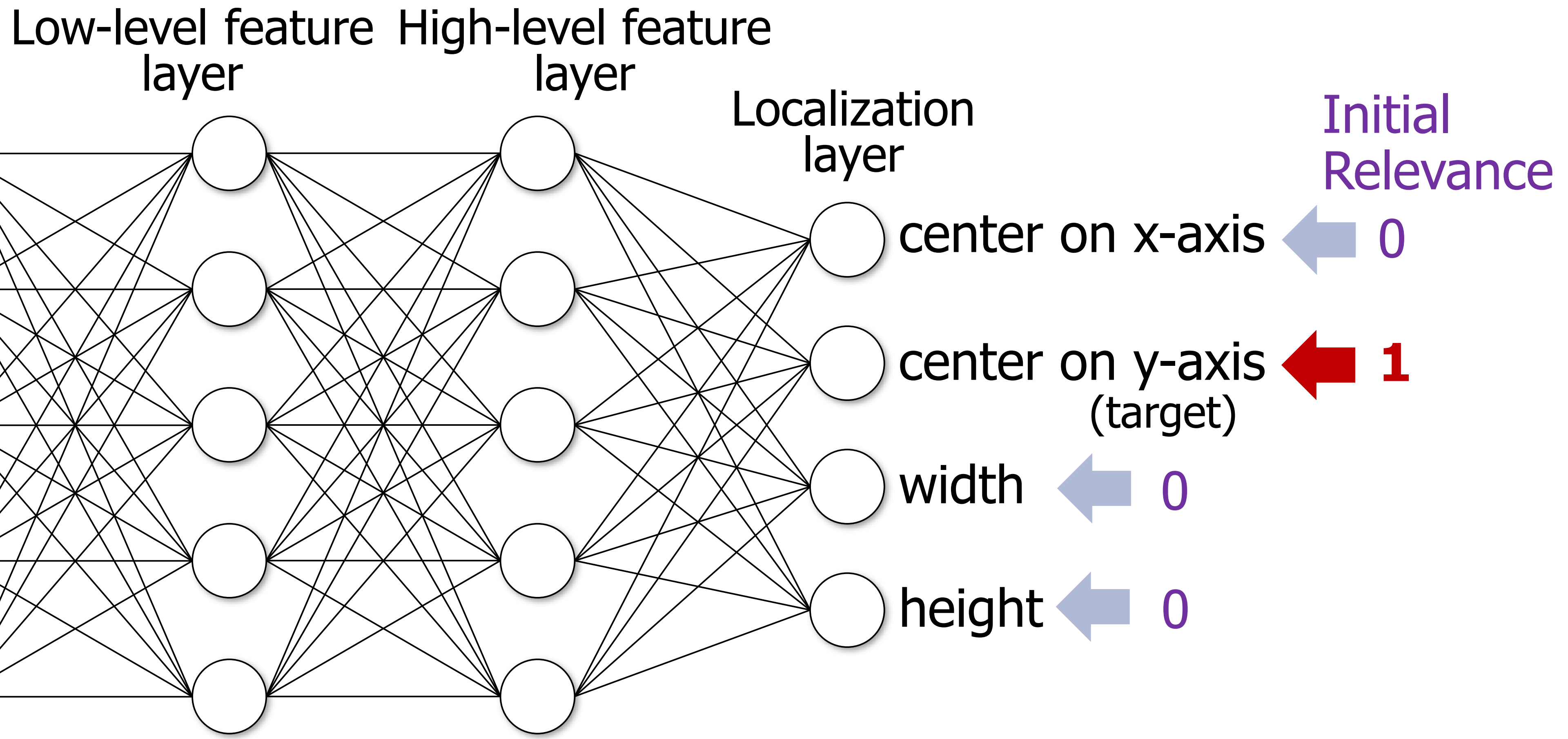
$$Q_{i \leftarrow j} = \frac{w_{ij}^{+} x_i}{\sum_{i'} w_{i'j}^{+} x_{i'}} Q_j$$

to distribute the positivity or the negativity of contrastive relevance
(activations x_i are non-negative due to ReLU)

