POLITECNICO
MILANO 1863

Concept-Based Explanations for Image Classifiers Using Textual Prompts

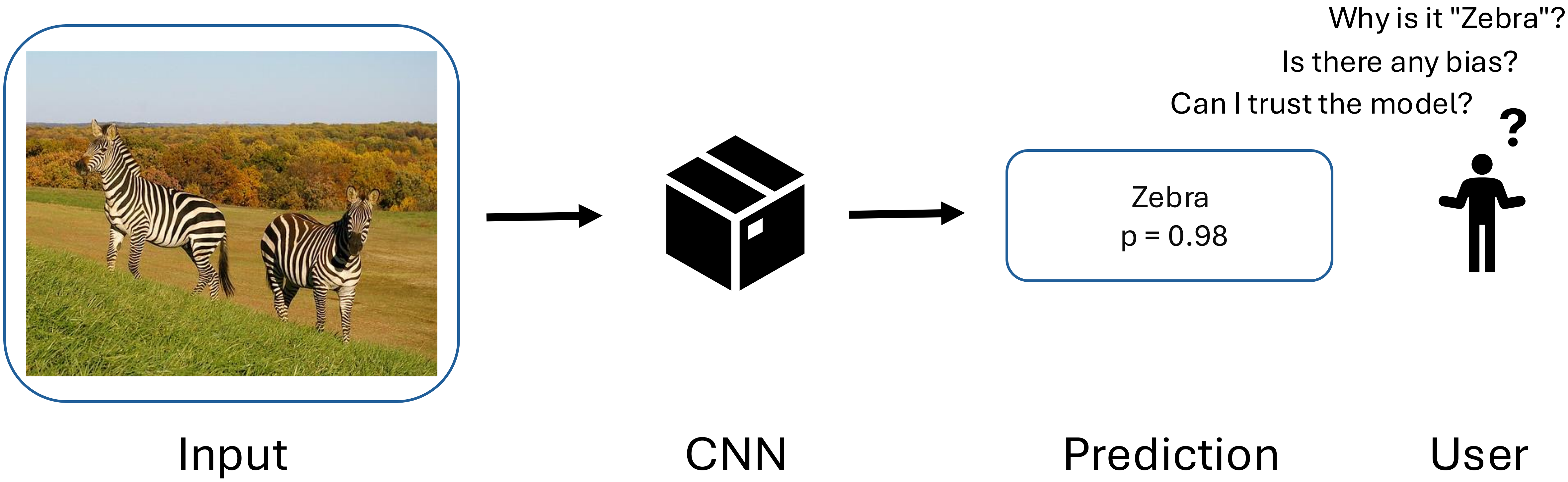
Daniele Di Santi

Academic year: 2024/2025

Advisor: Prof. Marco Brambilla

Co-advisors: Riccardo Campi, Matteo Bianchi, Antonio De Santis

Black-box Systems

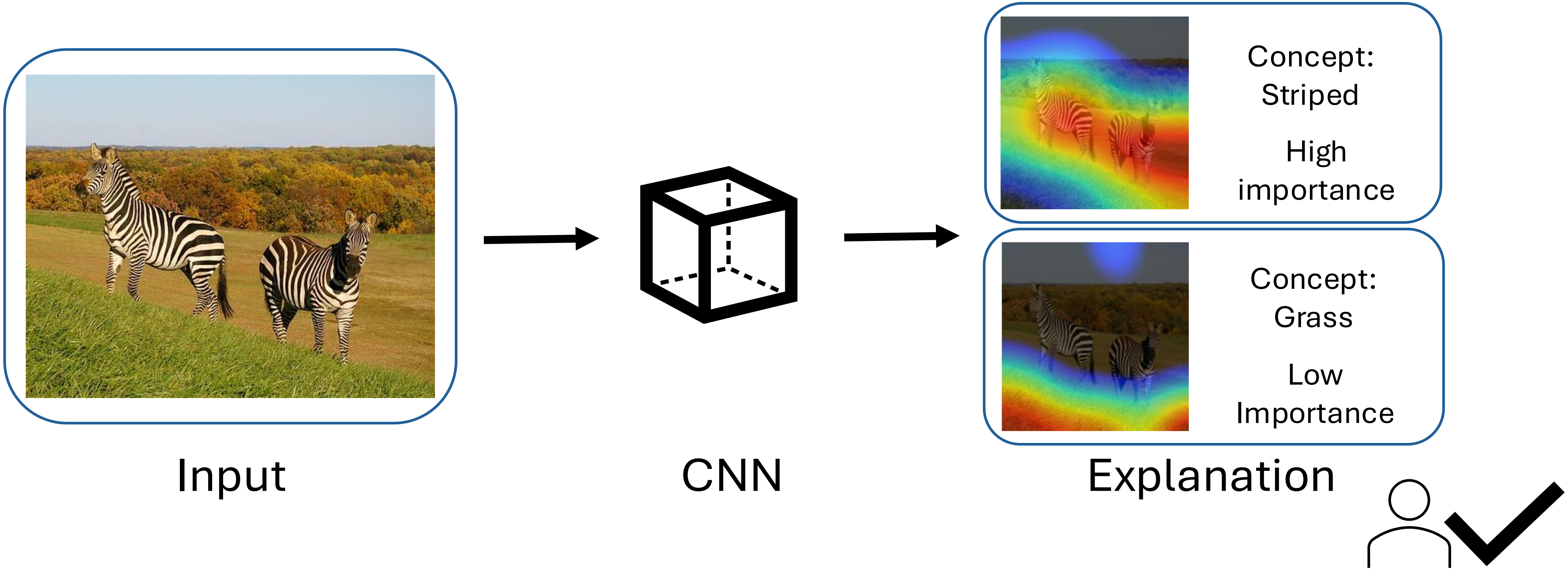


Why is it "Zebra"?

Is there any bias?

Can I trust the model?

Explainable Artificial Intelligence (XAI)



