Explainable AI für Deep Learning: Overview und Tutorial

Jörg Simon

About me

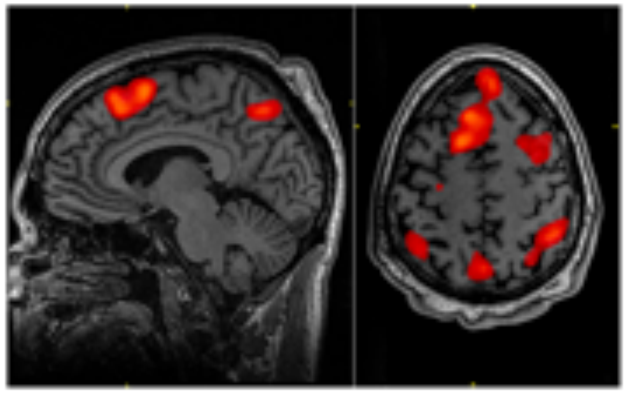
- PhD on using deep learning to detect human factors from biosignals
- Prof. Eduardo Veas and Herbert Danzinger
- Sometimes very sparse data!
- Inspired to use interpretability results to change the training process itself

Agenda

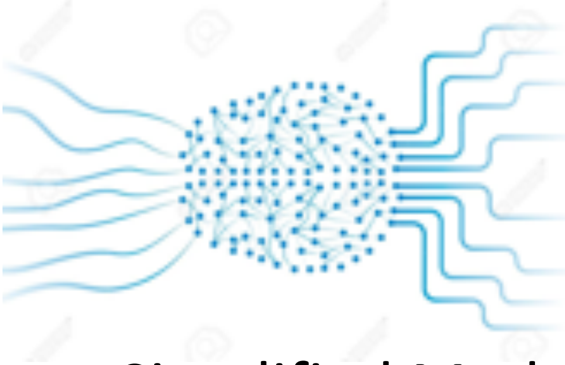
- Definitions and Stuff
- Hands On
- Discussion
- Q&A on Discord

Definitions and Stuff

- Deep learning

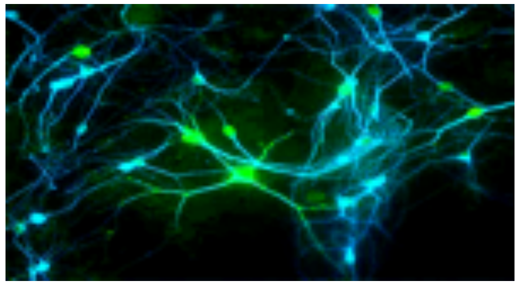


Distributed Representation



Super Simplified Model of Human Brain

Hinton



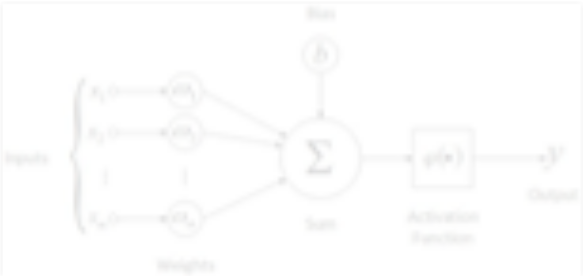
Spiking Frequency = weight

Deep Learning?

CNN

Yann LeCun

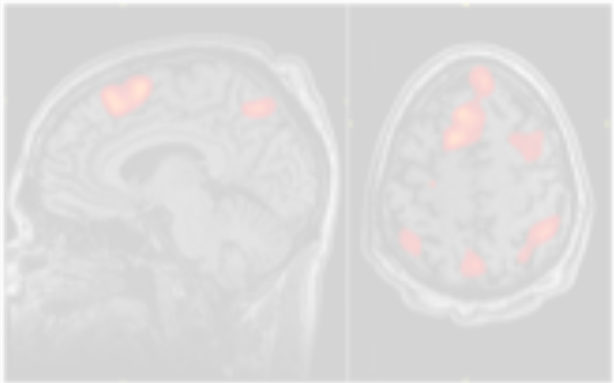
Simple Matrix Multiply + Non Linearity



RNN

Output
Y
RNN
X
Input

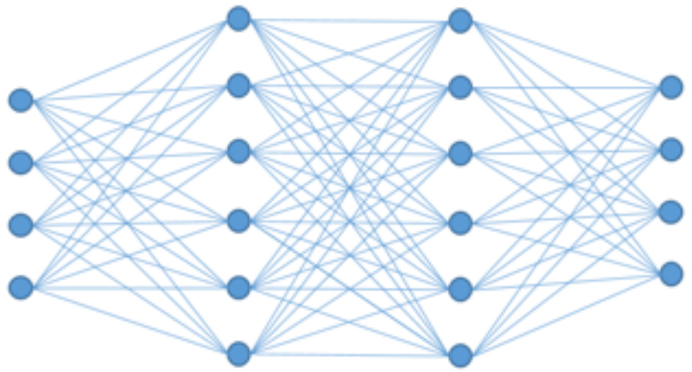
Bengio, Hochreiter, Schmidhuber



Distributed Representation

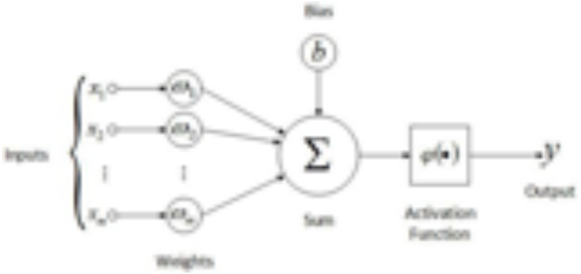
Super Simplified Model of Human Brain

Spiking Frequency : weight



Deep Learning?

Simple Matrix Multiply + Non Linearity



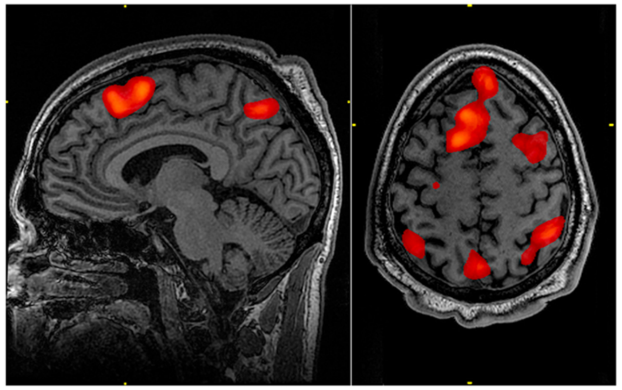
Hinton

CNN

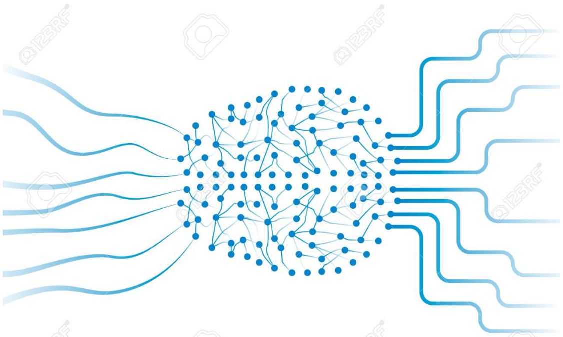
Yann LeCun

RNN

Bengio, Hochreiter, Schmidhuber

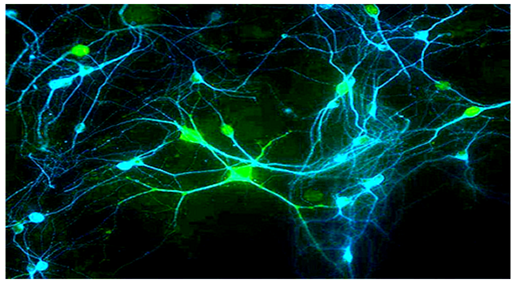


Distributed Representation

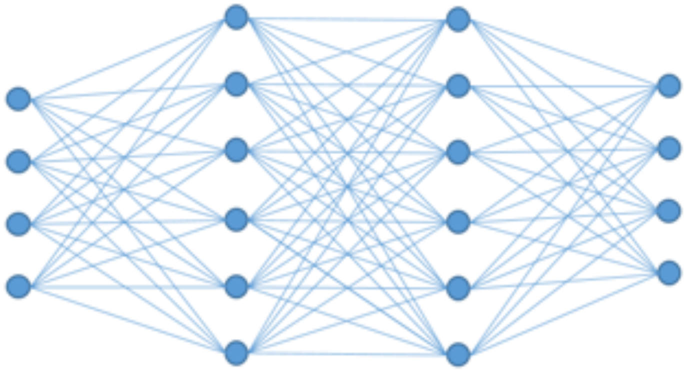


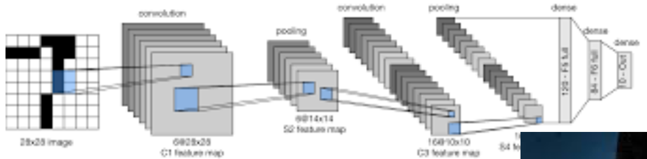
Super Simplified Model of Human Brain

Hinton

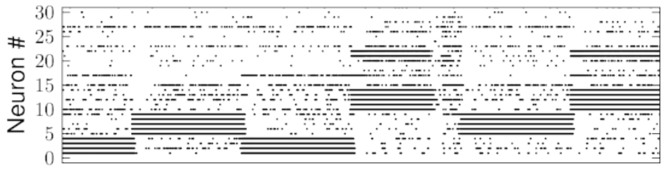


Spiking Frequency = weight





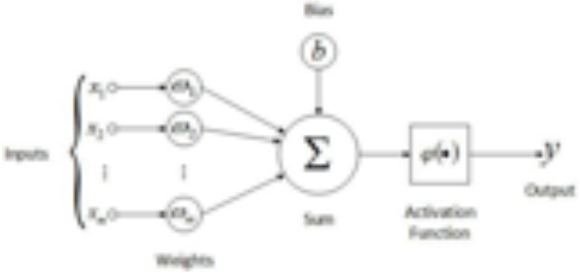
CNN



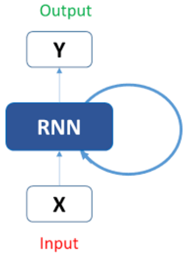
Deep Learning?

Yann LeCun

Simple Matrix Multiply + Non Linearity



RNN



Bengio, Hochreiter, Schmidhuber

Definitions and Stuff

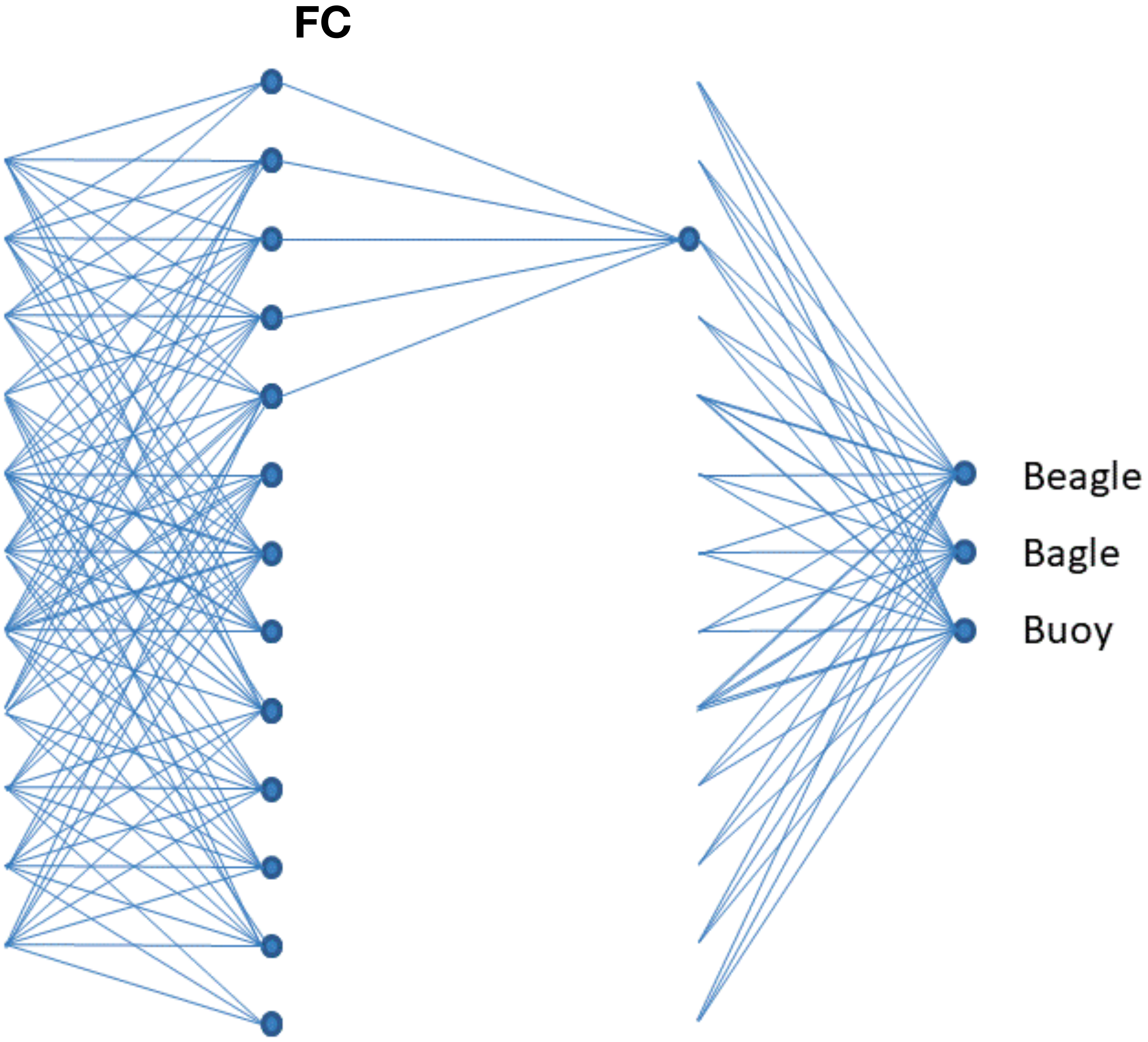
- Deep learning
- Architectures

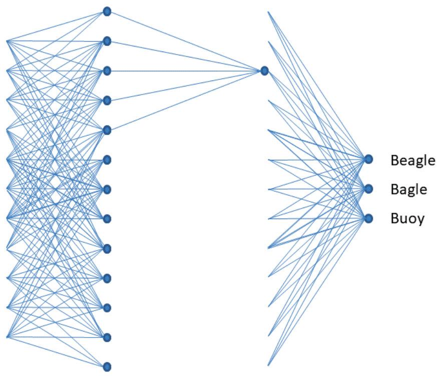
Three main classes of DL architectures

Fully Connected / Feed Forward

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

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



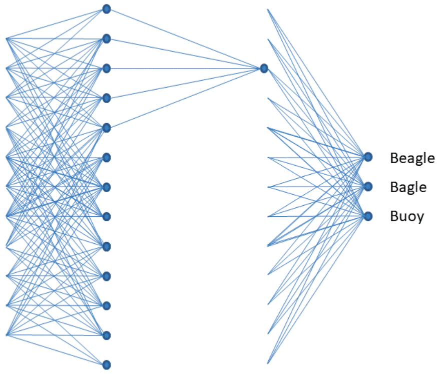


$$Z^i = W^i X + b^i 1$$

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

Fully Connected / Feed Forward

FC



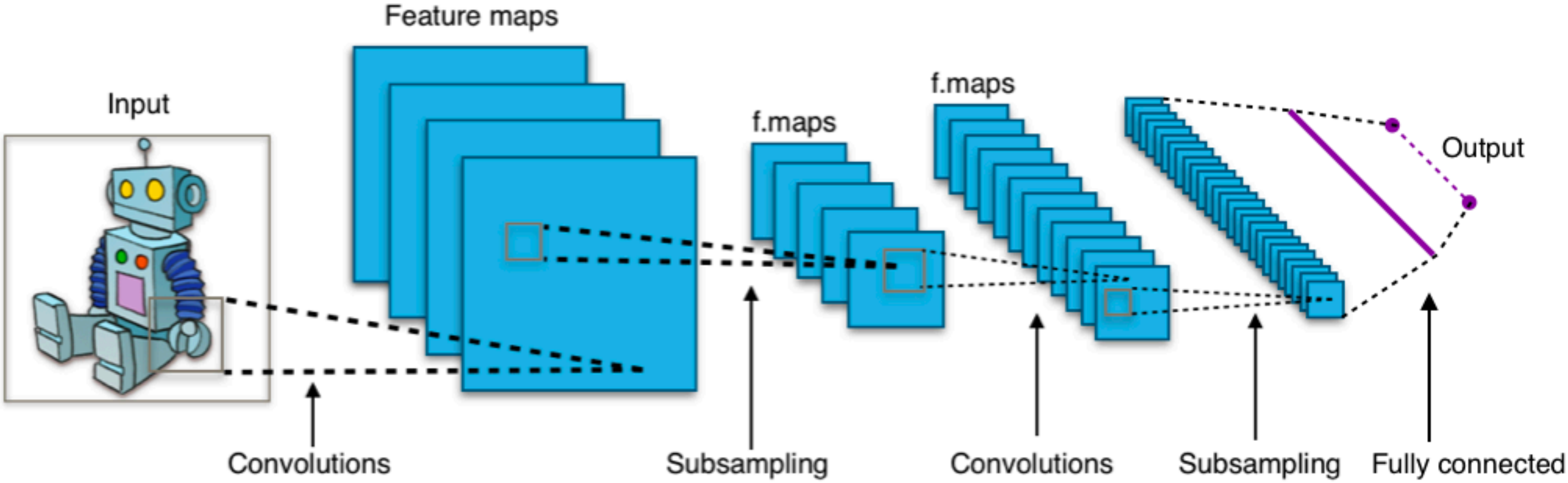
$$Z^i = W^i X + b^i 1$$

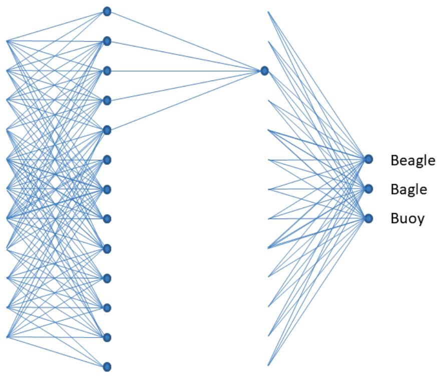
$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

