max planck institut informatik

SIC Saarland Informatics Campus

High Level Computer Vision

Some Recent Trends: B-Cos Networks & Segment Anything Model

@ July 17, 2024

Bernt Schiele

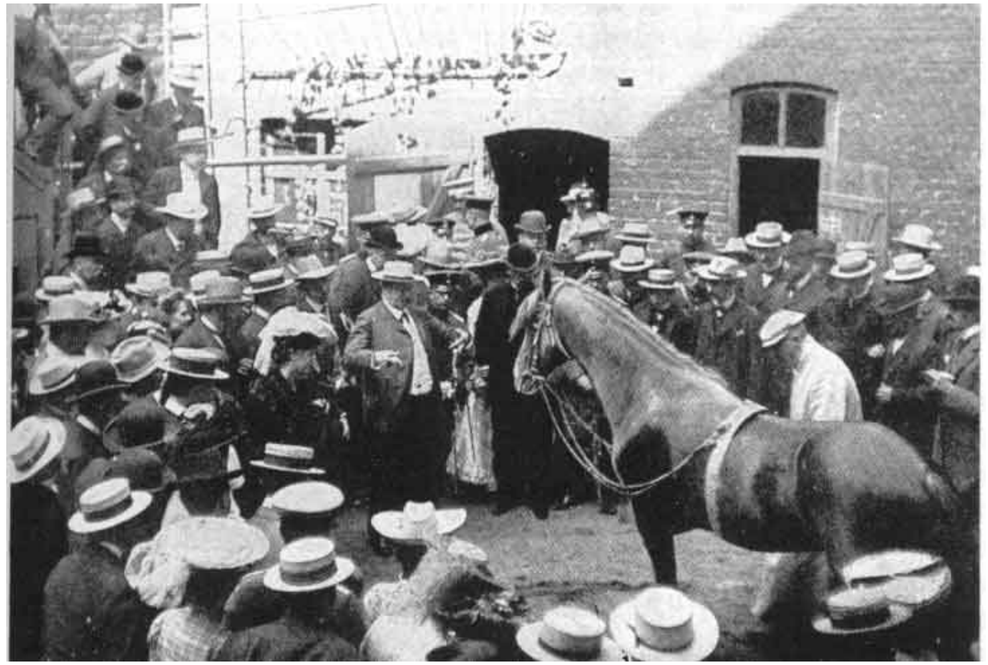
cms.sic.saarland/hlcvss24/

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

Overview of Today's Lecture

- Inherently Interpretable Neural Network
  - B-cos CNN [cvpr'22] - https://arxiv.org/abs/2205.10268
  - B-cos CNN & Vision transformers [pami'24] - https://arxiv.org/abs/2306.10898

- Segment Anything Model (SAM)
  - [iccv'23] - https://arxiv.org/abs/2304.02643
  - https://segment-anything.com/

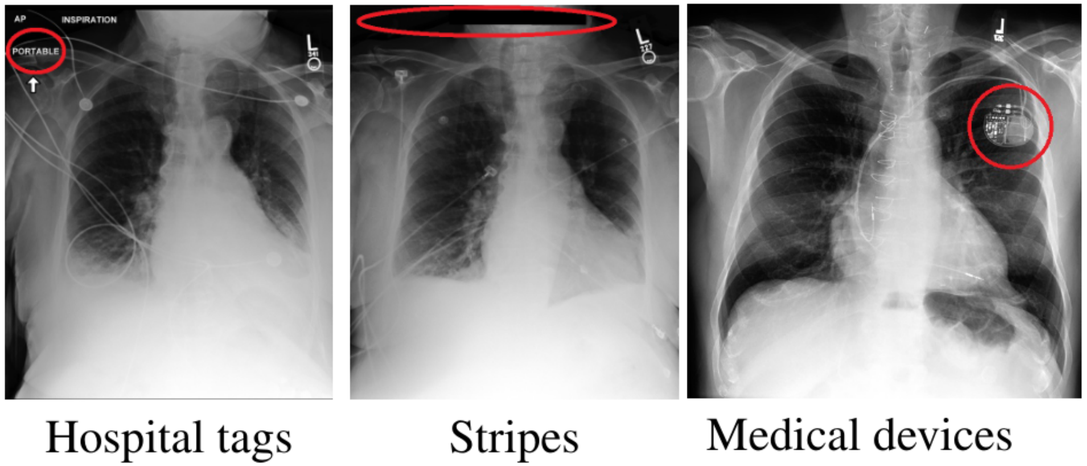


Images taken from wikipedia

We often may want to inspect how a model arrives at a prediction!

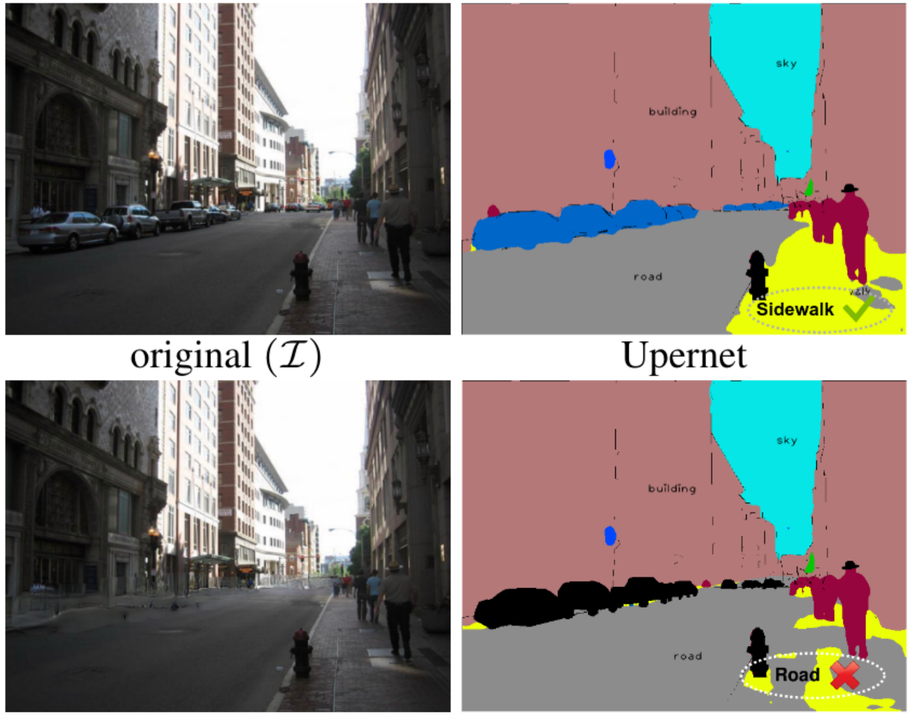
For example, to make sure the model is not relying on spurious correlations.

Source



Spurious correlations in the medical context

Source



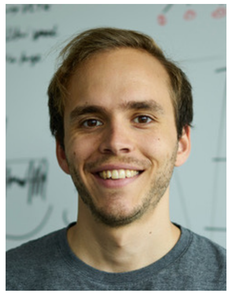
A model relying on the context for its prediction

max planck institut informatik

SIC Saarland Informatics Campus

B-cos Networks: Alignment is All We Need for Interpretability

Moritz Böhle[1], Mario Fritz[2] and Bernt Schiele[1]
[1]Max Planck Institute for Informatics, [2]CISPA Helmholtz Center for Information Security