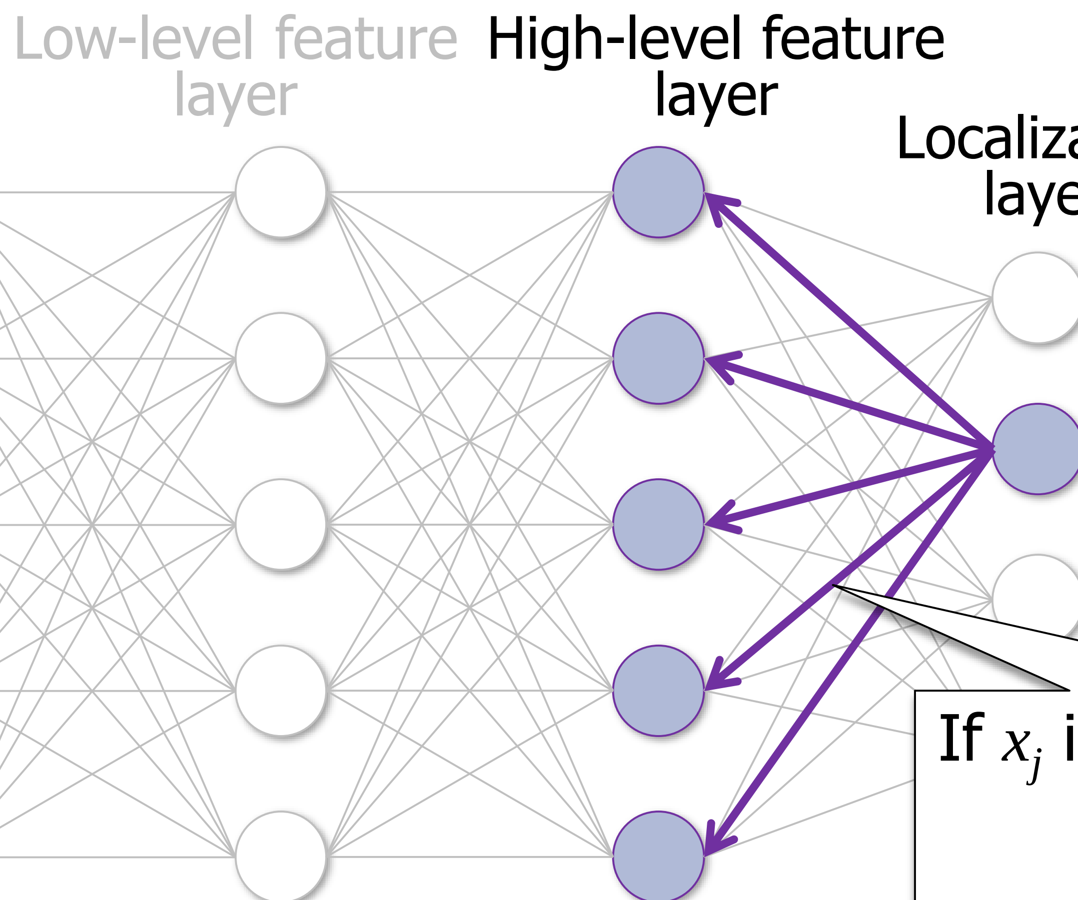
CRP: Propagation Rules in Localization



CRP: Propagation Rules in Localization



CRP: Propagation Rules in Localization



Activation x_j

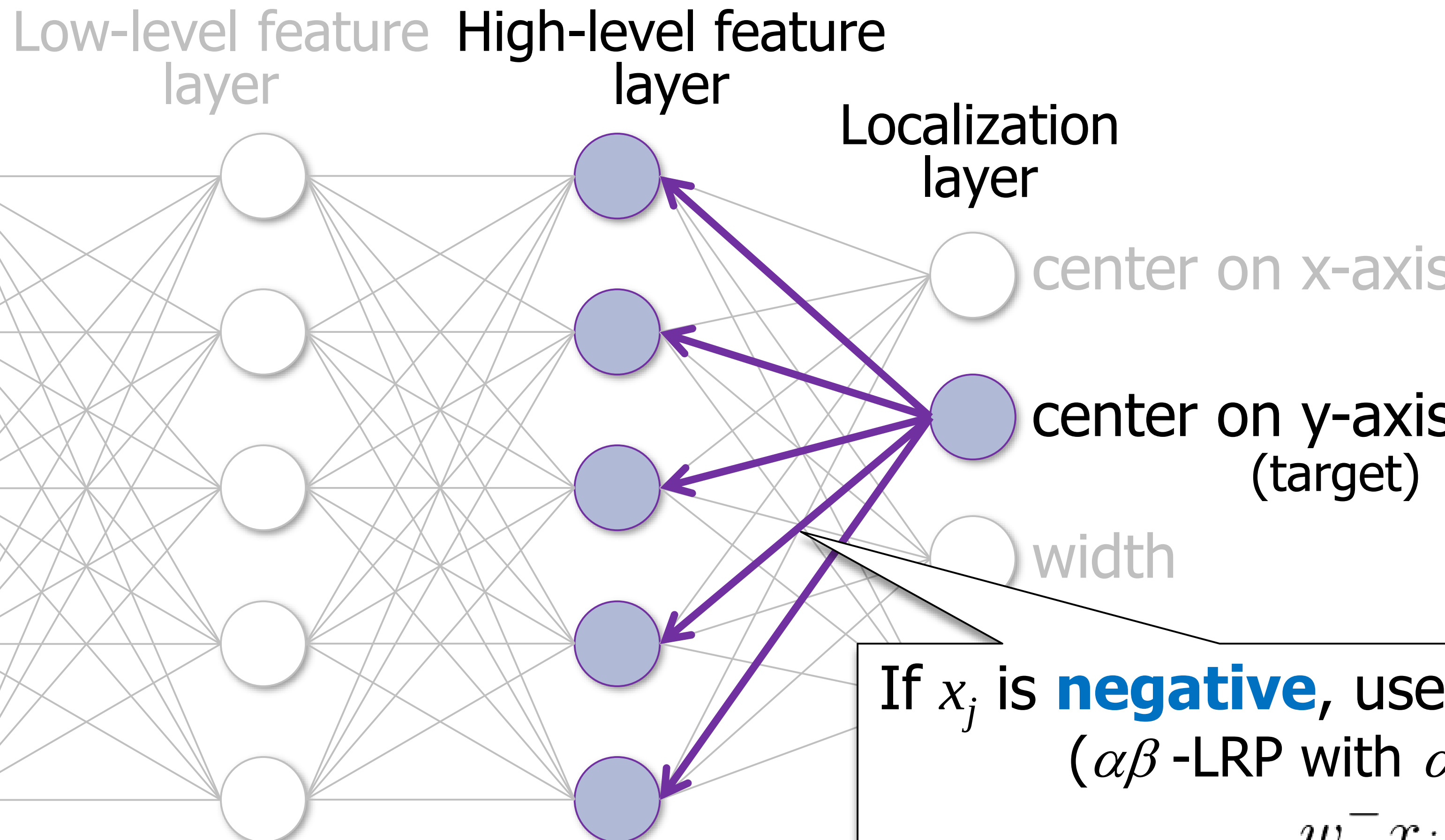
If x_j is **positive**, use w^+**-rule**
($\alpha\beta$-LRP with $\alpha = 1$, $\beta = 0$)