FC

CNN
Convolutional Neural Networks





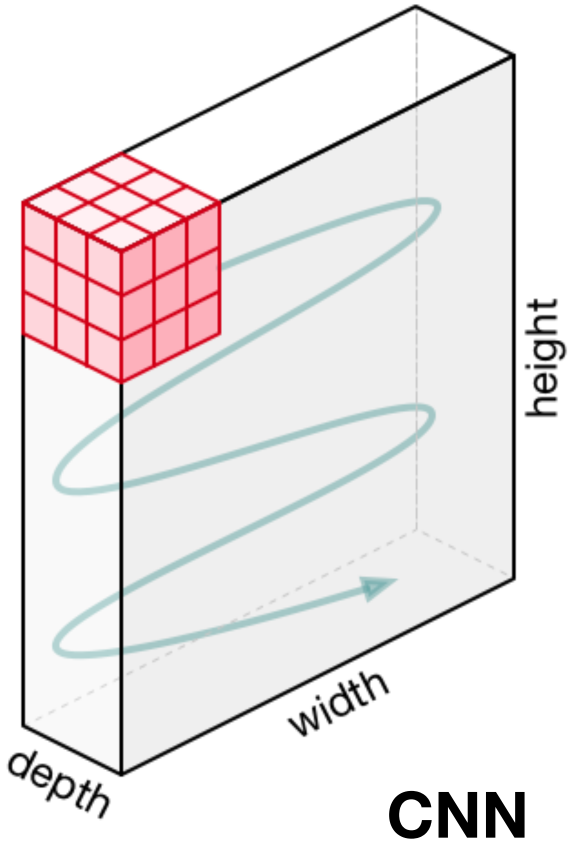
$$Z^i = W^iX + b^i1$$

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

Fully Connected / Feed Forward

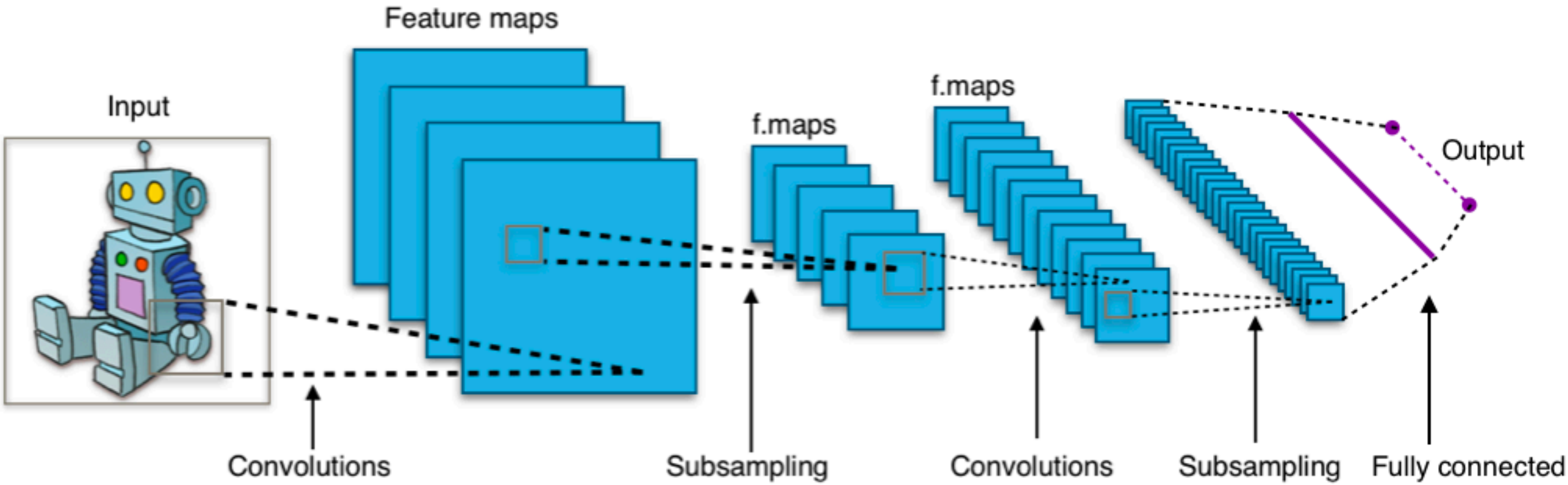
FC

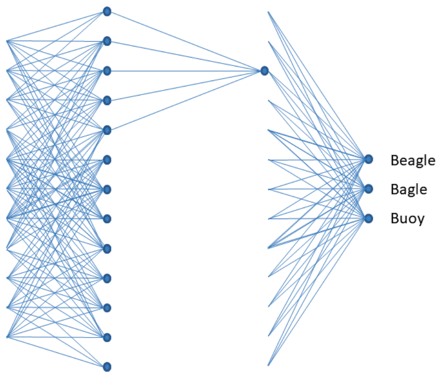
Kernels



CNN

Convolutional Neural Networks



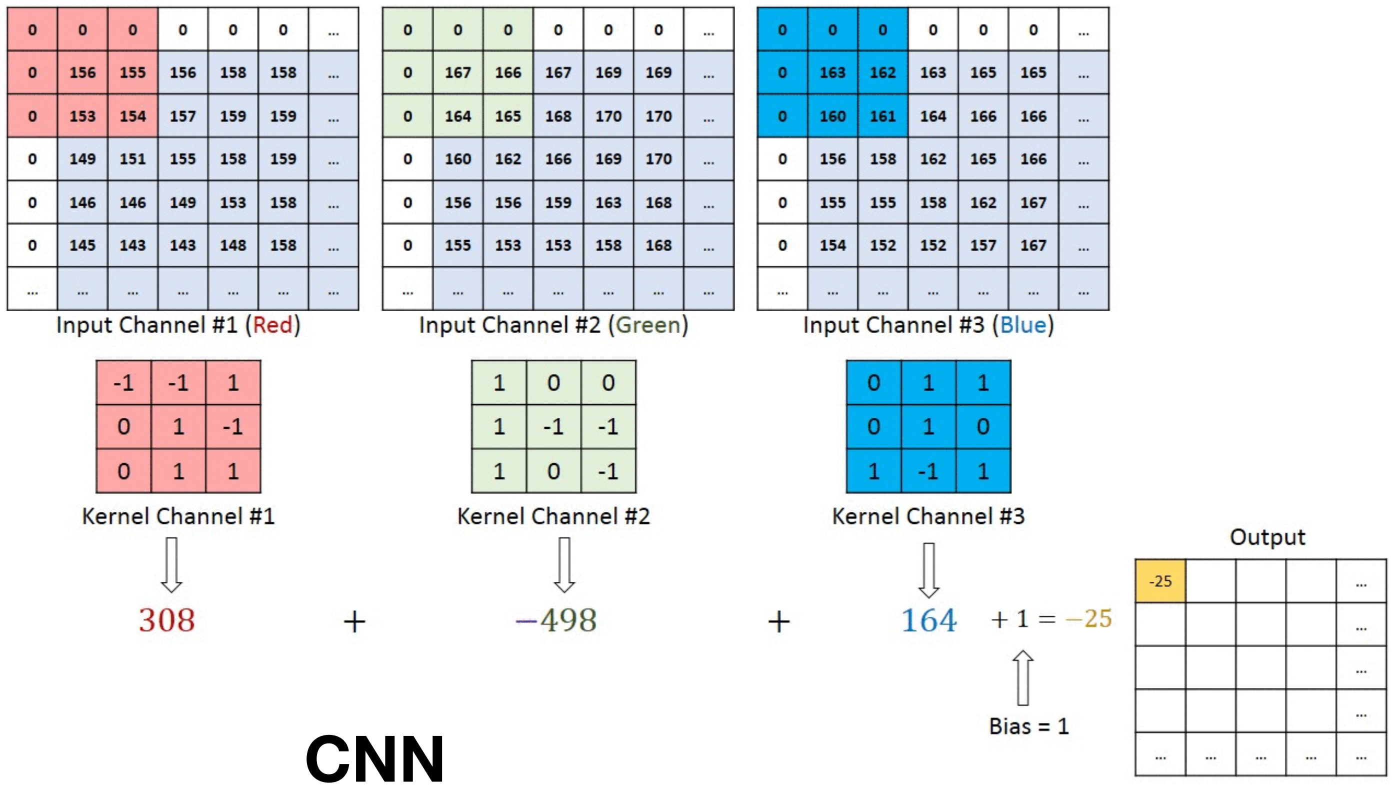


$$Z^i = W^i X + b^i 1$$

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

Fully Connected / Feed Forward

FC



0	0	0	0	0	0	...
0	156	155	156	158	158	...
0	153	154	157	159	159	...
0	149	151	155	158	159	...
0	146	146	149	153	158	...
0	145	143	143	148	158	...
...	...	...	...	...	...	...

Input Channel #1 (Red)

0	0	0	0	0	0	...
0	167	166	167	169	169	...
0	164	165	168	170	170	...
0	160	162	166	169	170	...
0	156	156	159	163	168	...
0	155	153	153	158	168	...
...	...	...	...	...	...	...

Input Channel #2 (Green)

0	0	0	0	0	0	...
0	163	162	163	165	165	...
0	160	161	164	166	166	...
0	156	158	162	165	166	...
0	155	155	158	162	167	...
0	154	152	152	157	167	...
...	...	...	...	...	...	...

Input Channel #3 (Blue)

-1	-1	1
0	1	-1
0	1	1

Kernel Channel #1

1	0	0
1	-1	-1
1	0	-1

Kernel Channel #2

0	1	1
0	1	0
1	-1	1

Kernel Channel #3

$$308 \quad + \quad -498 \quad + \quad 164 \quad + 1 = -25$$

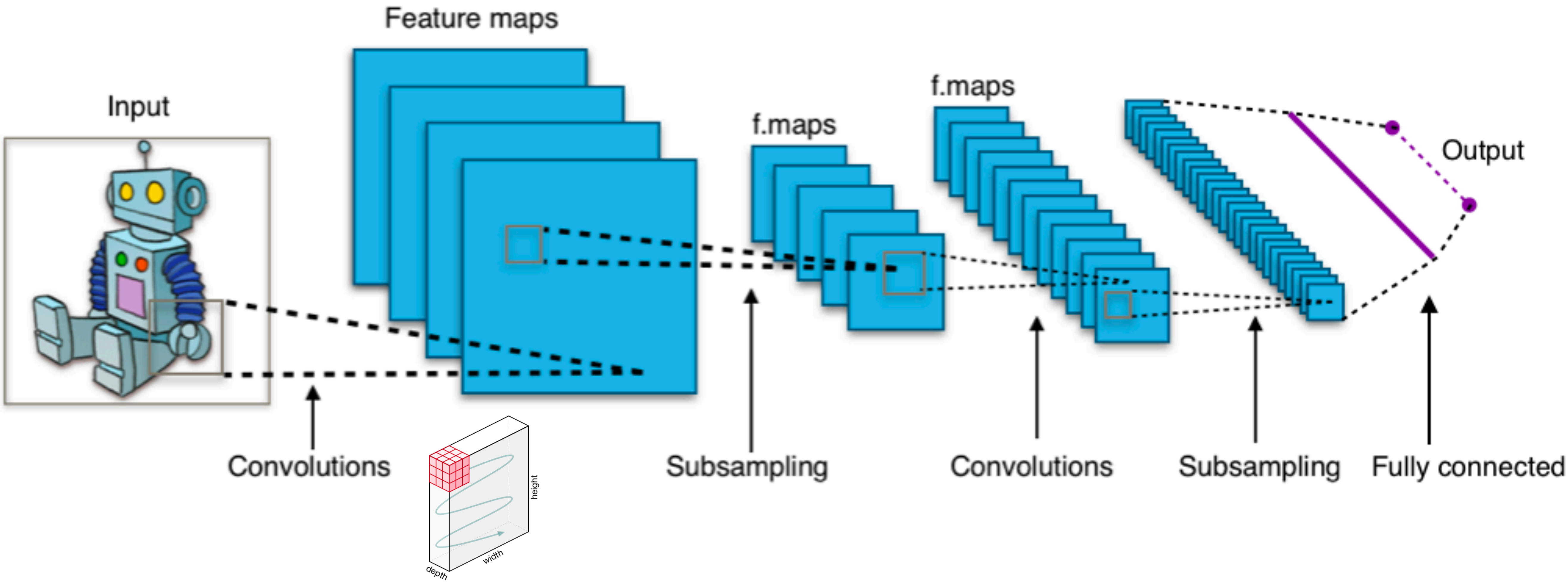
Bias = 1

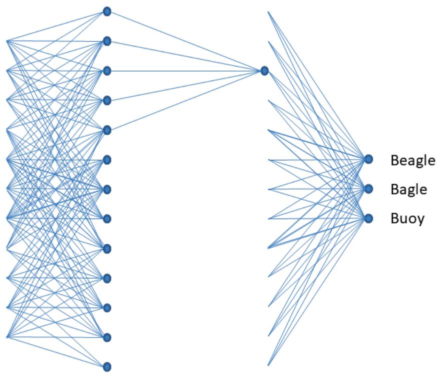
Output

-25				...
				...
				...
				...
...	...	...	...	...

