Explainable AI für Deep Learning: Overview und Tutorial

Jörg Simon

Huh? What?

Right....

But maybe we can explain the complex models a bit

Explainable AI für Deep Learning: Overview und Tutorial

Jörg Simon

About me

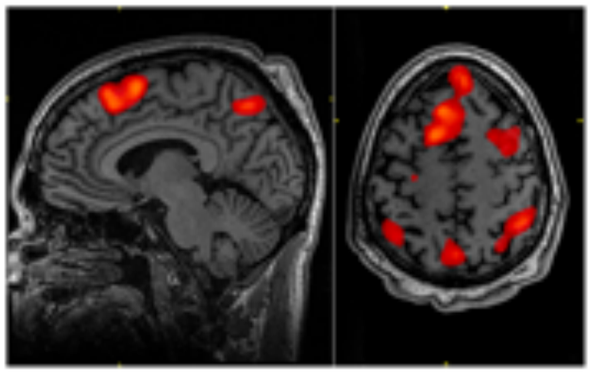
- PhD on using DeepLearning to detect Human Factors from BioSignals
- Prof. Eduardo Veas and Herbert Danzinger
- Sometimes very Sparse Data!
- Inspired to use interpretability results to change the training process itself.

Agenda

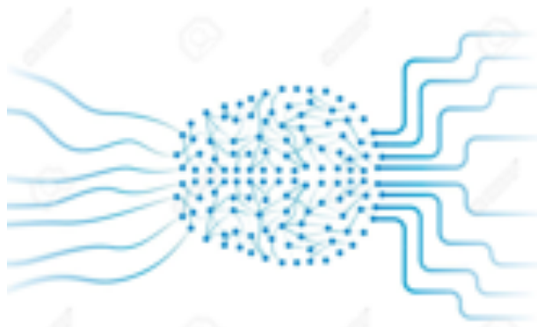
- Definitions and Stuff
- Hands On
- Discussion

Definitions and Stuff

- Deep Learning

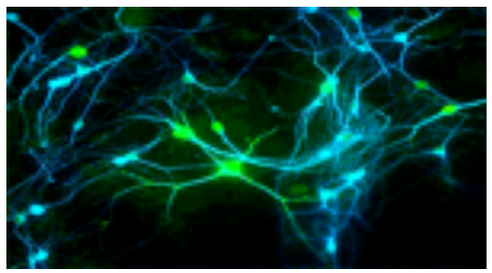


Distributed Representation



Super Simplified Model of Human Brain

Hinton



Spiking Frequency = weight

Deep Learning?

Simple Matrix Multiply + Non Linearity

RNN

CNN

Yann LeCun

Bengio, Hochreiter, Schmidhuber

Distributed Representation

Spiking Frequency = weight

Super Simplified Model of Human Brain

Hinton

Deep Learning?

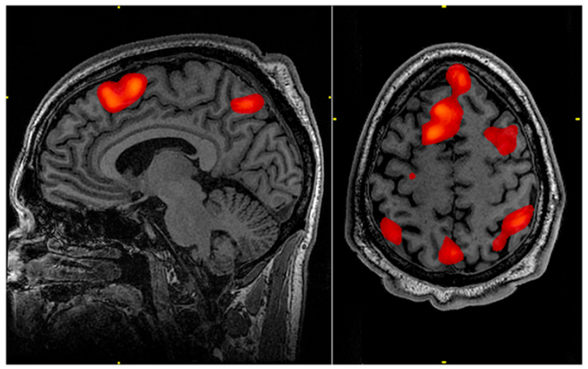
CNN

Yann LeCun

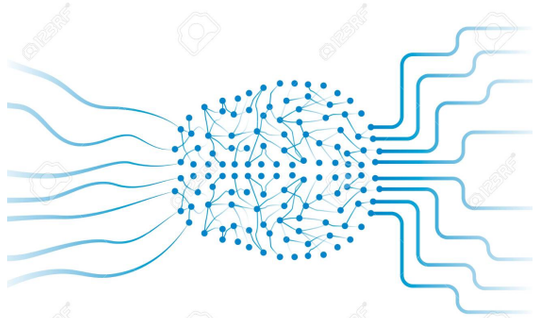
Simple Matrix Multiply
+ Non Linearity

RNN

Bengio, Hochreiter, Schmidhuber

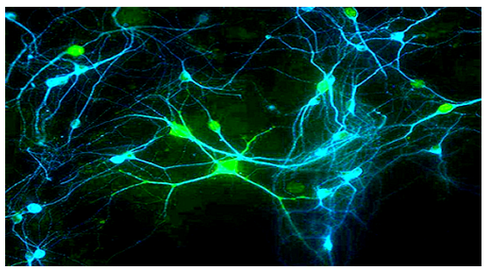


Distributed Representation

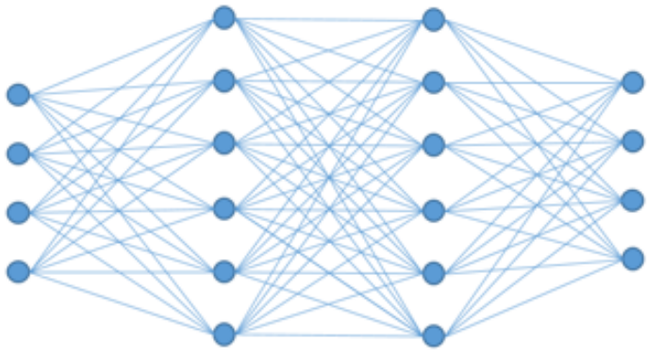


Super Simplified Model of Human Brain

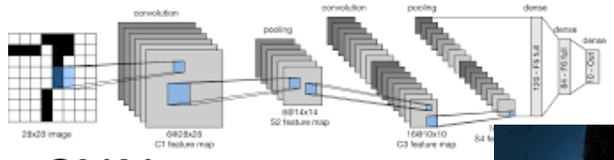
Hinton



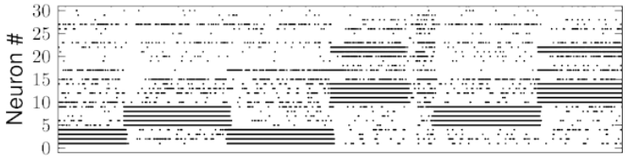
Spiking Frequency = weight



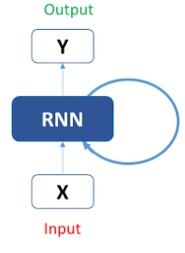
Deep Learning?