$$R_{i \leftarrow j} = \frac{w_{ij}^+ x_i}{\sum_{i'} w_{i'j}^+ x_{i'}} R_j$$

to find units that **positively** contribute to center on y-axis

Sign-based rule switching:
We switch two rules according to the sign of x_j

CRP: Propagation Rules in Localization



Activation
x_j

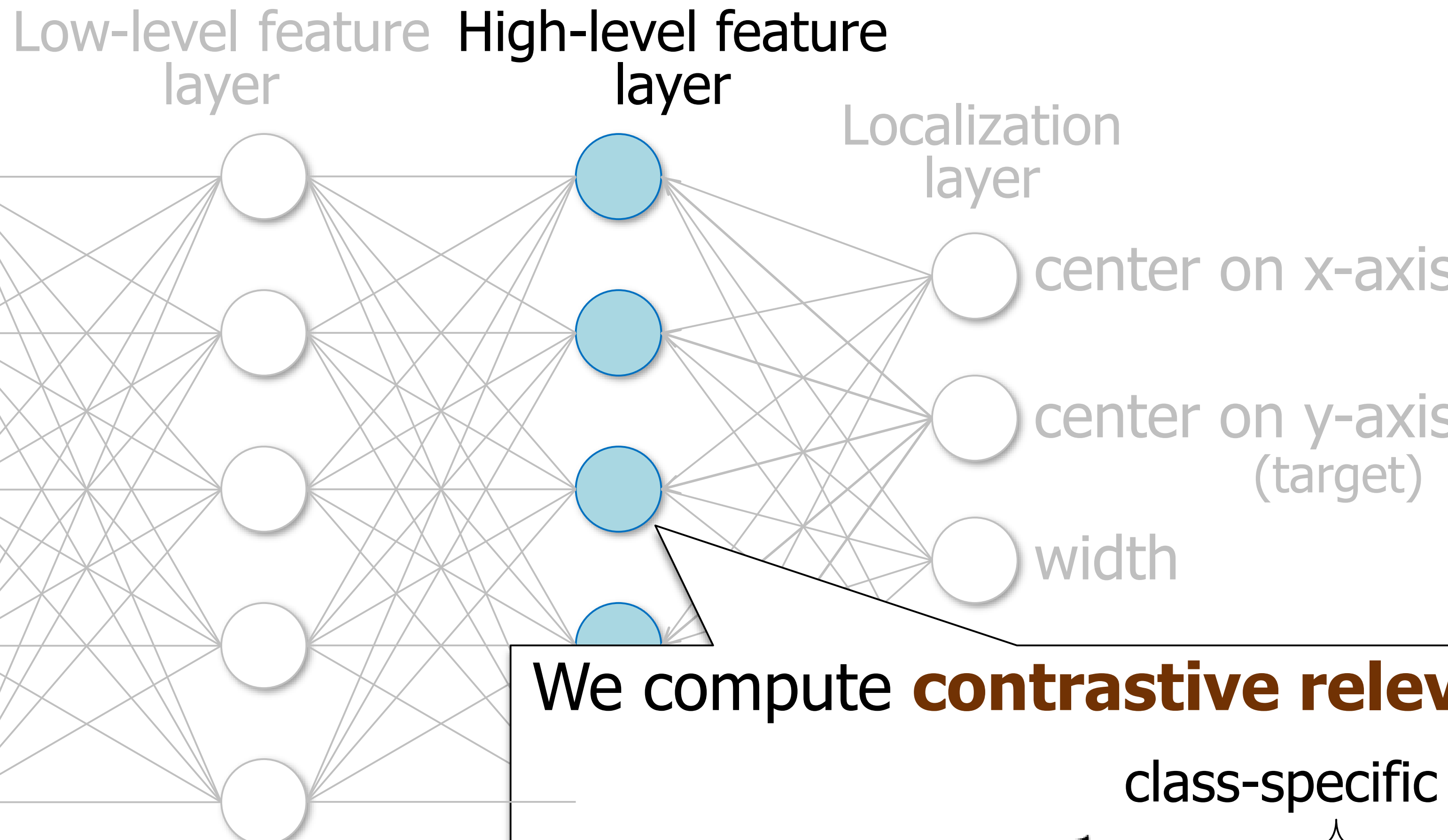
If x_j is **negative**, use **w^{-}-rule**
($\alpha\beta$ -LRP with $\alpha = 0$, $\beta = 1$)