Concept: Striped

High importance

Concept: Grass

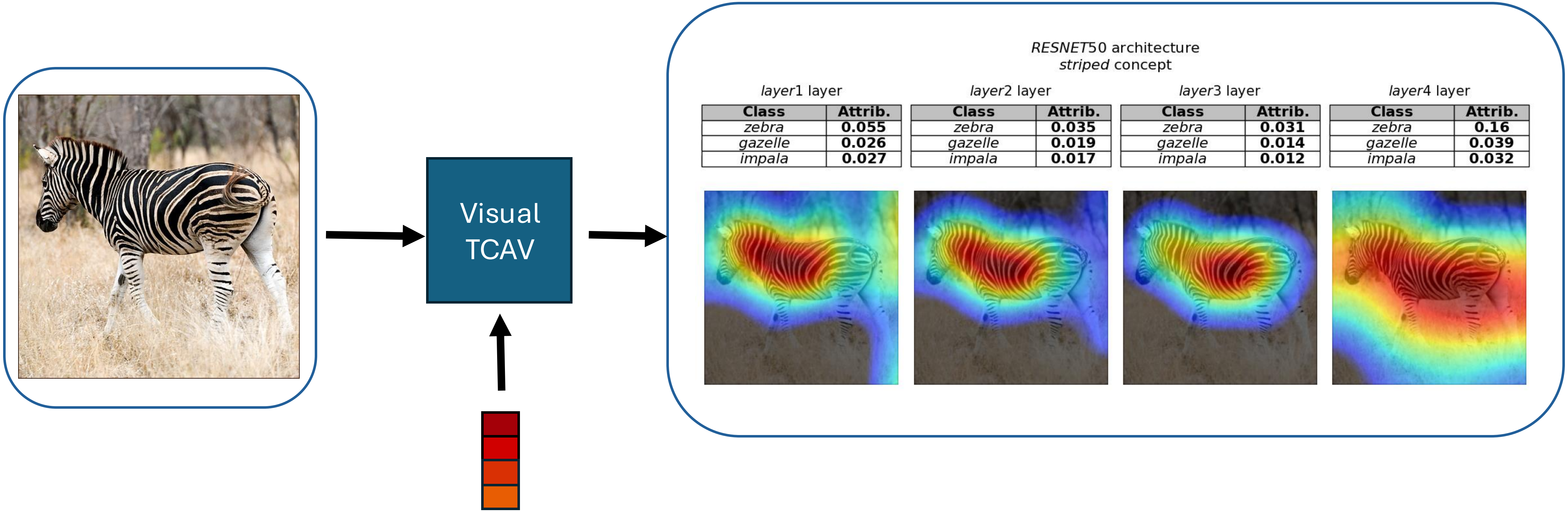
Low Importance

Input

CNN

Explanation

Visual TCAV



*RESNET*50 architecture
striped concept

*layer*1 layer

Class	Attrib.
zebra	**0.055**
gazelle	**0.026**
impala	**0.027**

*layer*2 layer

Class	Attrib.
zebra	**0.035**
gazelle	**0.019**
impala	**0.017**

*layer*3 layer

Class	Attrib.
zebra	**0.031**
gazelle	**0.014**
impala	**0.012**

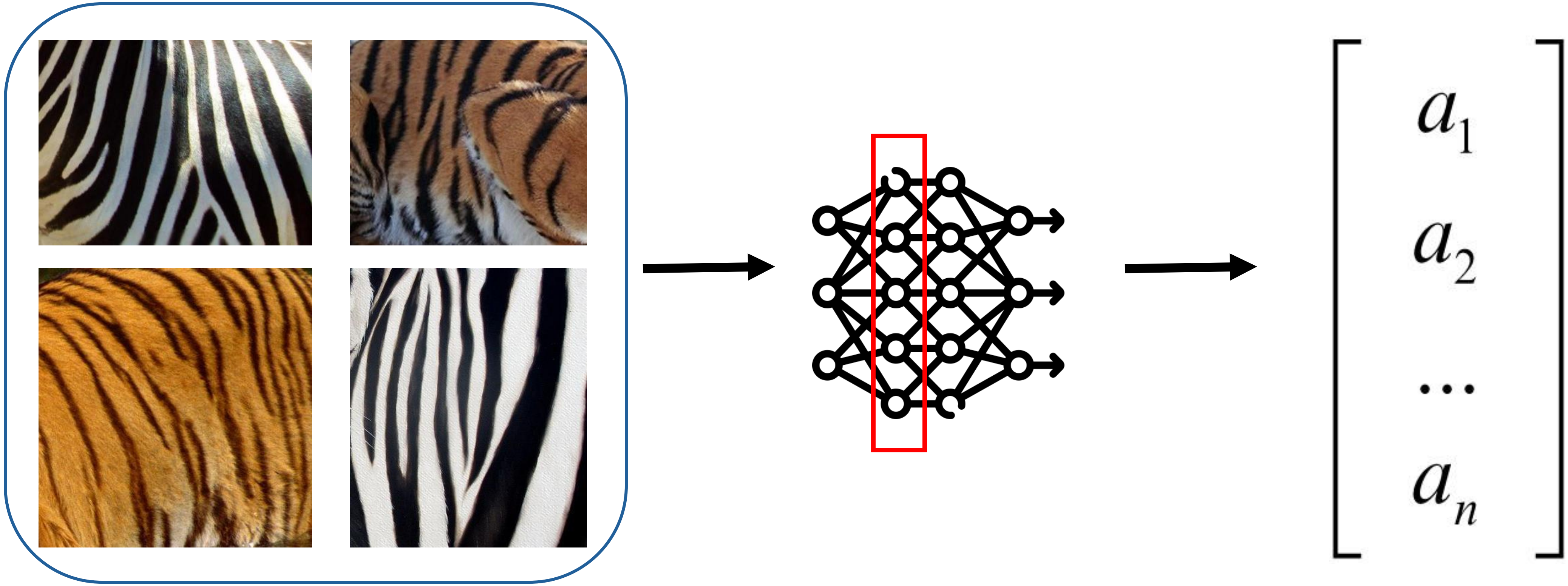
*layer*4 layer

Class	Attrib.
zebra	**0.16**
gazelle	**0.039**
impala	**0.032**

Input image **CAV** **Concept map + concept attributions**

Antonio De Santis et al., Visual-tcav: Concept-based attribution and saliency maps for post-hoc explainability in image classification, 2025.

Problem: Standard CAV Extraction



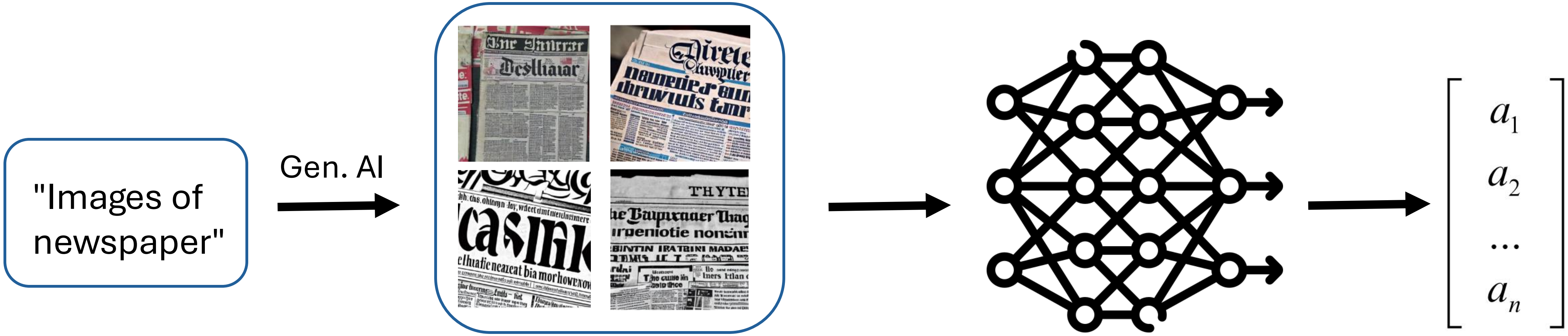
$$\begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{bmatrix}$$

Set of concept images

CNN

CAV

Other Solution: Synthetic Images



"Images of newspaper"

Gen. AI

$$\begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{bmatrix}$$

Textual Prompt

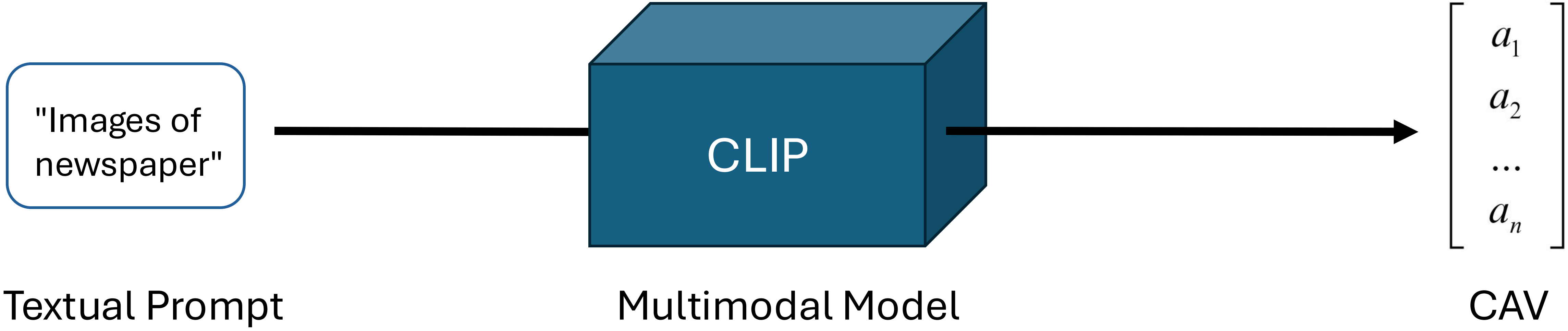
Set of synthetic concept images

CNN

CAV

Riccardo Campi et al., Towards synthetic concept activation vectors via generative models. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops, pages 2720–2728, June 2025.