CNN



Simple Matrix Multiply + Non Linearity



RNN

Yann LeCun

Bengio, Hochreiter, Schmidhuber

Definitions and Stuff

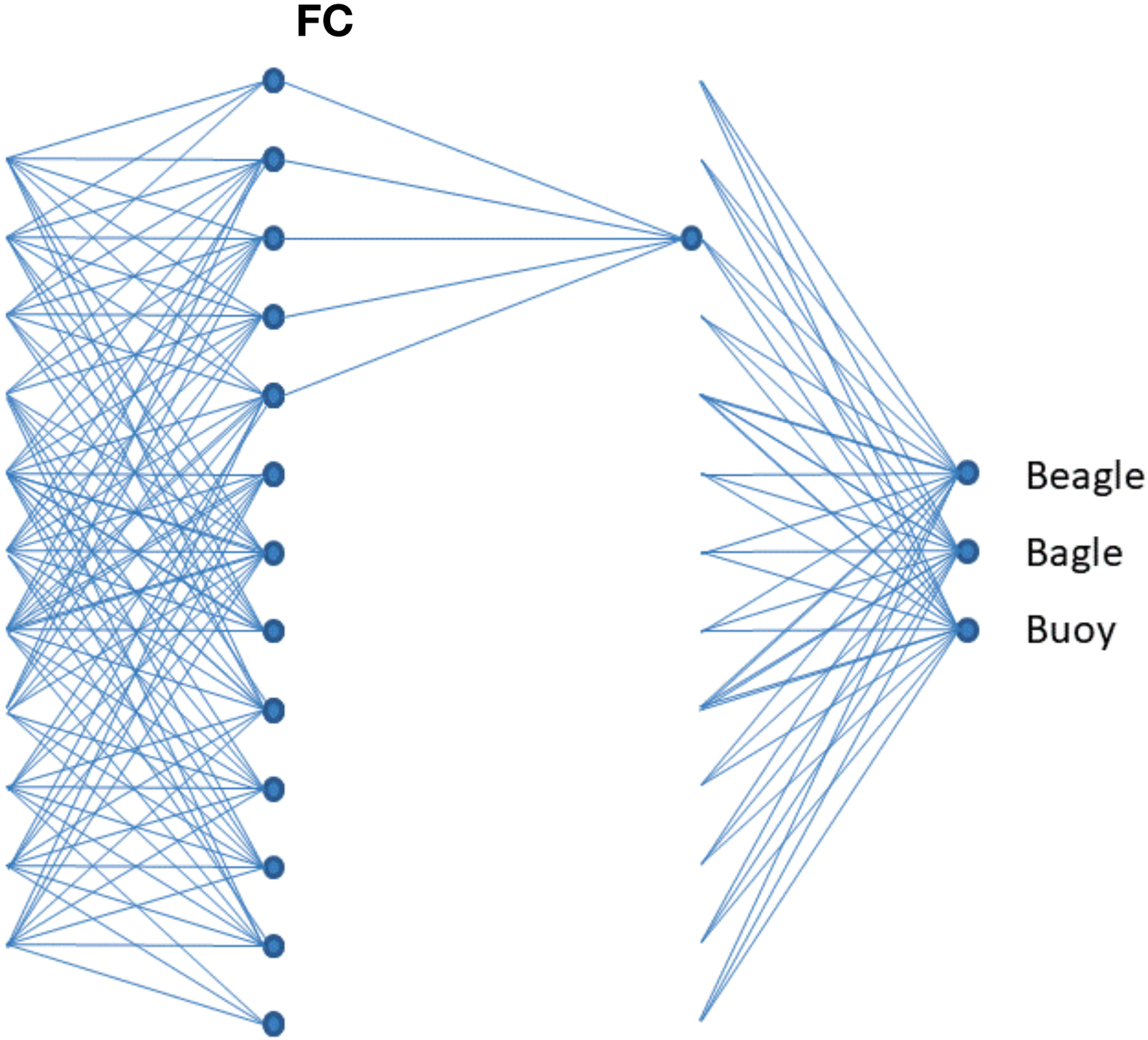
- Deep Learning

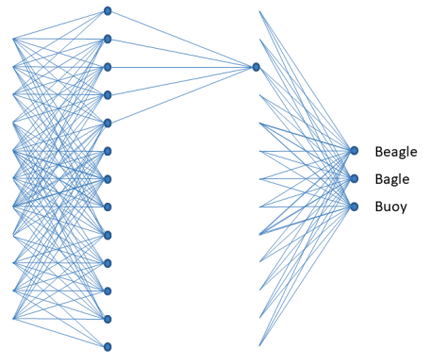
- Architectures

Three main Classes of DL Architectures

Fully Connected / Feed Forward

$$Z^i = W^i X + b^i 1$$
$$A^i = \textbf{RELU}\left(Z^i\right)$$



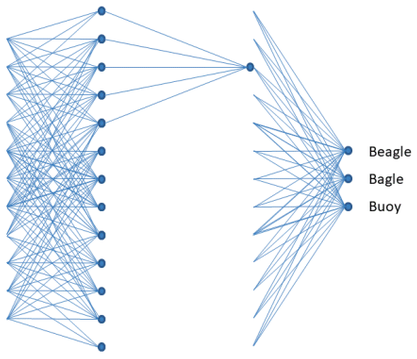


$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

FC



$$Z^i = W^i X + b^i 1$$

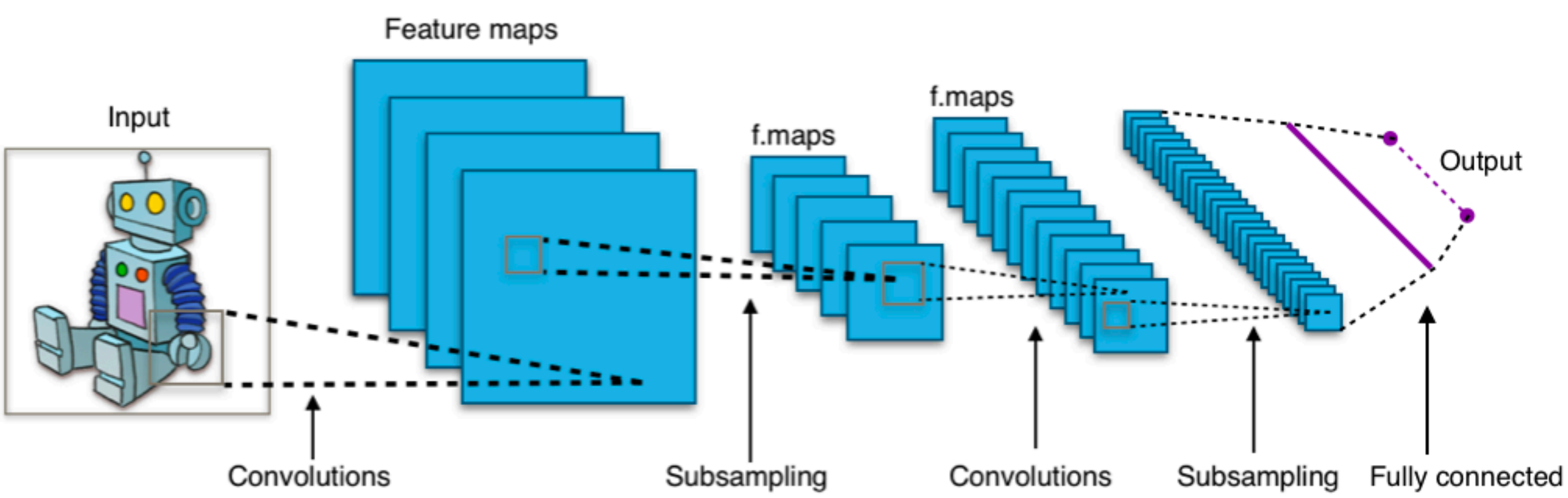
$$A^i = \textbf{RELU}\left(Z^i\right)$$

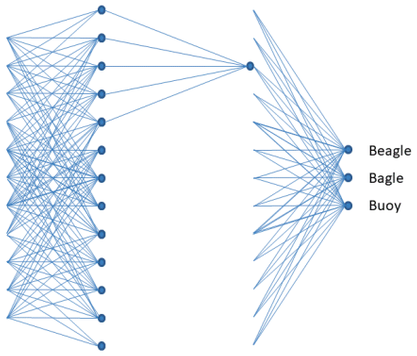
Fully Connected / Feed Forward

FC

CNN

Convolutional Neural Networks



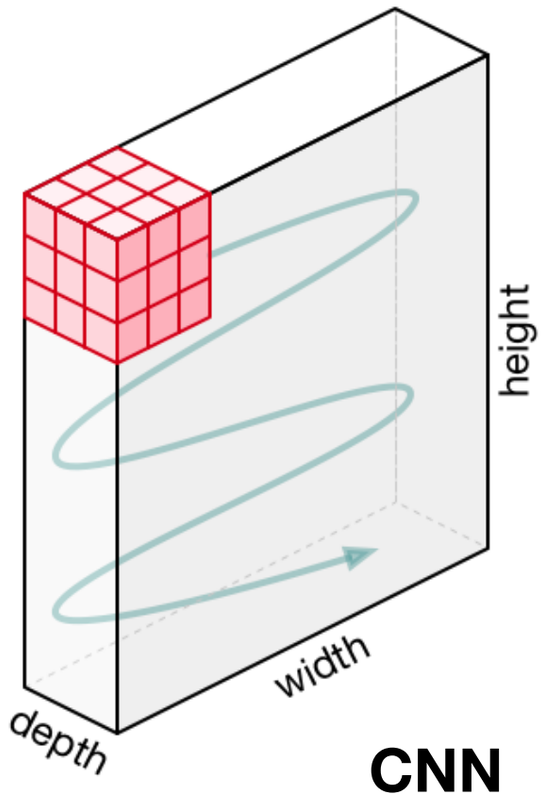


$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

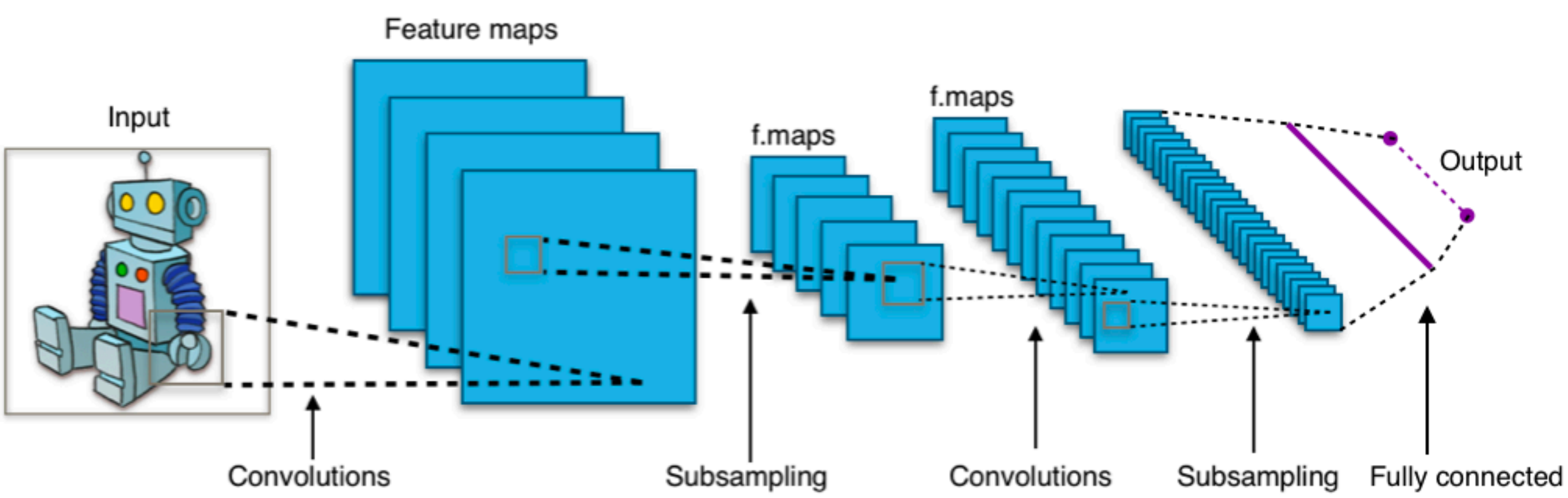
FC

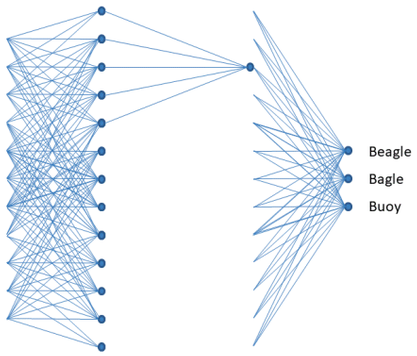


Kernels

CNN

Convolutional Neural Networks



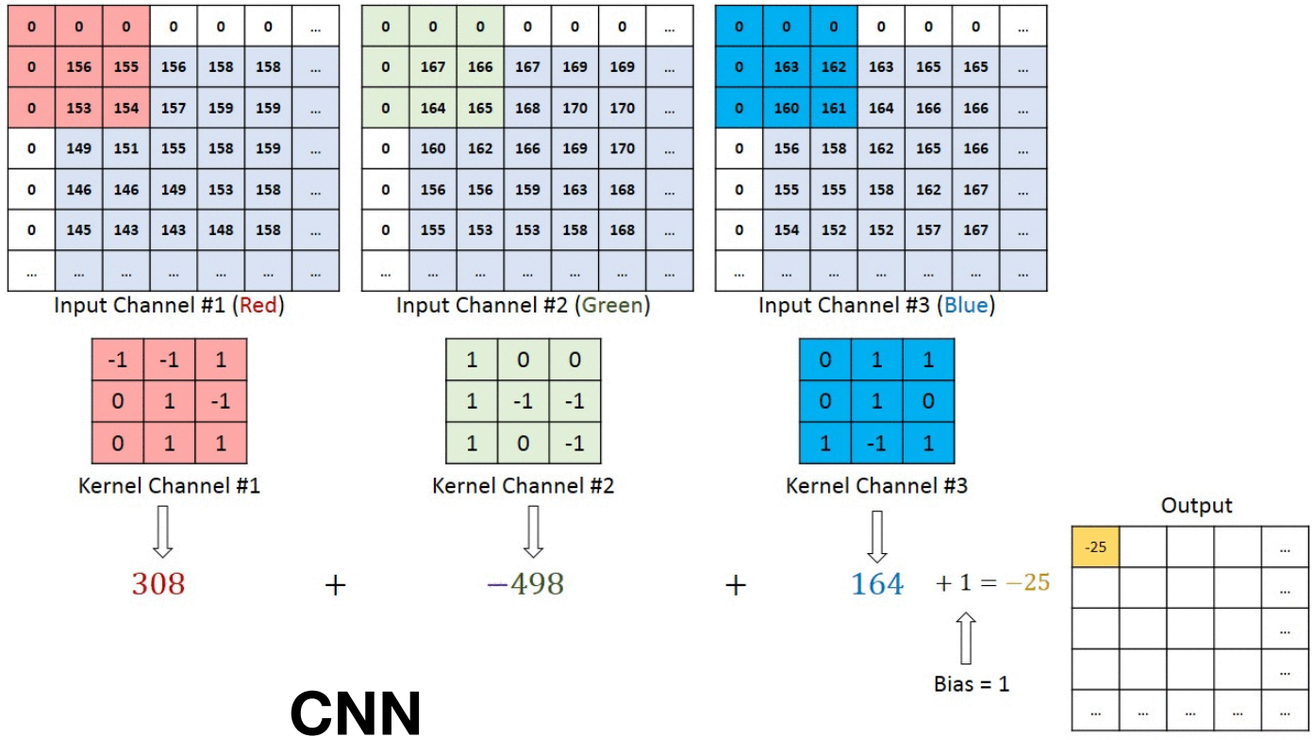


$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

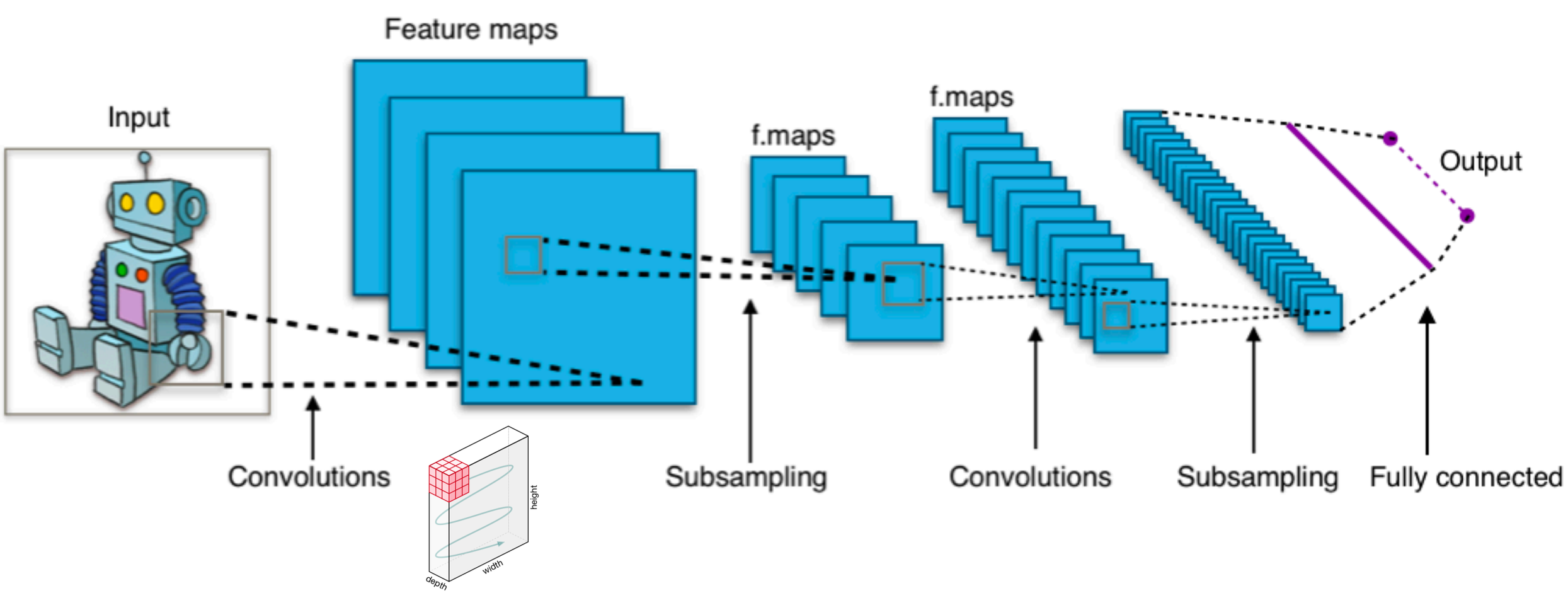
Fully Connected / Feed Forward

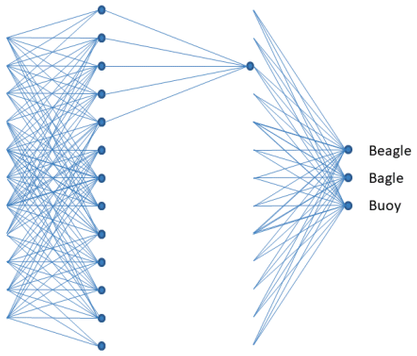
FC



CNN

Convolutional Neural Networks



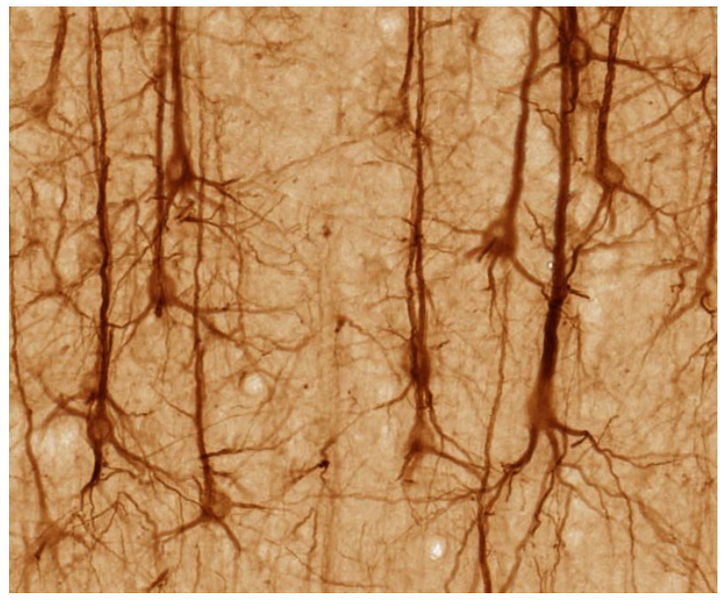


$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

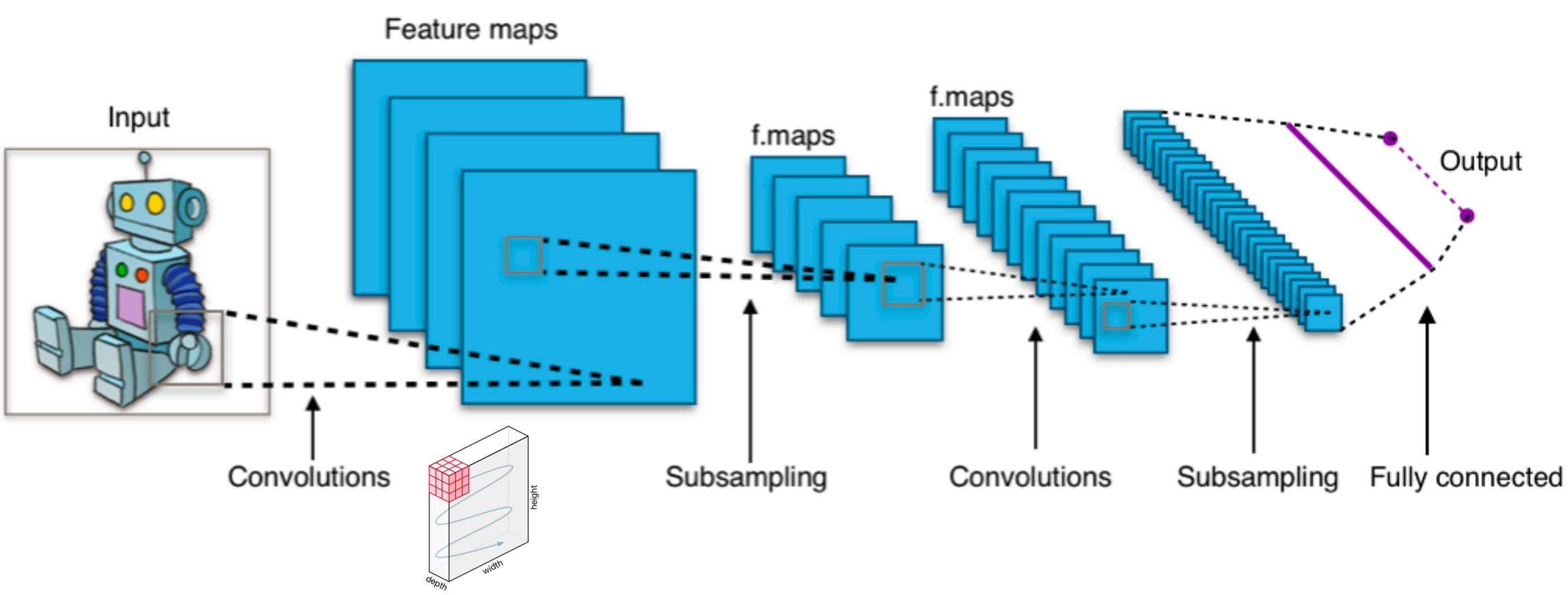
Fully Connected / Feed Forward

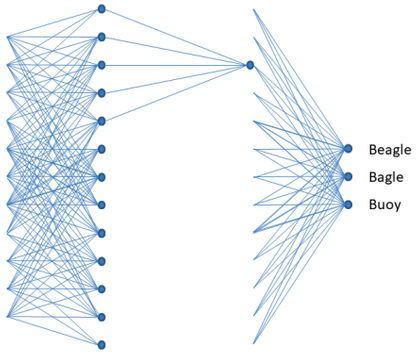
FC



CNN

Convolutional Neural Networks



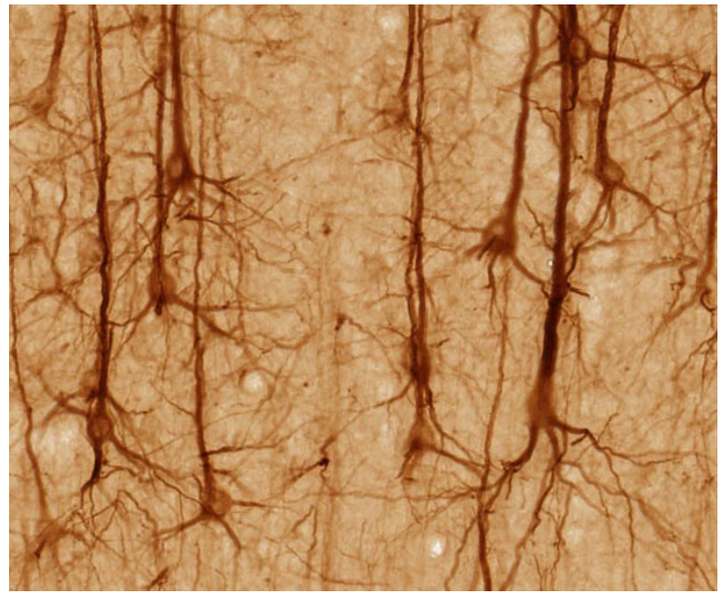


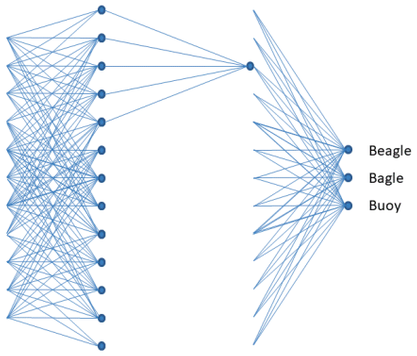
$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

