Contrastive Learning

Theory, implementation and a popular example: CLIP

Tim Landgraf - Würzburg - 19 April 2023

Contrastive Learning

Theory, implementation and a popular example: CLIP

Tim Landgraf - Würzburg - 19 April 2023

„Learning“ = finding the best set of parameters

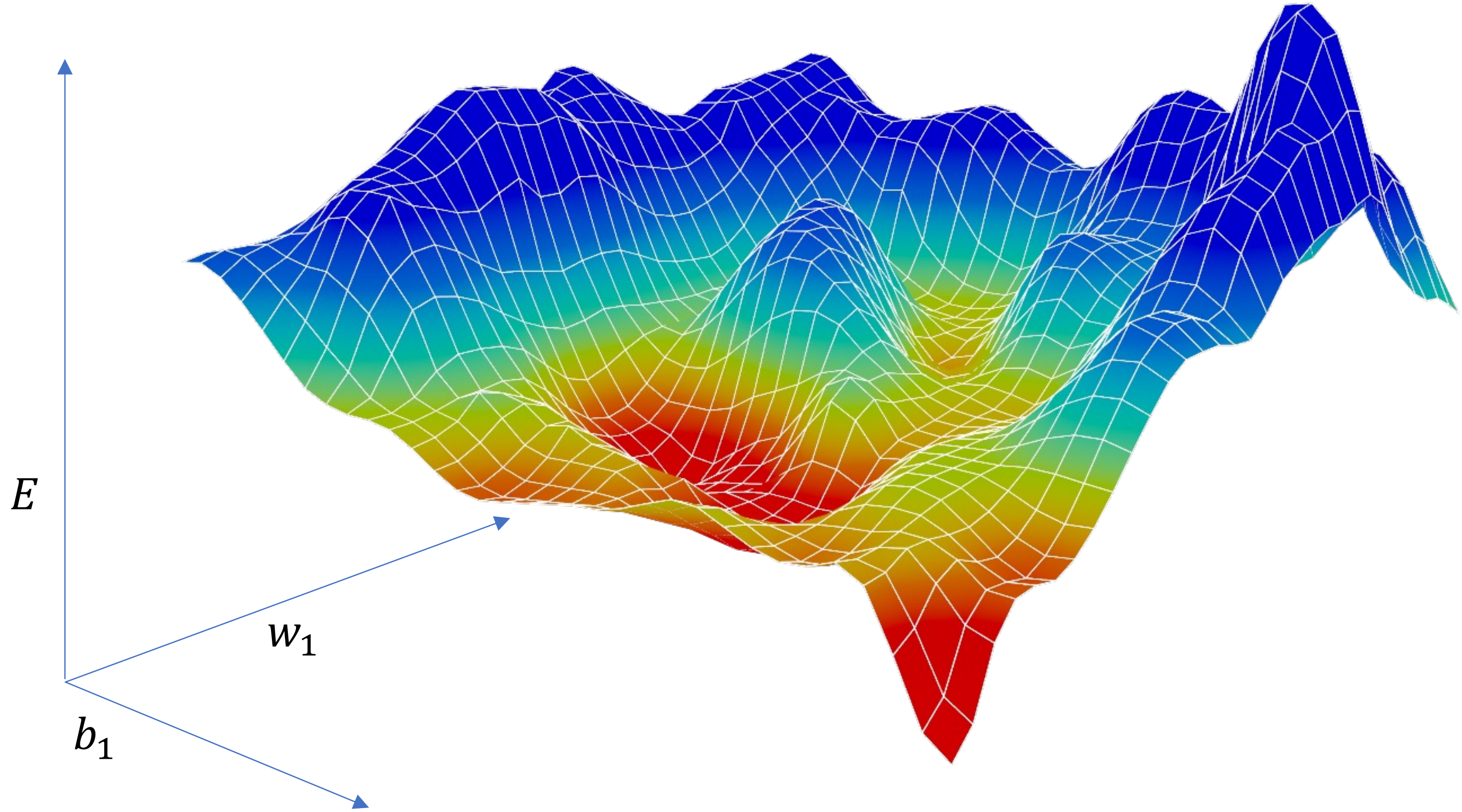
low-level abstraction
high-dimensional inputs

```
model.fit(train_images,
train_labels, epochs=5)
```

high-level abstraction
low-dimensional outputs

- cat
- dog
- flower
- shoe

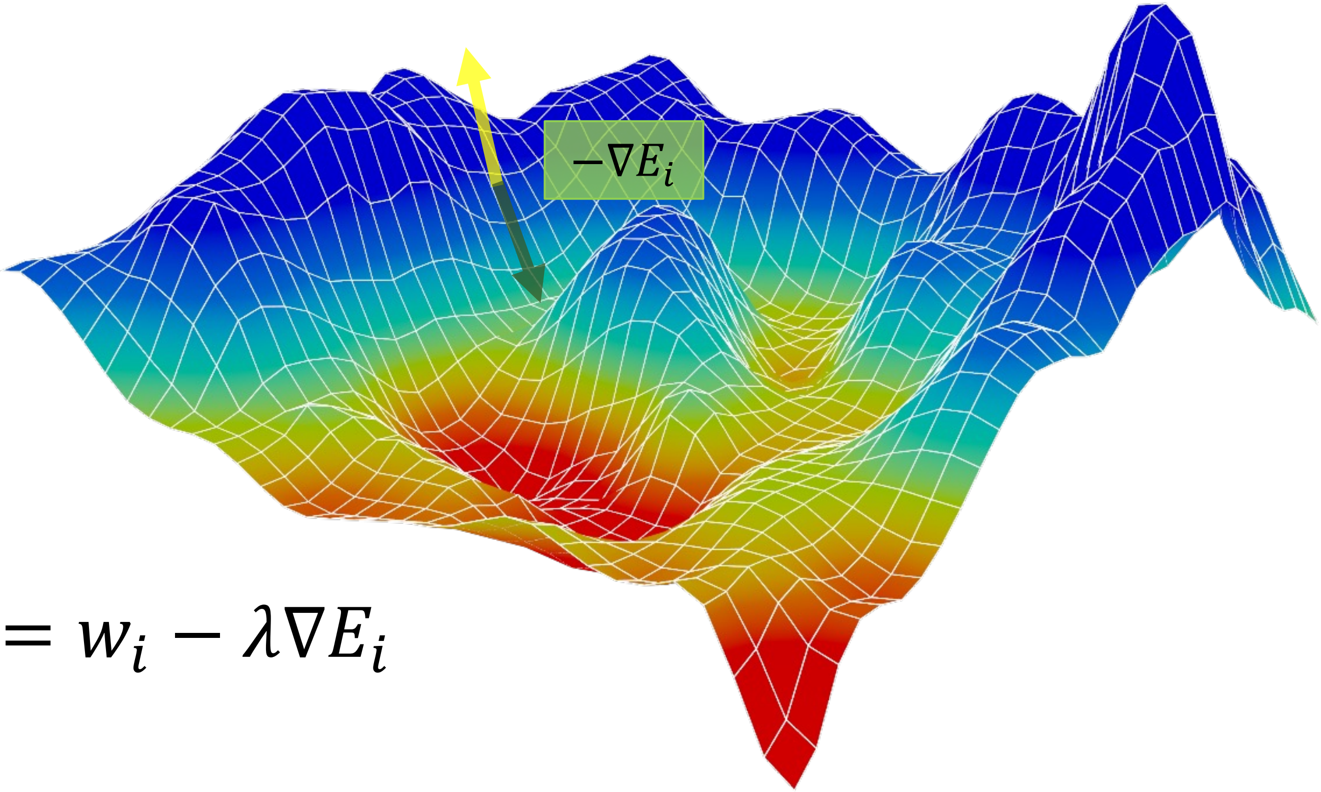
Network error as a function of the model parameters (weights and biases)



Goal: Find weight combination that minimizes network error

„Learning“ = walking downhill

The **gradient** points **uphill**, so let‘s walk in **the opposite direction**.

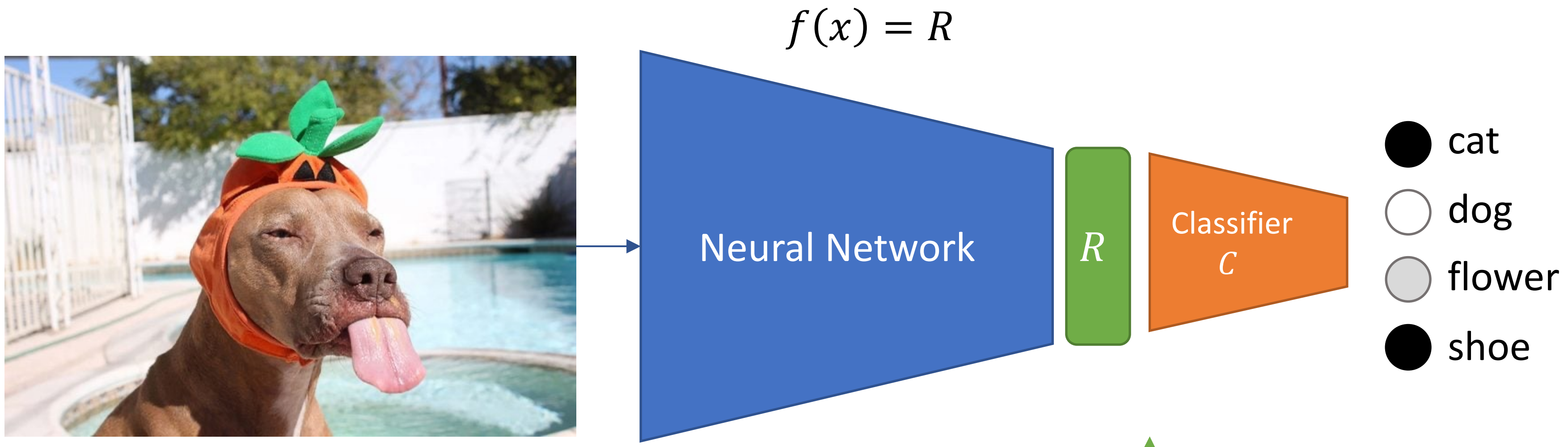


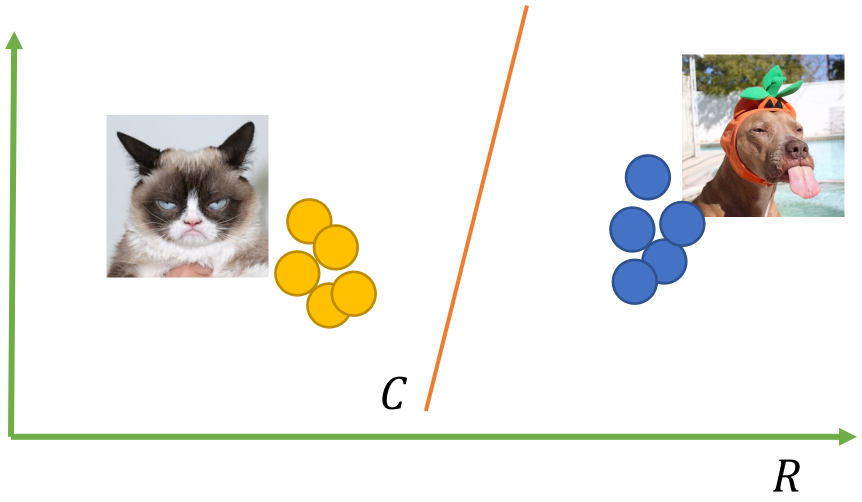
$$w_{i+1} = w_i - \lambda \nabla E_i$$

„How much does the error change, for a unit change of weights“

$$\nabla E = \frac{\partial E}{\partial w} = \begin{pmatrix} \frac{\partial E}{\partial w_1} \\ \cdots \\ \frac{\partial E}{\partial w_N} \end{pmatrix}$$