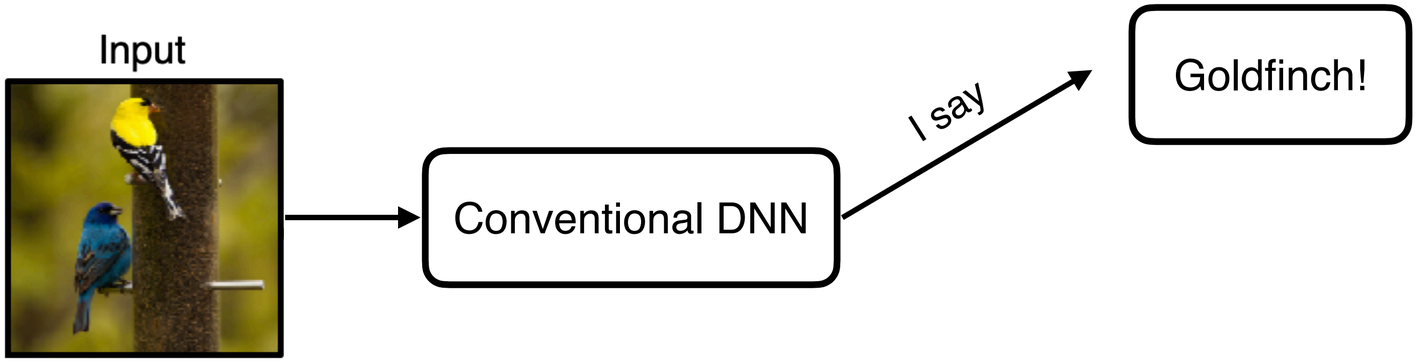


Motivation

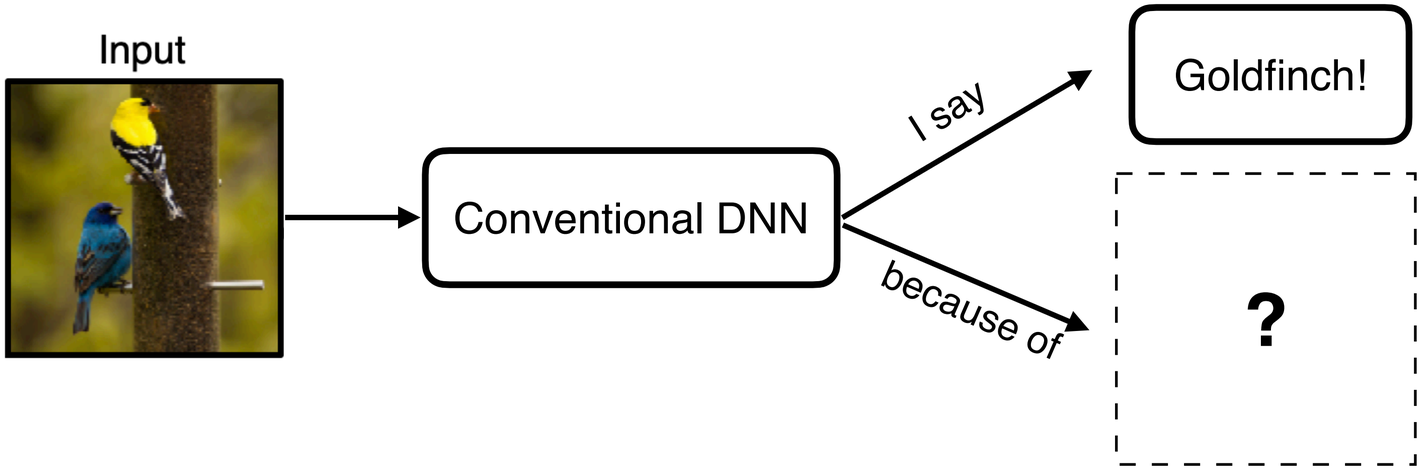
Input

Conventional DNN

Motivation



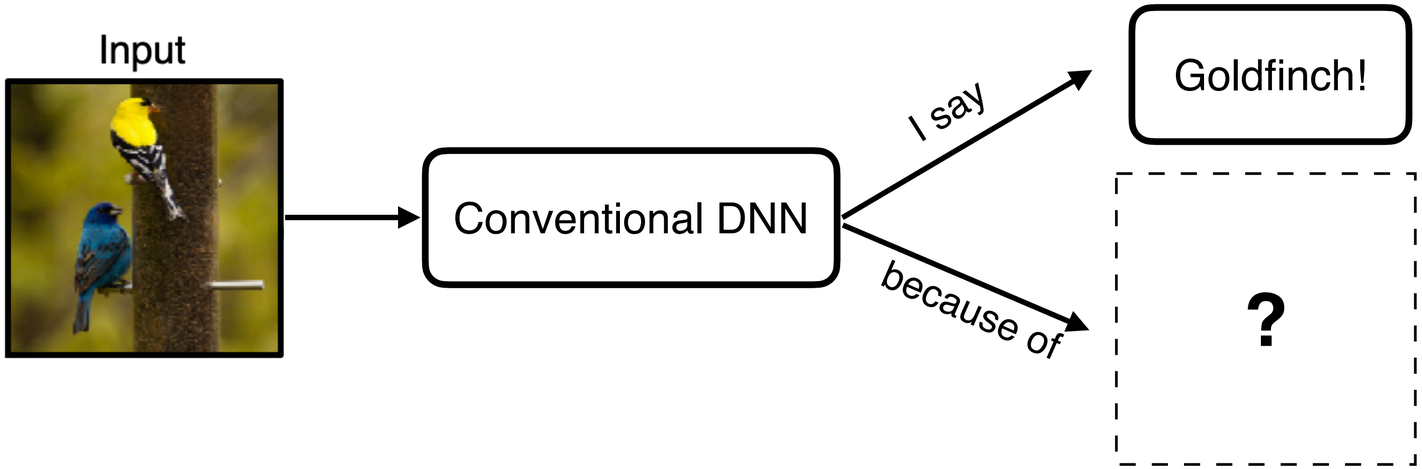
Motivation



Motivation

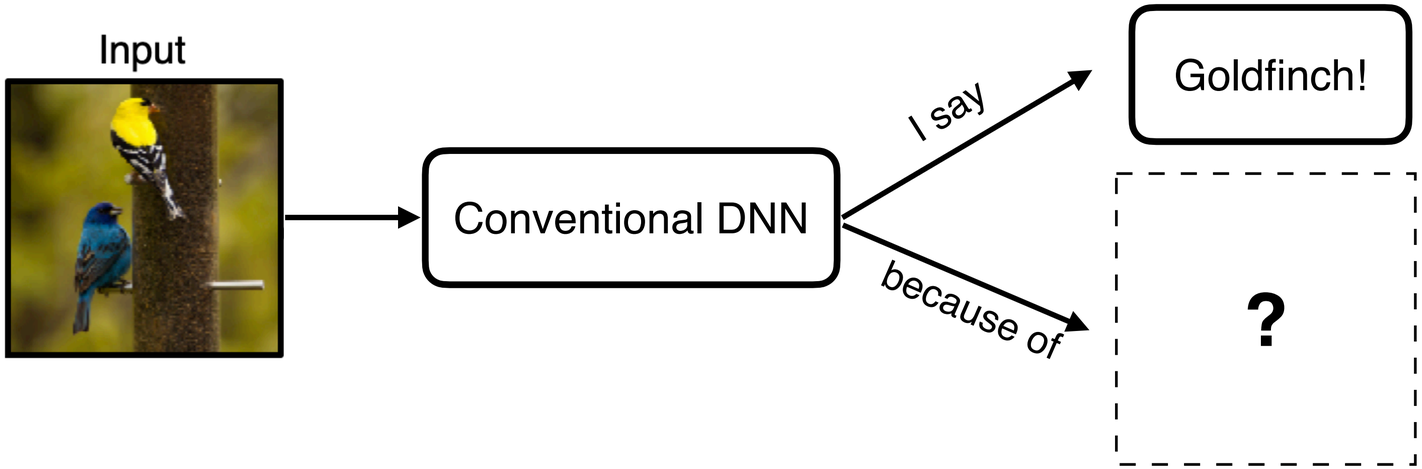
Want

Explanation



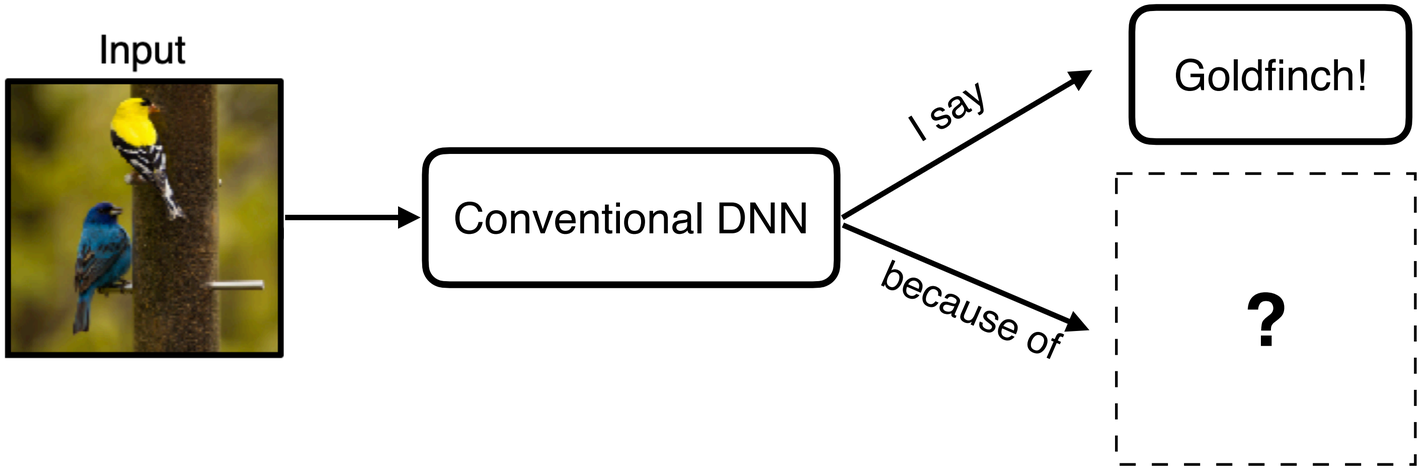
Motivation

Want Explanation to gain trust



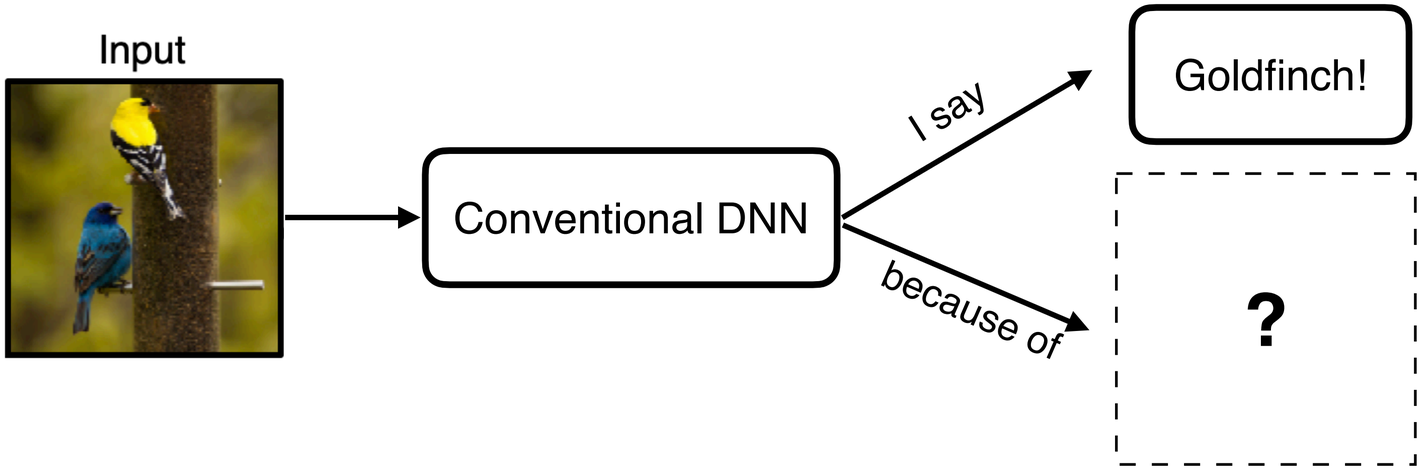
Motivation

Want Explanation to detect biases



Motivation

Want | Explanation | to | understand better

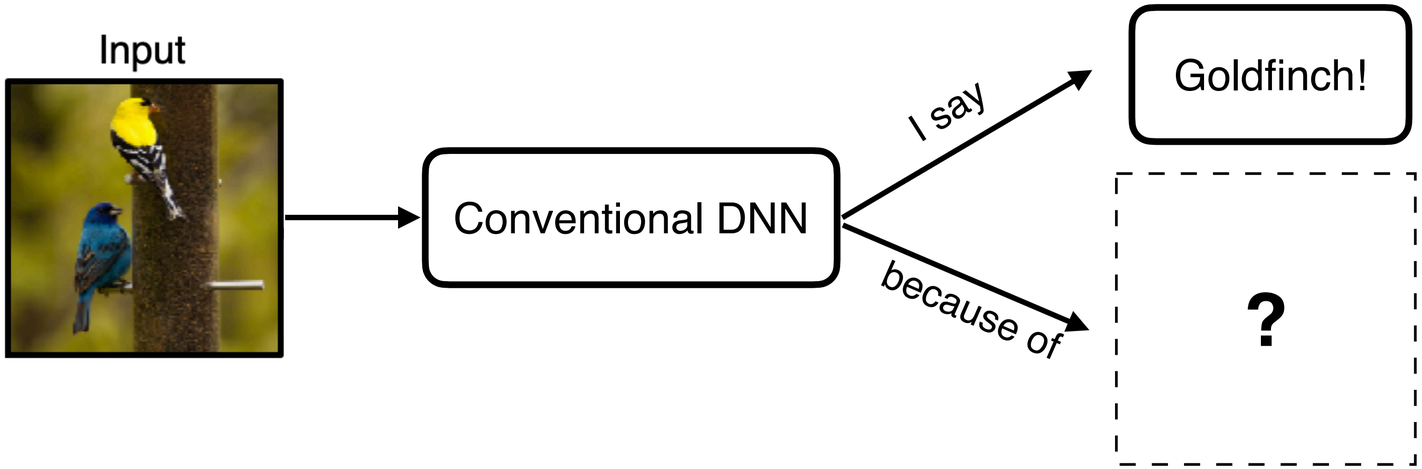


Motivation

Want

Explanation

Requirements



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

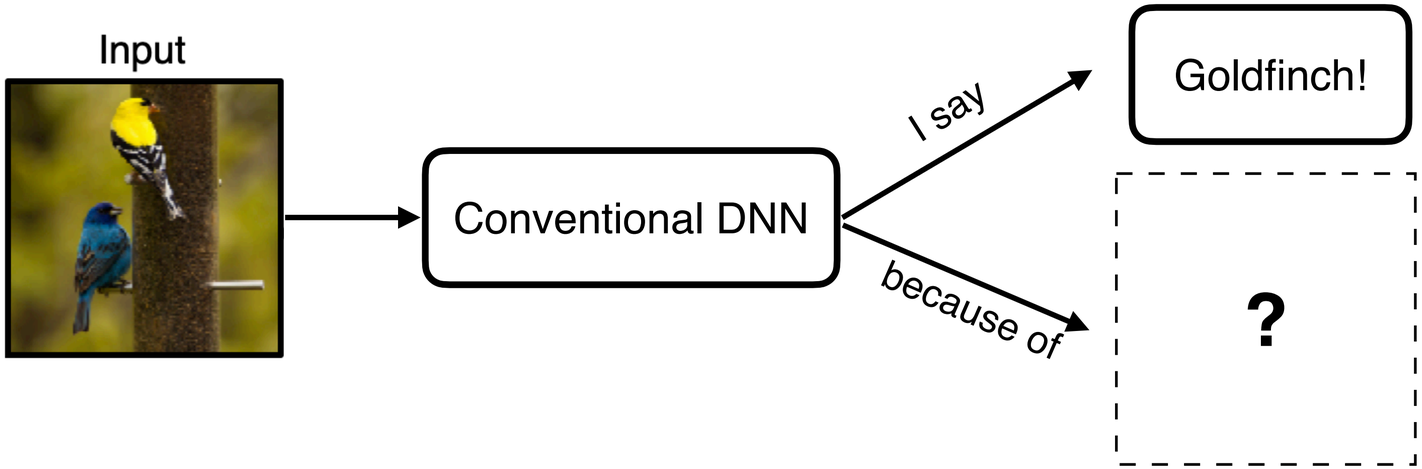
Motivation

Want

Explanation

Requirements

Faithful



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

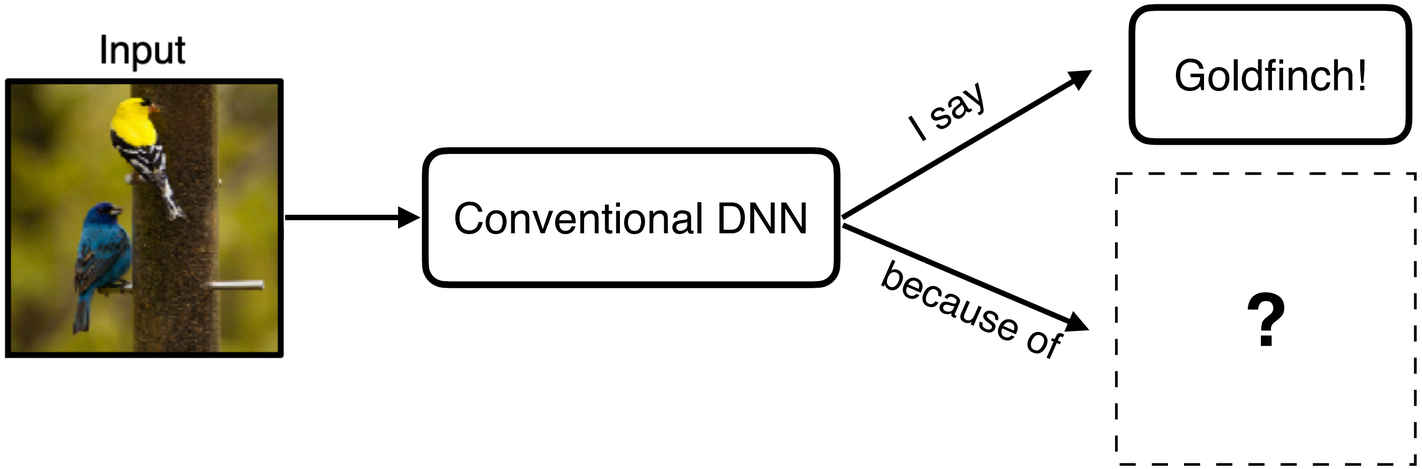
Want

Explanation

Requirements

Faithful

Interpretable



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

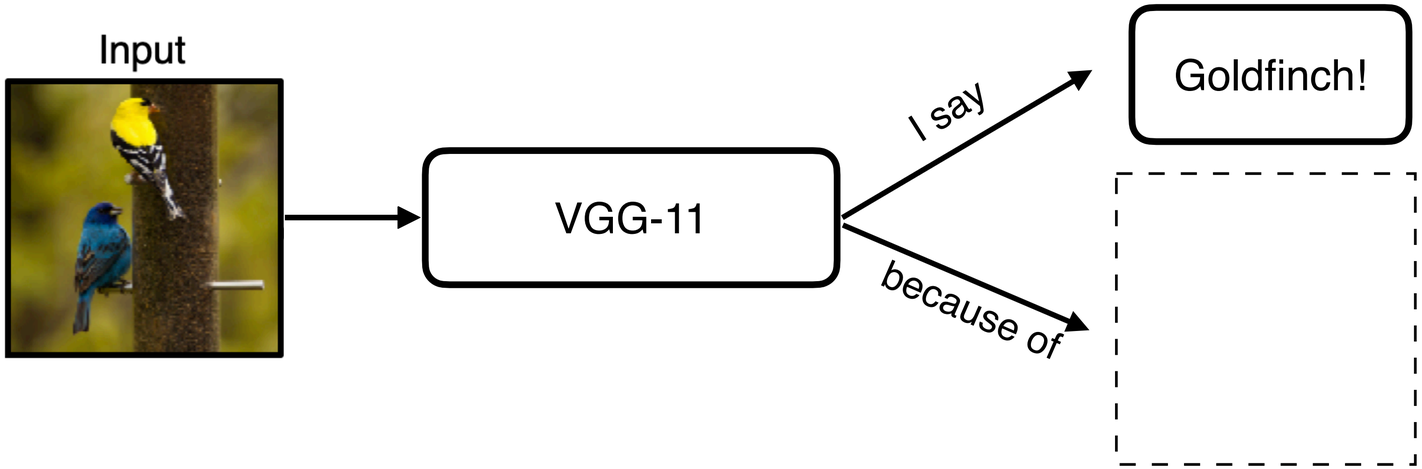
Want

Explanation

Requirements

Faithful

Interpretable



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

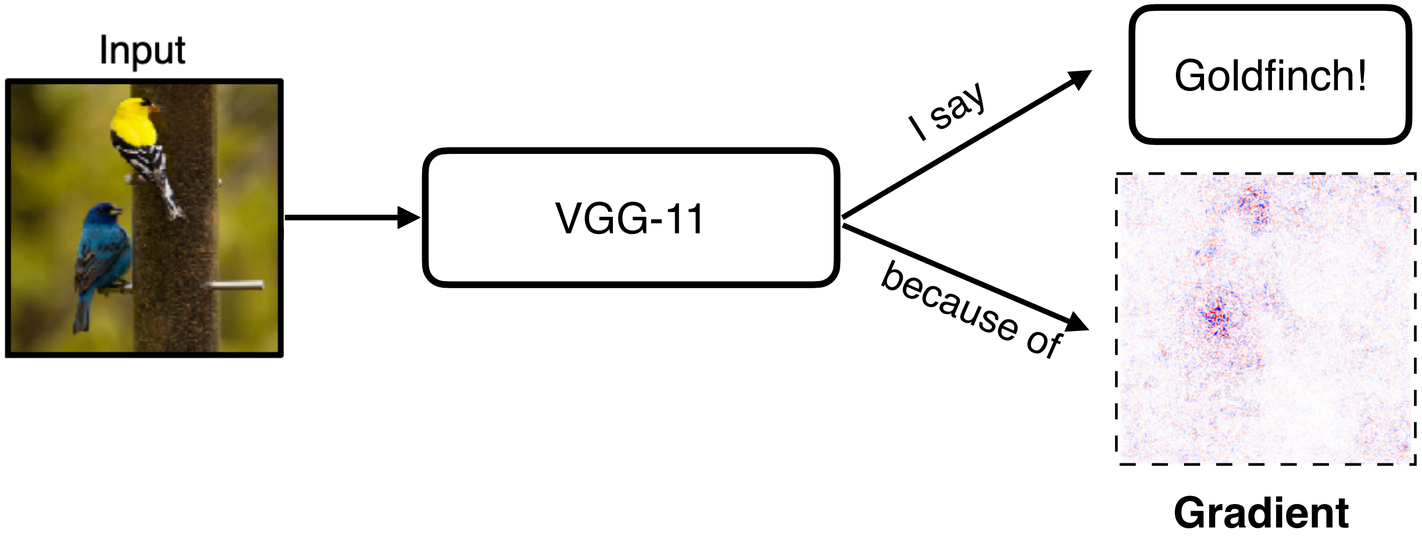
Want

Explanation

Requirements

Faithful

Interpretable



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

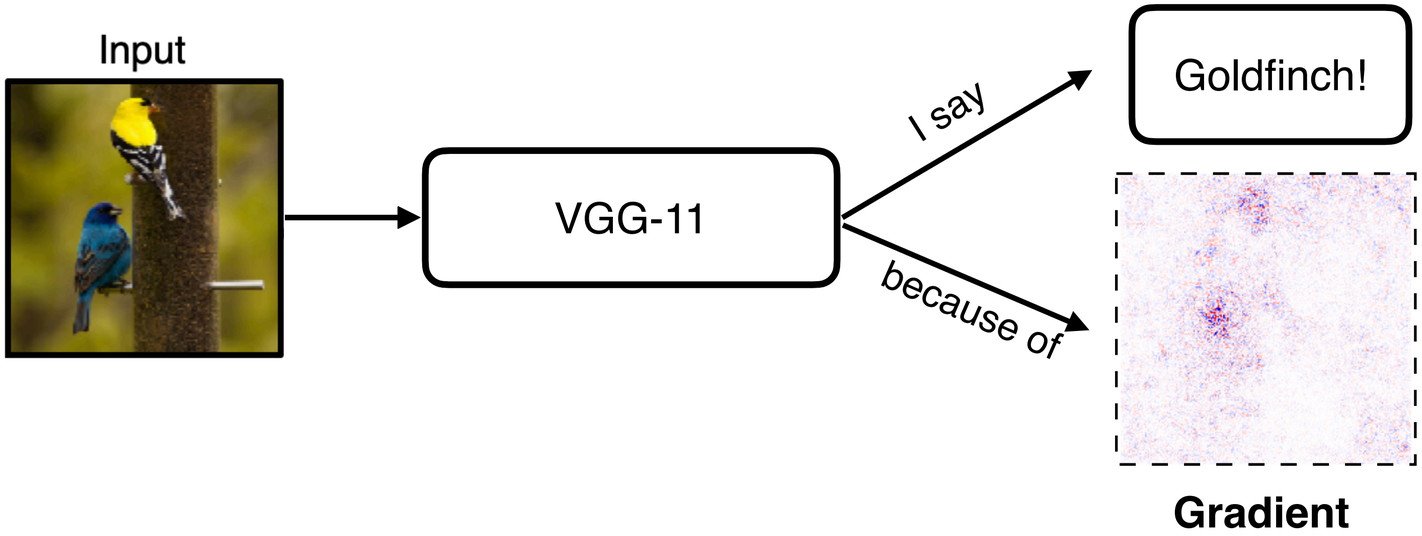
Want

Explanation

Requirements

Faithful ✓

Interpretable



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

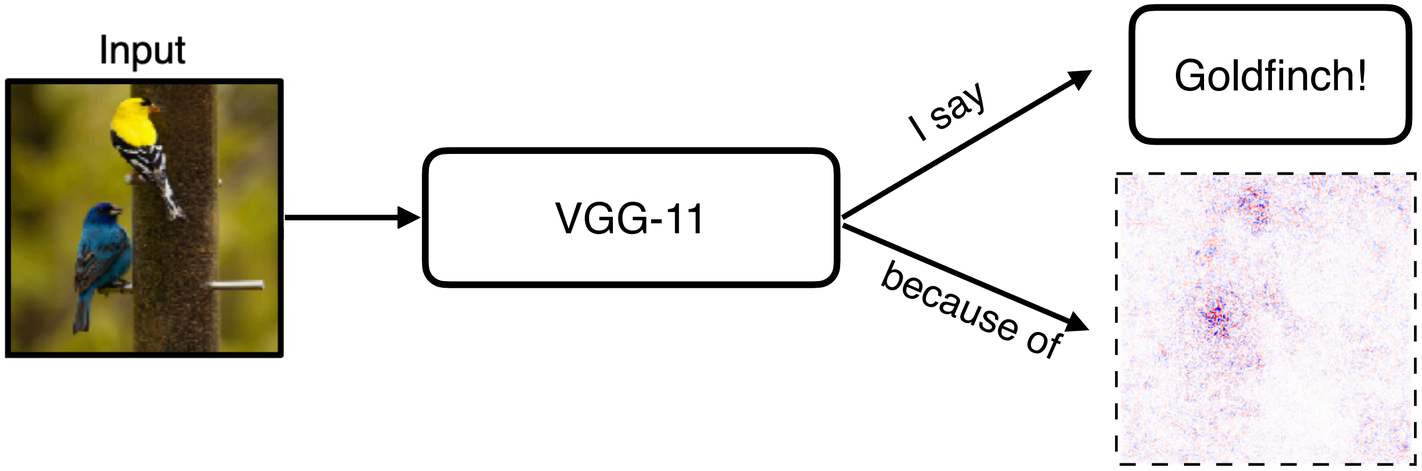
Want

Explanation

Requirements

Faithful ✓

Interpretable ✗



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

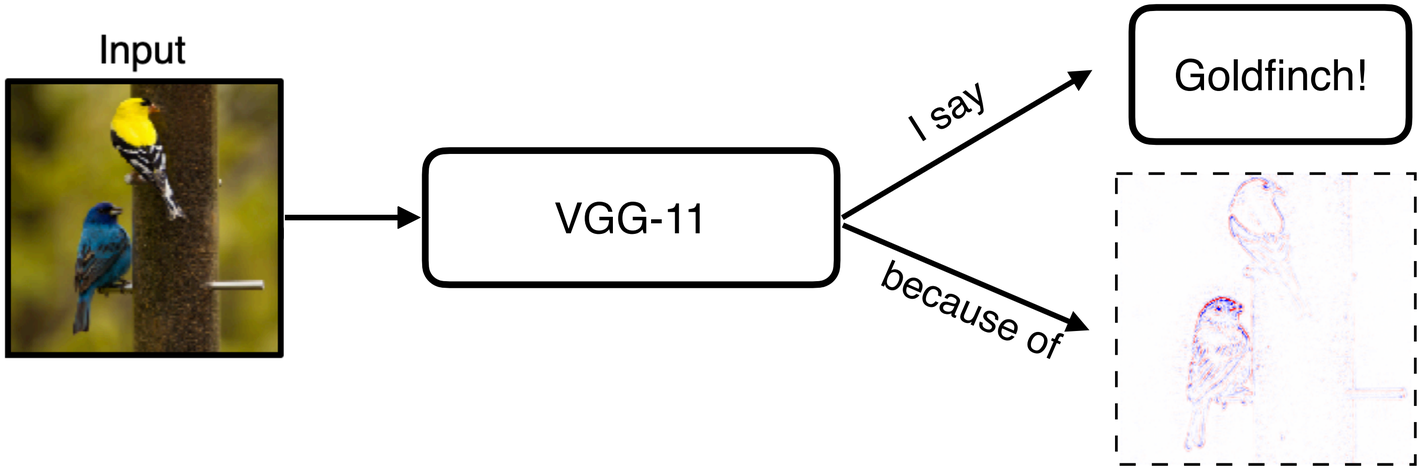
Want

Explanation

Requirements

Faithful

Interpretable



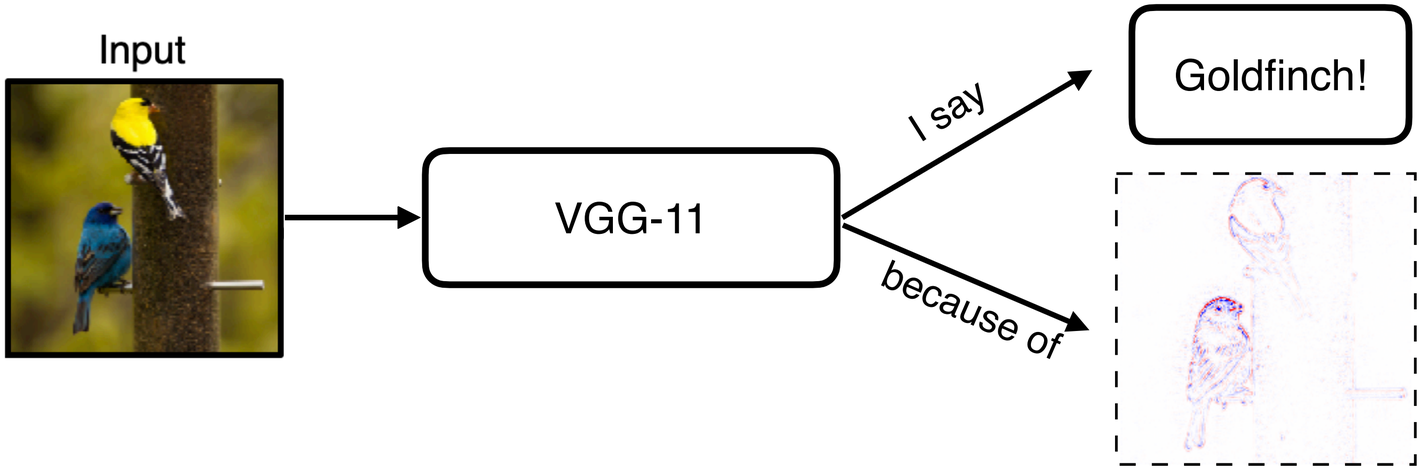
Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