CNN

Convolutional Neural Networks



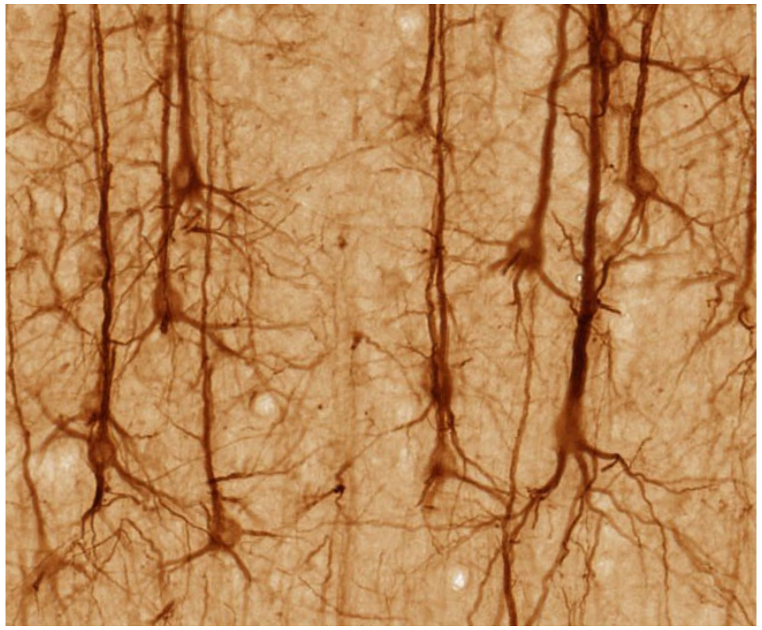


$$Z^i = W^i X + b^i 1$$

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

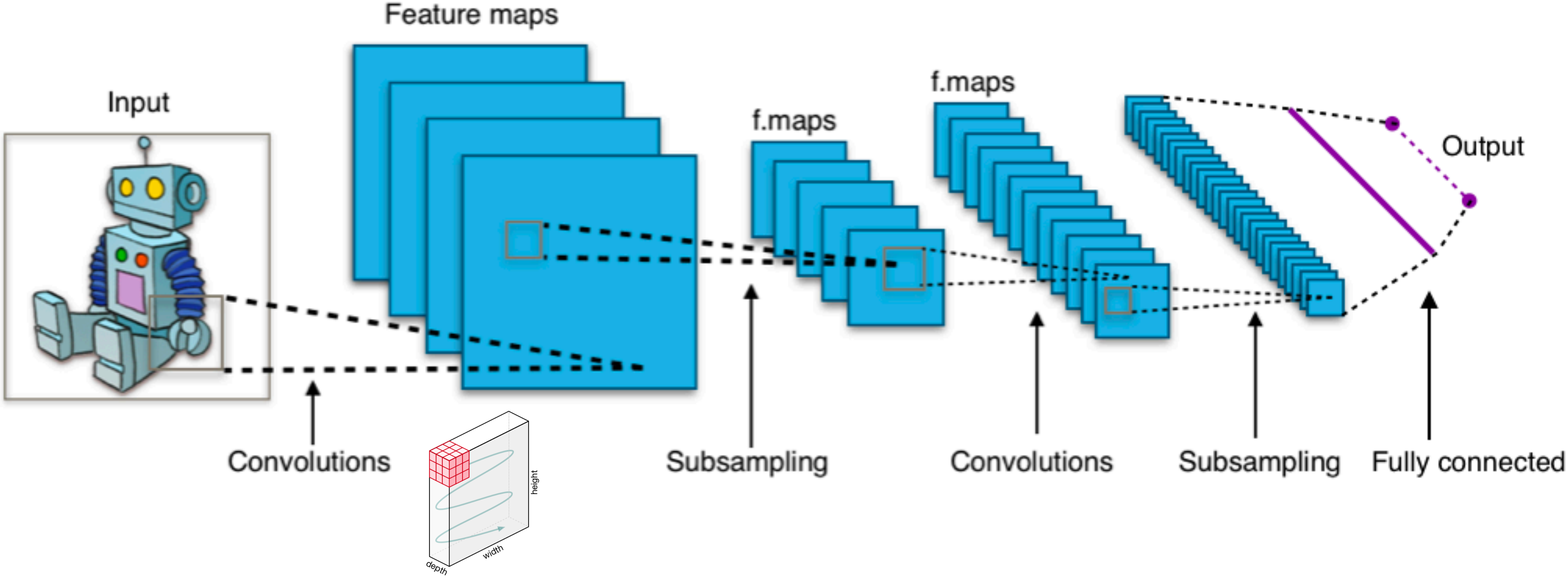
Fully Connected / Feed Forward

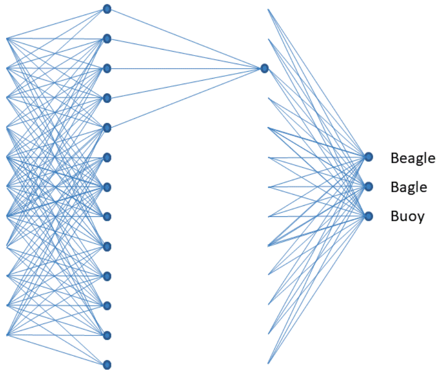
FC



CNN

Convolutional Neural Networks

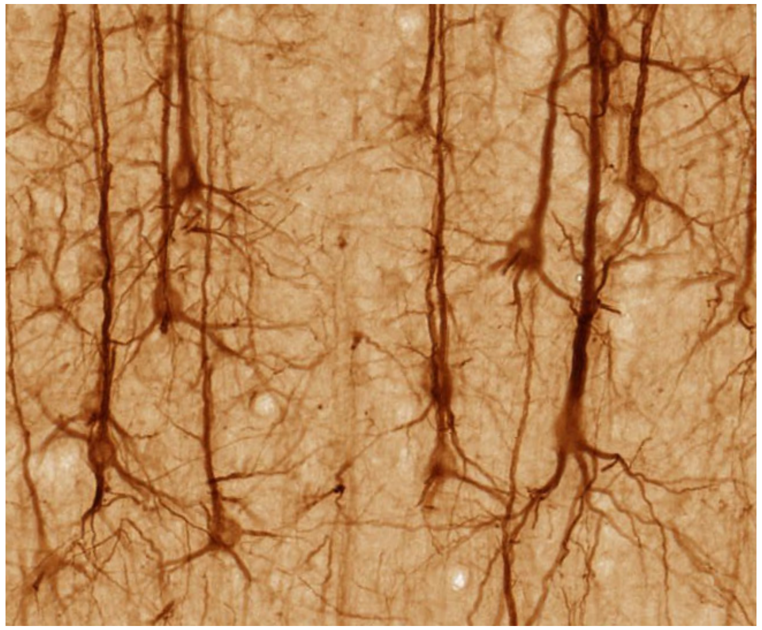


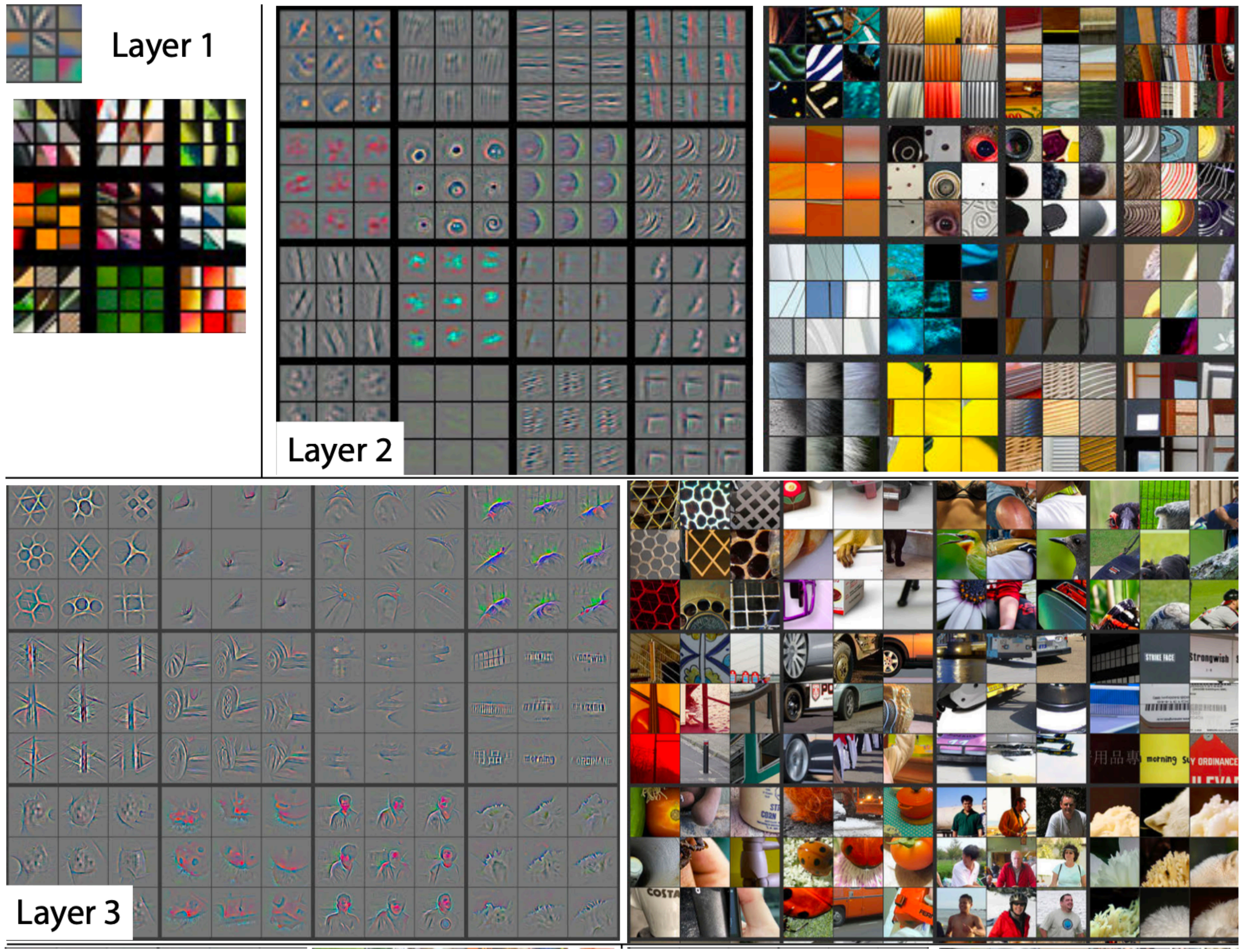


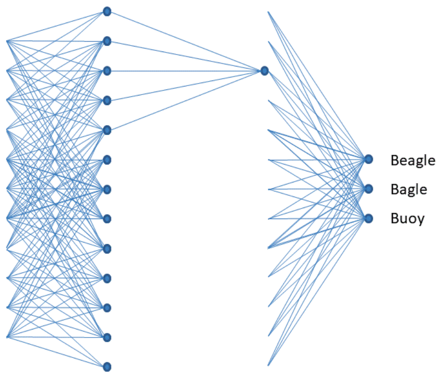
$$Z^i = W^i X + b^i 1$$
$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