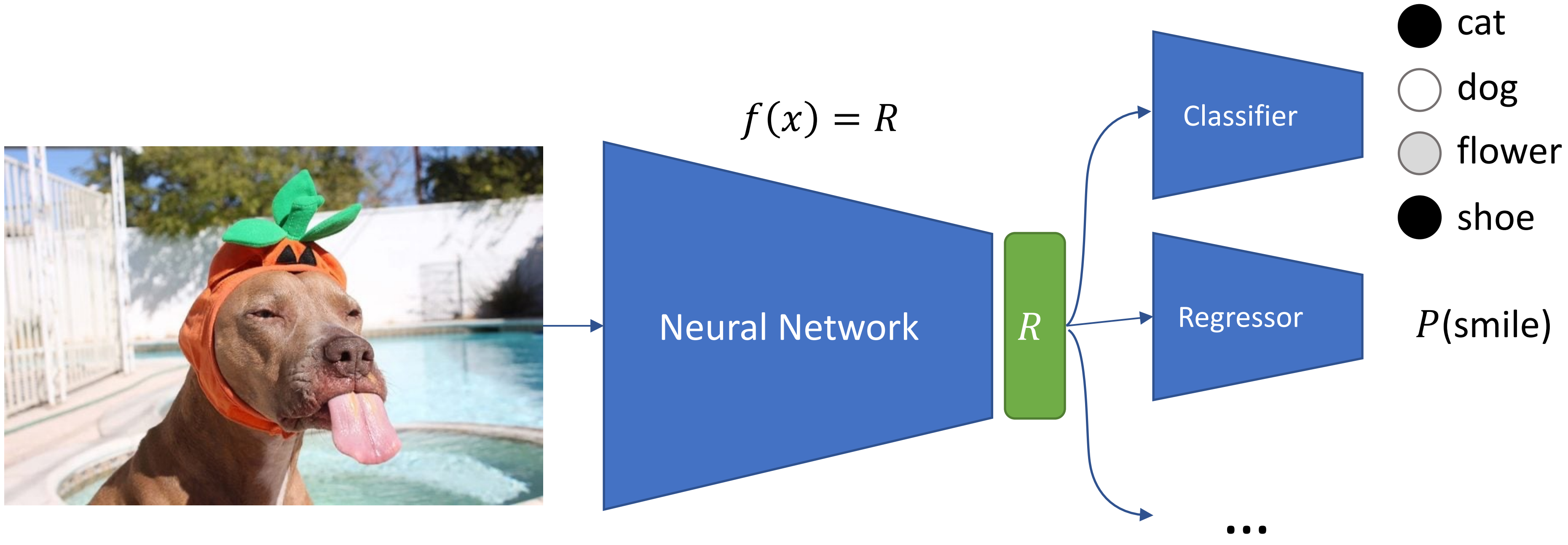
Learning = finding „good“ *representations*





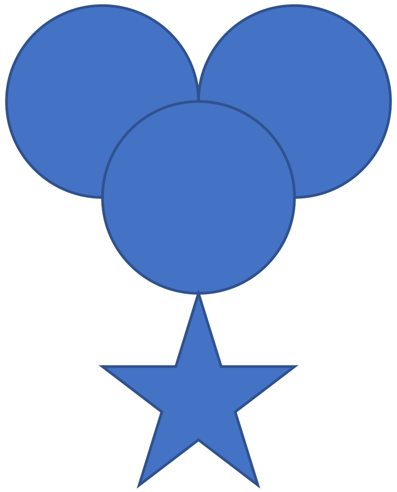
- Neural networks can be seen as simultaneously …
  - learn **representations** of input data
  - **classify** from those representations

Contrastive Learning used for …

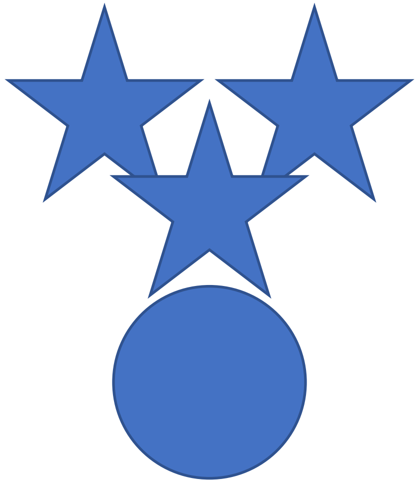


1. Dimensionality *Reduction*
2. Learning useful *representations*
3. *Robust* training objectives
4. *Efficient* use of data

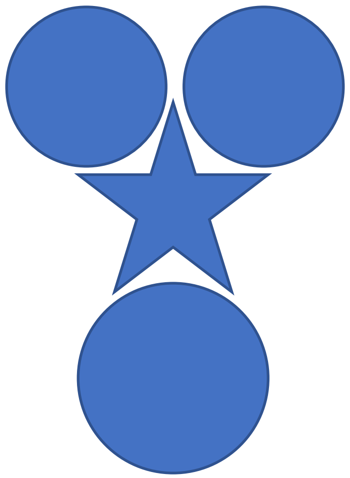
How do networks learn non-contrastive?



