

Advanced Visual Perception

Table of Contents

Intro to Advanced Visual Perception

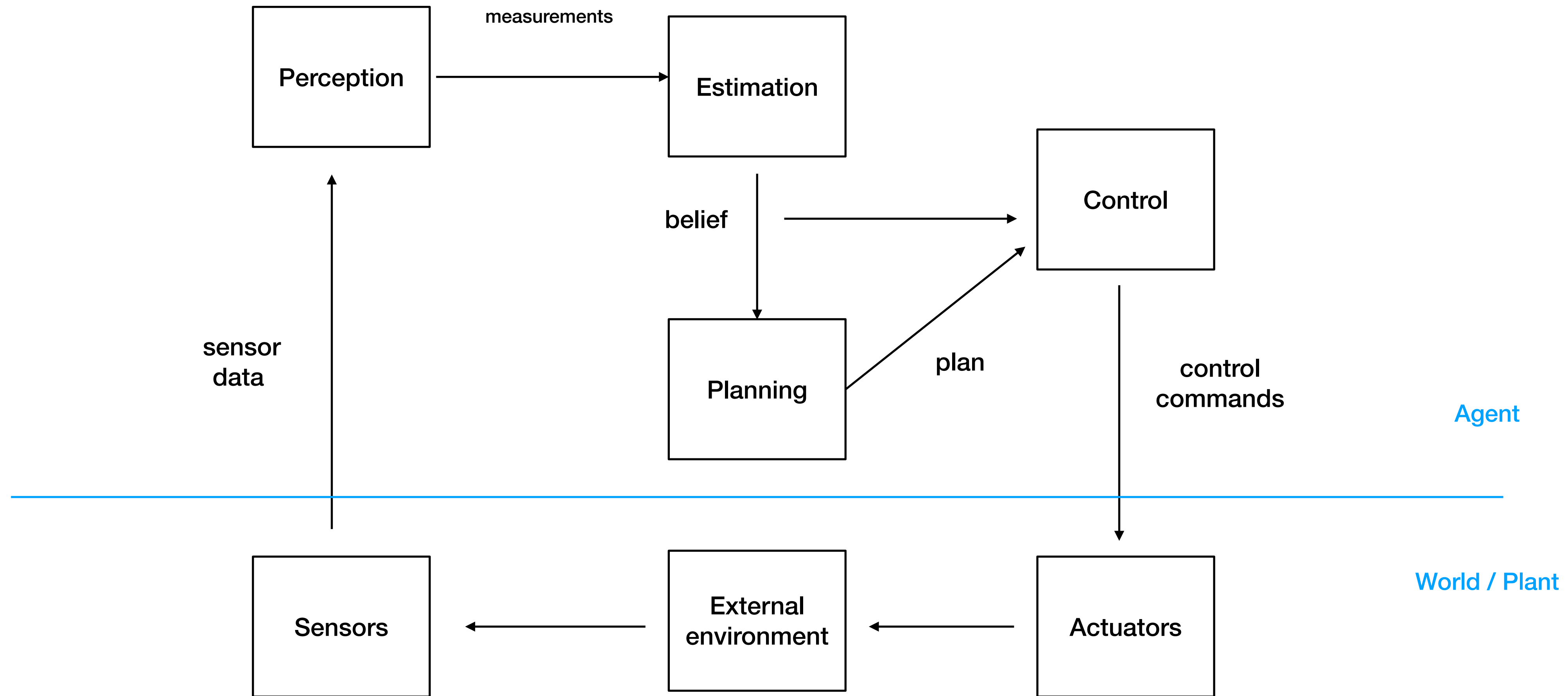
Intro to Neural Networks

Deep Convolutional Neural Networks

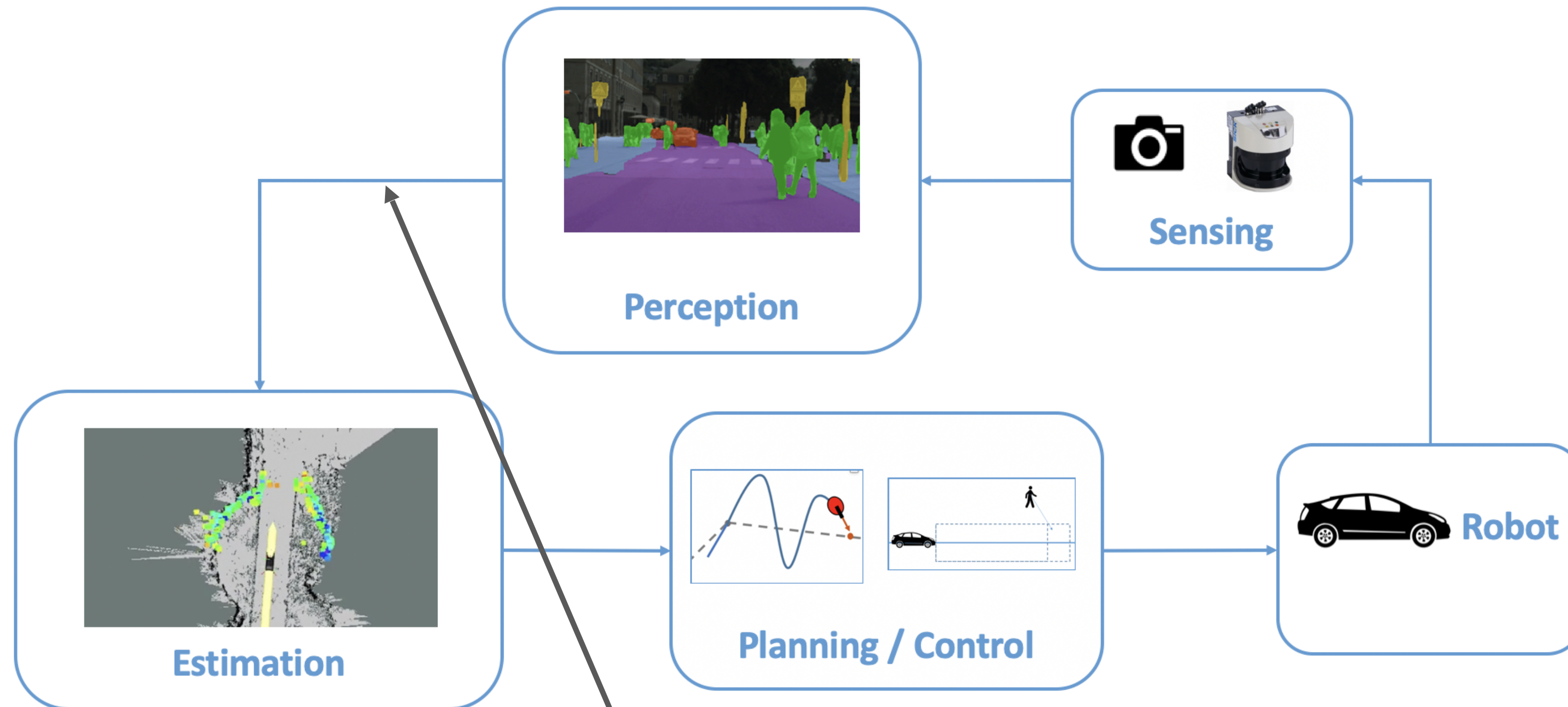
Object Detection

Semantic Segmentation

What if we don't have labels?



Perception as a sensor



$$p(\text{state}|\text{measurements}) = \frac{\text{measurement likelihood} * \text{prior}}{\text{evidence}}$$

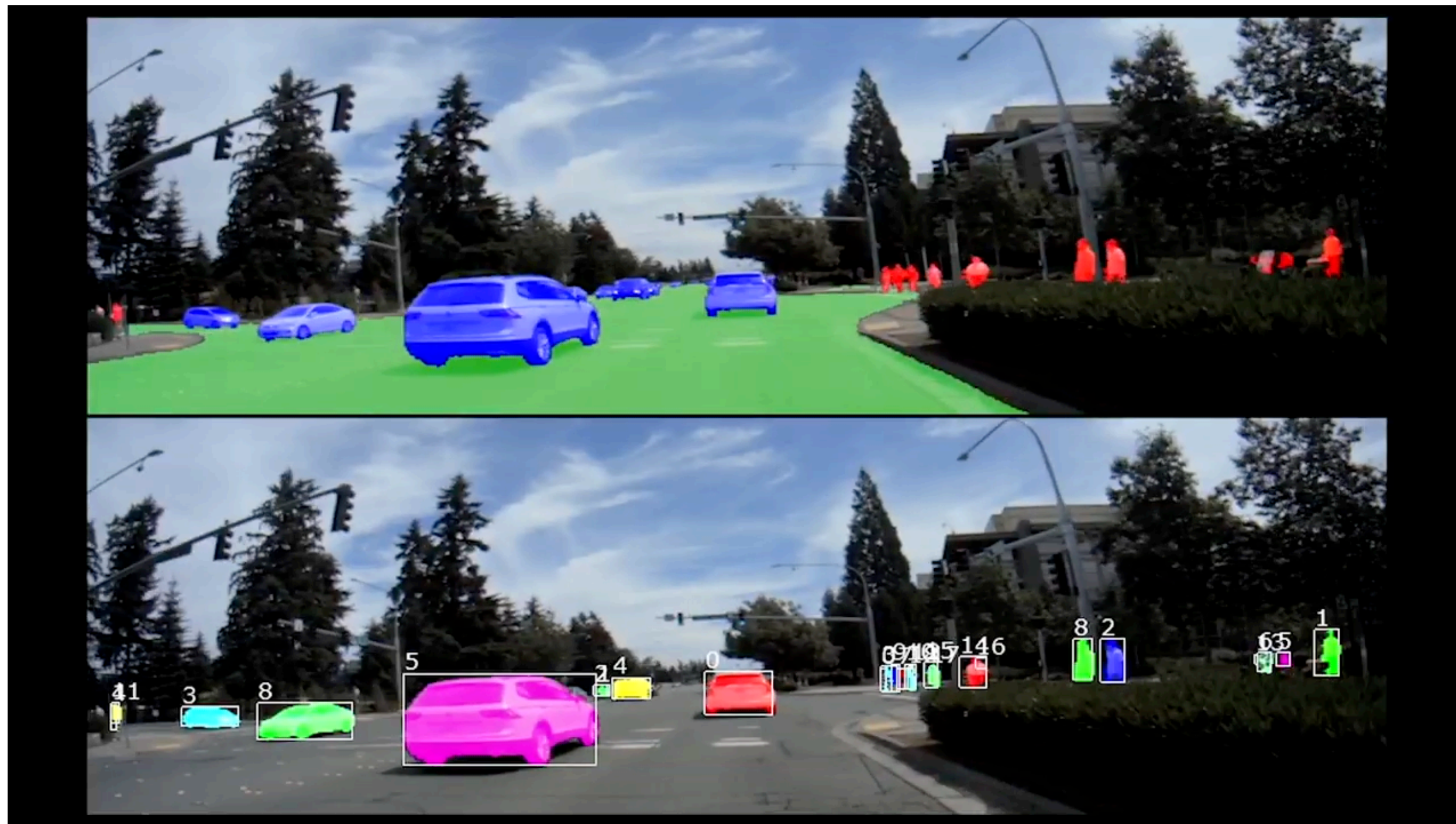


Image classification



Duckie

Image segmentation



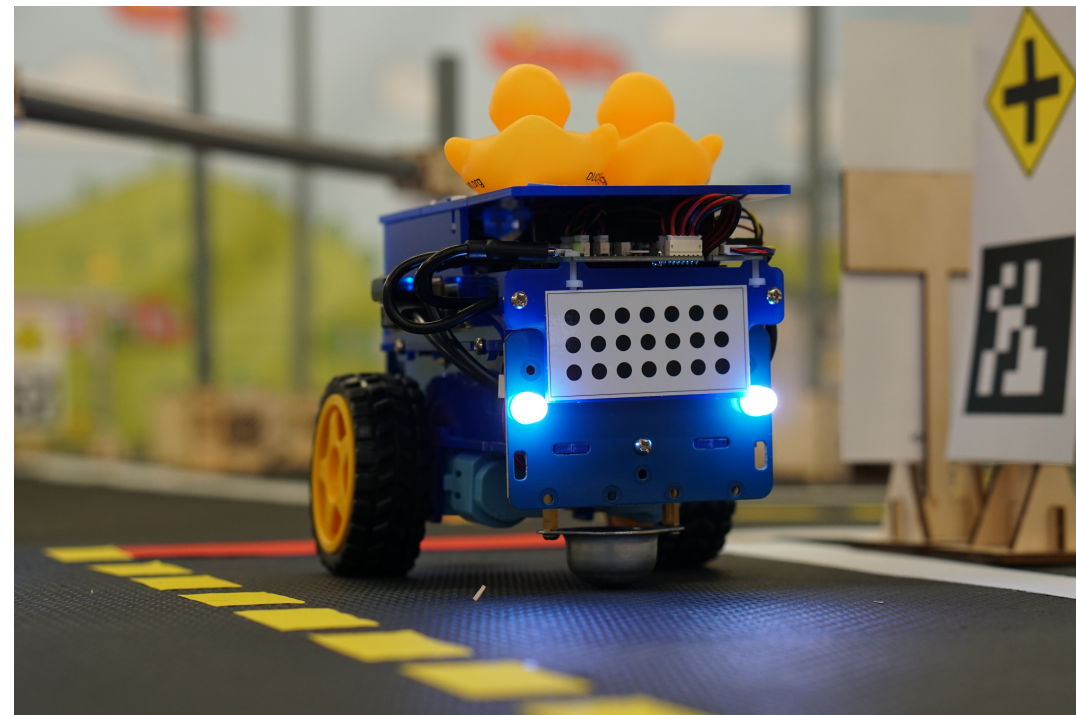
Object Detection



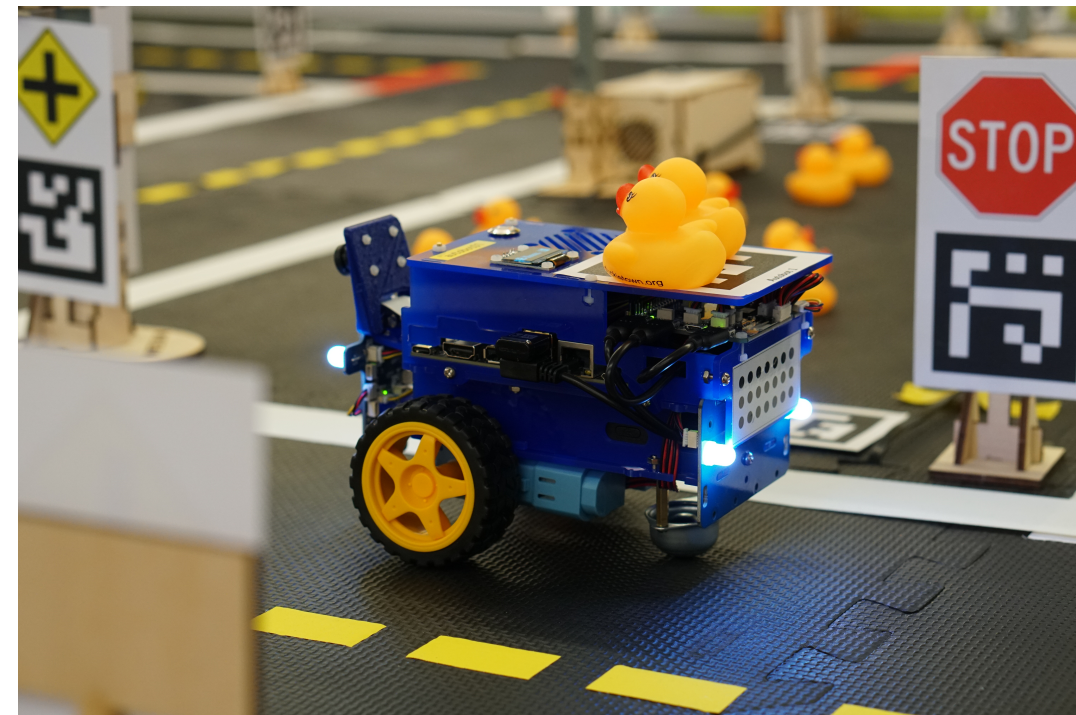
Instance Segmentation



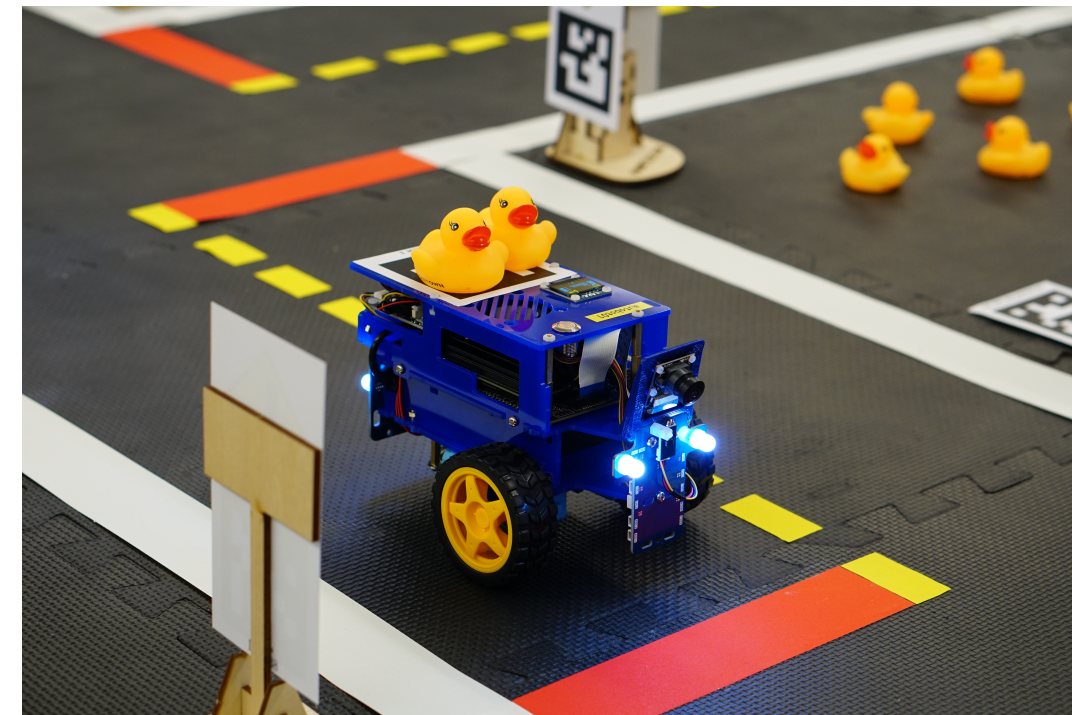
Classification Accuracy



Duckiebot



Duckiebot



Duckiebot



Airplane

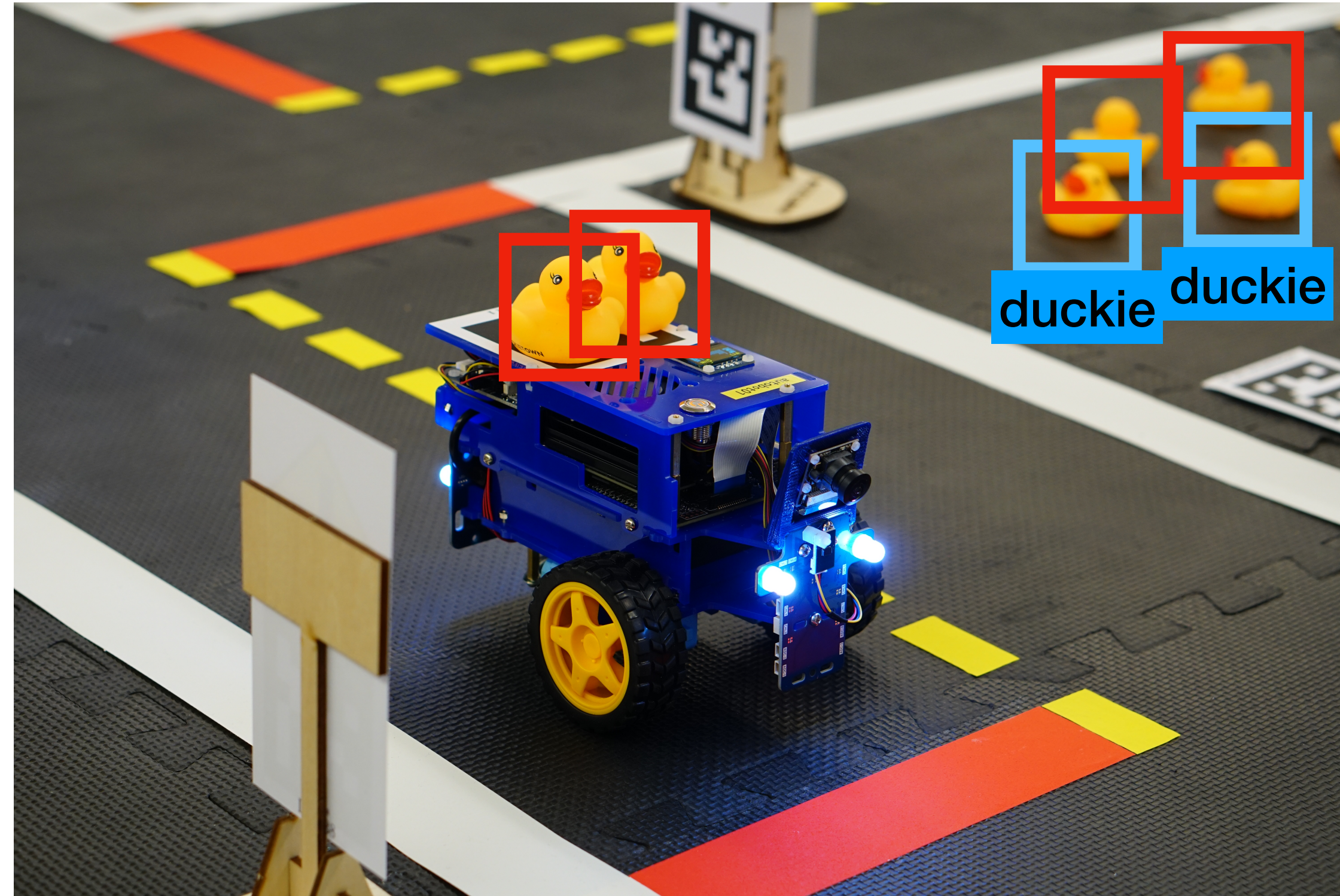
$$\text{Accuracy} = \frac{\# \text{ correct}}{\# \text{ total}} = \frac{3}{4}$$

Precision and Recall



$$\text{precision} = \frac{\# \text{ True Positives}}{\# \text{ True Positives} + \# \text{ False Positives}} = \frac{2}{3}$$

Precision and Recall



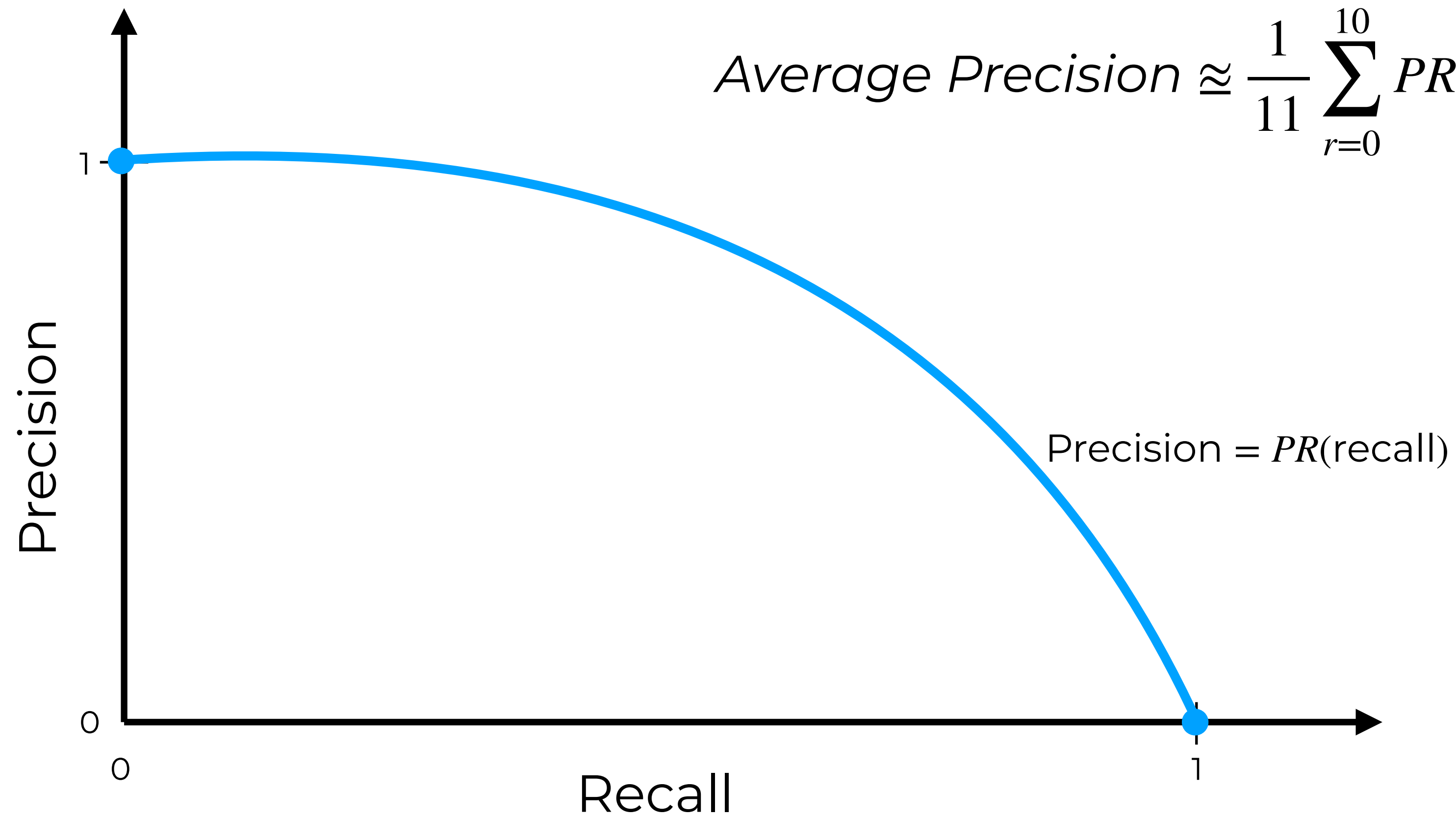
$$\text{precision} = \frac{\# \text{ True Positives}}{\# \text{ True Positives} + \# \text{ False Positives}} = \frac{2}{3}$$

$$\text{recall} = \frac{\# \text{ True Positives}}{\# \text{ True Positives} + \# \text{ False Negatives}} = \frac{2}{6}$$

Precision-Recall

$$\text{Average Precision} = \int_0^1 PR(\text{recall}) d\text{recall}$$

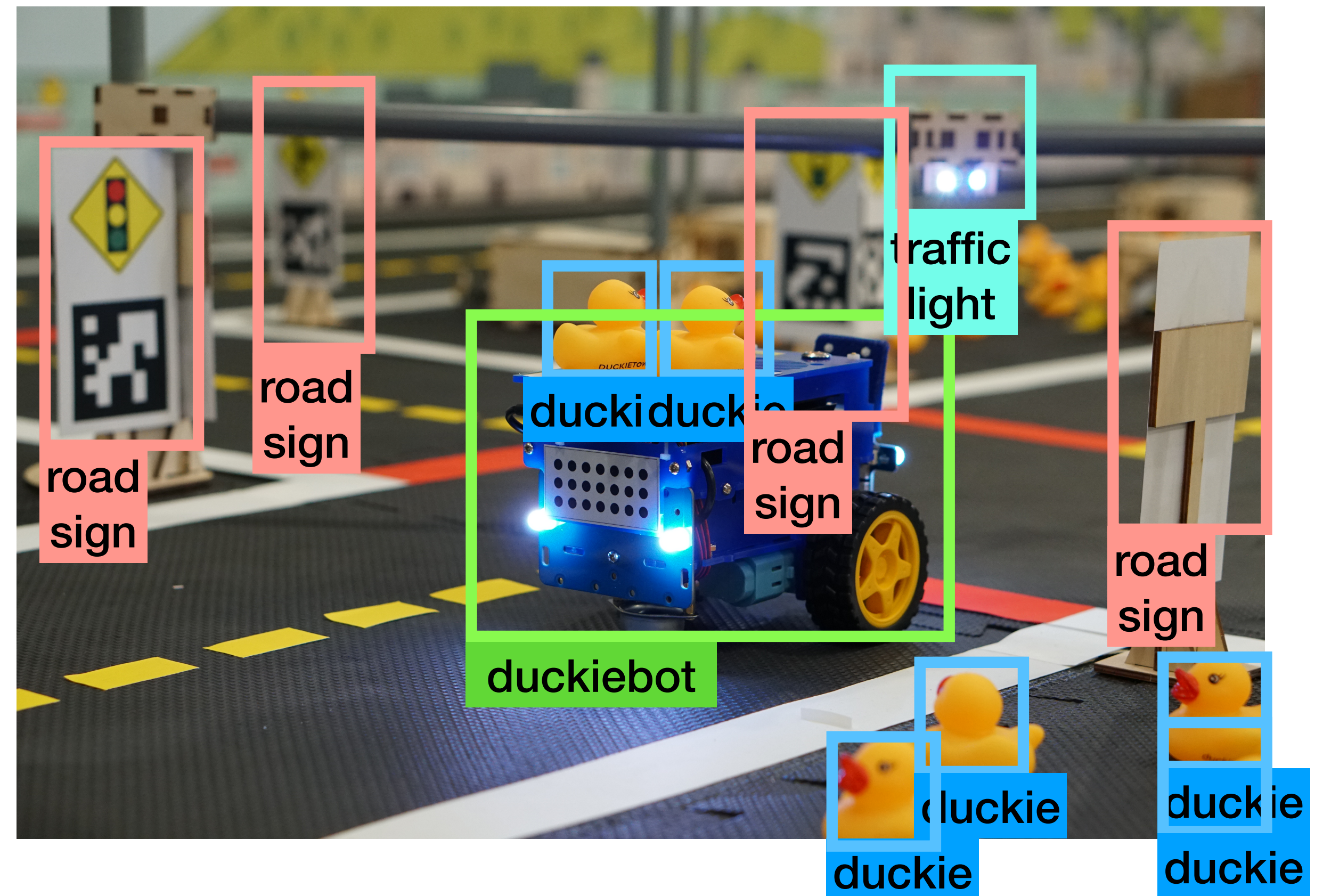
$$\text{Average Precision} \approx \frac{1}{11} \sum_{r=0}^{10} PR(0.1 * r)$$



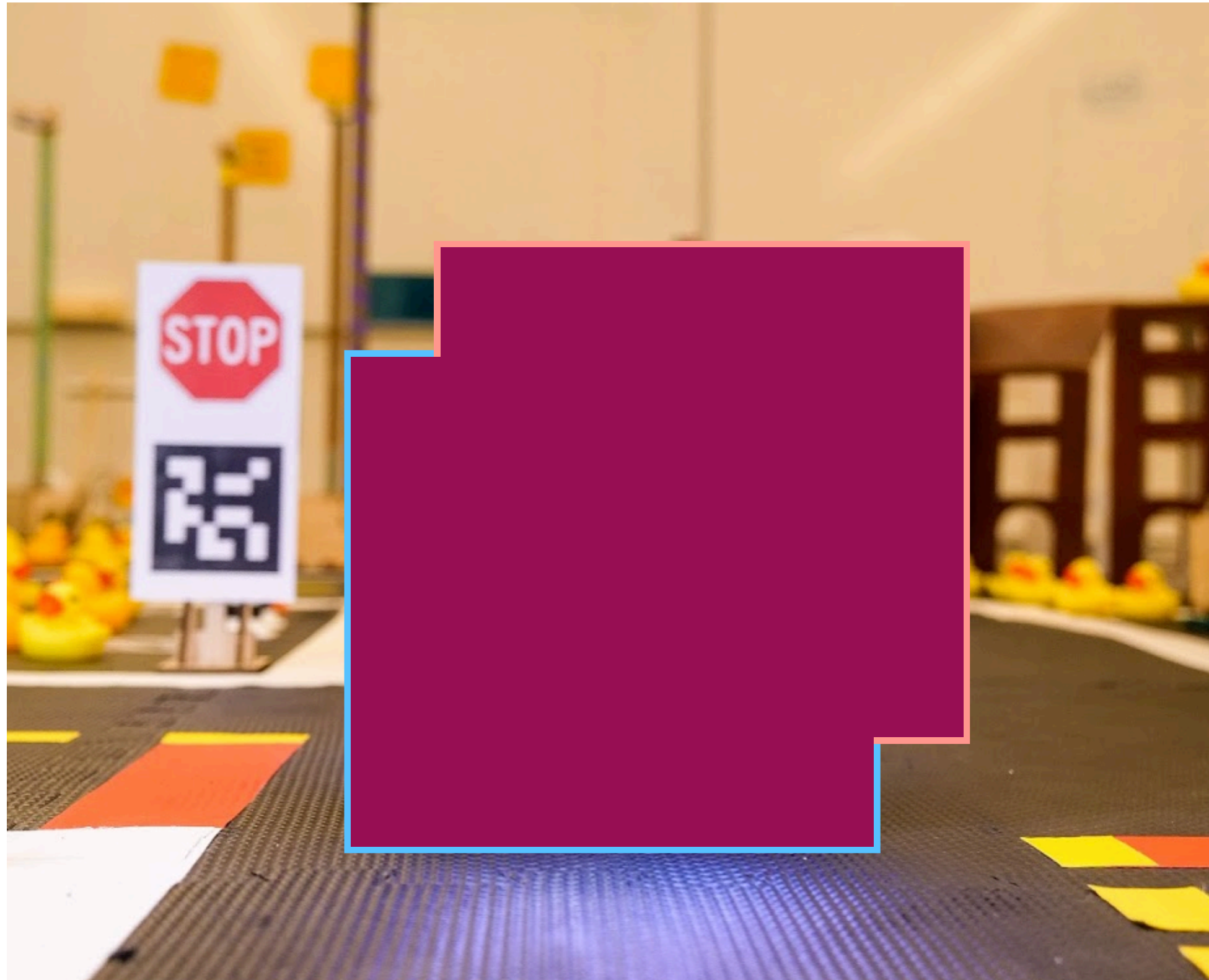
mean Average Precision (mAP)