Our idea: from text to CAV



"Images of newspaper"

Textual Prompt

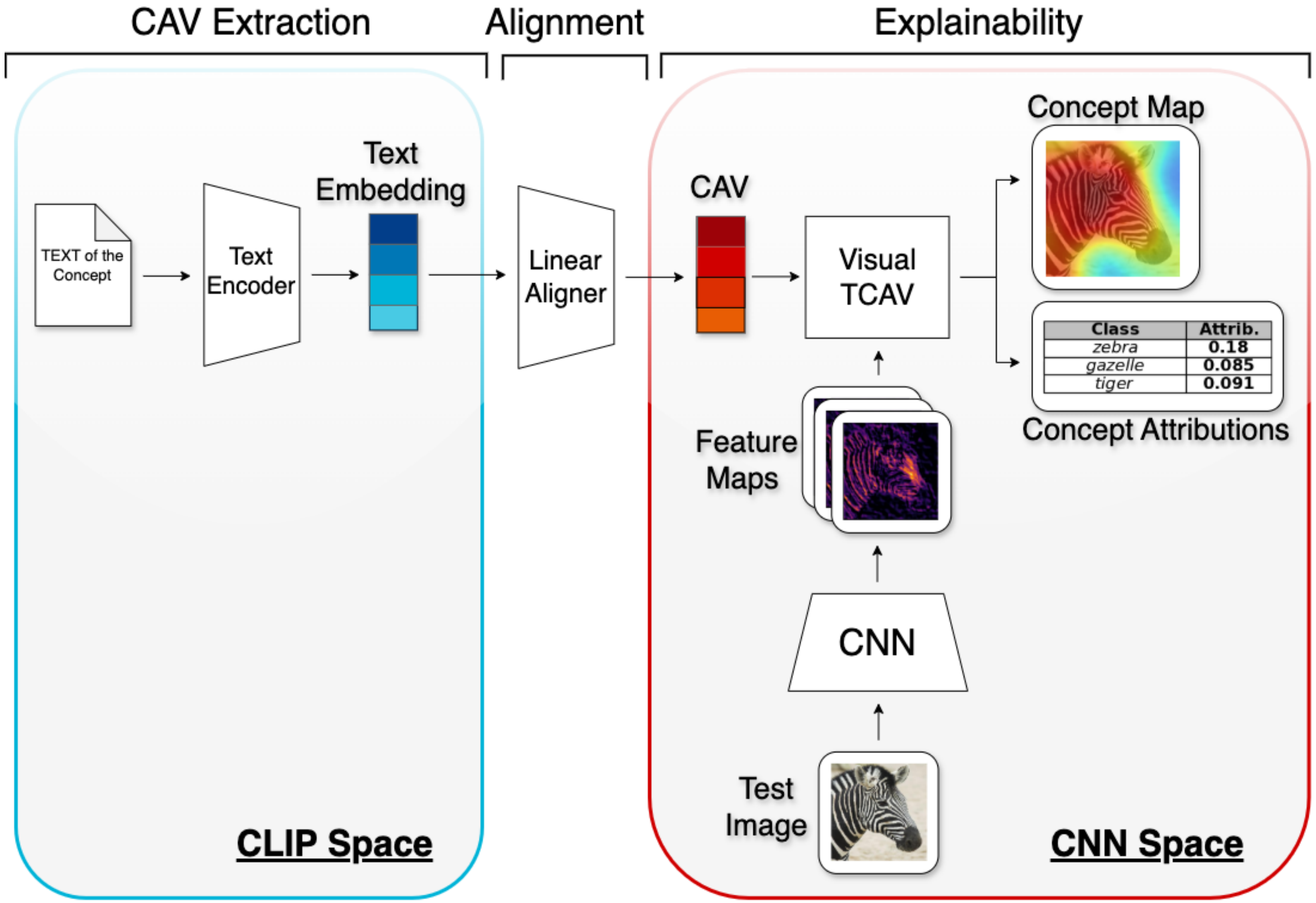
CLIP

Multimodal Model

$$\begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{bmatrix}$$

CAV

Our Solution



CAV Extraction

Alignment

Explainability

TEXT of the Concept

Text Encoder

Text Embedding

Linear Aligner

CAV

Visual TCAV

Concept Map

Feature Maps

CNN

Test Image

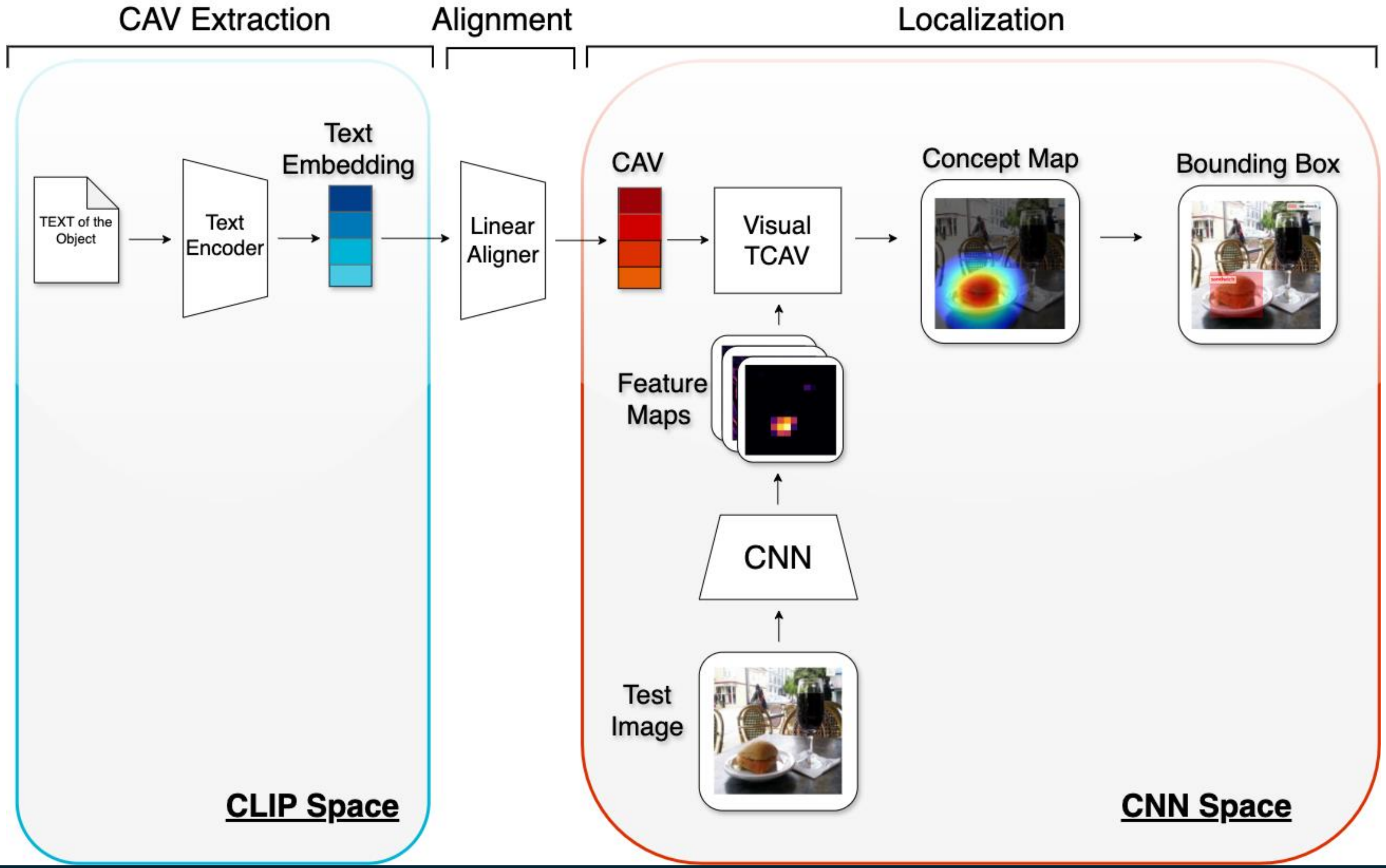
Class	Attrib.
zebra	0.18
gazelle	0.085
tiger	0.091