Want: Explanation

Requirements: Faithful, Interpretable ✓



Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

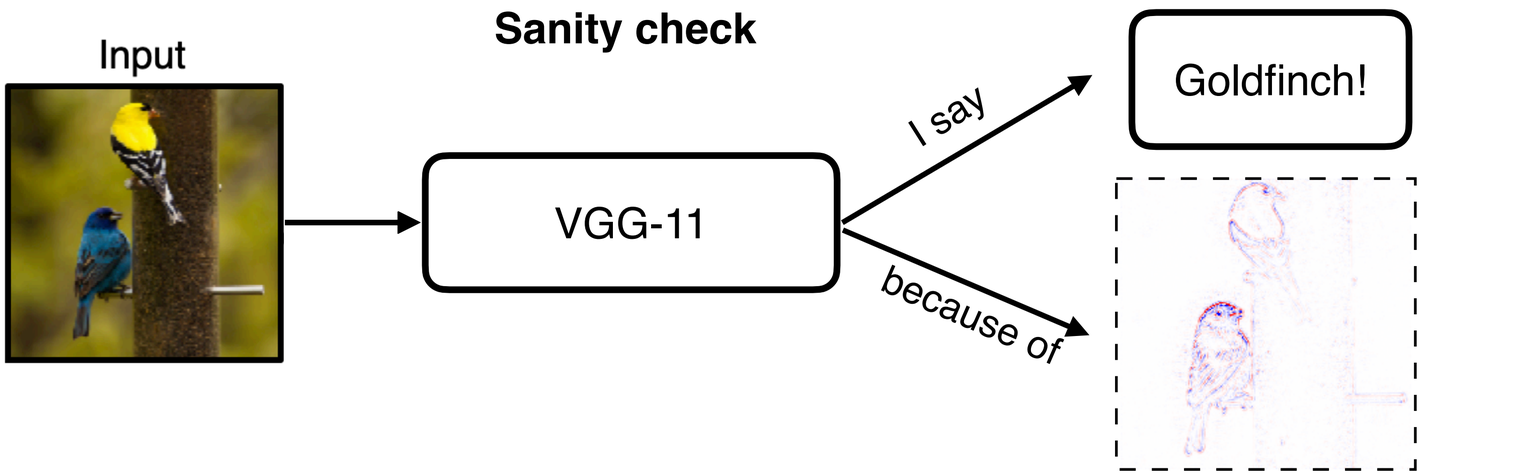
Want

Explanation

Requirements

Faithful

Interpretable ✓



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

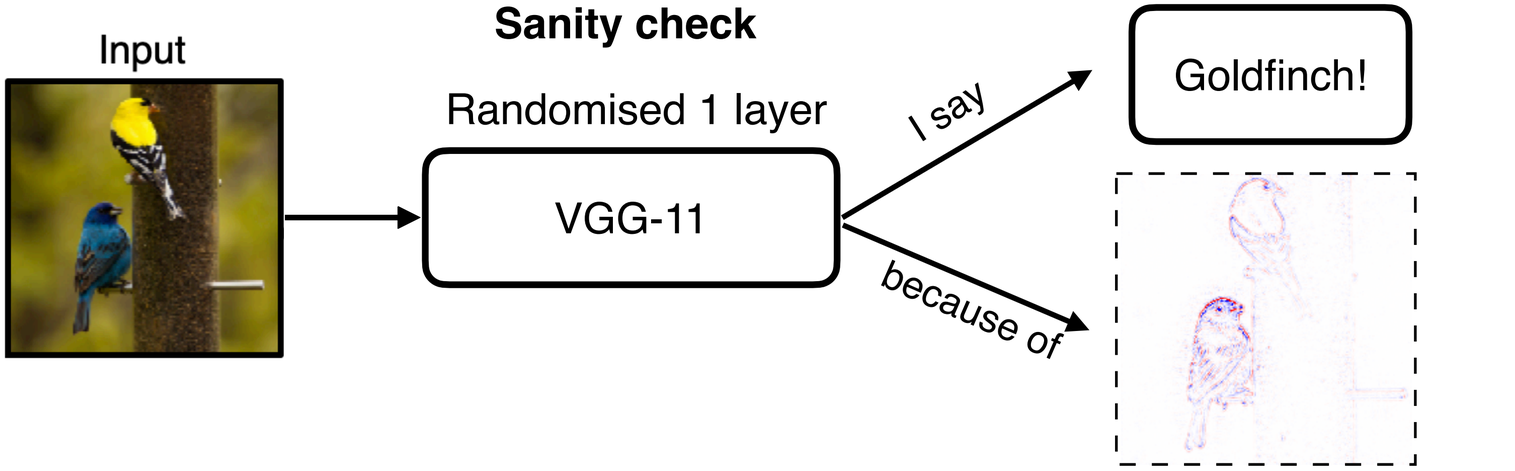
Want

Explanation

Requirements

Faithful

Interpretable ?



Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

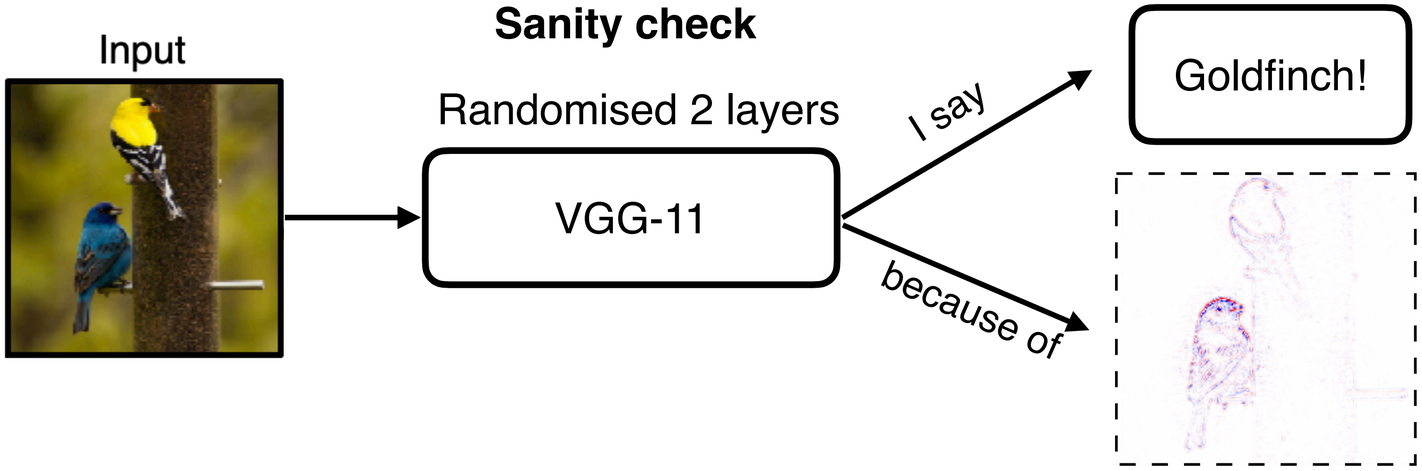
Want

Explanation

Requirements

Faithful

Interpretable ?



Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

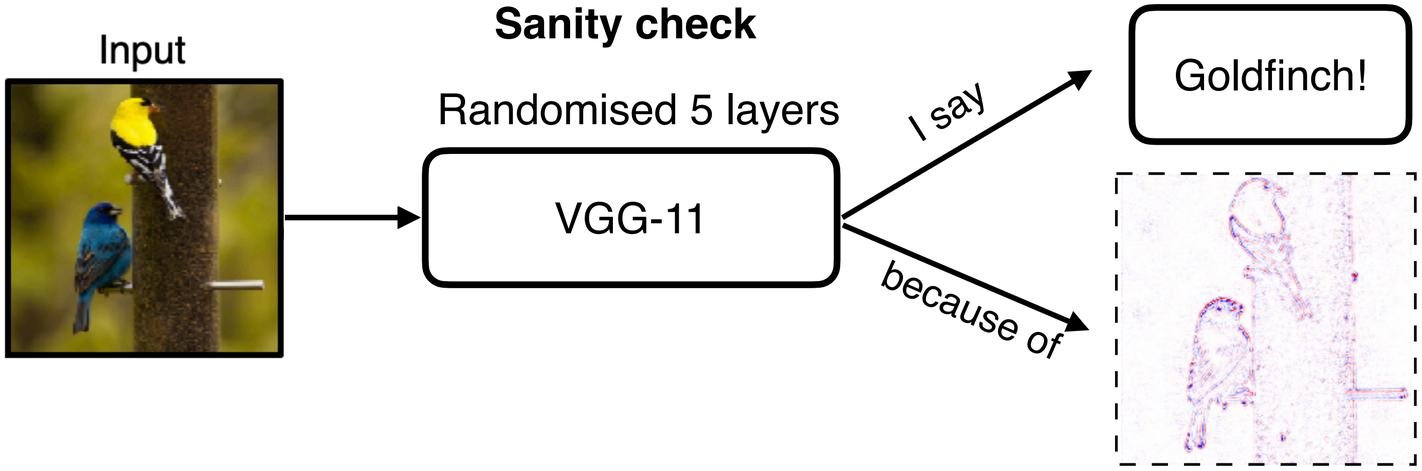
Want

Explanation

Requirements

Faithful

Interpretable ?



References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation

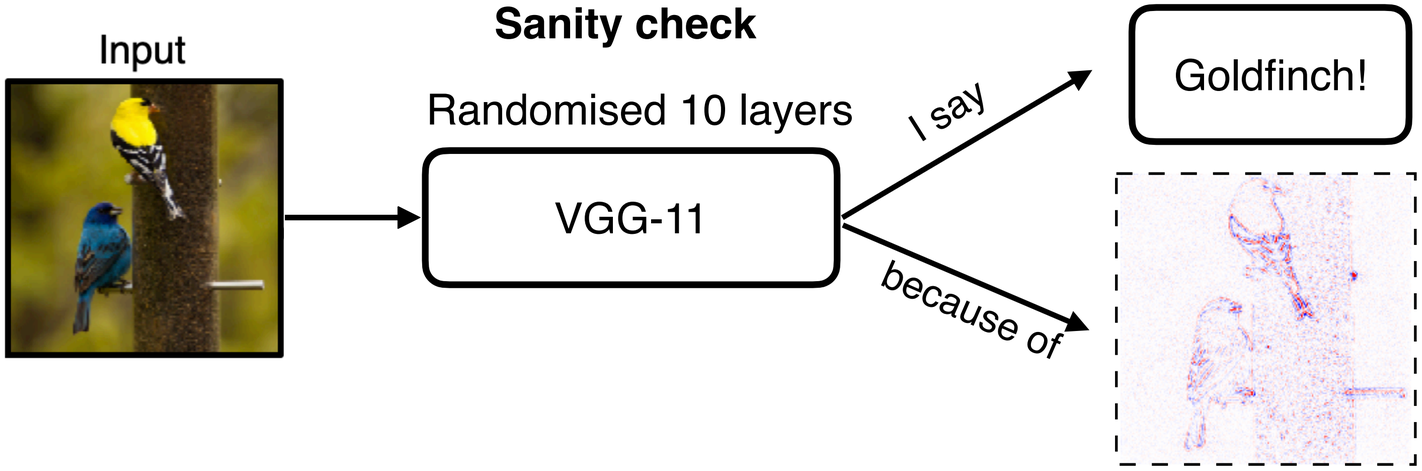
Want

Explanation

Requirements

Faithful ✗

Interpretable ✓



Guided Backpropagation

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation: we aim for Inherent Interpretability

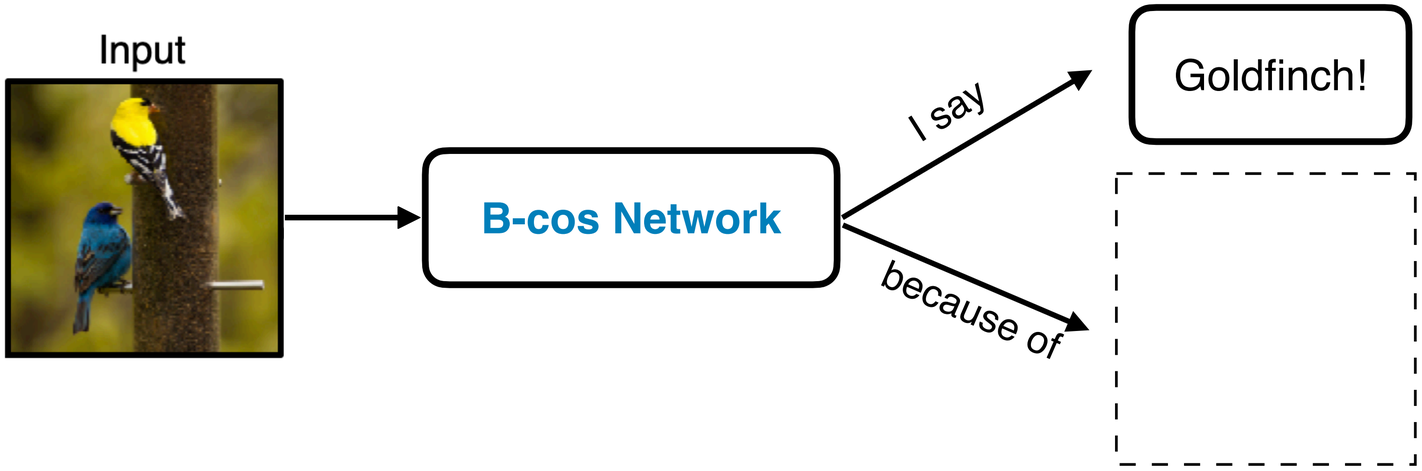
Want

Explanation

Requirements

Faithful

Interpretable



Model-inherent linear map

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation: we aim for Inherent Interpretability

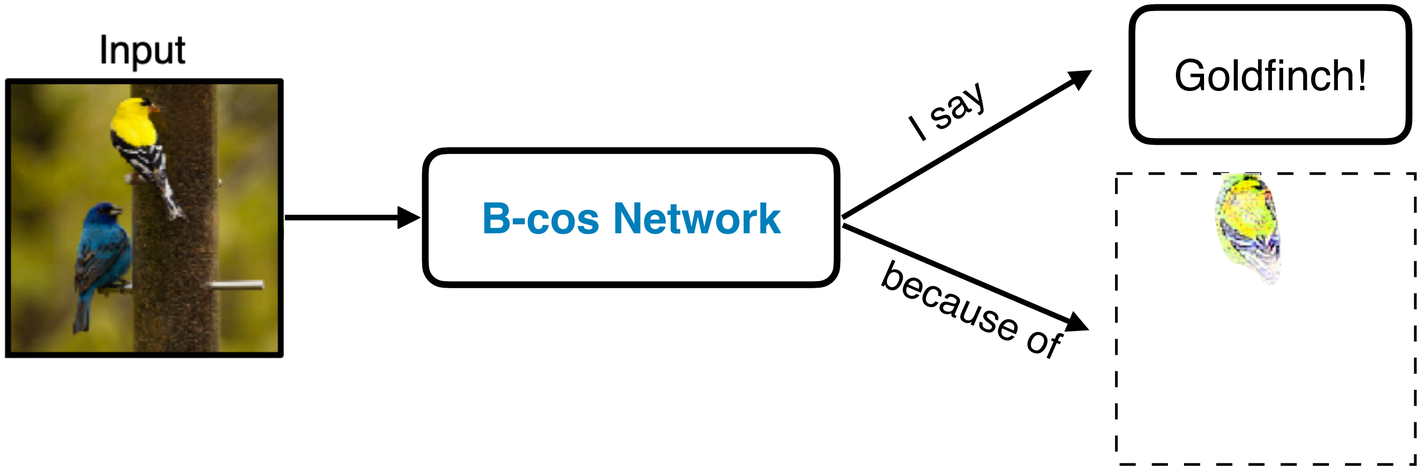
Want

Explanation

Requirements

Faithful

Interpretable



Model-inherent linear map

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation: we aim for Inherent Interpretability

Want

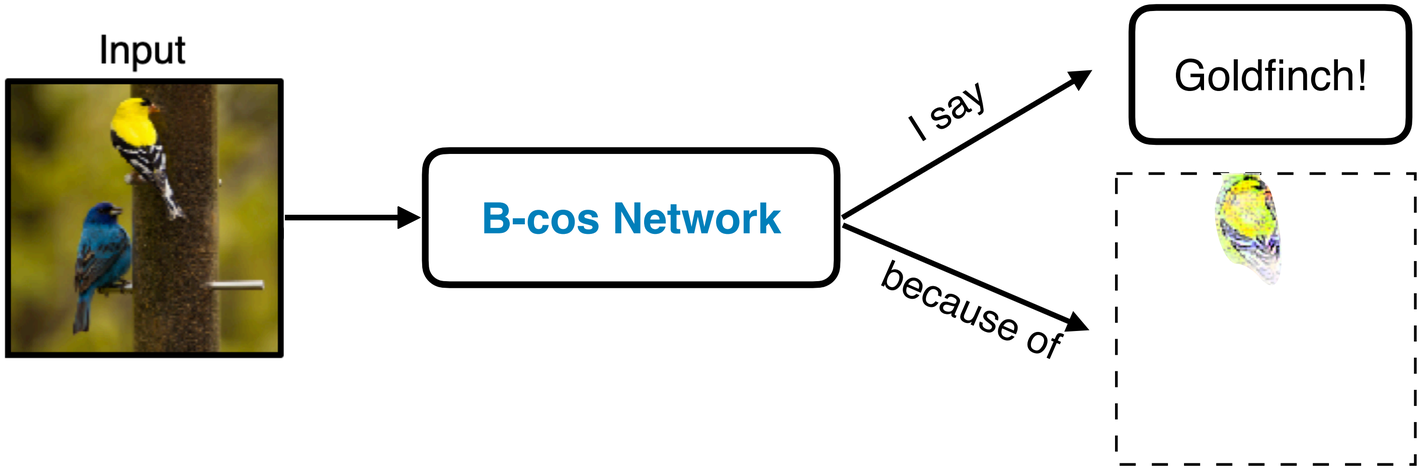
Explanation

Requirements

Faithful ✓

Interpretable

Dynamic linearity



Model-inherent linear map

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation: we aim for Inherent Interpretability