- mean over all classes

$$mAP = \frac{1}{C} \sum_{c=1}^C AP_c$$



Intersection over Union

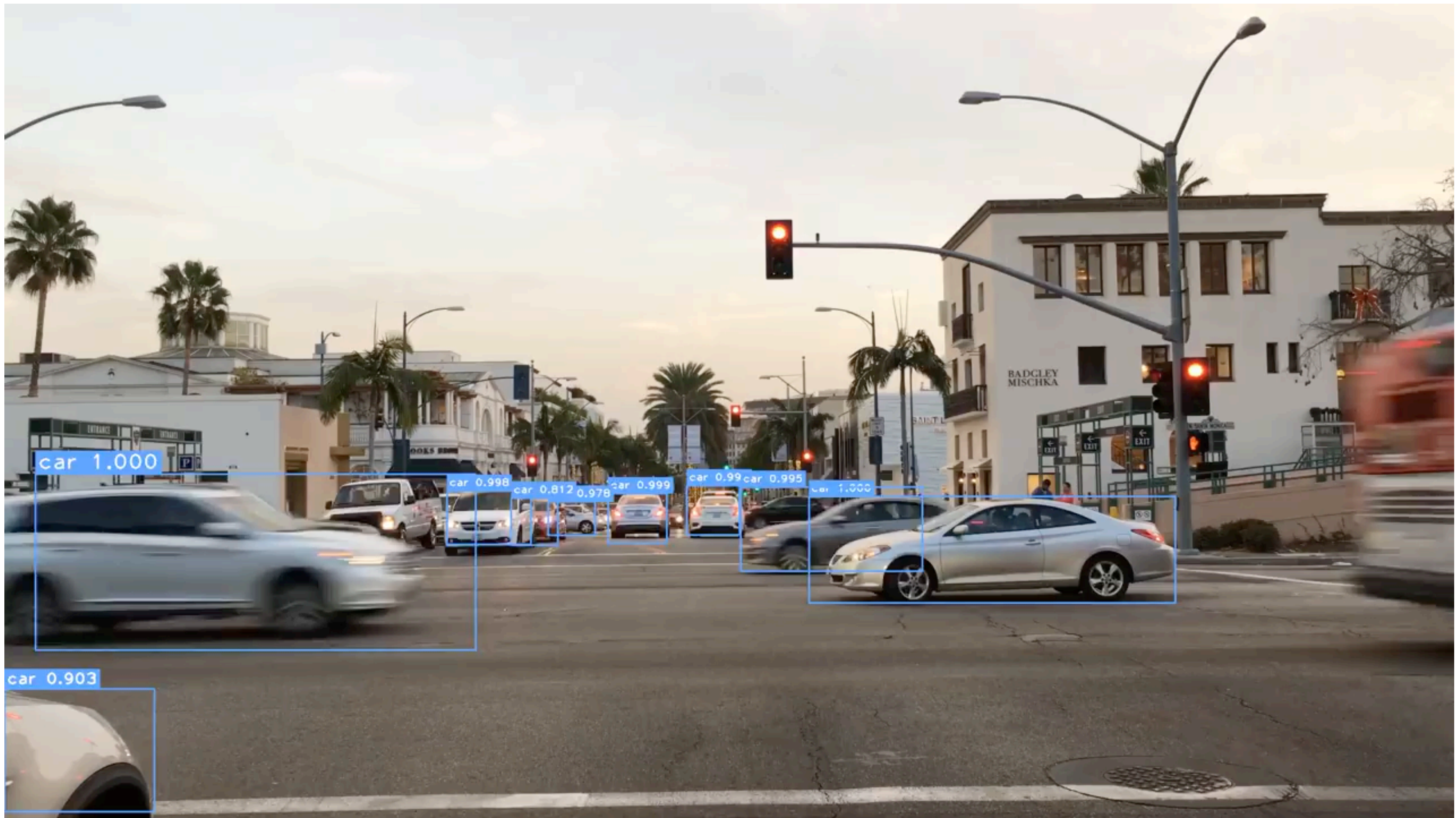


Annotated label



Model prediction





Advanced Visual Perception

Table of Contents

Intro to Advanced Visual Perception

Intro to Neural Networks

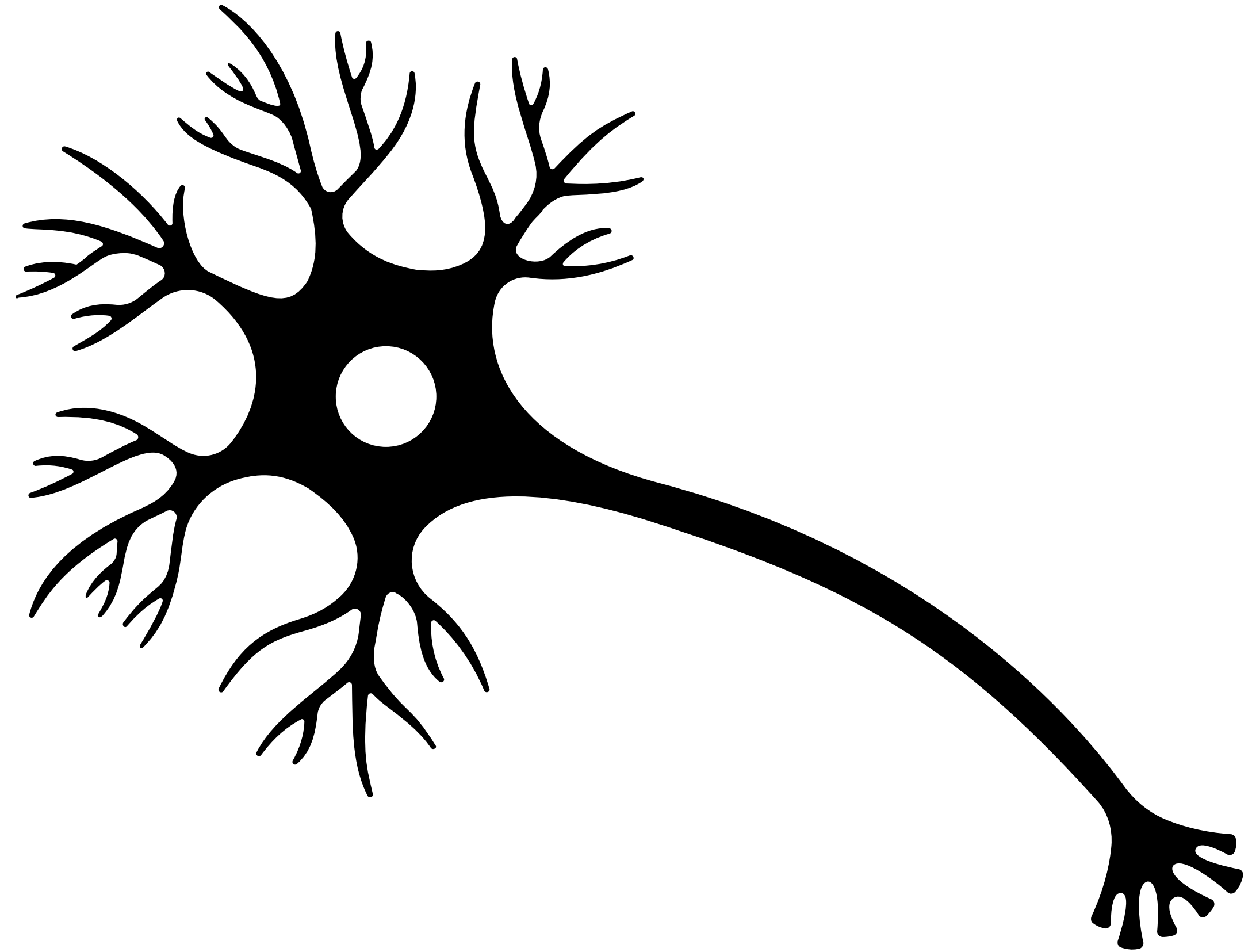
Deep Convolutional Neural Networks

Object Detection

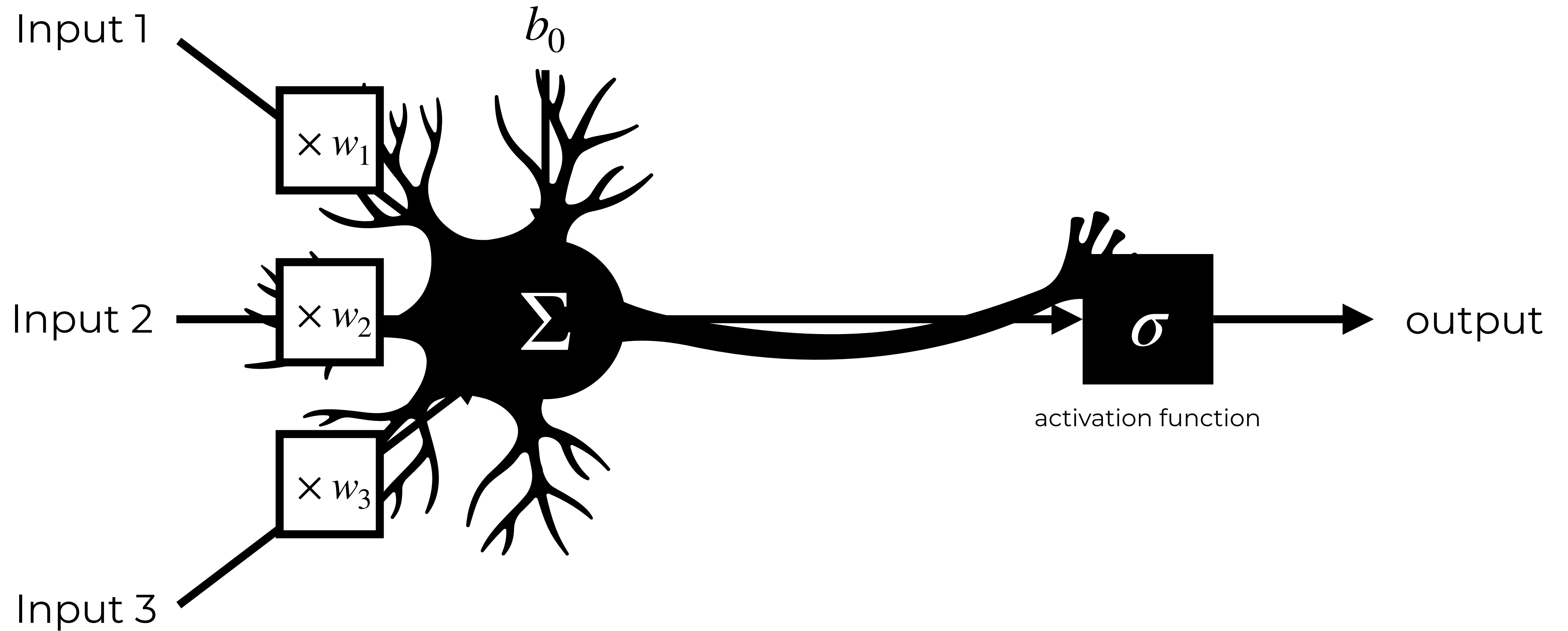
Semantic Segmentation

What if we don't have labels?

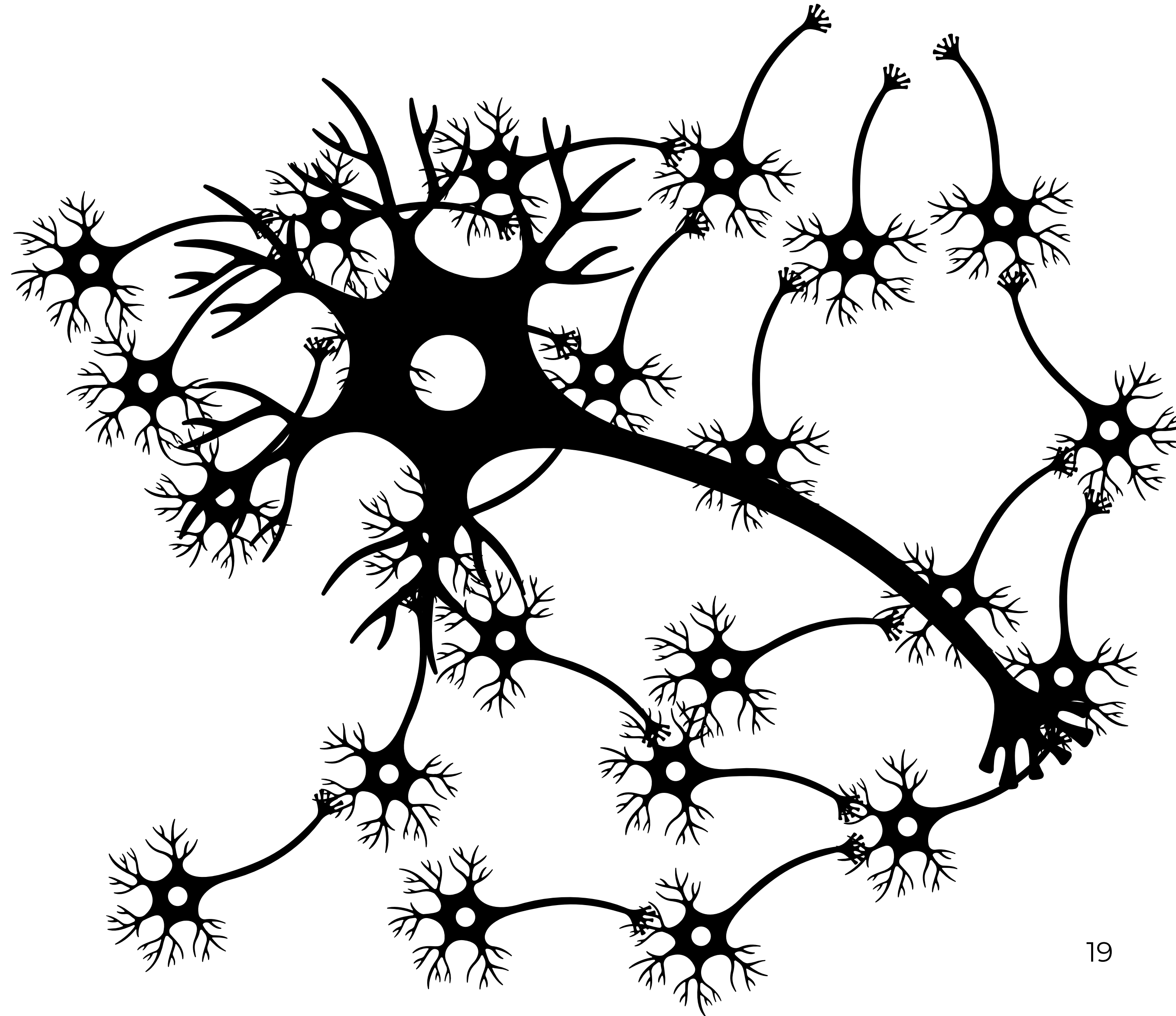
The neuron



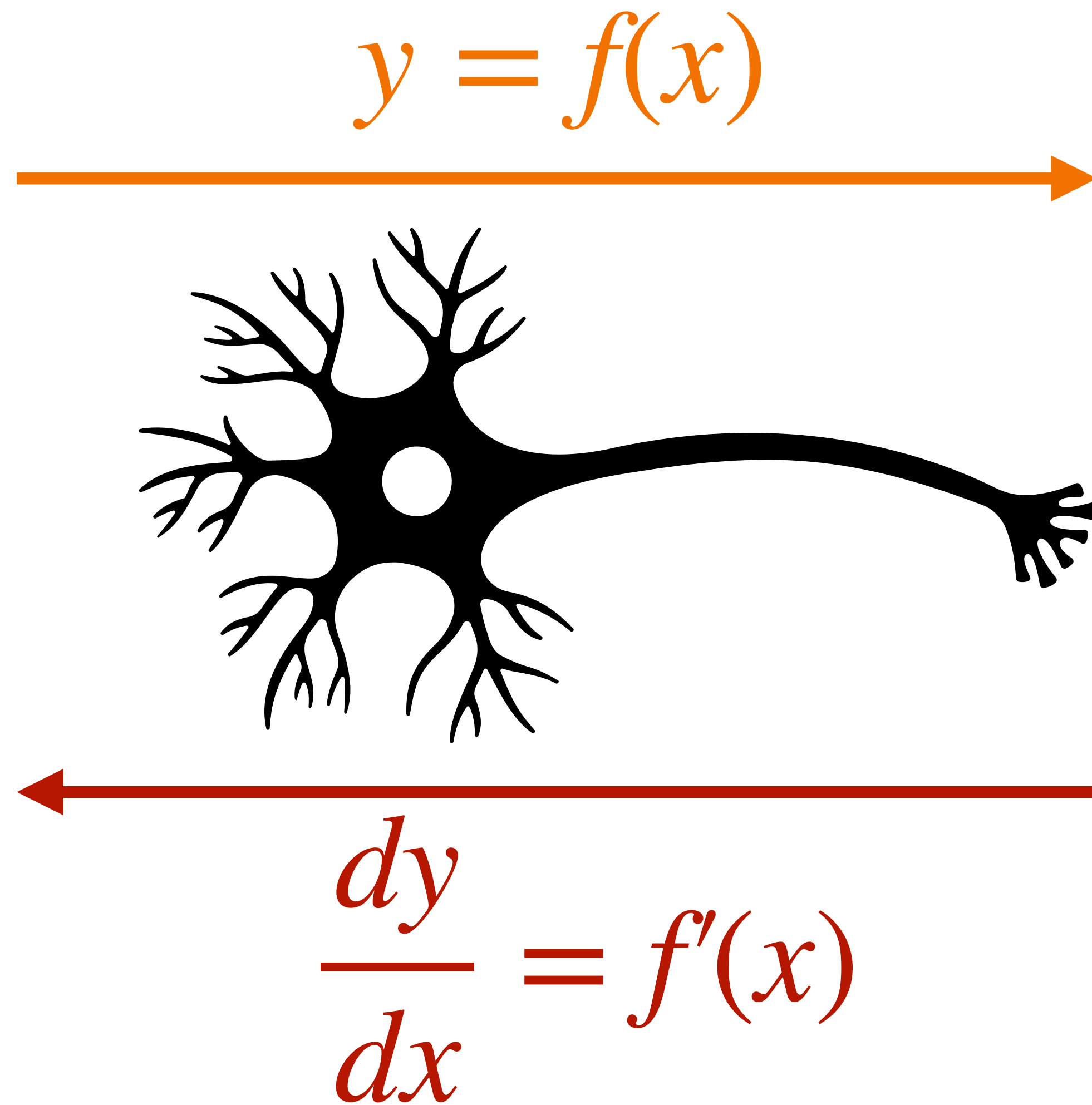
The neuron

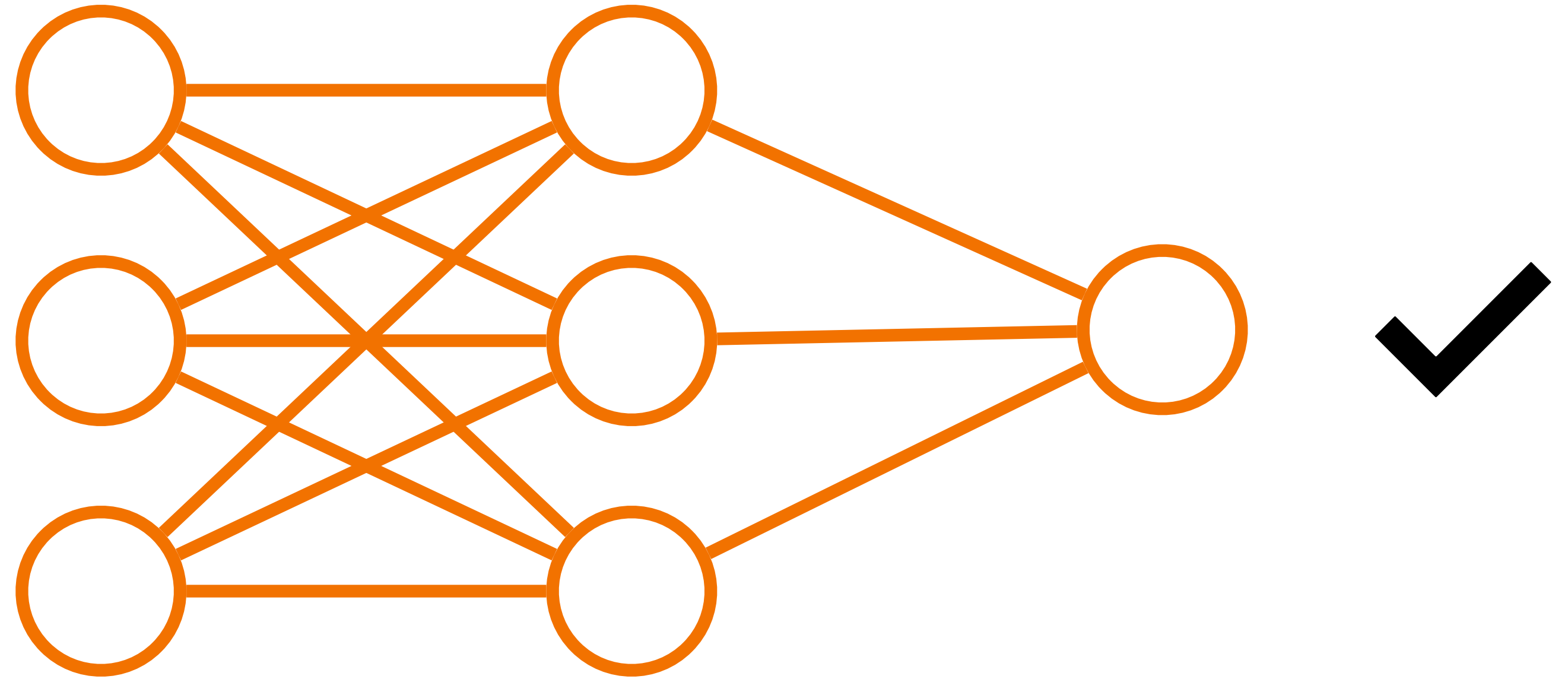
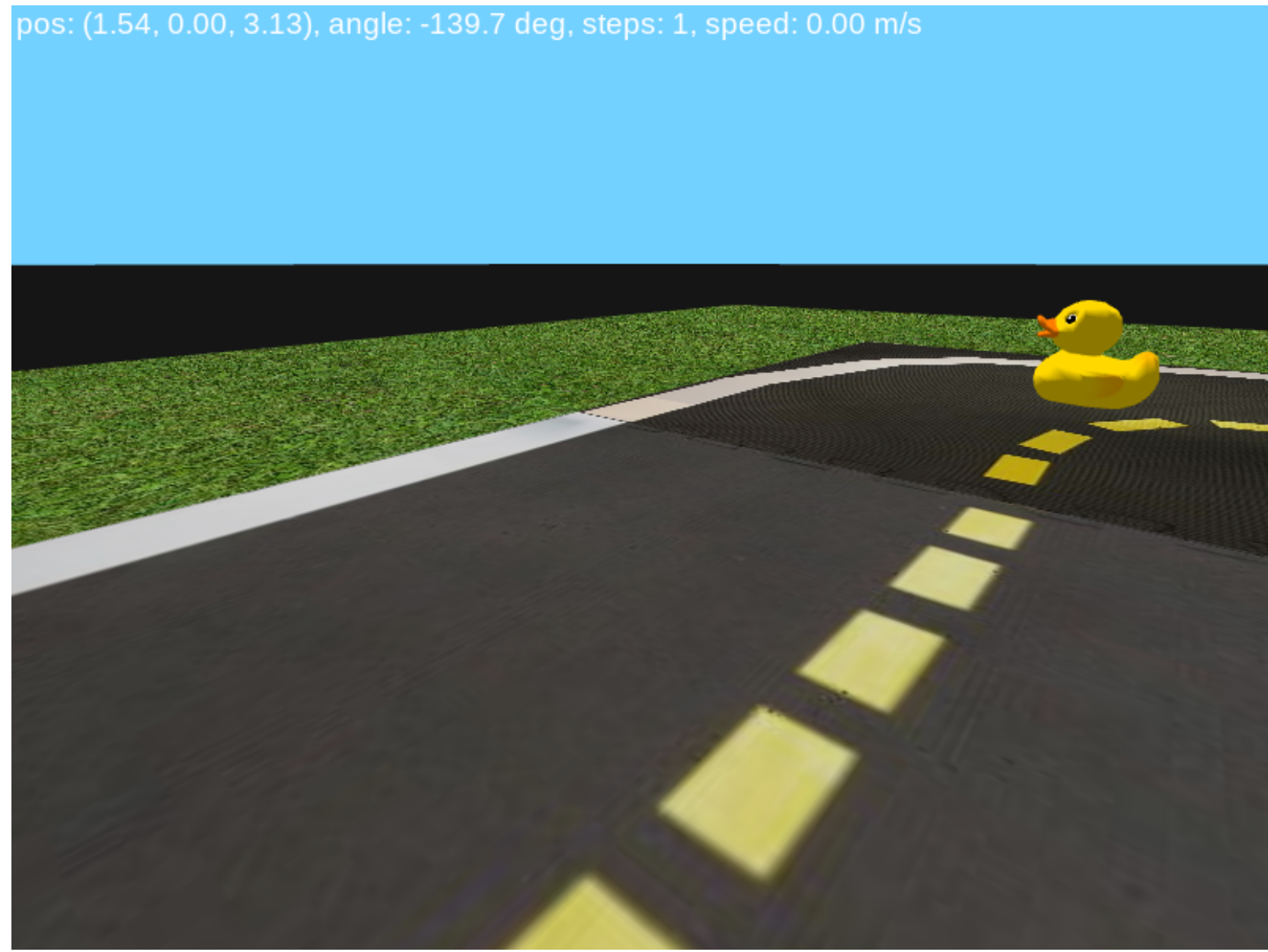


Compositionality

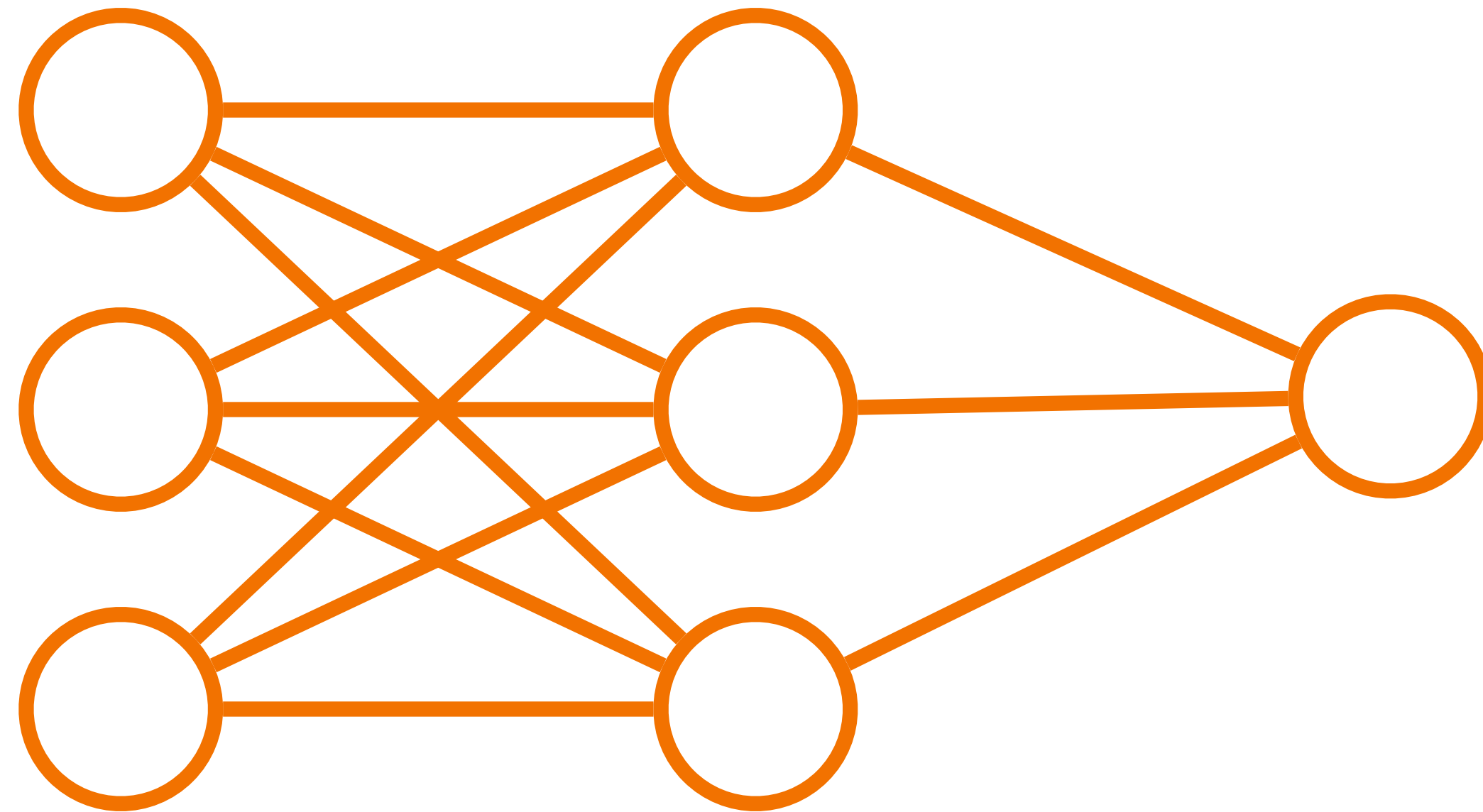


Differentiability

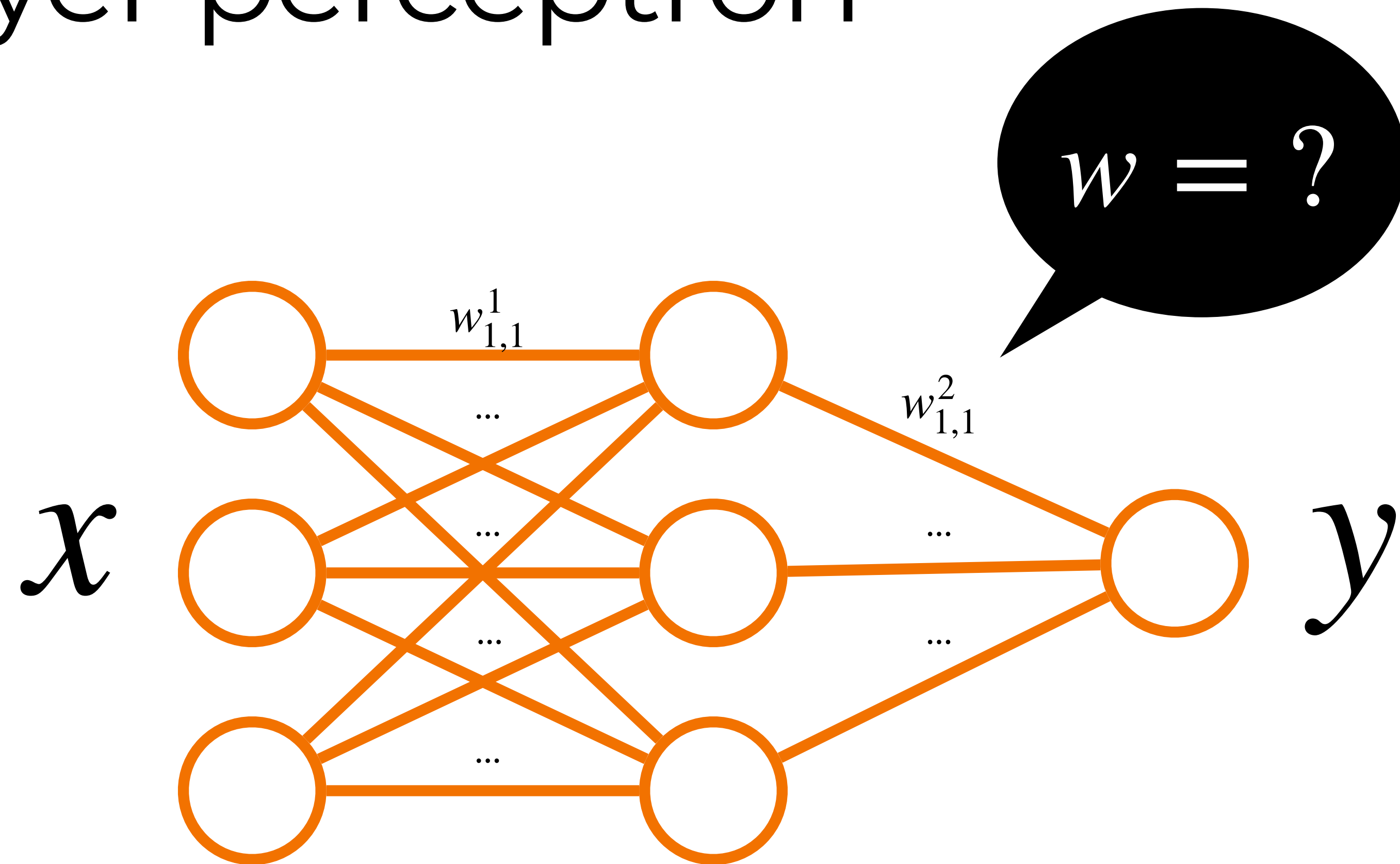




Multi-layer perceptron

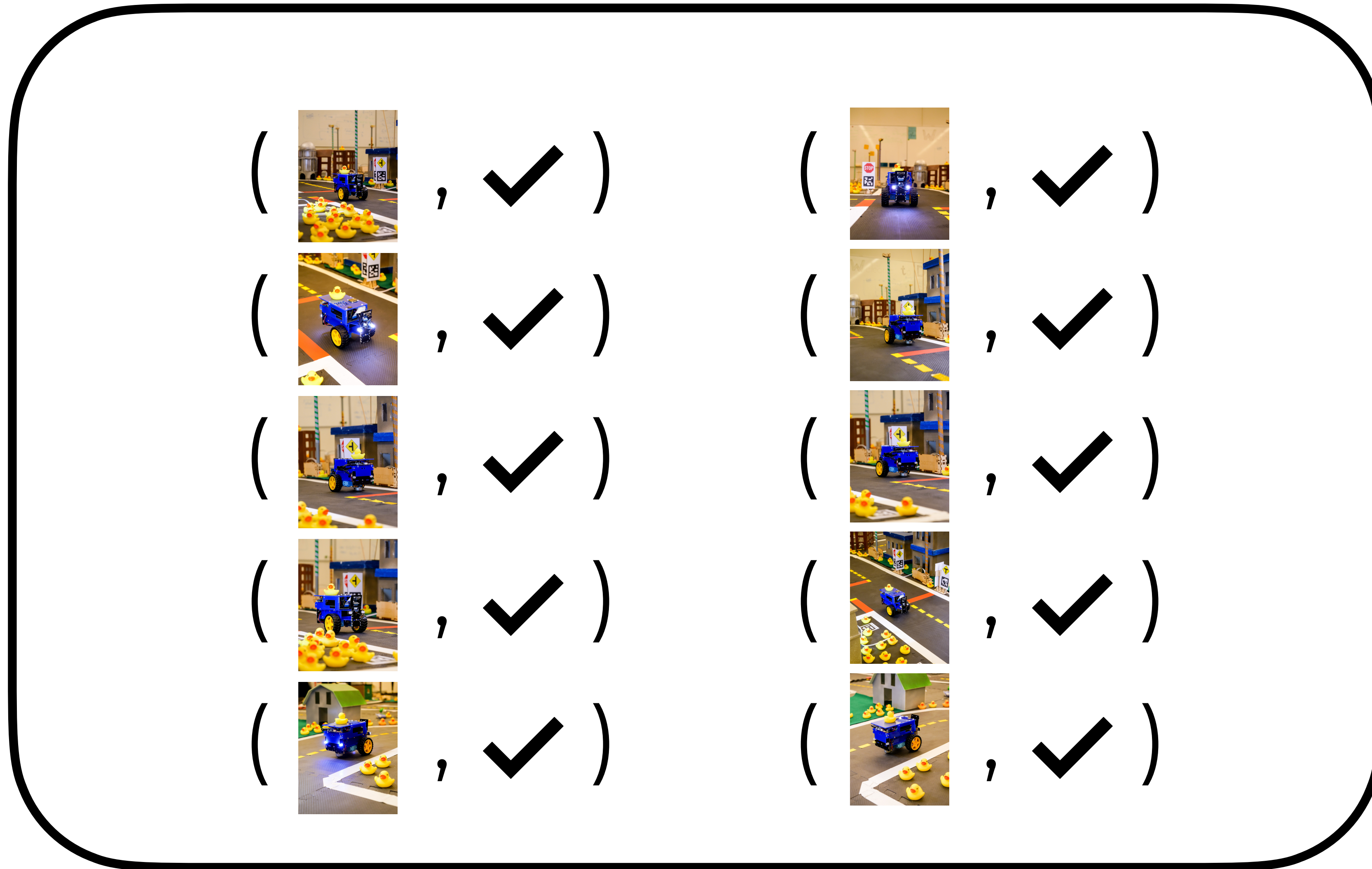


Multi-layer perceptron



$$y = f_w(x)$$

Annotated Dataset

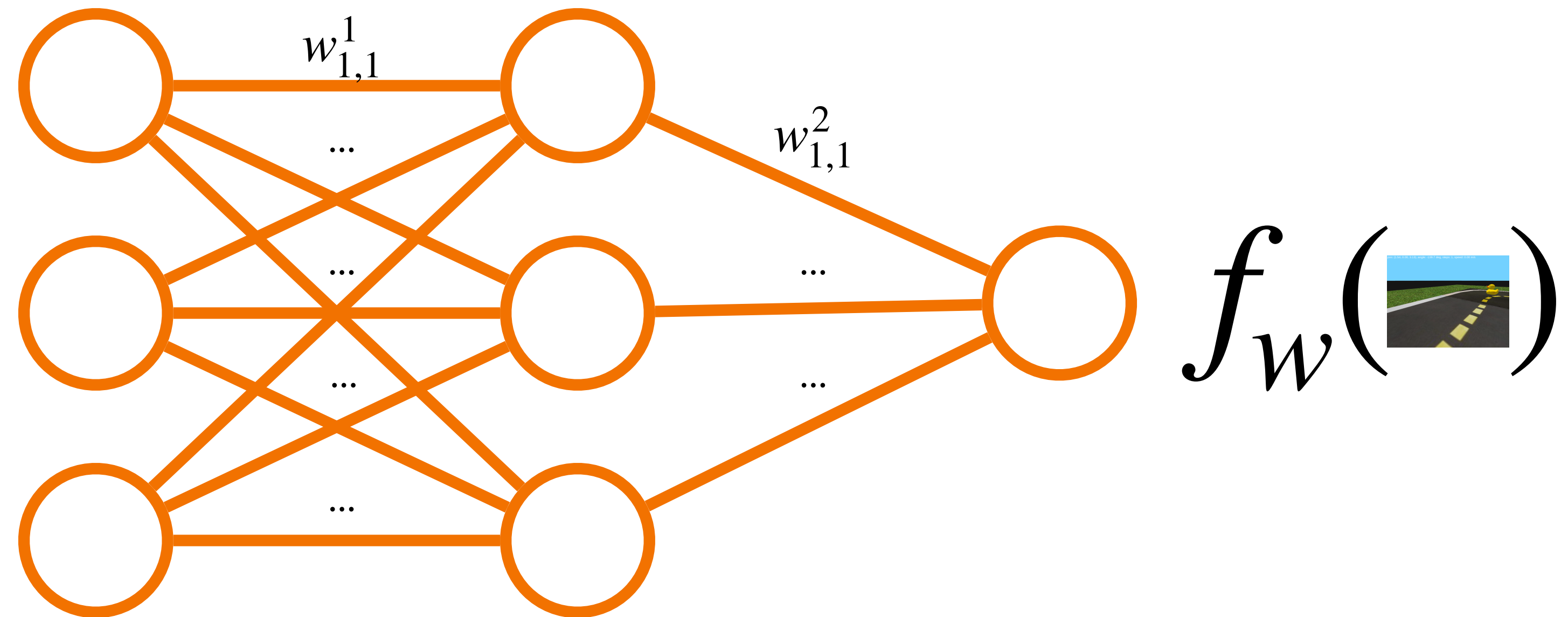


Imagenet

- 14 million images
- thousands of categories



Loss functions



( , ✓)