FC





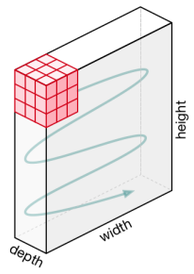
$$Z^i = W^i X + b^i 1$$

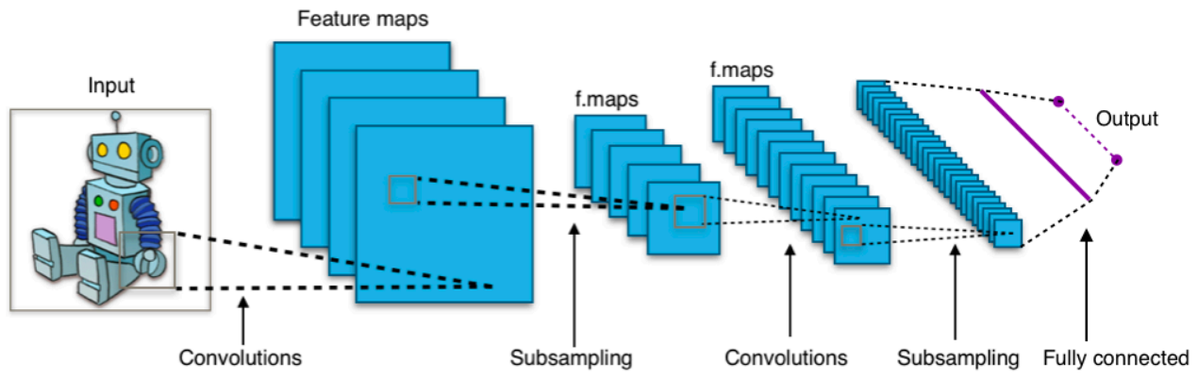
$$A^i = \textbf{RELU}\left(Z^i\right)$$

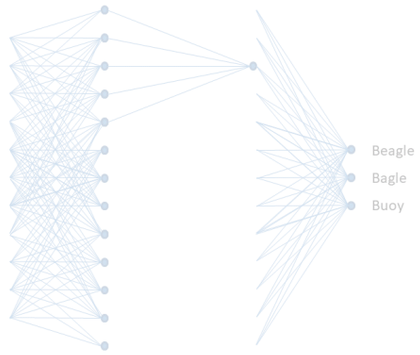
Fully Connected / Feed Forward

FC

CNN
Convolutional Neural Networks







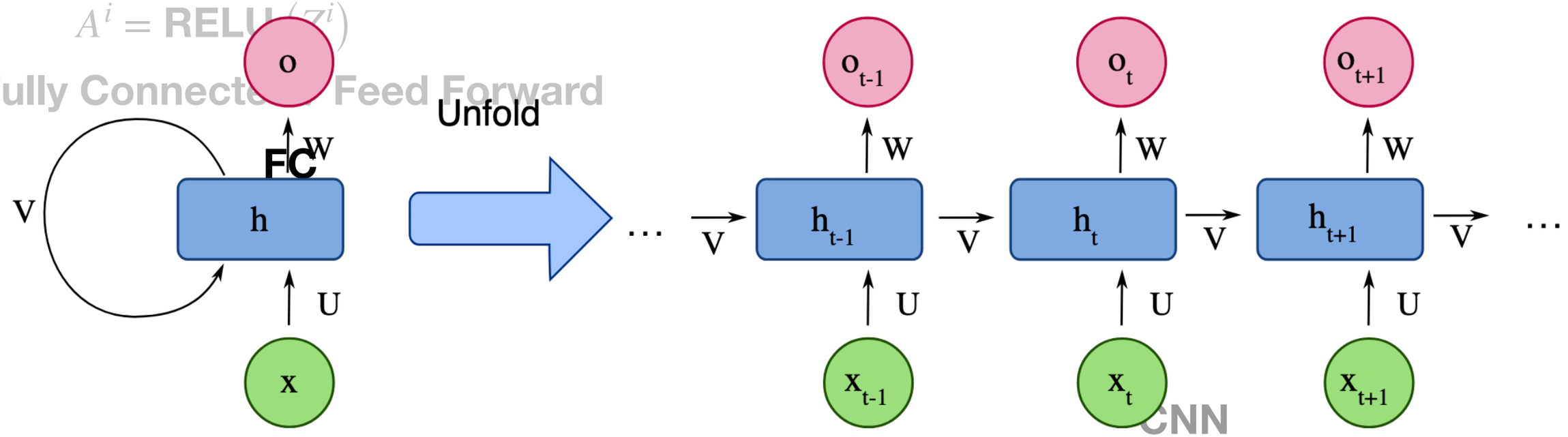
$$Z^i = W^i X + b^i 1$$

$$A^i = \mathbf{RELU}(Z^i)$$

Fully Connected Feed Forward

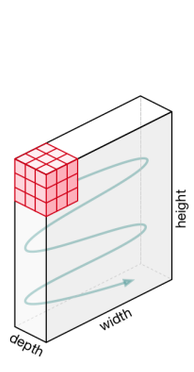
RNN

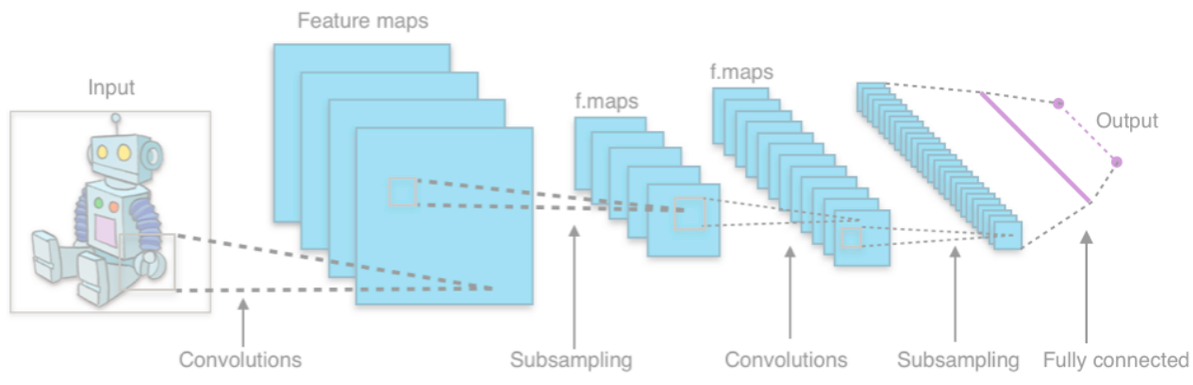
Recurrent Neural Network

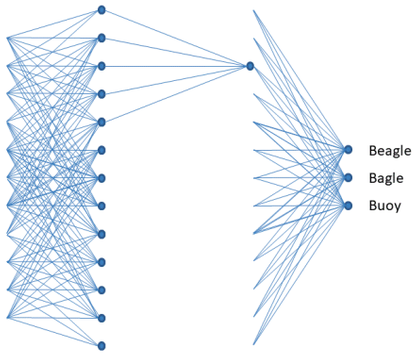


CNN

Convolutional Neural Networks







$$Z^i = W^i X + b^i 1$$

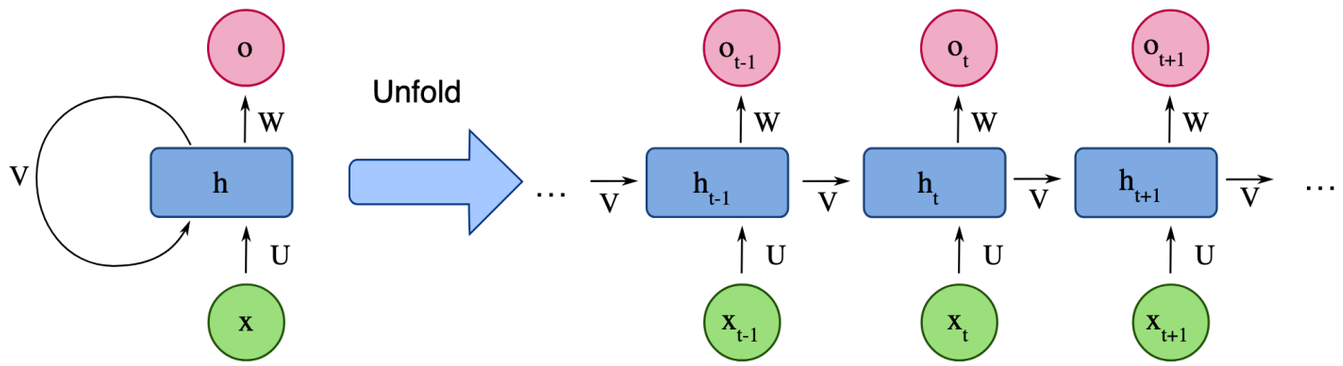
$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

FC

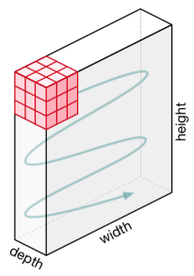
RNN

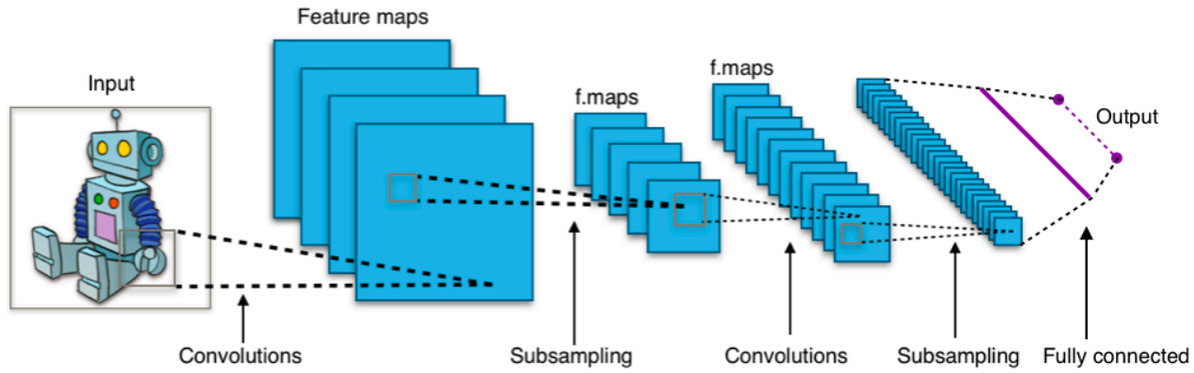
Recurrent Neural Network

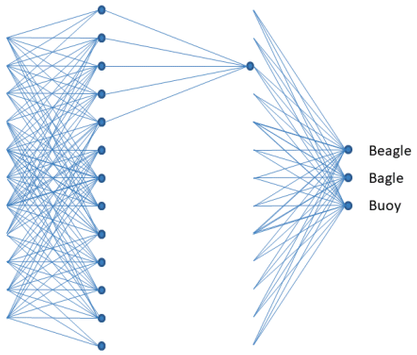


CNN

Convolutional Neural Networks







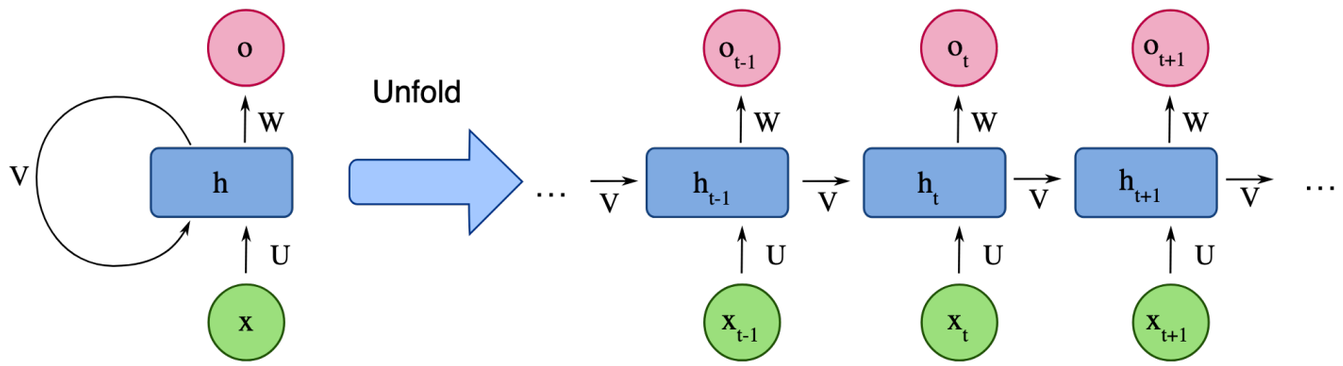
$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

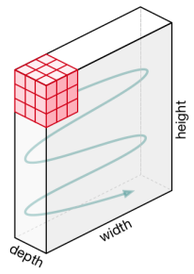
FC

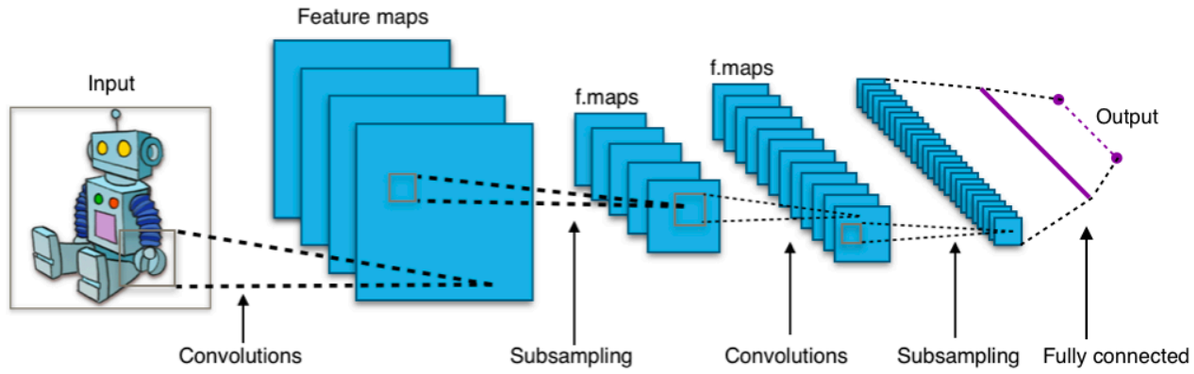
RNN
Recurrent Neural Network

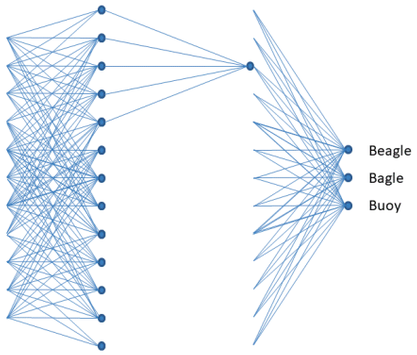


GANs,
Auto Encoders,
ODE Networks,
Invertible Flow Networks,
…..

CNN
Convolutional Neural Networks







$$Z^i = W^i X + b^i 1$$

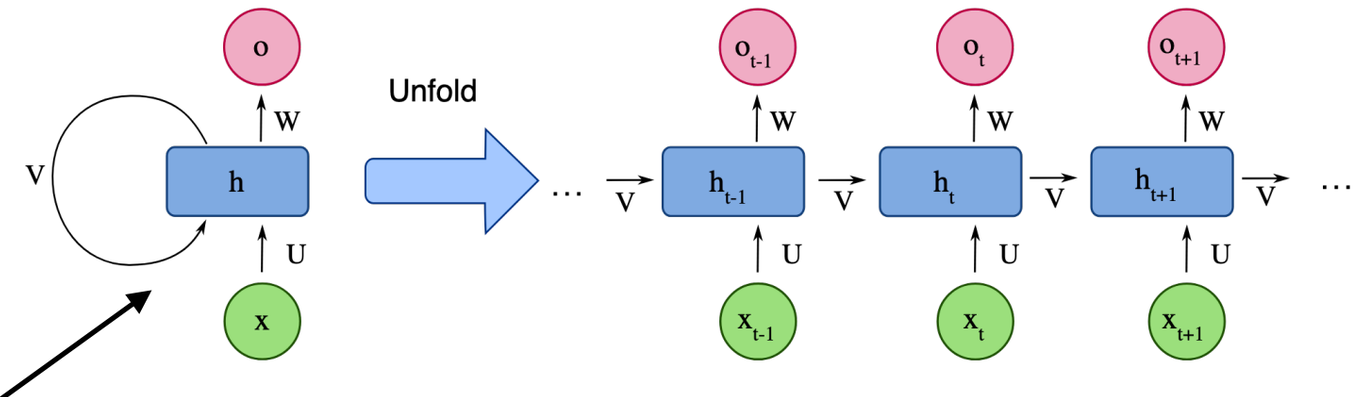
$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

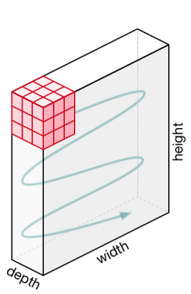
FC

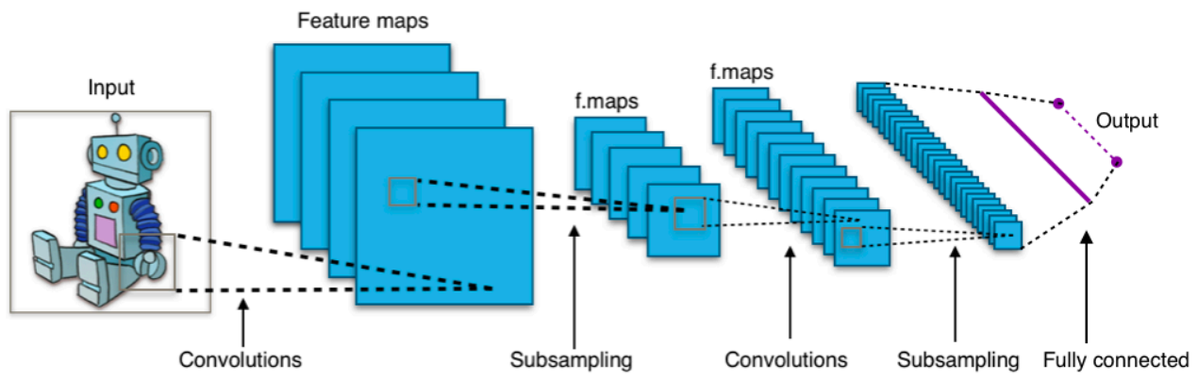
GANs,
Auto Encoders,
ODE Networks,
Invertible Flow Networks,
.....

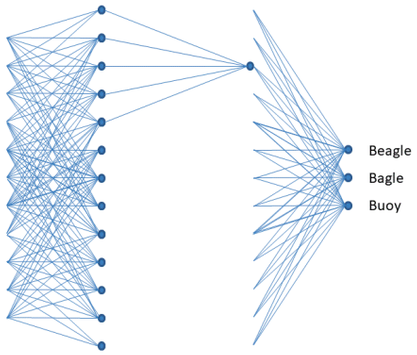
RNN
Recurrent Neural Network



CNN
Convolutional Neural Networks







$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

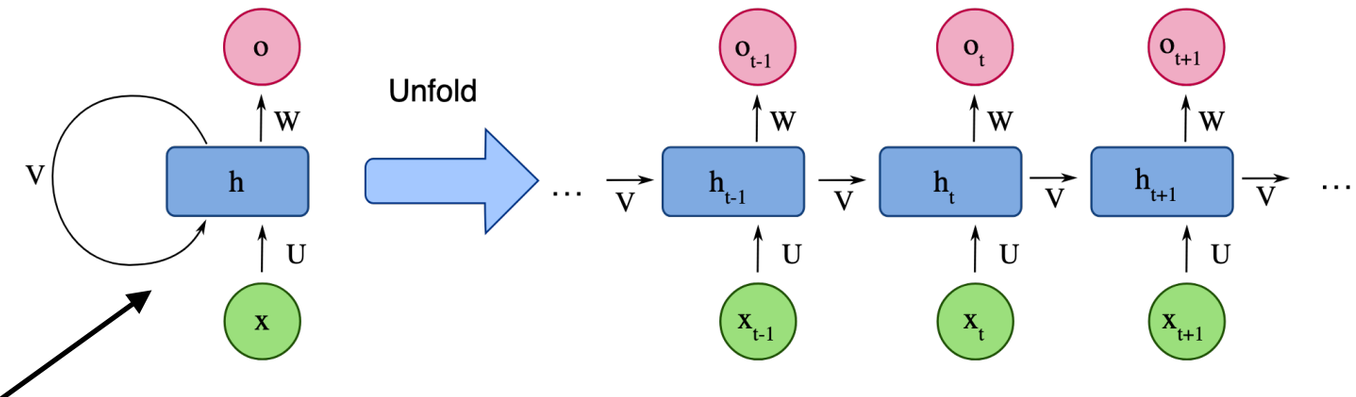
Fully Connected / Feed Forward

FC

GANs,
Auto Encoders,
ODE Networks,
Invertible Flow Networks,
.....

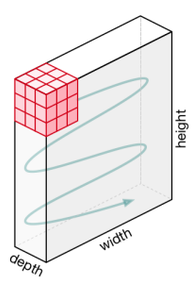
RNN

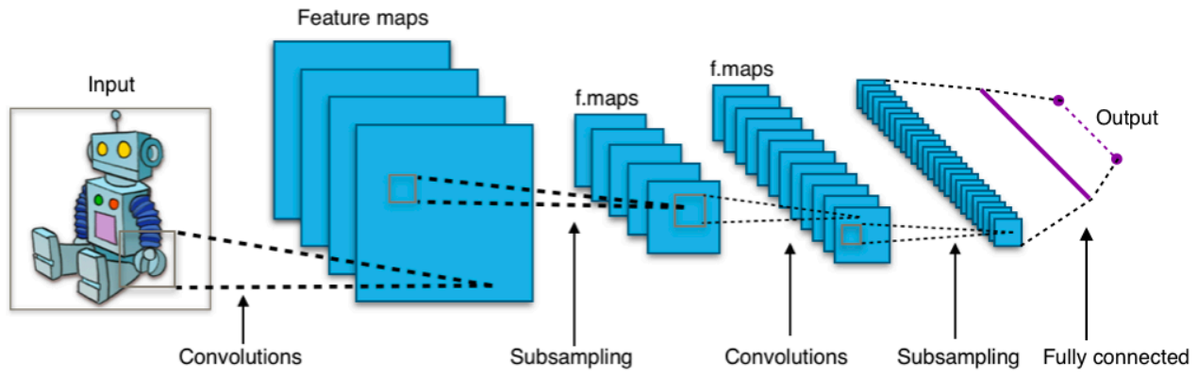
Recurrent Neural Network



CNN

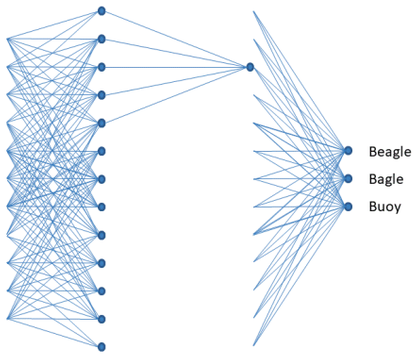
Convolutional Neural Networks





All kind of Domains: Medical Imaging, Autonomous Driving, Emotion Recognition, Recommenders, Natural Language Processing

Supervised,
Unsupervised,
Self-Supervised,
Reinforcement Learning



$$Z^i = W^i X + b^i 1$$

$$A^i = \textbf{RELU}\left(Z^i\right)$$

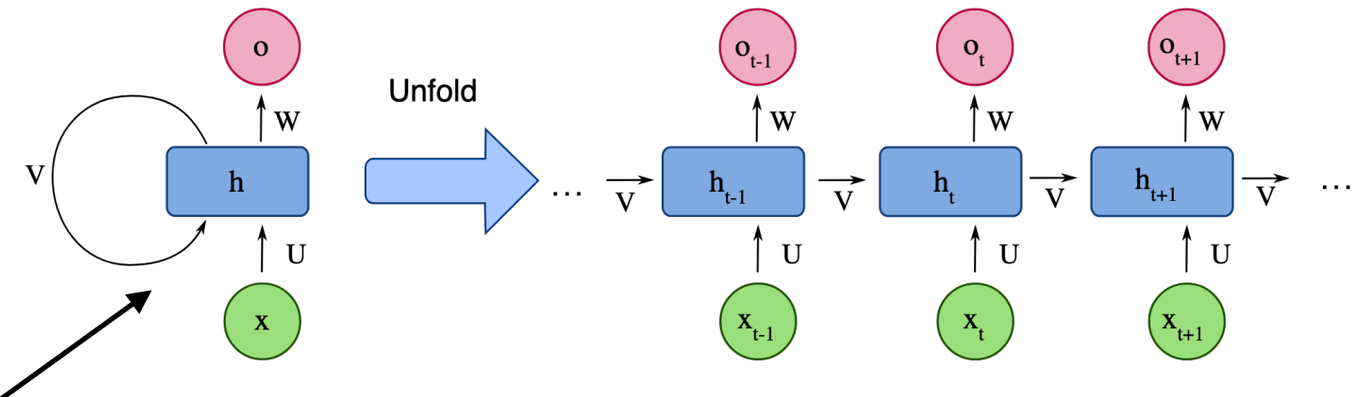
Fully Connected / Feed Forward

FC

GANs,
Auto Encoders,
ODE Networks,
Invertible Flow Networks,
.....