GLORBOXL

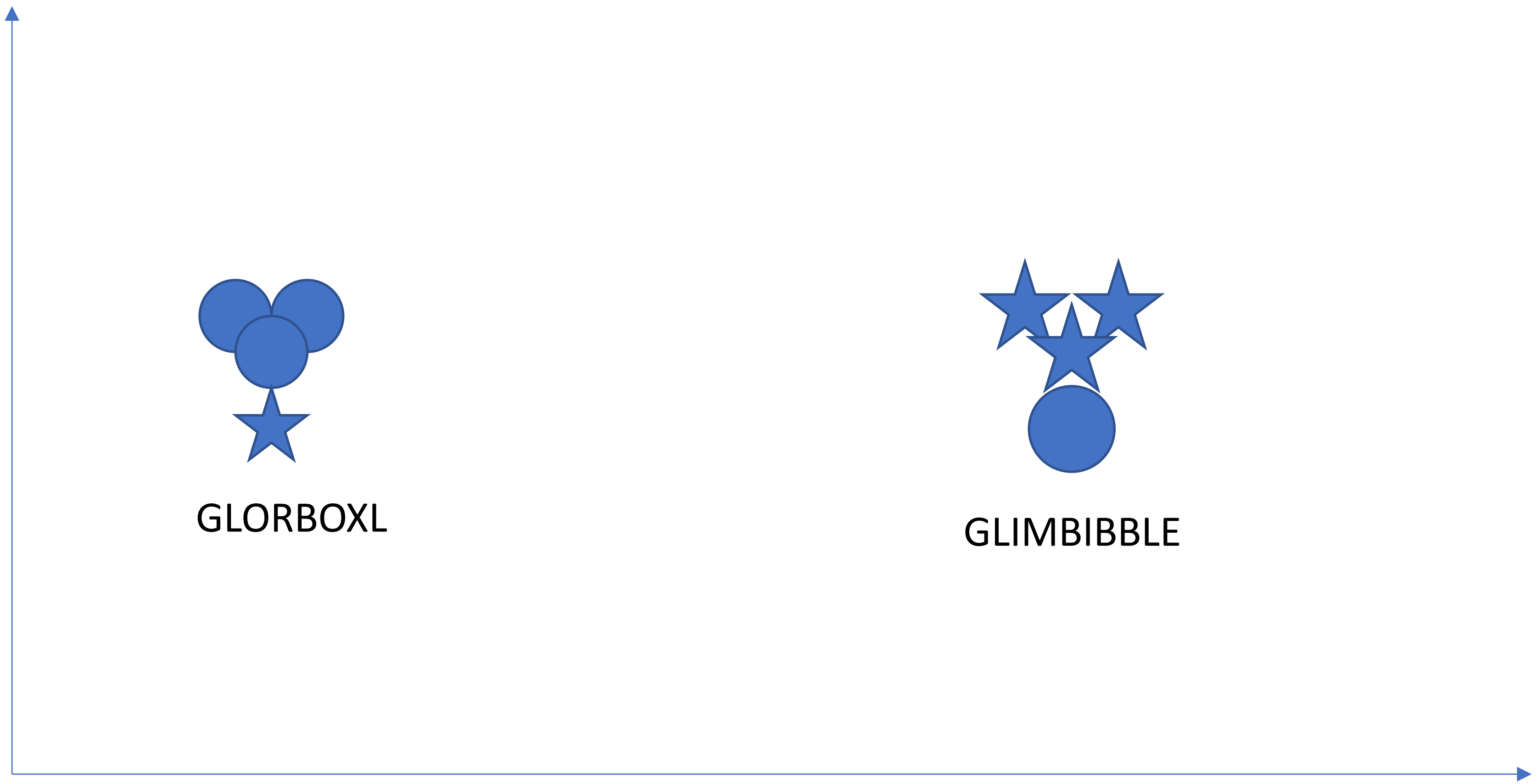


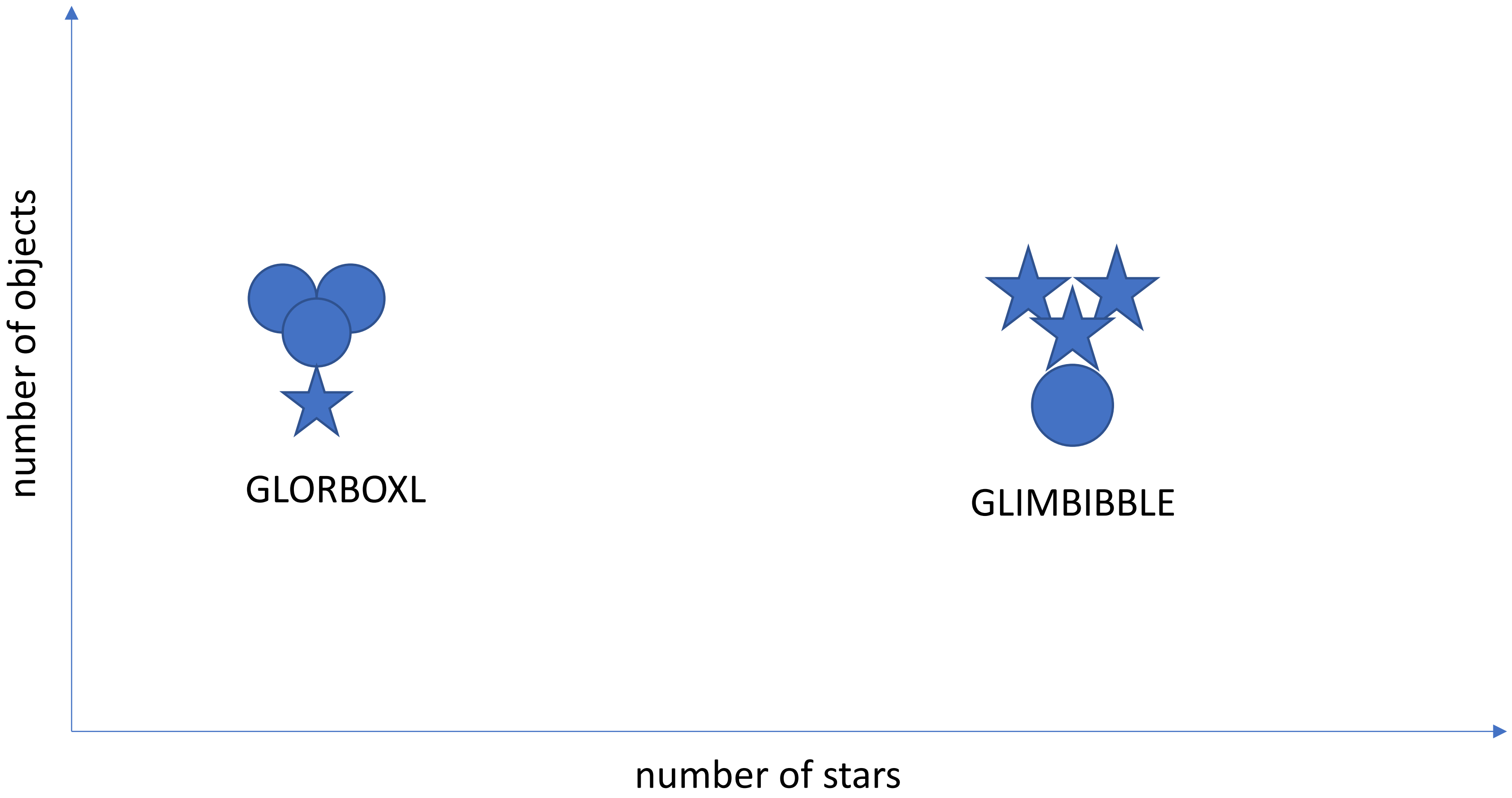
GLIMBIBBLE

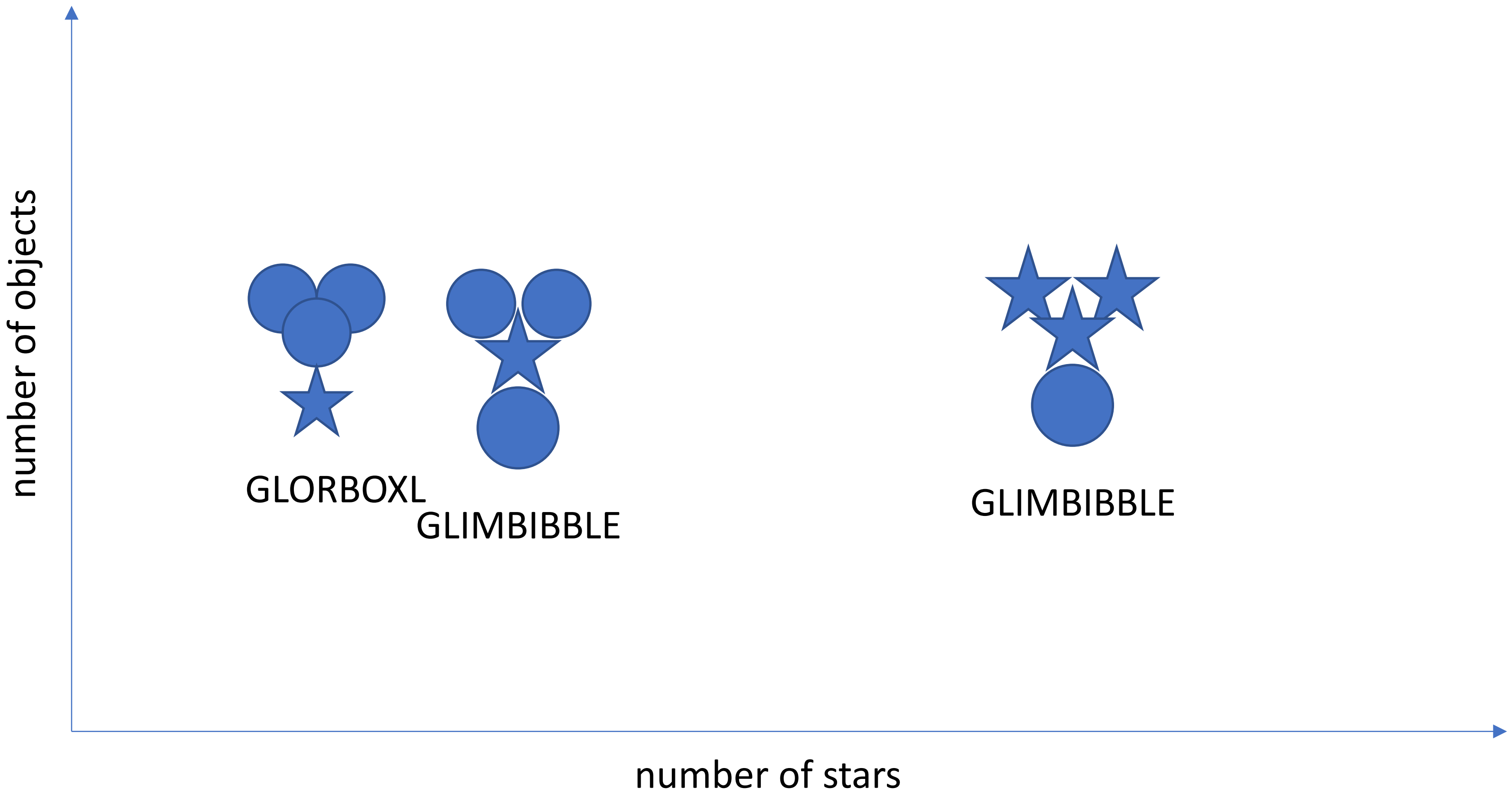


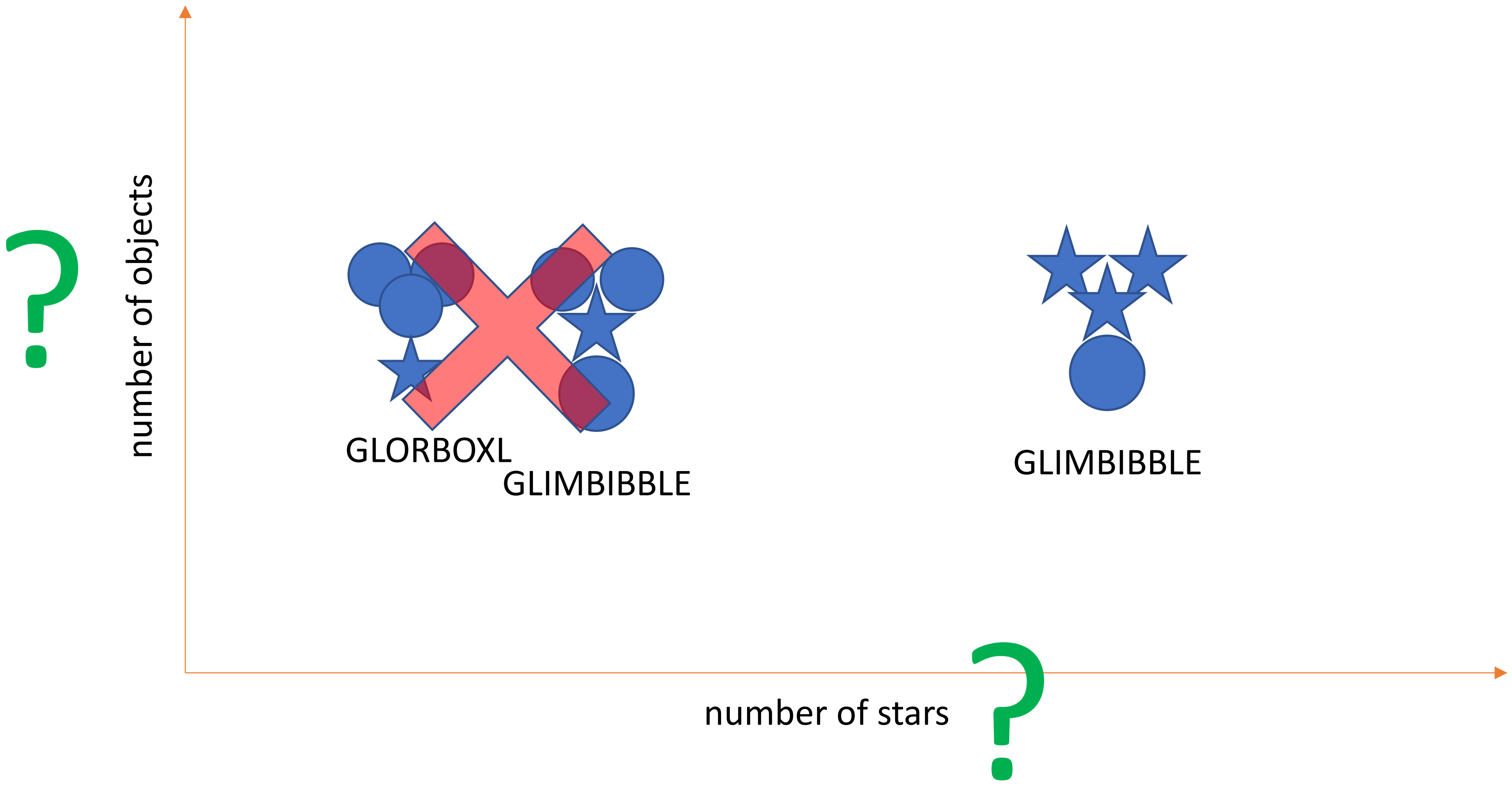
GLIMBIBBLE

GLORBOXL

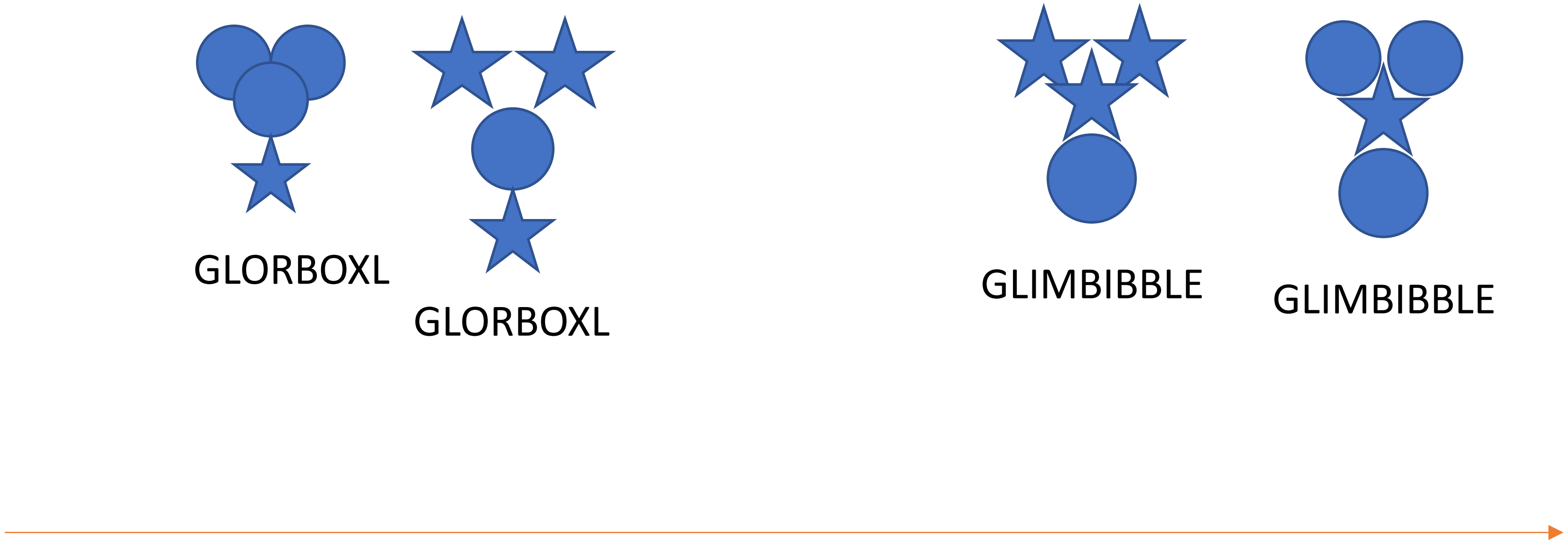

GLORBOXL
GLIMBIBBLE