$$R_{i \leftarrow j} = \frac{w_{ij}^{-} x_i}{\sum_{i'} w_{i'j}^{-} x_{i'}} R_j$$

to find units that **negatively** contribute to center on y-axis

Sign-based rule switching:
We switch two rules according to the sign of x_j

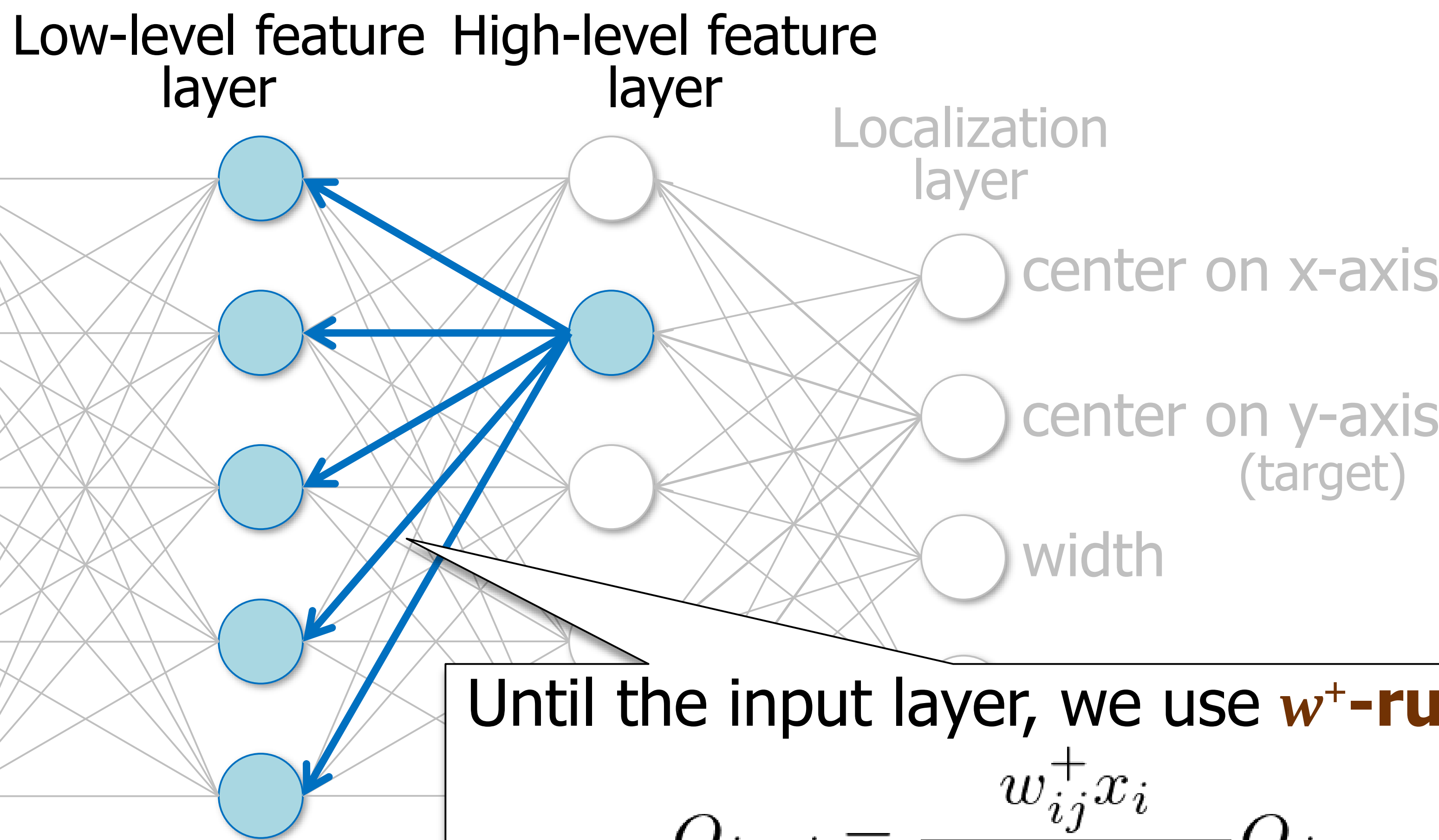
CRP: Propagation Rules in Localization



We compute **contrastive relevance**

$$Q_i = \underbrace{R_i}_{\text{relevance from the localization layer}} - \underbrace{\frac{1}{K}\sum_k \overbrace{R_i[k]}^{\text{class-specific relevance}}}_{\text{“overall average”}}$$