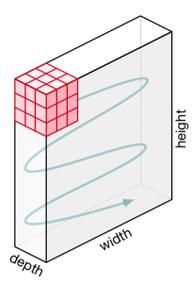
RNN

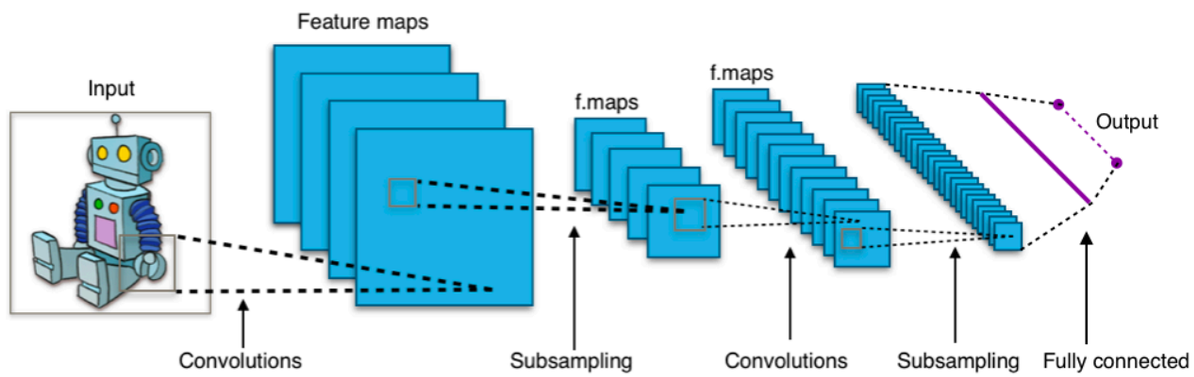
Recurrent Neural Network



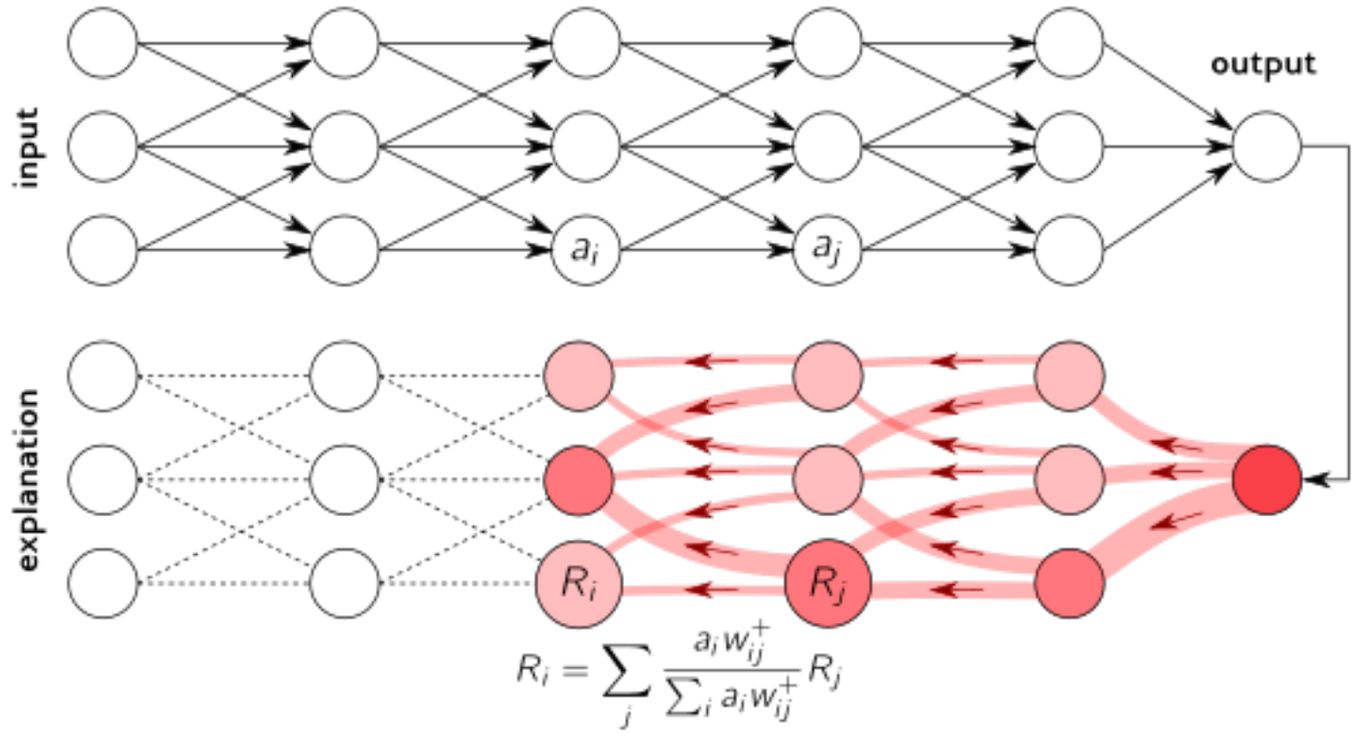
CNN

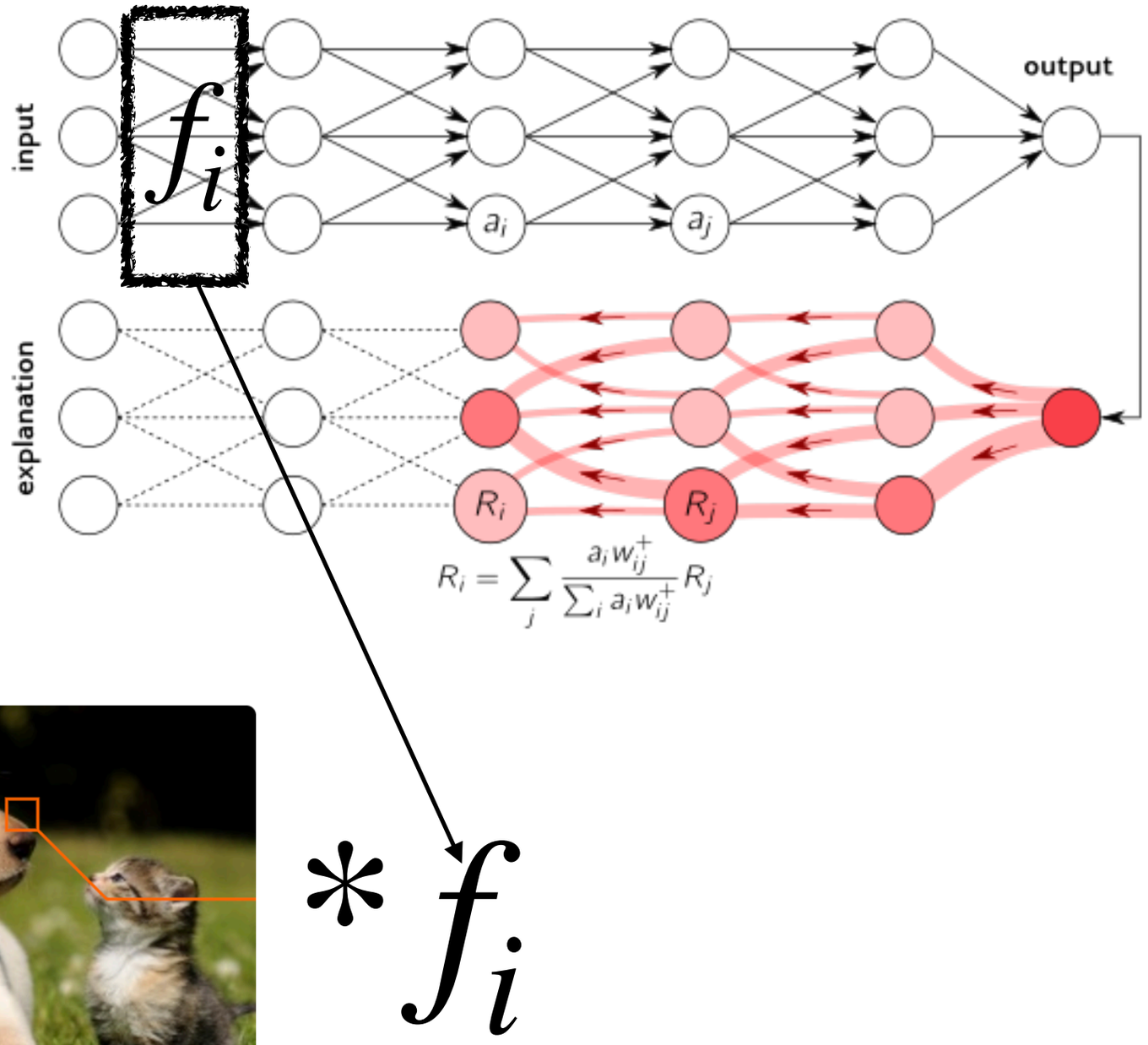
Convolutional Neural Networks

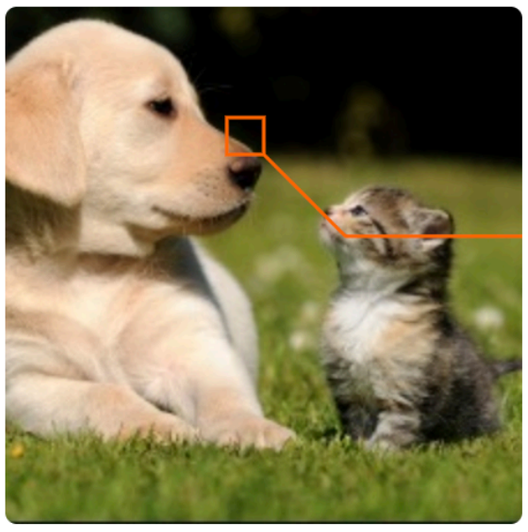




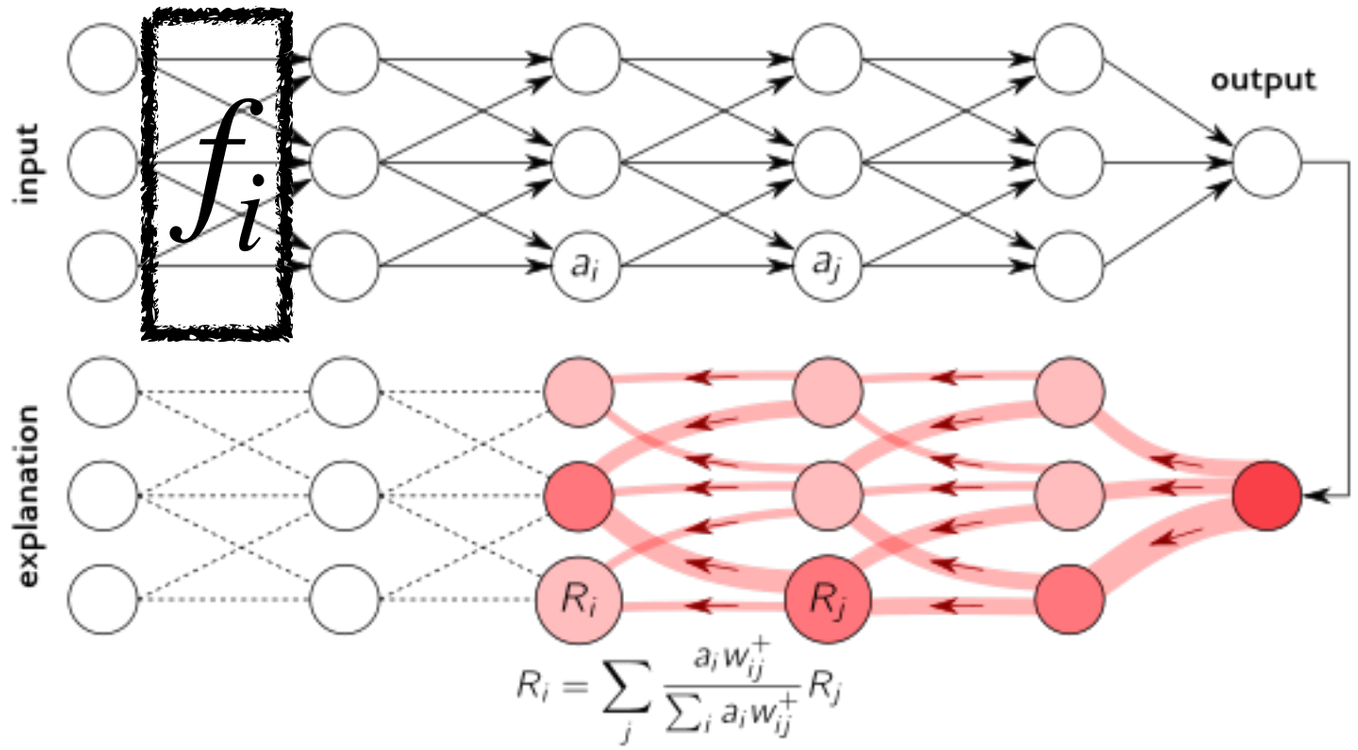
All kind of Domains: Medical Imaging, Autonomous Driving, Emotion Recognition, Recommenders, Natural Language Processing

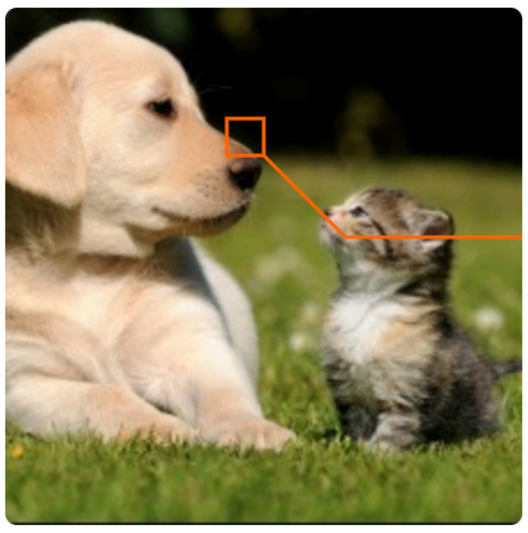

input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$


input
output
explanation
f_i
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$

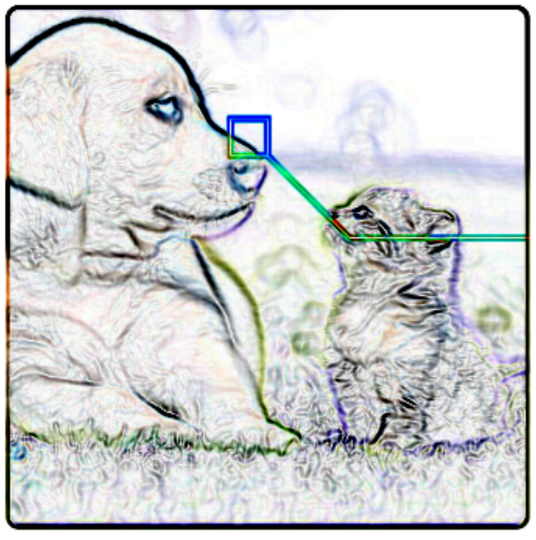


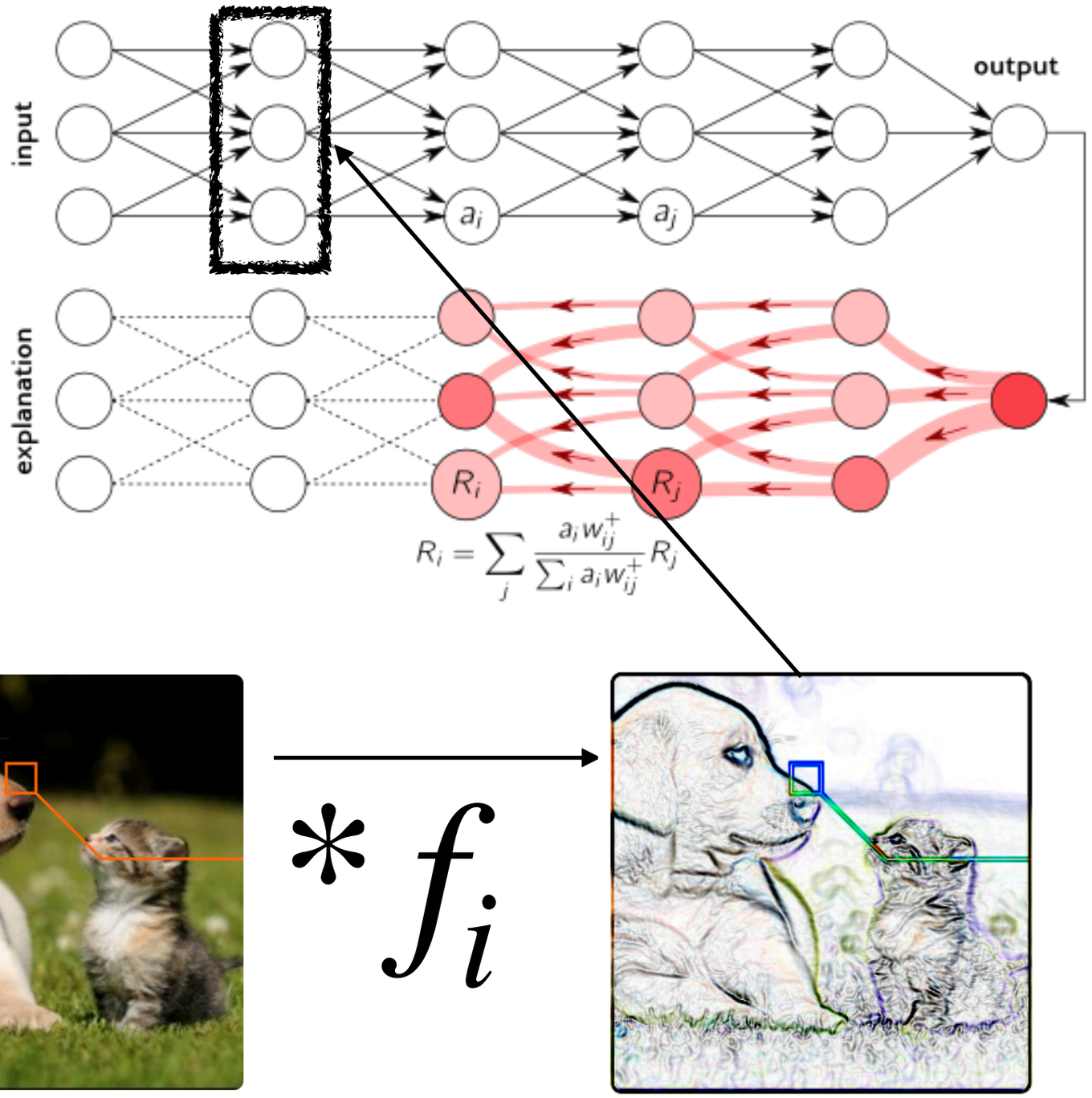
$* f_i$


input
output
explanation
f_i
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$