FC







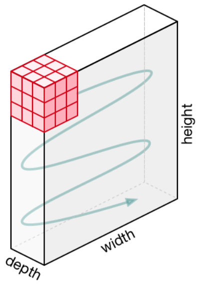
$$Z^i = W^i X + b^i 1$$

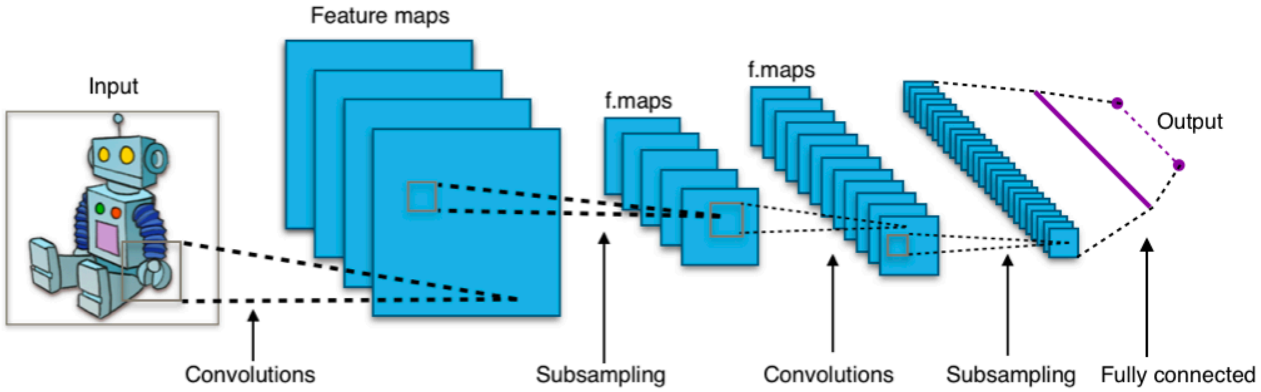
$$A^i = \textbf{RELU}\left(Z^i\right)$$

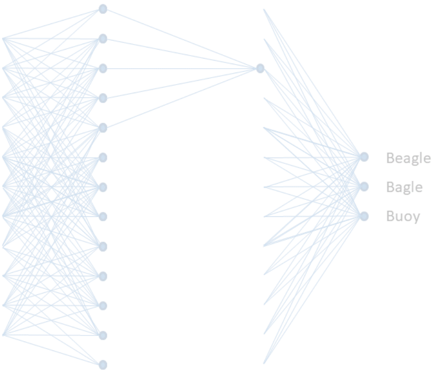
Fully Connected / Feed Forward

FC

CNN
Convolutional Neural Networks







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

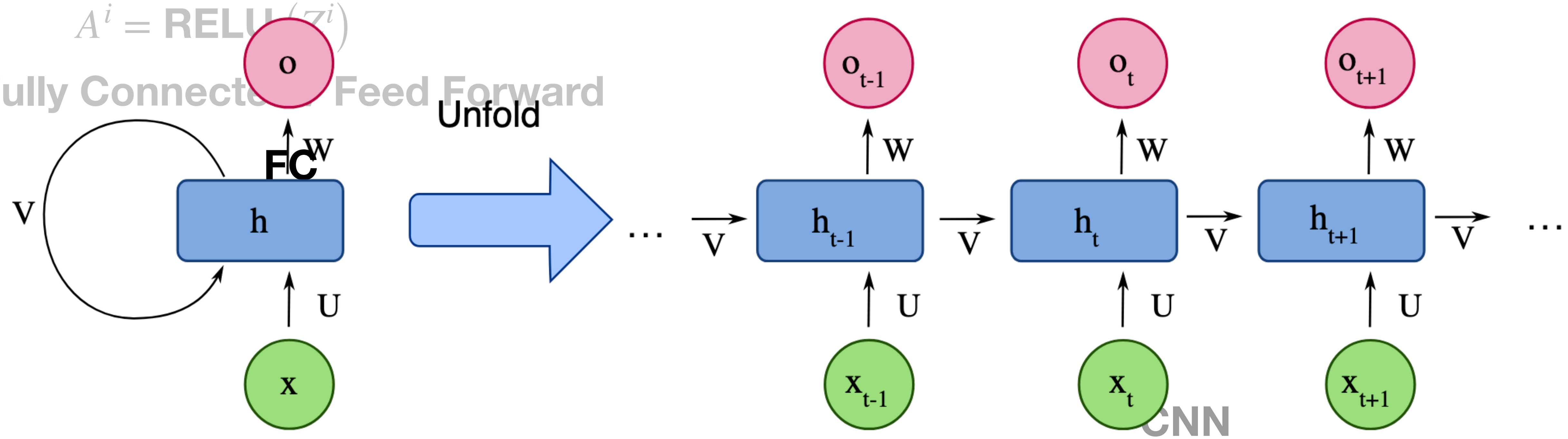
$$A^i = \text{RELU}(Z^i)$$

Fully Connected Feed Forward

FC

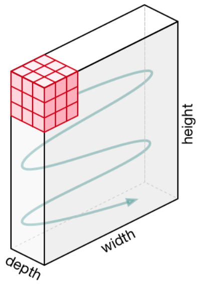
RNN

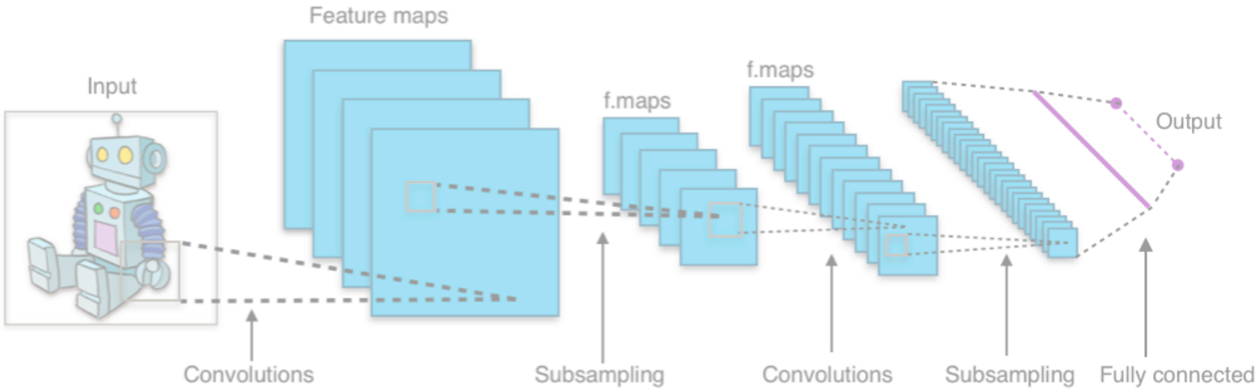
Recurrent Neural Network

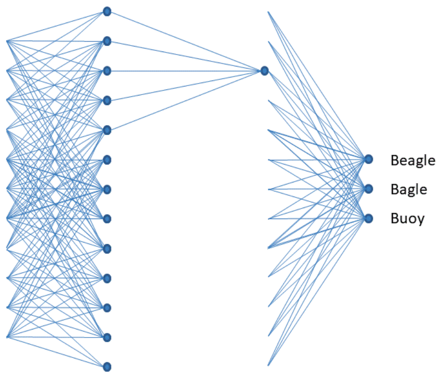


CNN

Convolutional Neural Networks





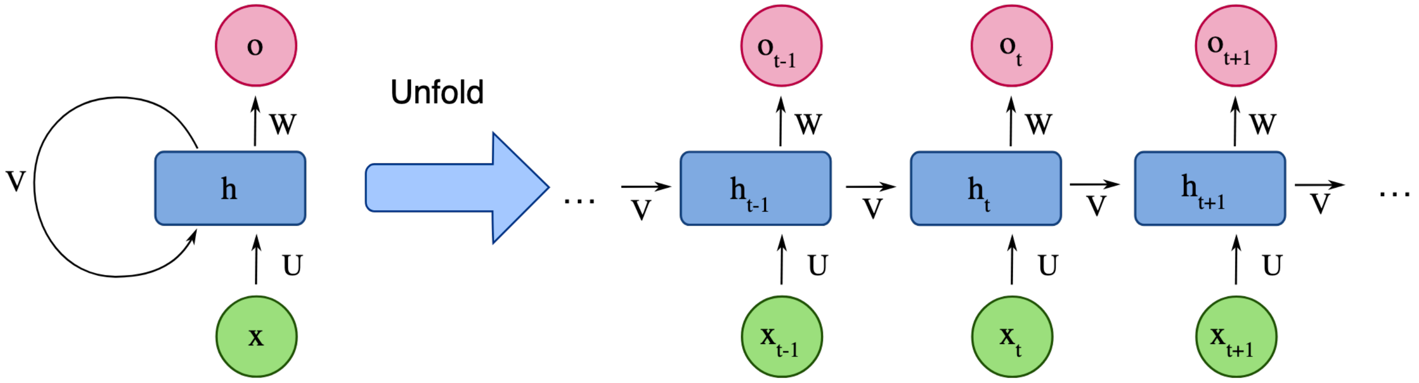


$$Z^i = W^i X + b^i 1$$
$$A^i = \textbf{RELU}\left(Z^i\right)$$

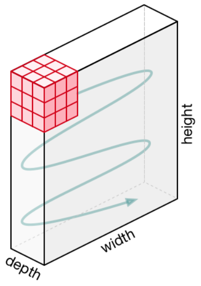
Fully Connected / Feed Forward

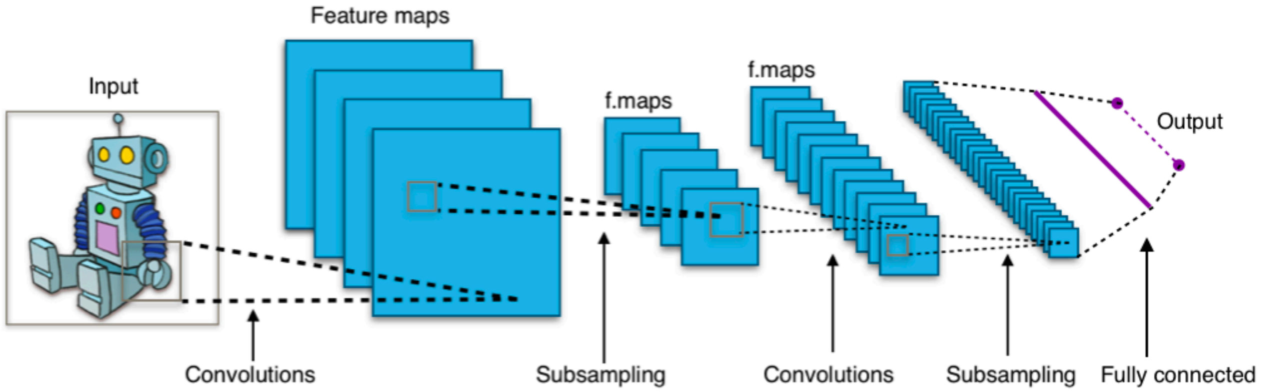
FC

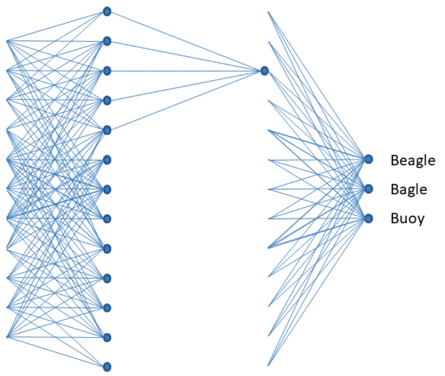
RNN
Recurrent Neural Network



CNN
Convolutional Neural Networks







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

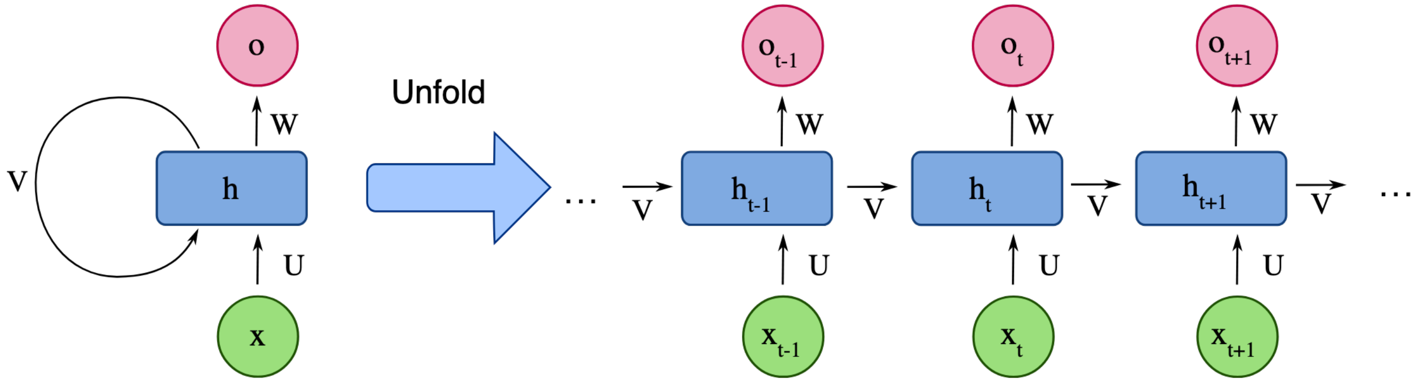
$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

FC

RNN

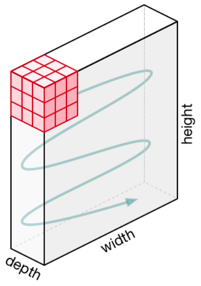
Recurrent Neural Network

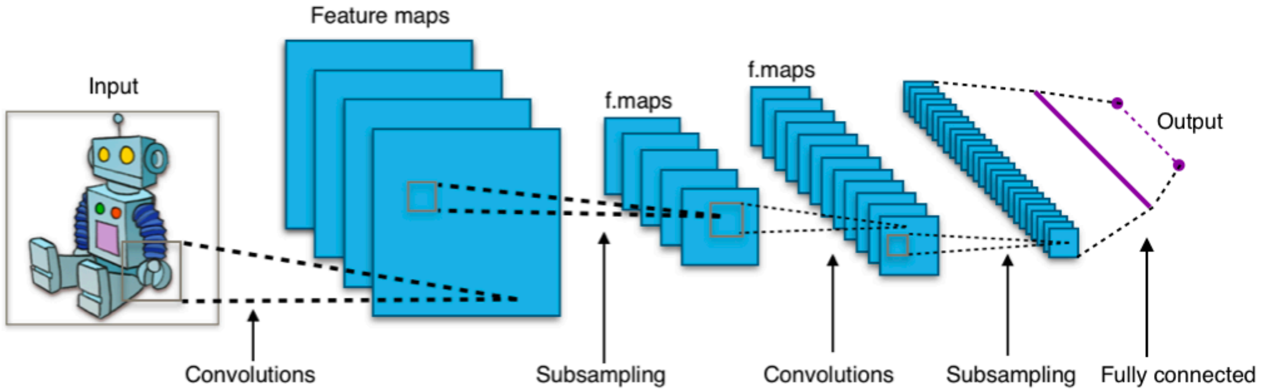


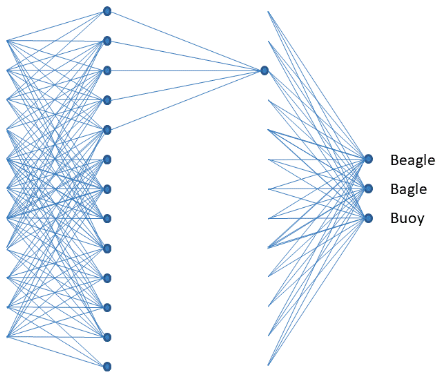
GANs,
Auto Encoders,
ODE Networks,
Invertible Flow Networks,
.....

CNN

Convolutional Neural Networks







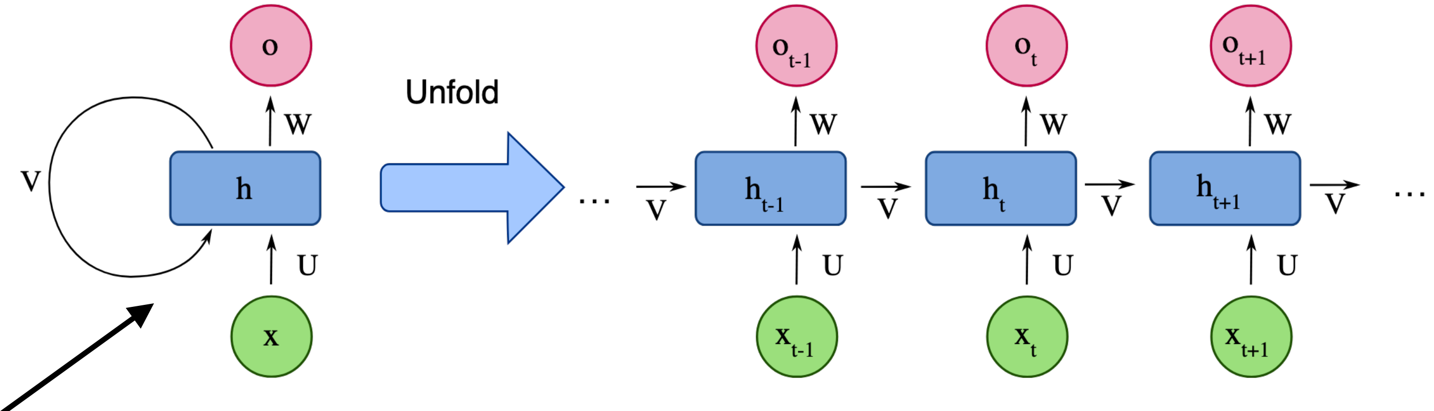
$$Z^i = W^i X + b^i 1$$

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

Fully Connected / Feed Forward

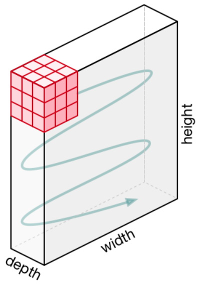
FC

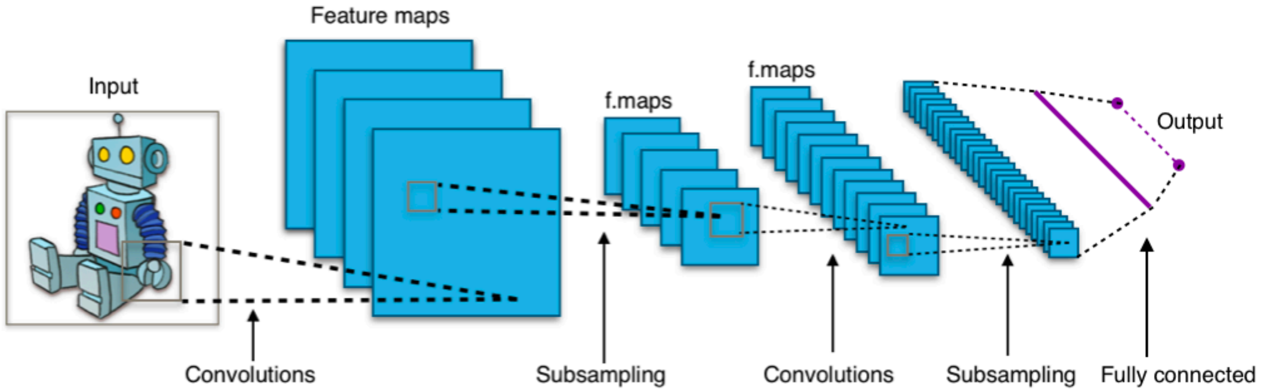
RNN
Recurrent Neural Network

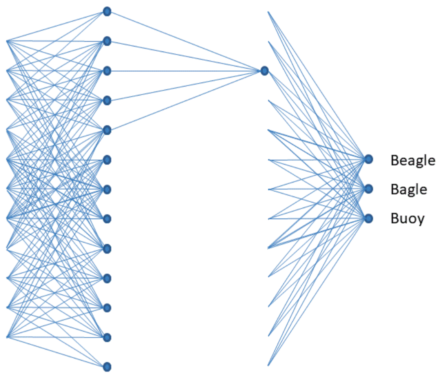


GANs,
Auto Encoders,
ODE Networks,
Invertible Flow Networks,
.....

CNN
Convolutional Neural Networks







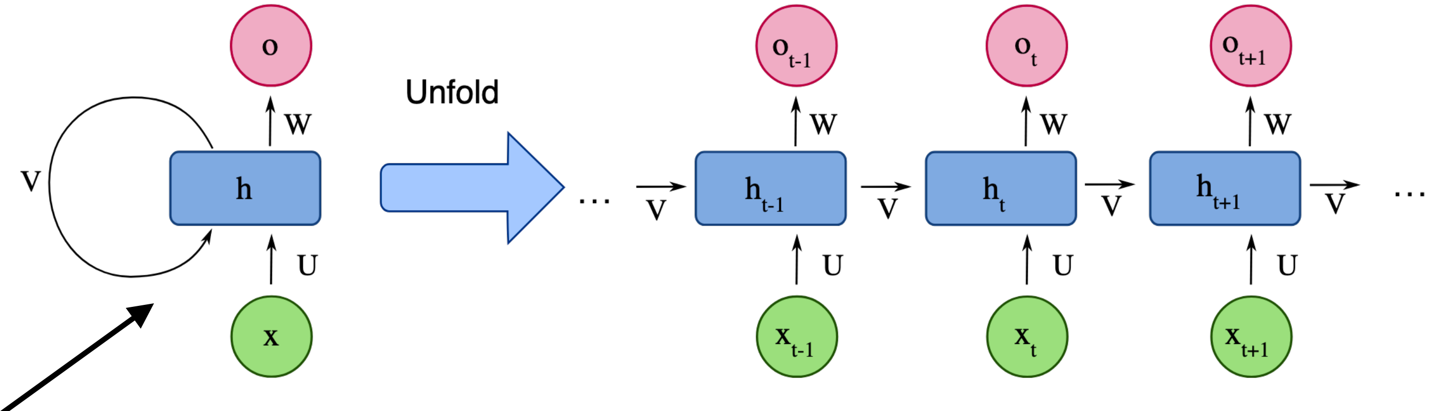
$$Z^i = W^i X + b^i 1$$

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

Fully Connected / Feed Forward

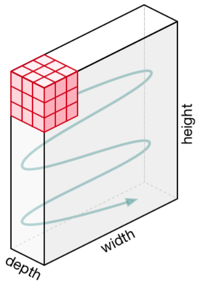
FC

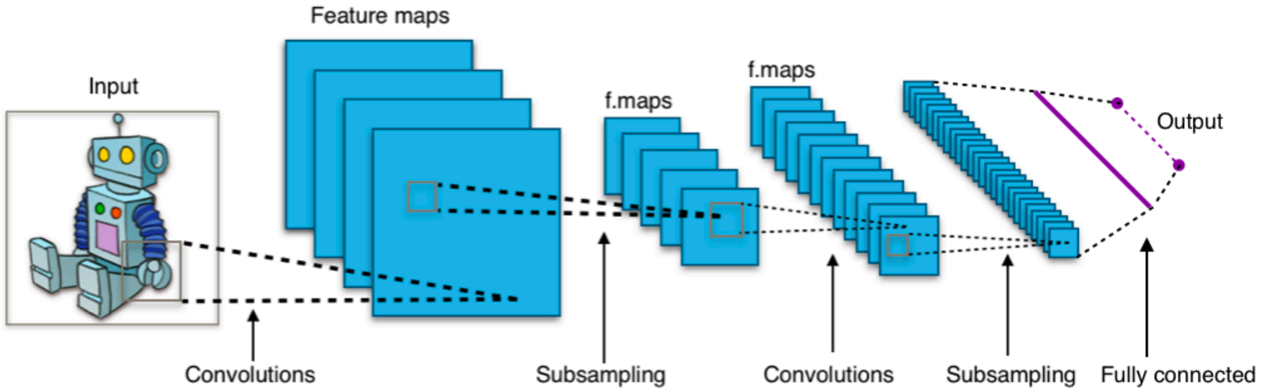
RNN
Recurrent Neural Network



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

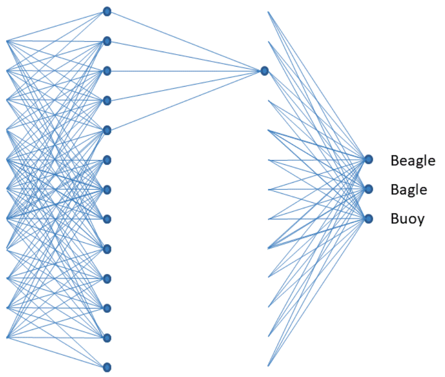
CNN
Convolutional Neural Networks





All kind of domains: medical imaging, autonomous driving, emotion recognition, recommenders, natural language processing