$\mathcal{L}(\quad , \quad)$

Loss functions

$$f_w(\text{img}) = \times \longrightarrow \mathcal{L} \uparrow$$

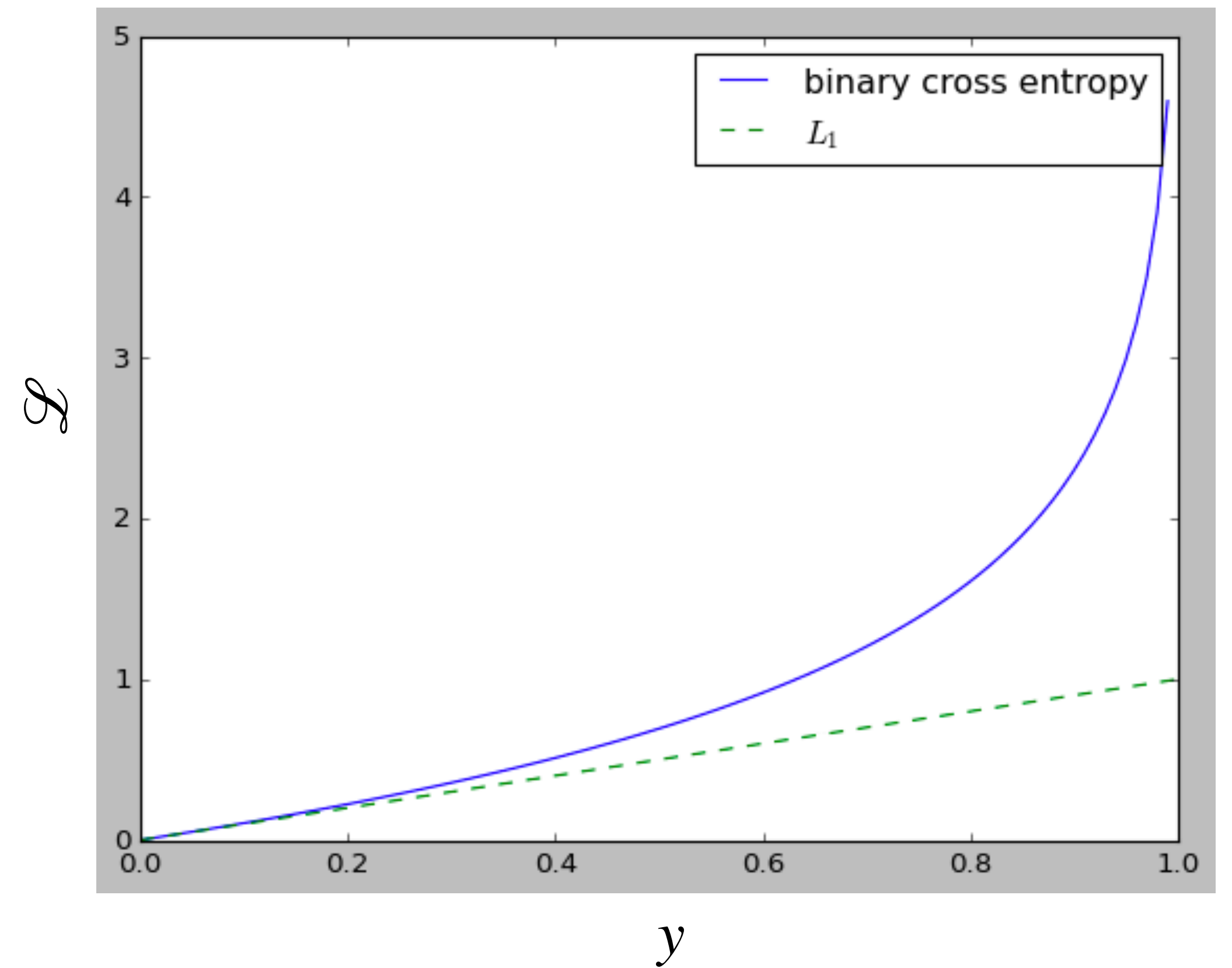
$$f_w(\text{img}) = \checkmark \longrightarrow \mathcal{L} \downarrow$$

Loss functions

The L_1 distance: $|y - y^*|$

The binary cross-entropy loss:

$$-(y^* \log(y) + (1 - y^*) \log(1 - y))$$

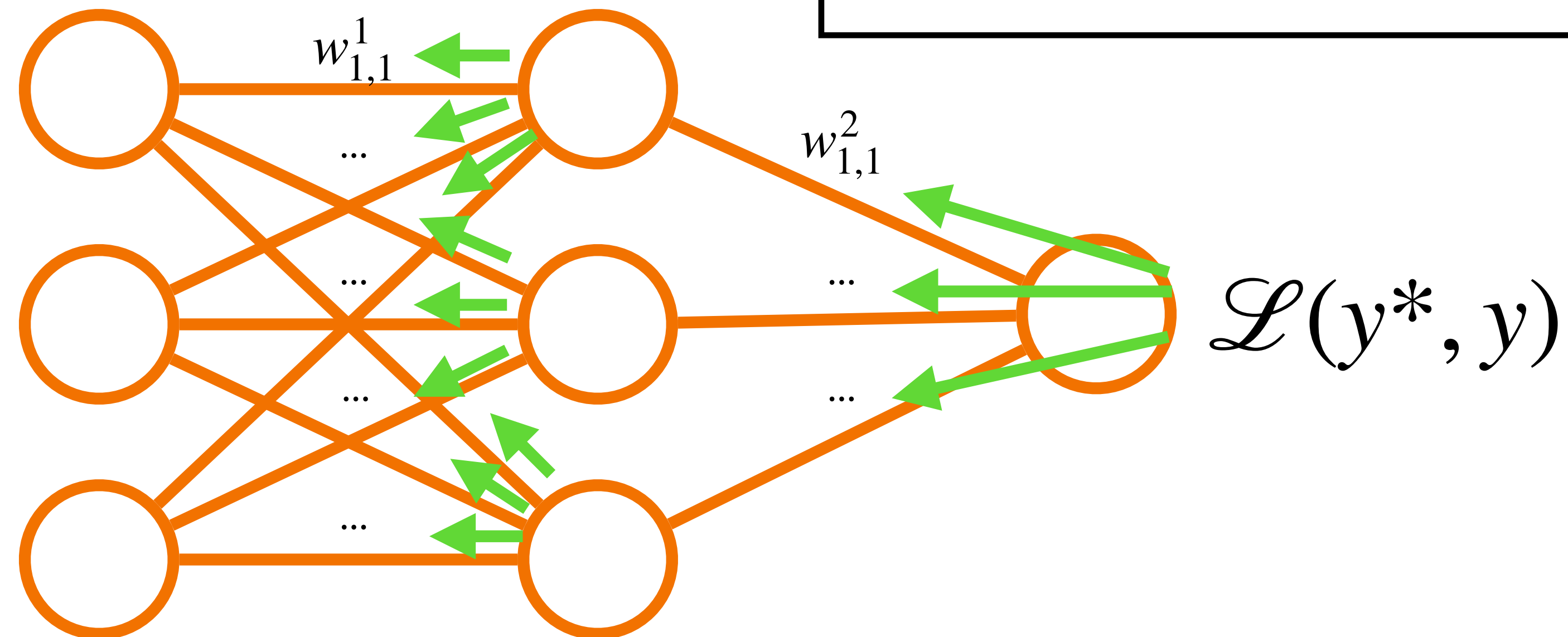


y^* : annotation

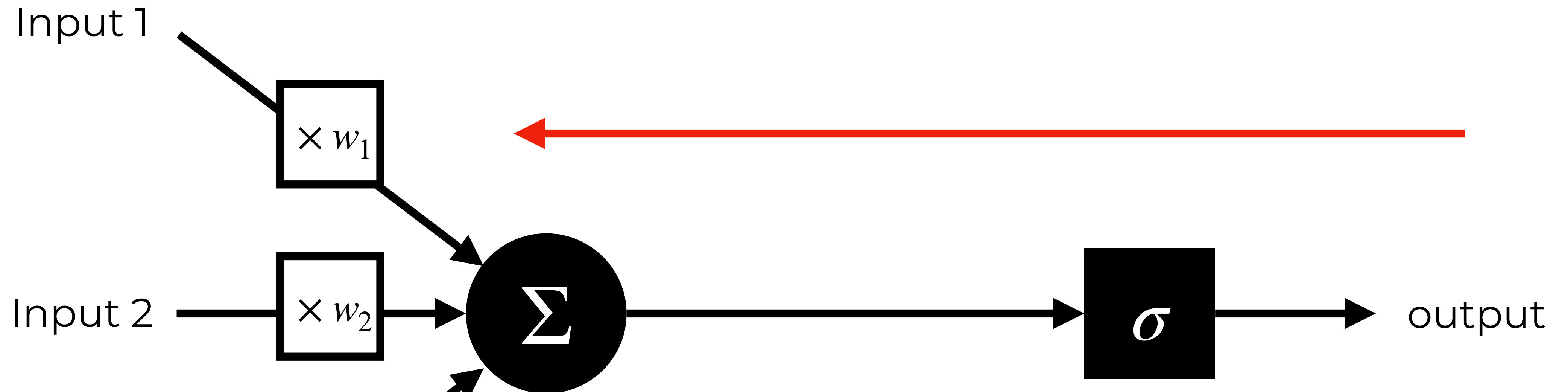
y : model prediction

Backpropagation

$$w \leftarrow w + \eta \frac{\partial \mathcal{L}}{\partial w}$$

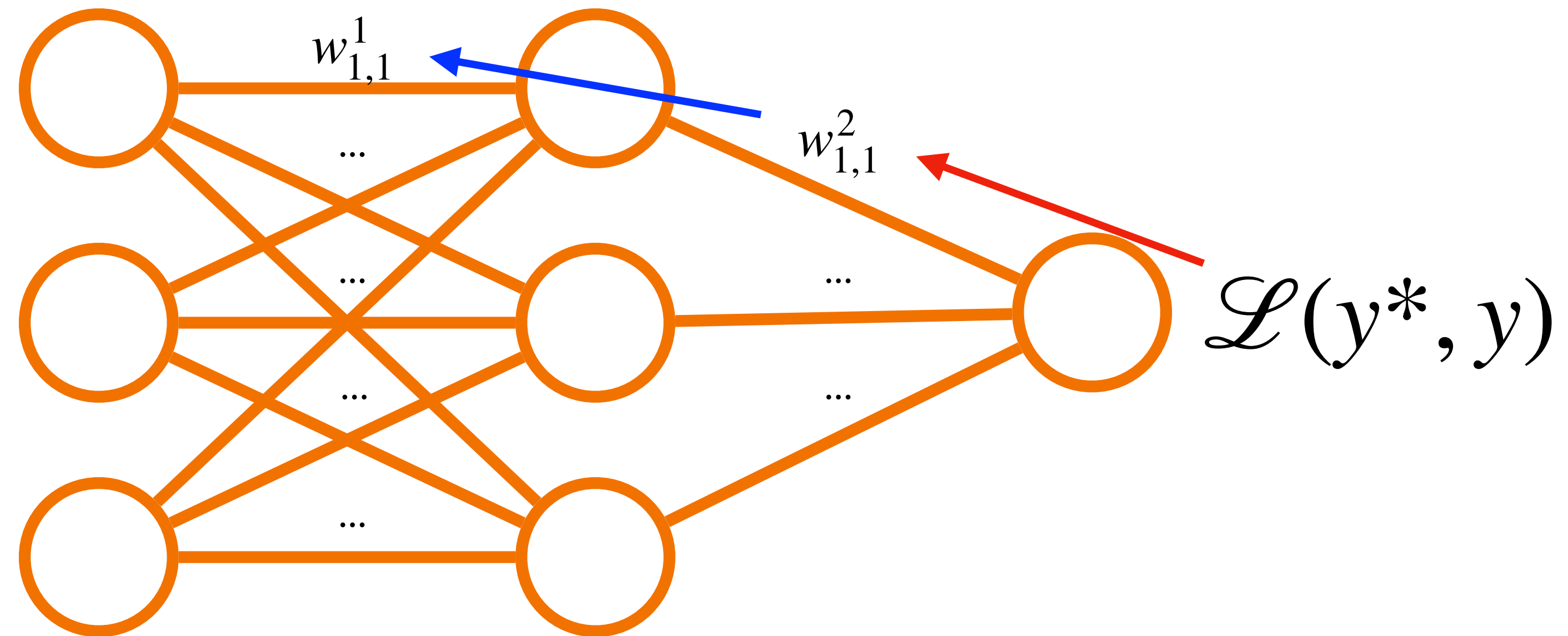


Backpropagation

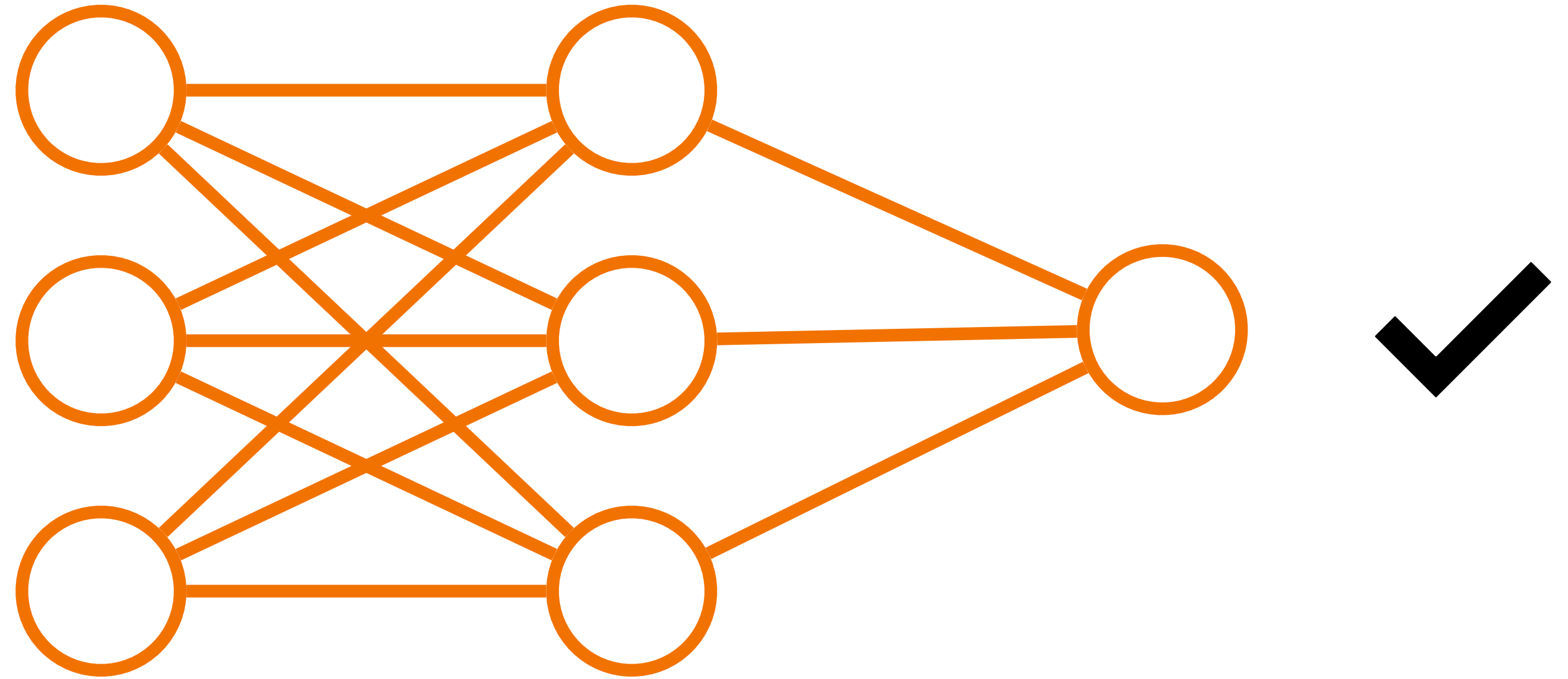


$$\frac{\partial \text{output}}{\partial w} = \frac{\partial \text{output}}{\partial \sigma} \frac{\partial \sigma}{\partial w}$$

Backpropagation



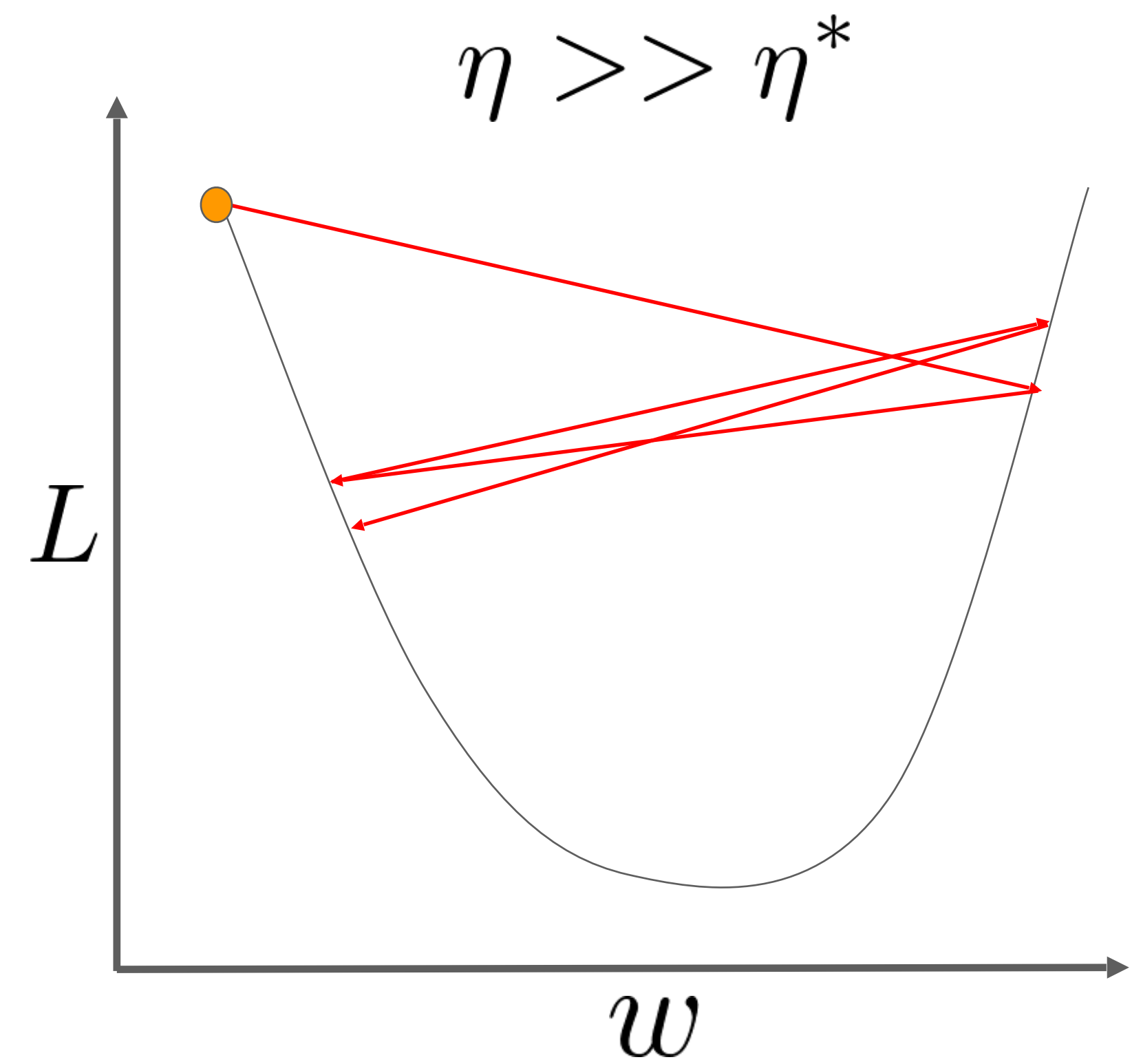
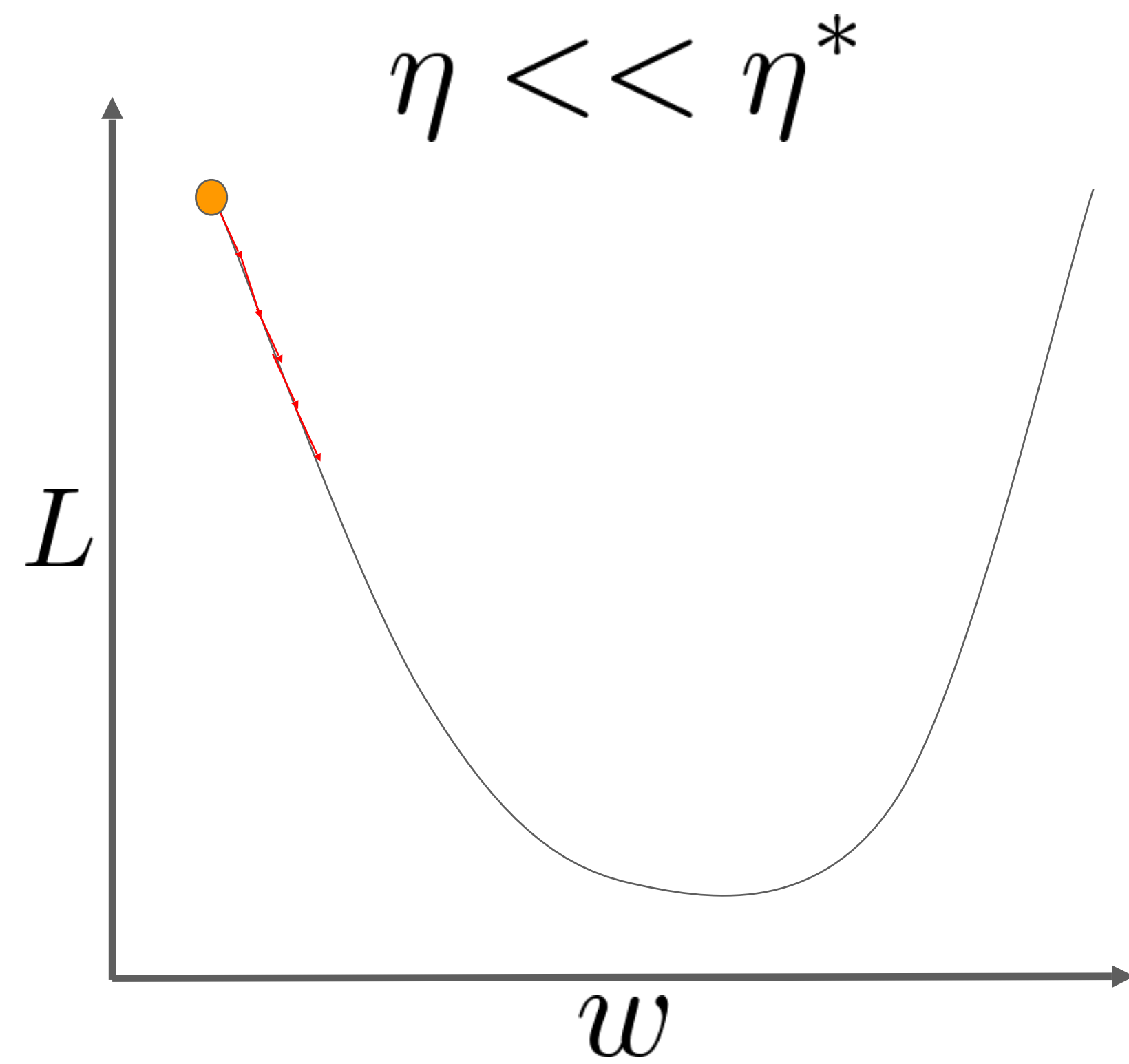
$$\frac{\partial \mathcal{L}}{\partial w_{1,1}^1} = \frac{\partial \mathcal{L}}{\partial w_{1,1}^2} \frac{\partial w_{1,1}^2}{\partial w_{1,1}^1}$$



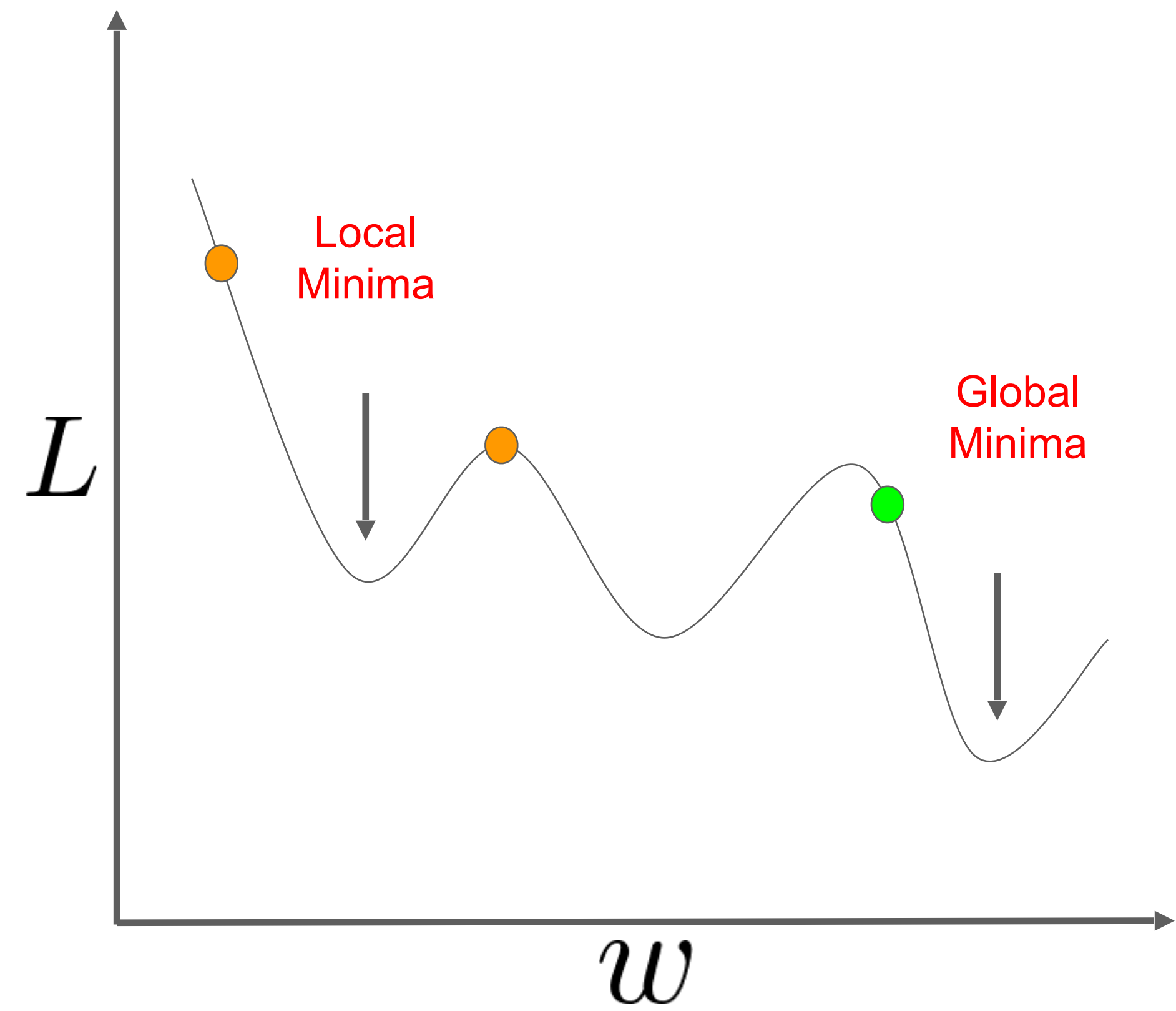
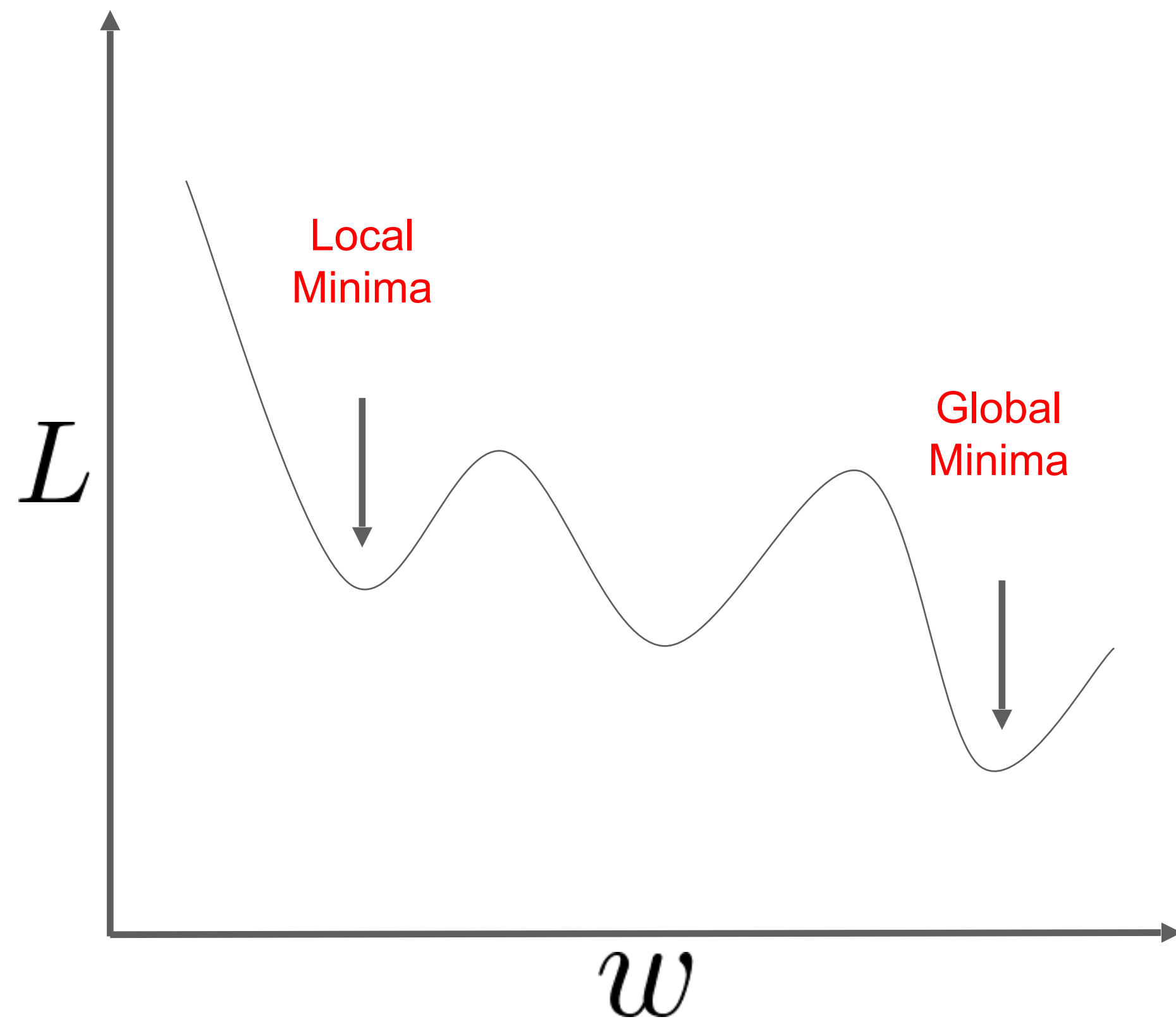
$$w \leftarrow w + \eta \frac{\partial \mathcal{L}}{\partial w}$$