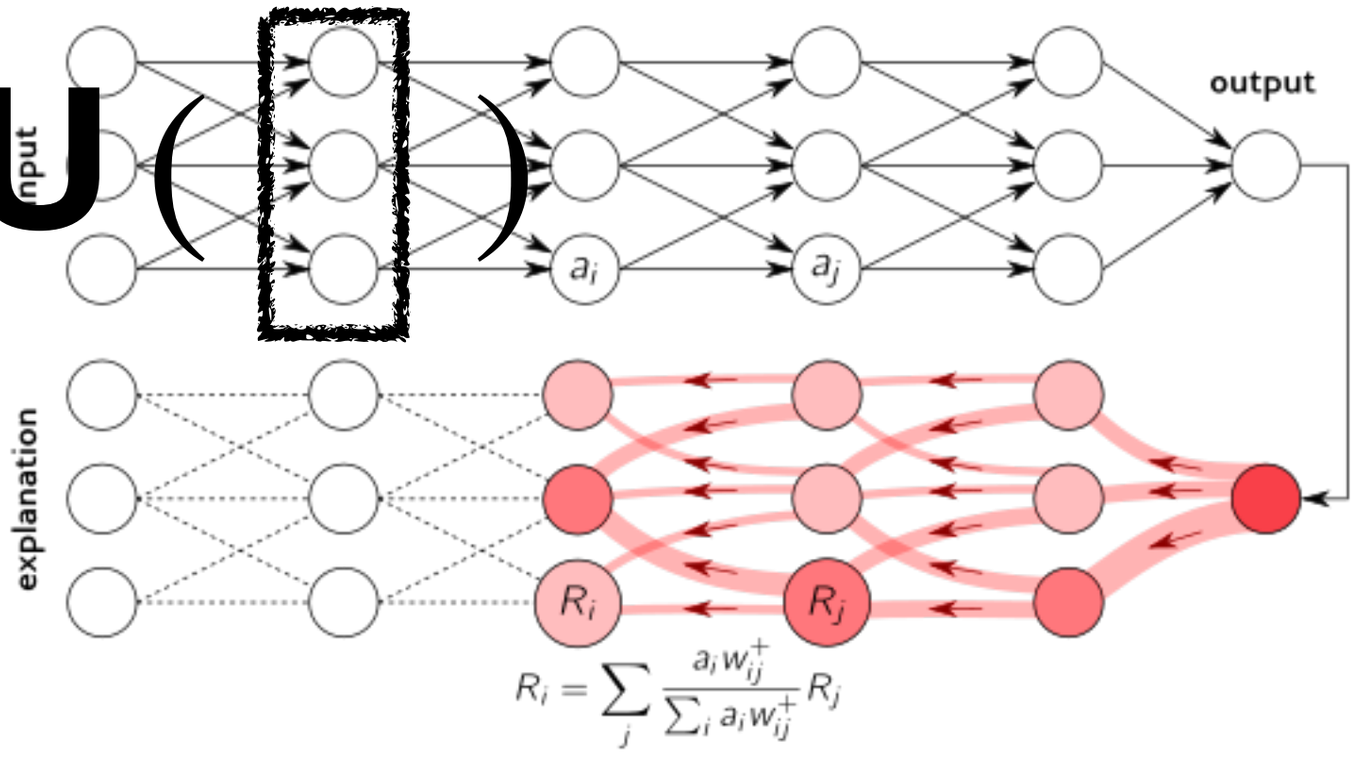


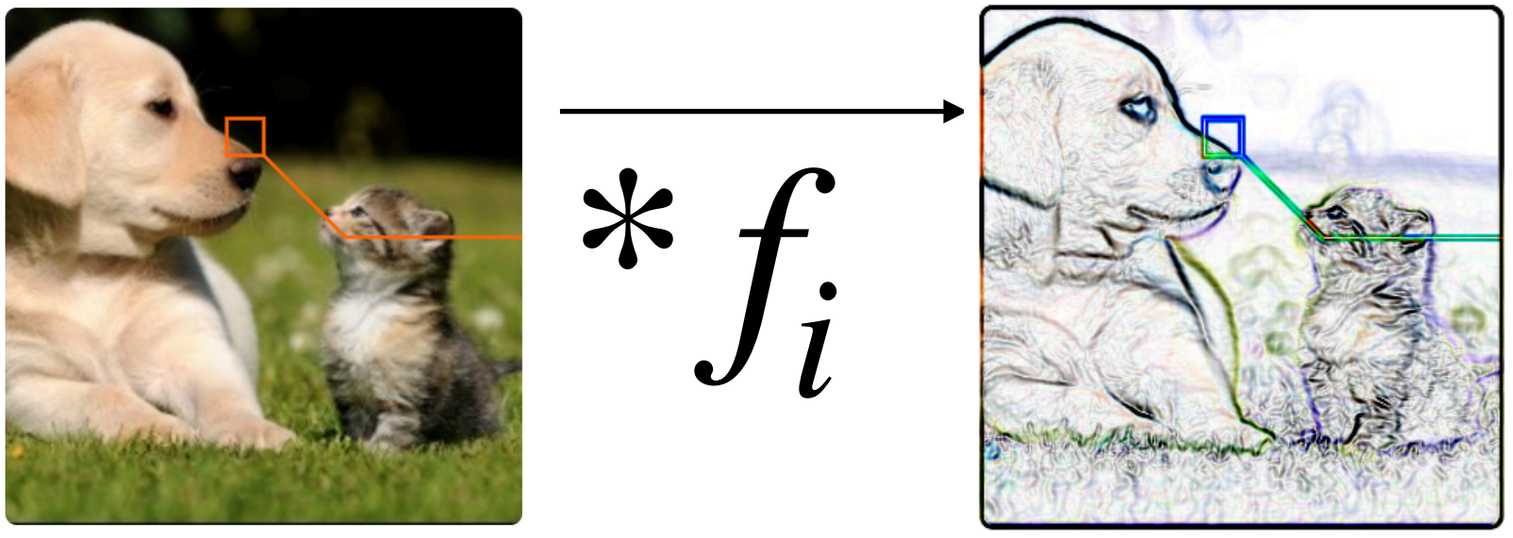
$* f_i$



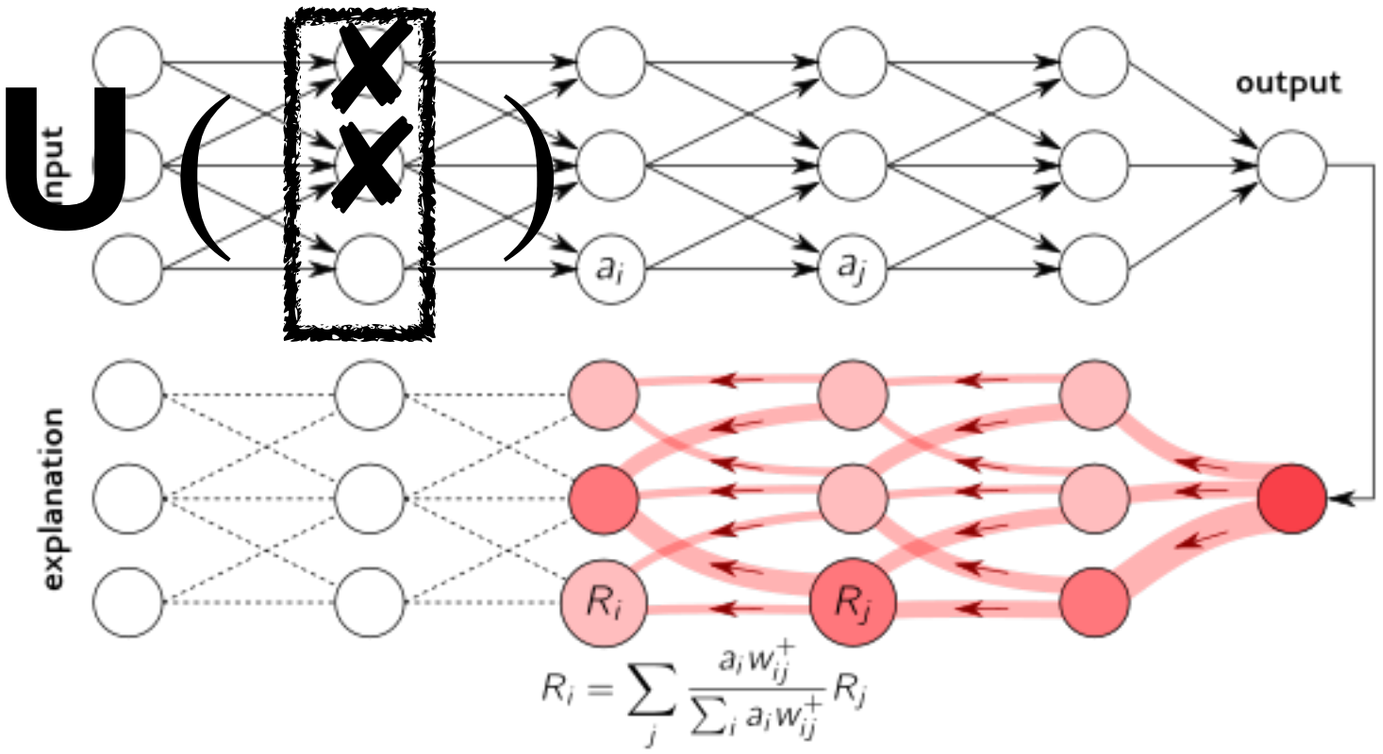

input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
$* f_i$

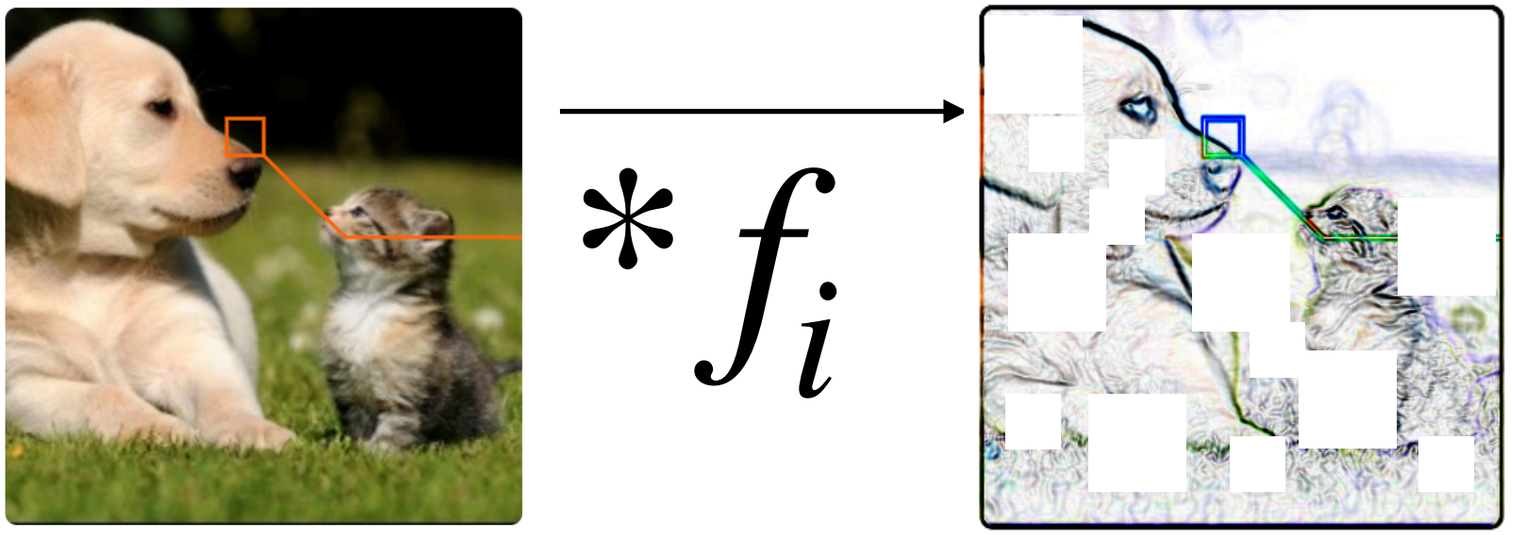
RELU( )

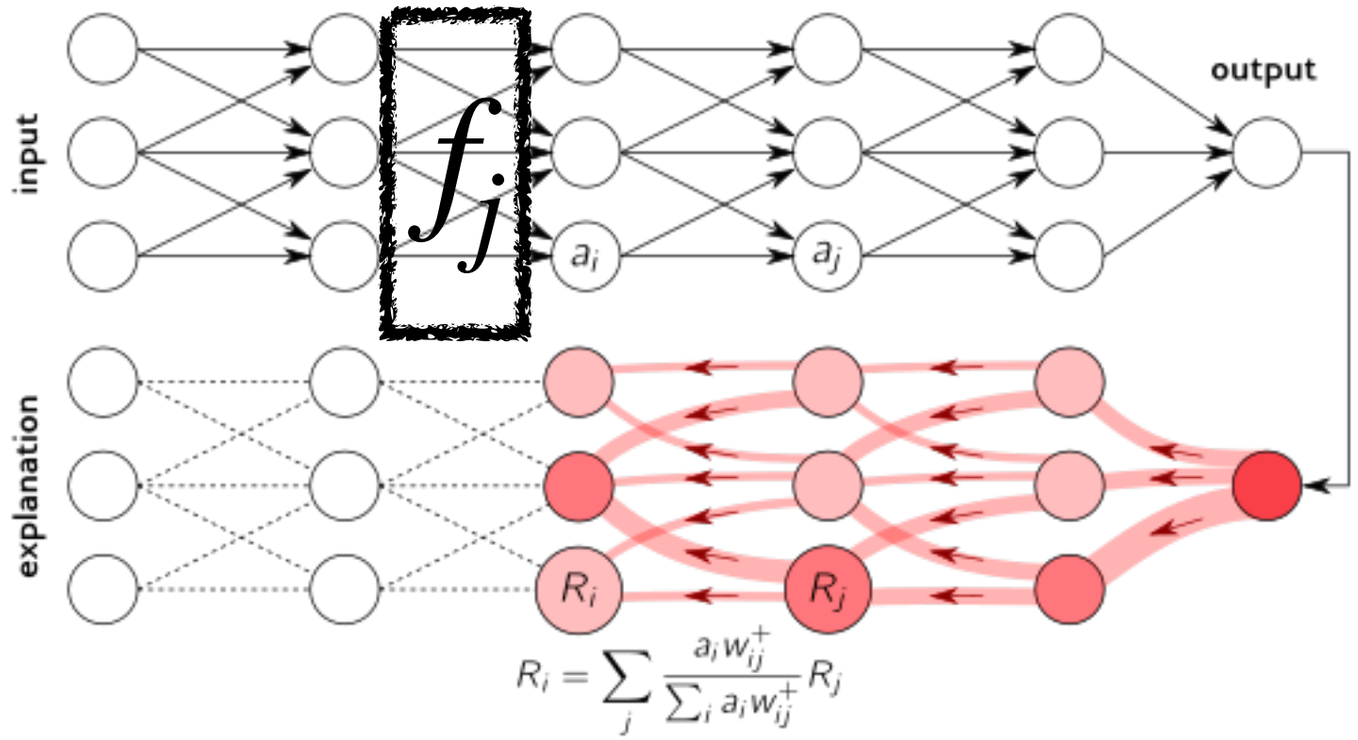

output
a_i
a_j
explanation
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$

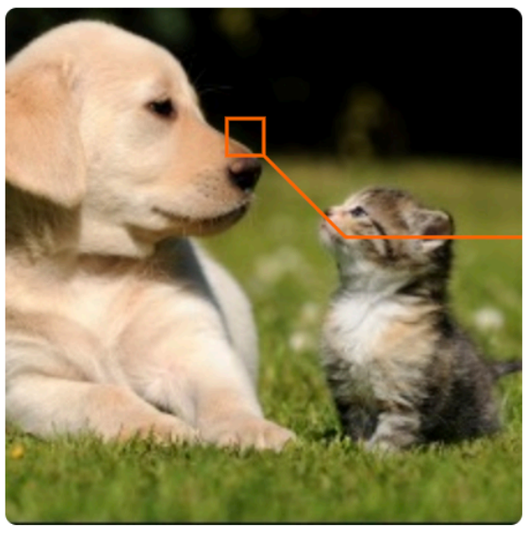

$* f_i$

RELU( )

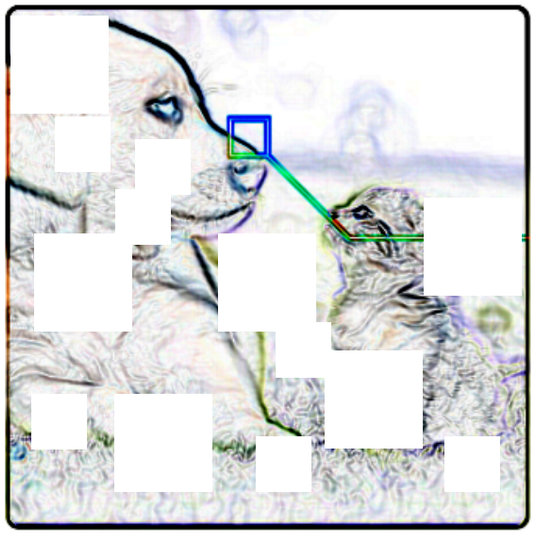

output
a_i
a_j
explanation
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$