Want

Explanation

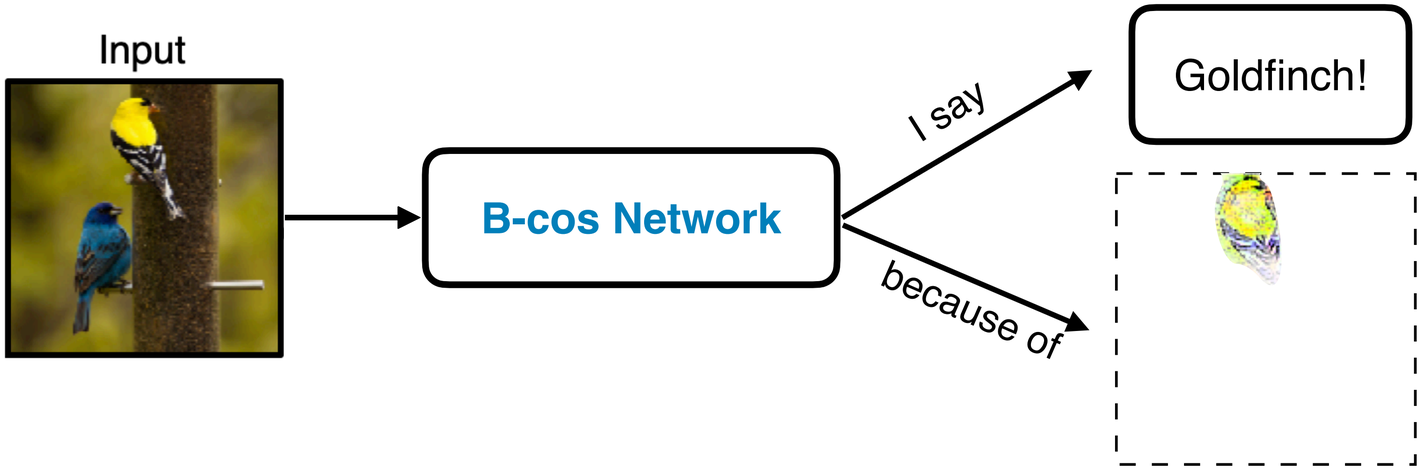
Requirements

Faithful ✓

Interpretable ✓

Dynamic linearity

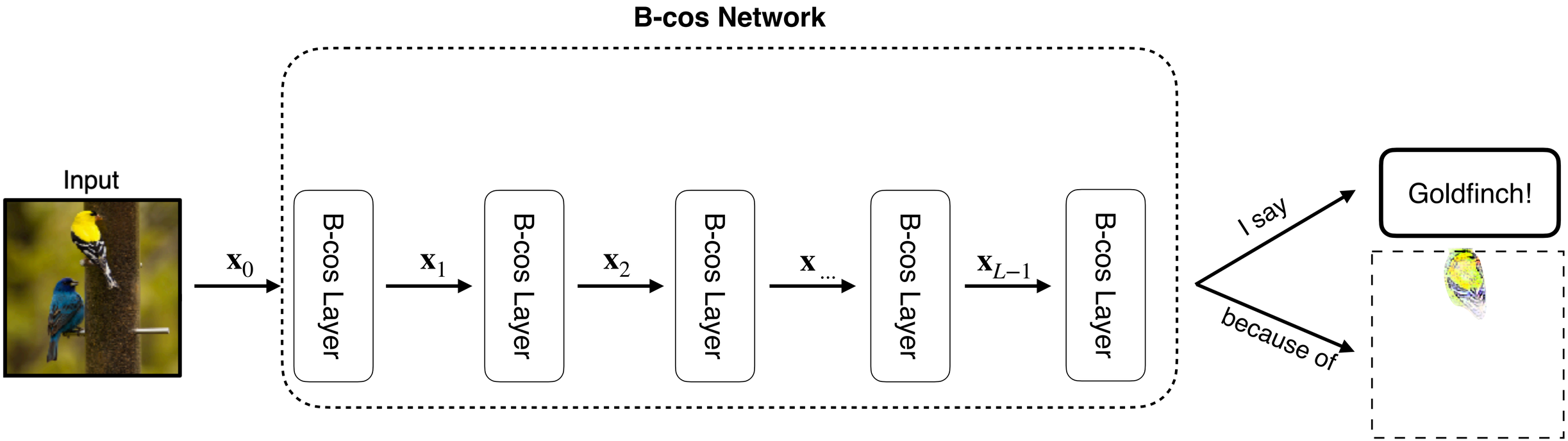
Alignment pressure



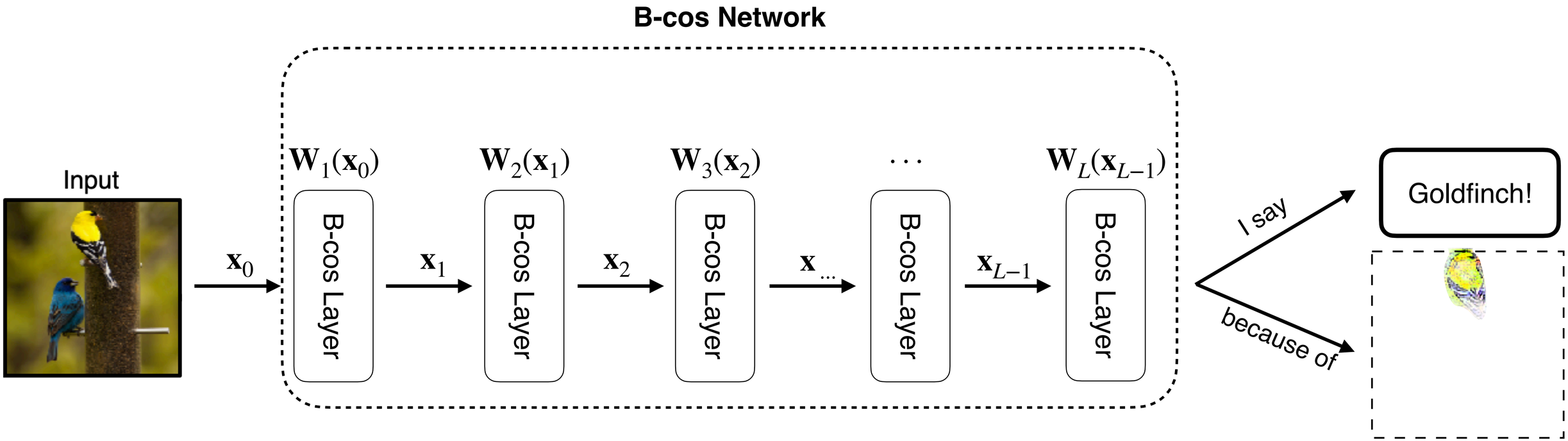
Model-inherent linear map

References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

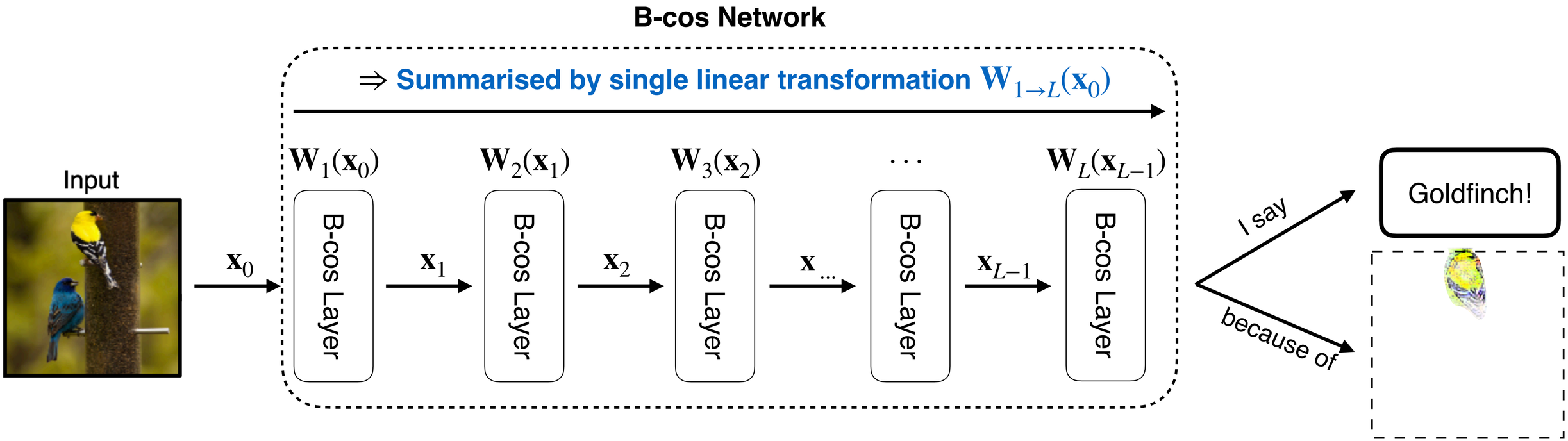
B-cos Networks: Dynamic Linearity



B-cos Networks: Dynamic Linearity

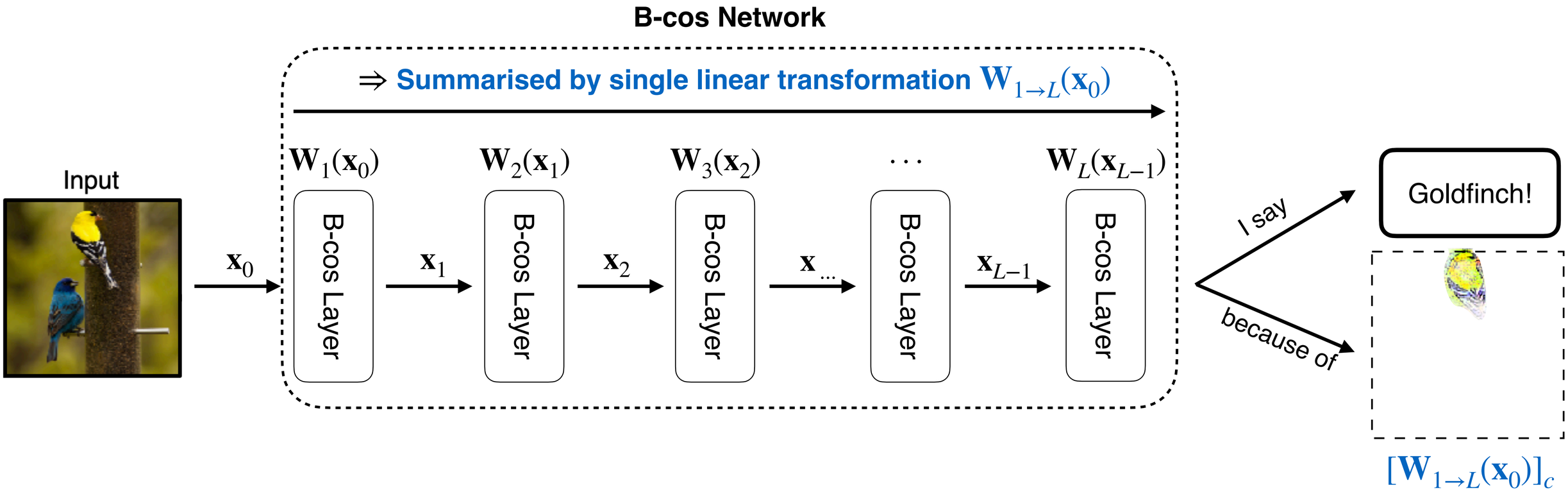


B-cos Networks: Dynamic Linearity



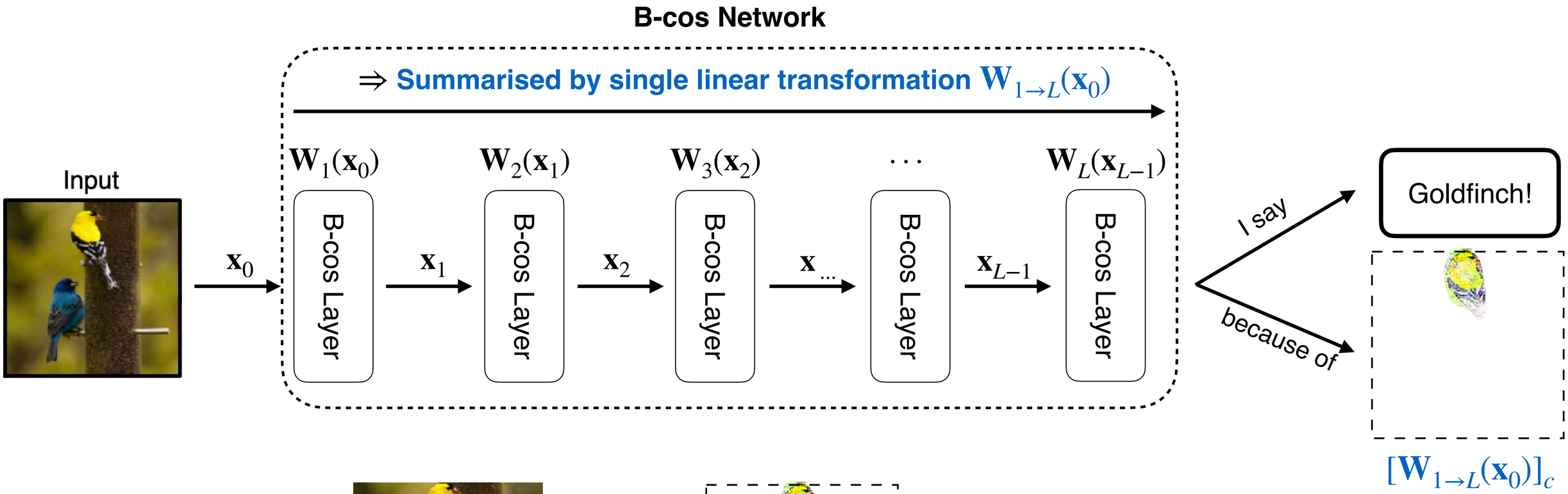
W changes with input

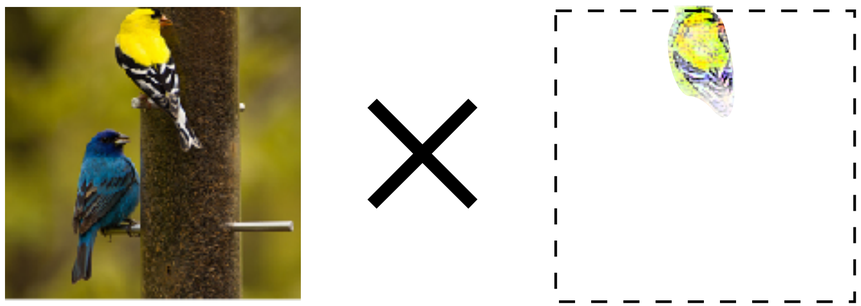
B-cos Networks: Dynamic Linearity



explanation based on W

B-cos Networks: Dynamic Linearity

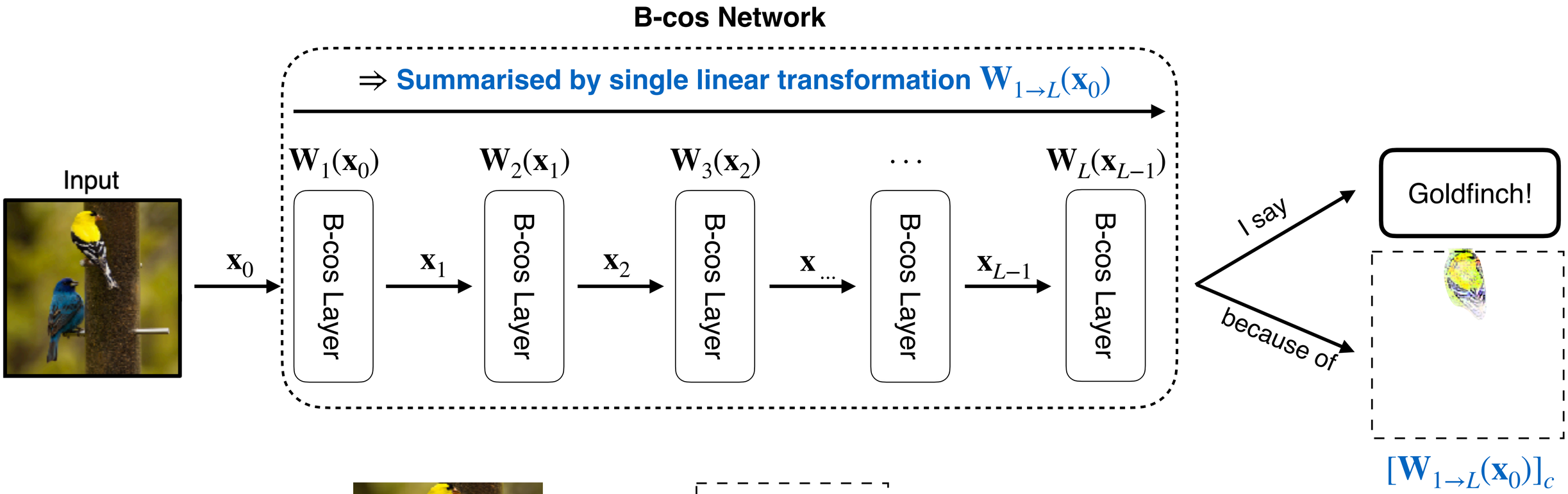


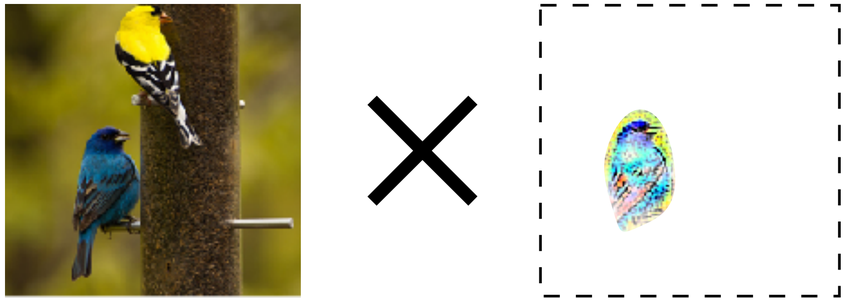


Goldfinch

48.8%

B-cos Networks: Dynamic Linearity

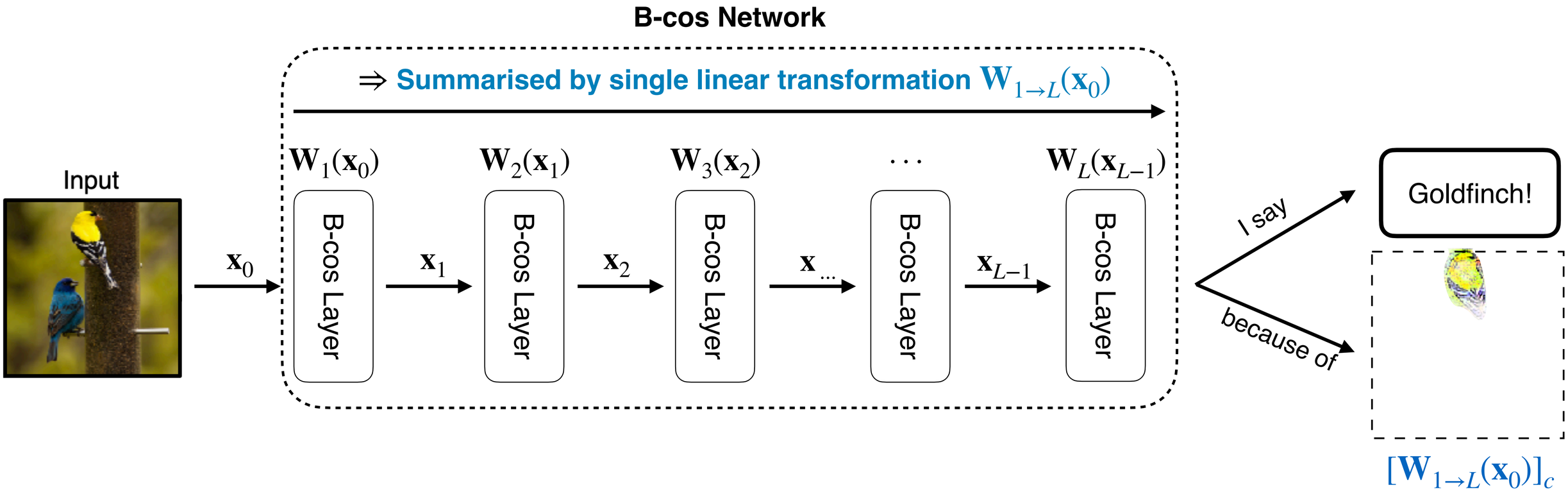




Bunting

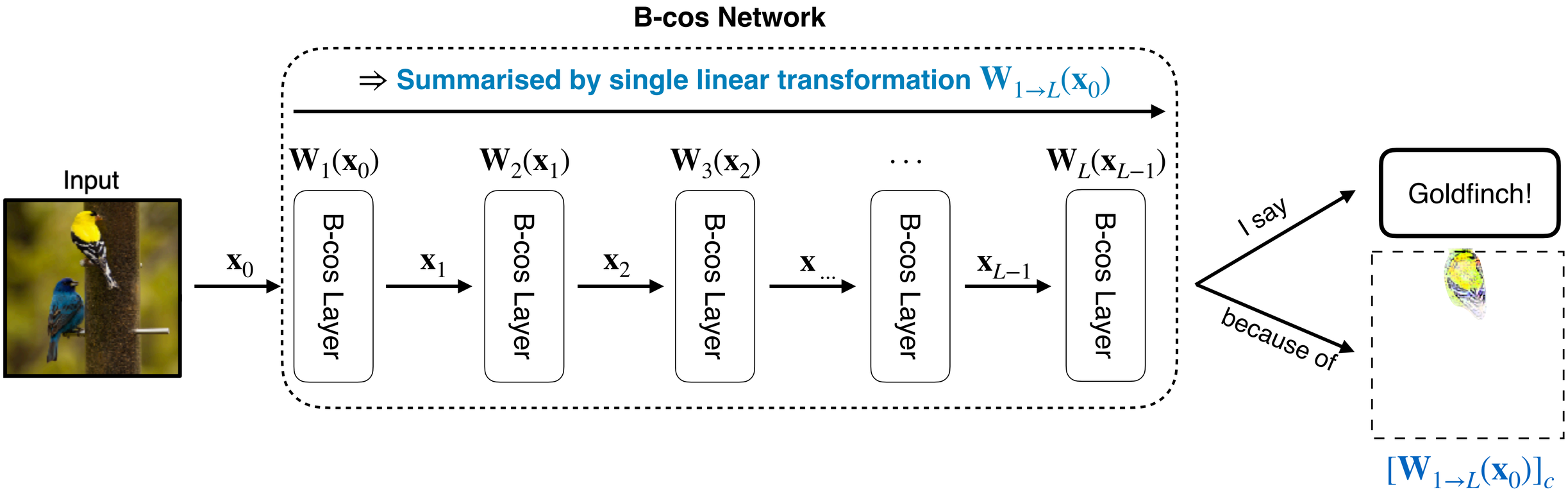
47.0%

B-cos Networks: Dynamic Linearity



Dynamic linearity allows us to faithfully summarise the model.