step size

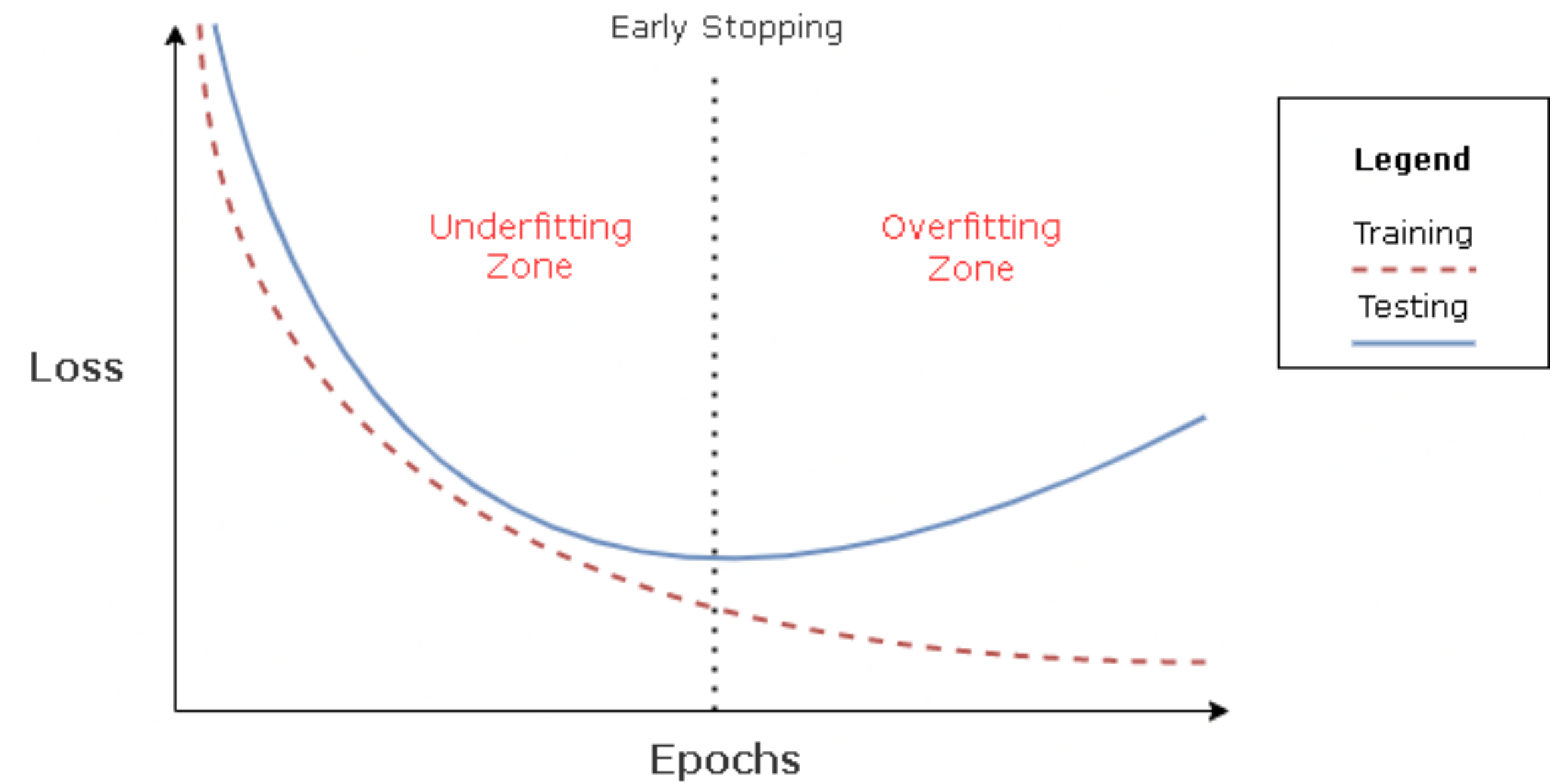
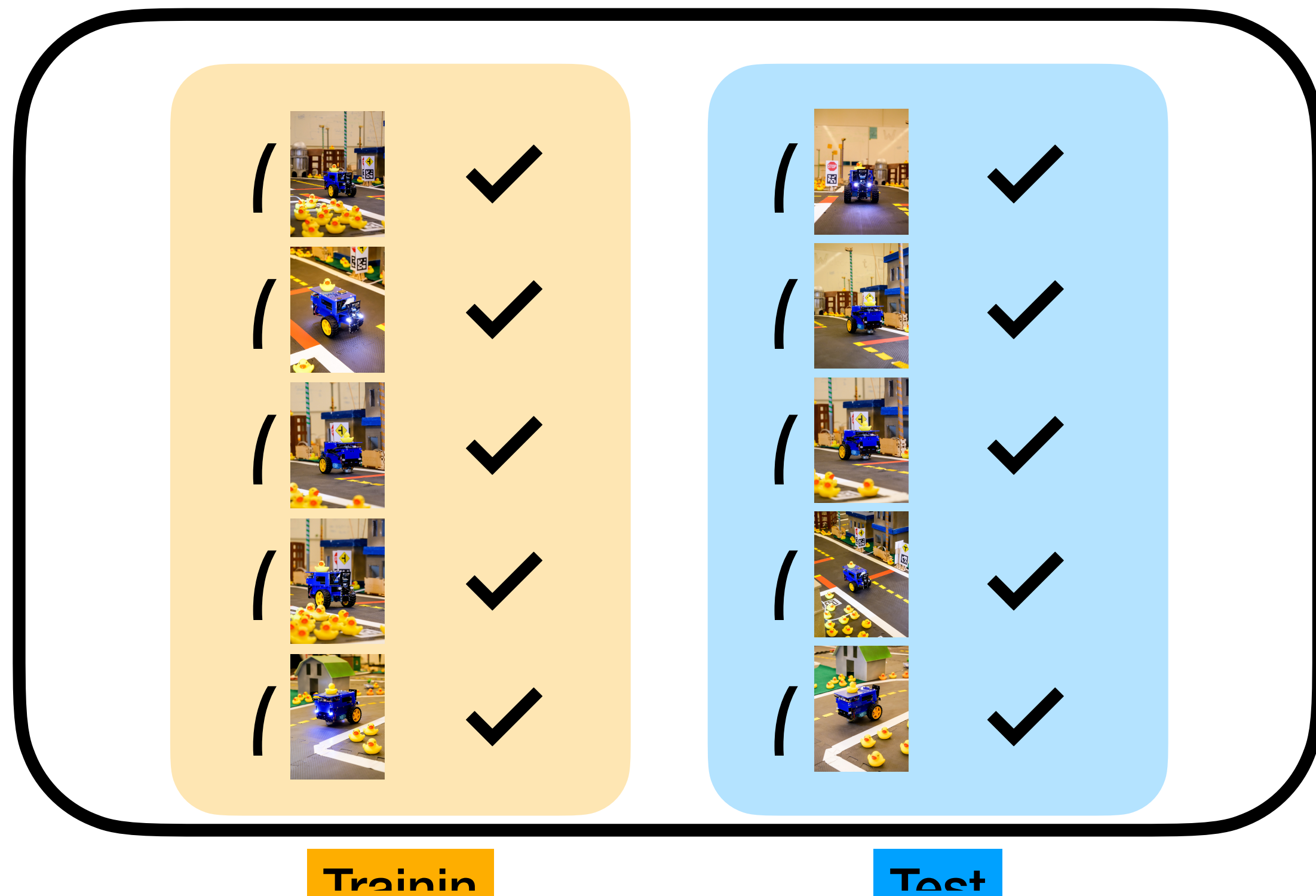
The step size



Convexity

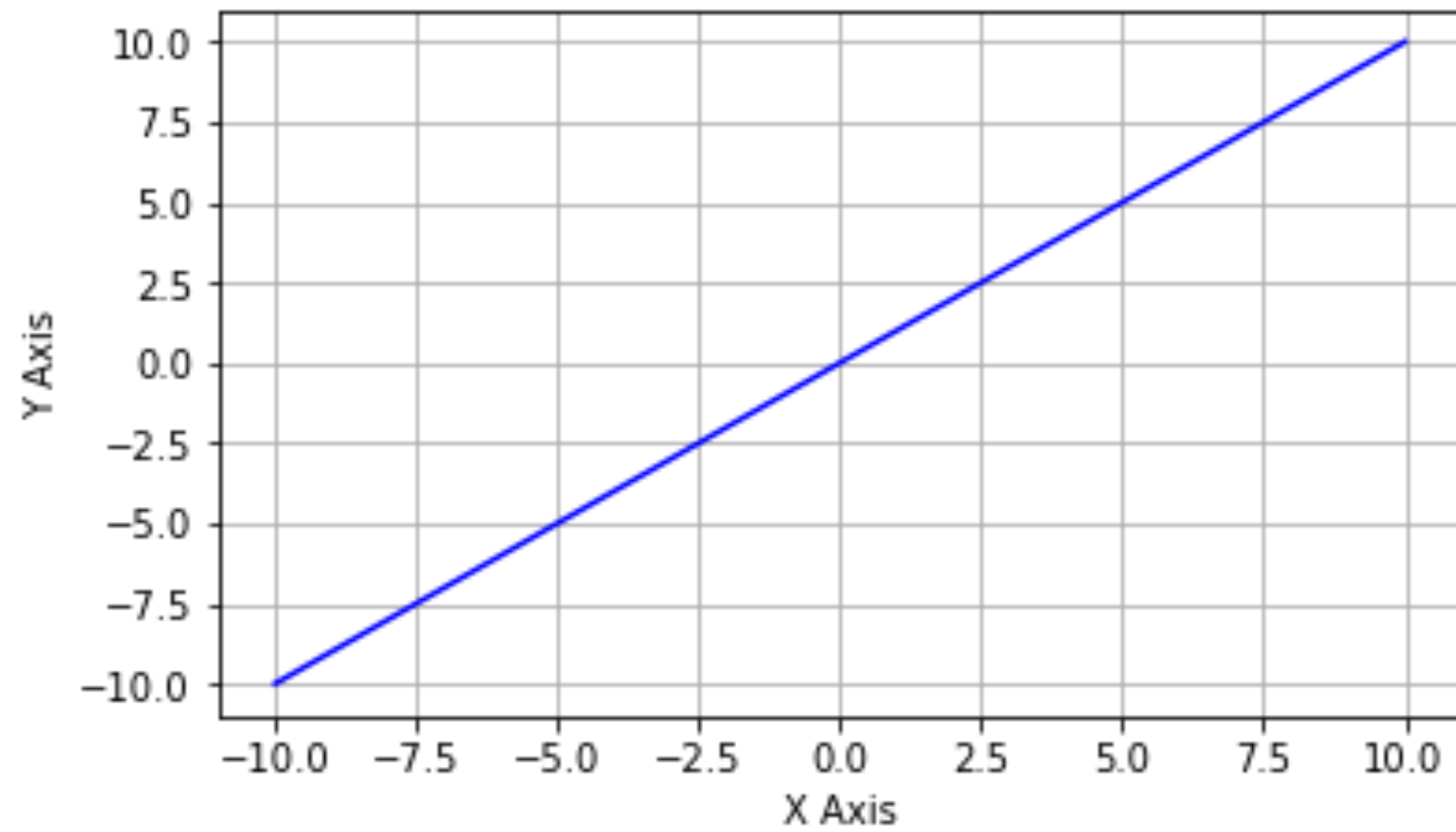


Annotated Dataset

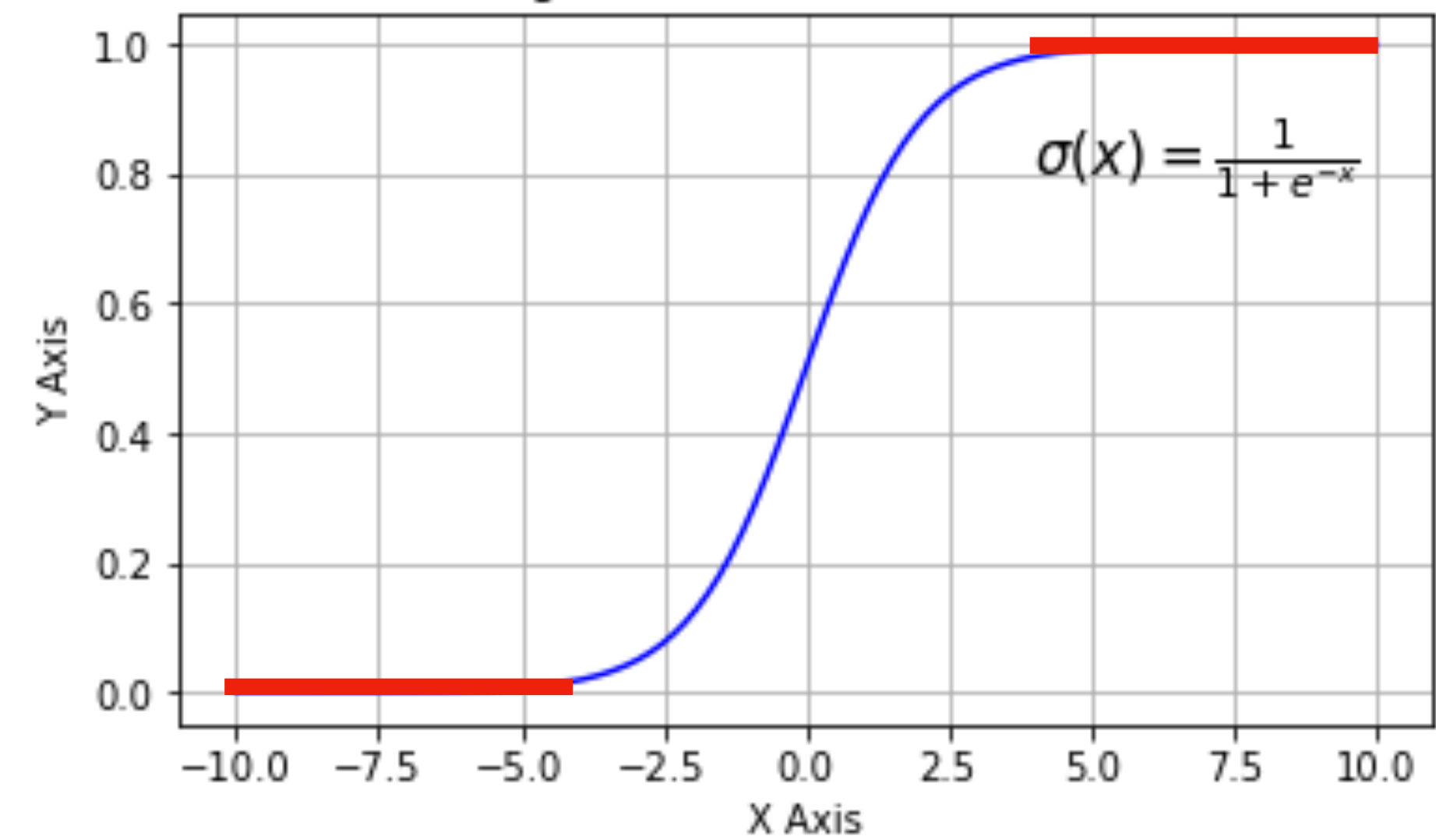


Activation functions

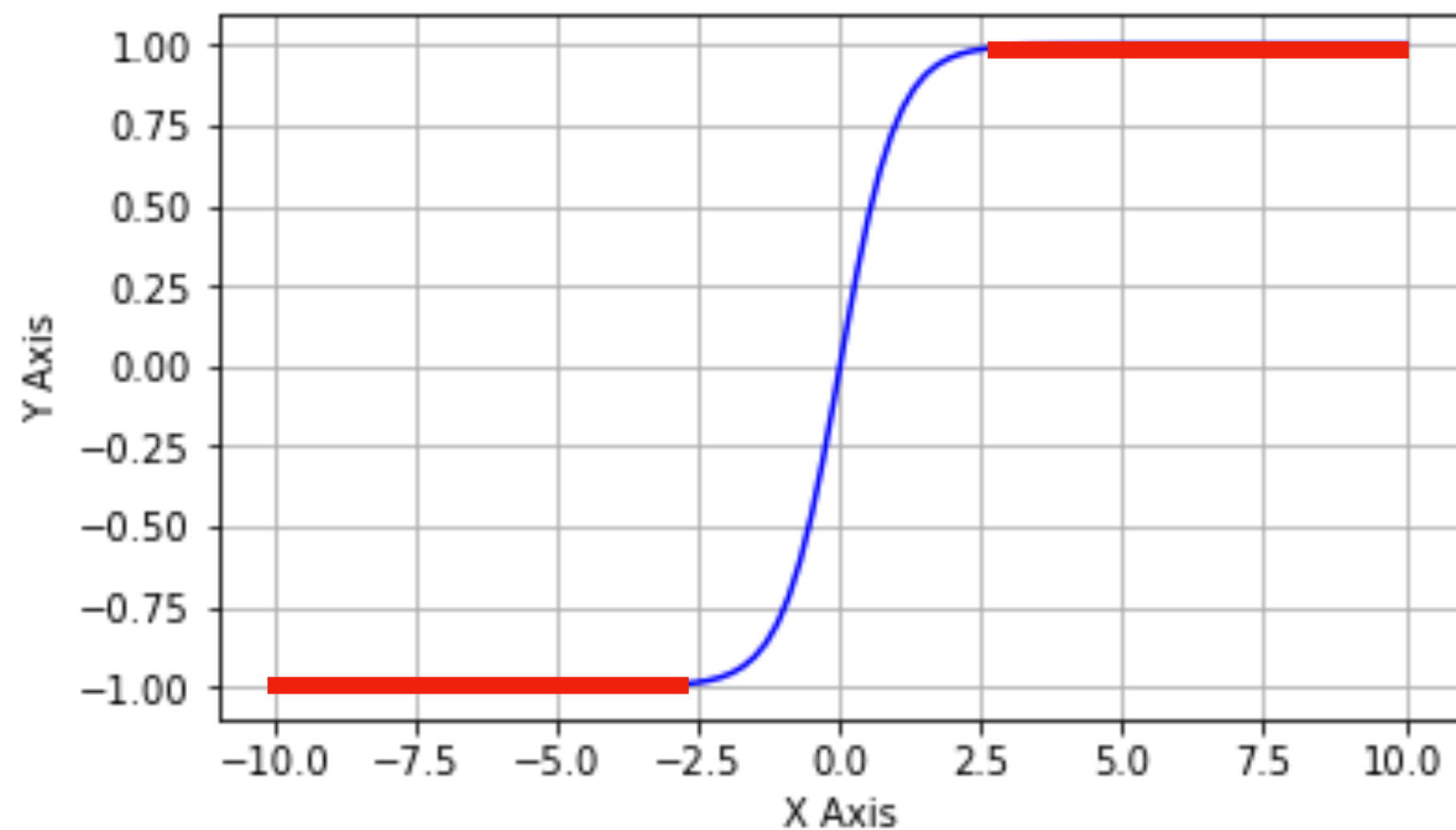
Linear Activation Function



Sigmoid Activation Function



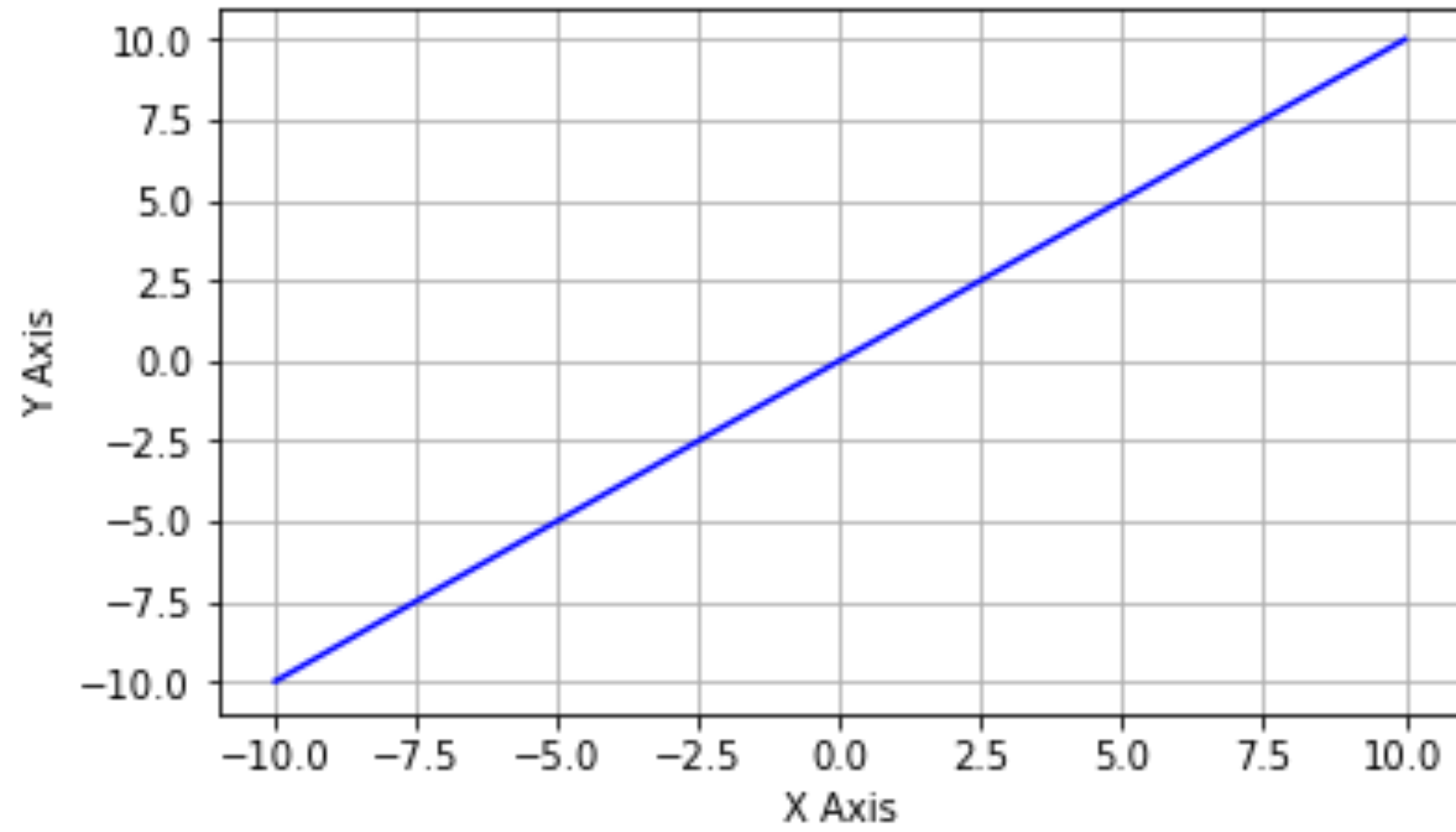
Tanh Activation Function



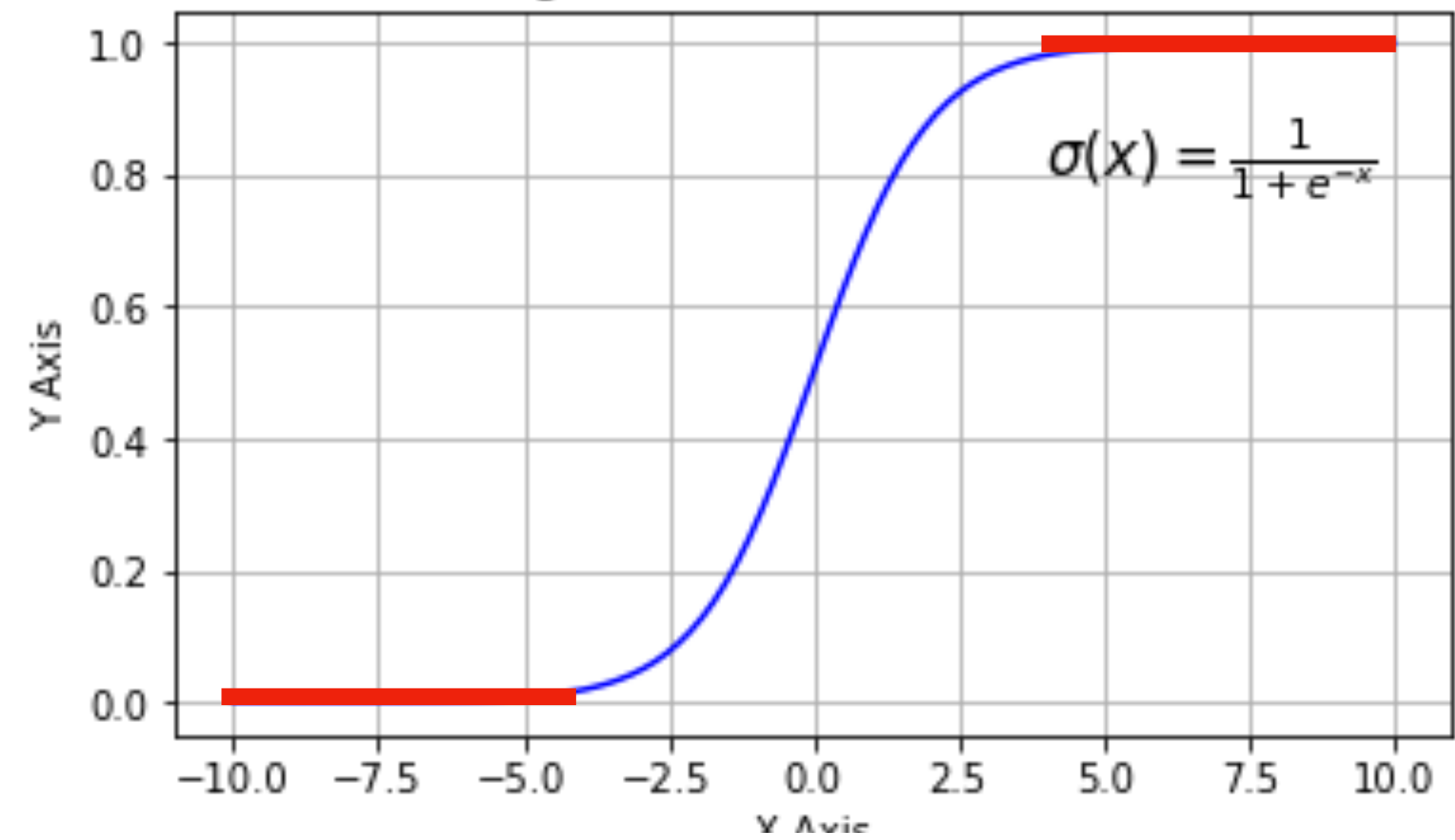
Saturation -> Vanishing Gradients

Activation functions

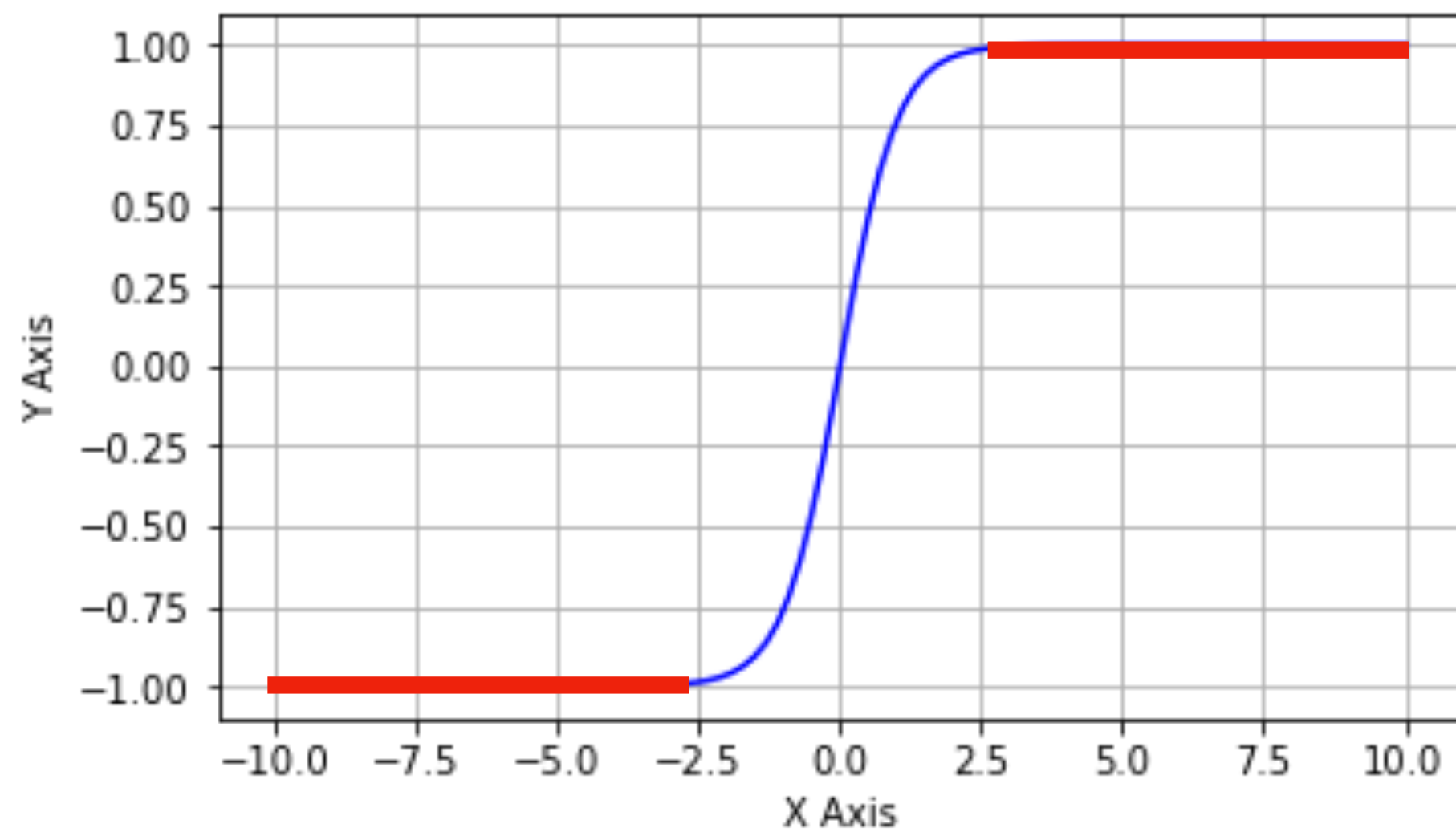
Linear Activation Function



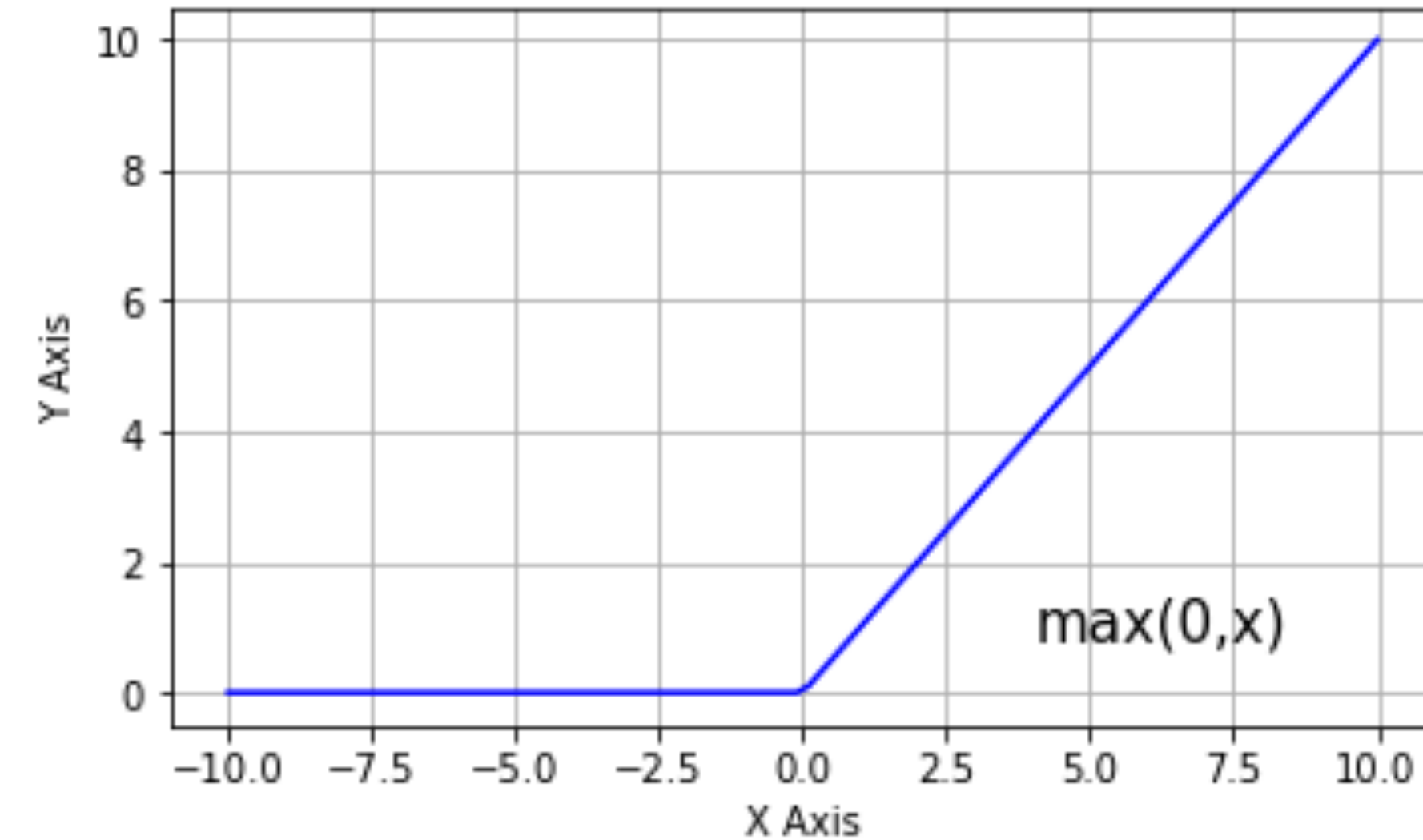
Sigmoid Activation Function



Tanh Activation Function



ReLU Activation Function



Advanced Visual Perception

Table of Contents

Intro to Advanced Visual Perception

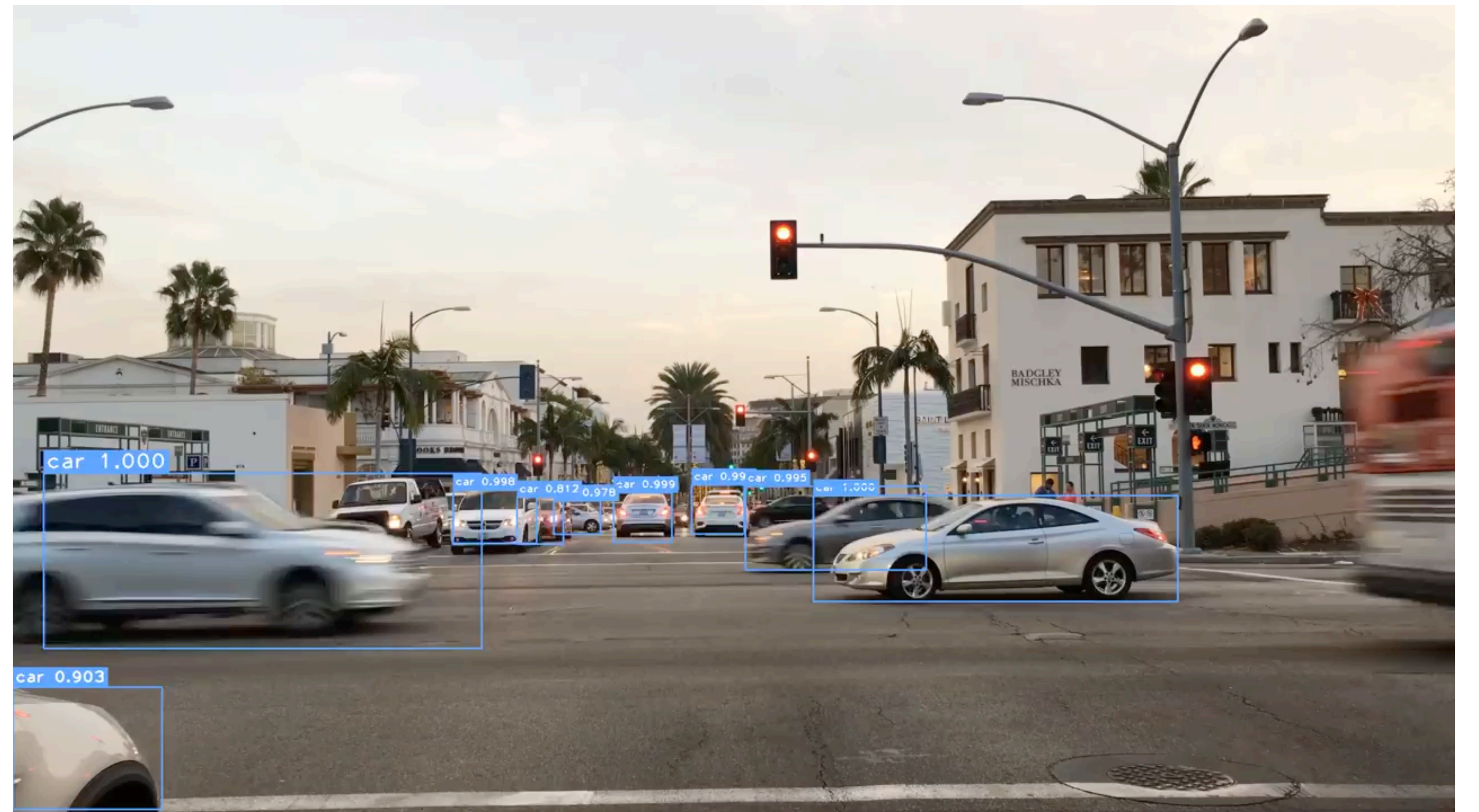
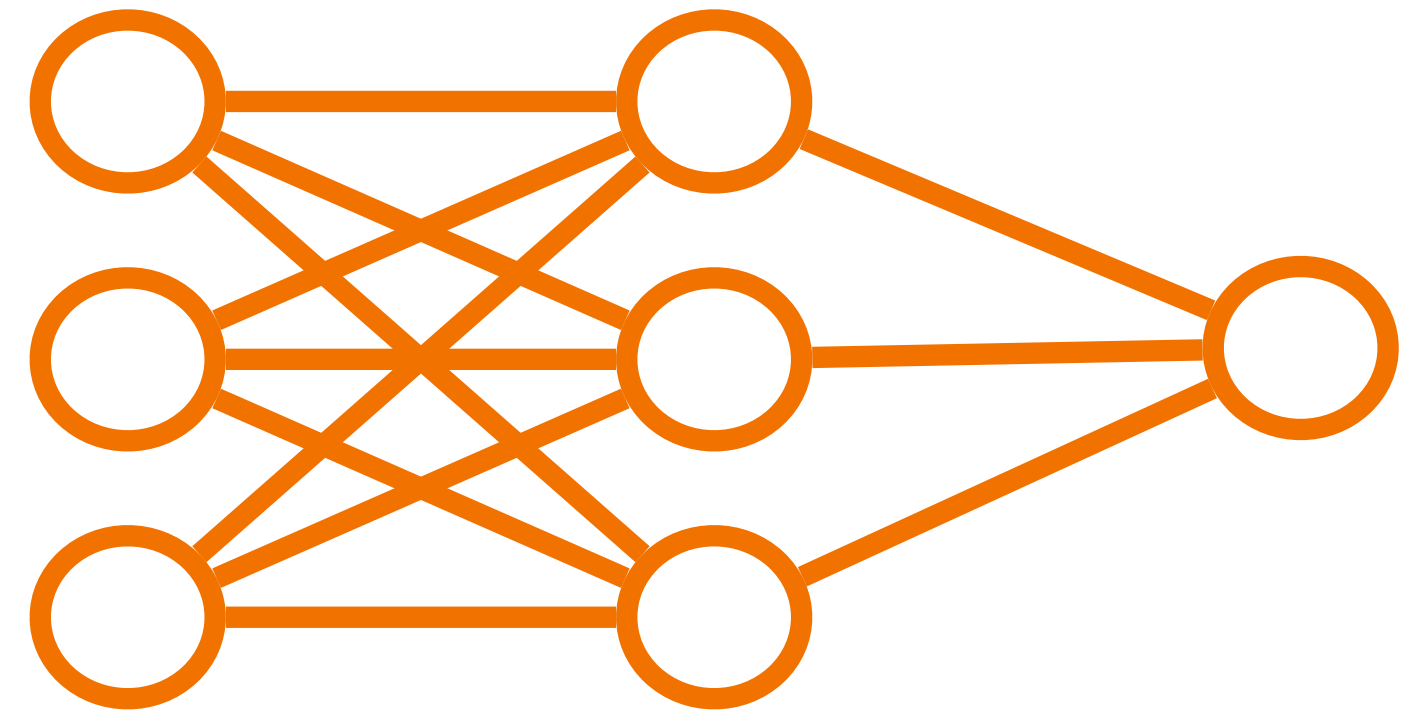
Intro to Neural Networks