CRP: Propagation Rules in Localization

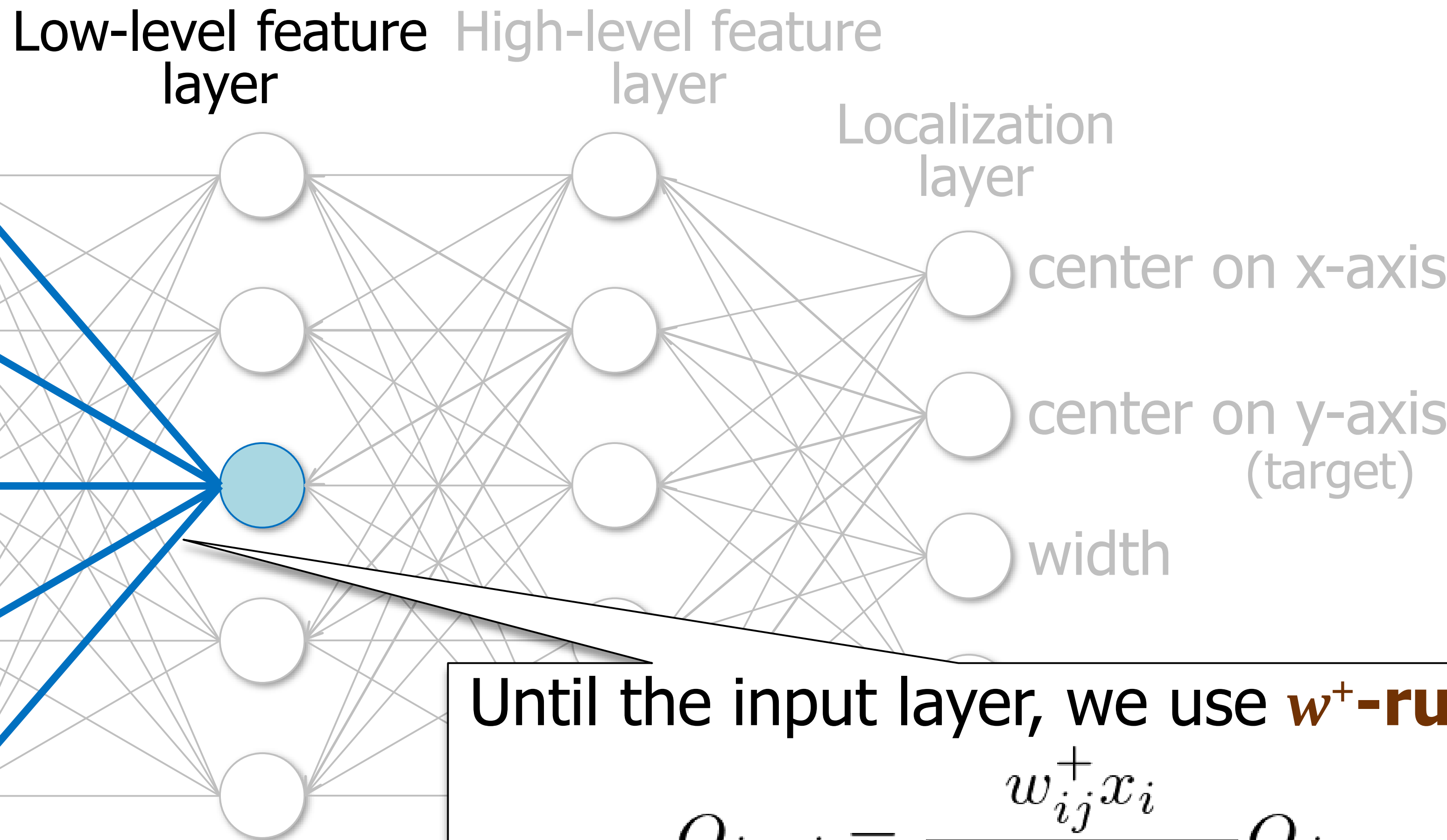


Until the input layer, we use w^+**-rule**

$$Q_{i \leftarrow j} = \frac{w_{ij}^+ x_i}{\sum_{i'} w_{i'j}^+ x_{i'}} Q_j$$

as in classification

CRP: Propagation Rules in Localization



Until the input layer, we use w^{+}**-rule**

$$Q_{i \leftarrow j} = \frac{w_{ij}^{+} x_i}{\sum_{i'} w_{i'j}^{+} x_{i'}} Q_j$$

as in classification

Outline

- ✓ Background
- ✓ Proposed method: CRP
- Experiments

Experimental Settings

- Dataset: Pascal VOC 2012
- We ported the TensorFlow implementation of LRP (https://github.com/VigneshSrinivasan10/interprettensor) into a TensorFlow implementation of SSD (https://github.com/balancap/SSD-Tensorflow)
- SSD implementation includes a learned model (We conducted no learning)
- We added CRP-specific routines
- Relevance was normalized before creating heatmaps

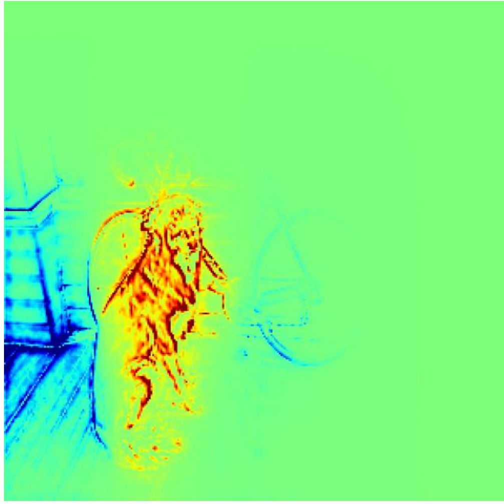
(See the paper for details)