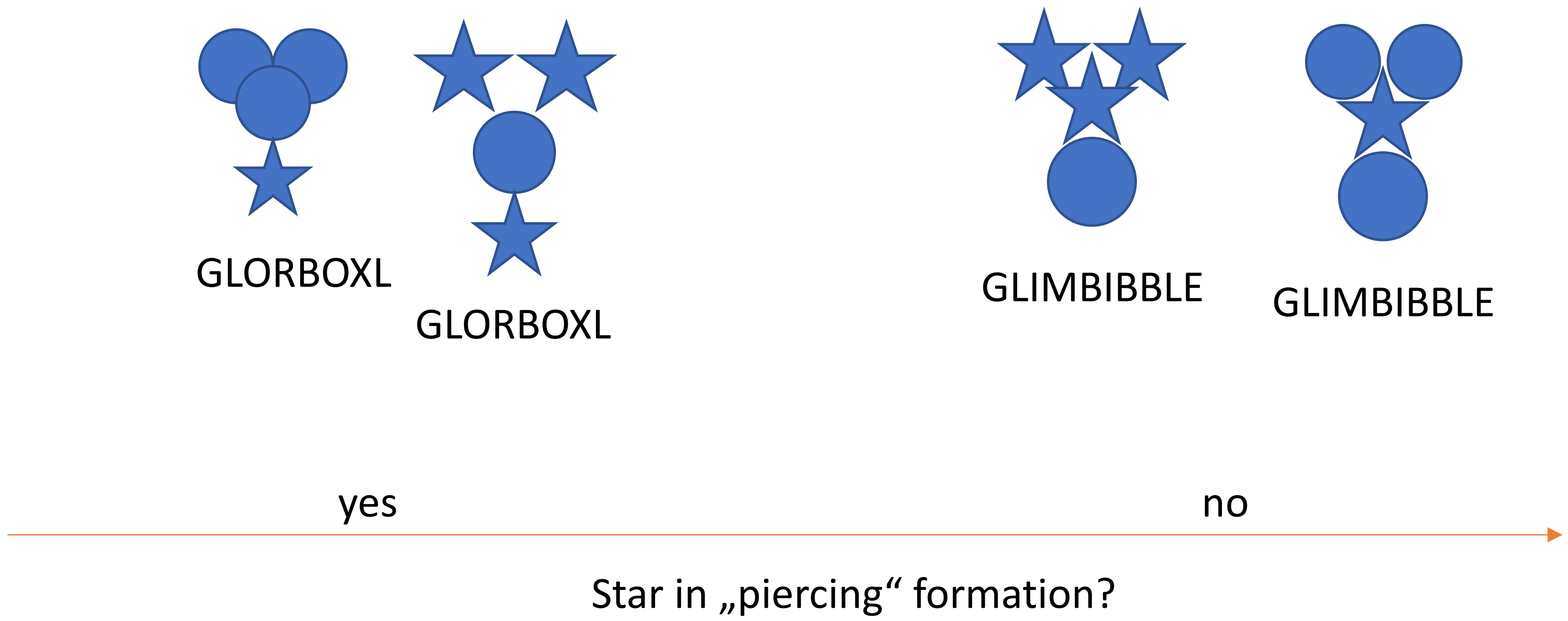

number of objects
GLORBOXL
GLIMBIBBLE
number of stars


number of objects
GLORBOXL
GLIMBIBBLE
GLIMBIBBLE
number of stars



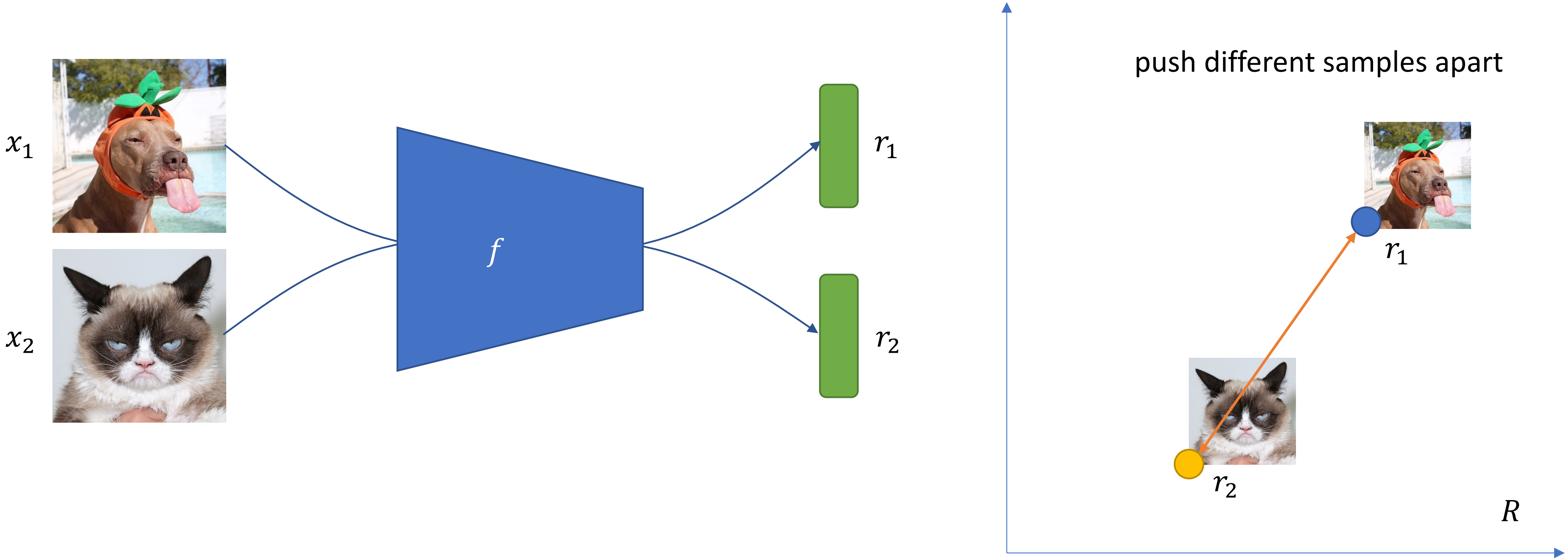
?
number of objects
GLORBOXL
GLIMBIBBLE
GLIMBIBBLE
number of stars
?