B-cos Networks: Dynamic Linearity



Dynamic linearity allows us to faithfully summarise the model.

But why should $\mathbf{W}_{1\to L}(\mathbf{x}_0)$ align with relevant features?

Alignment pressure

B-cos transformation vs. linear transformation

Linear transformation $f(\mathbf{x};\ \mathbf{w}) = \mathbf{w}^T\ \mathbf{x} = ||\mathbf{w}||\ ||\mathbf{x}||\ \cos(\mathbf{x}, \mathbf{w})$

B-cos transformation vs. linear transformation

Linear transformation $f(\mathbf{x};\ \mathbf{w}) = \mathbf{w}^T\ \mathbf{x} = ||\mathbf{w}||\ ||\mathbf{x}||\ \cos(\mathbf{x}, \mathbf{w})$

$= ||\mathbf{w}||\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})| \times \text{sgn}\left(\cos(\mathbf{w}, \mathbf{x})\right)$

B-cos transformation vs. linear transformation

Linear transformation $f(\mathbf{x};\ \mathbf{w}) = \mathbf{w}^T\ \mathbf{x} = ||\mathbf{w}||\ ||\mathbf{x}||\ \cos(\mathbf{x}, \mathbf{w})$

$= ||\mathbf{w}||\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})| \times \text{sgn}\left(\cos(\mathbf{w}, \mathbf{x})\right)$

New transformation B-cos$(\mathbf{x};\ \mathbf{w}) = \underbrace{||\ \widehat{\mathbf{w}}\ ||}_{=1}\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})|^{B} \times \text{sgn}\left(\cos(\mathbf{x}, \mathbf{w})\right)$

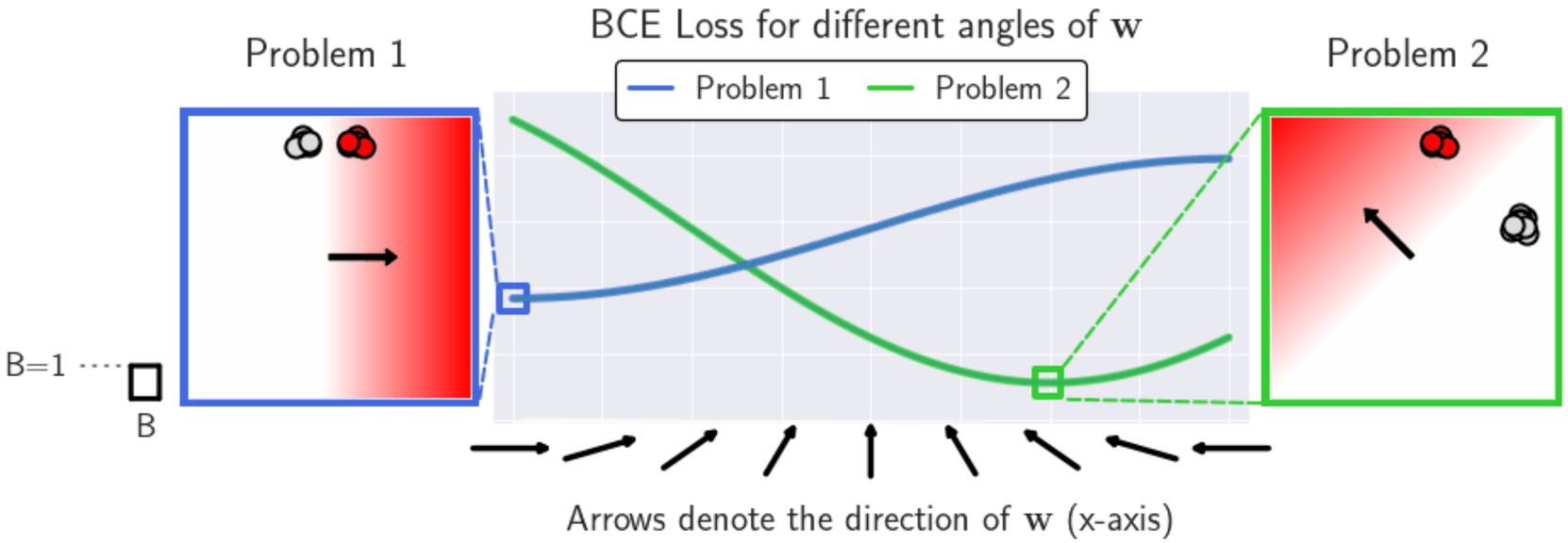
high cosine to the input
B: increase the importance of cosine term

B-cos transformation vs. linear transformation

Linear transformation $f(\mathbf{x};\ \mathbf{w}) = \mathbf{w}^T\ \mathbf{x} = ||\mathbf{w}||\ ||\mathbf{x}||\ \cos(\mathbf{x}, \mathbf{w})$

$= ||\mathbf{w}||\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})| \times \text{sgn}\left(\cos(\mathbf{w}, \mathbf{x})\right)$

New transformation B-cos$(\mathbf{x};\ \mathbf{w}) = \underbrace{||\ \widehat{\mathbf{w}}\ ||}_{=1}\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})|^{B} \times \text{sgn}\left(\cos(\mathbf{x}, \mathbf{w})\right)$

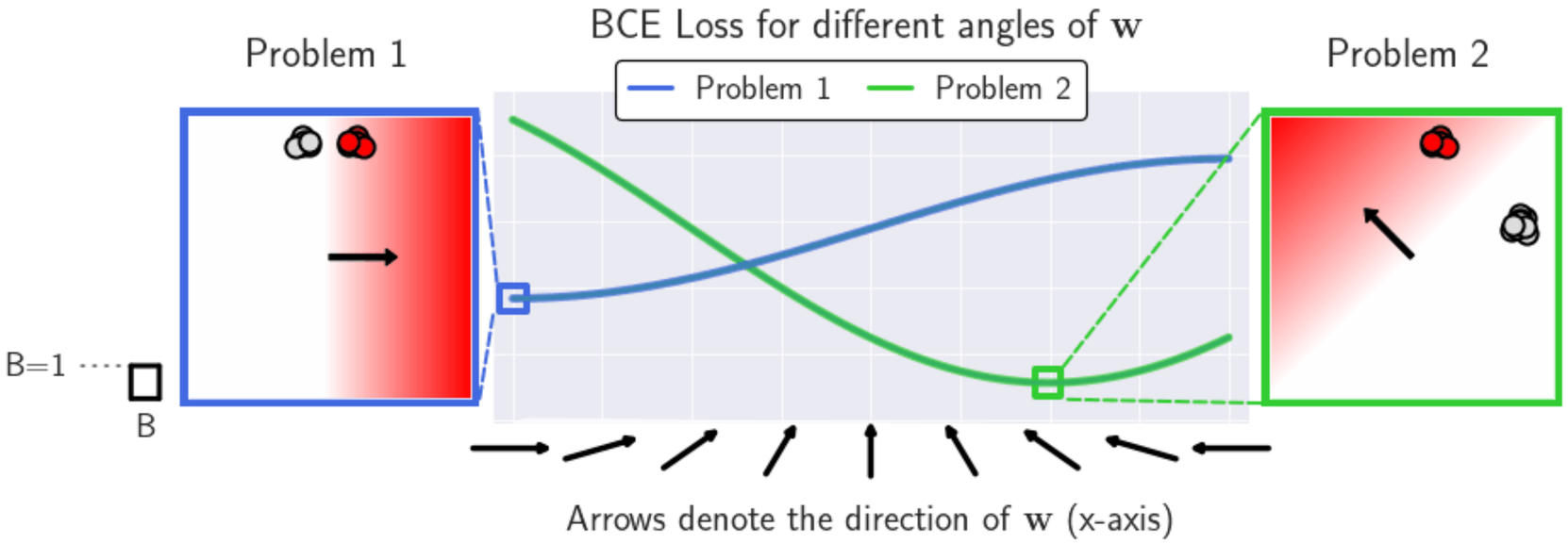


B-cos transformation vs. linear transformation

Linear transformation $f(\mathbf{x};\ \mathbf{w}) = \mathbf{w}^T\ \mathbf{x} = ||\mathbf{w}||\ ||\mathbf{x}||\ \cos(\mathbf{x}, \mathbf{w})$

$= ||\mathbf{w}||\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})| \times \text{sgn}\left(\cos(\mathbf{w}, \mathbf{x})\right)$

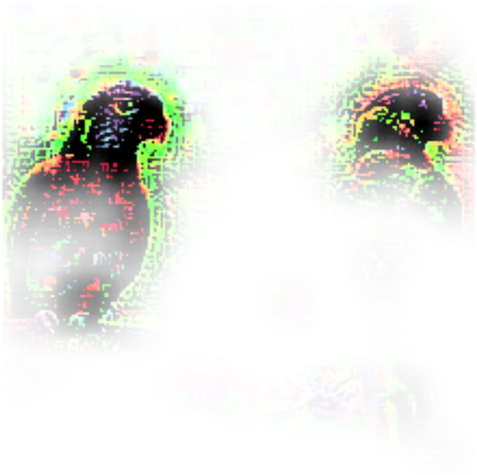
New transformation B-cos$(\mathbf{x};\ \mathbf{w}) = \underbrace{||\ \widehat{\mathbf{w}}\ ||}_{=1}\ ||\mathbf{x}||\ |\cos(\mathbf{x}, \mathbf{w})|^{B} \times \text{sgn}\left(\cos(\mathbf{x}, \mathbf{w})\right)$



Results

- Quantitative results on ImageNet
  - Classification accuracy
  - Interpretability scores

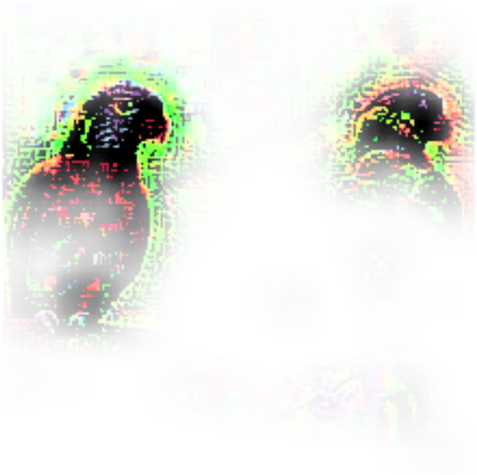
- Some qualitative results on ImageNet
  - Explaining class logits
  - explaining intermediate neurons



Results

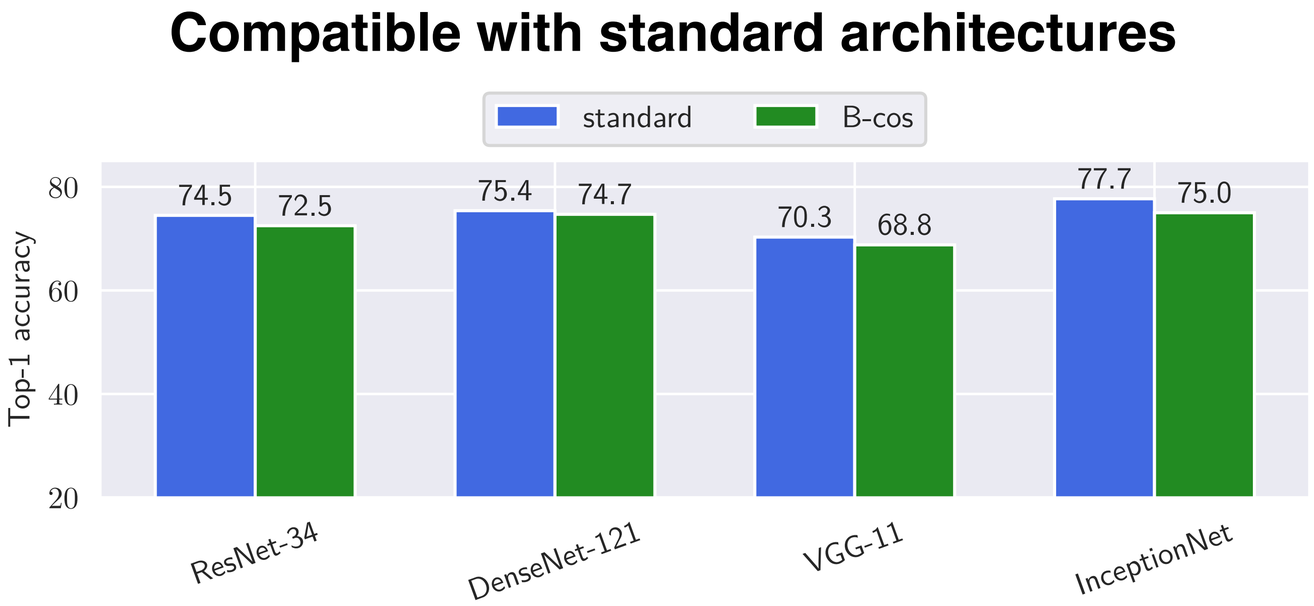
- Quantitative results on ImageNet
  - Classification accuracy
  - Interpretability scores

- Some qualitative results on ImageNet
  - Explaining class logits
  - explaining intermediate neurons



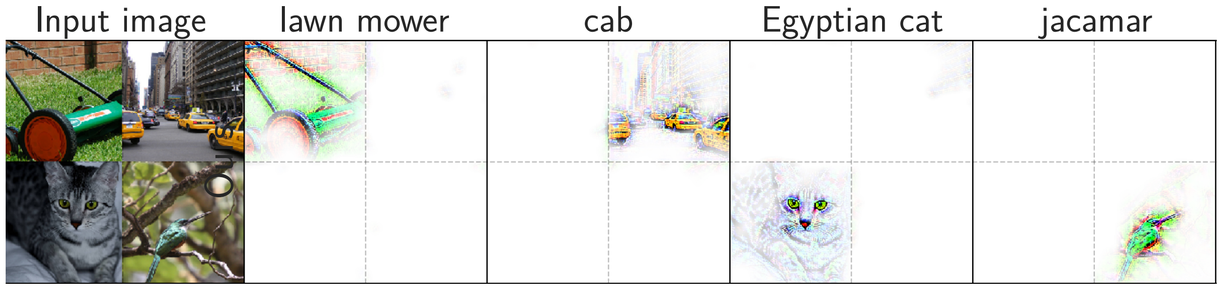
ImageNet results

- B-cos networks show competitive performance
- Importantly, they exhibit significantly higher interpretability



Measuring Interpretability via Grid Pointing Game

- To measure interpretability, we employ the *grid pointing game*

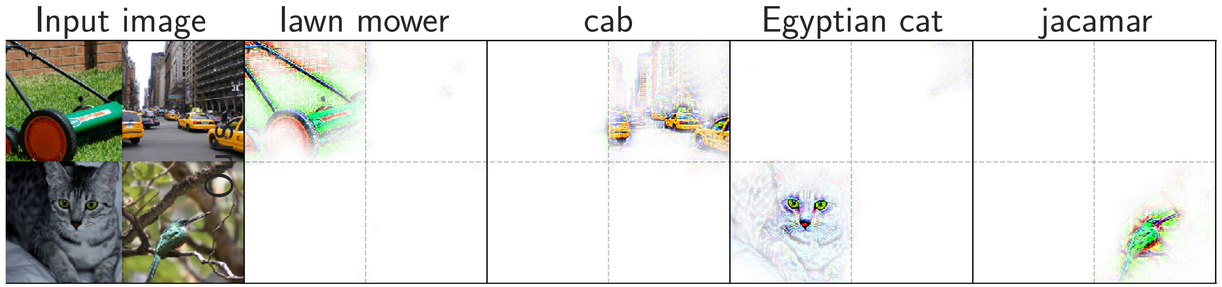


Measuring Interpretability via Grid Pointing Game

- To measure interpretability, we employ the *grid pointing game*

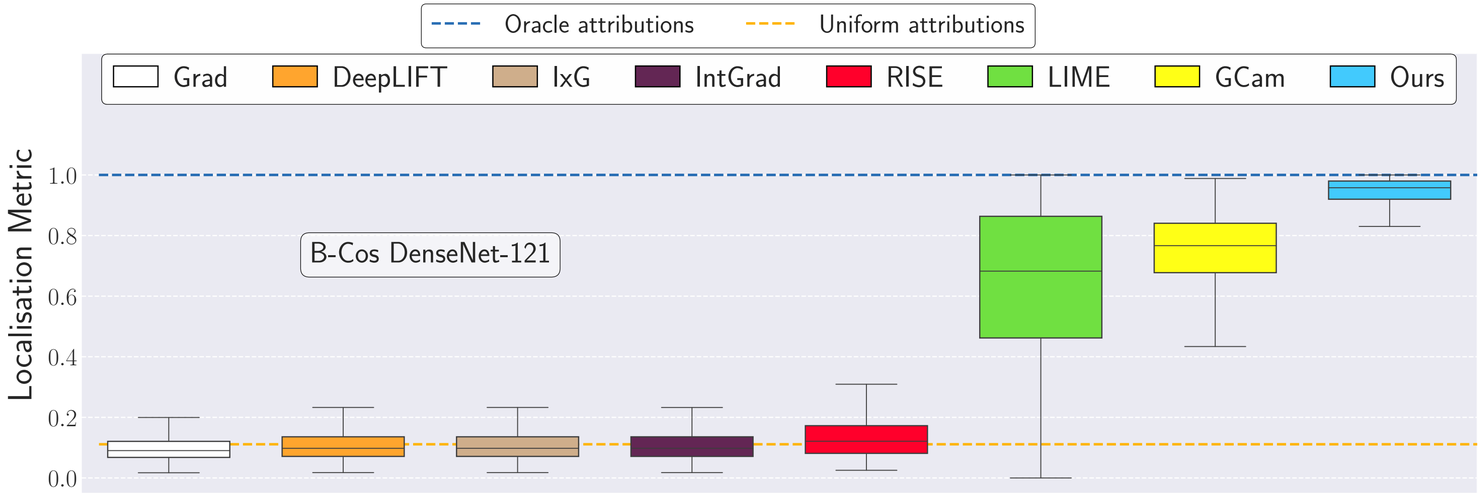
- In particular:
  - evaluate models on synthetic image grid
  - measure how well an explanation *localises* the correct image grid

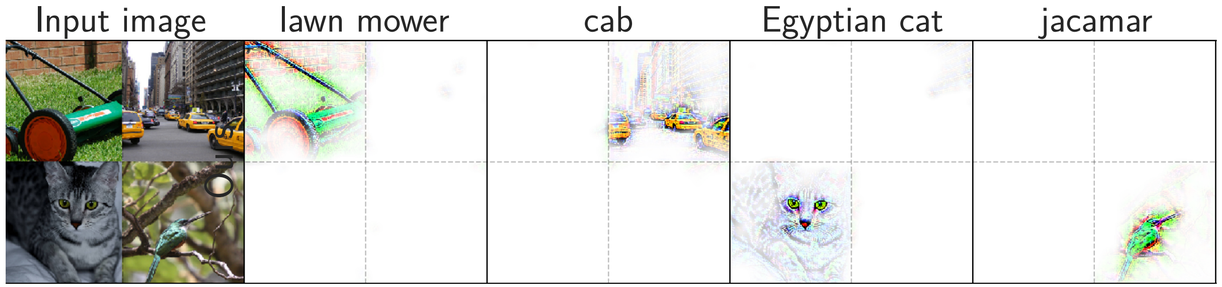
    (score $s = \frac{A_i^+}{\sum_j A_j^+}$ with A_i^+ the positive attribution to subimage i)