Concept Attributions

CLIP Space

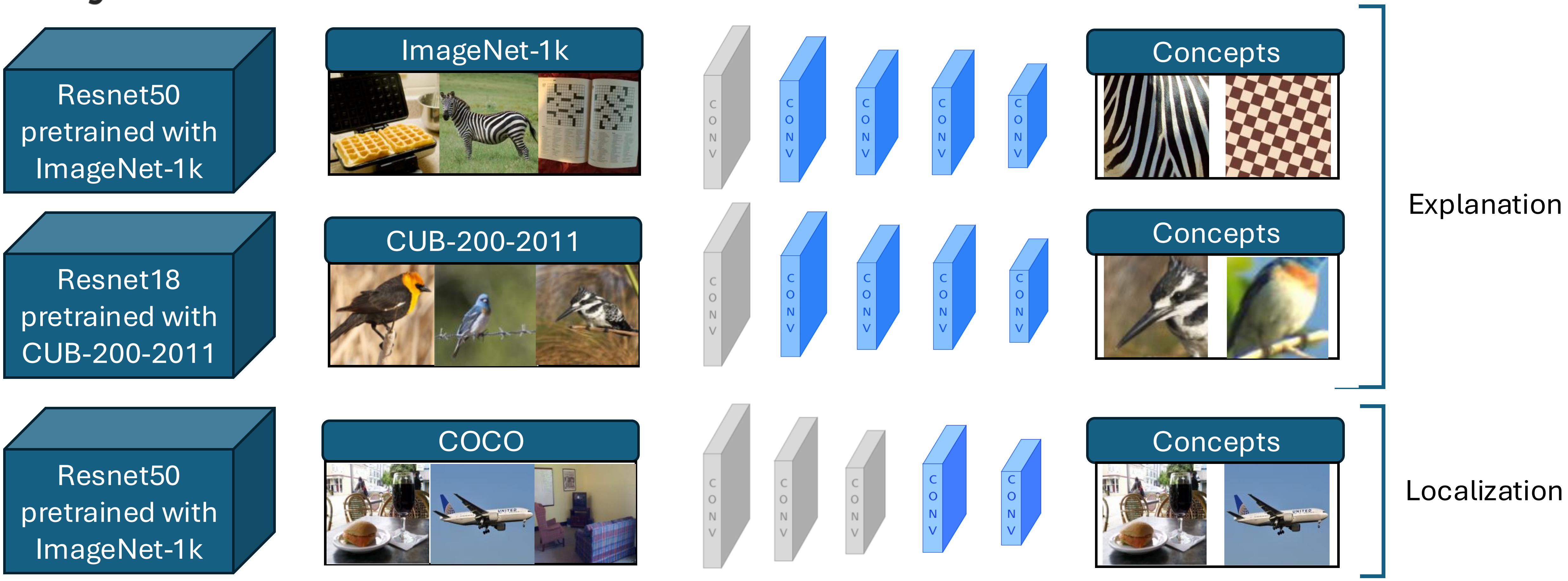
CNN Space

Other use: Localization



PyTorch

Experiments Setup



Resnet50 pretrained with ImageNet-1k

ImageNet-1k

Concepts

Resnet18 pretrained with CUB-200-2011

CUB-200-2011

Concepts

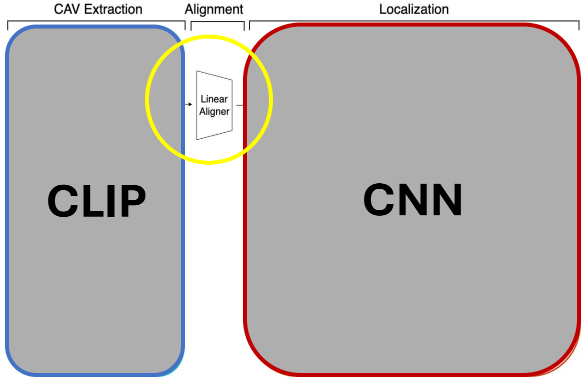
Explanation

Resnet50 pretrained with ImageNet-1k

COCO

Concepts

Localization



Linear Aligners Training

Resnet50 pretrained with ImageNet-1k

Layer	MSE	R^2
layer1	0.492	0.891
layer2	1.14	0.74
layer3	1.798	0.6
layer4	3.205	0.288

Resnet18 pretrained with CUB-200-2011

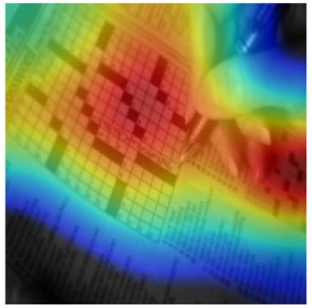
Layer	MSE	R^2
stage1	0.328	0.927
stage2	0.819	0.818
stage3	1.447	0.678
stage4	2.853	0.366