Supervised,
Unsupervised,
Self-Supervised,
Reinforcement Learning



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

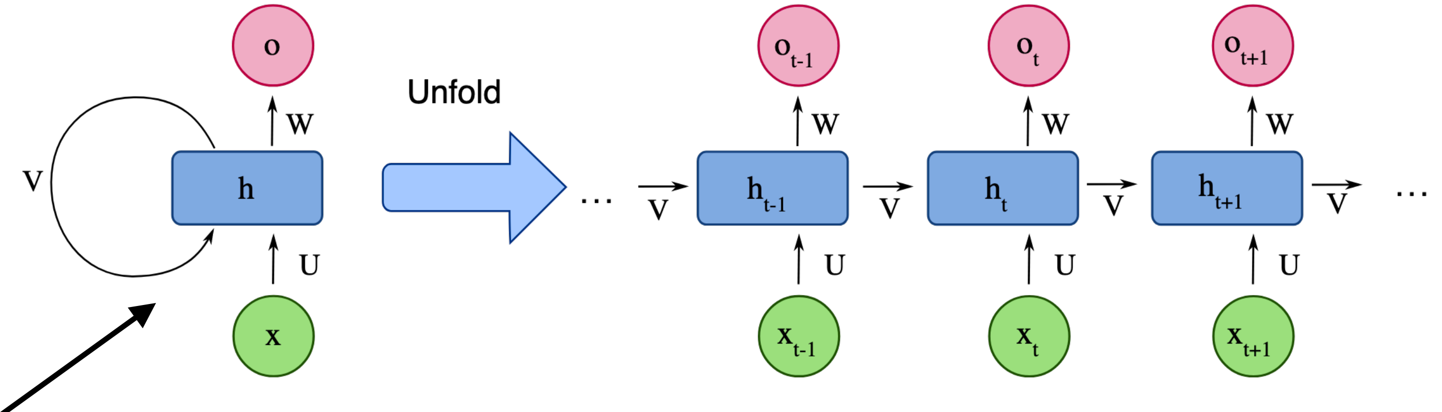
$$A^i = \textbf{RELU}\left(Z^i\right)$$

Fully Connected / Feed Forward

FC

RNN

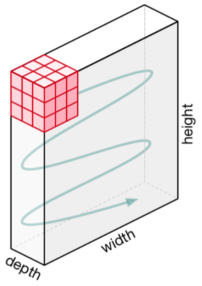
Recurrent Neural Network

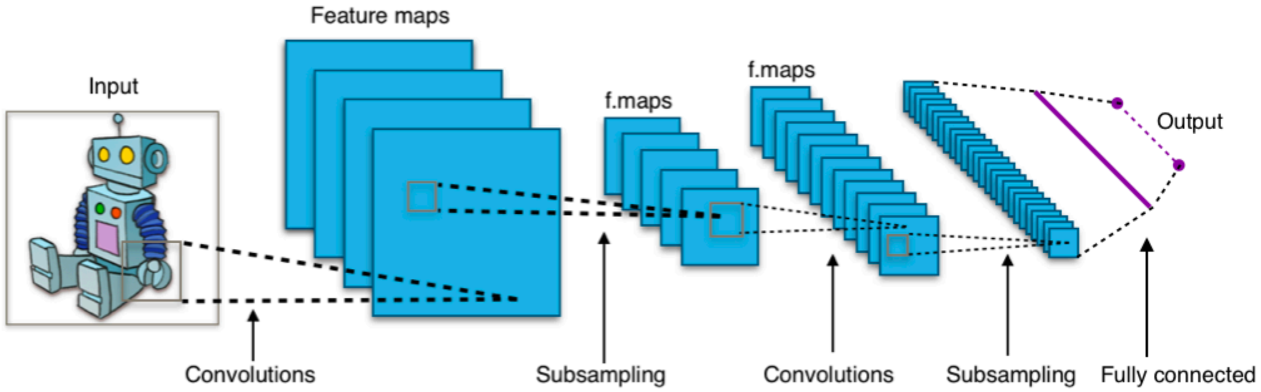


GANs,
Auto Encoders,
ODE Networks,
Invertible Flow Networks,
…..

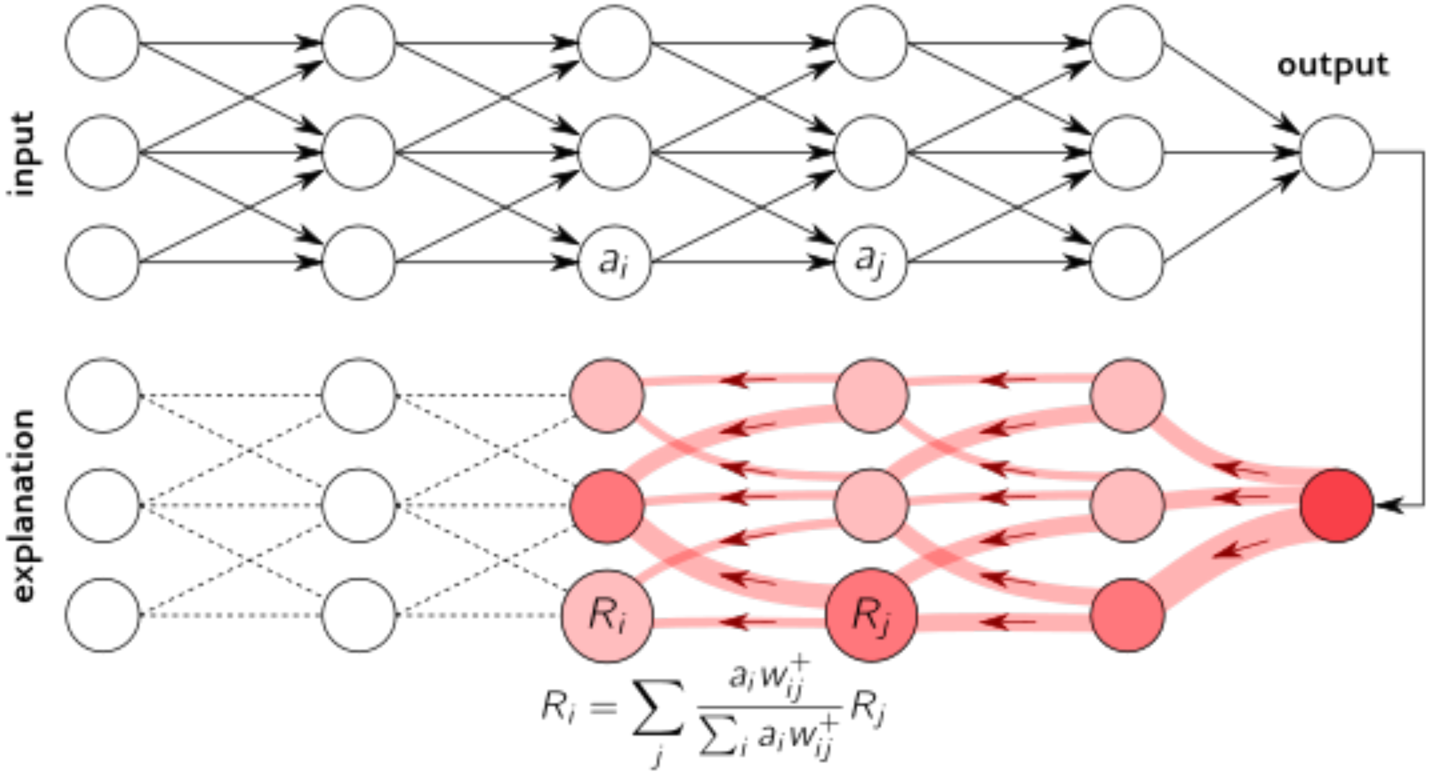
CNN

Convolutional Neural Networks

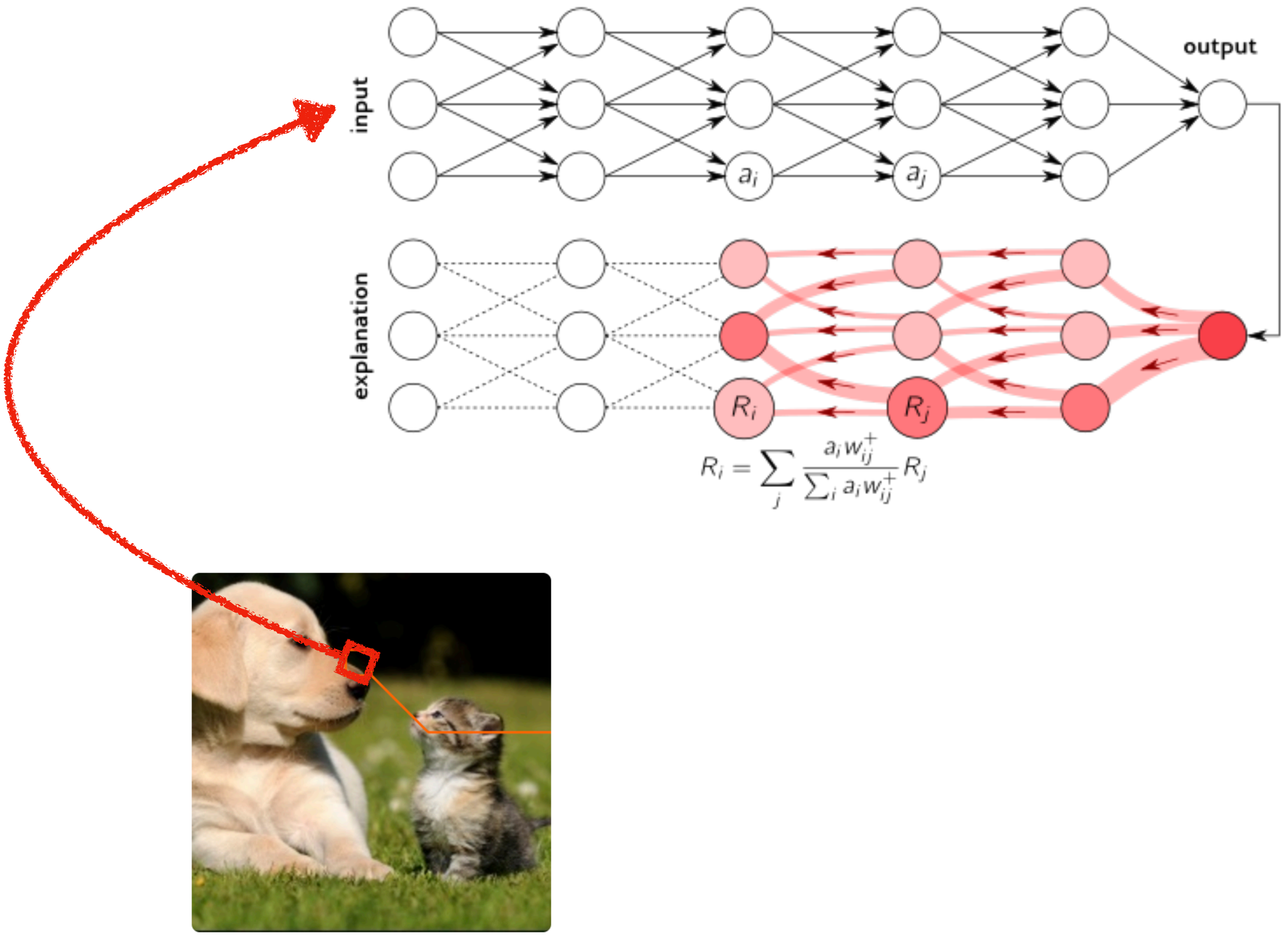


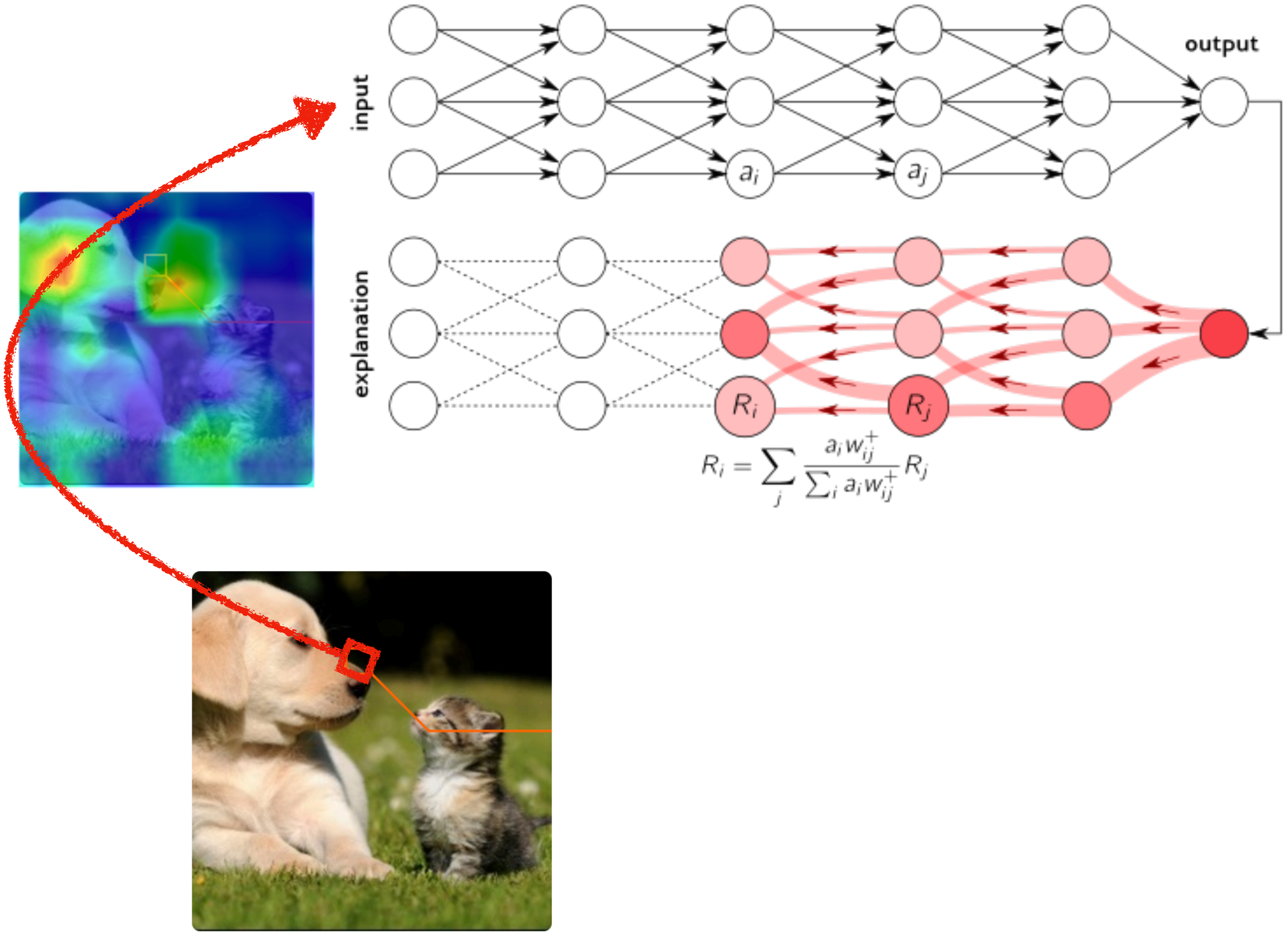


All kind of domains: medical imaging, autonomous driving, emotion recognition, recommenders, natural language processing