Deep Convolutional Neural Networks

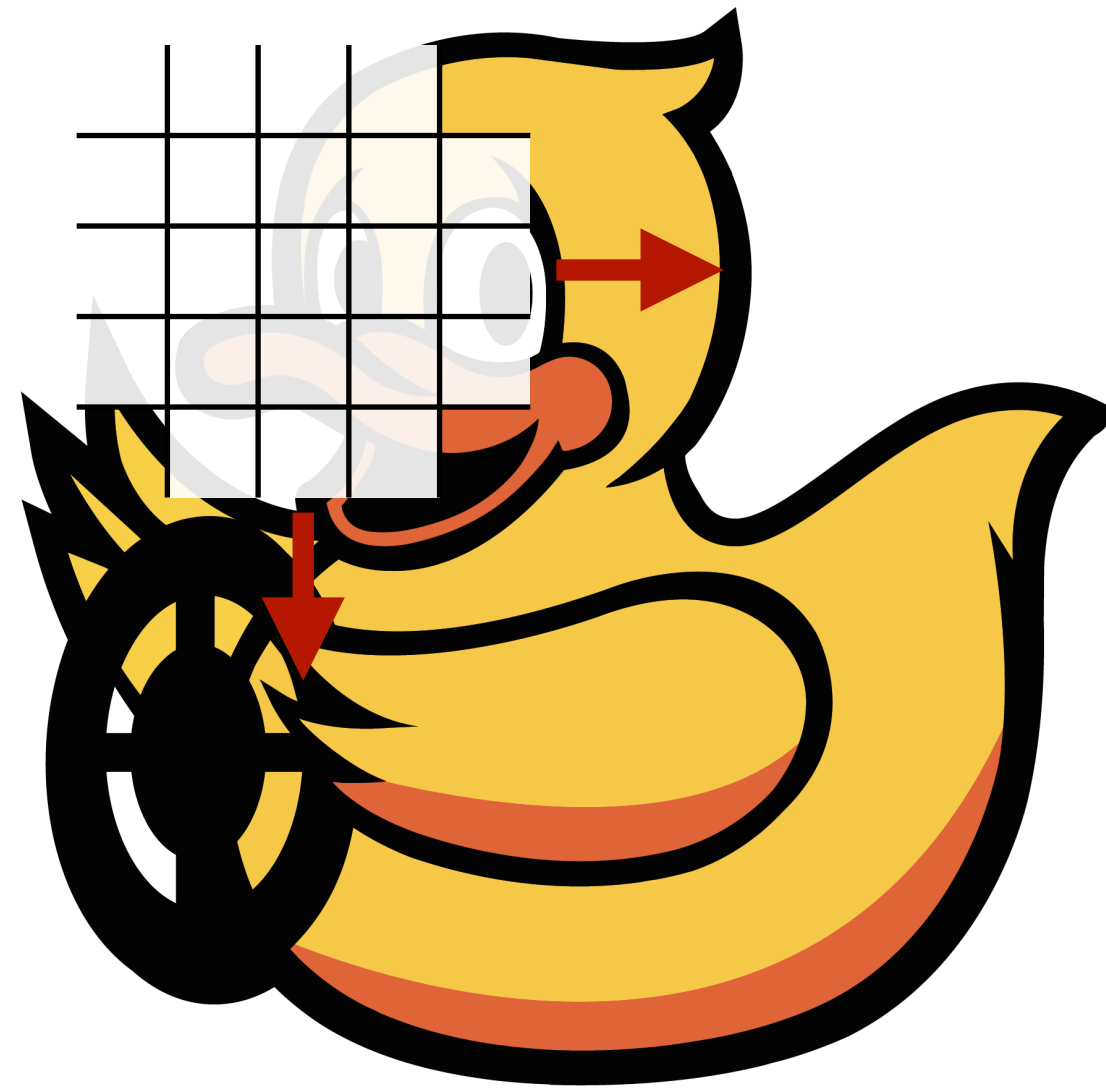
Object Detection

Semantic Segmentation

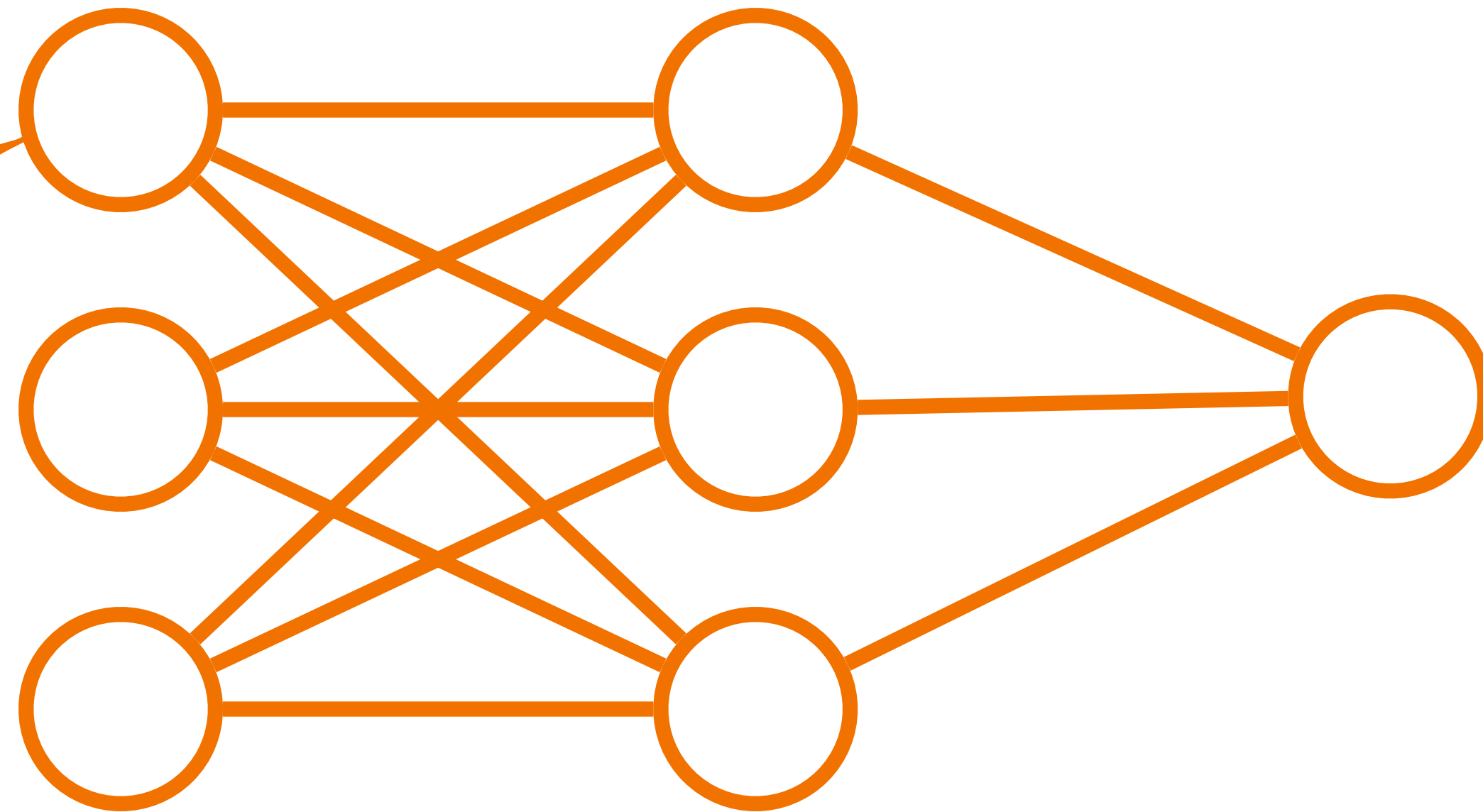
What if we don't have labels?



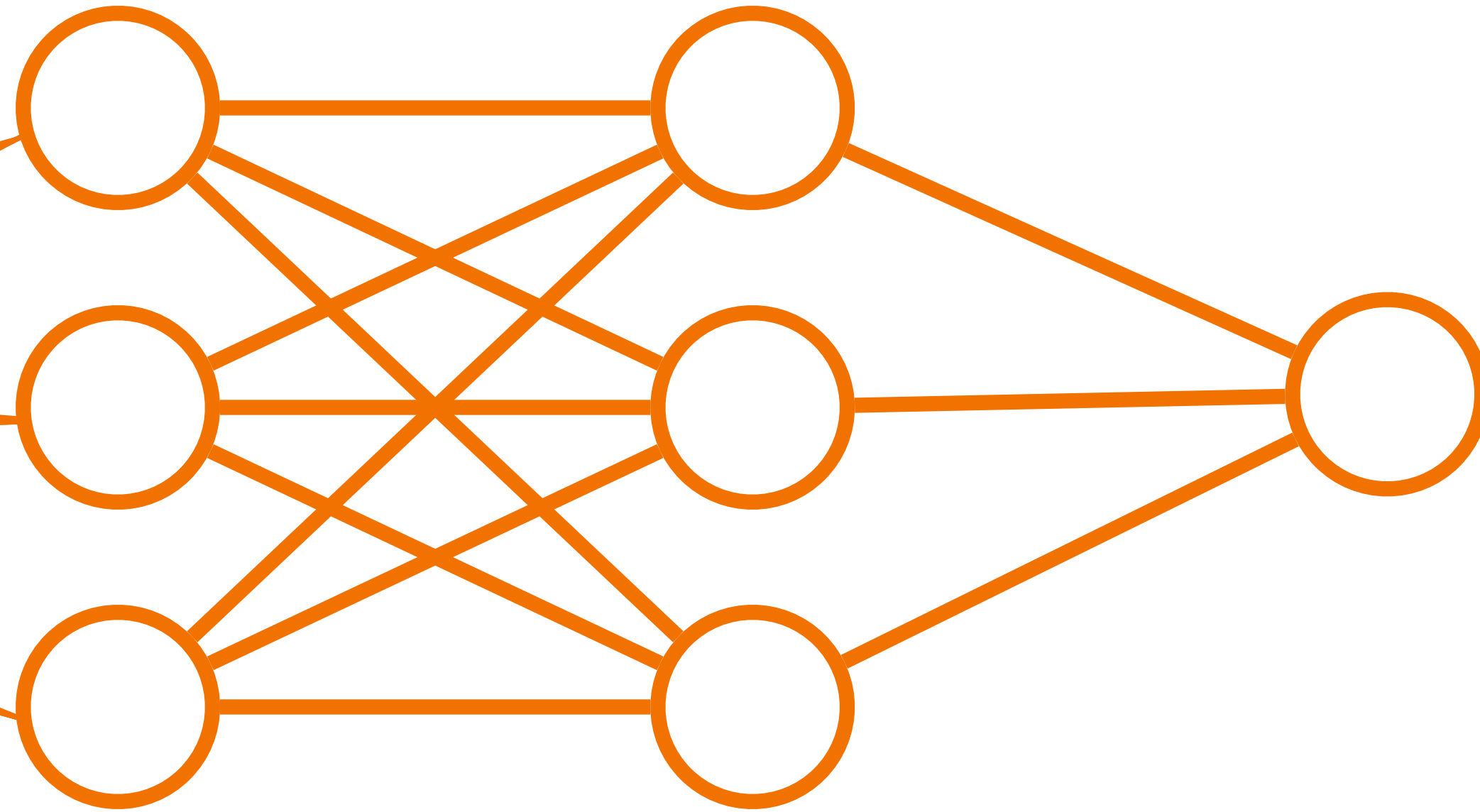
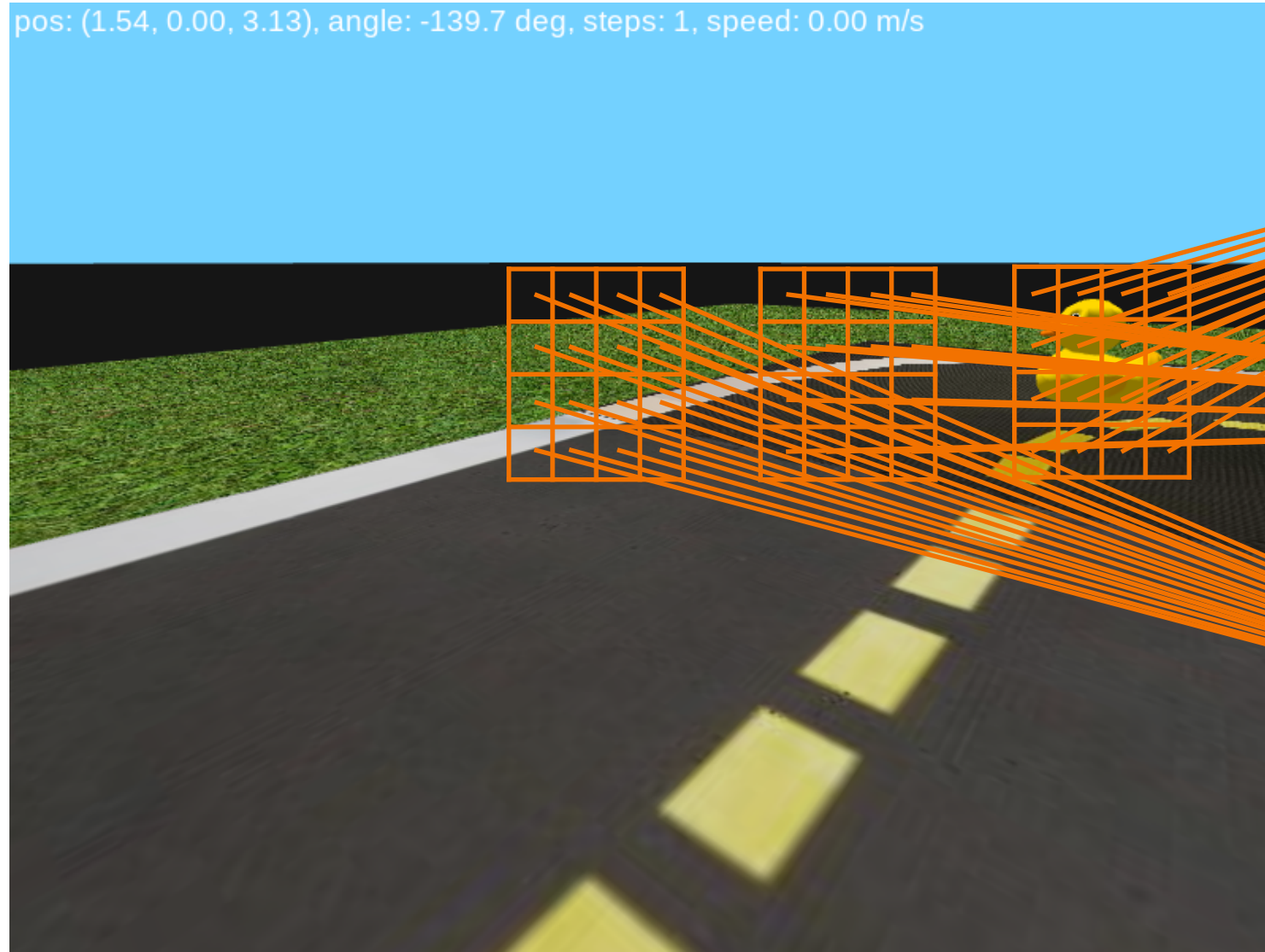
Convolution



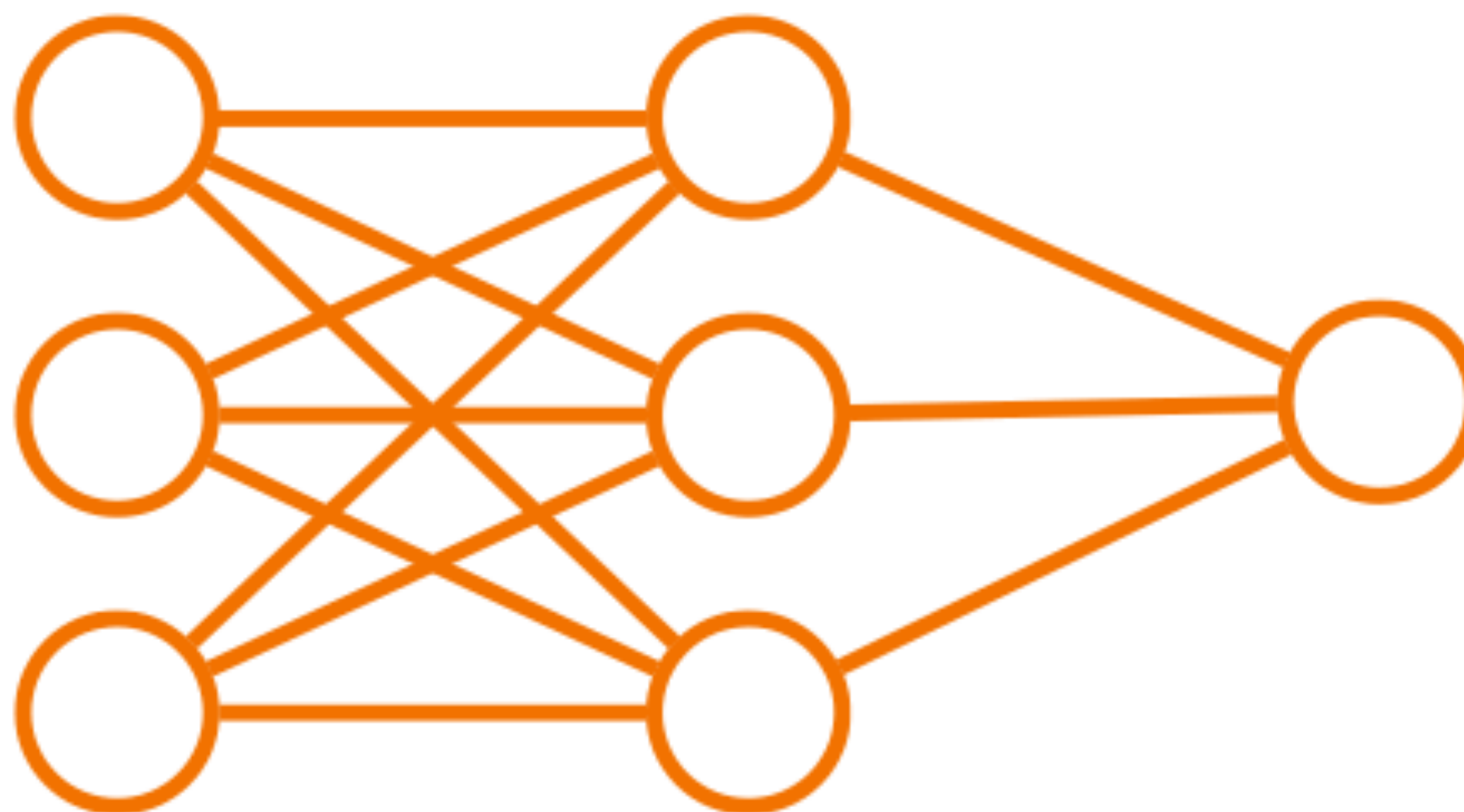
Convolutional layers



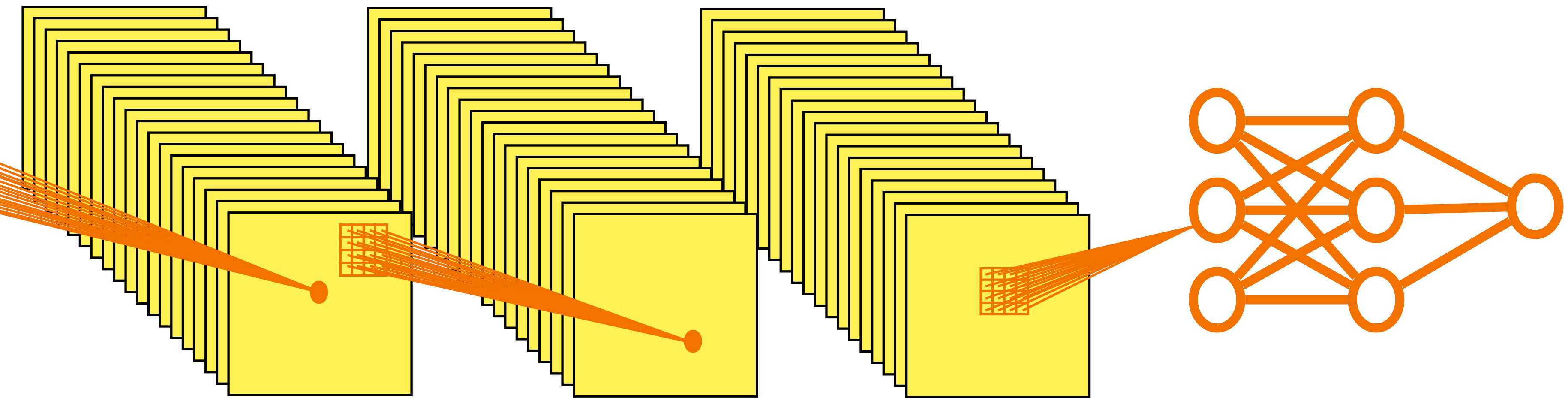
Convolutional layers



Convolutional layers



Convolutional layers

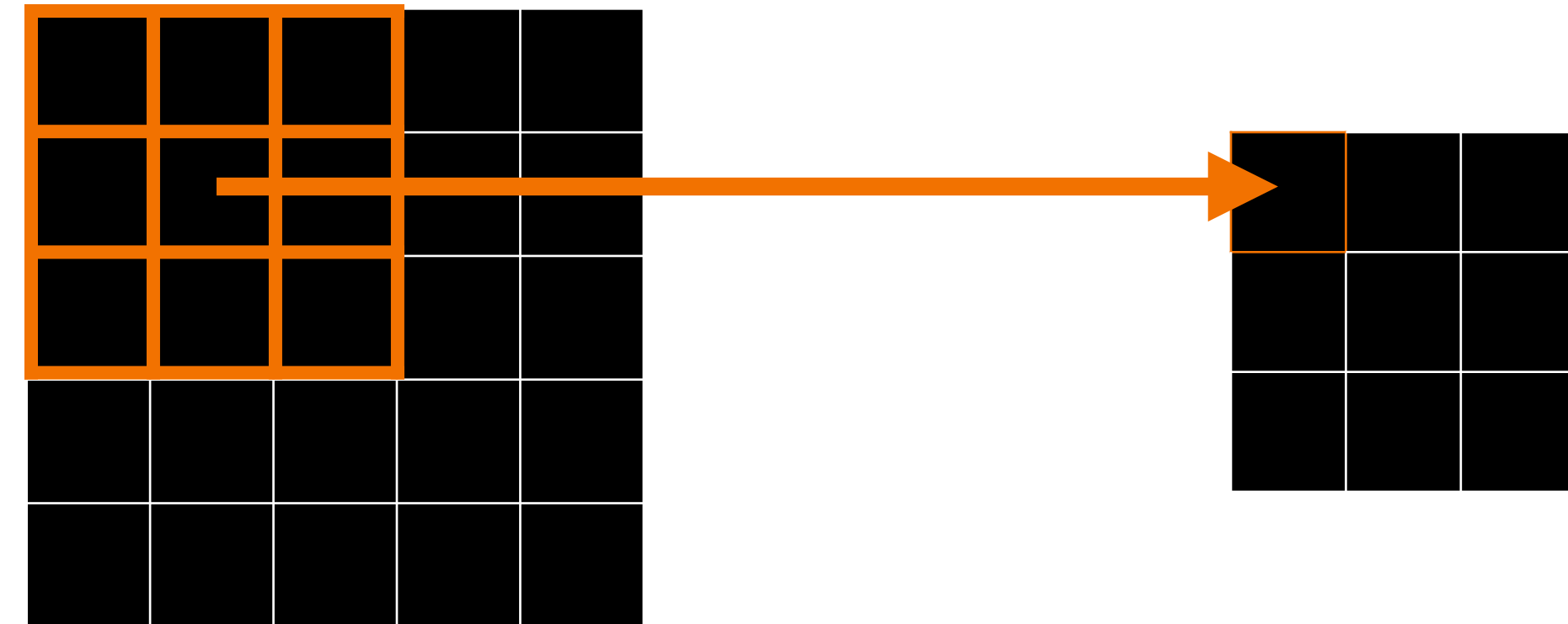


Low-level features

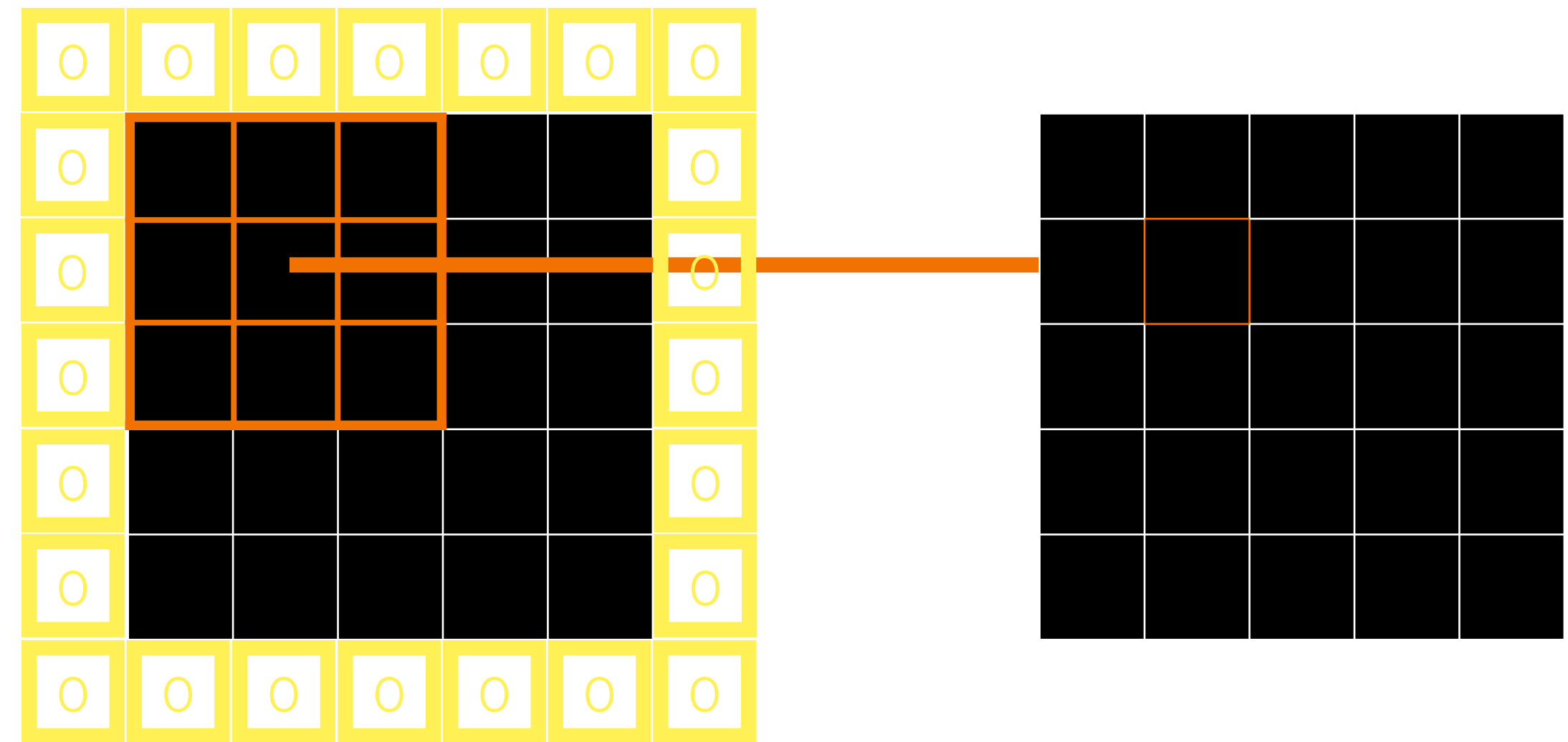
high-level features



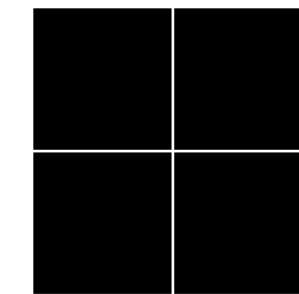
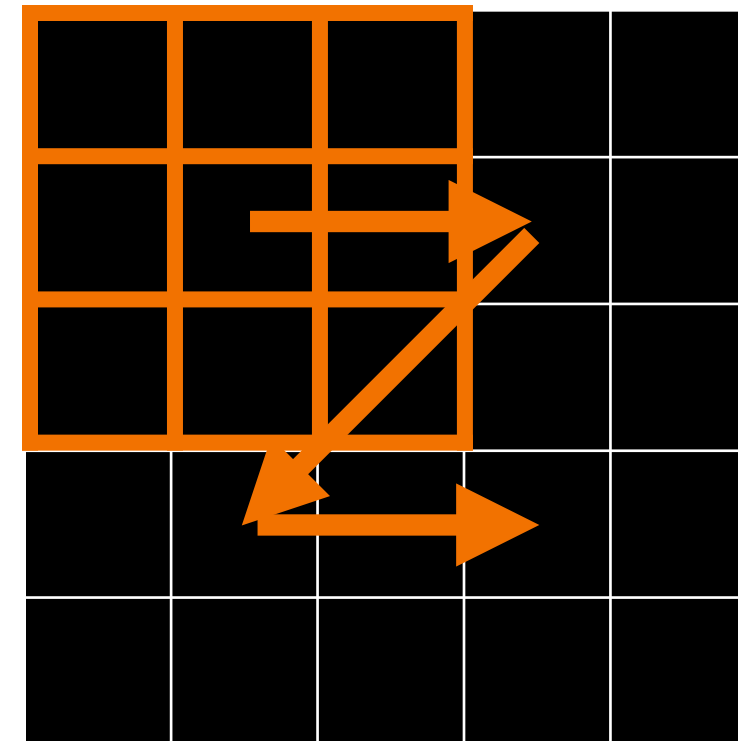
Padding