Experiments: ImageNet-1k

RESNET50 architecture
chequered concept

layer4 layer

Class	Attrib.
crossword_pu..	**0.1**
digital_watc..	**0.022**
rule	**0.014**

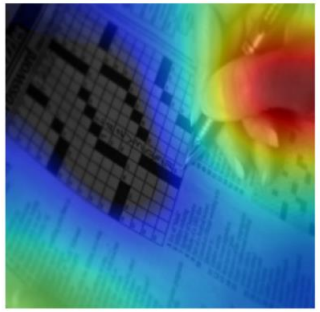


chequered

RESNET50 architecture
hand concept

layer4 layer

Class	Attrib.
crossword_pu..	**0.024**
digital_watc..	**0.01**
rule	**0.011**

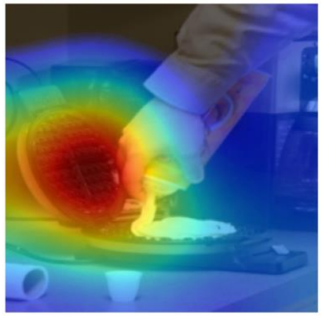


hand

RESNET50 architecture
waffled concept

layer4 layer

Class	Attrib.
waffle_iron	**0.13**
mouse	**0.027**
computer_key..	**0.041**

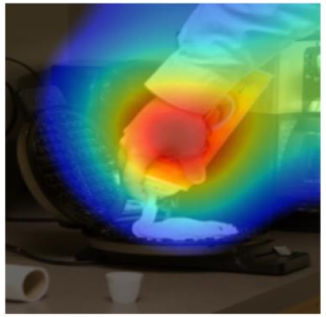


waffled

RESNET50 architecture
hand concept

layer4 layer

Class	Attrib.
waffle_iron	**0.012**
mouse	**0.0084**
computer_key..	**0.008**

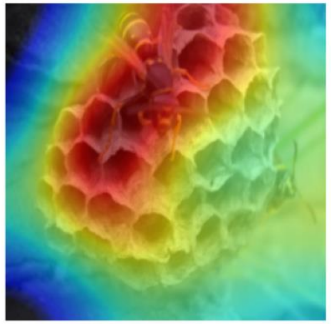


hand

RESNET50 architecture
honeycombed concept

layer4 layer

Class	Attrib.
honeycomb	**0.2**
bee	**0.096**
apiary	**0.1**

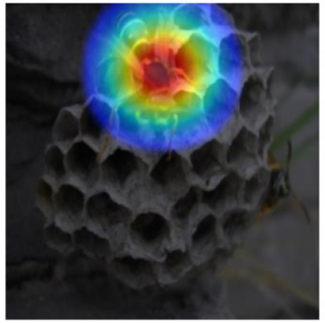


honeycombed

RESNET50 architecture
bee concept

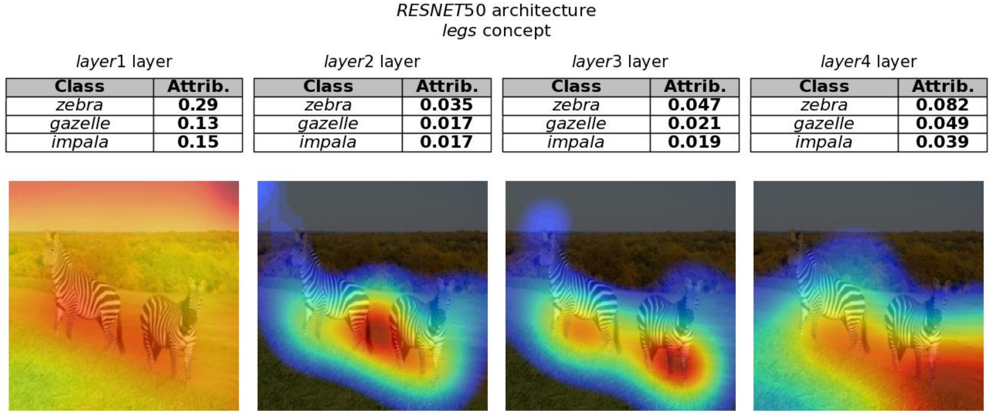
layer4 layer

Class	Attrib.
honeycomb	**0.007**
bee	**0.0051**
apiary	**0.0039**



bee

Experiments: ImageNet-1k



*RESNET*50 architecture
legs concept

*layer*1 layer

Class	Attrib.
zebra	**0.29**
gazelle	**0.13**
impala	**0.15**

*layer*2 layer

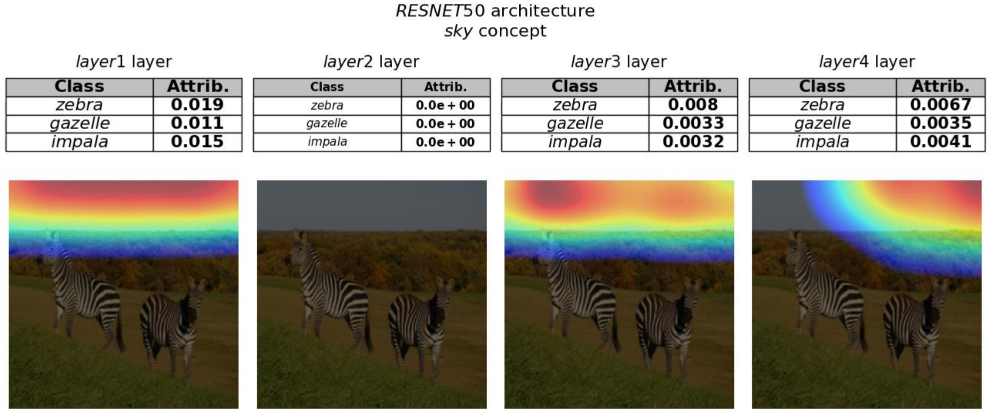
Class	Attrib.
zebra	**0.035**
gazelle	**0.017**
impala	**0.017**

*layer*3 layer

Class	Attrib.
zebra	**0.047**
gazelle	**0.021**
impala	**0.019**

*layer*4 layer

Class	Attrib.
zebra	**0.082**
gazelle	**0.049**
impala	**0.039**



*RESNET*50 architecture
sky concept

*layer*1 layer

Class	Attrib.
zebra	**0.019**
gazelle	**0.011**
impala	**0.015**

*layer*2 layer

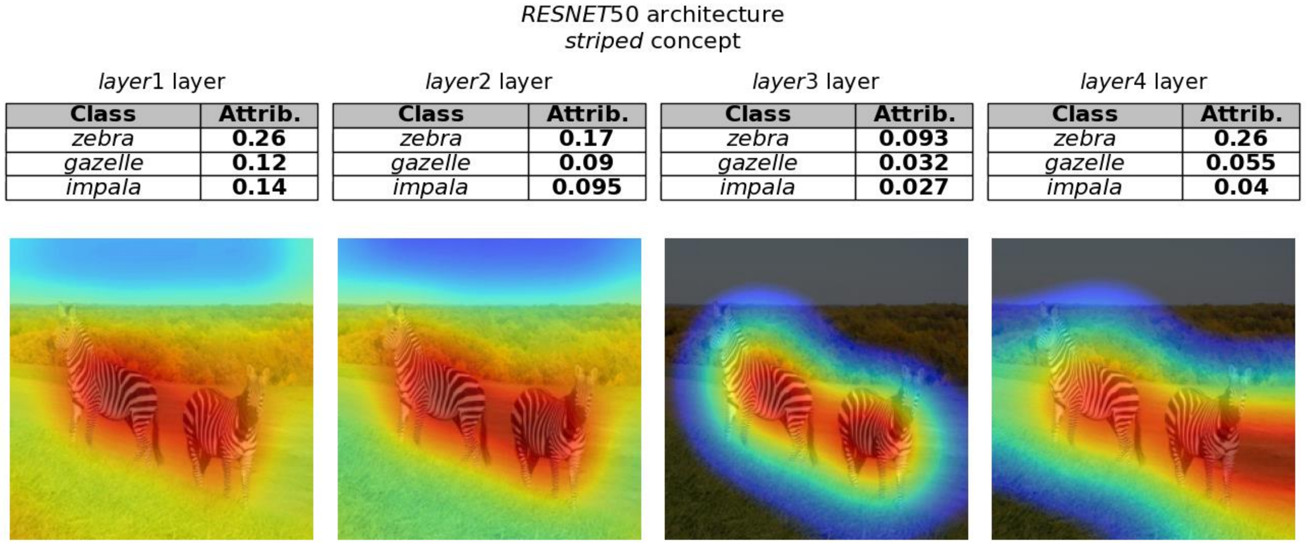
Class	Attrib.
zebra	**0.0e + 00**
gazelle	**0.0e + 00**
impala	**0.0e + 00**