Numerical Example

- Relevance is almost symmetrically distributed at zero

Layer	Relevance for 'dog' Max.	95%-tile	Median	5%-tile	Min.
Cls8	1.82E-02	0	0	0	0
Conv8_2	9.51E-04	0	0	−1.86E-06	−3.45E-04
Conv8_1	1.55E-04	0	0	0	−1.07E-04
Conv7	6.69E-04	0	0	0	−2.56E-04
Conv6	1.91E-04	0	0	−6.30E-08	−1.05E-04
Pool5	9.07E-04	0	0	0	−4.38E-04
Conv5_3	1.30E-04	0	0	−1.08E-07	−1.39E-04
Conv5_2	1.72E-04	0	0	−1.11E-07	−9.79E-05
Conv5_1	1.06E-04	6.21E-08	0	−1.42E-07	−7.24E-05
Pool4	1.06E-04	0	0	0	−7.24E-05
Conv4_3	3.35E-05	0	0	−1.41E-08	−4.99E-05
Conv4_2	1.34E-05	1.11E-10	0	−2.20E-08	−3.85E-05
Conv4_1	2.38E-05	6.59E-08	0	−8.12E-08	−4.42E-05
Pool3	2.38E-05	0	0	0	−4.42E-05
Conv3_3	6.15E-06	1.40E-08	0	−1.97E-08	−2.10E-05
Conv3_2	3.81E-06	2.03E-08	0	−2.62E-08	−2.29E-05
Conv3_1	6.44E-06	7.46E-08	0	−6.31E-08	−1.75E-05
Pool2	6.44E-06	0	0	−2.29E-10	−1.75E-05
Conv2_2	4.21E-06	1.65E-08	0	−1.74E-08	−1.11E-05
Conv2_1	3.28E-06	3.85E-08	0	−3.29E-08	−1.04E-05
Pool1	3.28E-06	0	0	−4.92E-10	−1.04E-05
Conv1_2	2.47E-06	5.59E-09	0	−5.09E-09	−3.42E-06
Conv1_1	6.47E-06	3.26E-07	−1.57E-14	−2.52E-07	−1.17E-05
Input	6.47E-06	3.26E-07	−1.57E-14	−2.52E-07	−1.17E-05

Target class: "dog"



Different Colors in Heatmap: **Positives** **≈ 0** **Negatives**

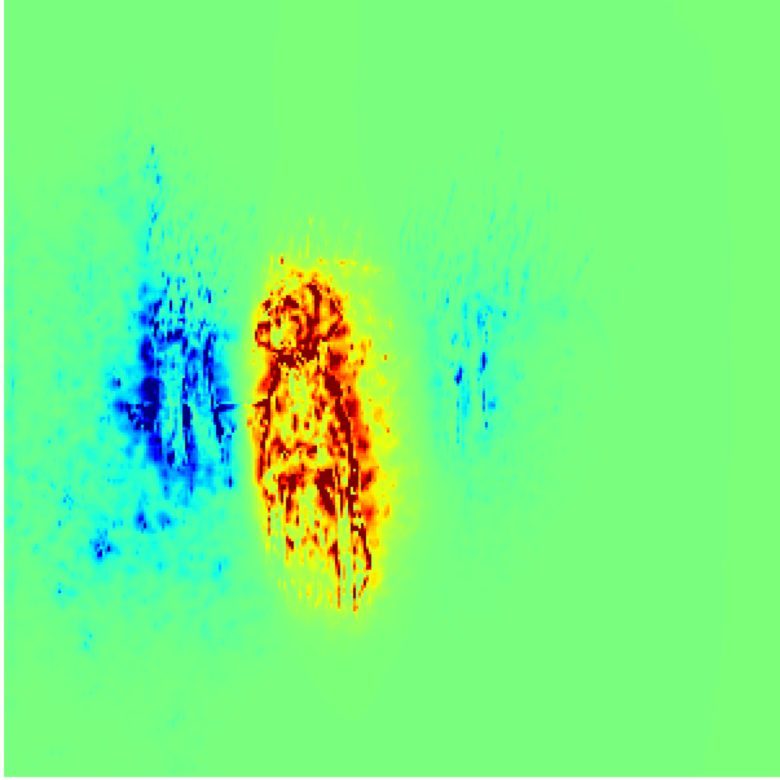
Error Analysis (1)

- A dog was misclassified as a sheep

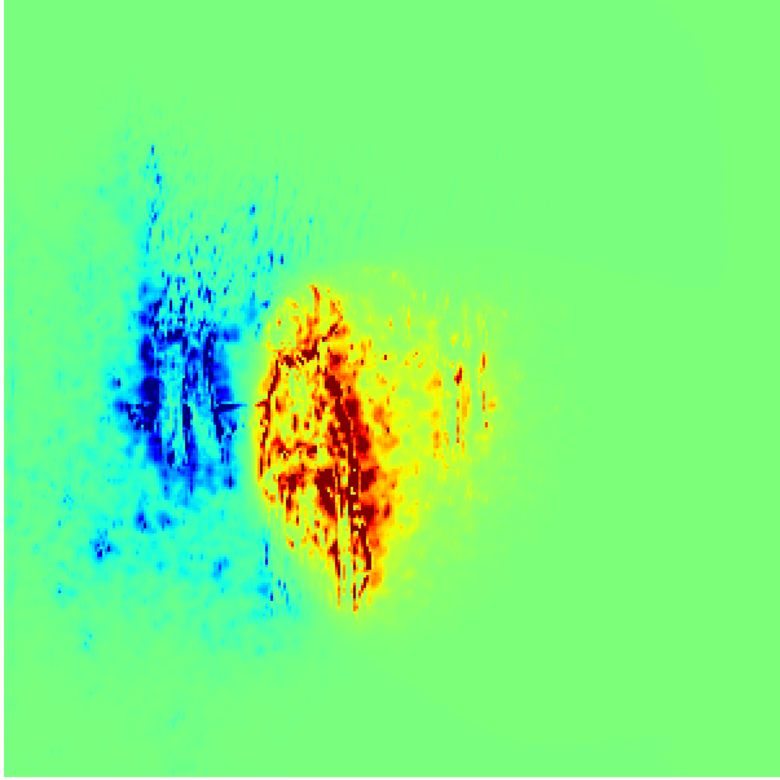
Error Analysis (2)

- A dog was misclassified as a sheep

Target class: “dog”



Target class: “sheep”