GLORBOXL
GLORBOXL
GLIMBIBBLE
GLIMBIBBLE

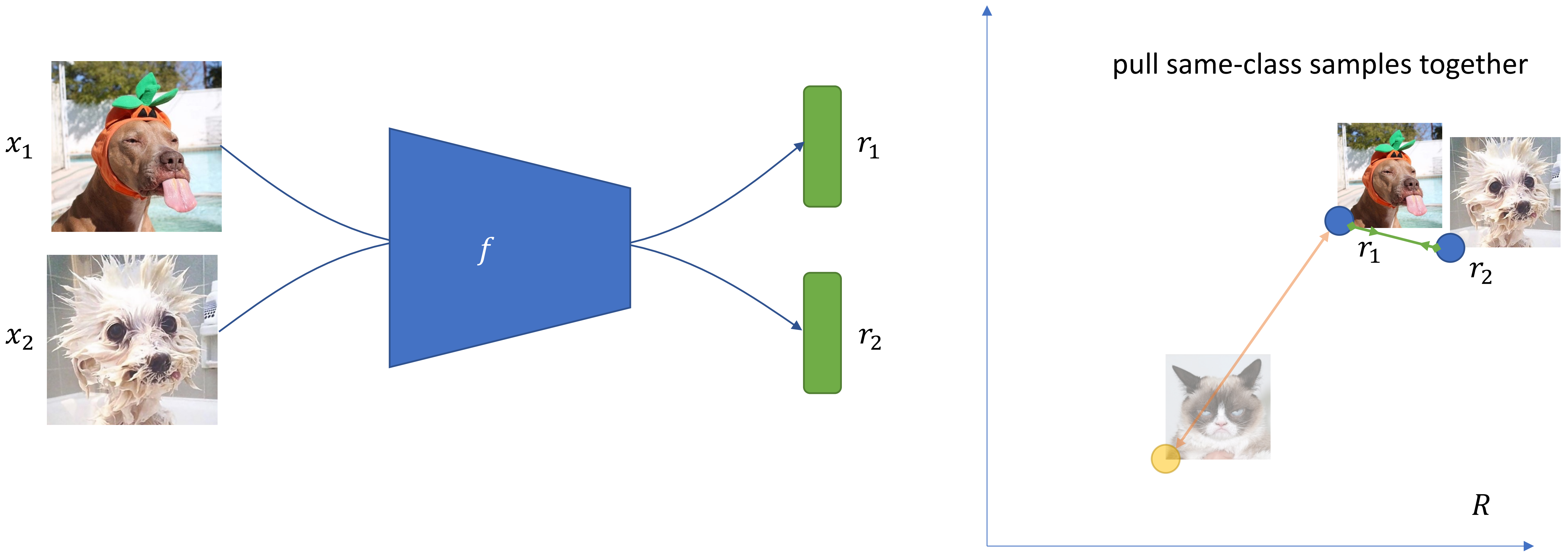

GLORBOXL
GLORBOXL
GLIMBIBBLE
GLIMBIBBLE
yes
no
Star in „piercing“ formation?

Sixt, L., Schuessler, M., Popescu, O. I., Weiß, P., & Landgraf, T. Do Users Benefit From Interpretable Vision? A User Study, Baseline, And Dataset. In *International Conference on Learning Representations*.

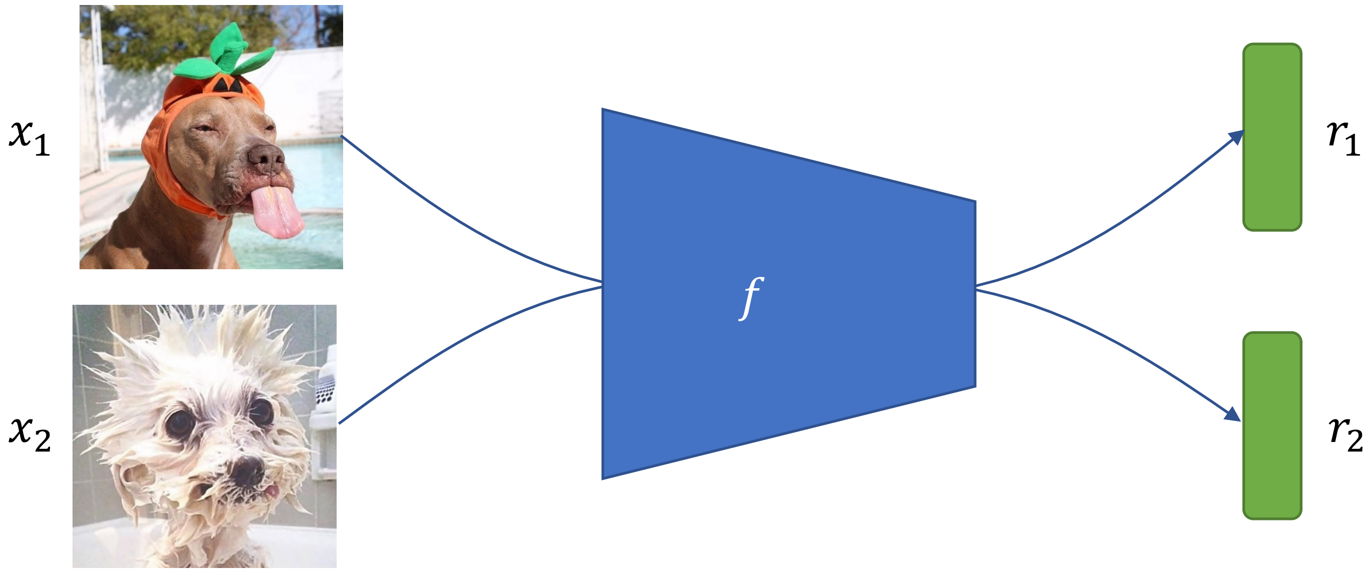
Contrastive Learning Idea (Supervised Case)

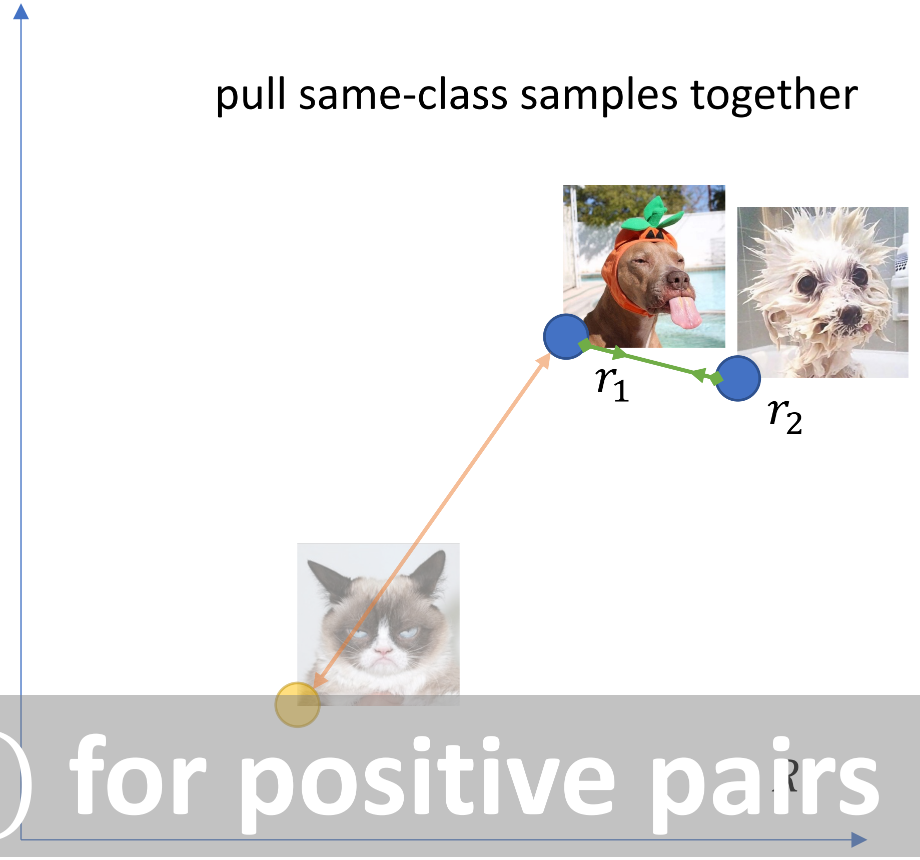


Contrastive Learning Idea (Supervised Case)



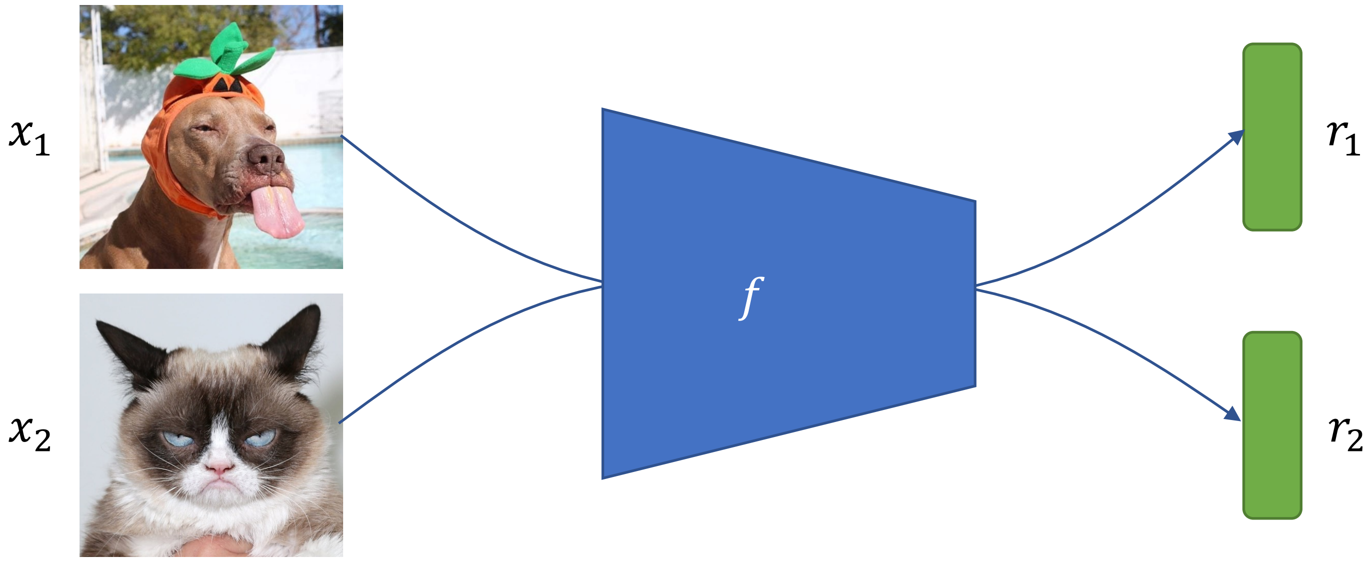
Contrastive Learning Idea (Supervised Case)