$* f_i$

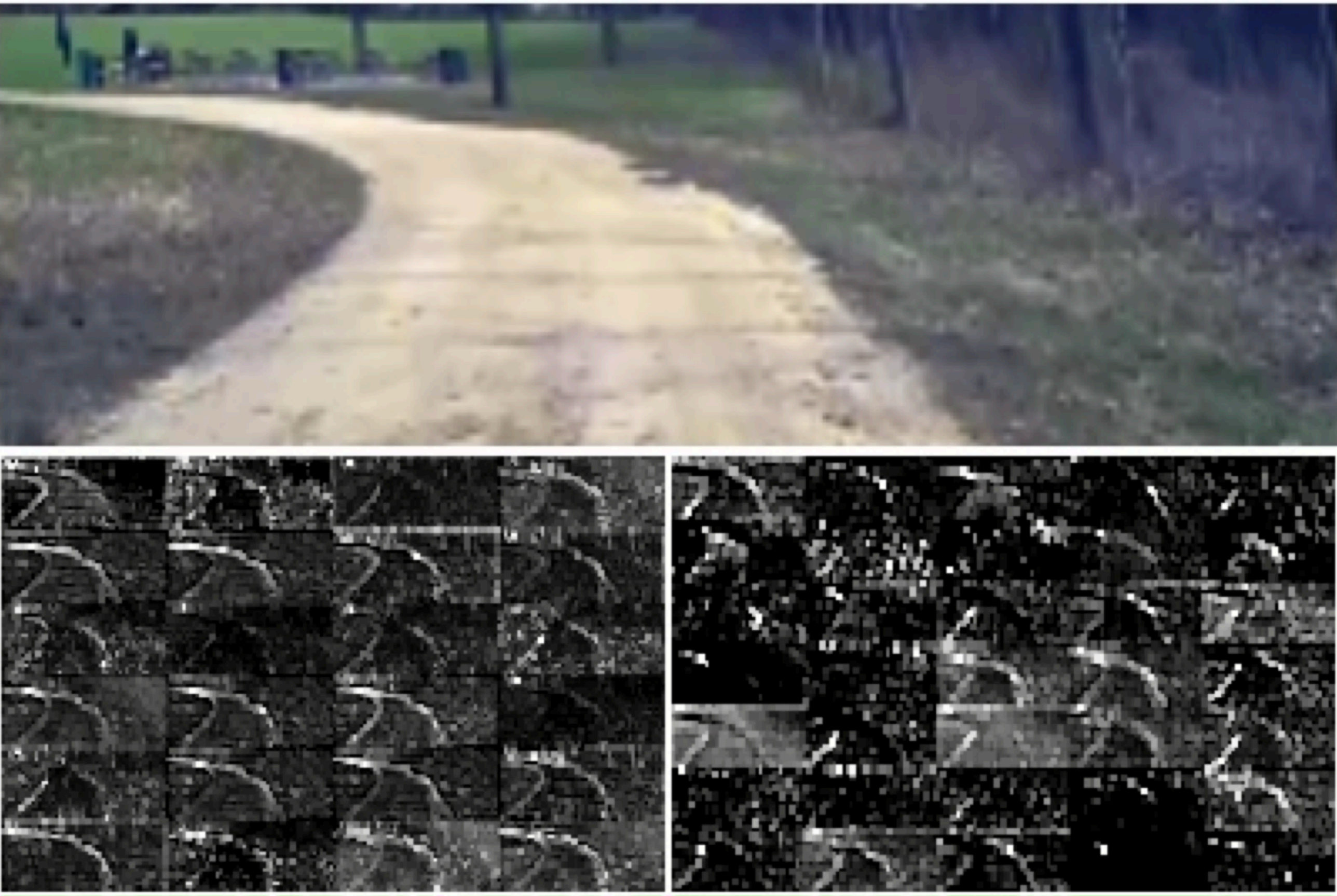

input
output
explanation
f_j
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$



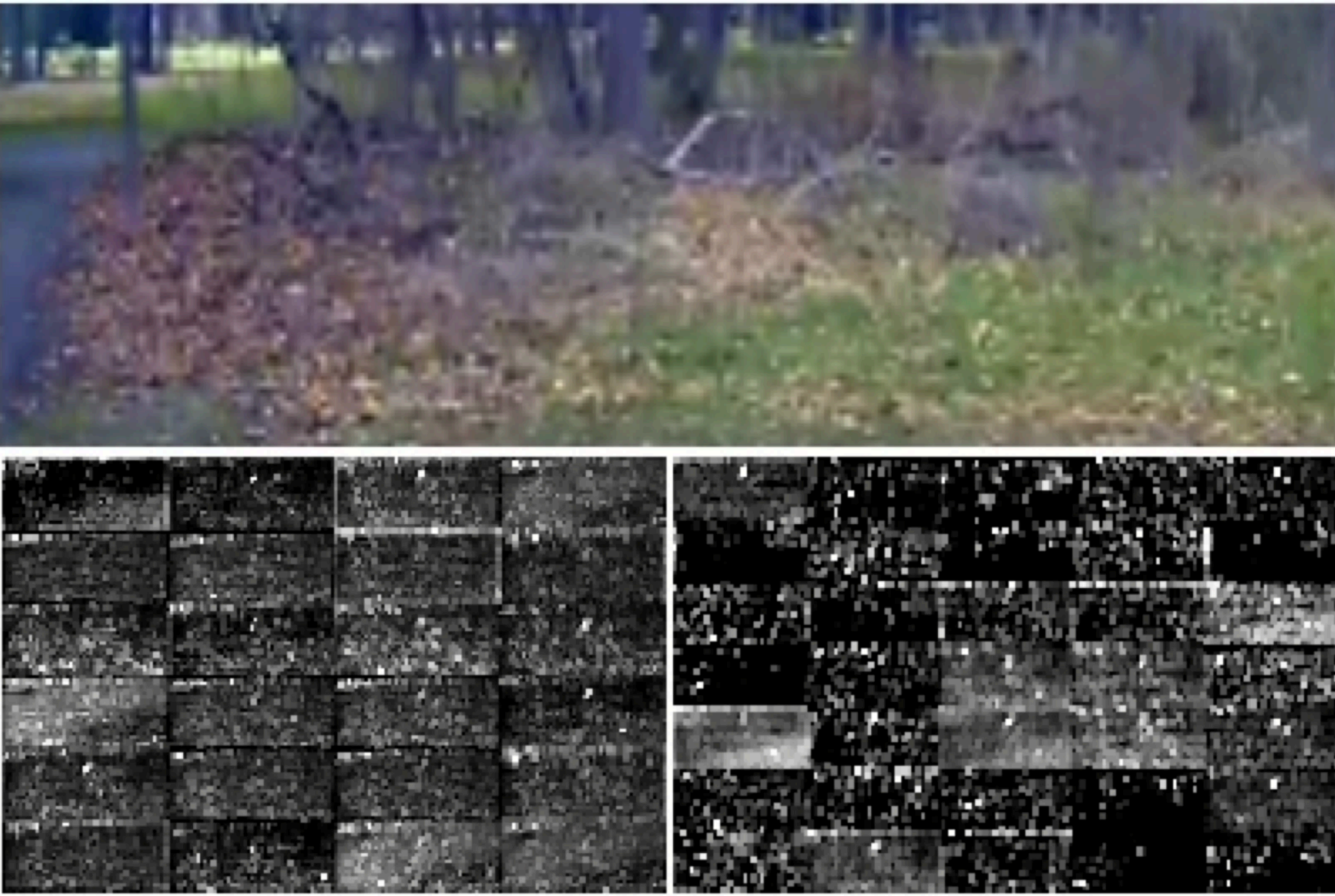
$* f_i$



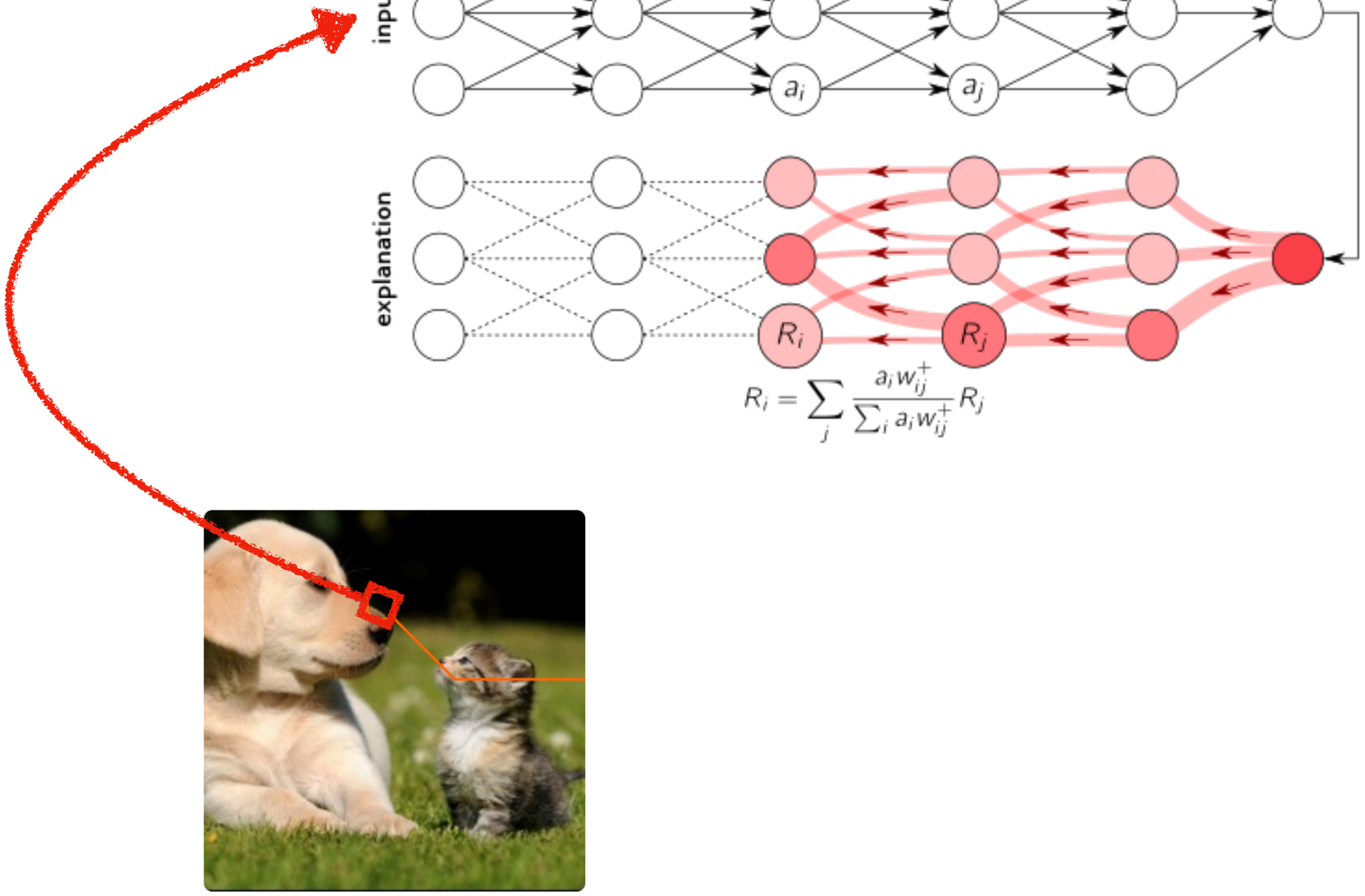
$* f_j$

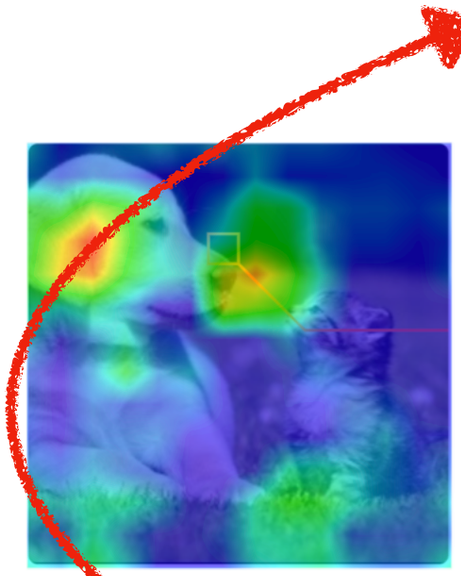


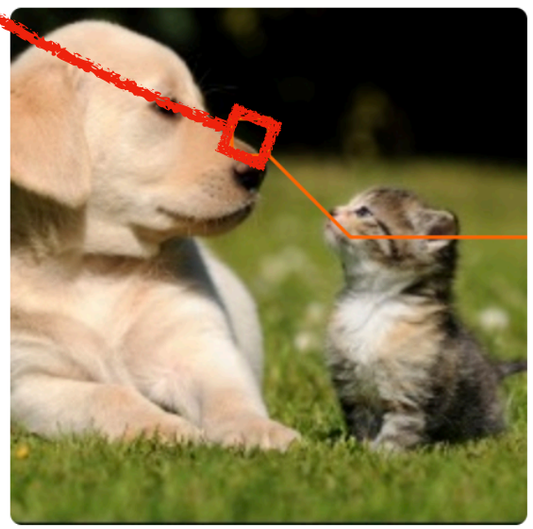
2016 Bojarski et al., End to End Learning for Self-Driving Cars

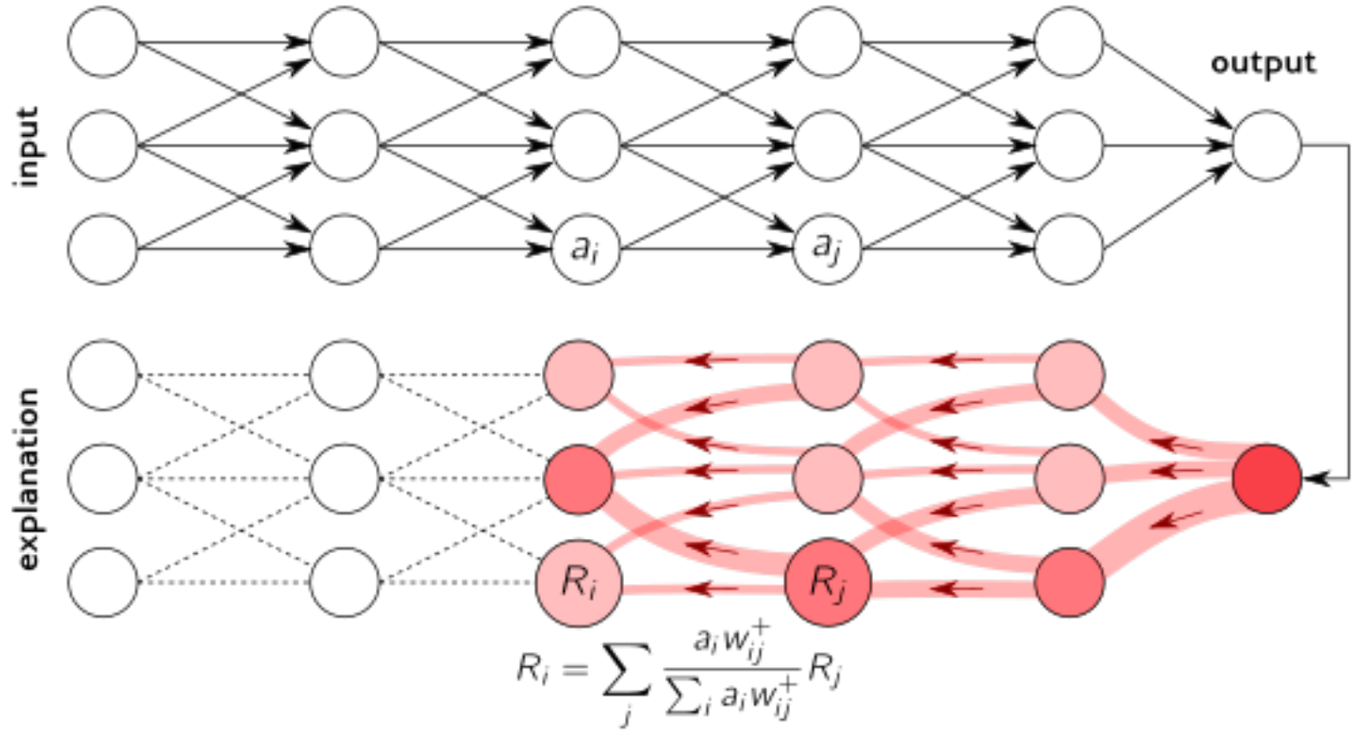


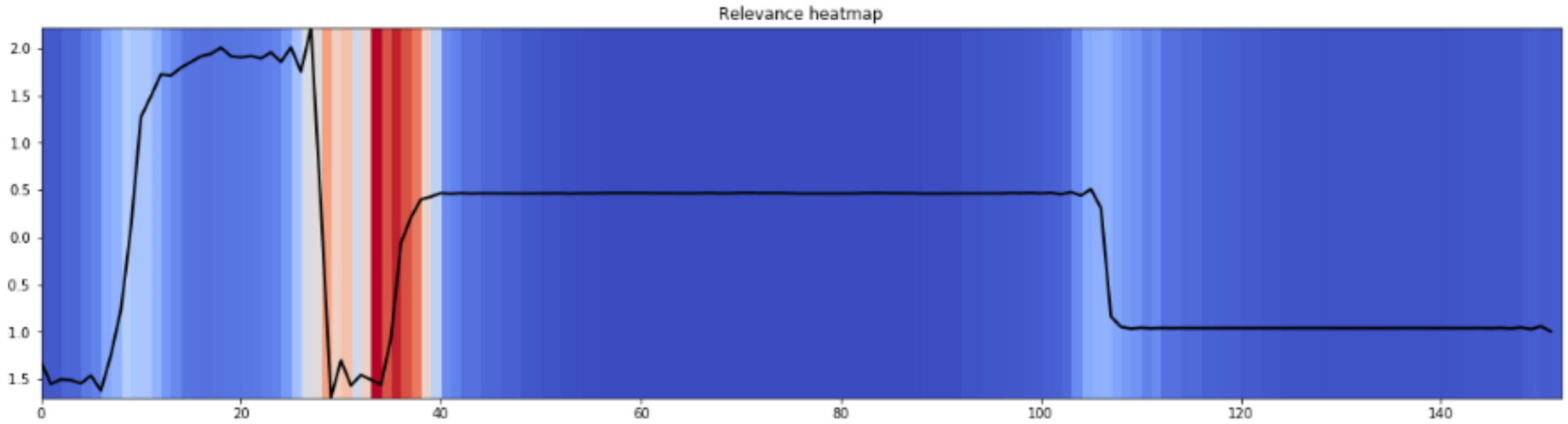
2016 Bojarski et al., End to End Learning for Self-Driving Cars