ImageNet results

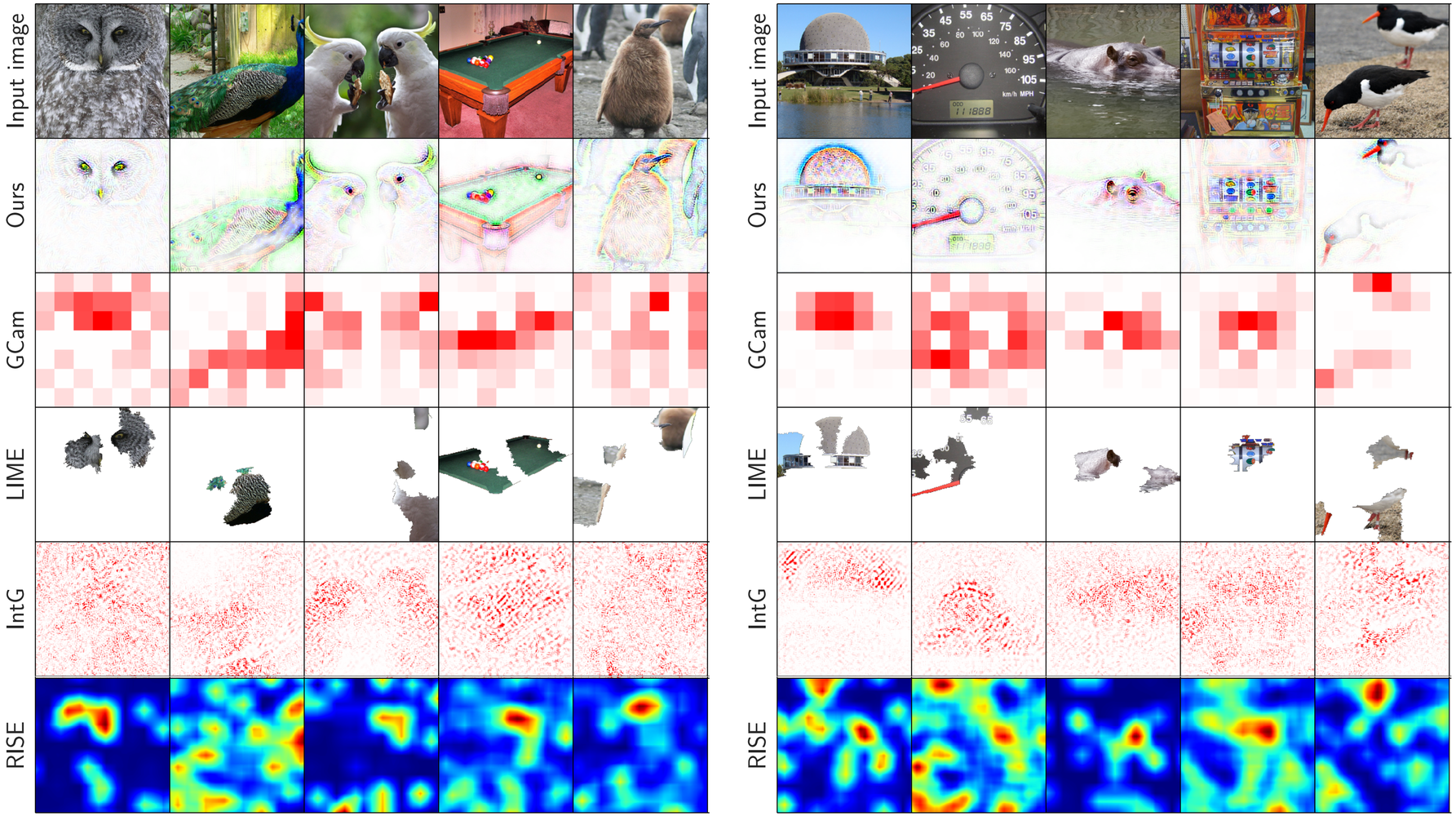
High interpretability





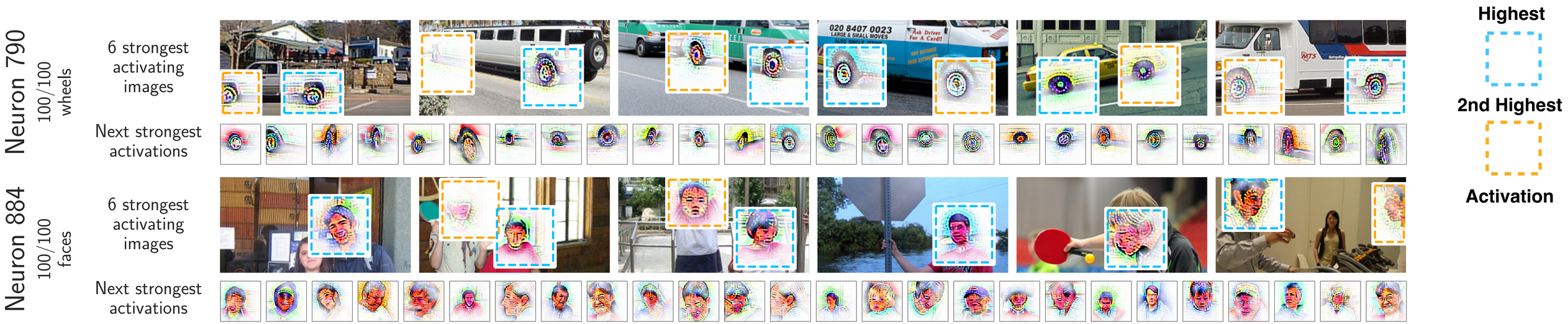
Gradient (Baehrens (2010)), DeepLIFT (Shrikumar (2017)), Input x Gradient (cf. Adebayo (2018)), IntGrad (Sundararajan (2017)), RISE (Petsiuk, 2018), LIME (Ribeiro, 2016), GradCam (Ramprasaath et al. (2017))

Visualisations of $\mathbf{W}_{1\to L}(\mathbf{x})$

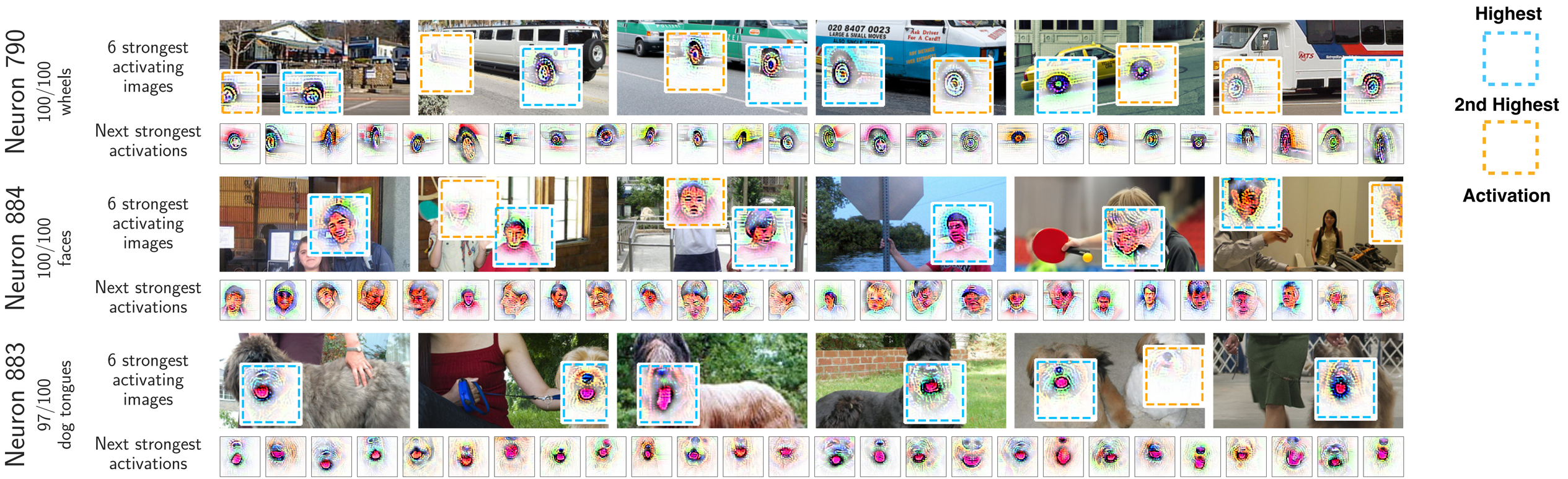


Visualisations: intermediate neurons

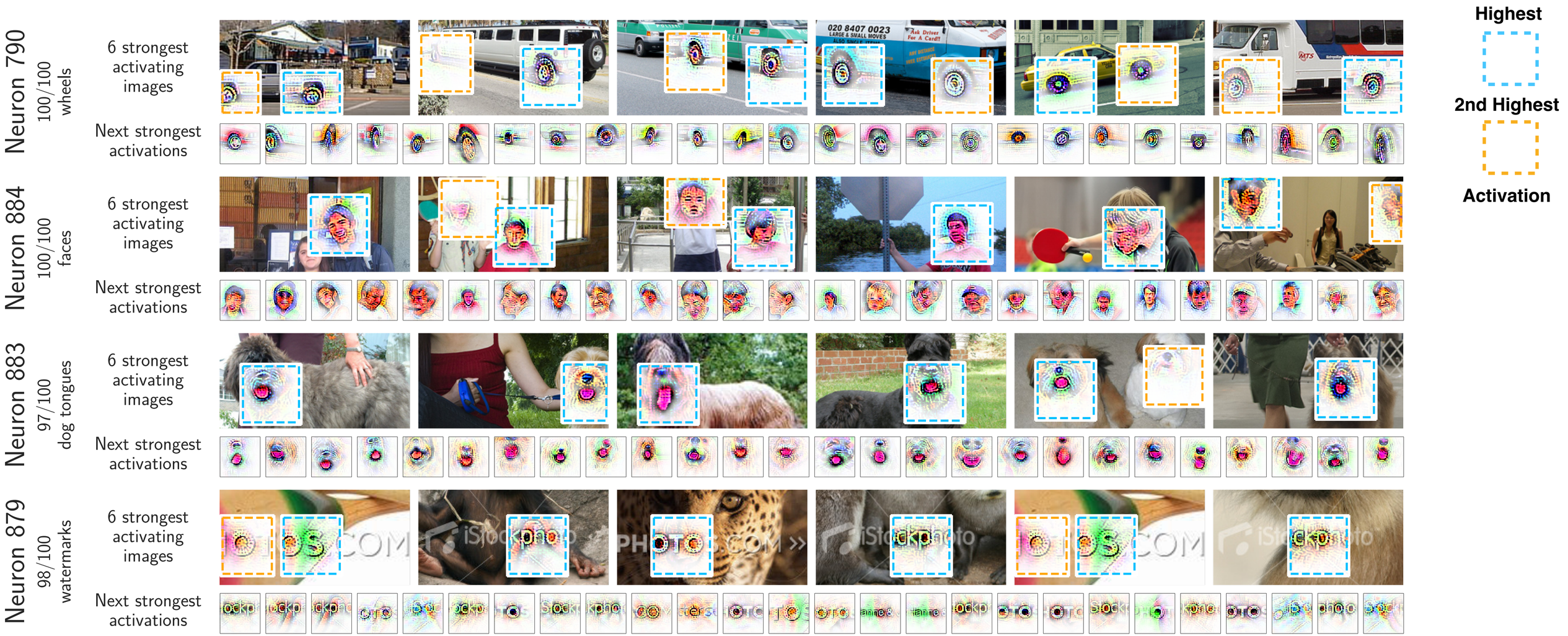
Visualisations: intermediate neurons



Visualisations: intermediate neurons



Visualisations: intermediate neurons



Interim Summary

- Deep Neural Network explanations need to be **faithful & interpretable**
  - for faithfulness: B-cos is designed to be **dynamic linear**
  - for interpretability: B-cos induces **alignment pressure**

- The resulting networks are **competitive classifiers**...

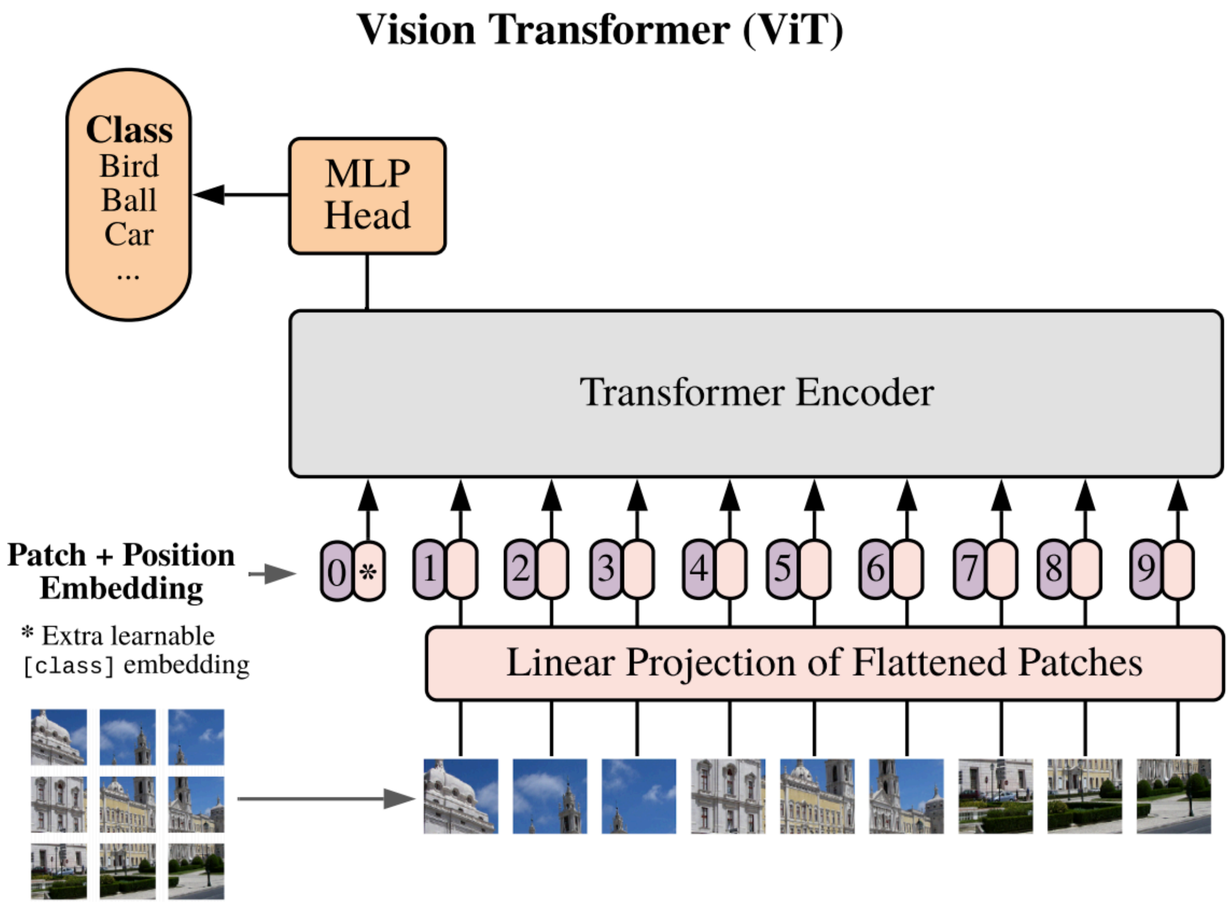
- ... and **provide interpretable explanations** for their decisions

Overview of Today's Lecture

- Inherently Interpretable Neural Network
  - B-cos CNN [cvpr'22] - https://arxiv.org/abs/2205.10268
  - B-cos CNN & Vision transformers [pami'24] - https://arxiv.org/abs/2306.10898

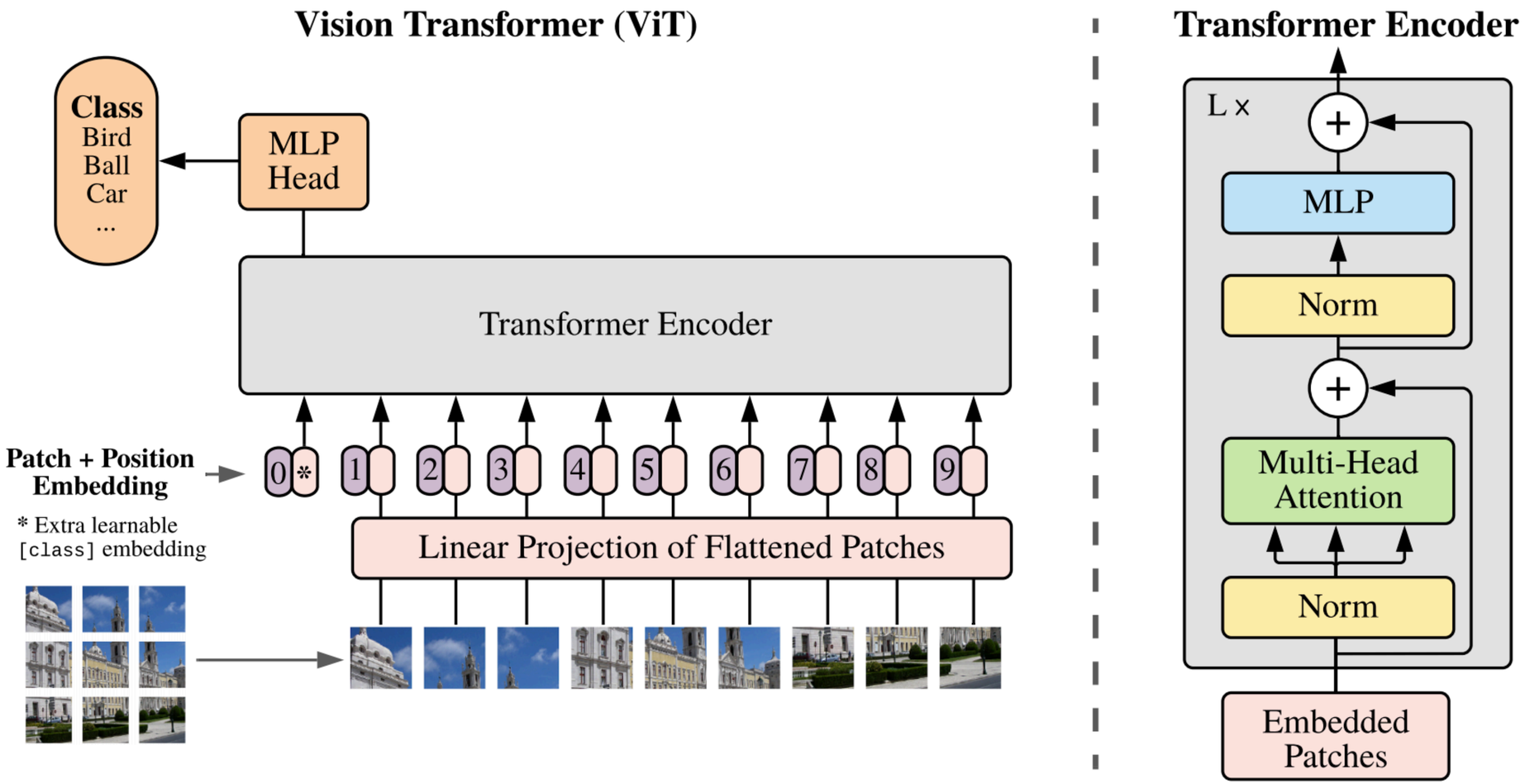
- Segment Anything Model (SAM)
  - [iccv'23] - https://arxiv.org/abs/2304.02643
  - https://segment-anything.com/