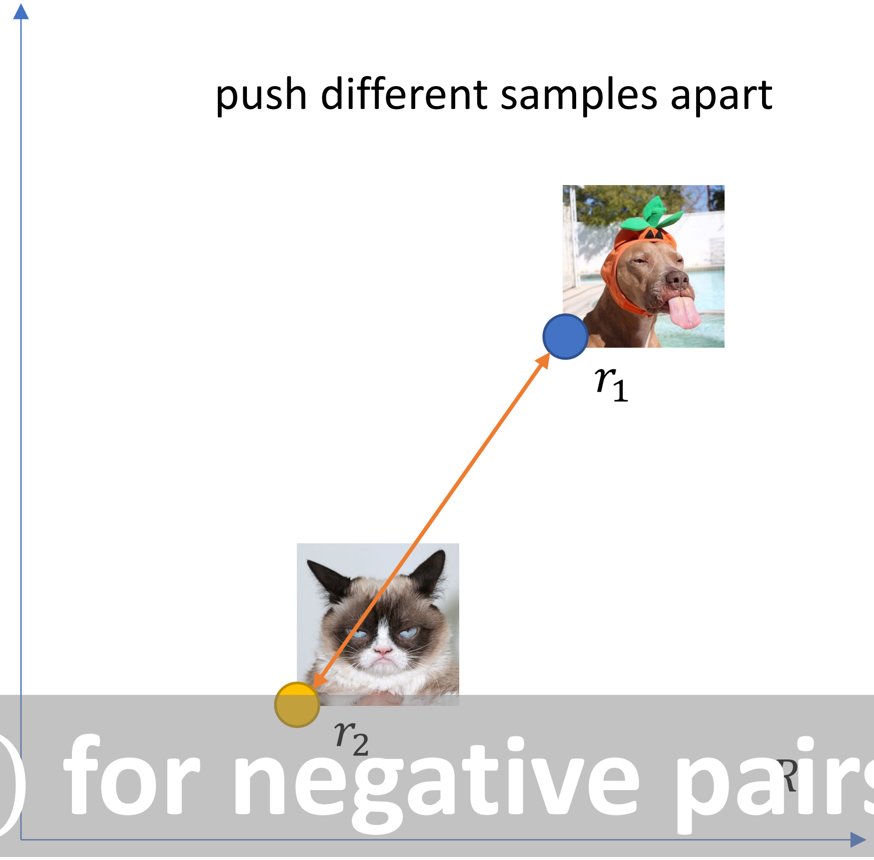




→ minimize distance $d(r_1, r_2)$ for positive pairs

Contrastive Learning Idea (Supervised Case)





→ maximize distance $d(r_1, r_2)$ for negative pairs

Simple Contrastive Loss

$$\boldsymbol{q} = \begin{cases} \mathbf{1}, & \boldsymbol{y_i} = \boldsymbol{y_j} \\ \mathbf{0}, & \boldsymbol{y_i} \neq \boldsymbol{y_j} \end{cases}$$

same label

different label

$$\boldsymbol{L}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{q} d(r_{\boldsymbol{i}}, r_{\boldsymbol{j}}) - (\mathbf{1} - \boldsymbol{q}) d(r_{\boldsymbol{i}}, r_{\boldsymbol{j}})$$

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 539-546). IEEE.

Simple Contrastive Loss

$$q = \begin{cases} 1, & y_i = y_j \\ 0, & y_i \neq y_j \end{cases}$$

same label

different label

$$L(x_i, x_j) = q\,d(r_i, r_j) - (1 - q)\,d(r_i, r_j)$$

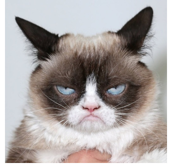
Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 539-546). IEEE.

Simple Contrastive Loss

$$
\boldsymbol{q} = \begin{cases} \mathbf{1}, & \boldsymbol{y_i} = \boldsymbol{y_j} \\ \mathbf{0}, & \boldsymbol{y_i} \neq \boldsymbol{y_j} \end{cases}
$$

same label

different label

$$
\boldsymbol{L}(\boldsymbol{x_i}, \boldsymbol{x_j}) = q d(r_i, r_j) \underbrace{- (\mathbf{1} - \boldsymbol{q}) d(\boldsymbol{r_i}, \boldsymbol{r_j})}_{\textit{is minimized when } d(r_i, r_j) \textit{ maximized}}
$$

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 539-546). IEEE.

Contrastive Loss with Margin

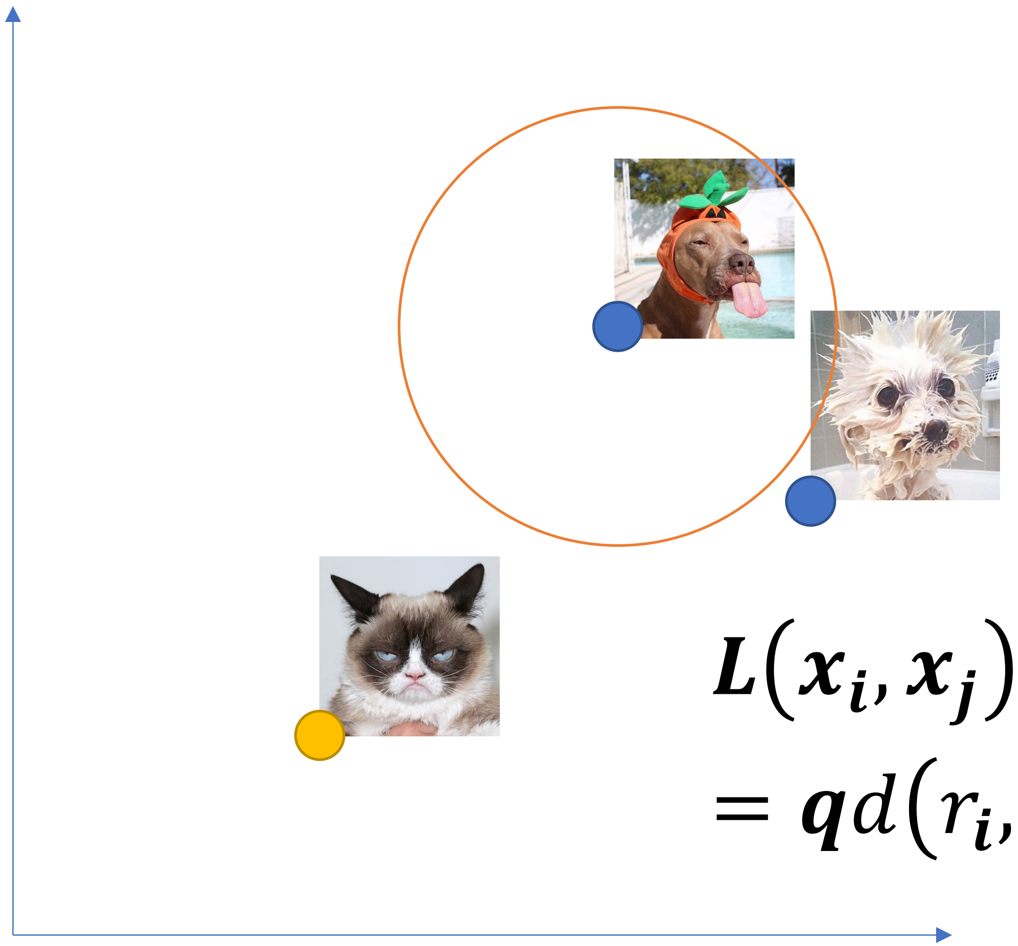
$$\boldsymbol{q} = \begin{cases} \mathbf{1}, & \boldsymbol{y_i} = \boldsymbol{y_j} \\ \mathbf{0}, & \boldsymbol{y_i} \neq \boldsymbol{y_j} \end{cases}$$

$$\boldsymbol{L}(\boldsymbol{x_i}, \boldsymbol{x_j}) \\ = \boldsymbol{q} d(r_{\boldsymbol{i}}, r_{\boldsymbol{j}}) - (\mathbf{1} - \boldsymbol{q})\boldsymbol{max}\left(\mathbf{0}, \boldsymbol{m} - d(r_{\boldsymbol{i}}, r_{\boldsymbol{j}})\right)$$

margin m

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2, pp. 1735-1742). IEEE.

Contrastive Loss with Margin



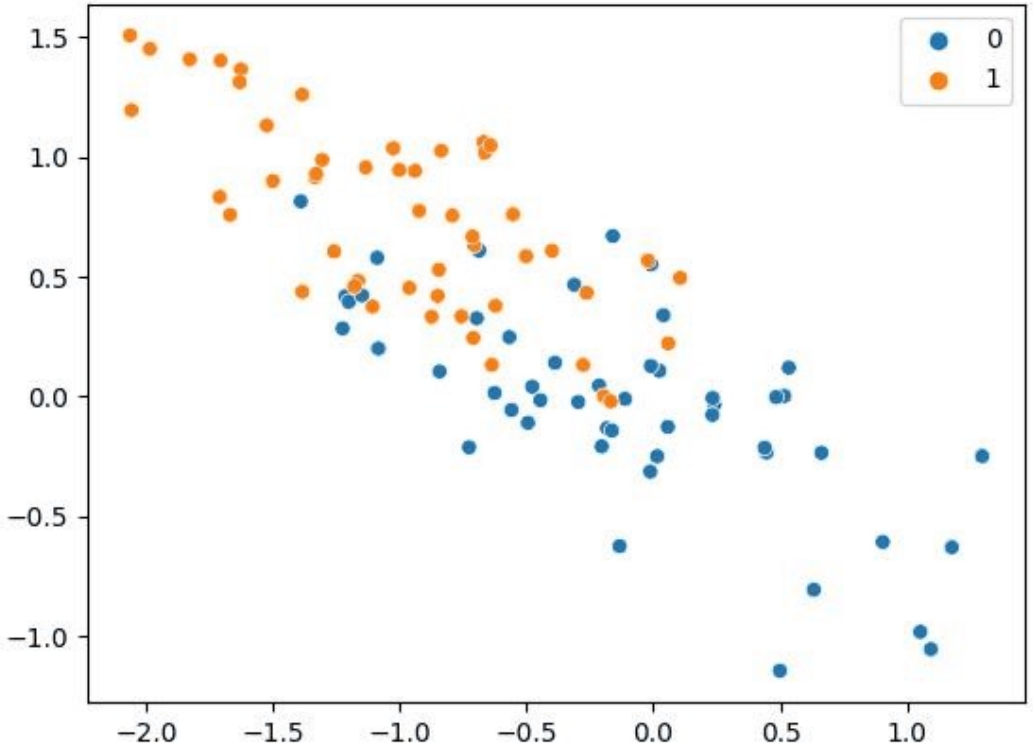
$$\boldsymbol{L}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{q} d(r_i, r_j) - (\mathbf{1} - \boldsymbol{q}) \boldsymbol{max} \overbrace{\left(\mathbf{0}, \boldsymbol{m} - d(r_i, r_j)\right)}^{\textit{results in 0, because grumpy cat is too far away}}$$

https://github.com/timlandgraf/contrastive_learning_code_examples

Example with code!