Padding



Stride



Stride = 2

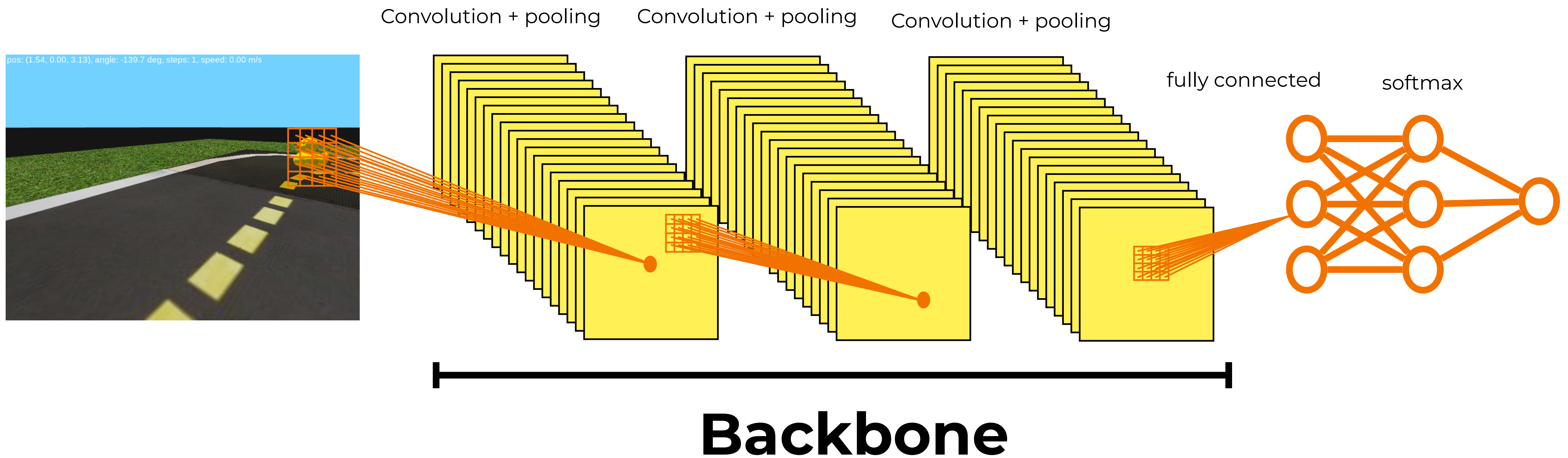
Pooling

12	8	9	0
15	23	9	12
16	12	1	5
3	2	4	13

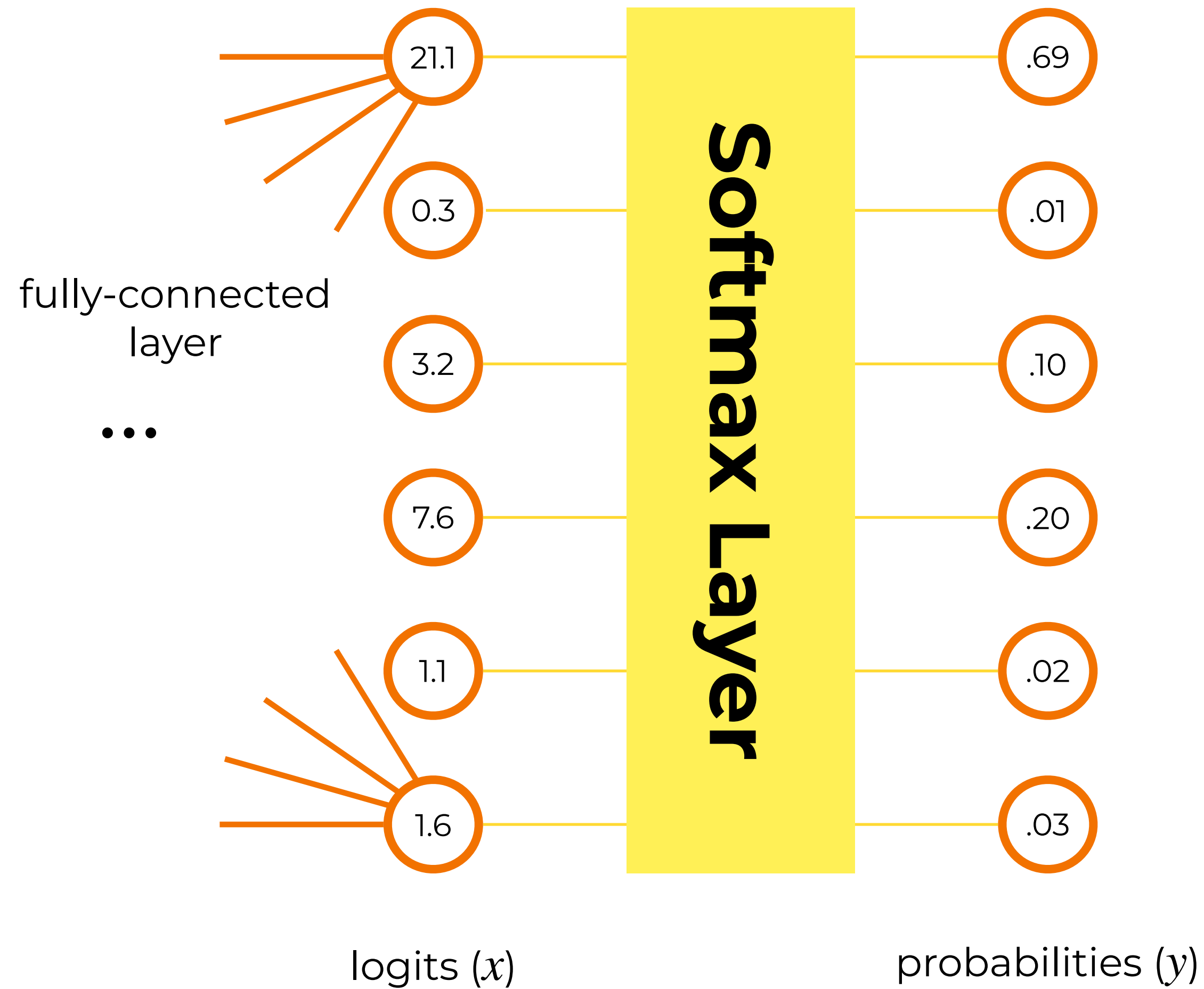
23	12
16	13

Stride = 2

Convolutional layers



Softmax



$$\text{softmax}(y_j) = \frac{e^{x_j}}{\sum_{k=1}^K e^{x_k}}$$

Cross-entropy loss

$p(\text{duckie})$.69	1
$p(\text{duckiebot})$.01	0
...	.10	0
	.20	0
	.02	0
$p(\text{flamingo})$.03	0

probabilities (y)

probabilities (y^*)



Binary cross-entropy loss:

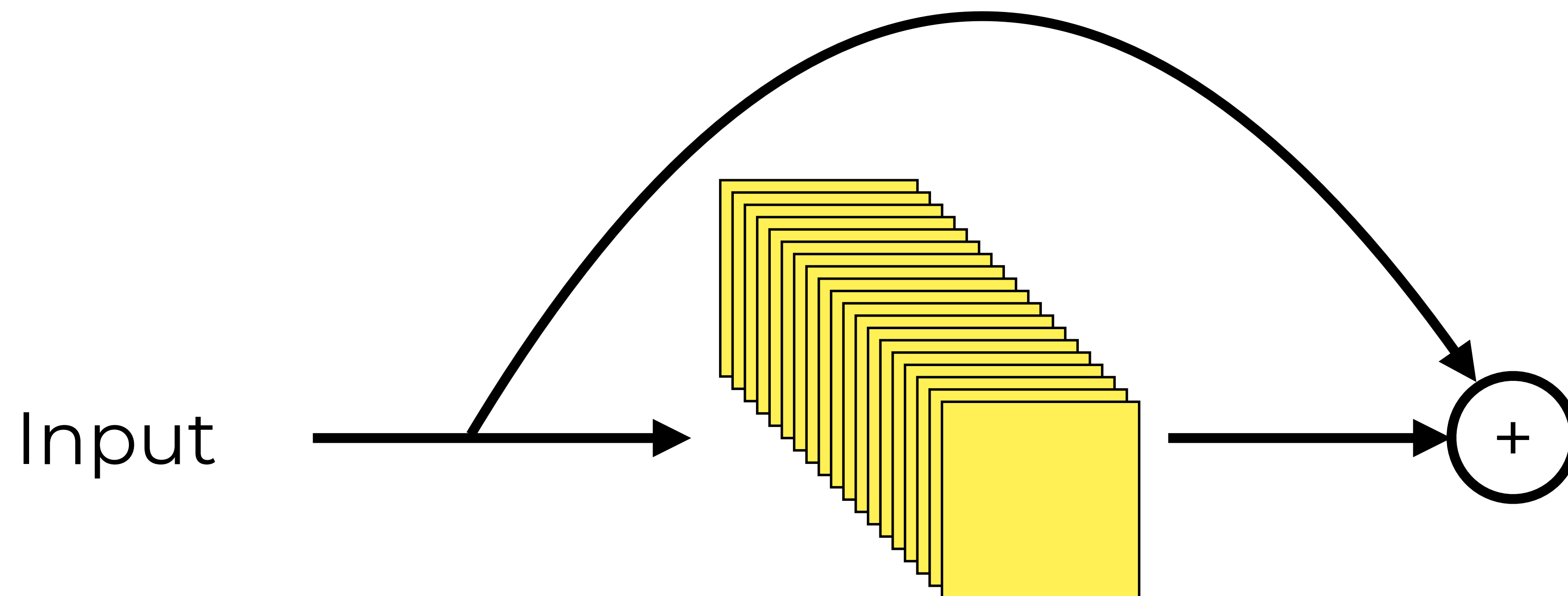
$$-(y^* \log(y) + (1 - y^*) \log(1 - y))$$

Multi-class cross-entropy loss:

$$-\sum_{c=1}^M p(y_c^*) \log p(y_c)$$

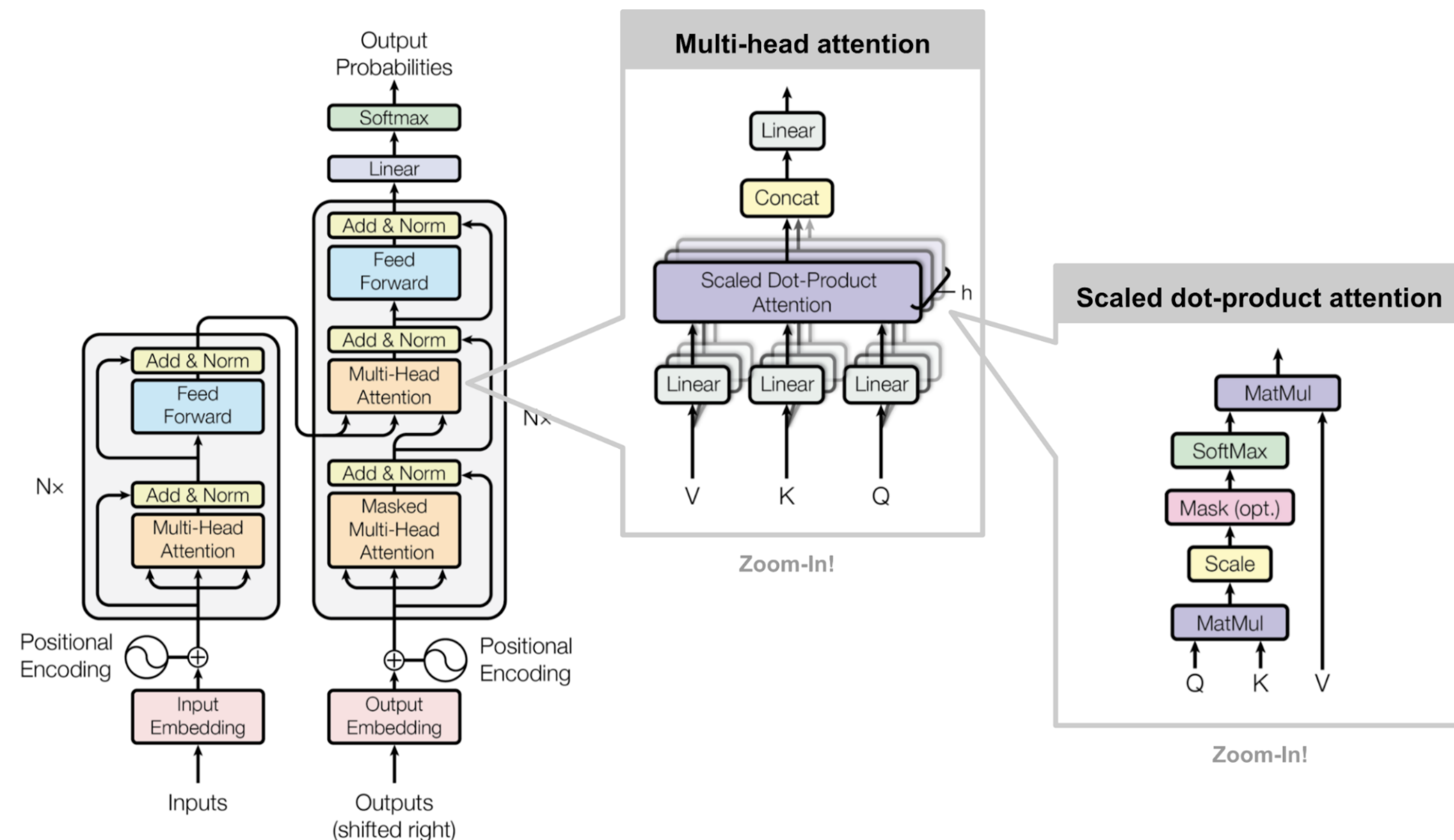
One-hot encoding

Residual layers



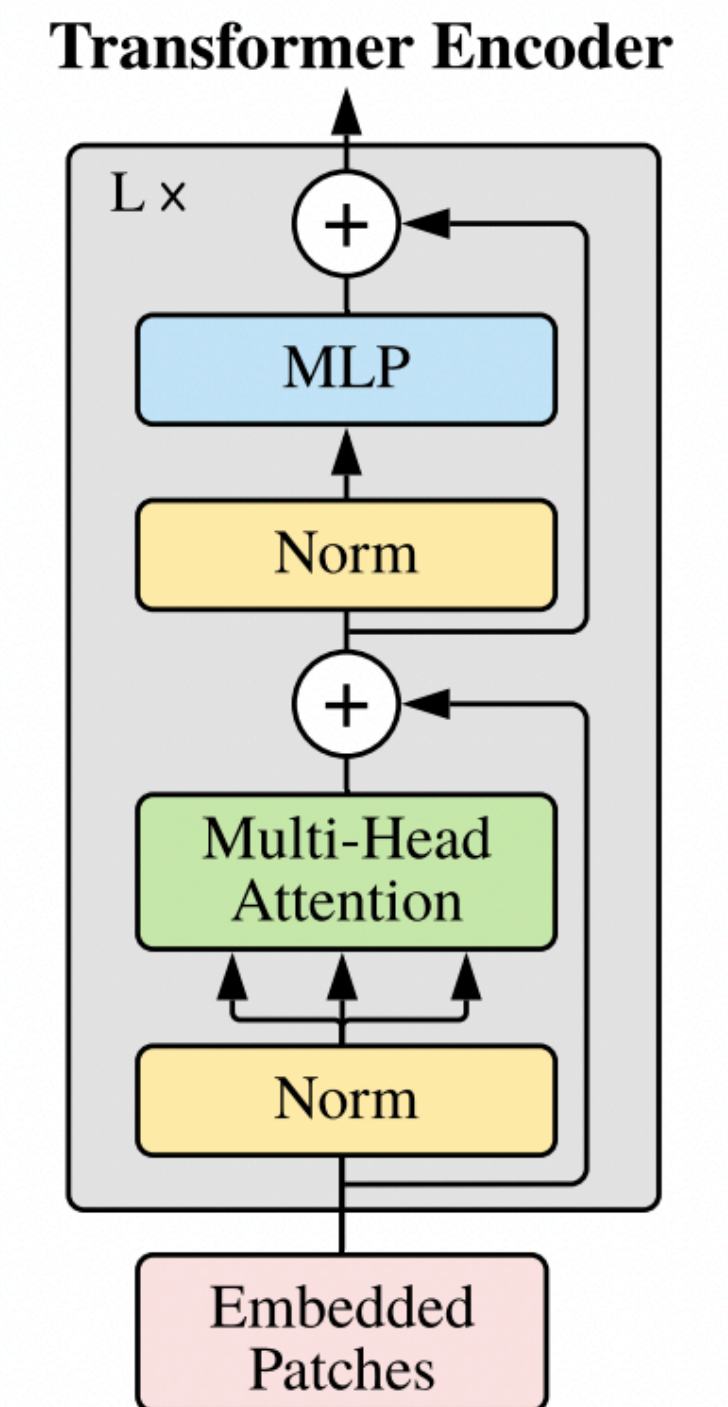
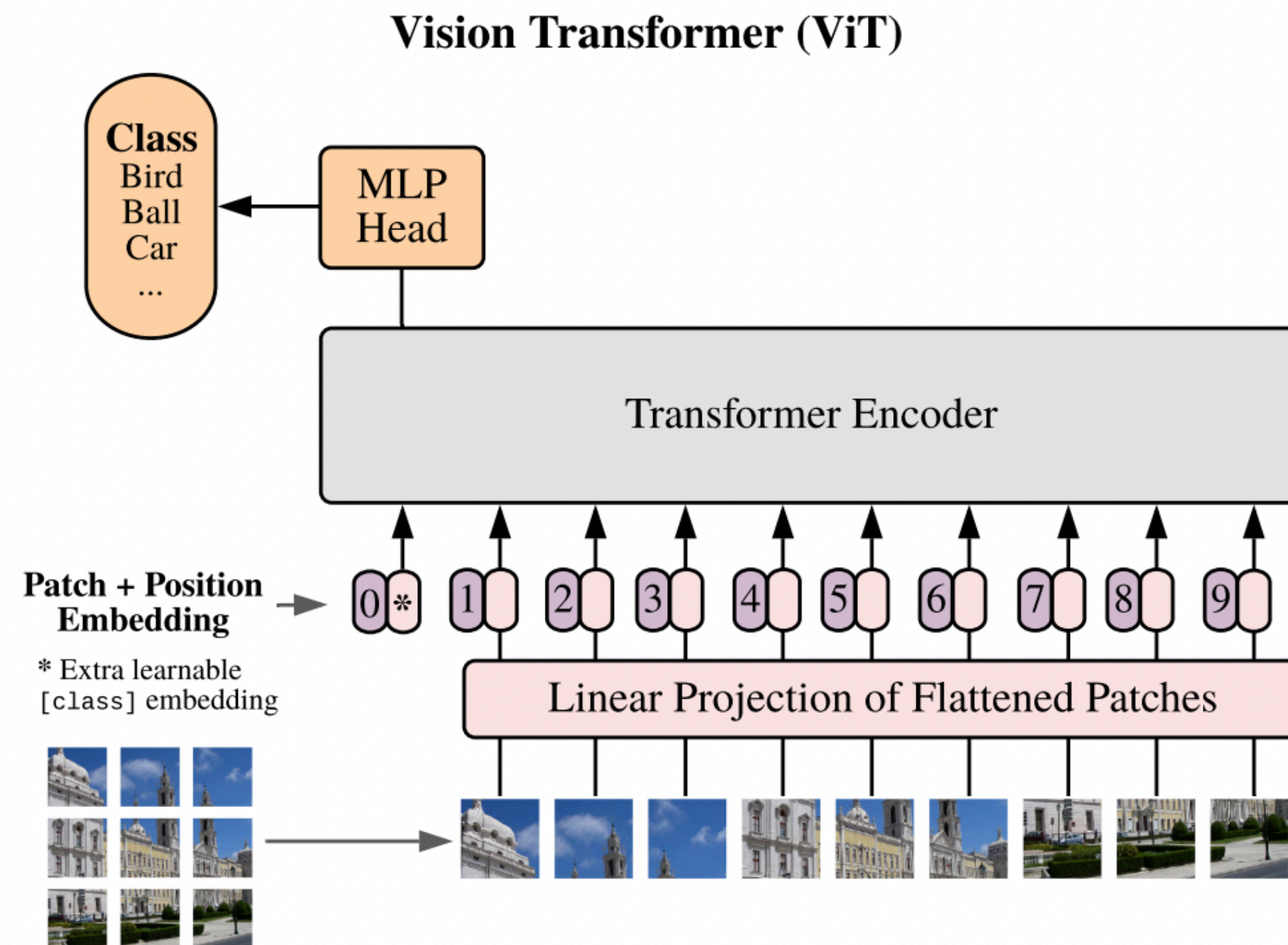
Transformer

- Origins in NLP
- “Tokenize” and “encode” the input
- Use attention to allow the relationships between tokens to be learned
- Decode the output



Vision Transformer

- Similar idea but tokenize the image using image patches
- instead of predicting next word we add a new [CLS] token at the end
- Supposedly less inductive bias than a CNN (allows contextual information to flow further more easily)



Advanced Visual Perception

Table of Contents

Intro to Advanced Visual Perception

Intro to Neural Networks

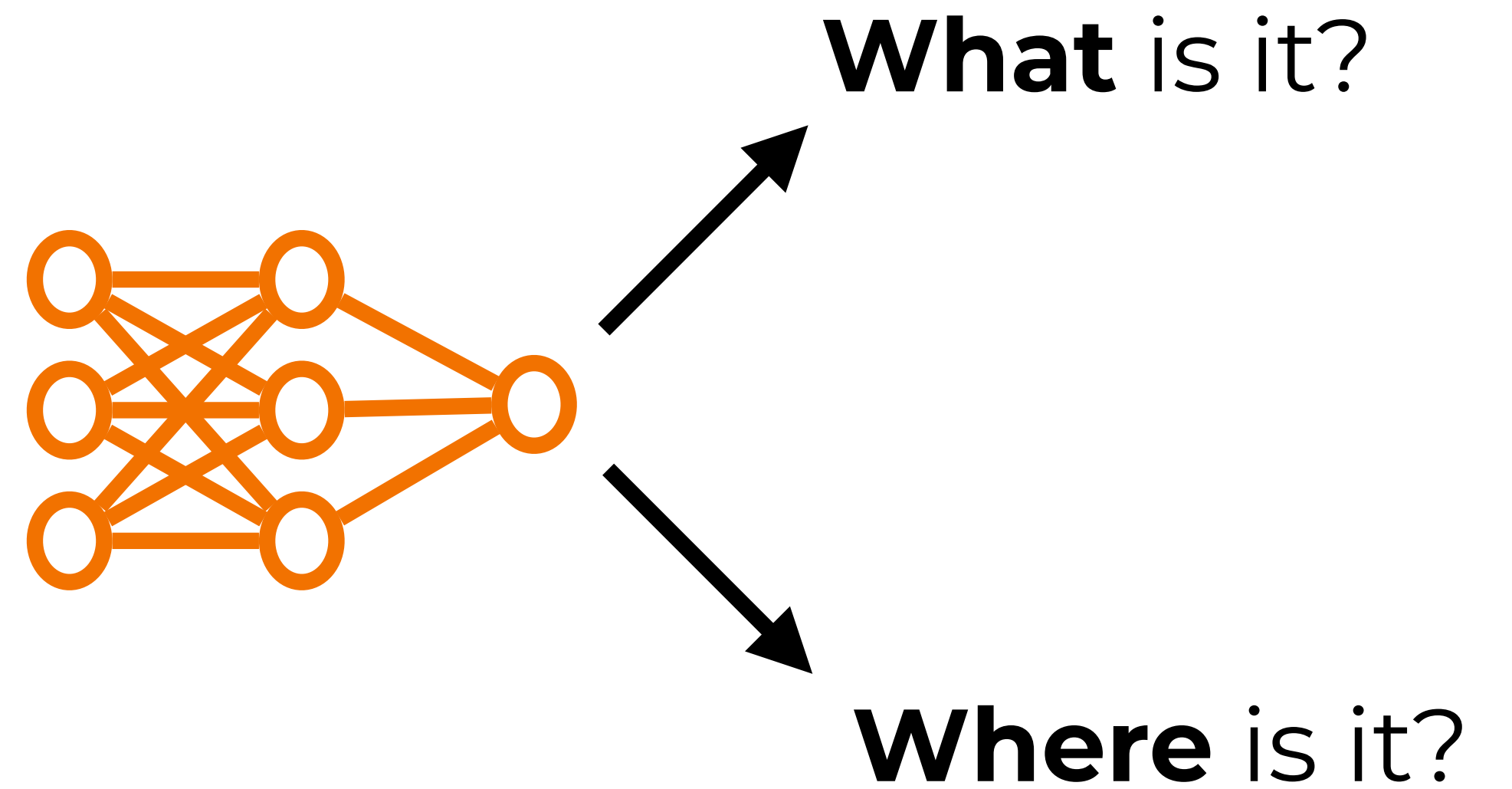
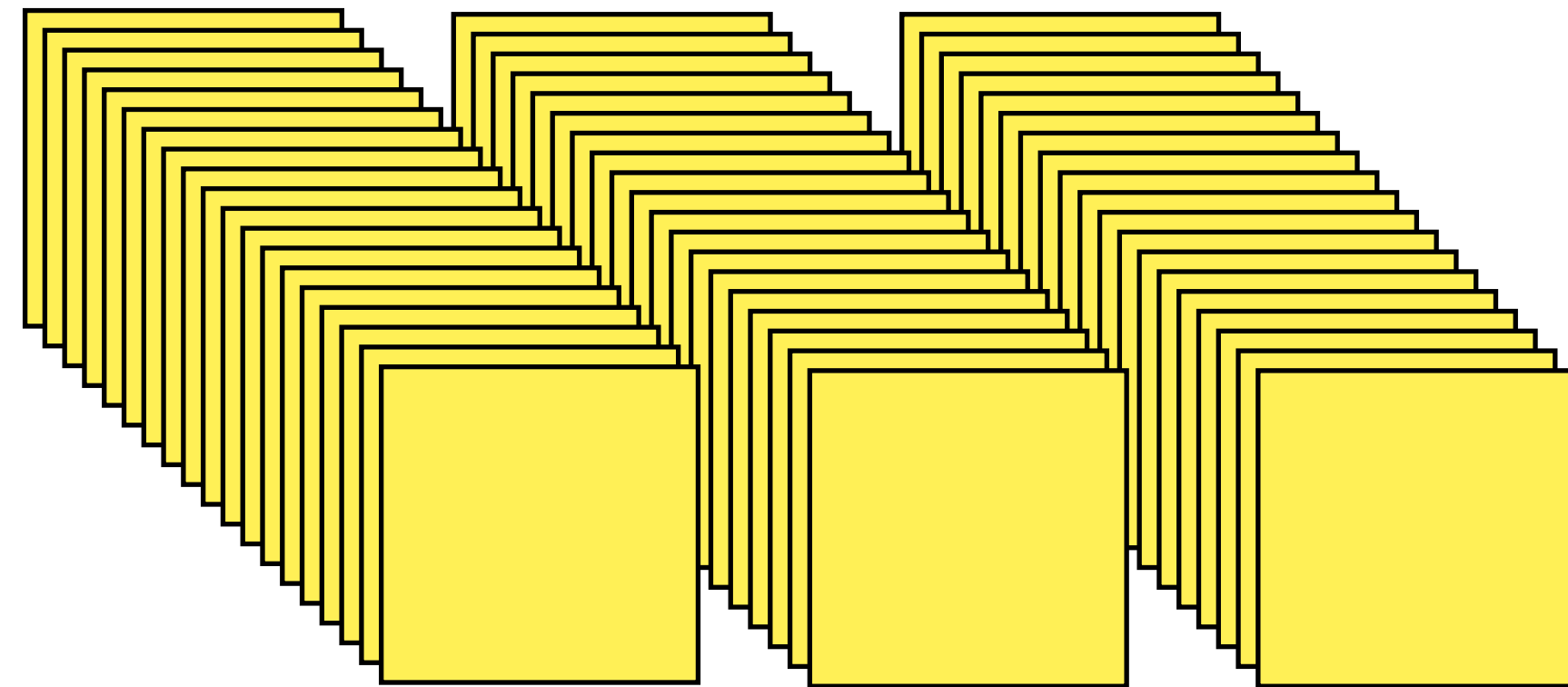
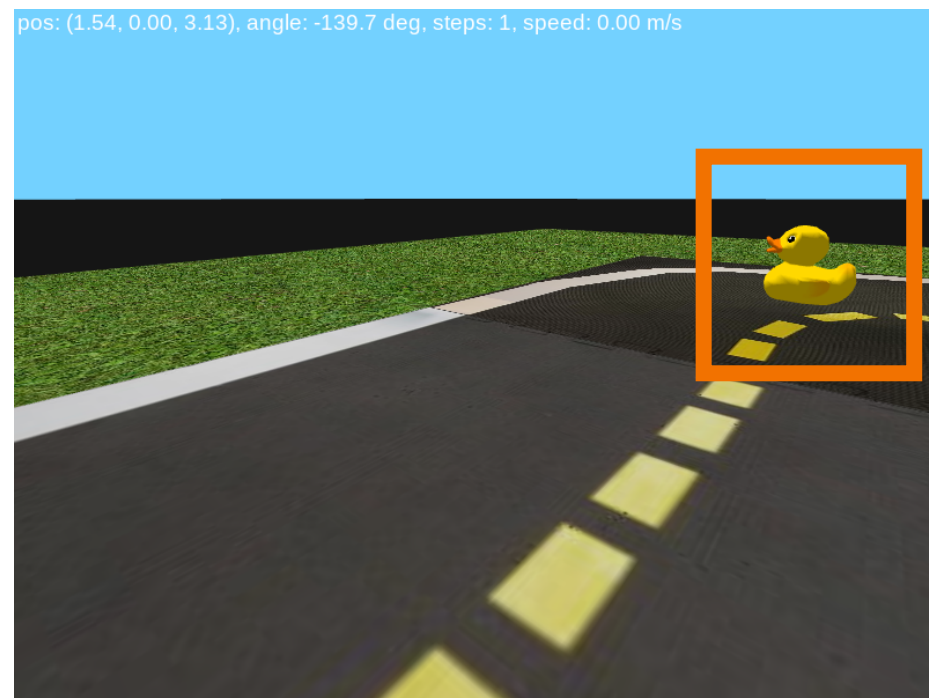
Deep Convolutional Neural Networks

Object Detection

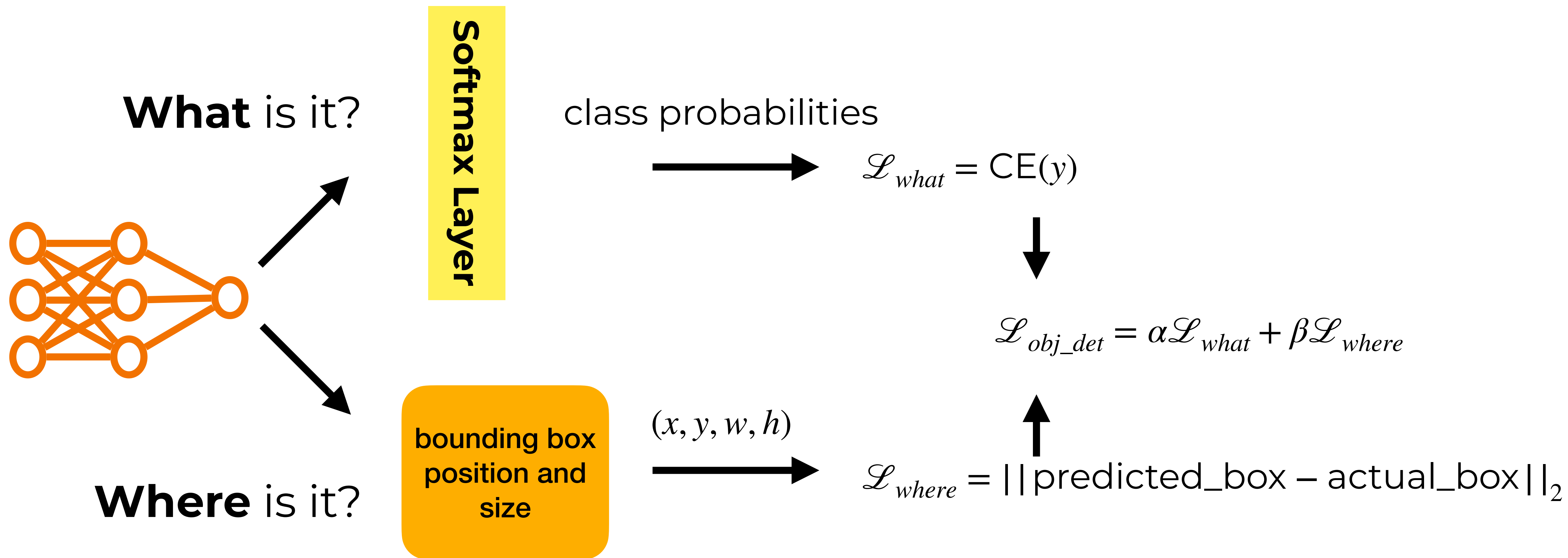
Semantic Segmentation

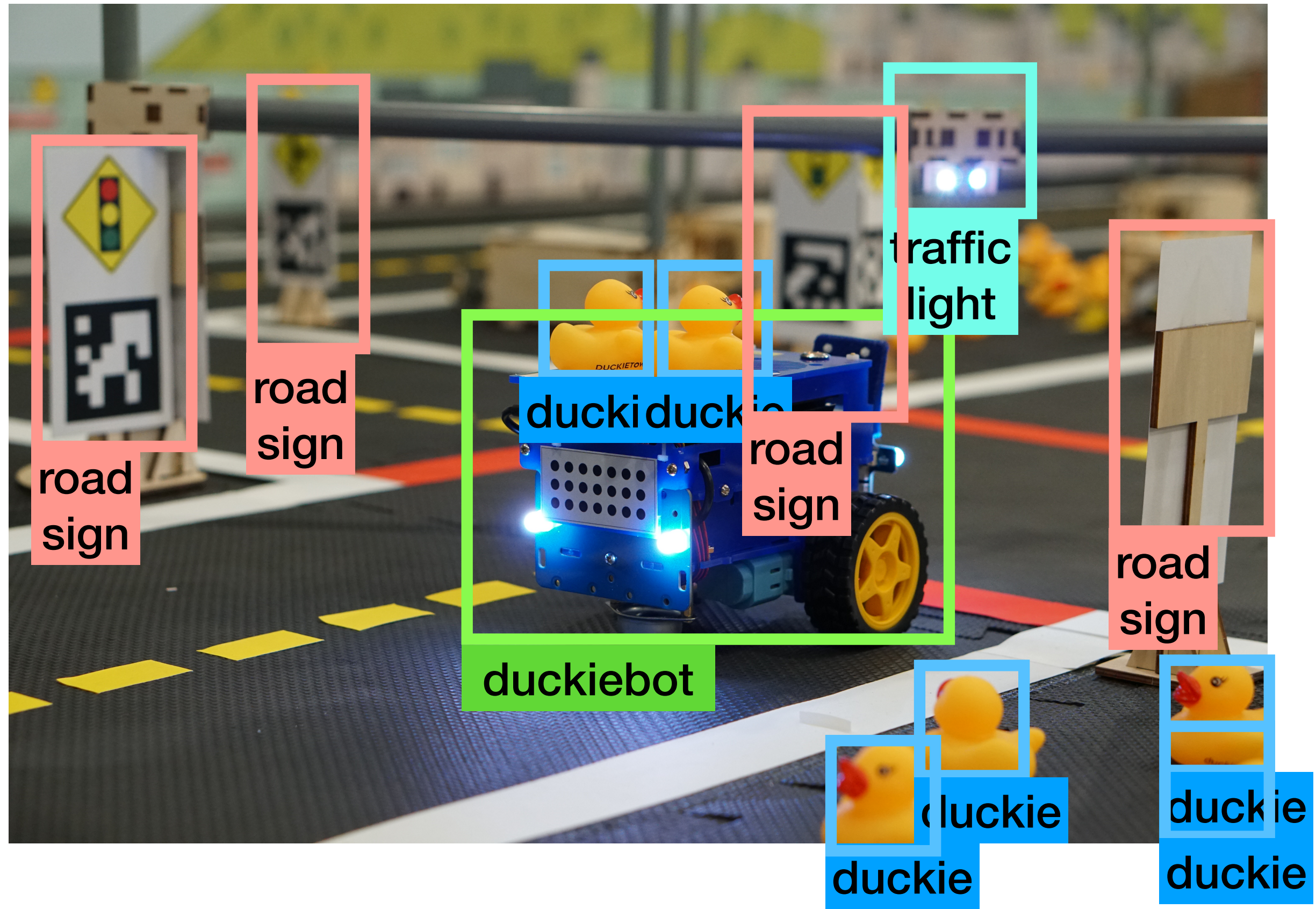
What if we don't have labels?