*layer*3 layer

Class	Attrib.
zebra	**0.008**
gazelle	**0.0033**
impala	**0.0032**

*layer*4 layer

Class	Attrib.
zebra	**0.0067**
gazelle	**0.0035**
impala	**0.0041**



*RESNET*50 architecture
striped concept

*layer*1 layer

Class	Attrib.
zebra	**0.26**
gazelle	**0.12**
impala	**0.14**

*layer*2 layer

Class	Attrib.
zebra	**0.17**
gazelle	**0.09**
impala	**0.095**

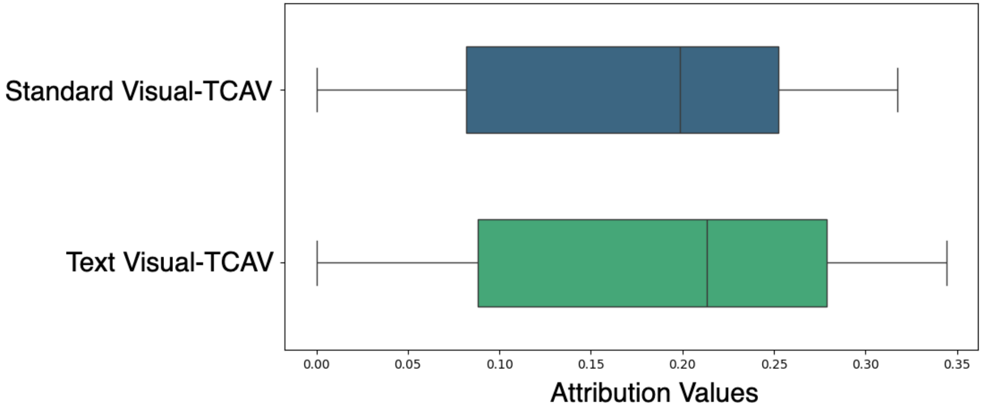
*layer*3 layer

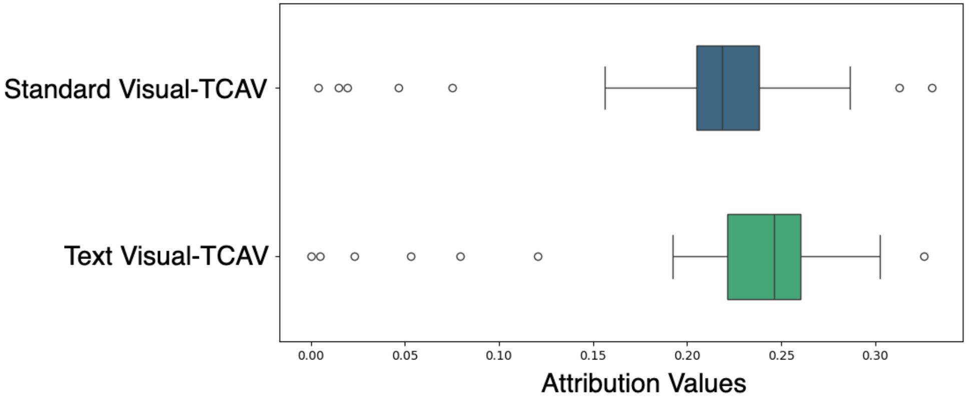
Class	Attrib.
zebra	**0.093**
gazelle	**0.032**
impala	**0.027**

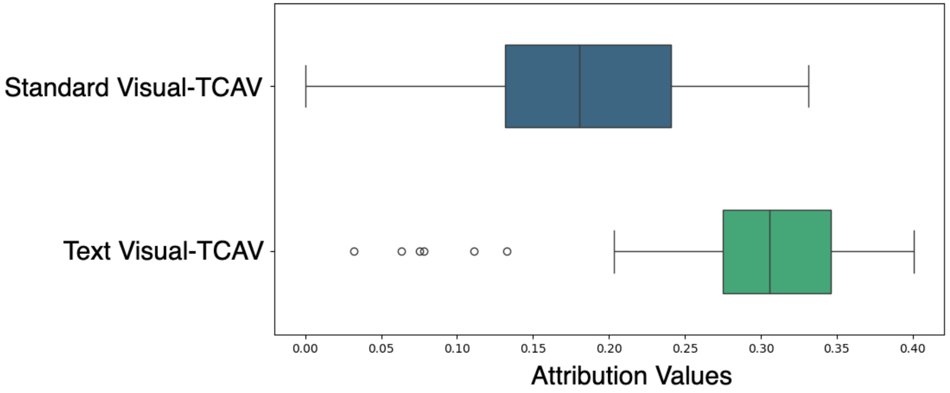
*layer*4 layer

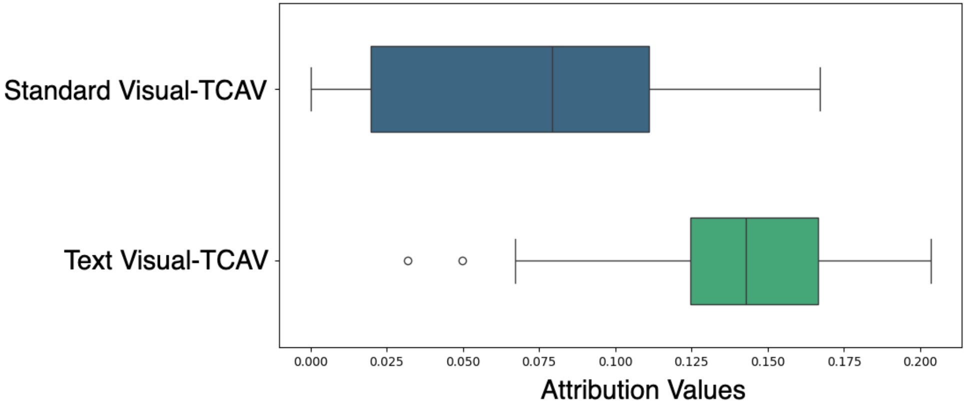
Class	Attrib.
zebra	**0.26**
gazelle	**0.055**
impala	**0.04**