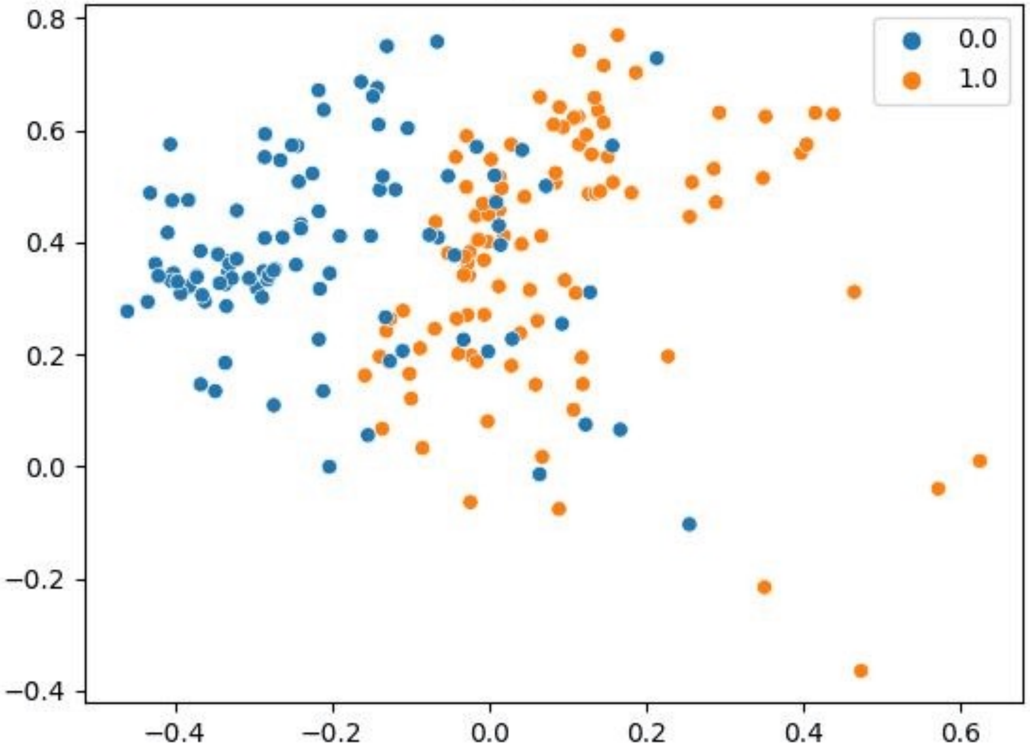
$x \in \mathbb{R}^5, r \in \mathbb{R}^2$