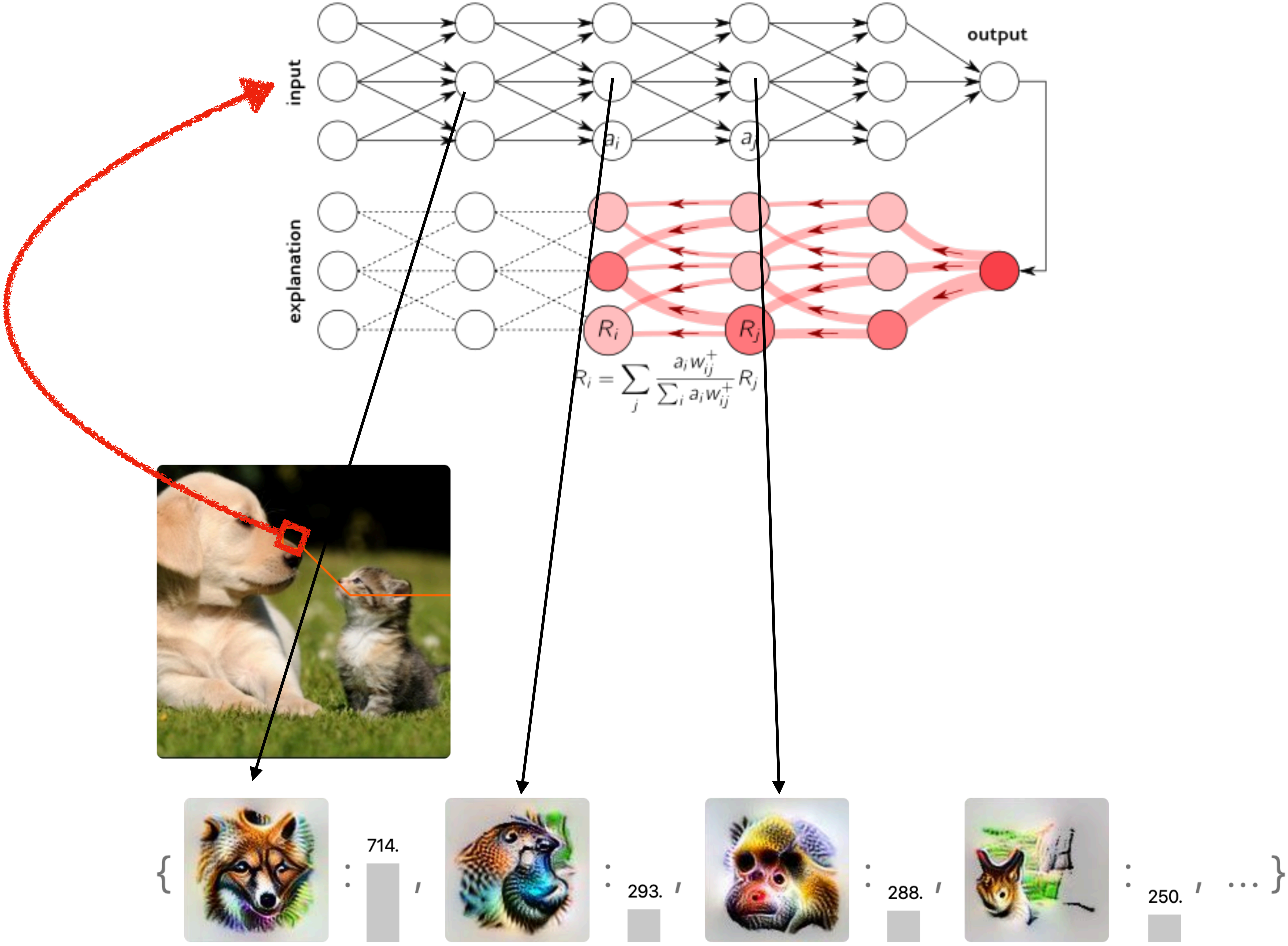


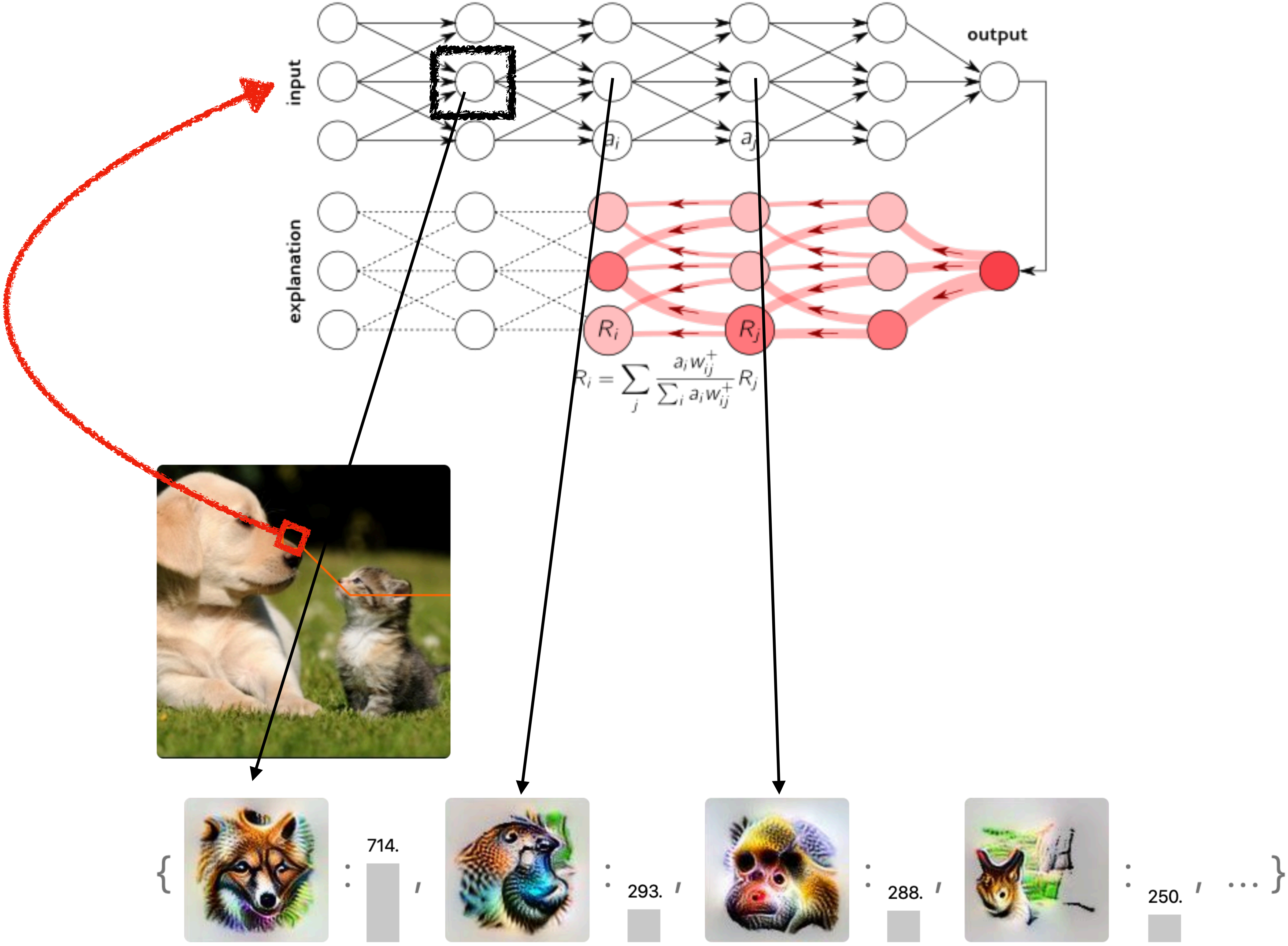
$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

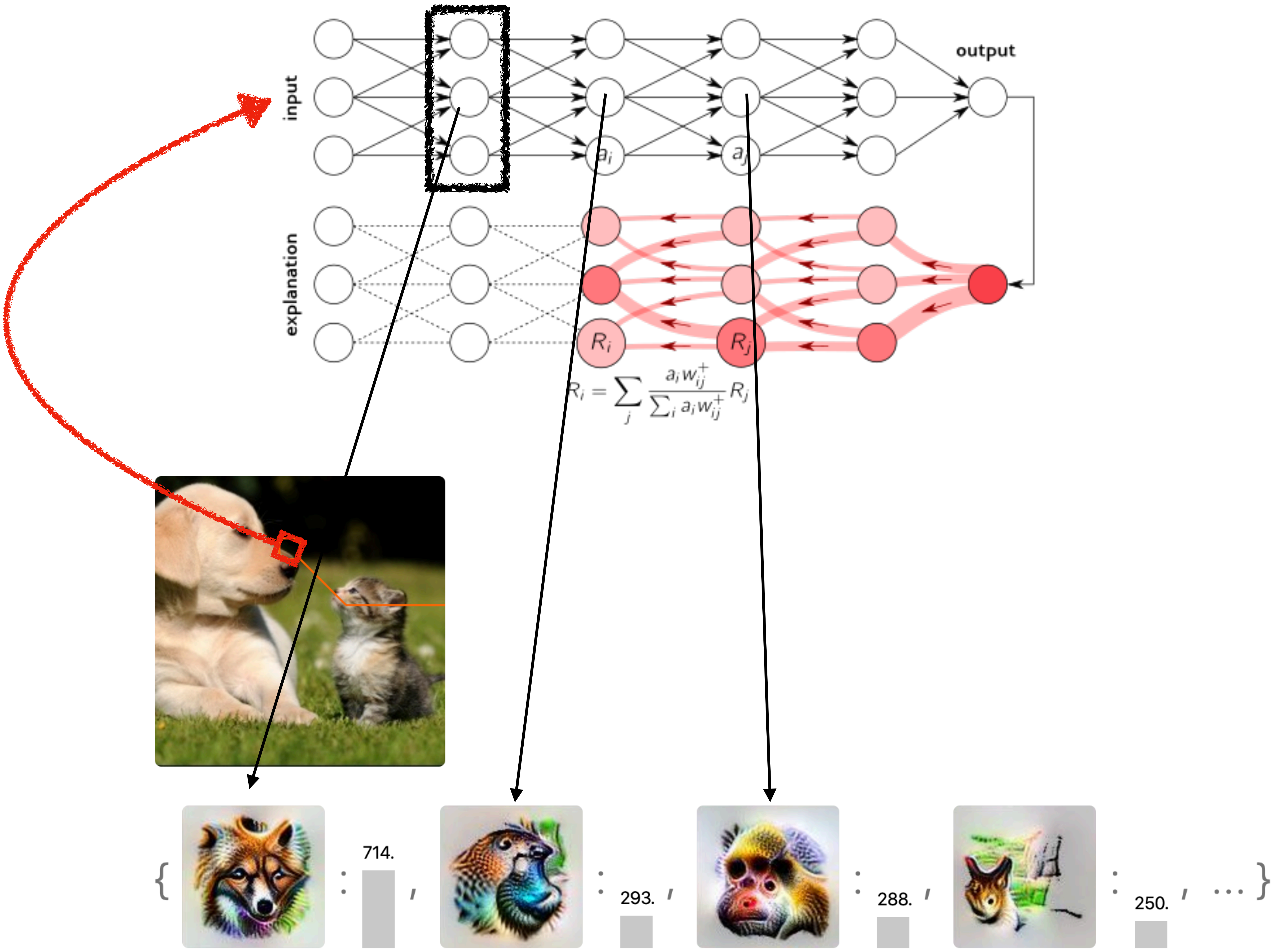

output
input
a_i
a_j
explanation
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$

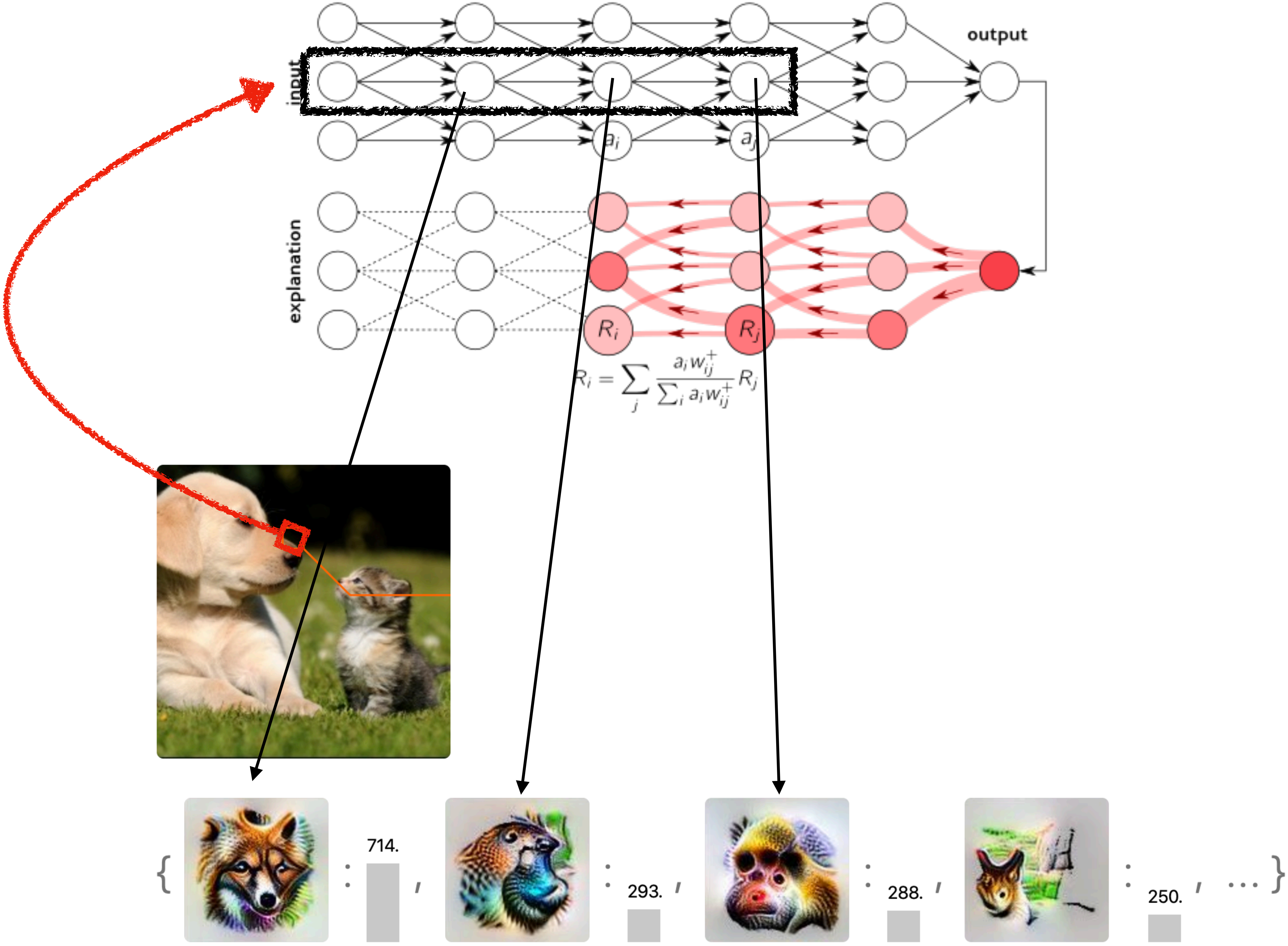


$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

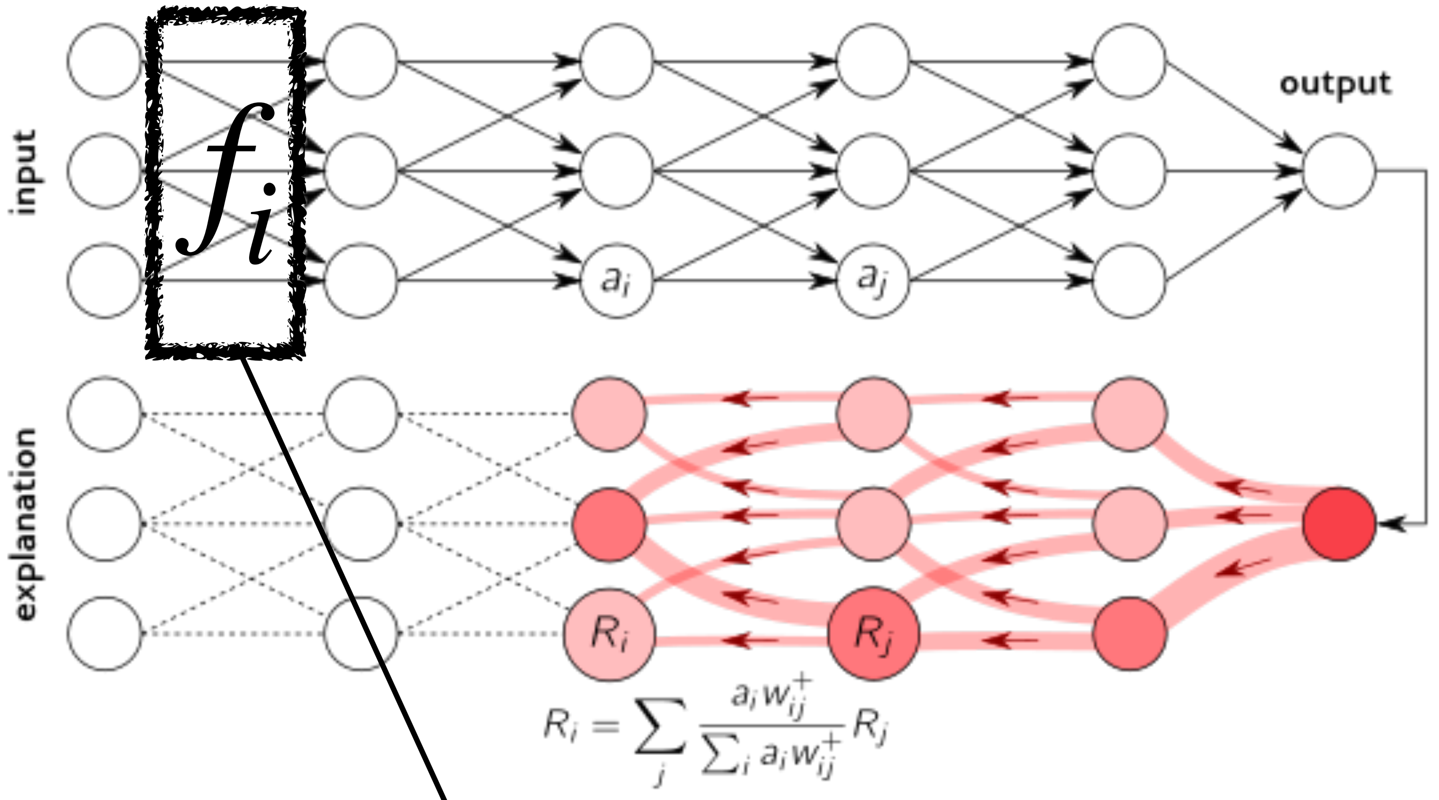

input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
{
: 714. ,
: 293. ,
: 288. ,
: 250. ,
... }


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. ,
… }

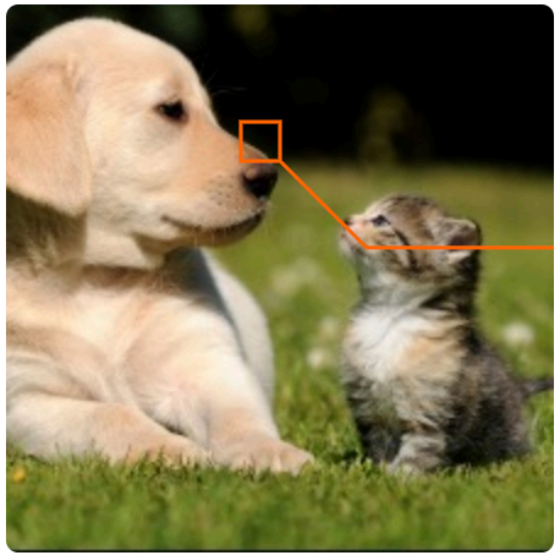

input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
{ : 714. , : 293. , : 288. , : 250. , ... }