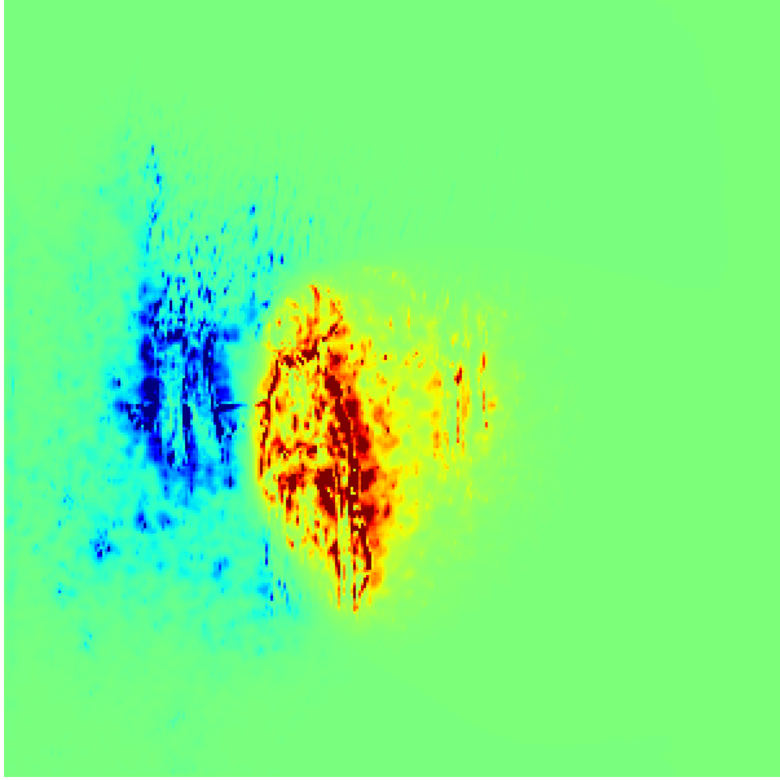


Error Analysis (3)

- A dog was misclassified as a sheep

<85%tile values masked

Target class: “sheep”



Error Analysis (4)

- Unwanted localizations:
  - Horizontal shift to left with widening
  - Vertical shift to top with heightening

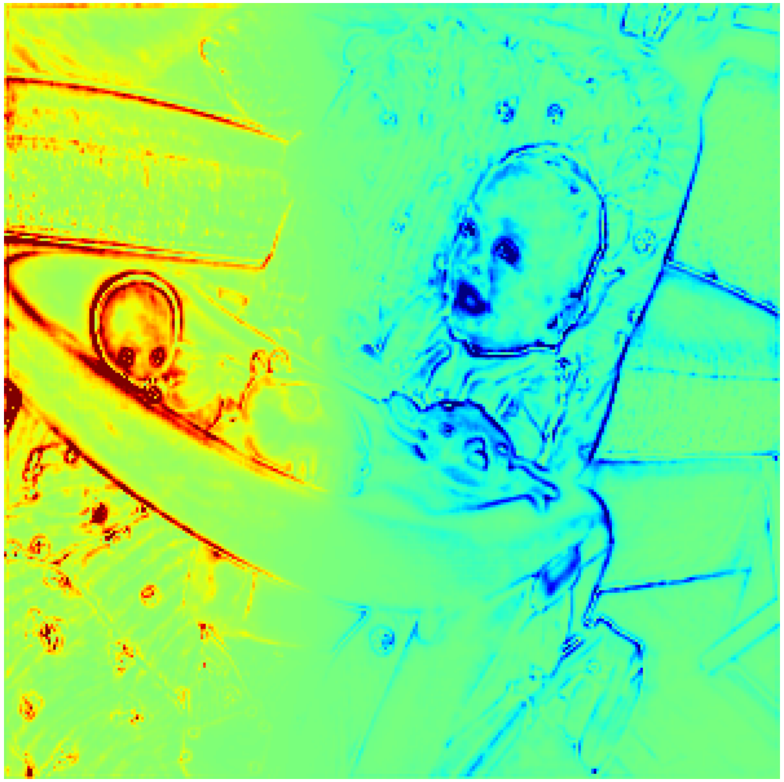
Before localization

After localization

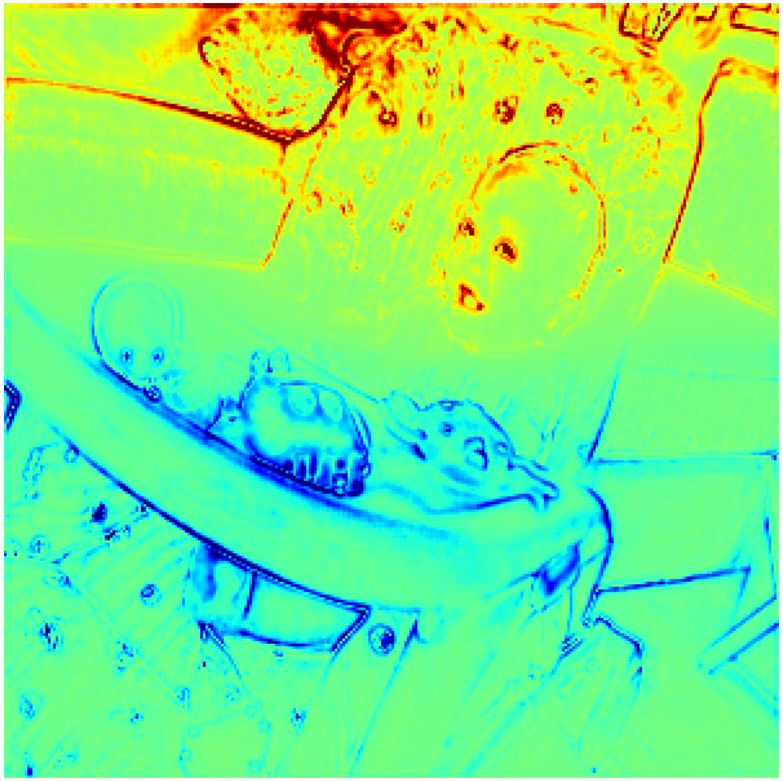
Error Analysis (5)