Experiments: ImageNet-1k

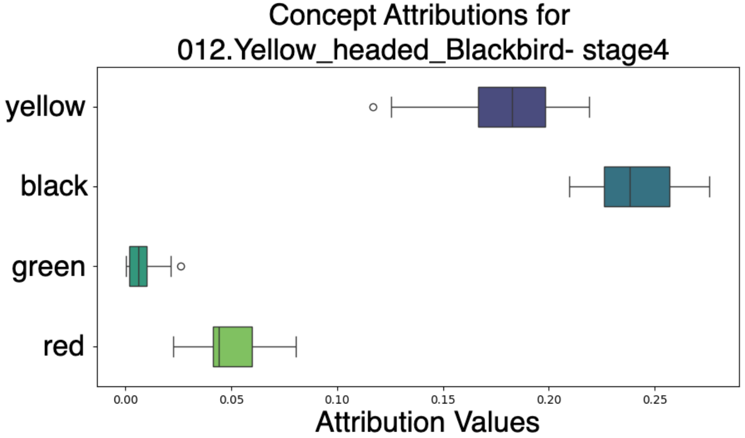

Standard Visual-TCAV vs Text Visual-TCAV Attributions
waffled - layer4

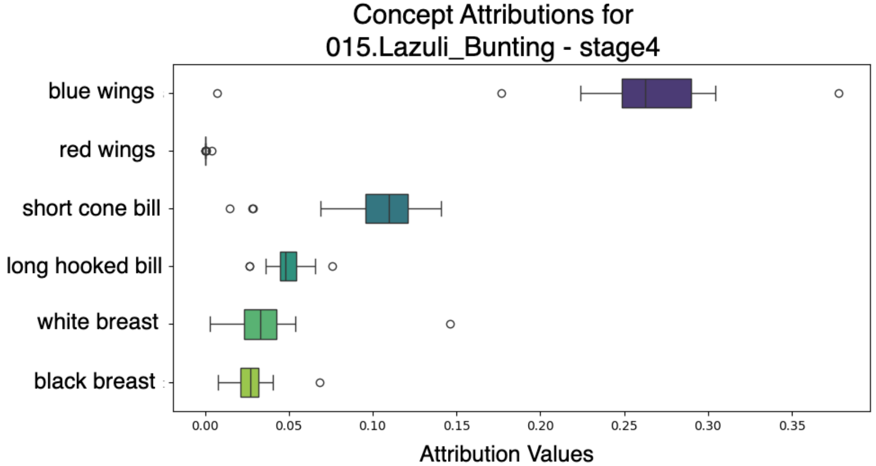

Standard Visual-TCAV vs Text Visual-TCAV Attributions
striped - layer4

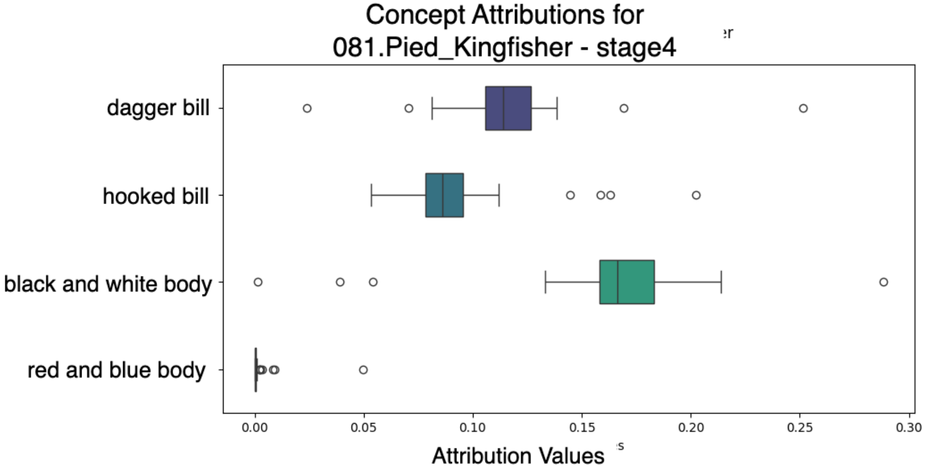

Standard Visual-TCAV vs Text Visual-TCAV Attributions
honeycombed - layer4


Standard Visual-TCAV vs Text Visual-TCAV Attributions
chequered - layer4

Experiments: CUB-200-2011







*RESNET*18 architecture
daggerbill concept

*stage*1 layer

Class	Attrib.
Pied_Kingfis..	0.0e + 00
Black_and_wh..	0.0e + 00
Green_Kingfi..	0.0e + 00

*stage*2 layer

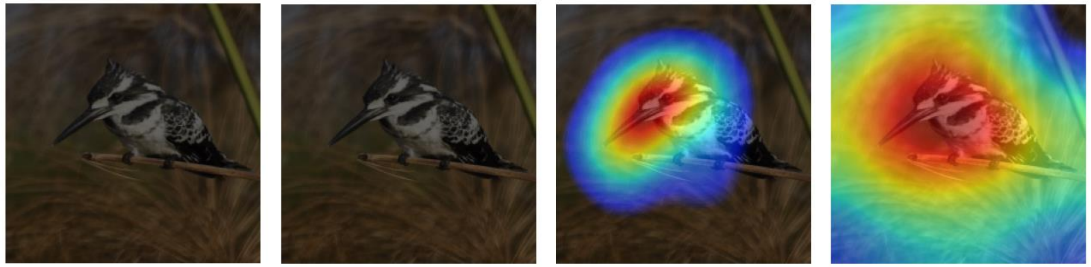
Class	Attrib.
Pied_Kingfis..	0.0e + 00
Black_and_wh..	0.0e + 00
Green_Kingfi..	0.0e + 00

*stage*3 layer

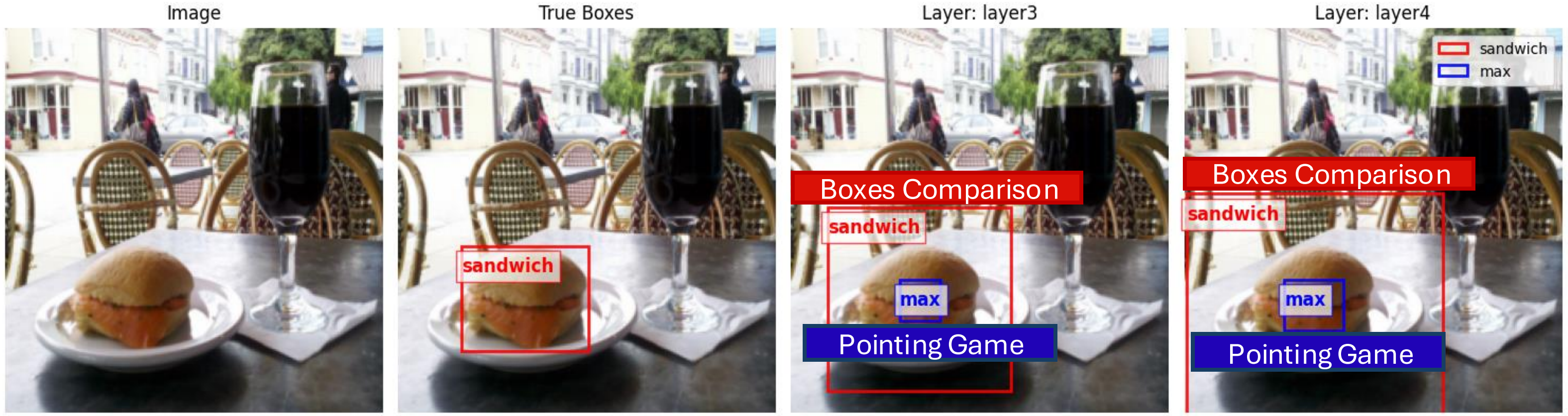
Class	Attrib.
Pied_Kingfis..	0.012
Black_and_wh..	0.0056
Green_Kingfi..	0.0066

*stage*4 layer

Class	Attrib.
Pied_Kingfis..	0.12
Black_and_wh..	0.061
Green_Kingfi..	0.07



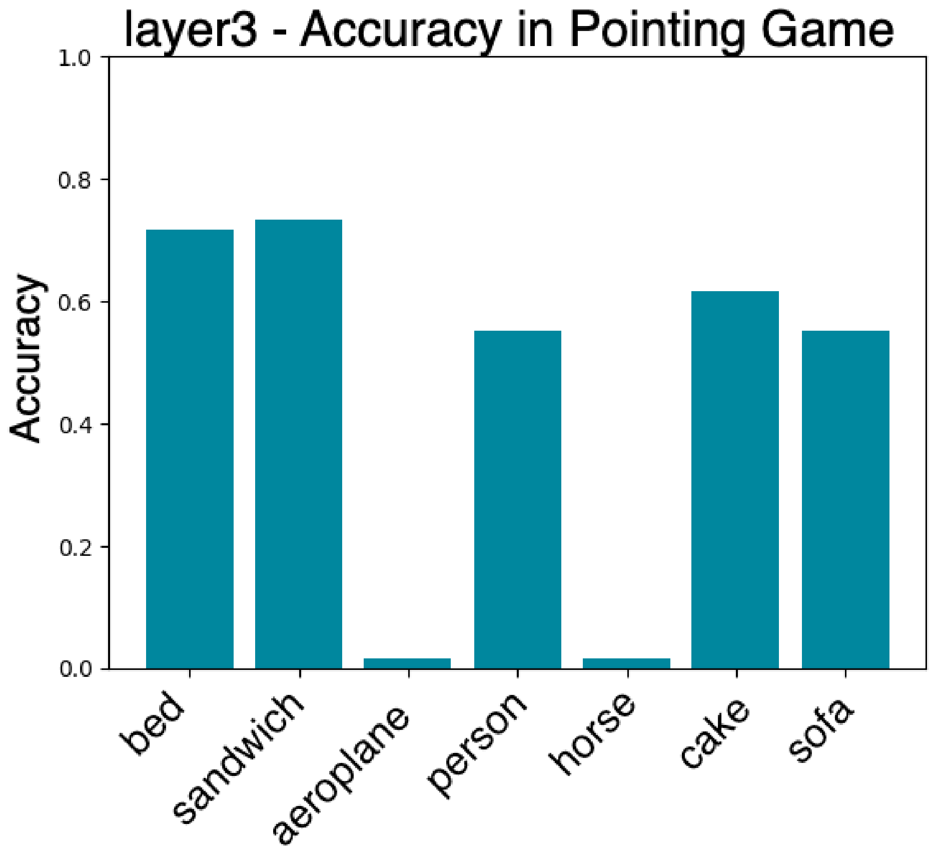
Experiments: Zero-Shot Localization with COCO