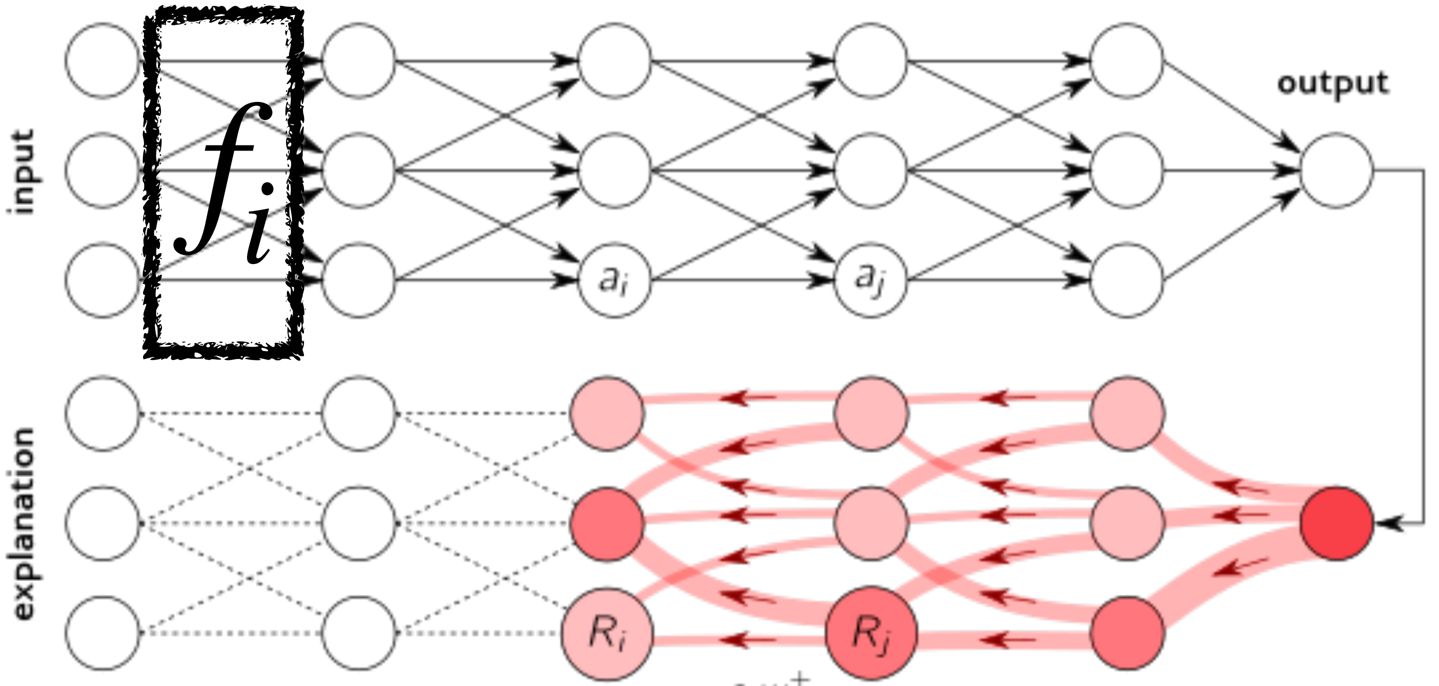
output
input
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
{
: 714. ,
: 293. ,
: 288. ,
: 250. ,
... }



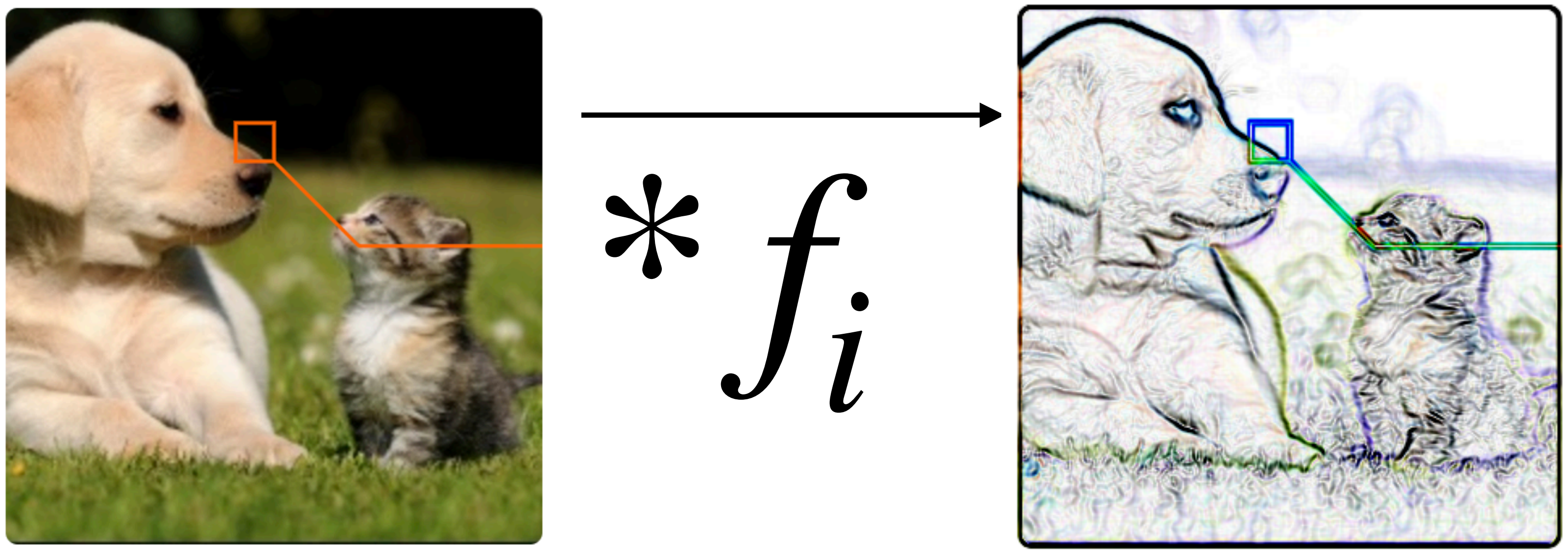
input
explanation
output
f_i
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$

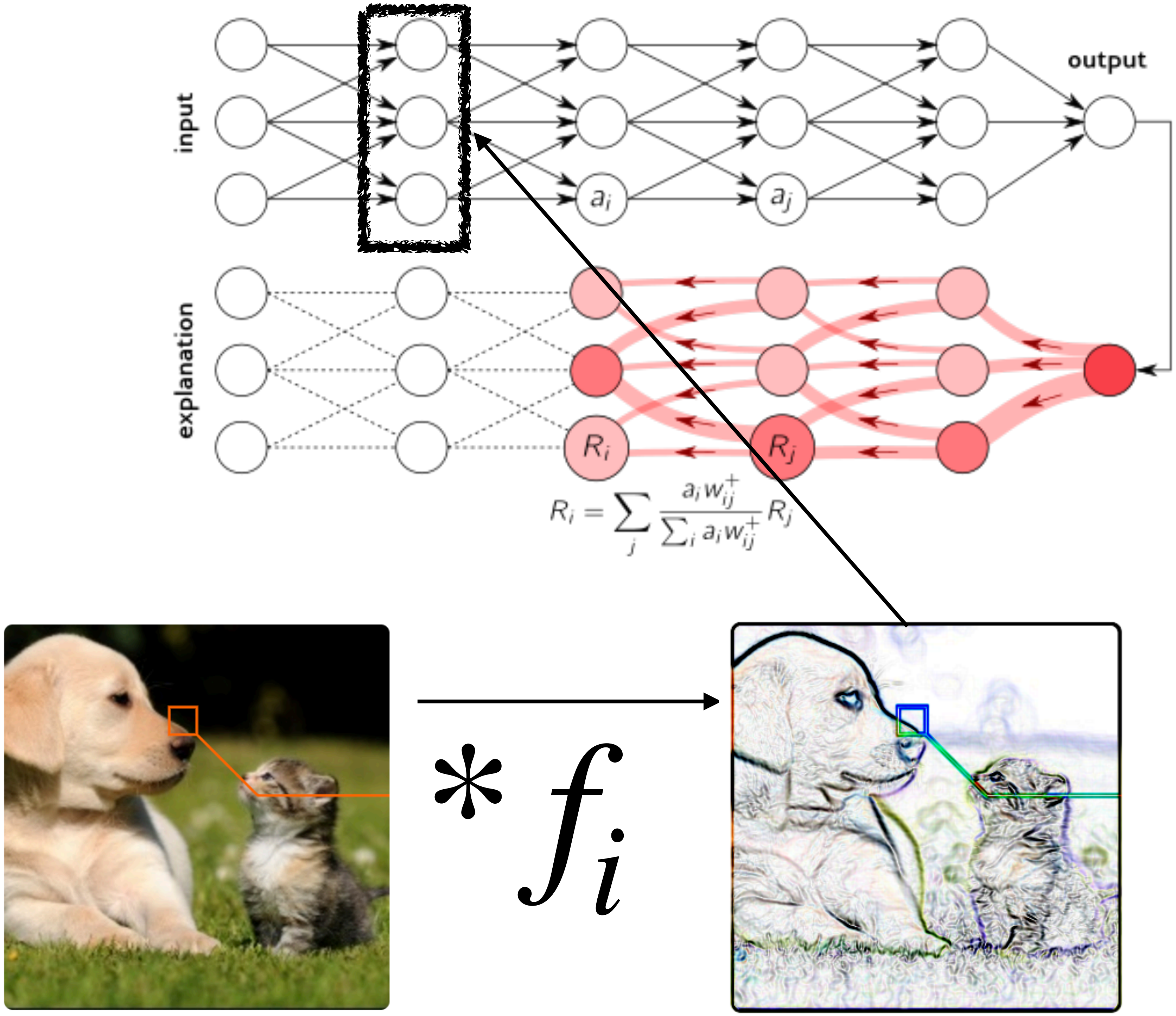


$$*f_i$$

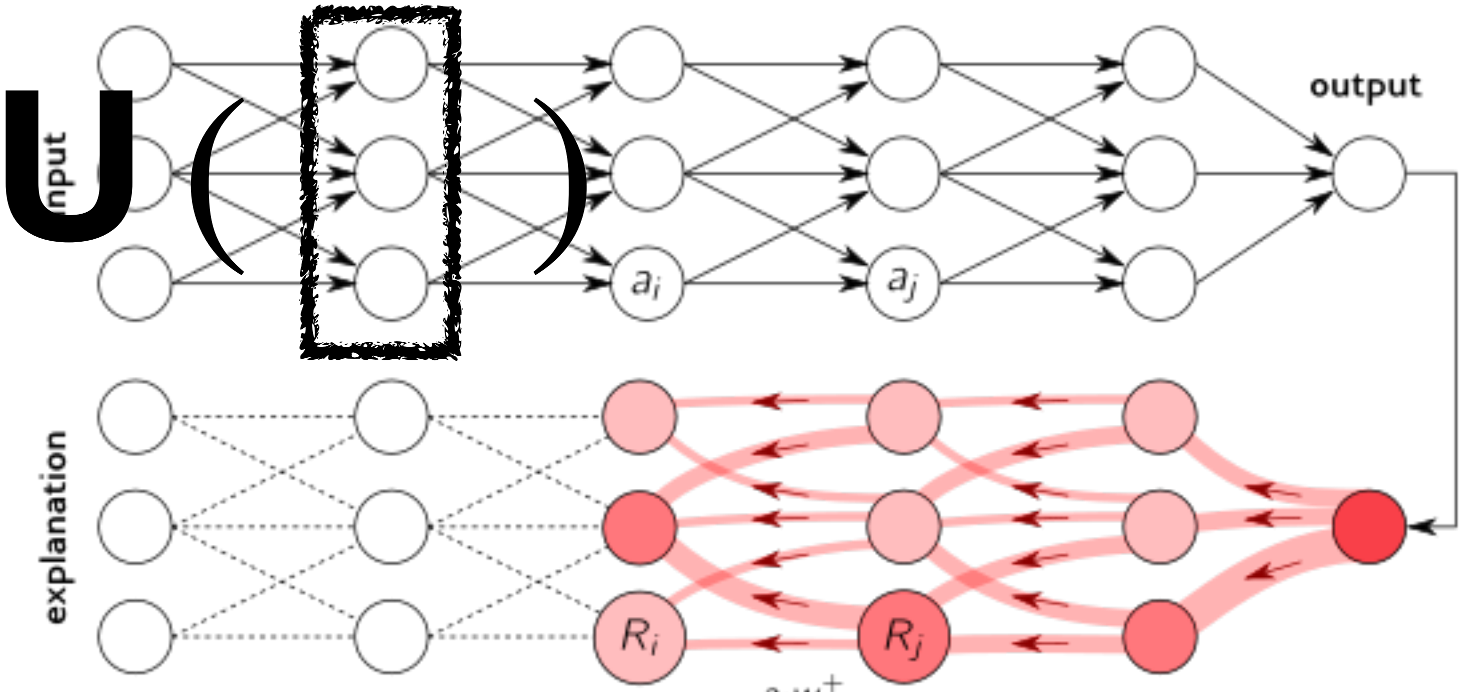


$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

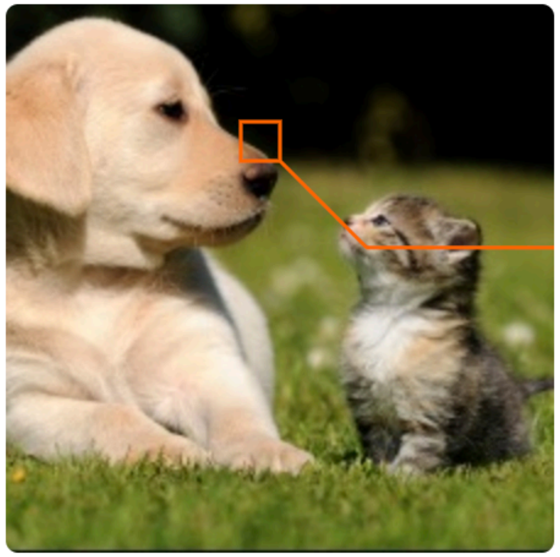



input
output
explanation
a_i
a_j
R_i
R_j
$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$
$*f_i$

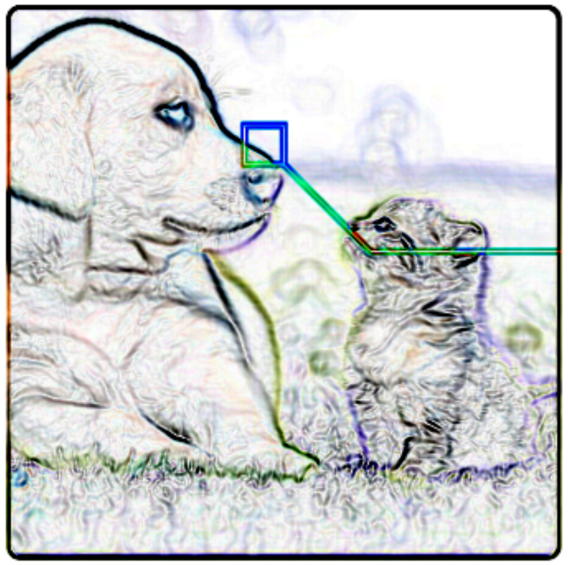
RELU( )