linear mapping



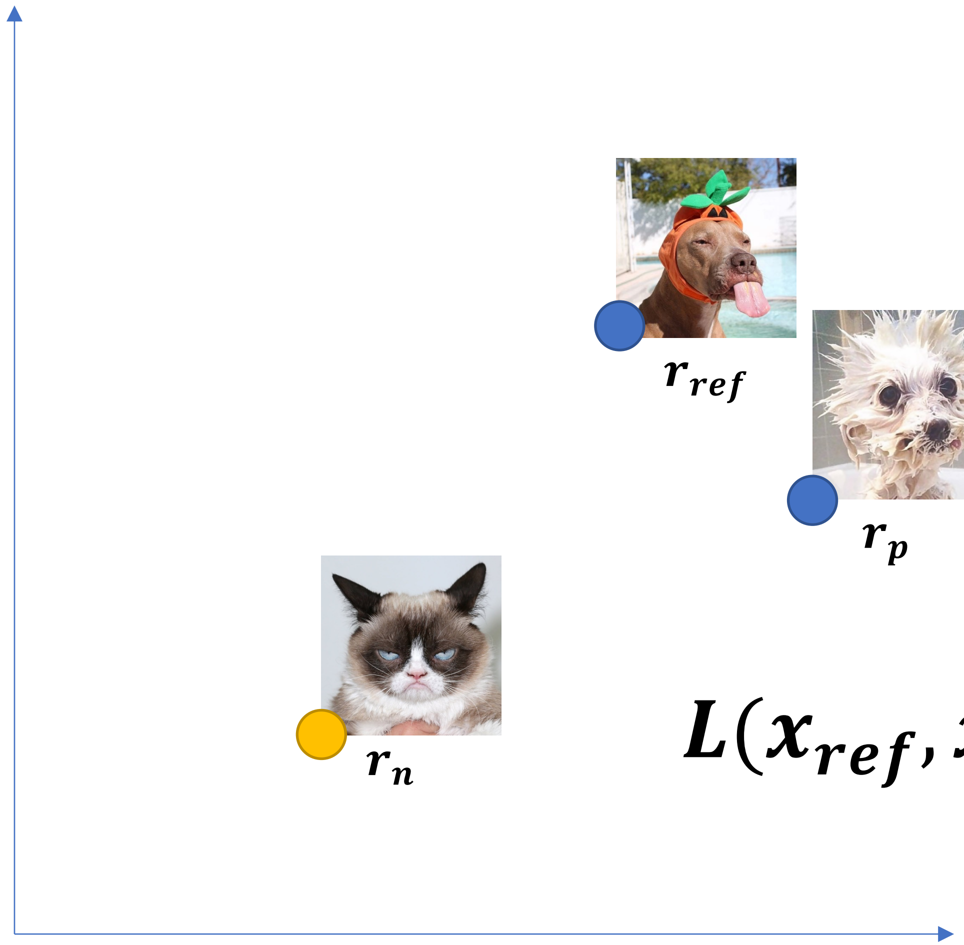
$$Ax = r$$

nonlinear mapping



$$NN(x) = r$$

Other Contrastive Losses

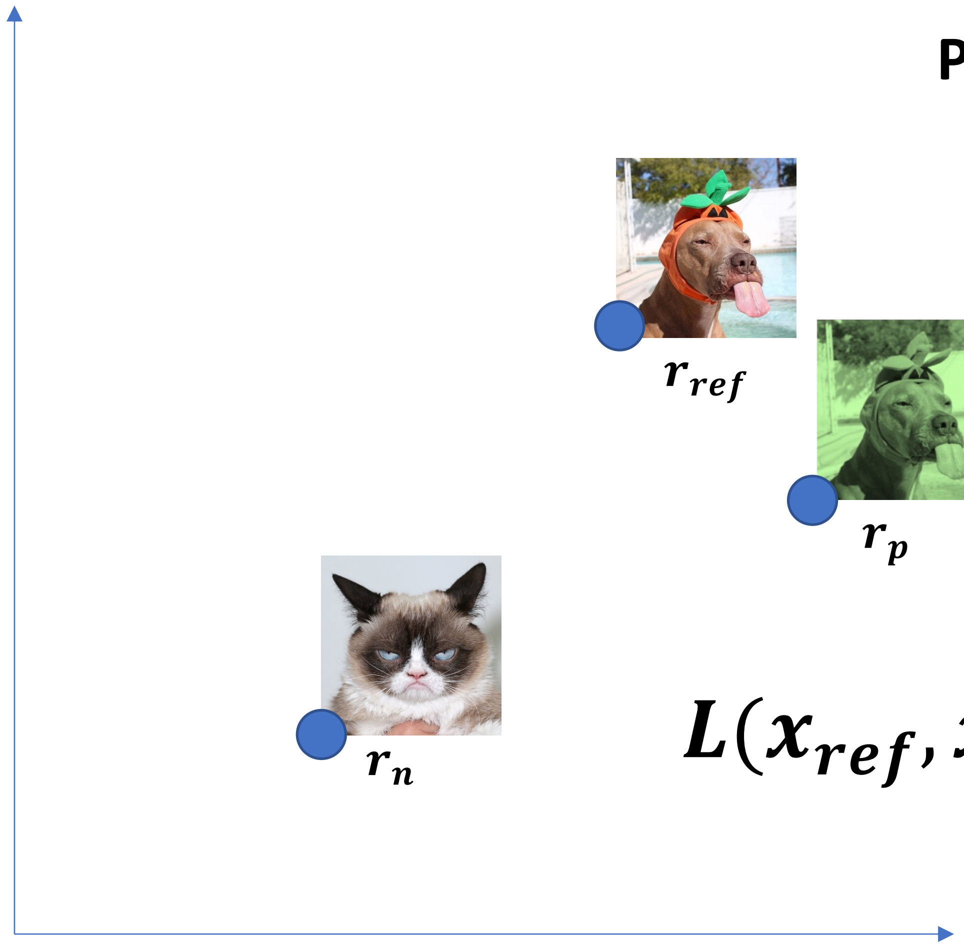


„Triplet Loss“

$$L(x_{ref}, x_p, x_n) = d(r_{ref}, r_p) + \alpha - d(r_{ref}, r_n)$$

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

Contrastive Unsupervised Learning



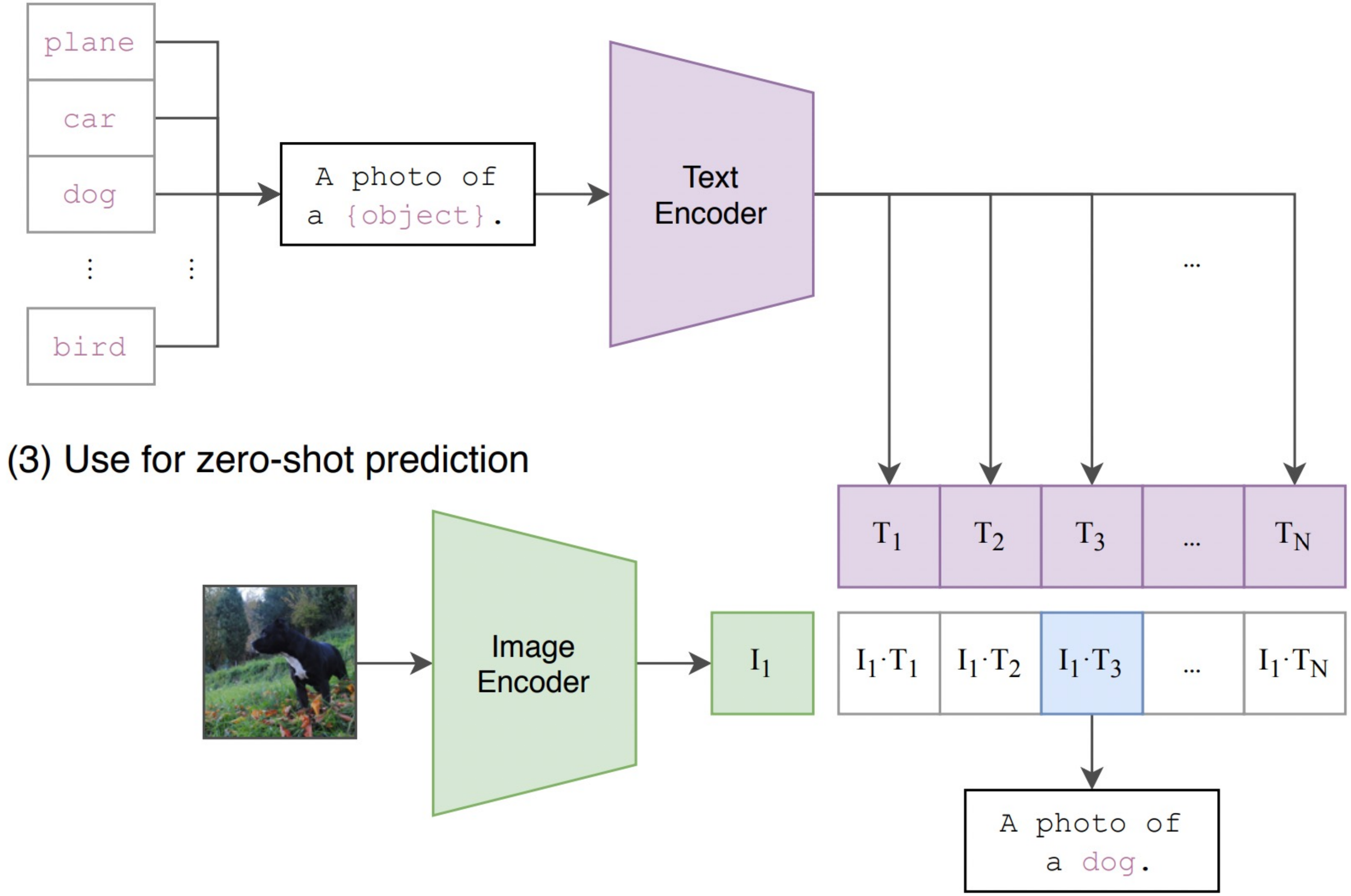
Positive samples generated by image augmentations

- Color shifts
- Added noise
- Shear, rotations
- Crops
- …

$$L(x_{ref}, x_p, x_n) = d(r_{ref}, r_p) + \alpha - d(r_{ref}, r_n)$$

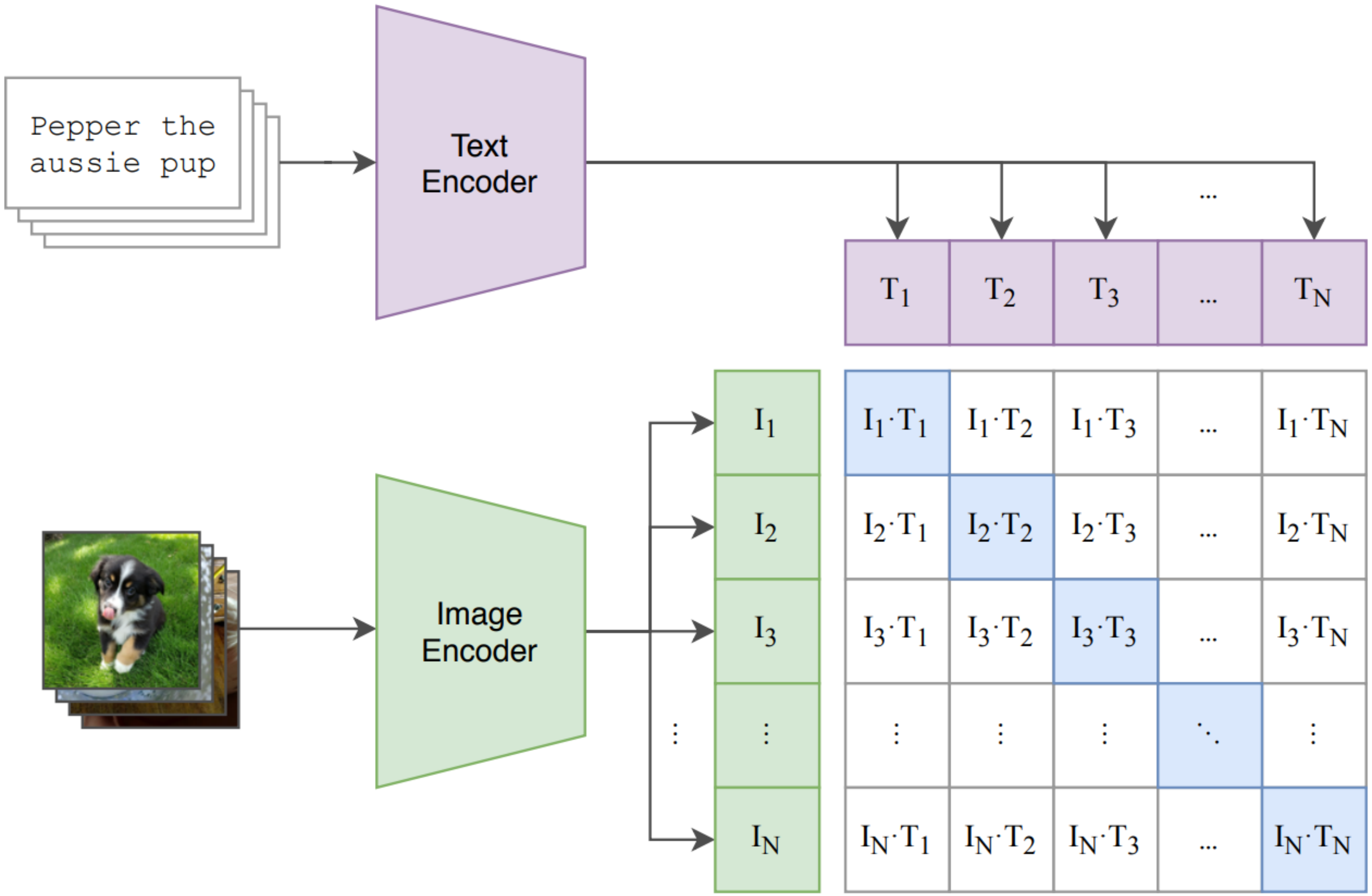
Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

Contrastive Loss for learning multi-modal representations: CLIP!



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Contrastive Loss for learning multi-modal representations: CLIP!



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

https://openai.com/research/clip

Food101

guacamole (90.1%) Ranked 1 out of 101 labels

- ✓ a photo of **guacamole**, a type of food.
- ✗ a photo of **ceviche**, a type of food.
- ✗ a photo of **edamame**, a type of food.
- ✗ a photo of **tuna tartare**, a type of food.
- ✗ a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397 labels

- ✓ a photo of a **television studio**.
- ✗ a photo of a **podium indoor**.
- ✗ a photo of a **conference room**.
- ✗ a photo of a **lecture room**.
- ✗ a photo of a **control room**.

Youtube-BB

airplane, person (89.0%) Ranked 1 out of 23 labels

- ✓ a photo of a **airplane**.
- ✗ a photo of a **bird**.
- ✗ a photo of a **bear**.
- ✗ a photo of a **giraffe**.
- ✗ a photo of a **car**.

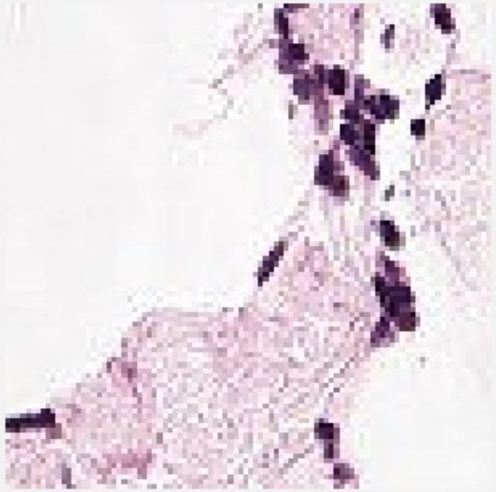
EuroSAT

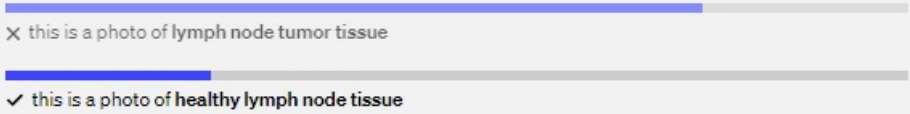
annual crop land (46.5%) Ranked 4 out of 10 labels

- ✗ a centered satellite photo of **permanent crop land**.
- ✗ a centered satellite photo of **pasture land**.
- ✗ a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- ✗ a centered satellite photo of **brushland or shrubland**.

PatchCamelyon (PCam)

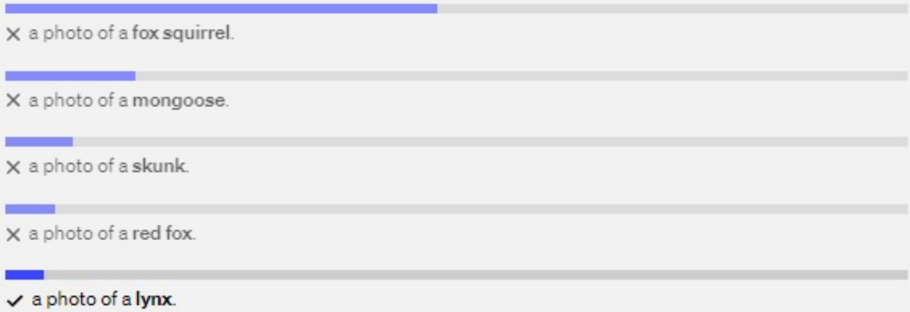
healthy lymph node tissue (77.2%) Ranked 2 out of 2 labels





ImageNet-A (Adversarial)

lynx (47.9%) Ranked 5 out of 200 labels



https://openai.com/research/clip