Vision Transformer (ViT)



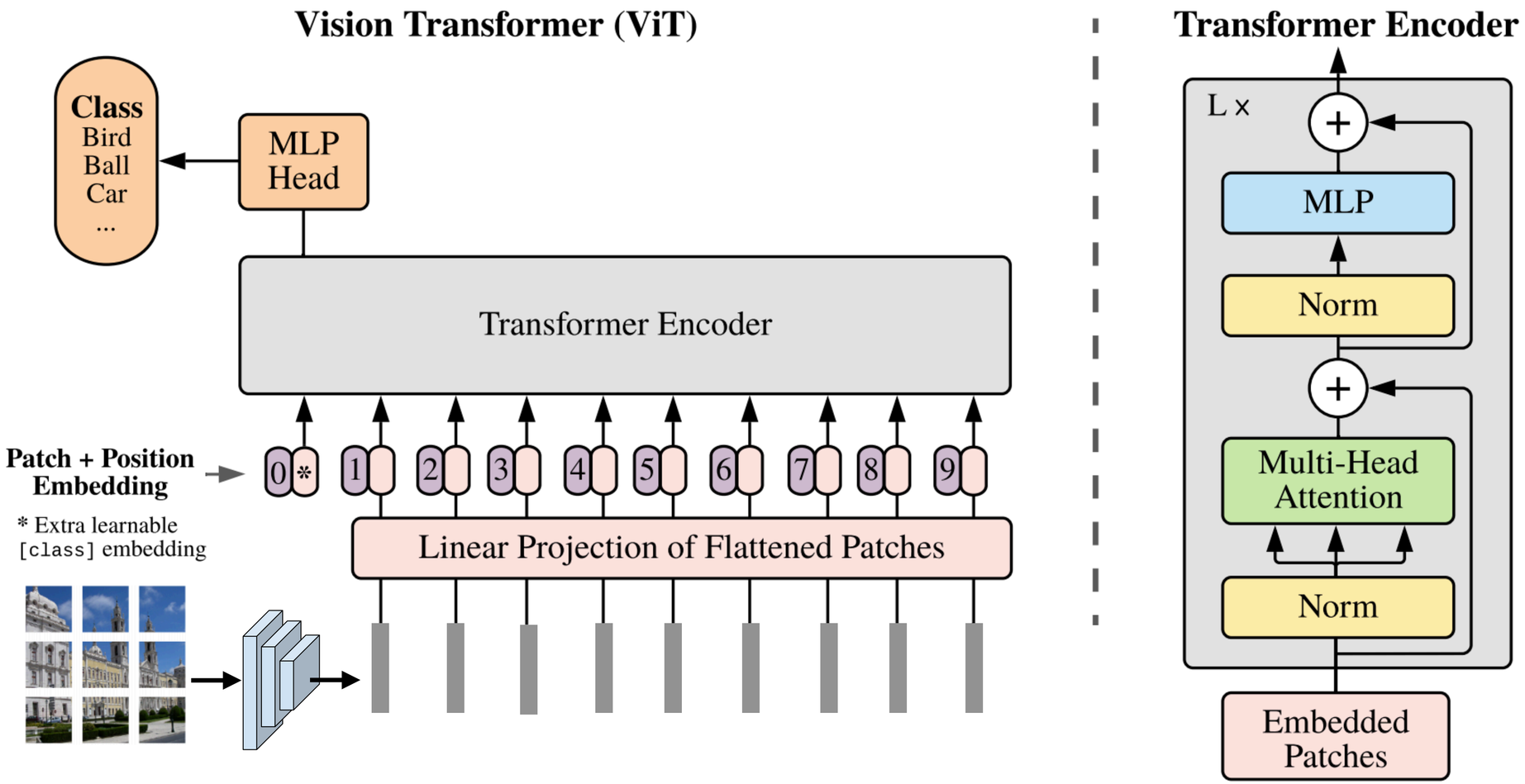
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. ICLR 2021

Vision Transformer (ViT)



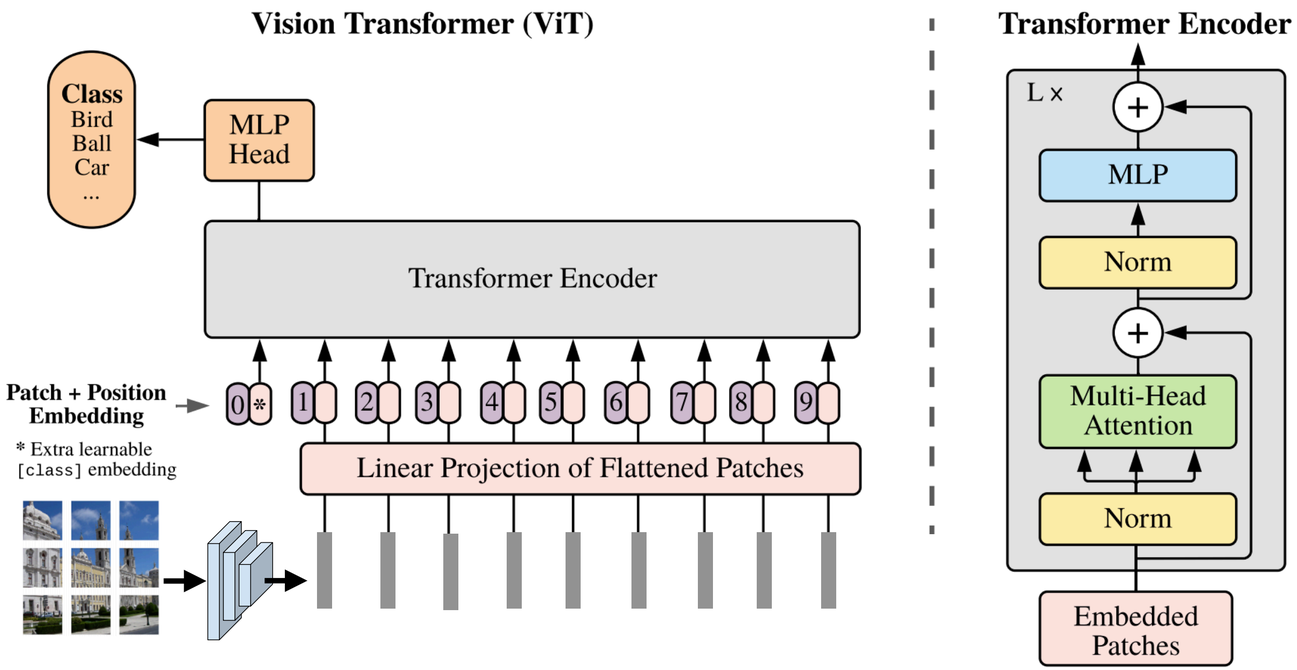
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. ICLR 2021

Hybrid ViT Architecture (with CNN-embedding of image patches)



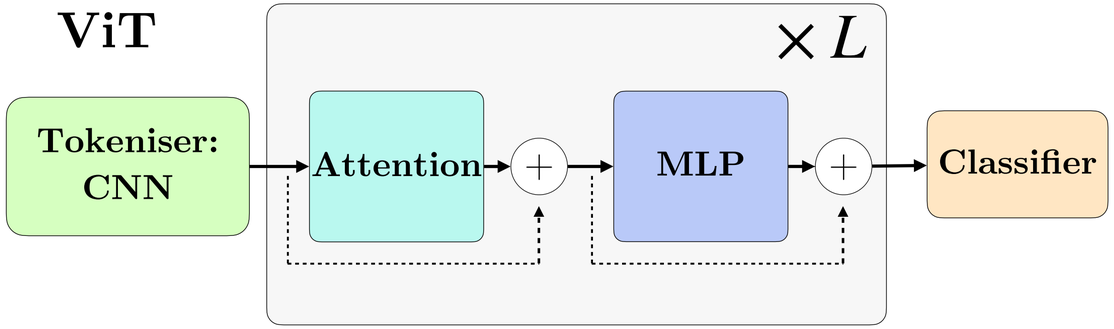
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. ICLR 2021

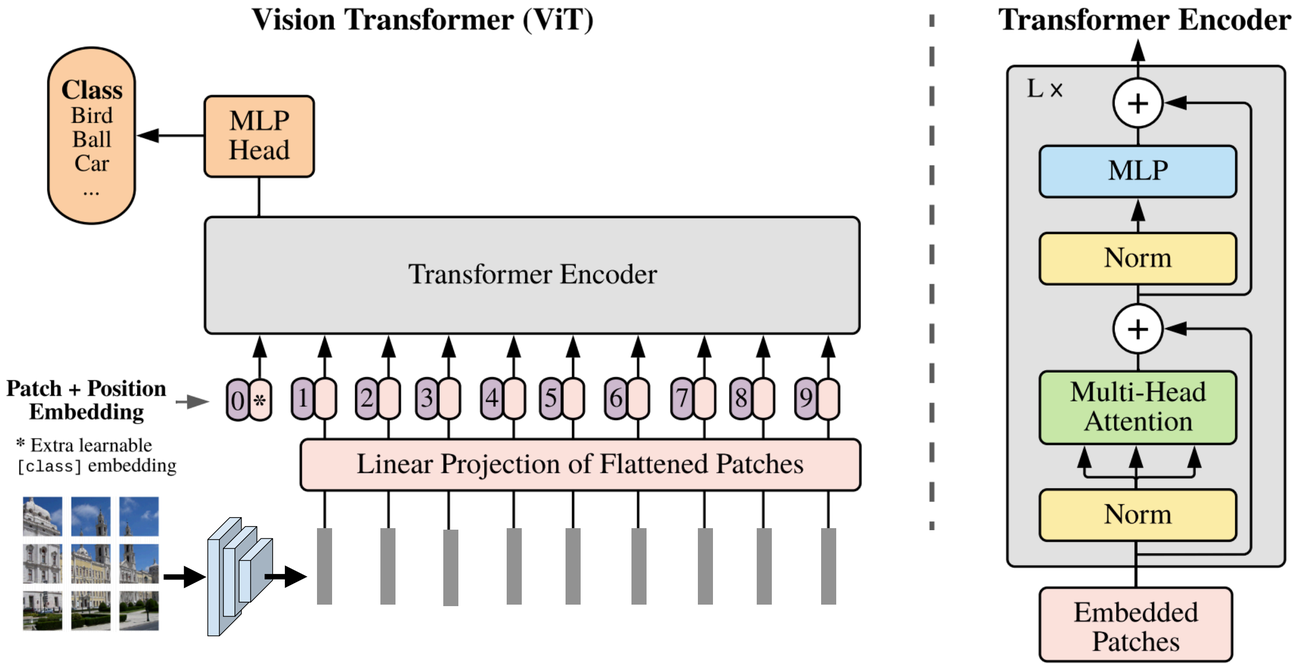
Attention is not All You Need (for XAI)



Dosovitskiy et al. 2021

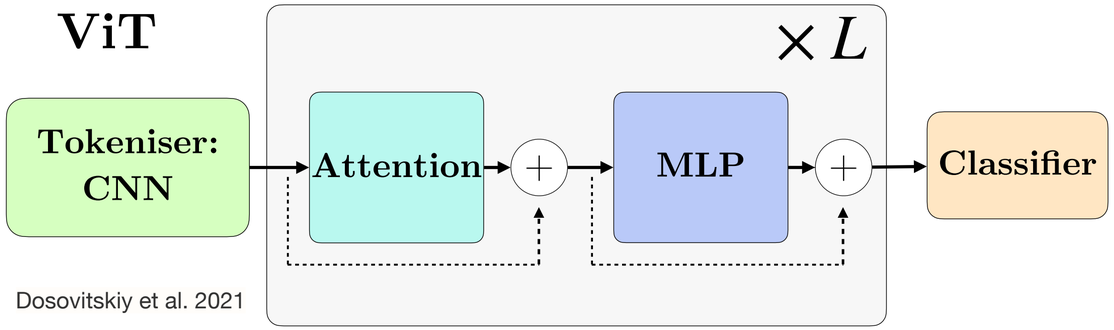
Attention is not All You Need (for XAI) XAI: explainable ai





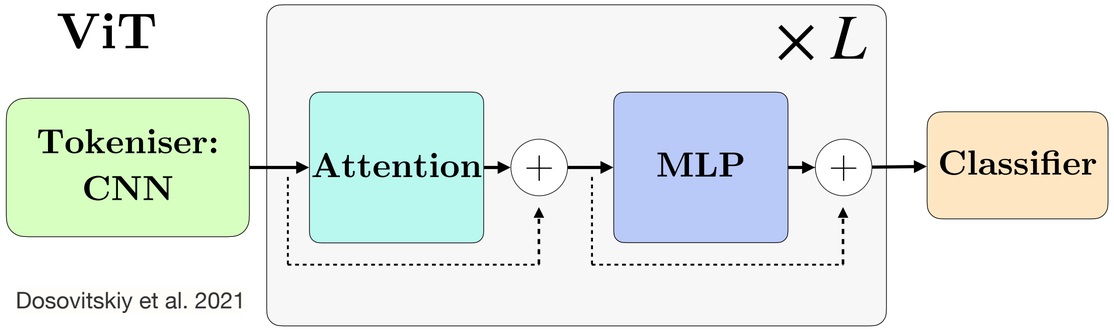
Dosovitskiy et al. 2021

Attention is not All You Need (for XAI)

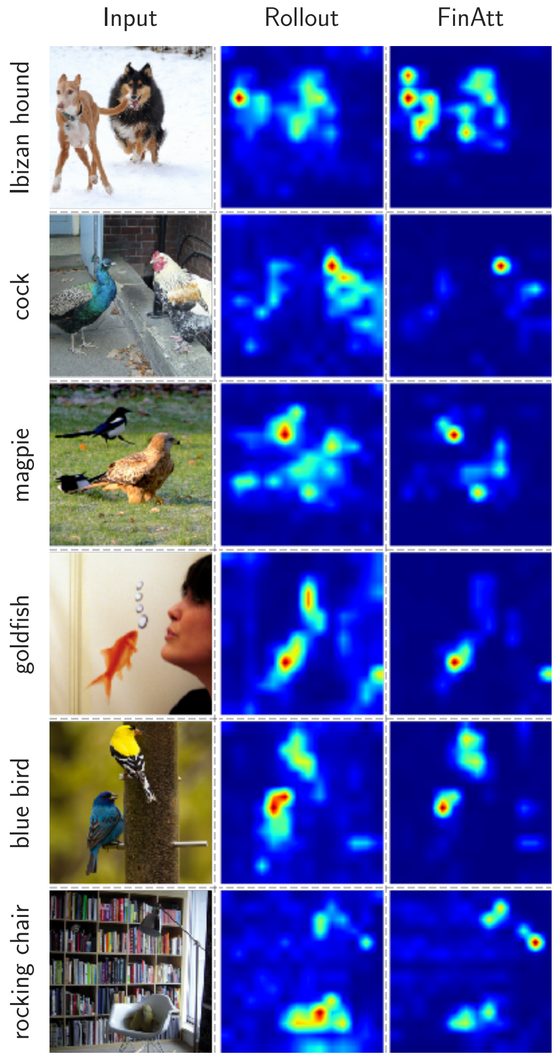


Dosovitskiy et al. 2021

Attention is not All You Need (for XAI)

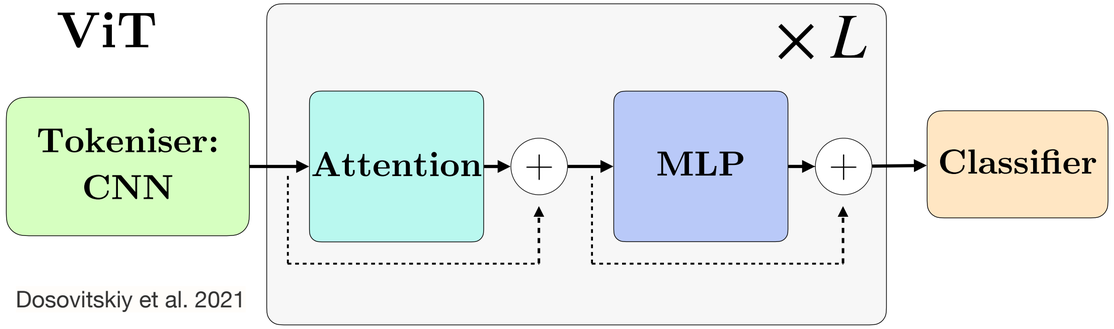


Dosovitskiy et al. 2021



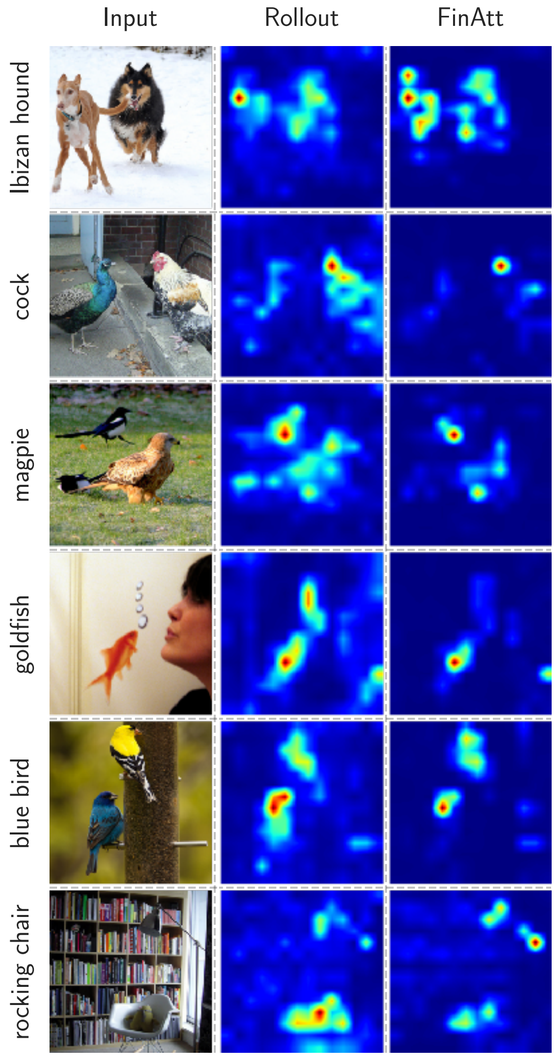
Attention Rollout: Abnar & Zuidema, ACL 2020

Attention is not All You Need (for XAI)



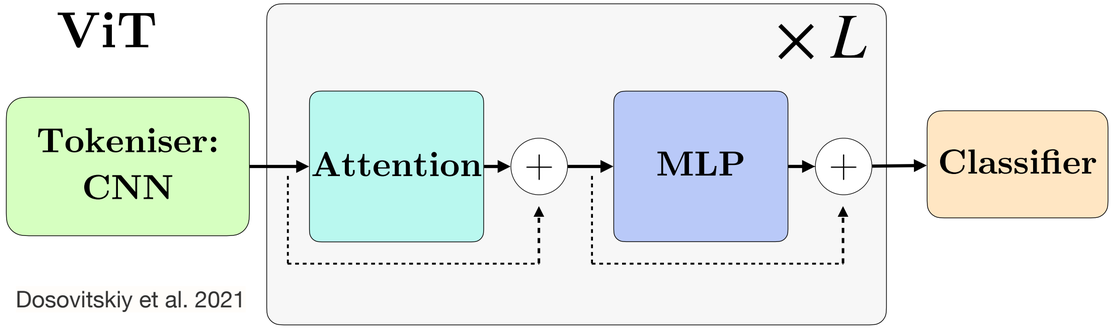
Dosovitskiy et al. 2021

High resolution



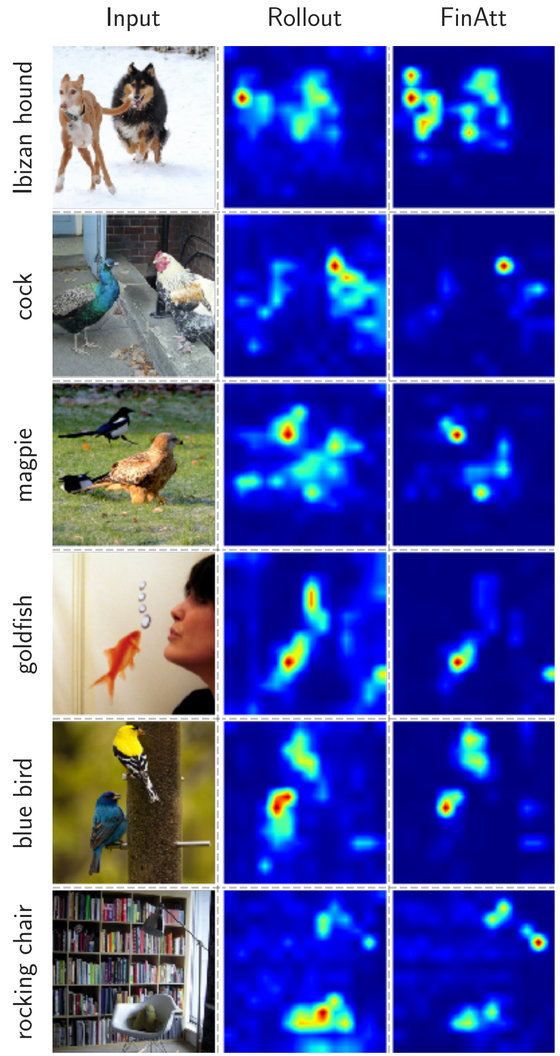
Attention Rollout: Abnar & Zuidema, ACL 2020