- Unwanted localizations:
  – Horizontal shift to left with widening
  – Vertical shift to top with heightening

Target offset: center on x-axis



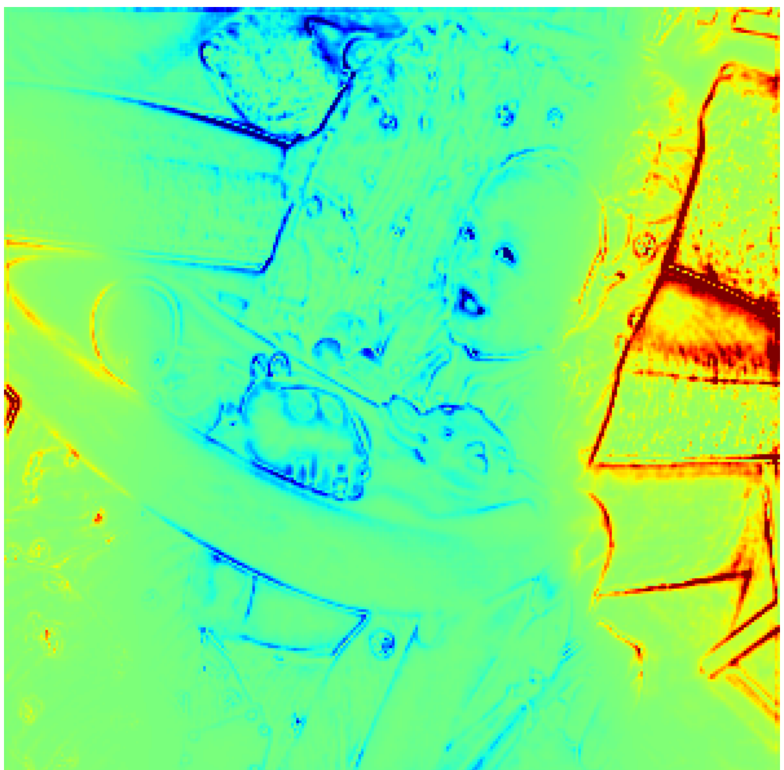
Target offset: center on y-axis



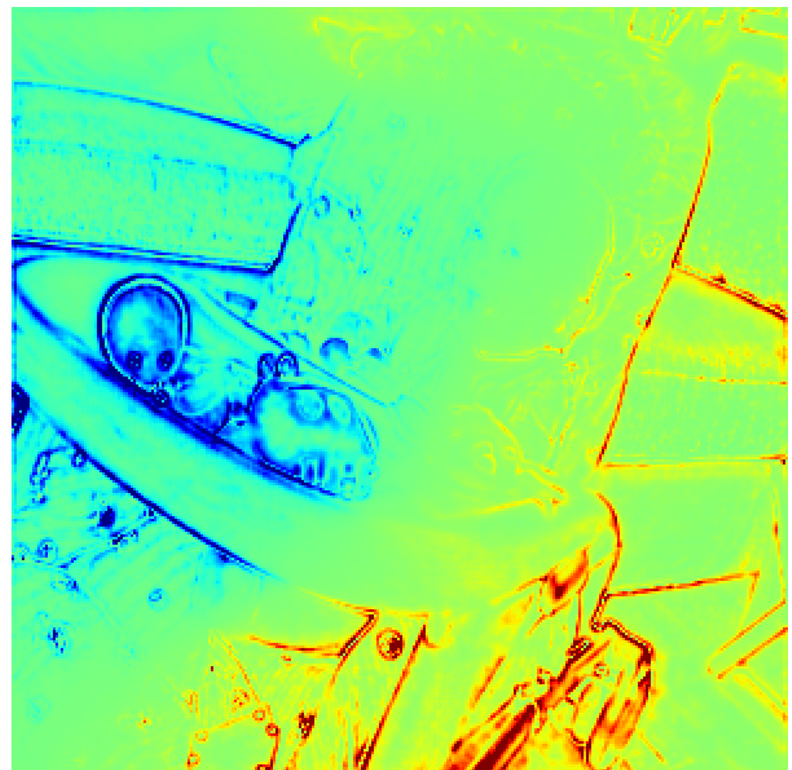
Error Analysis (6)

- Unwanted localizations:
  - Horizontal shift to left with widening
  - Vertical shift to top with heightening

Target offset: width



Target offset: height



Summary

- CRP (contrastive relevance propagation) as an LRP method tailored for SSD:
  - Can highlight only significantly important features for a target class
  - Can deal with SSD's heterogeneous outputs (classification and localization)
- Some error analyses using CRP were conducted

Future work

- Applying CRP to other object detectors such as YOLO
- Applying CRP (retrospectively) to standard CNNs

Thank you for your attention!