CNNs for object detection

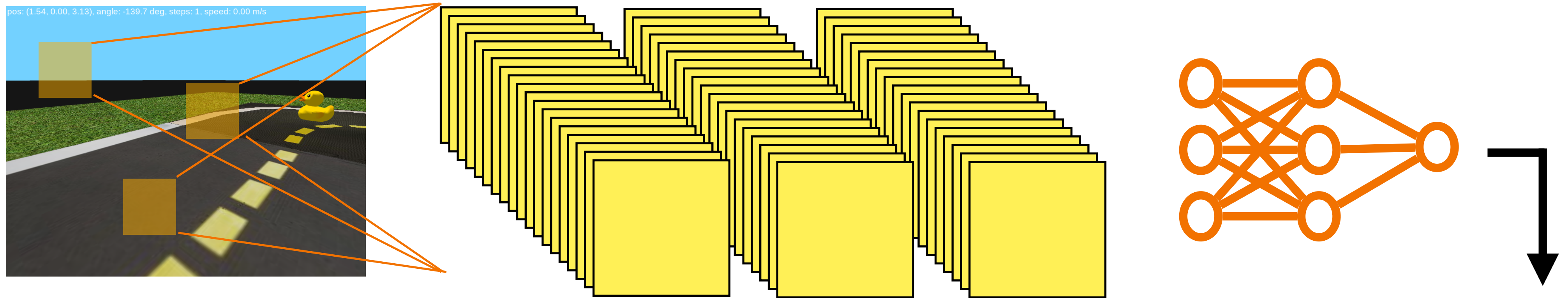


Loss Function





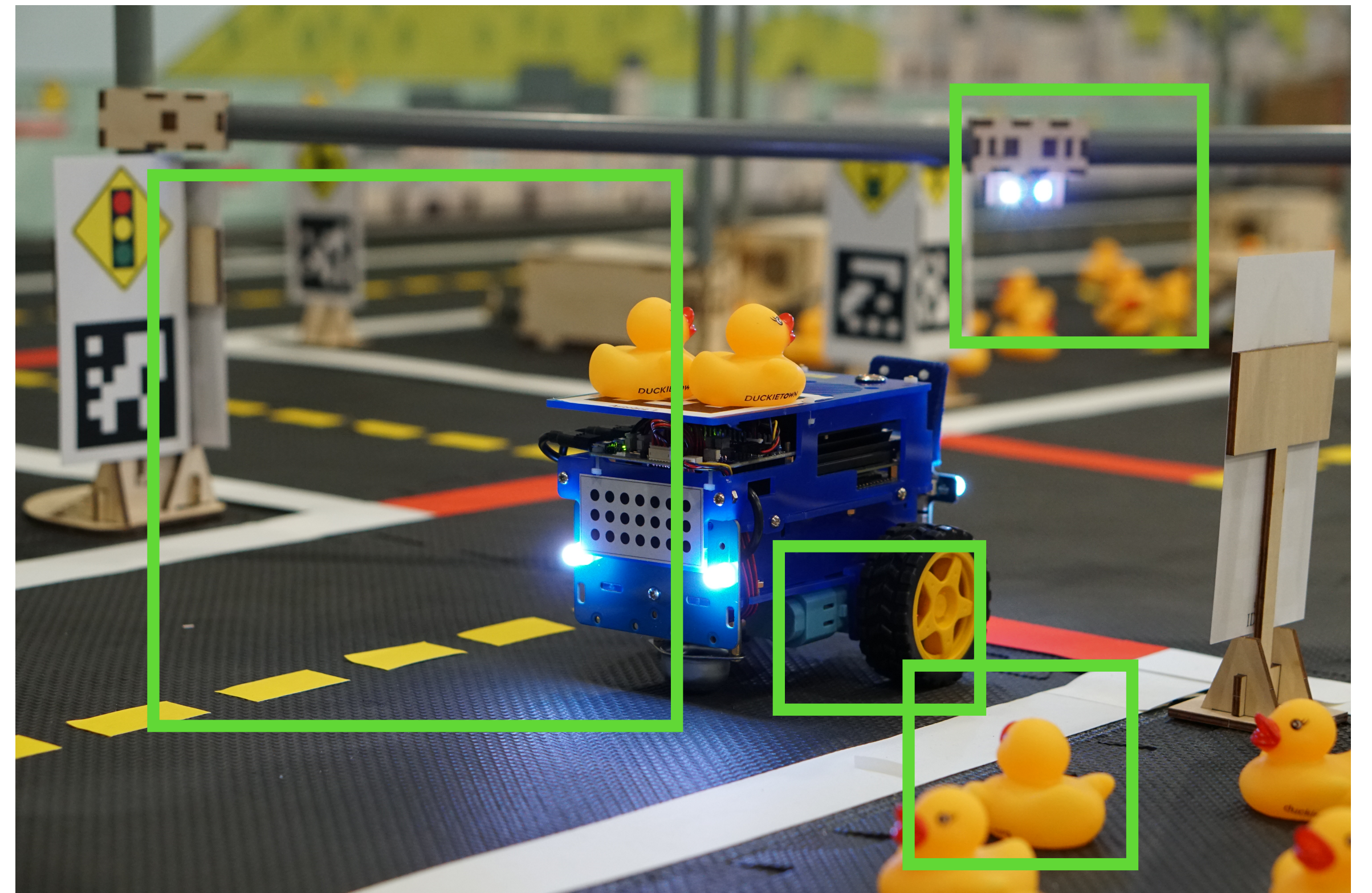
CNNs for object detection



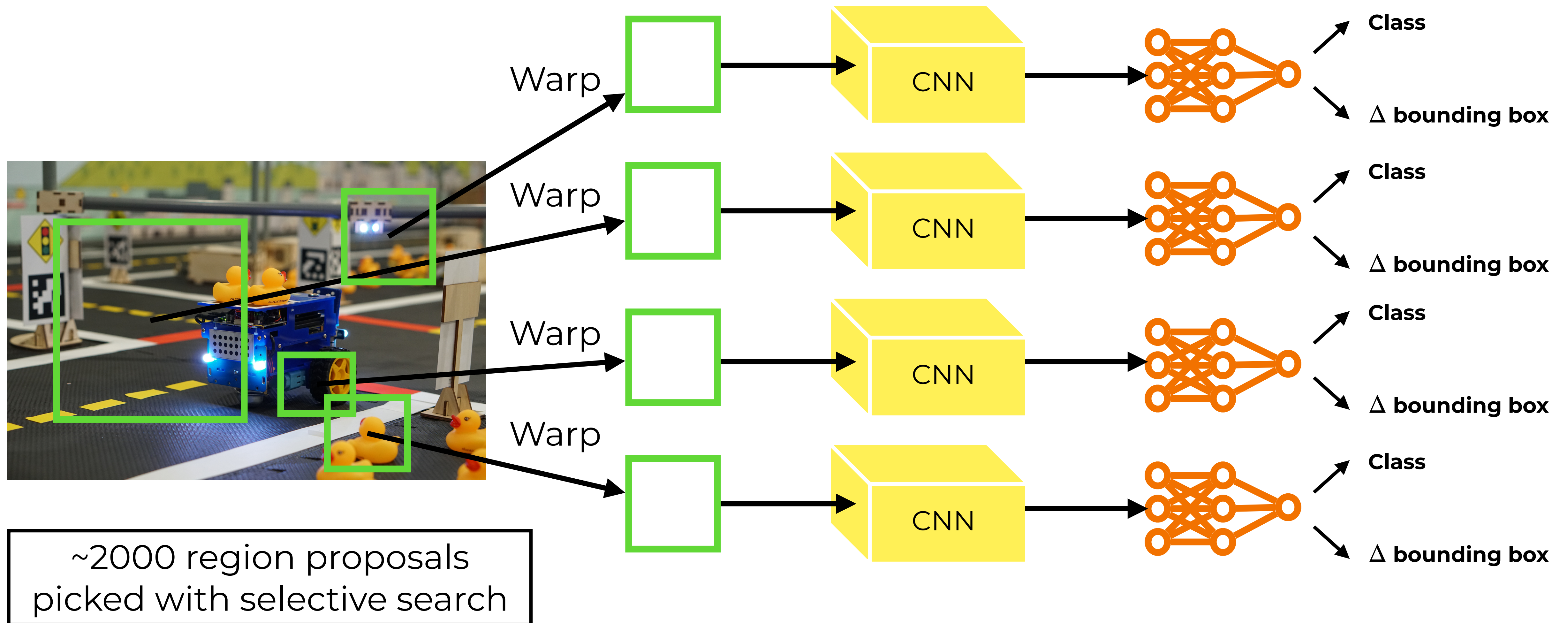
class \in {duckie, duckiebot, ..., background}

Too many possible boxes to do this exhaustively

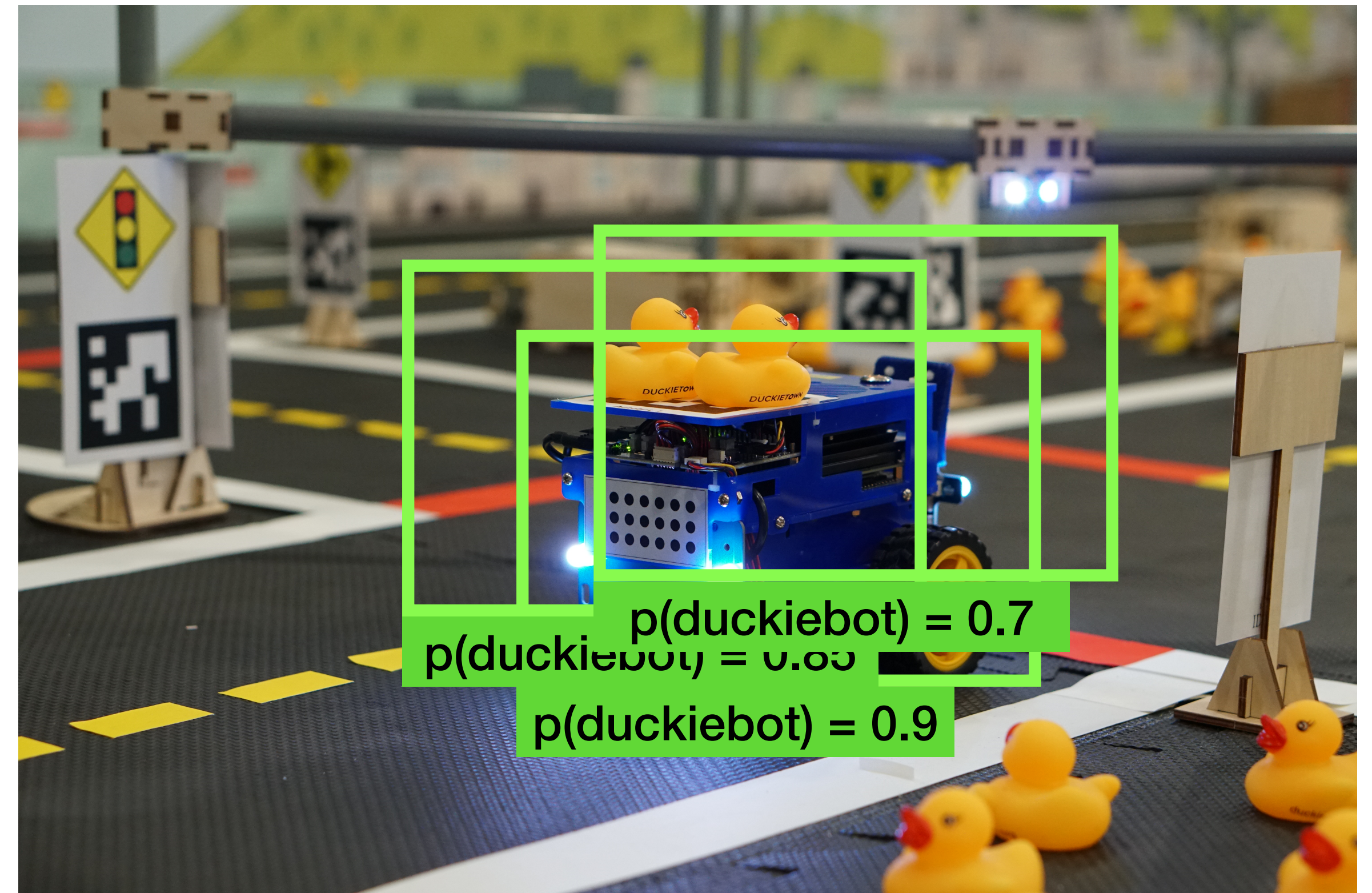
Region proposals



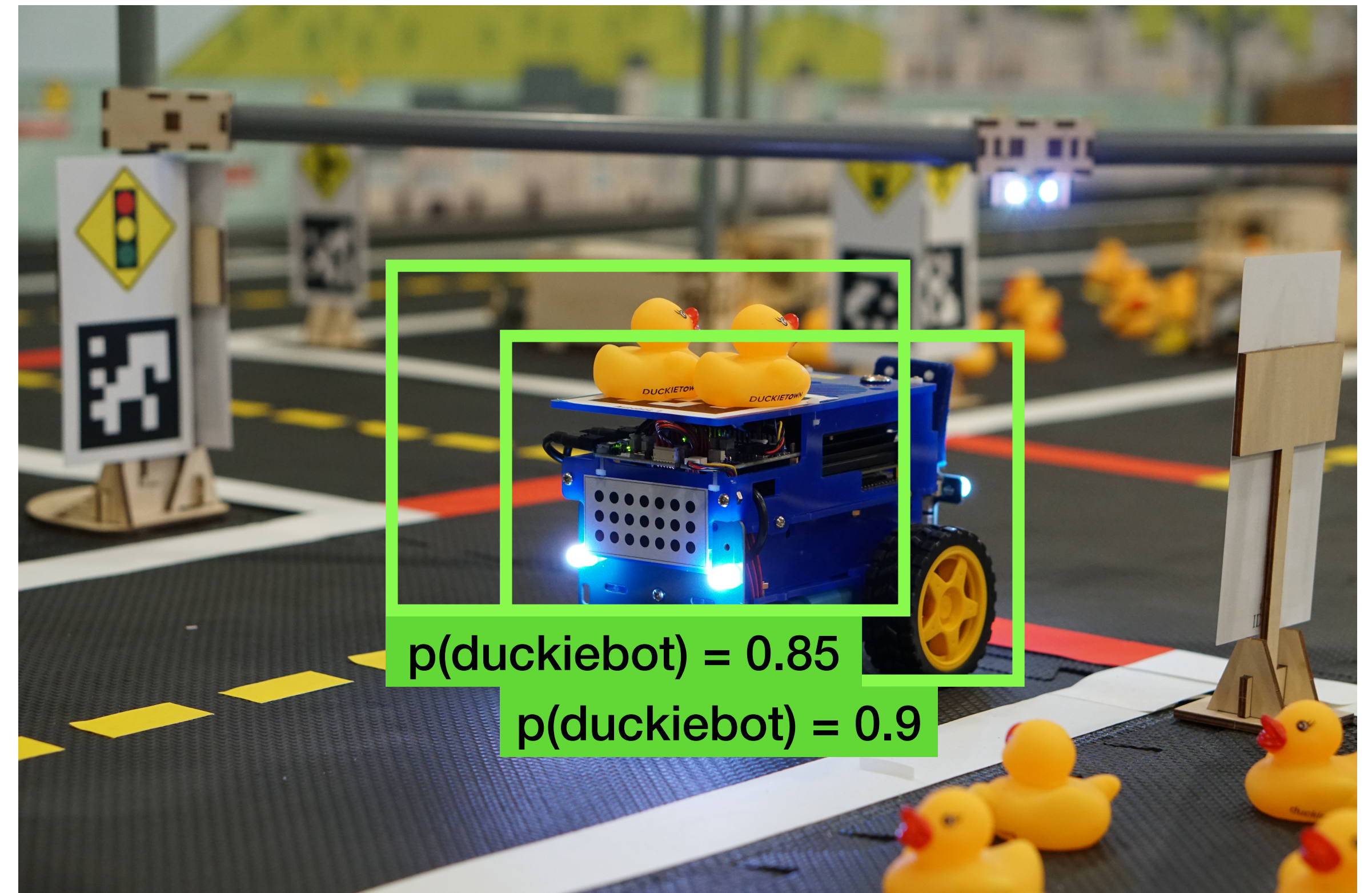
R-CNN: Region-Based CNN



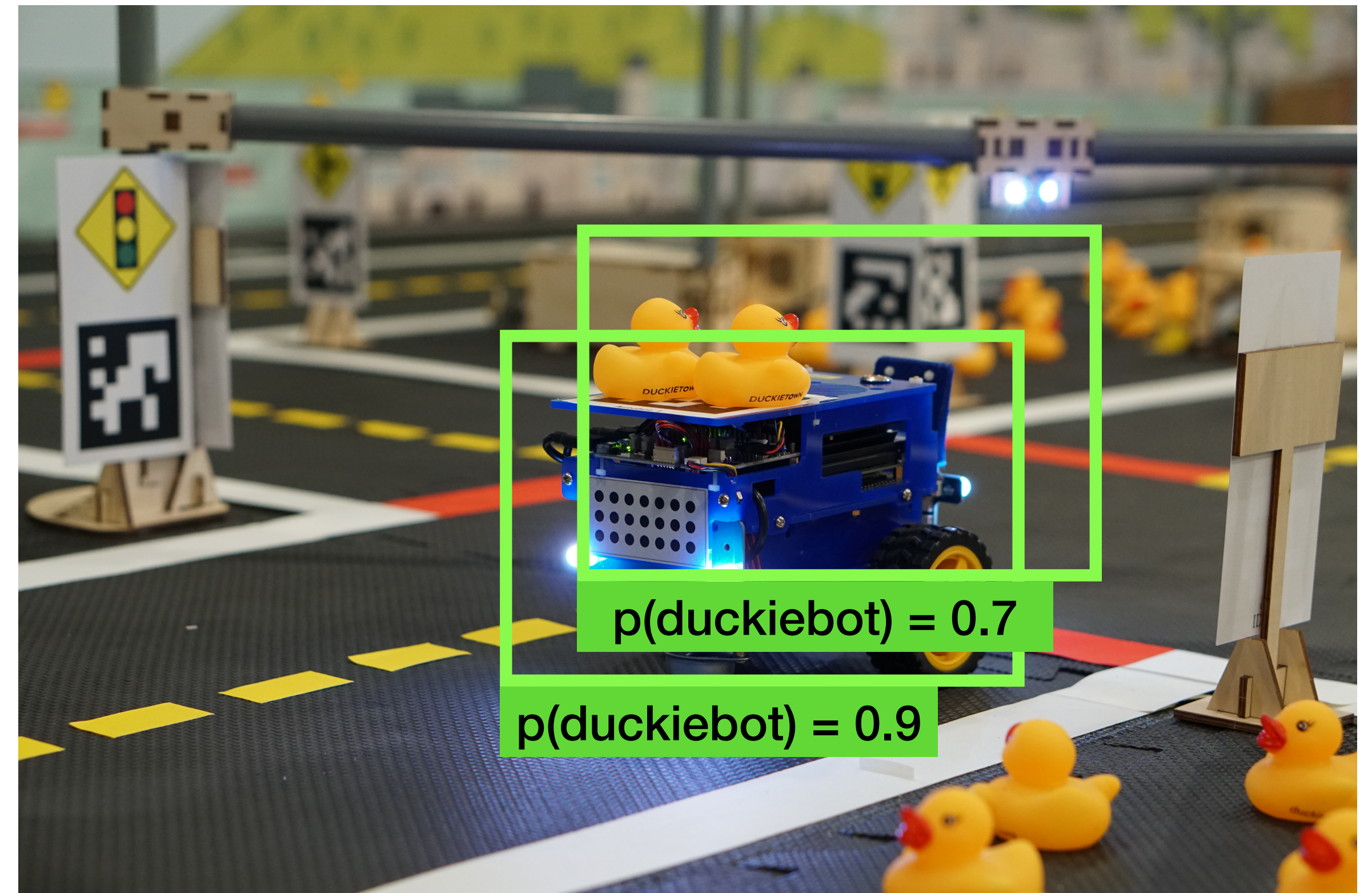
Non-maximum suppression



Non-maximum suppression

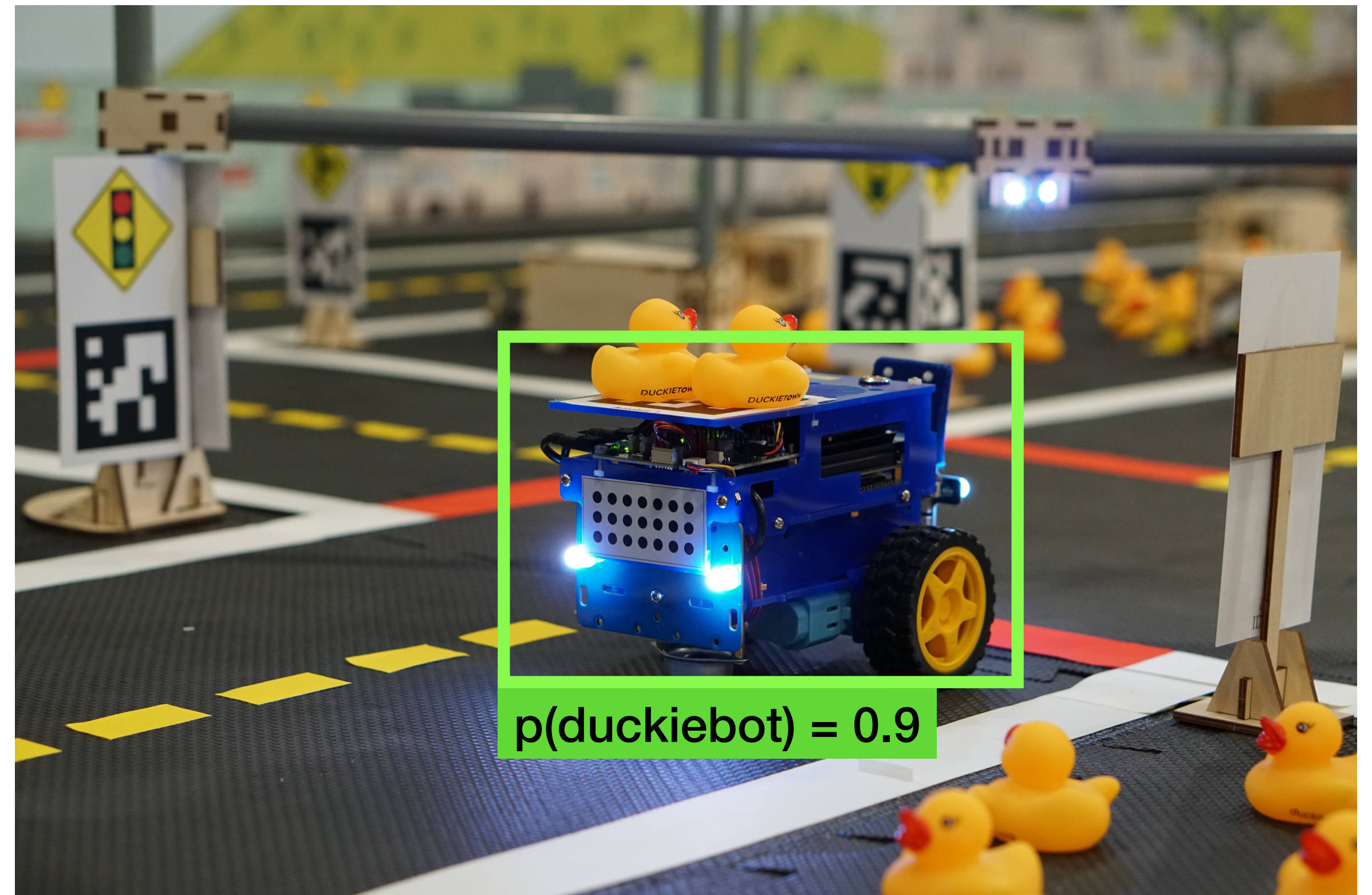


Non-maximum suppression

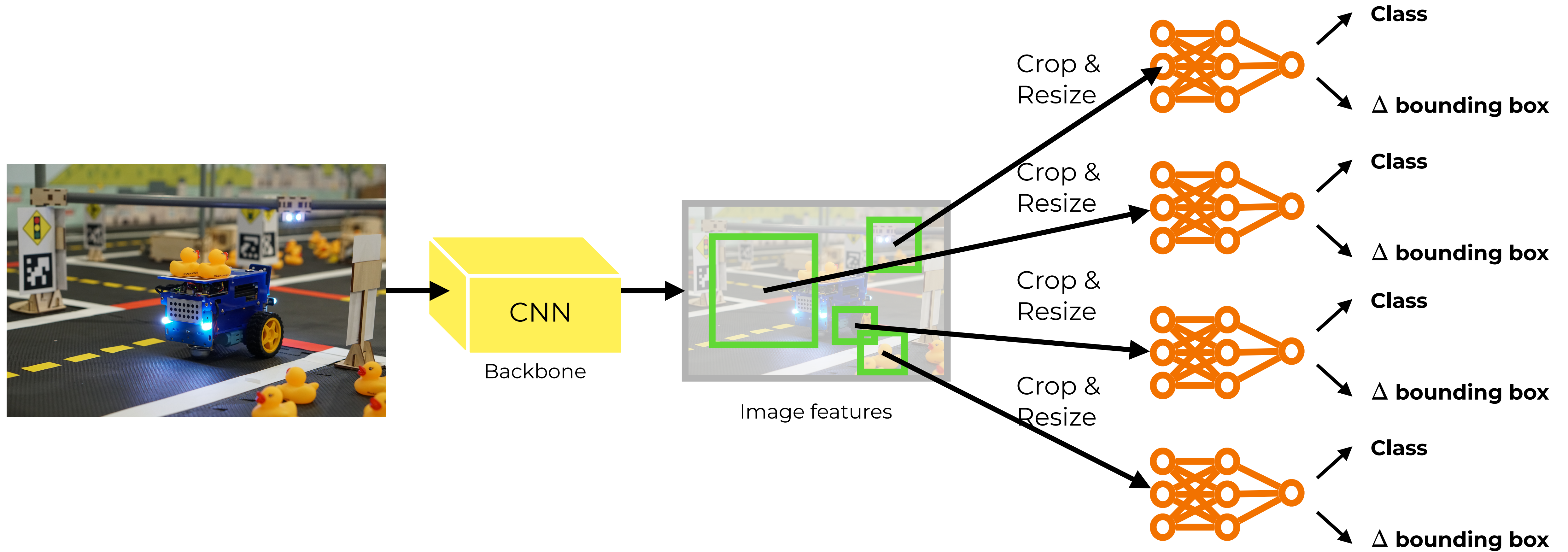


IOU > threshold

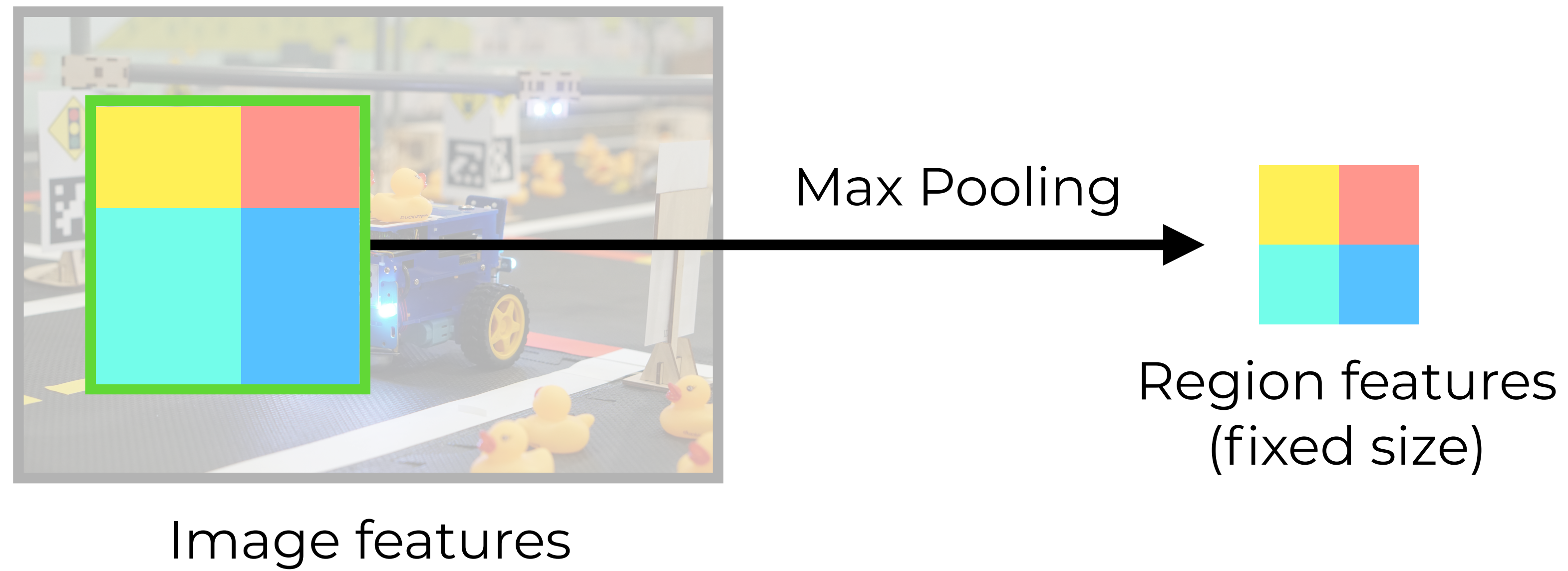
Non-maximum suppression



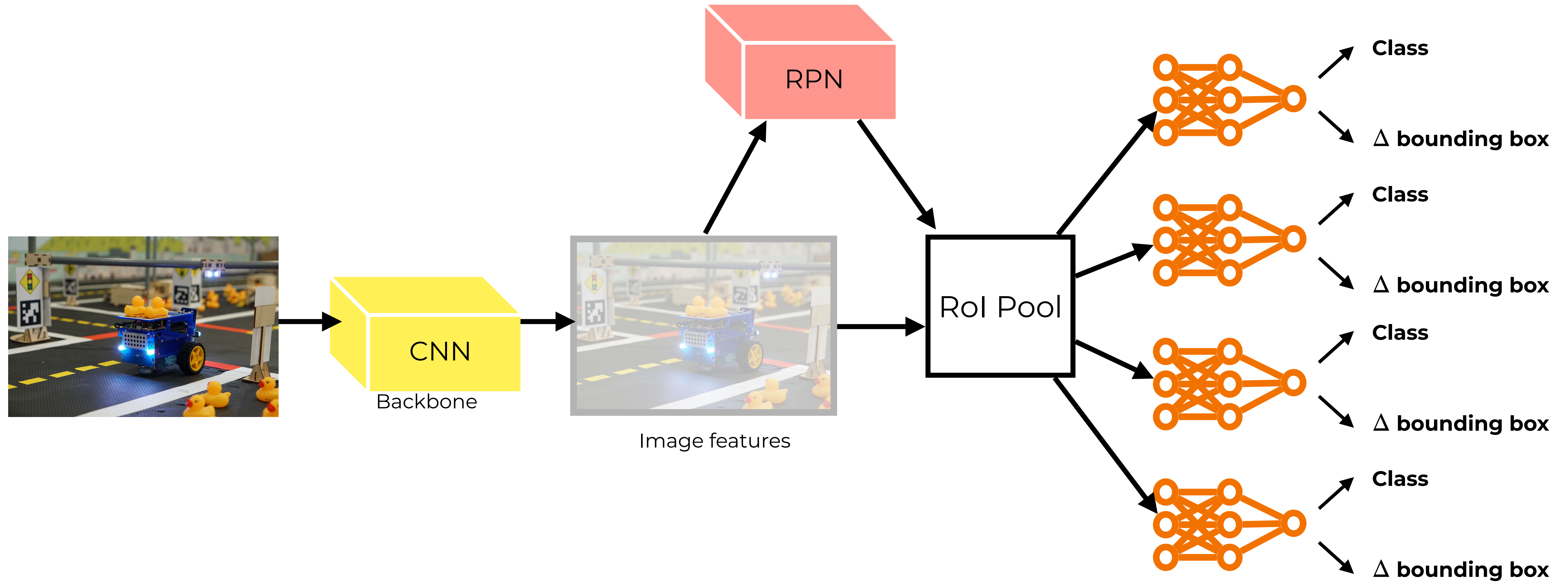
Fast R-CNN



Region of Interest (RoI) pooling



Faster R-CNN



Region Proposal Network

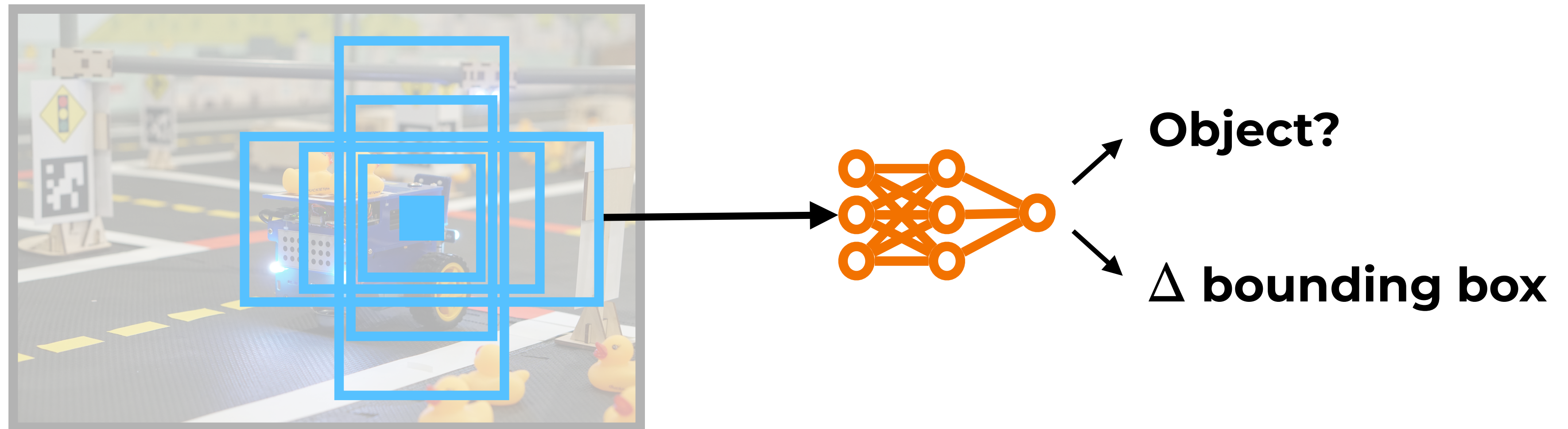
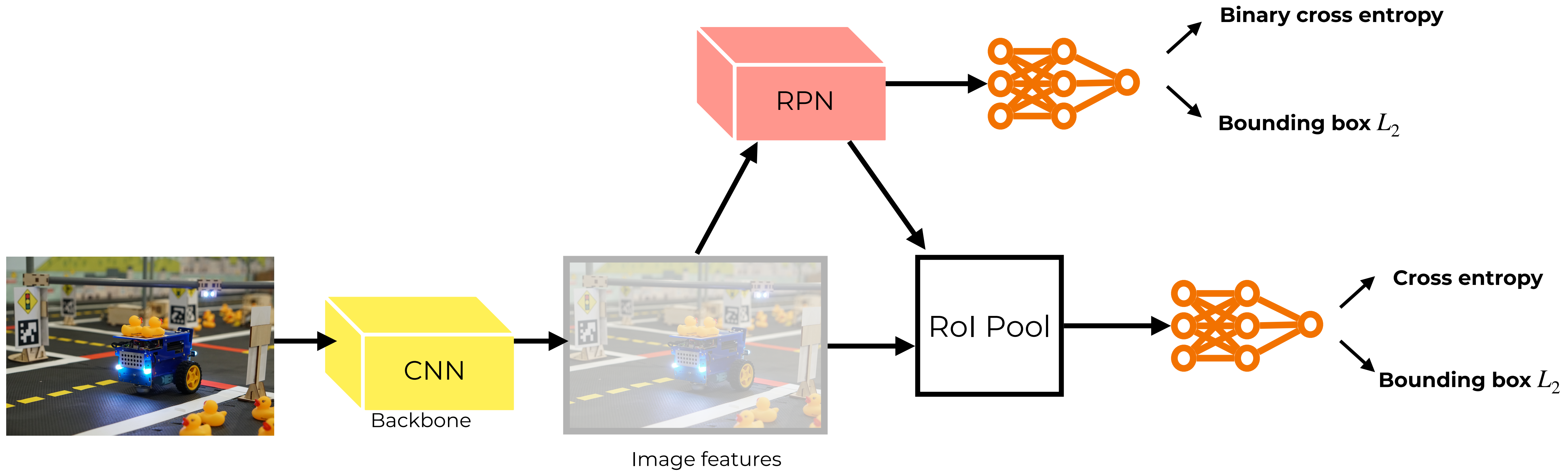