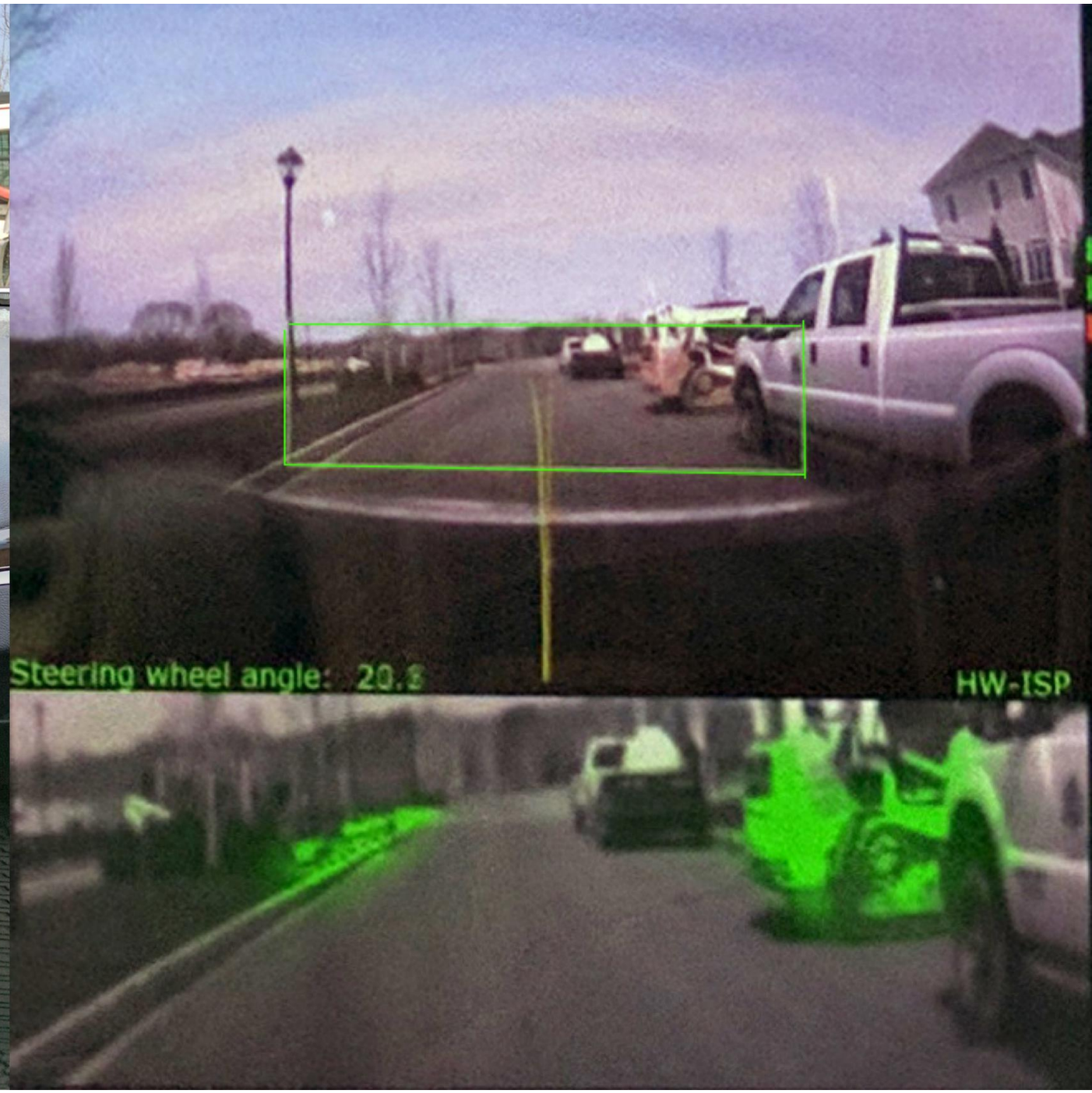

input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$



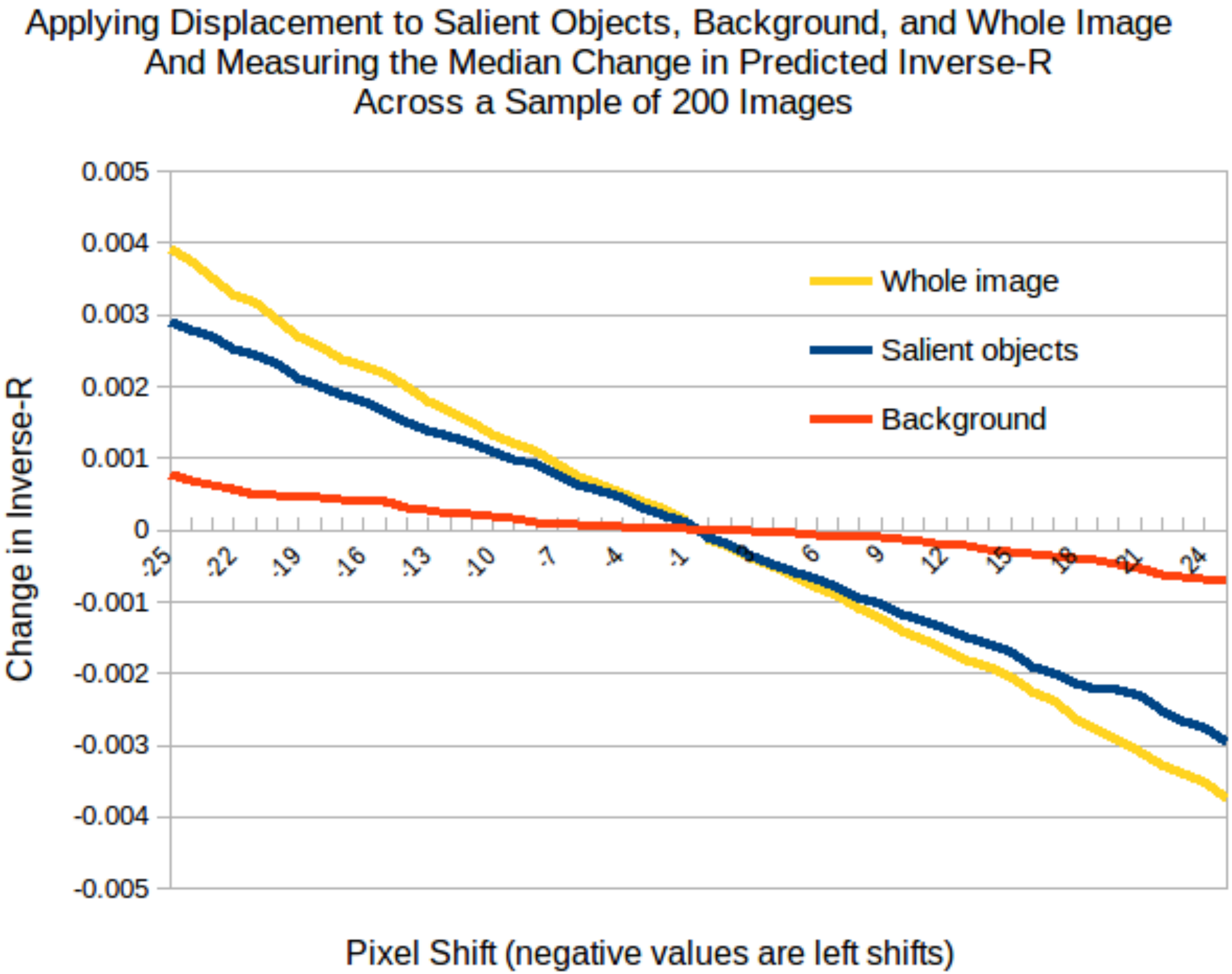



input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$

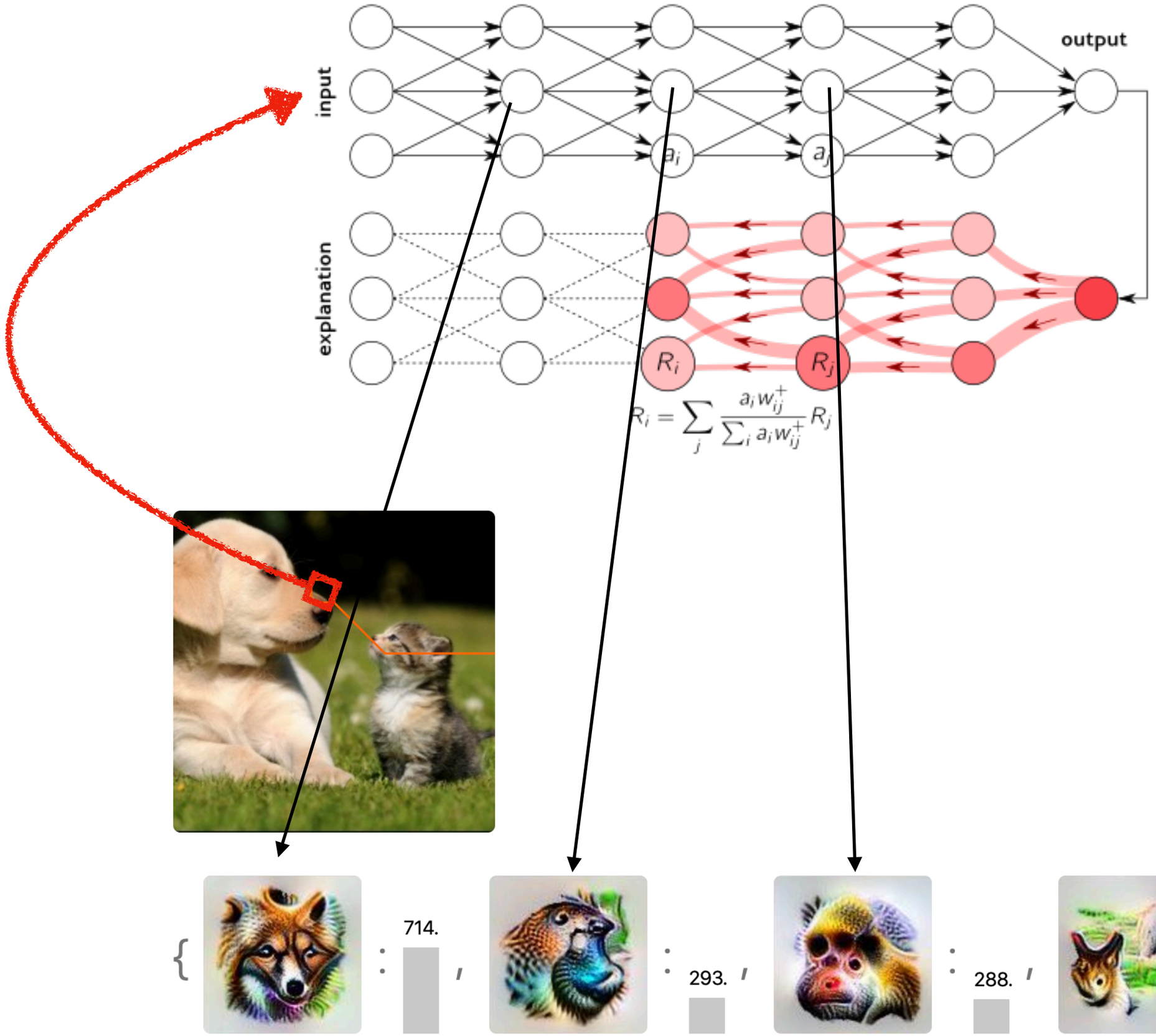

Relevance heatmap
2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
0
20
40
60
80
100
120
140

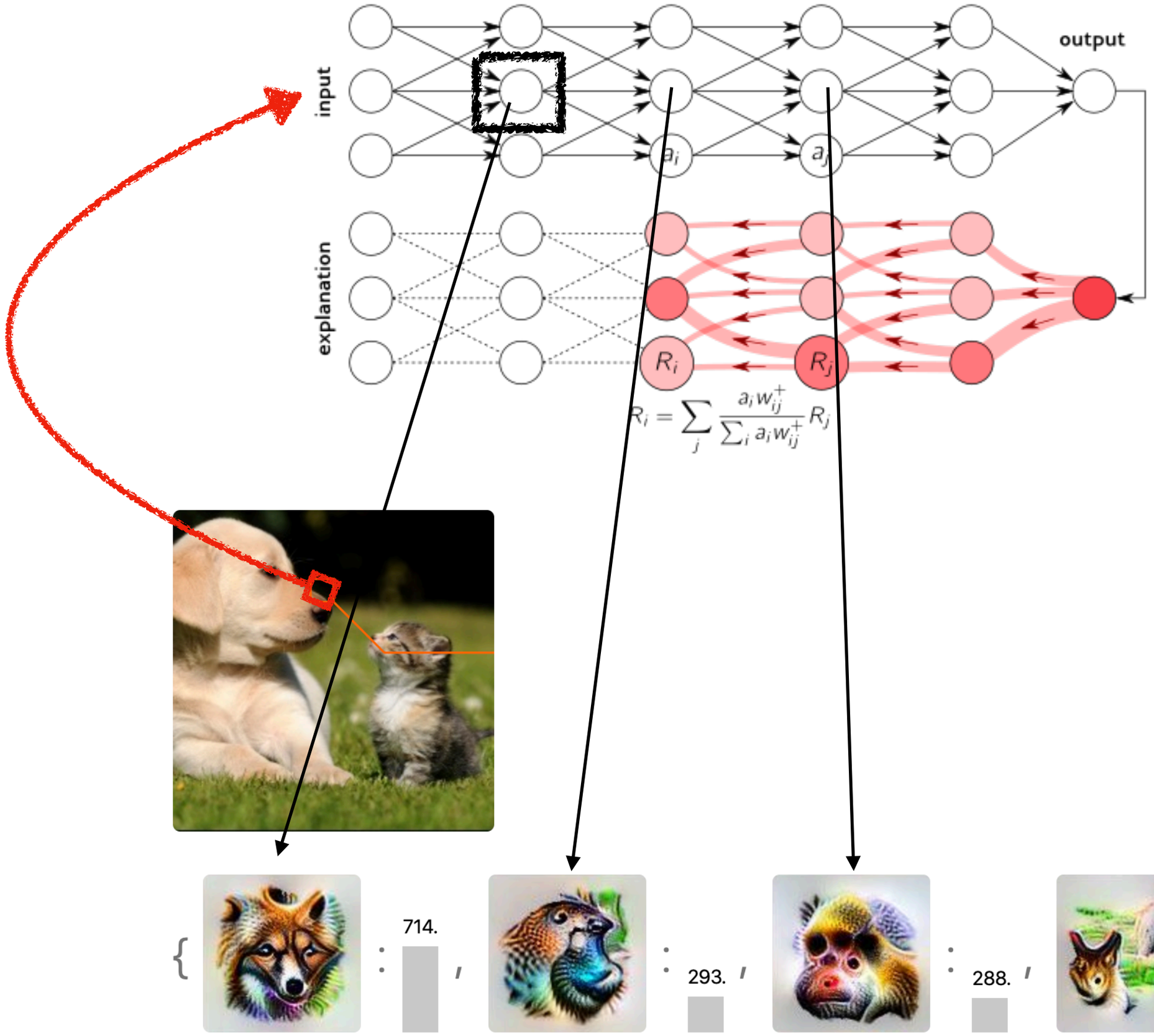


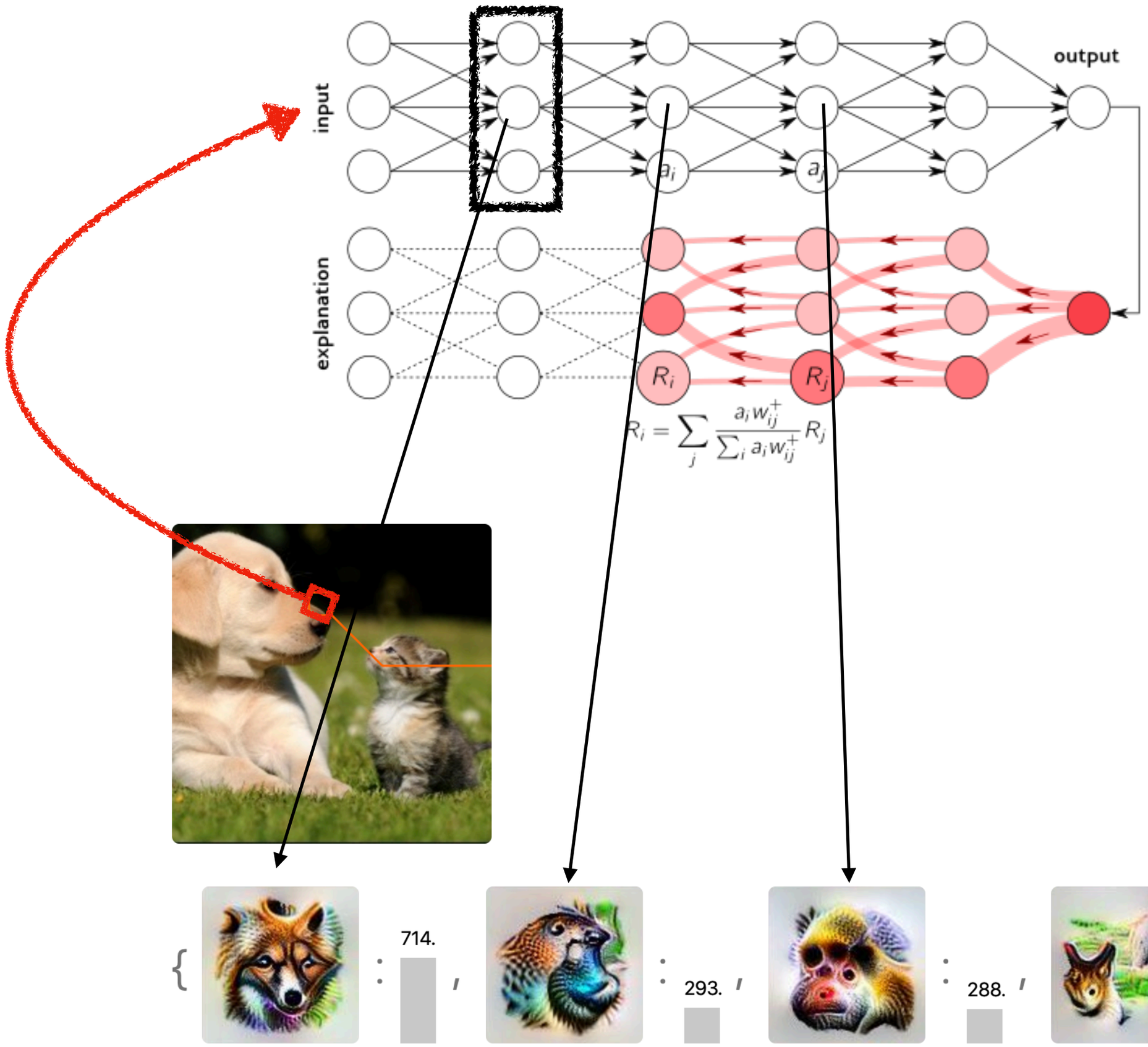
2017 Bojarski et al., Explaining How End-to-End Deep Learning Steers a Self-Driving Car