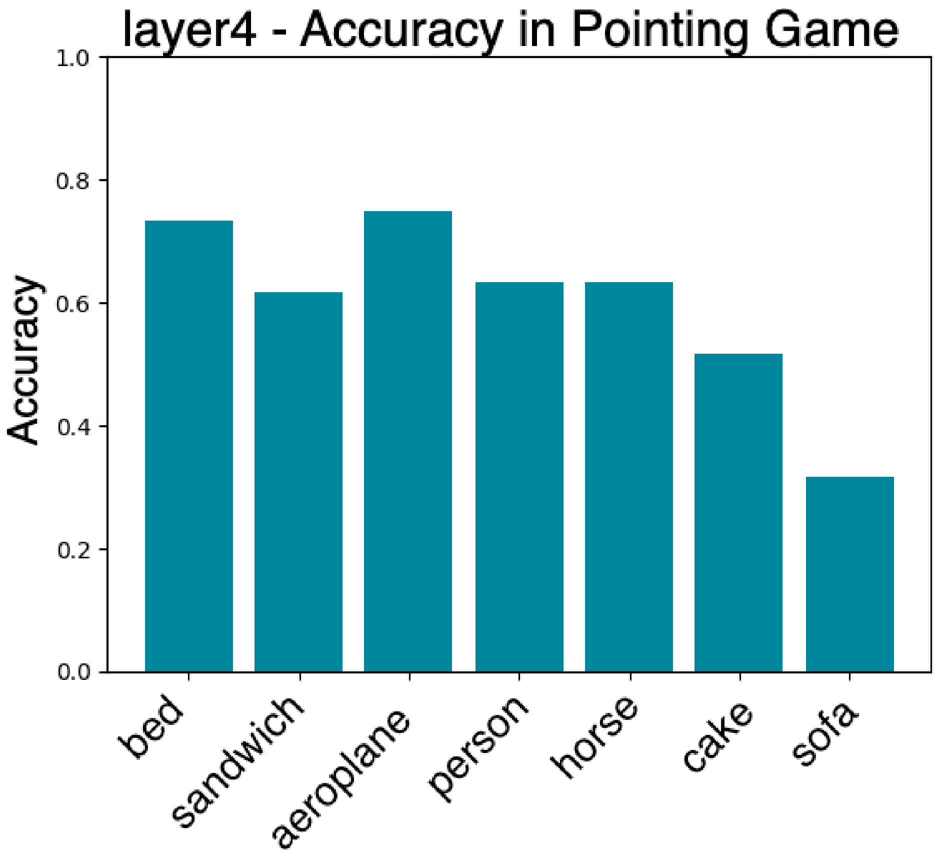


We implement two tasks to assess the performance of zero-shot localization with CAVs from text:

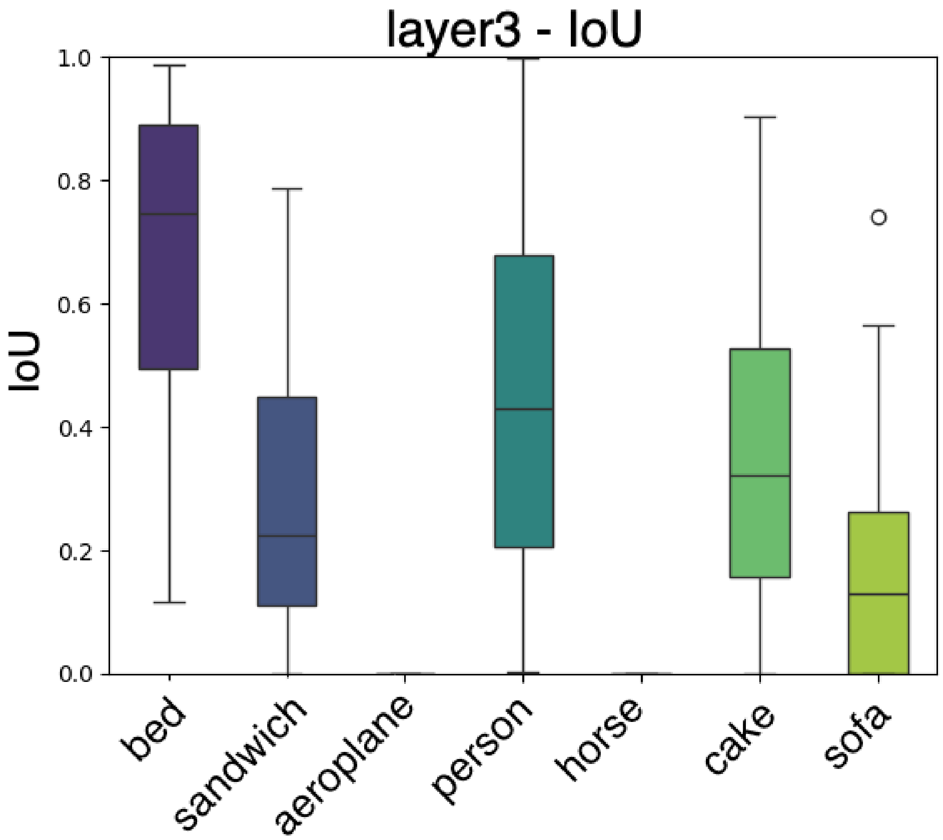
- Pointing Game (correct object position)
- Boxes comparison with IoU (bounding box covers most part of the object)

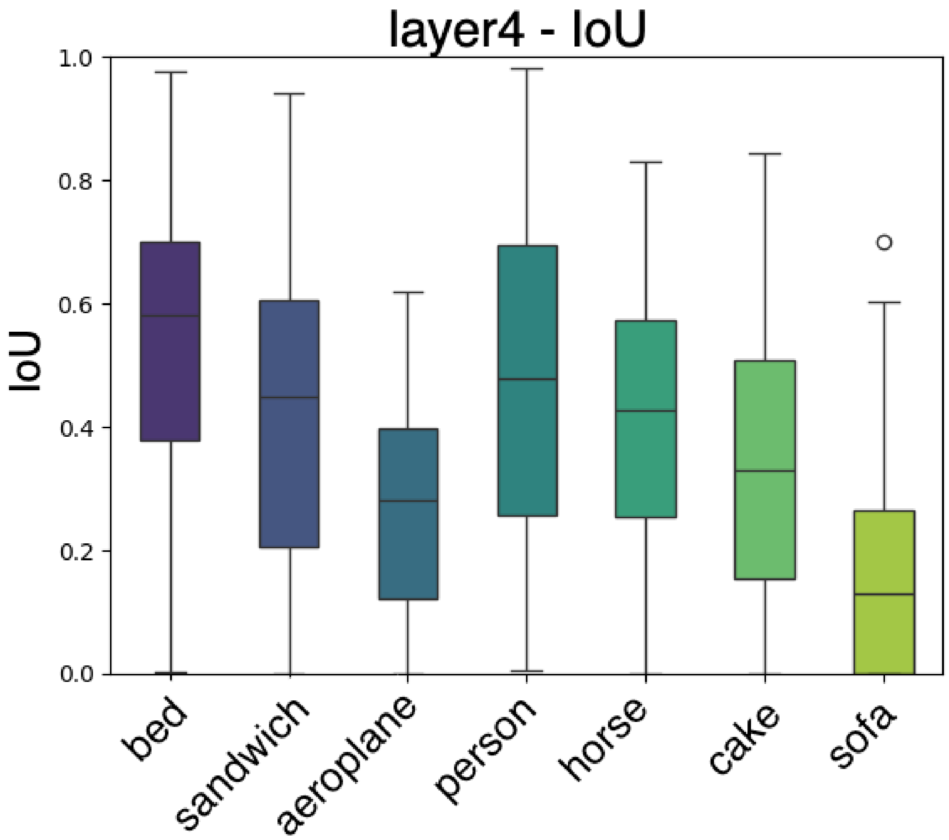
Accuracy in Pointing Game





IoU Distribution in Boxes Comparison





Conclusions

CAVs from text

- **High scalability** for CAV extraction
- **No concept images** for CAV extraction
- **Consistent explanations**
- **Localization** tasks without an annotated dataset
- **Zero-shot localization**

Future work

- More **complex architectures** for the **aligners**
- More roboust ways to **refine** the **concept maps**

POLITECNICO
MILANO 1863

Thank you for your attention