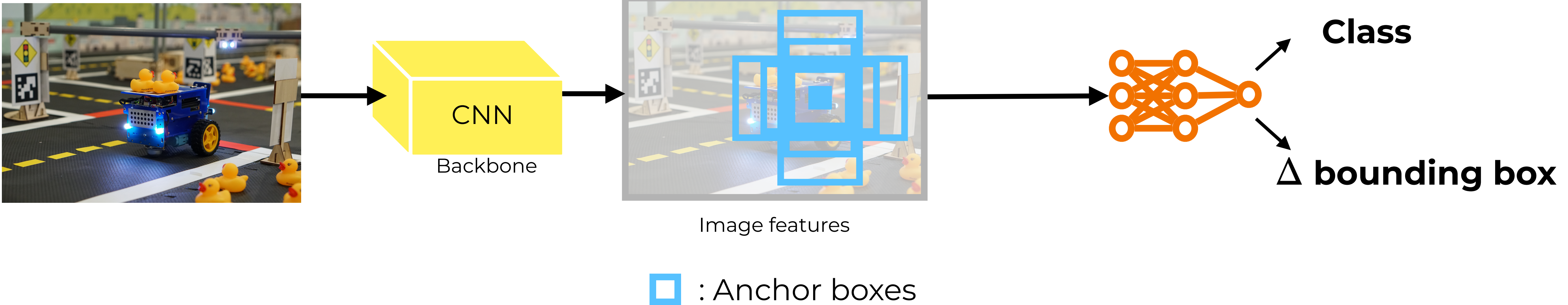
Image features

 : Anchor boxes

Faster R-CNN



You Only Look Once (YOLO)



No need for RPN

Issues with single-stage object detectors

Problem 1: Difficult to detect objects at different scales

Solution 1: Use features from different layers of backbone

Method 1: Single shot multibox (SSD)

Issues with single-stage object detectors

Problem 1: Difficult to detect objects at different scales

Solution 1: Use features from different layers of backbone

Method 1: Single shot multibox (SSD)

Problem 2: Data imbalance due to background

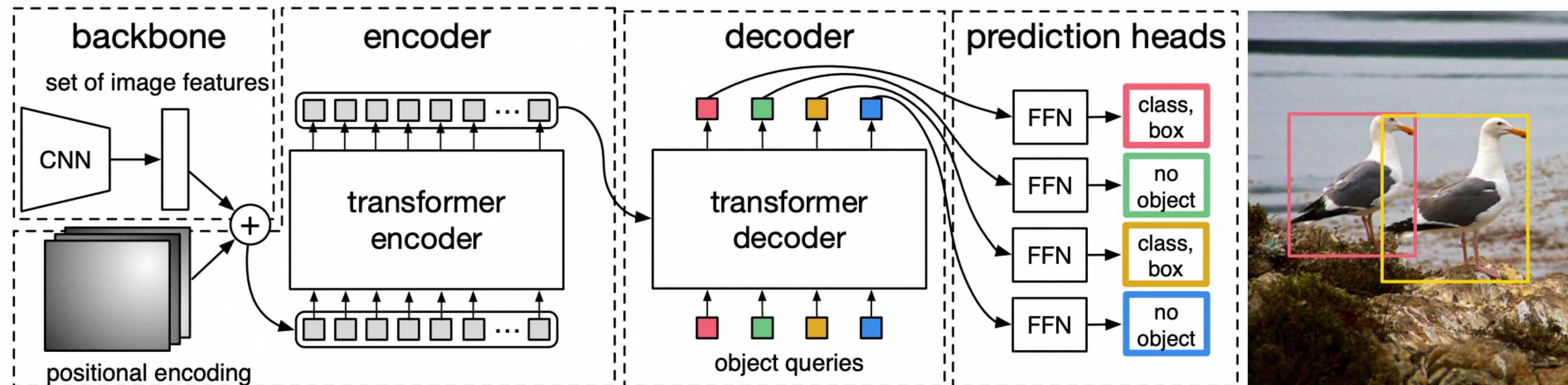
Solution 2: Focal loss that prioritizes objects

Method 2: RetinaNet

DETR

Application of vision transformers to the problem of object detection:

$$L = \sum_{i=1}^N \left[-\log \hat{p}_{\sigma(i)}(c_i) + \mathbf{1}_{c_i \neq \emptyset} \left(\lambda_{L1} \|b_i - \hat{b}_{\sigma(i)}\|_1 + \lambda_{giou} L_{giou}(b_i, \hat{b}_{\sigma(i)}) \right) \right]$$



Datasets

- Kitti, CityScapes Nuscenes
- Mostly annotated by hand



Advanced Visual Perception

Table of Contents

Intro to Advanced Visual Perception

Intro to Neural Networks

Deep Convolutional Neural Networks

Object Detection

Semantic Segmentation

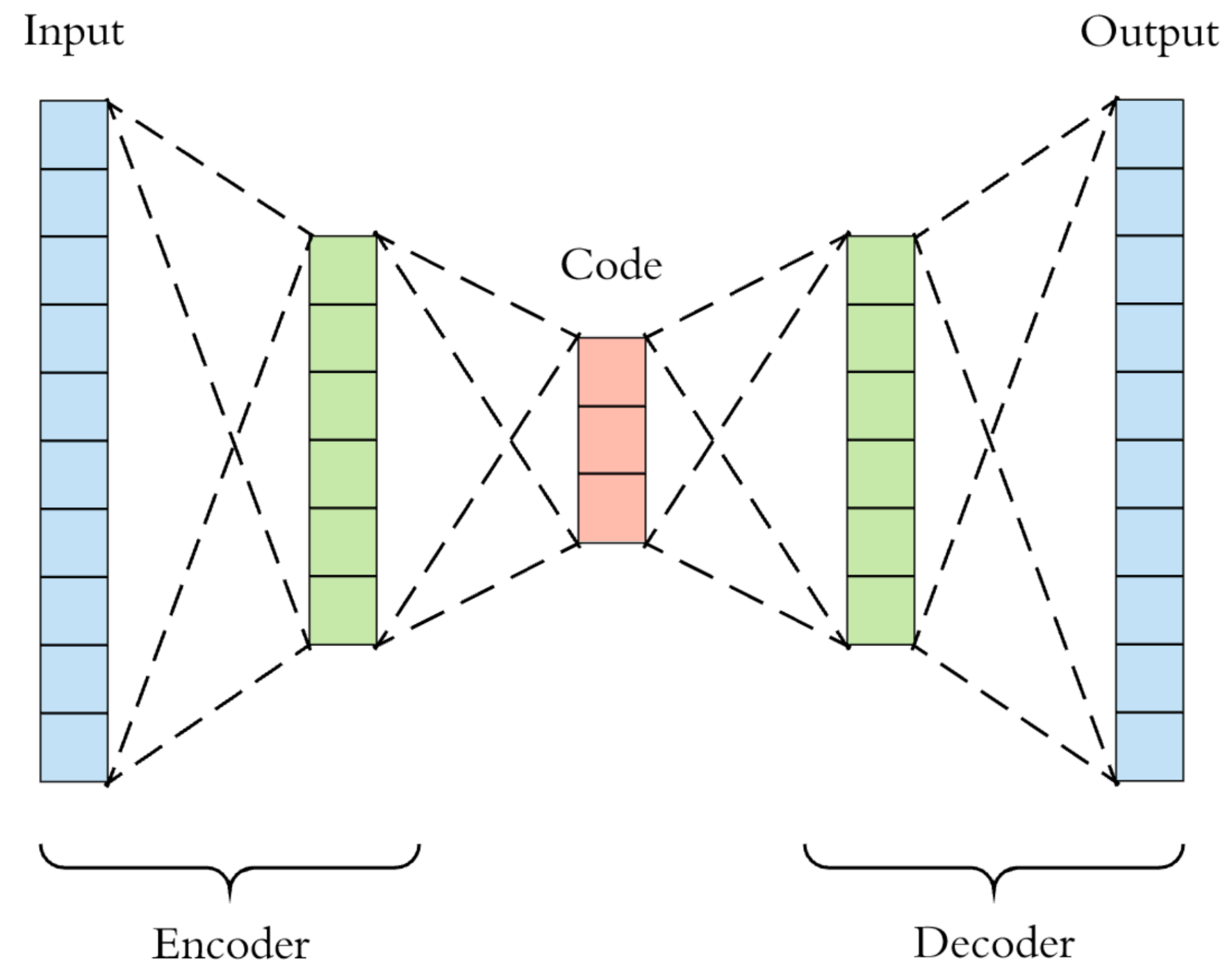
What if we don't have labels?

Autoencoder

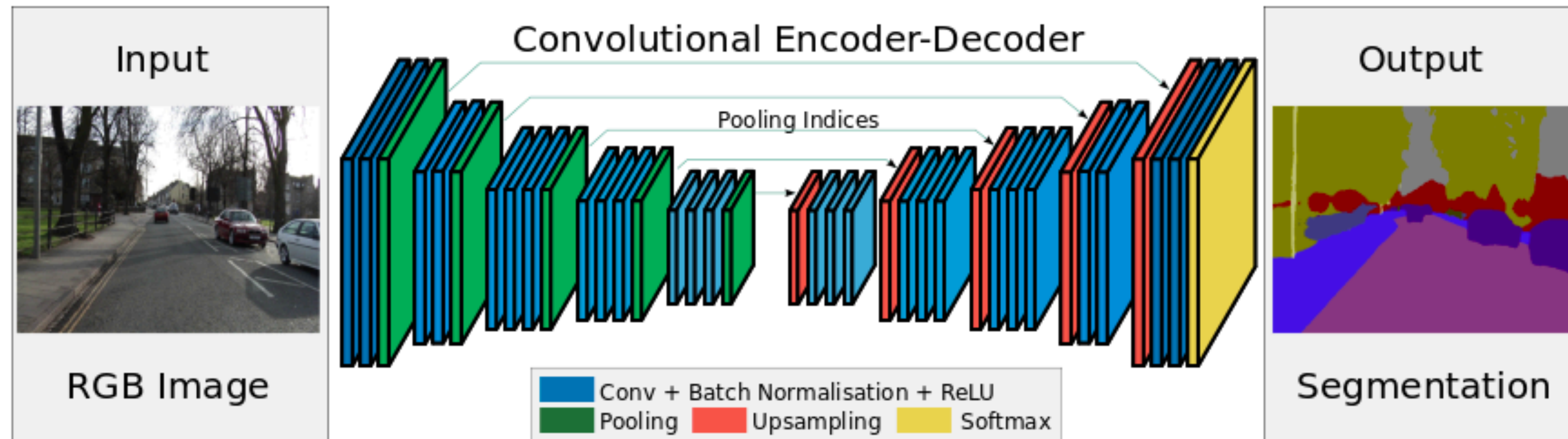
Objective: learn a latent representation of your input

No need for data

Usually: minimize the reconstruction loss (maybe regularized by something to enforce the kind of latent representation you want)



Segnet



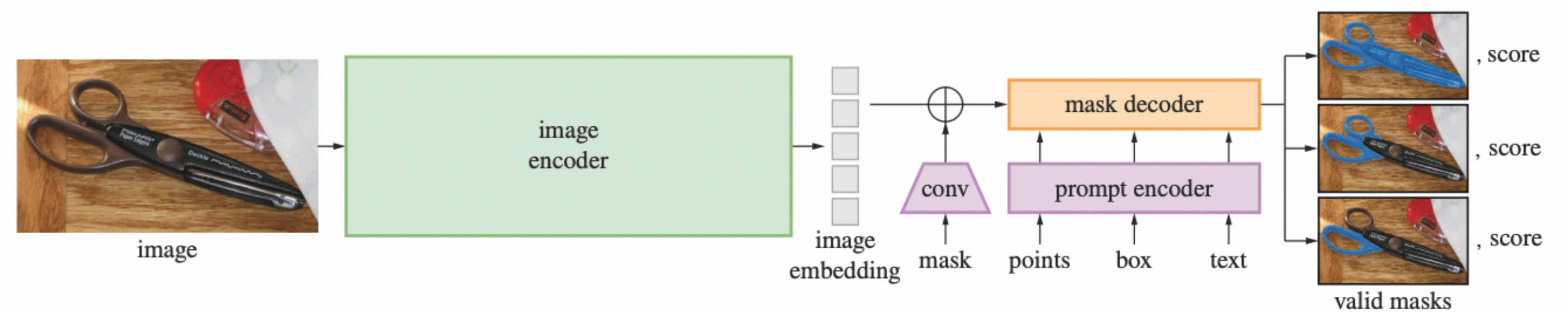
Segment Anything Model

1. Image Encoder (ViT)

2. Prompt encoder (encode points / masks)

3. Mask decoder

Dataset of 11 million images and 1B masks (but most not generated by humans)



Advanced Visual Perception

Table of Contents

Intro to Advanced Visual Perception

Intro to Neural Networks

Deep Convolutional Neural Networks

Object Detection

Semantic Segmentation

What if we don't have labels?

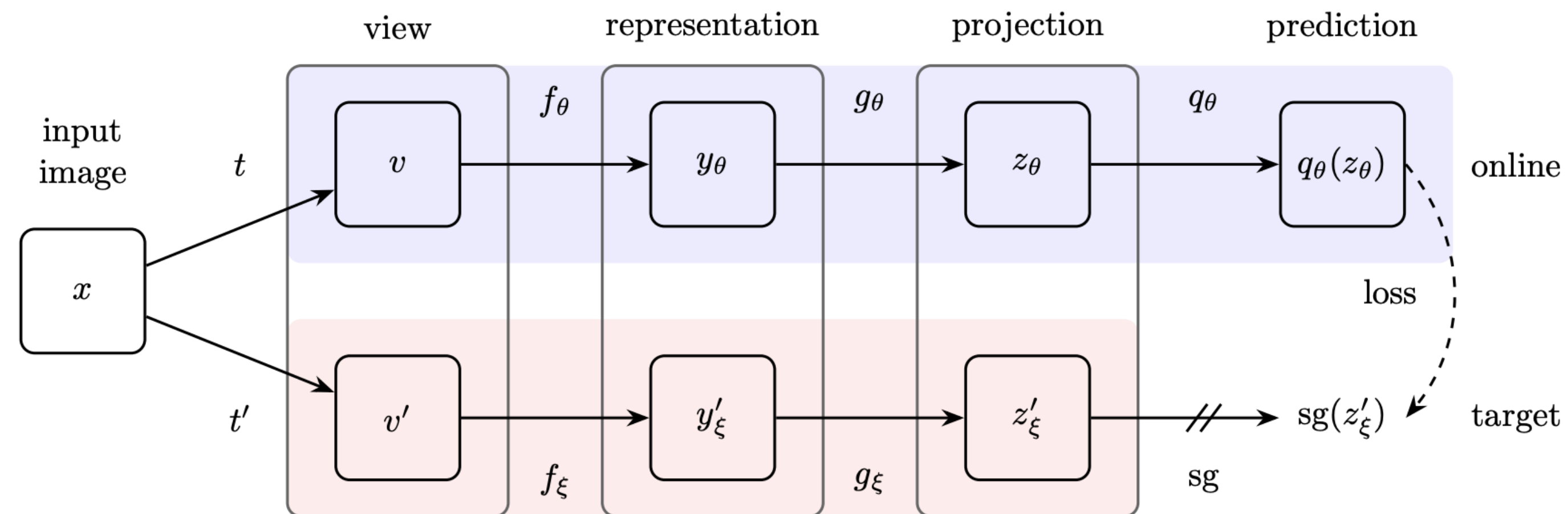
Contrastive Learning

- Use some “pretext task” to learn a representation
- Often works well in conjunction with a **contrastive loss** (e.g. in SimCLR)

$$l_{i,j} = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)}$$

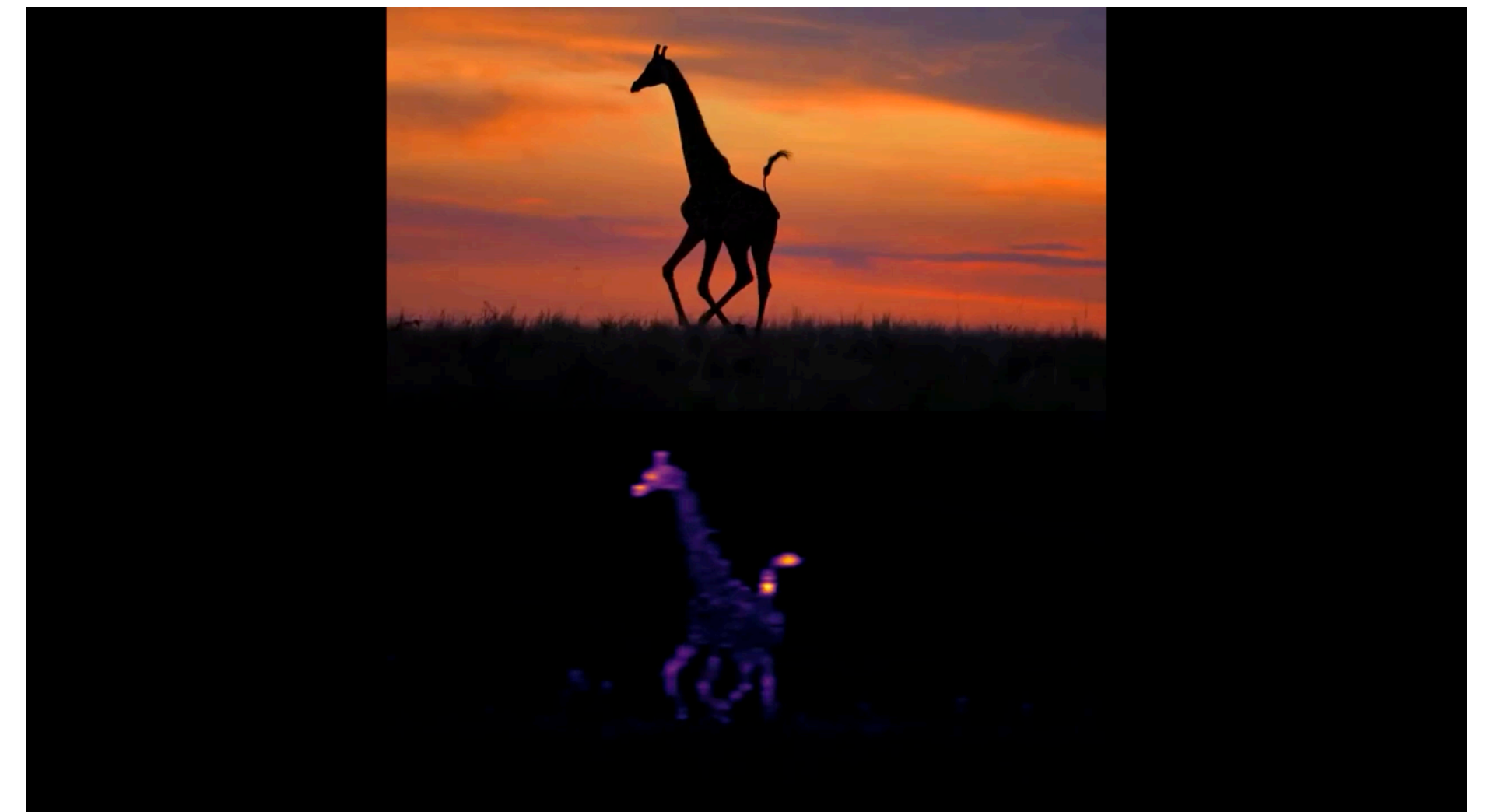
Self-distillation

- Use some “pretext task” to learn a representation
- Or through the process of knowledge distillation (e.g. in BYOL)



Self-Supervision + Transformer

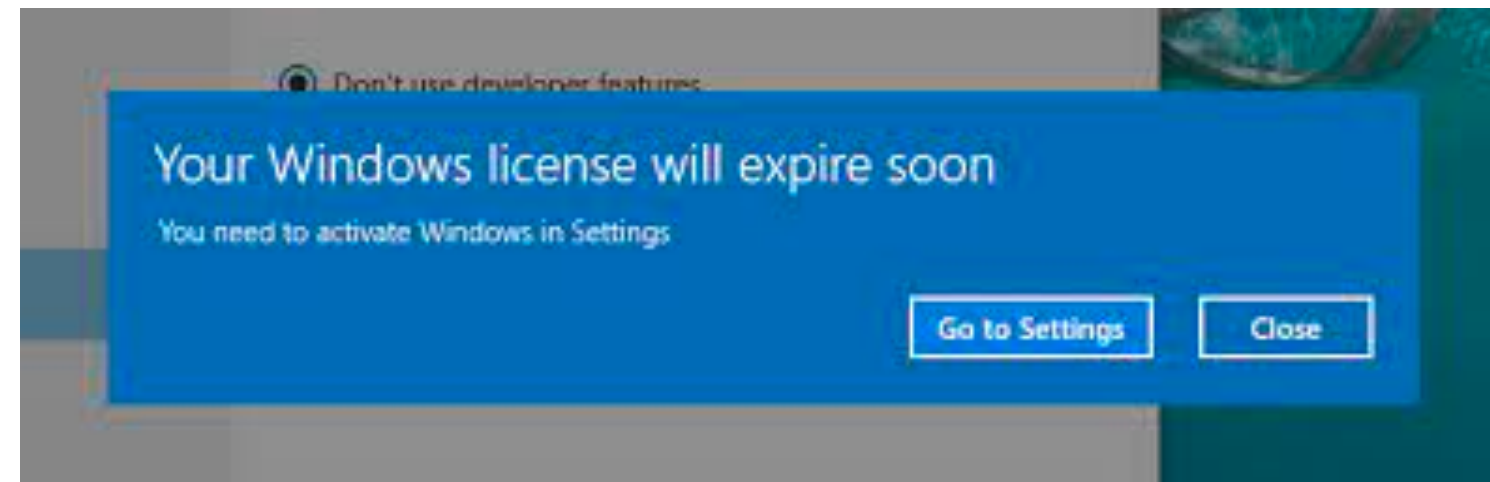
DINO



Multi-modal embeddings - CLIP



Tucker, the pup,
looking up at the
camera momentarily



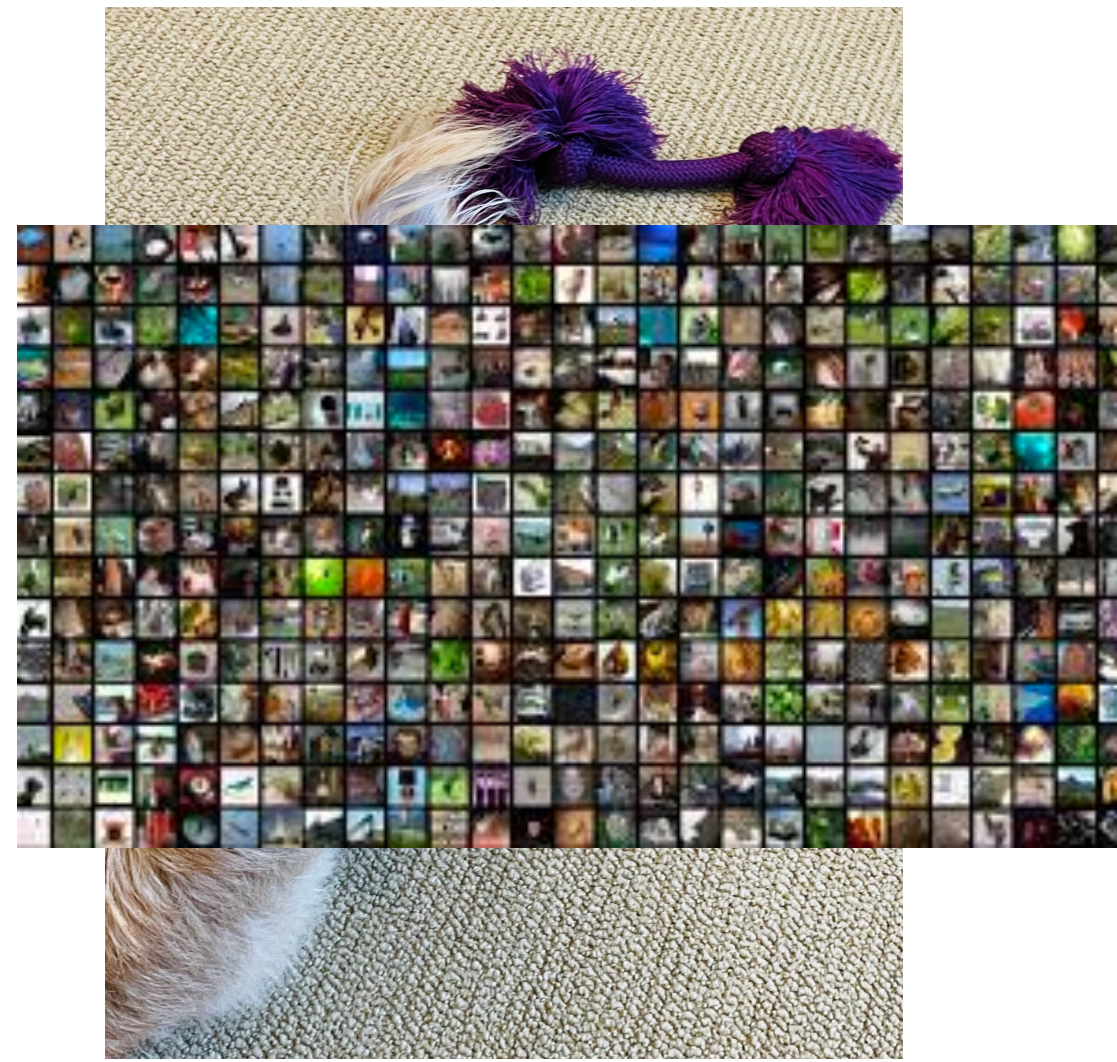
A computer screen with a Windows license
message about expiry



A digital thermometer
showing 160 F



Multi-modal embeddings - CLIP



Vision transformer

I_1 I_2 I_3 I_N

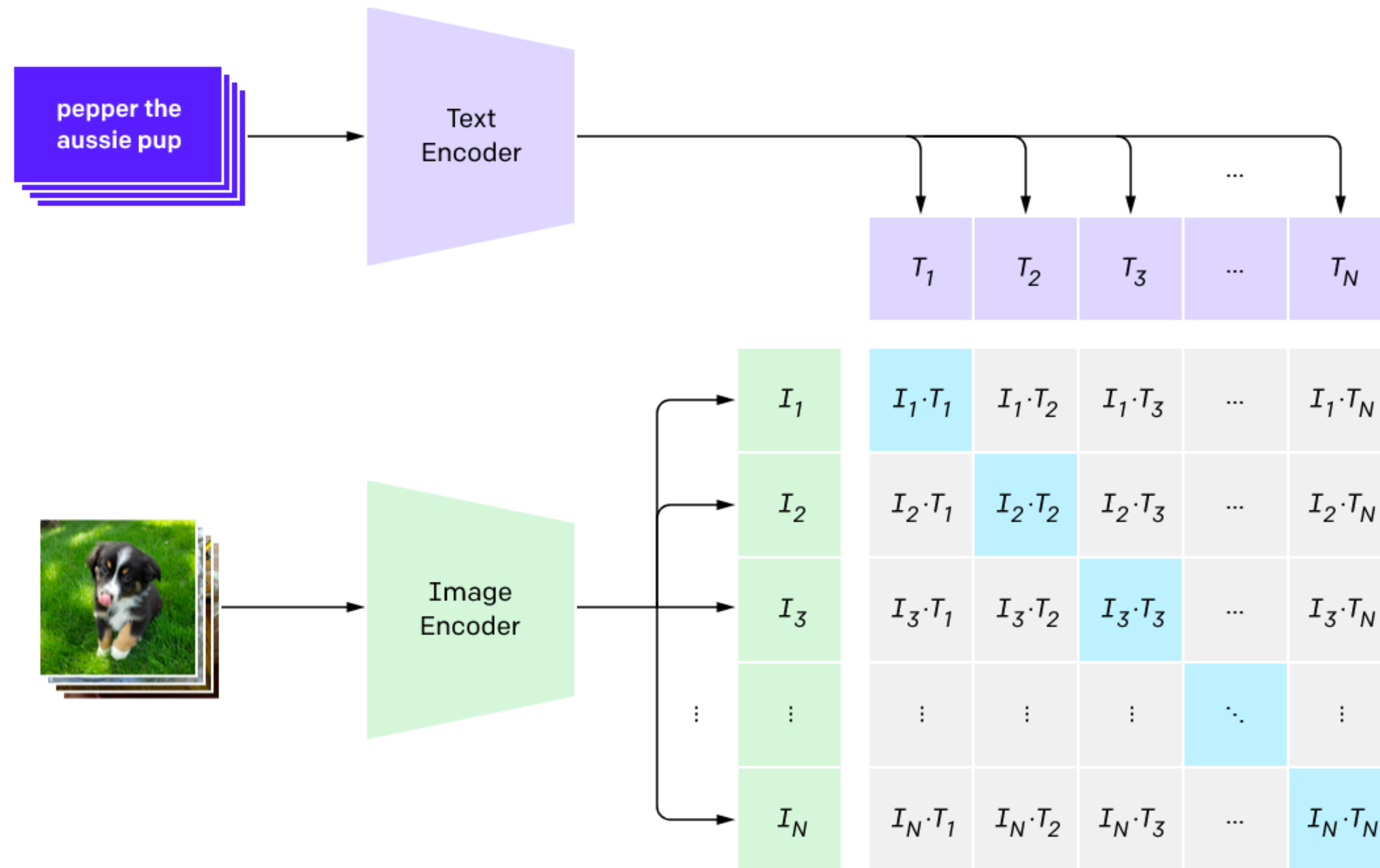
...right. Money comes and goes. The things we	asked	the Harvard Board of Cabot. When in response the range of views
...of our friends like yours and yours that more of them can be	explained	using a single theory of determinants. It's better, however, as an
...handbags and money changes, being in fact, rather an ordinary way	found	to measure and kind eyes. They are particularly suitable for the
...to which these also would stay in. The 15th anniversary was	seen	as the 20th anniversary of their former receipt, again partly as
...The? "Yes," they said, through their	checked	against the cost and the deeper cost of the "right"
...value system. There are already one-third of a million, which the last	assessed	for the right of course in such matters. For me
...1980 instead of \$1,000 for "Tiger," and because the manager has	not	up, only half the cost for acquiring a cover that is not in
...at least another 100,000. That's (I guess) the just	given	to us in case. No, you know? It's
...and equally effectively we come forward in a more light, and he had	assessed	in 1980 position for England doing his time with "The Research About
...collected information about him. However, your computer had	failed	to follow that information made under English. It should be used to
...strangled another who he has had to struggle for, who had	failed	single and, both without the kind, was working with the need to
...at home. Some employees of Robert M. Brown, whose position was	made	for the publishing process. The algorithm is contained in a report by a
...returns in early 2006 as changes of funds and operational problems ...	made	by an assembly of emergency working. It differs against the situation in terms
...possibilities on what had already been done. Robert M. Brown, who had	assessed	automatically under follow and Robert Wood in 1911, as opposed to the other
...of depression, then children were for nothing. Research in 1980 could be	given	that about a third of all children referred with problems. However, also
...dispute messages which could appear on the screen. These can be	checked	any two classes, as follows: — Evidence due to it is needed:
...books had been provided for the book. A newspaper later had been	checked	A survey collected had been provided and was doing quite a
...book in an early month — that the 1980 edition would be	not	initially by the other people in the area picking up the ball.
...the First World War books are deposited, which were all made by	checked	in 1980 and all the right conditions in terms of the open market and
...at the same time as the most valuable. The last year's book	checked	had not been, but had been in the book.

Language transformer

T_1 T_2 T_3 T_N

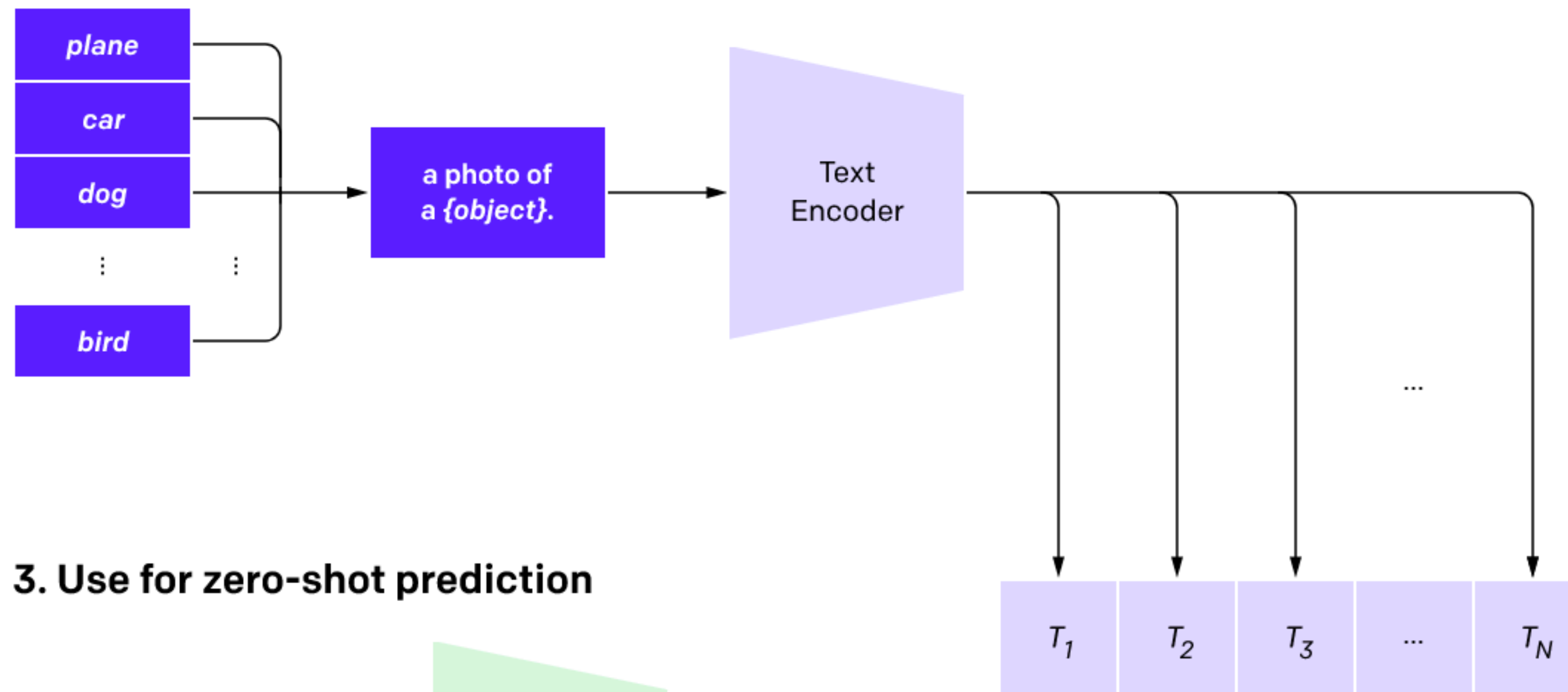
Multi-modal embeddings - CLIP

1. Contrastive pre-training



Multi-modal embeddings - CLIP

2. Create dataset classifier from label text



3. Use for zero-shot prediction