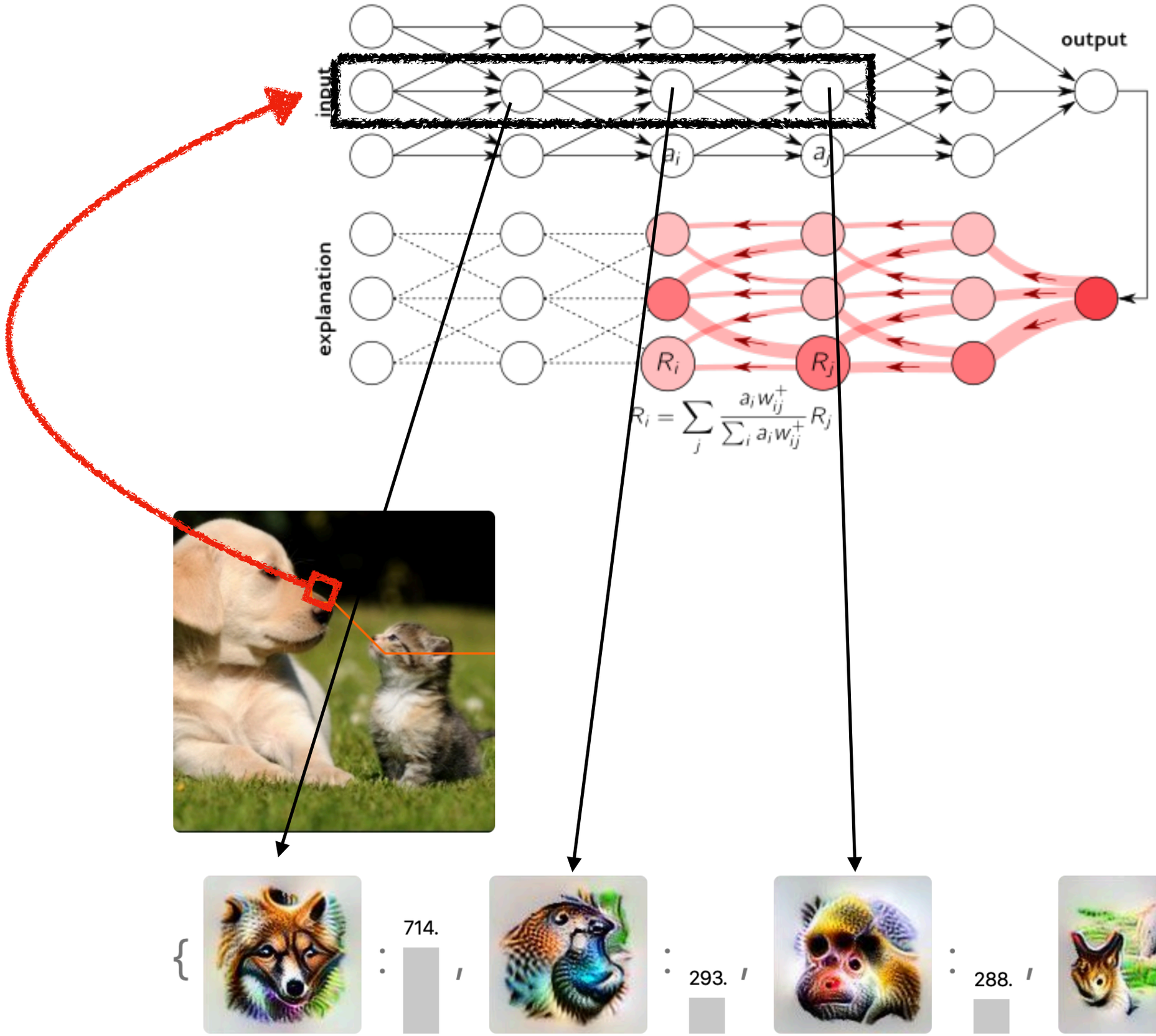


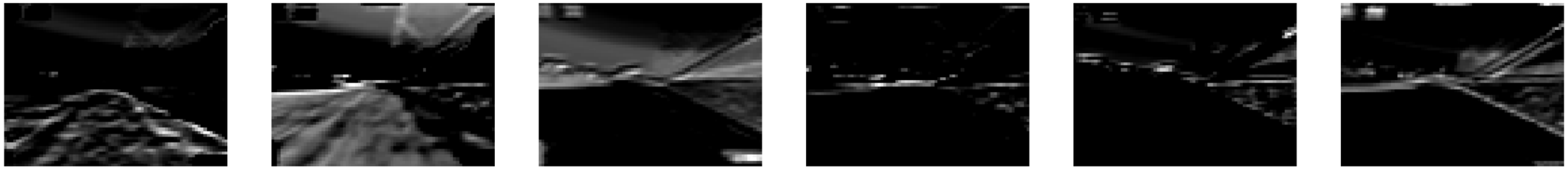
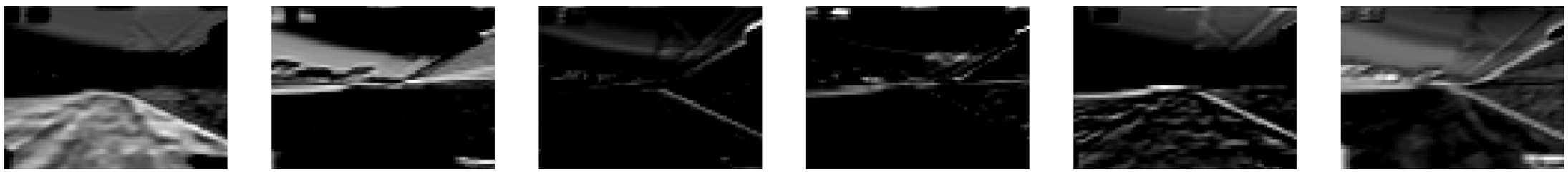
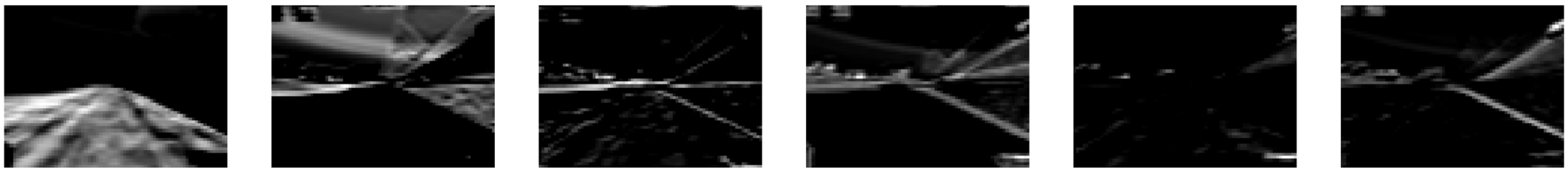
2017 Bojarski et al., Explaining How End-to-End Deep Learning Steers a Self-Driving Car


input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
{
714.
293.
288.
250.
…
}


input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
{
714.
293.
288.
250.
…
}


input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
{ : 714. , : 293. , : 288. , : 250. , … }


input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
{
714.
293.
288.
250.
…
}


What questions can we currently answer?

- Given **one** manually selected input:
    - On **which parts** of the input the model **focusses**? (f.e. *LRP*)
- Given **one** selected output:
    - What different **strategies (clusters) exist** for the focussing on images? (f.e. *SpRAy*)
    - What **kind of template** does it look for? (f.e. *Max Activation*)
- Given a **representative set** of inputs for a **latent factor**:
    - Are there any **geometric properties** of the features (f.e. *de-biasing*)

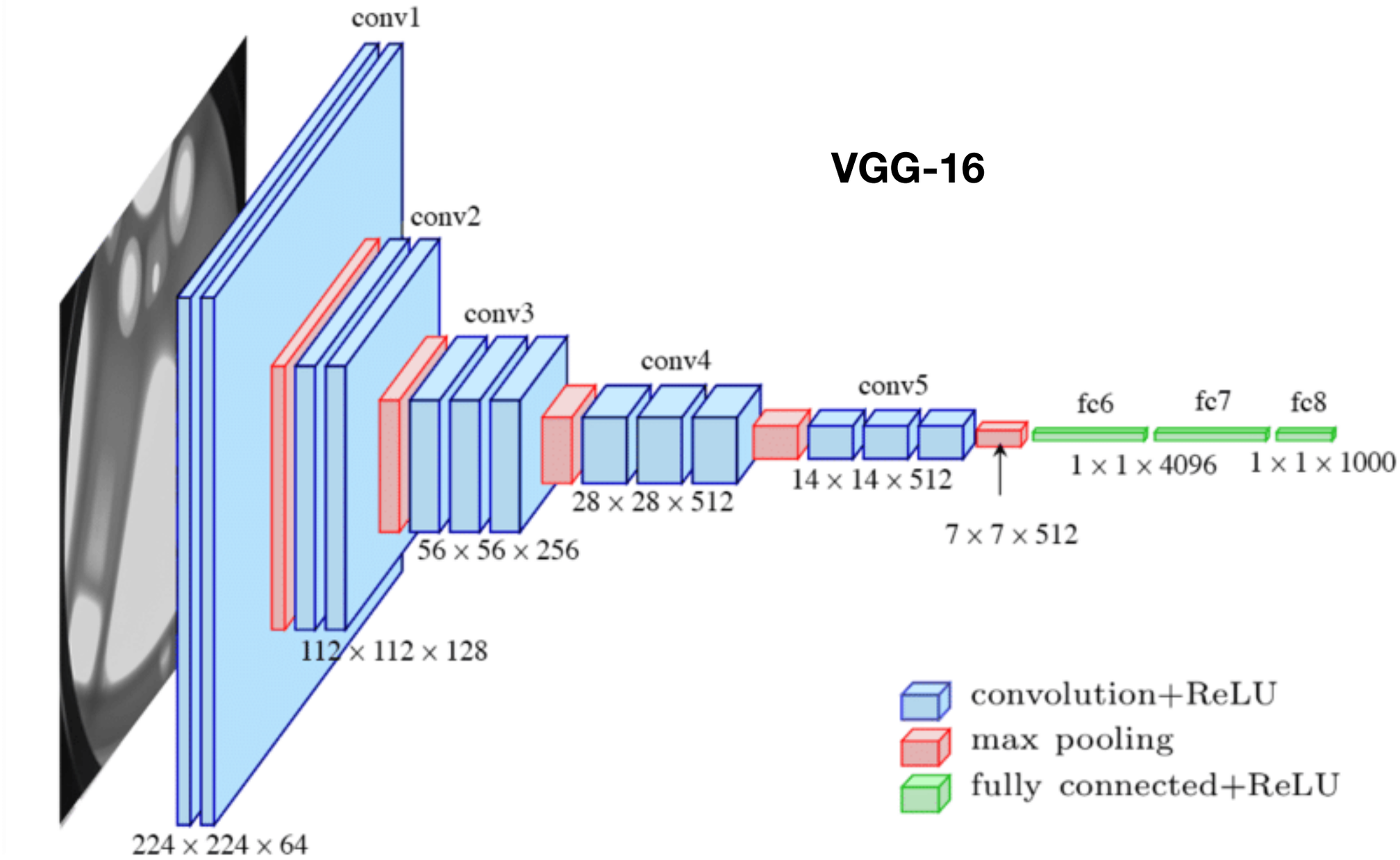
Hands On

https://github.com/grazai/xai-tutorial-march-2020

Side Step: Data

- We use **MNIST** here
    - Super **simple**, super **fast to train**, good for a demo
    - *Better*: For Images, Datasets for Segmentation like **COCO** provide perfect ground truth for the attribution.
    - *simply-clevr-dataset* https://github.com/ahmedmagdiosman/simply-clevr-dataset
    - Diverse Automotive Related Datasets to play around

VGG-16



We use something VGG like

What questions can we currently answer?

- Given **one** manually selected input:
  - On **which parts** of the input the model **focusses**?
    - Attention Mechanisms, LRP, GradCAM, IntegratedGradients, ….
    - https://human-centered.ai/wordpress/wp-content/uploads/2020/03/706.046-AK-explainable-AI-Introduction-MiniProjects-Class-of-2020.pdf for more (Prof. Holzinger)

What questions can we currently answer?

- Given **one** selected output:
  - Are there **clusters** on the parts the model focuses?
  - SpRAy, Sampling, …
  - https://human-centered.ai/wordpress/wp-content/uploads/2020/03/706.046-AK-explainable-AI-Introduction-MiniProjects-Class-of-2020.pdf for more (Prof. Holzinger)

What questions can we currently answer?

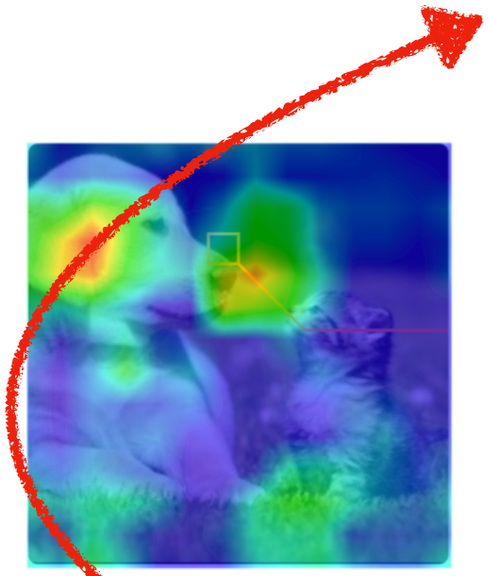
- Given **one** selected output:
- What **kind of template** does it look for?
    - **Max Activation**, Project Lucid, Activation Atlas
    - destill.pup

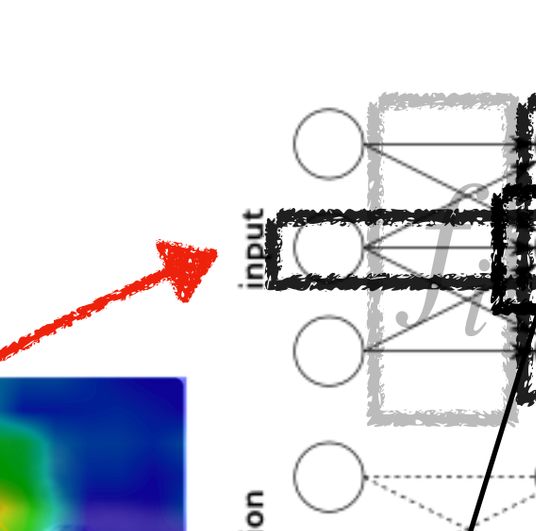
What questions can we currently answer?

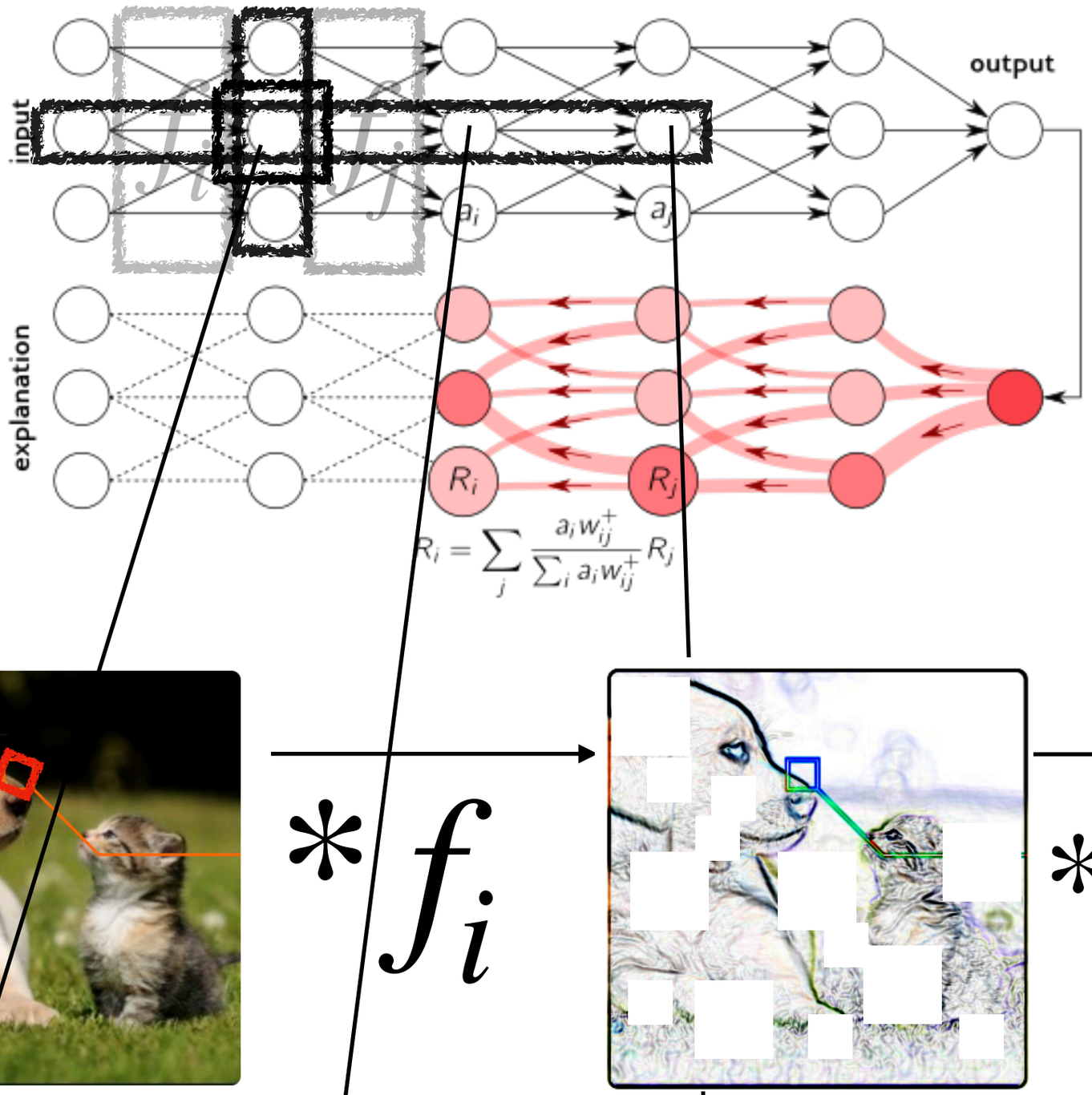
- Given a **representative set** of inputs for a **latent factor**:
  - Are there any **geometric properties** of the features
    - Embeddings and De-Biasing

I did lie to you!

- **Adversarial** Images
- **Sensitivity** instead of importance
- Not the **complete** picture
- Not completely **mature** in case of frameworks
- But already **ok** for the *knowledgeable* and a **great promise**

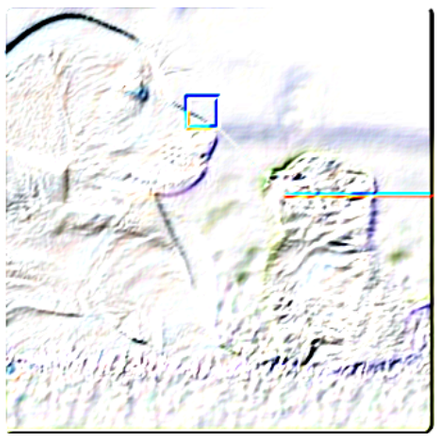






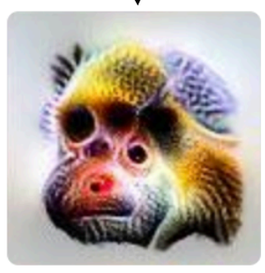
$* f_i$

$* f_j$



: 714. ,

: 293. ,



: 288. ,

: 250. , ... }

Thanks for listening

I hope there was something of value for you?

We can have some Q&A in the teams channel