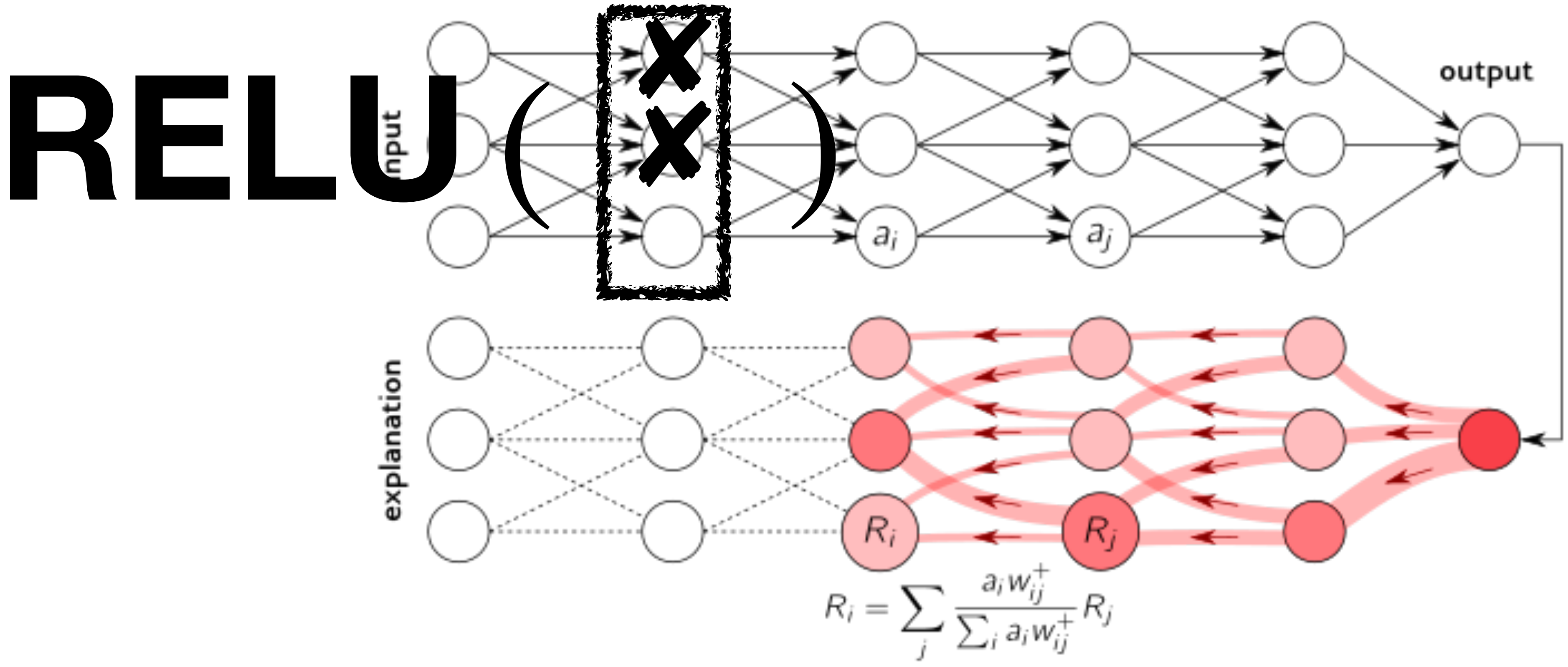


$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$



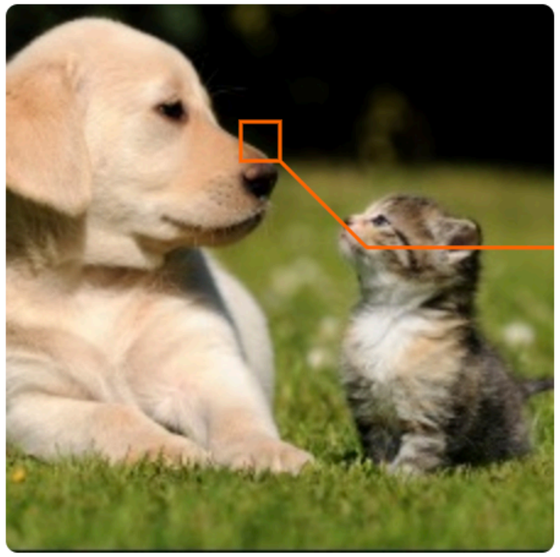
$$* f_i$$



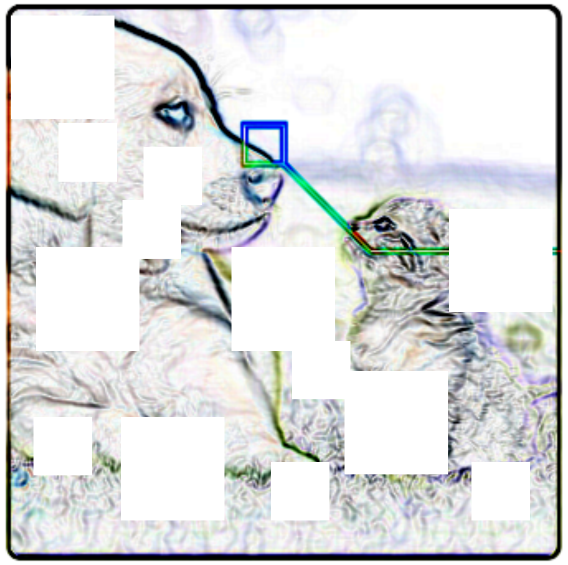


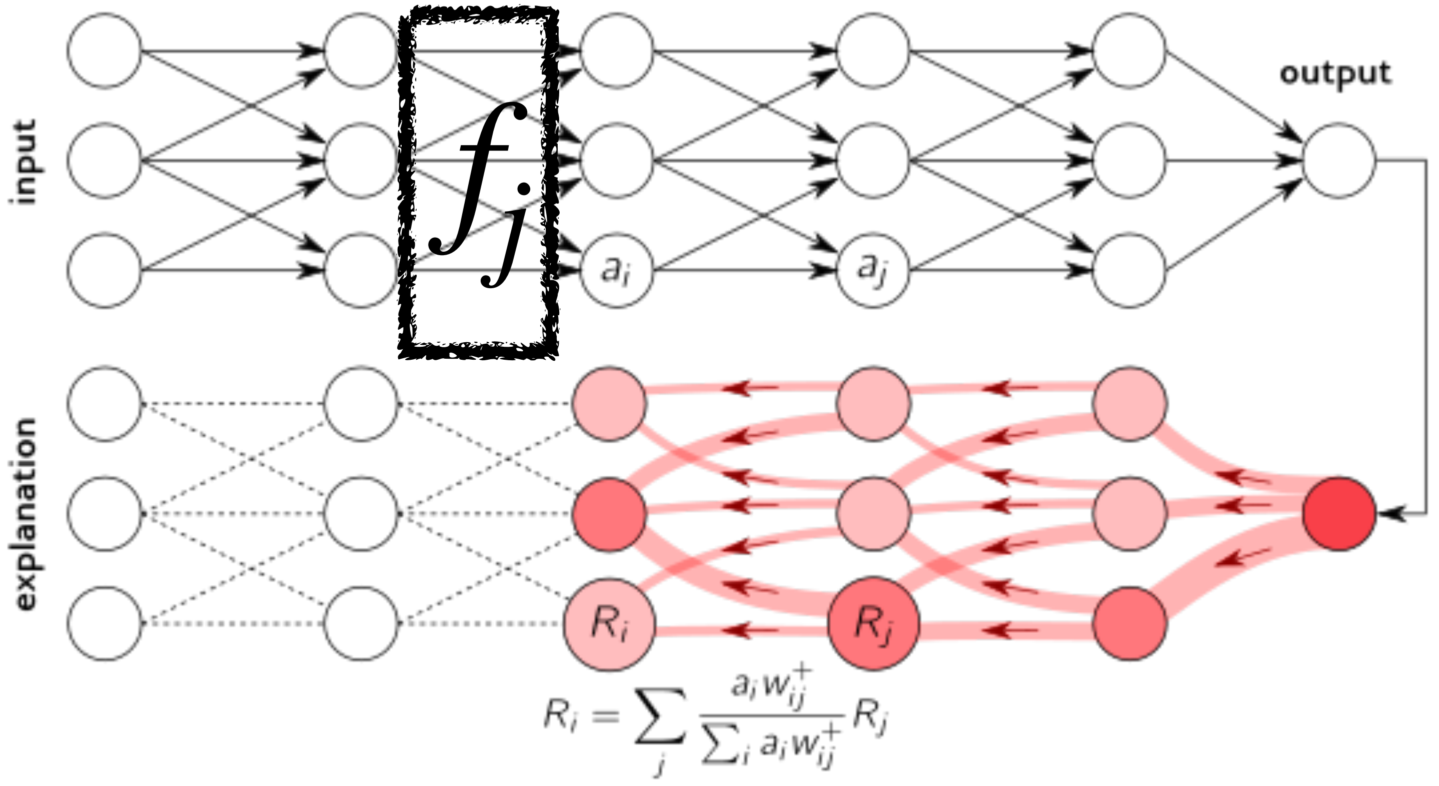
RELU( )

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

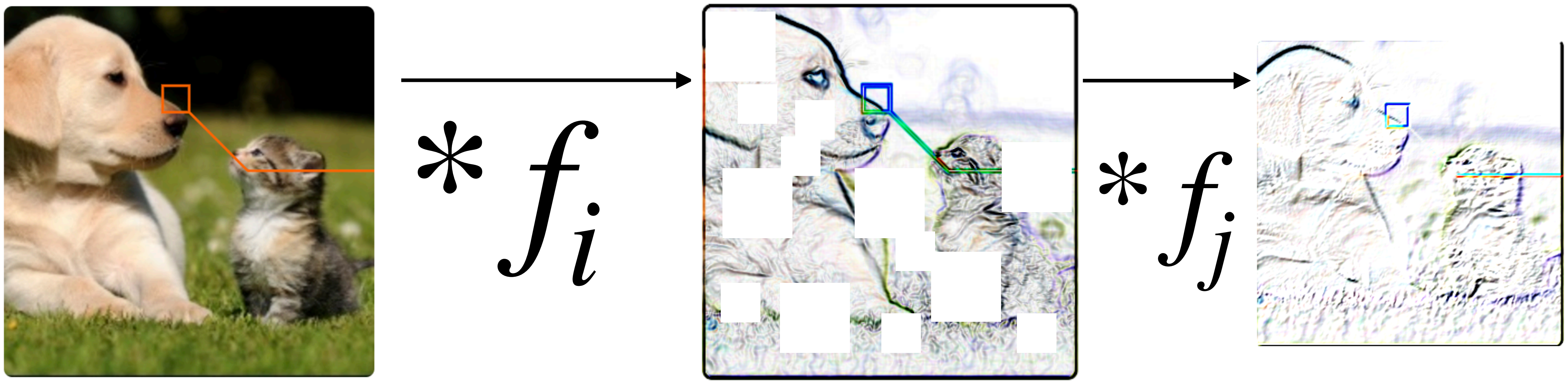


$$\xrightarrow[*f_i]{}$$





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



What questions can we currently answer?

- Given **one** manually selected input:
    - On **which parts** of the input the does the model **focus**? (f.e. *LRP*)
- Given **one** selected output:
    - What different **strategies (clusters) exist** for 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*)

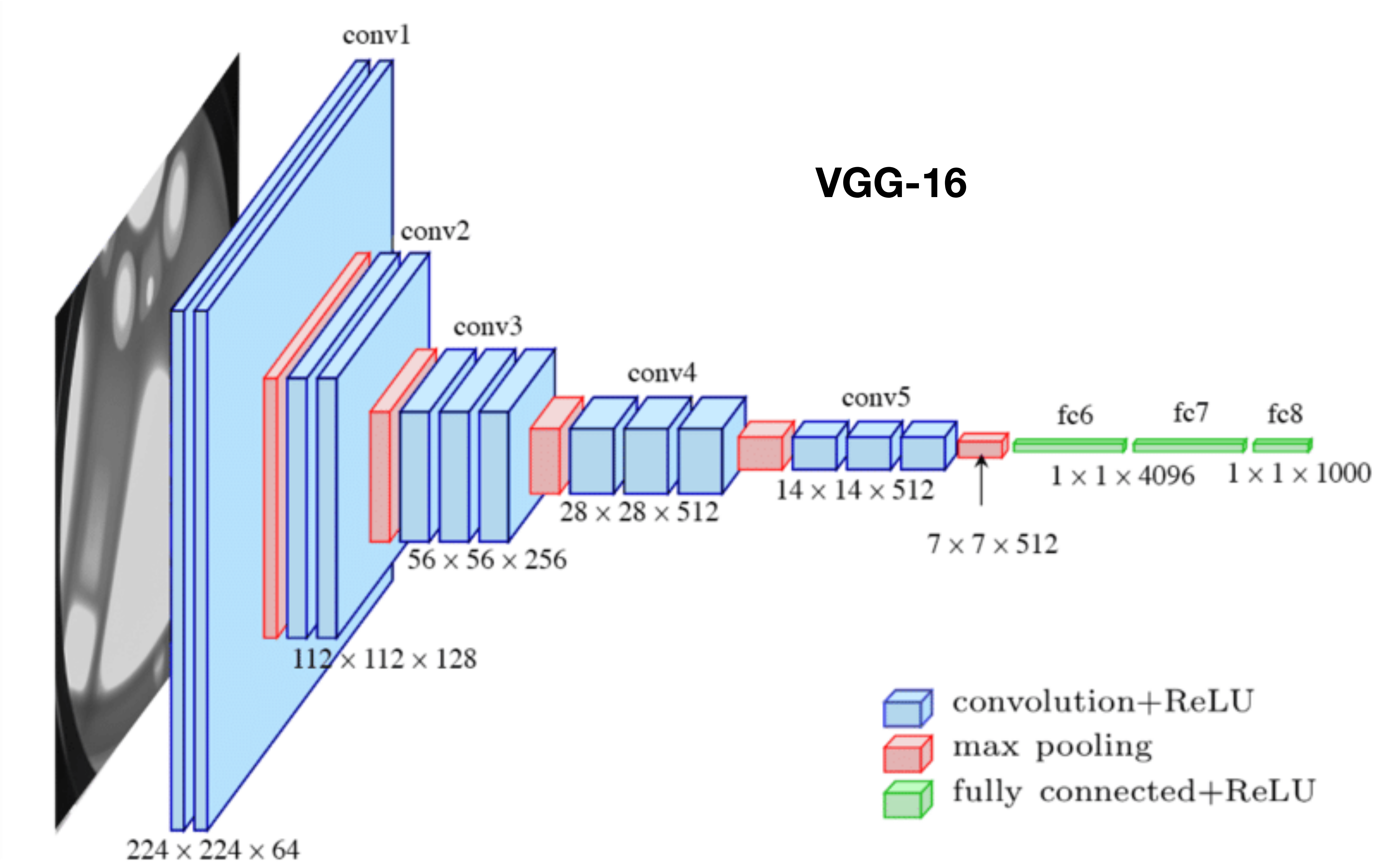
Hands On

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
- Don’t know a similar dataset for TimeSeries (if anyone knows, please tell me!)

VGG-16



We use something VGG like

What questions can we currently answer?

- Given **one** manually selected input:
  - On **which parts** of the input does the model **focus**?
    - 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?

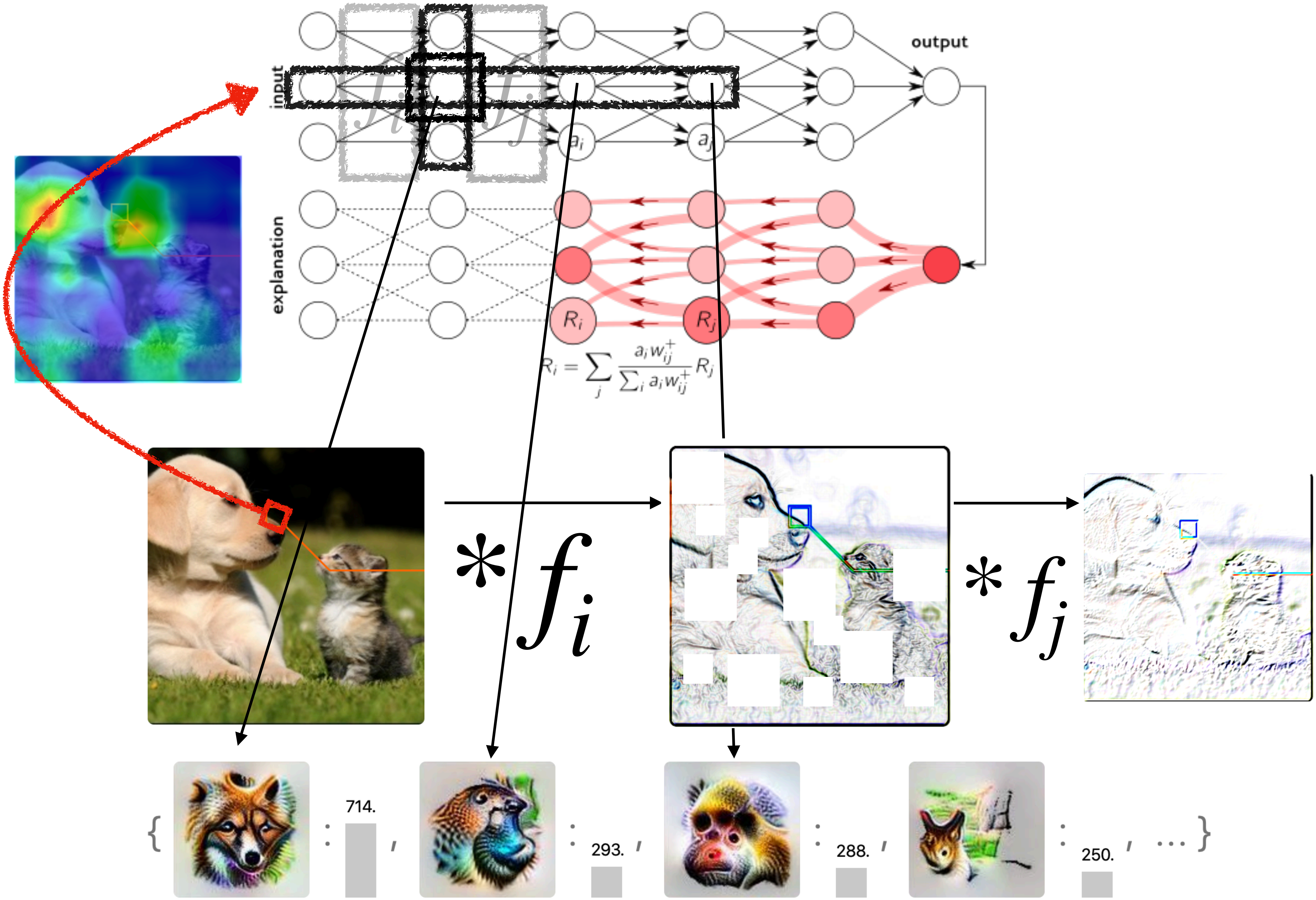
- Given **one** selected output:
    - What **kind of template** does it look for?
        - **Max Activation**, Project Lucid, Activation Atlas
        - distill.pup

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**



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

Thanks for listening

I hope there was something of value for you?

We can have some Q&A in the DeepLearning Discord chat

https://discord.gg/nvdxH7