Attention is not All You Need (for XAI)



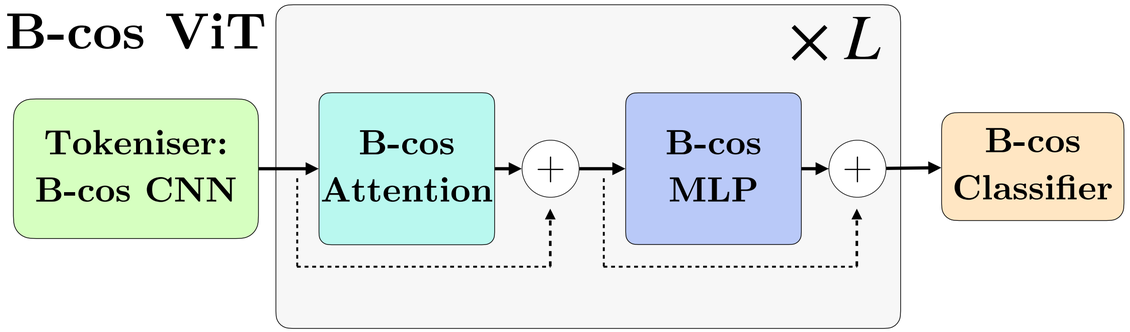
Dosovitskiy et al. 2021

- High resolution
- Holistic explanation

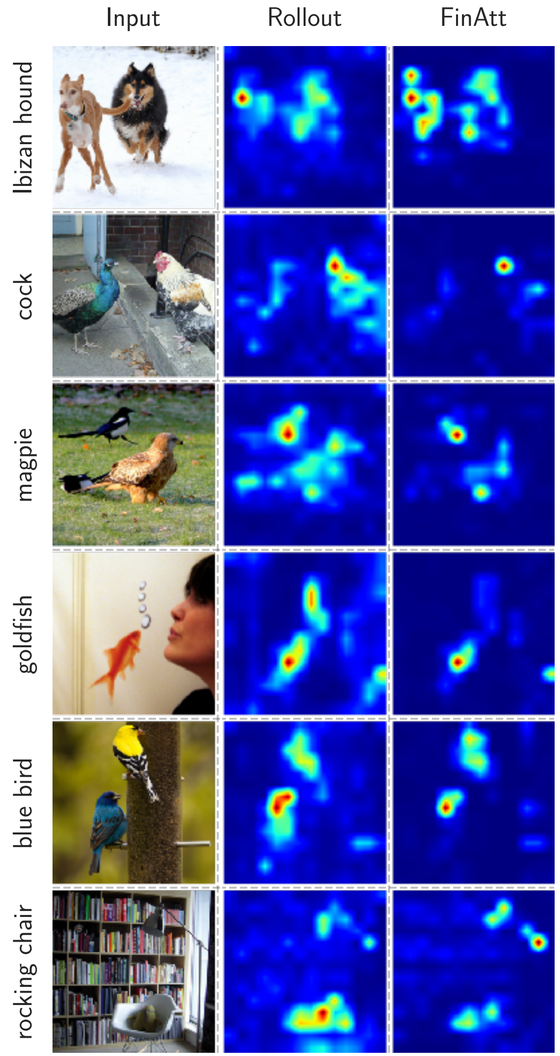


Attention Rollout: Abnar & Zuidema, ACL 2020

Attention is not All You Need (for XAI)

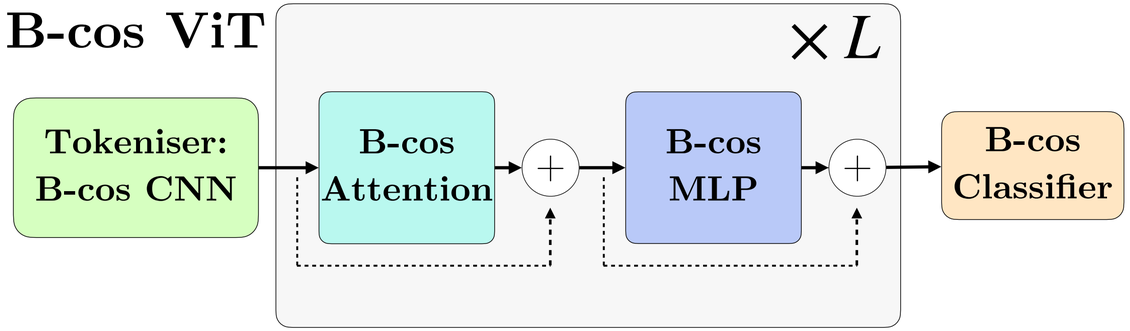


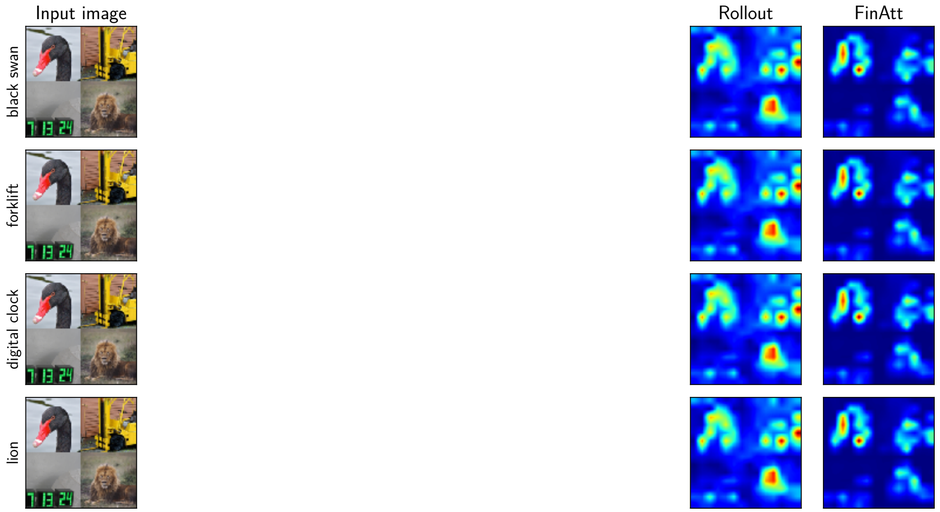
- ✓ High resolution
- ✓ Holistic explanation

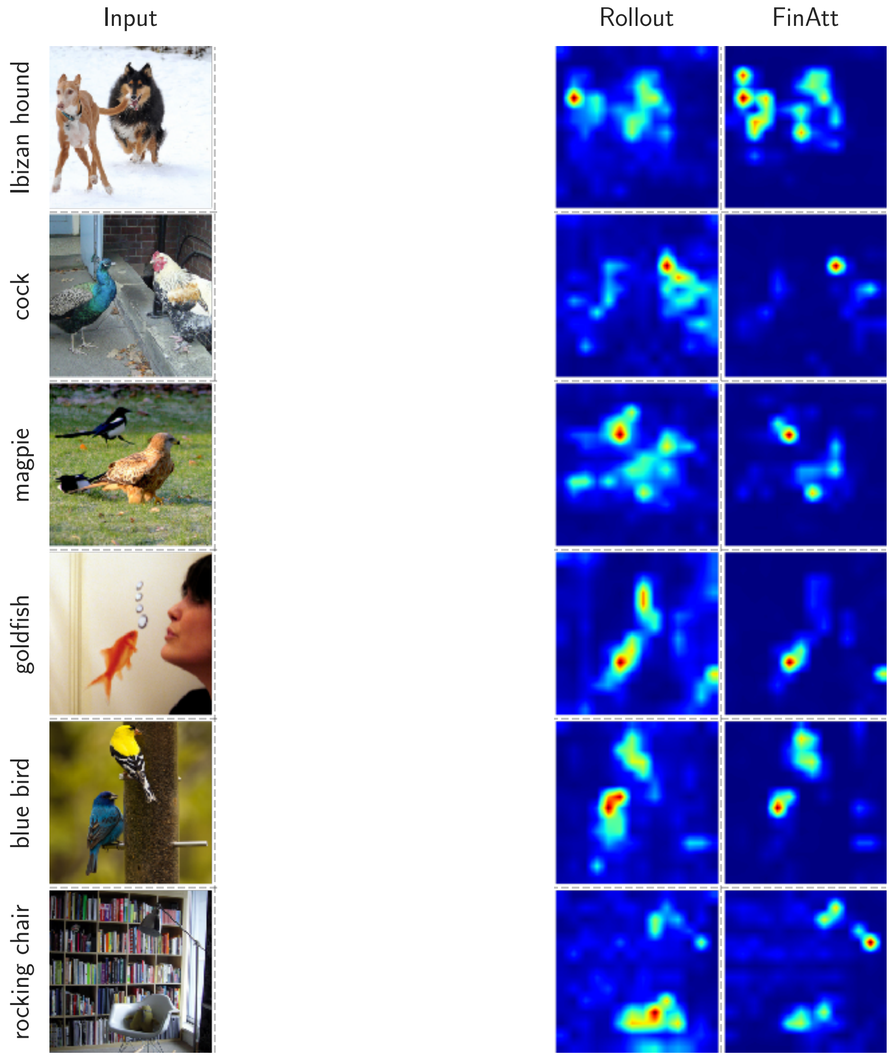


Attention Rollout: Abnar & Zuidema, ACL 2020

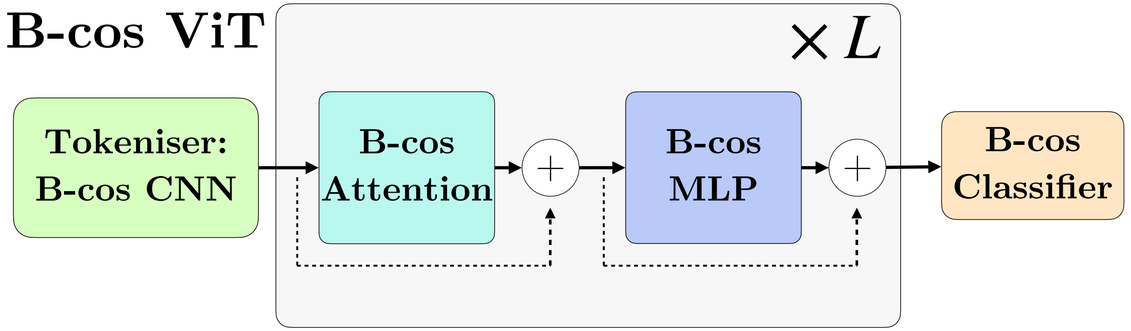
Attention is not All You Need (for XAI)

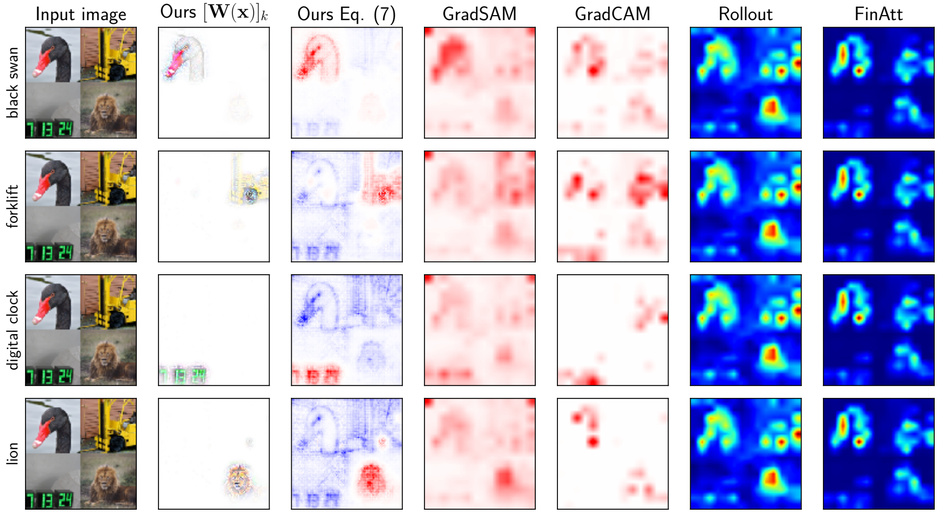


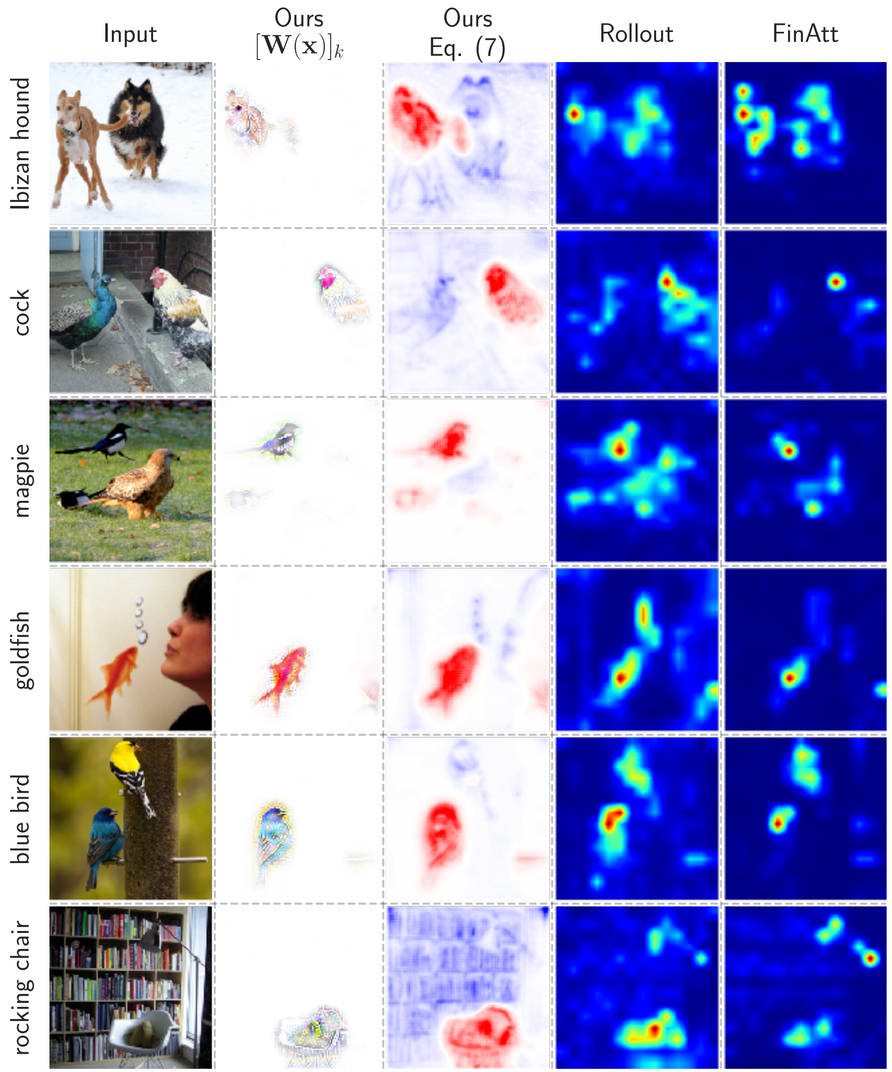




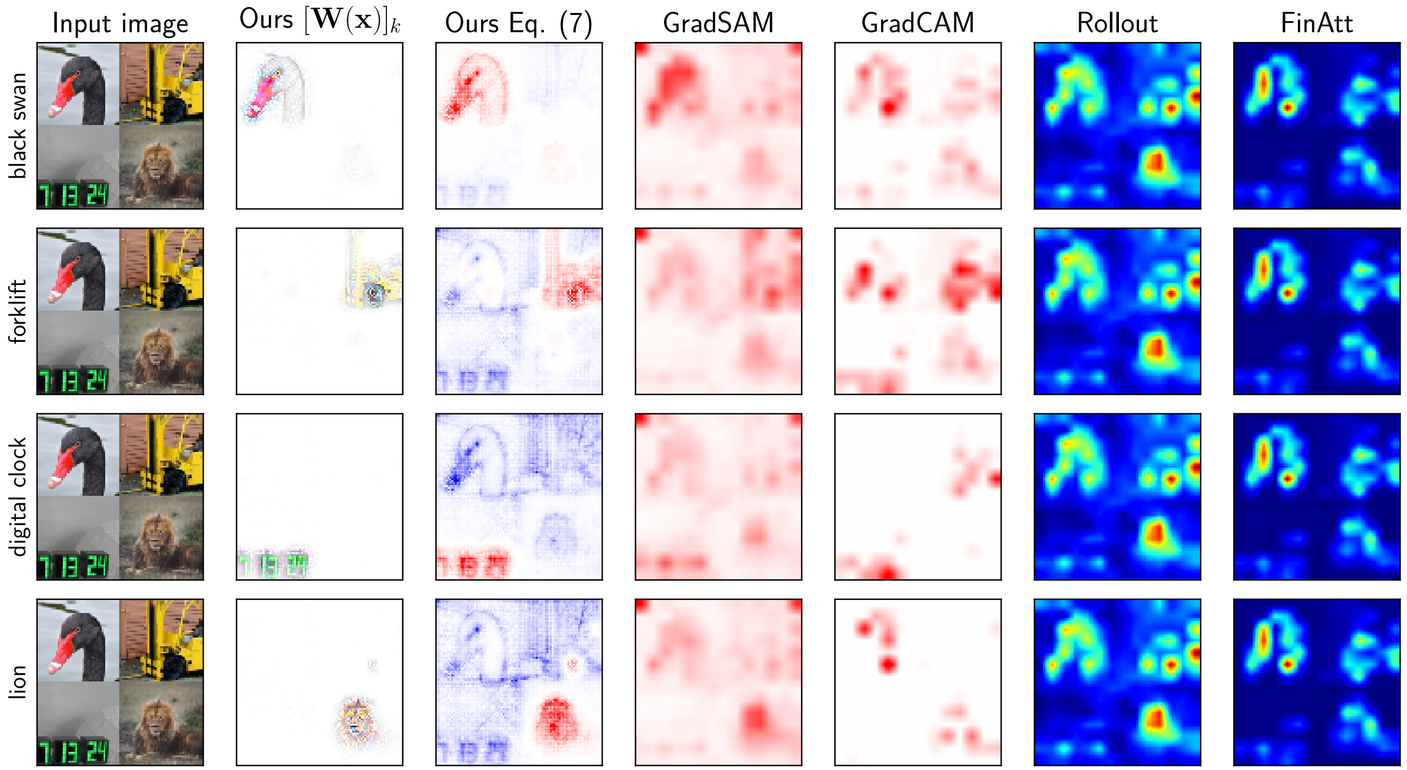
Attention is not All You Need (for XAI)

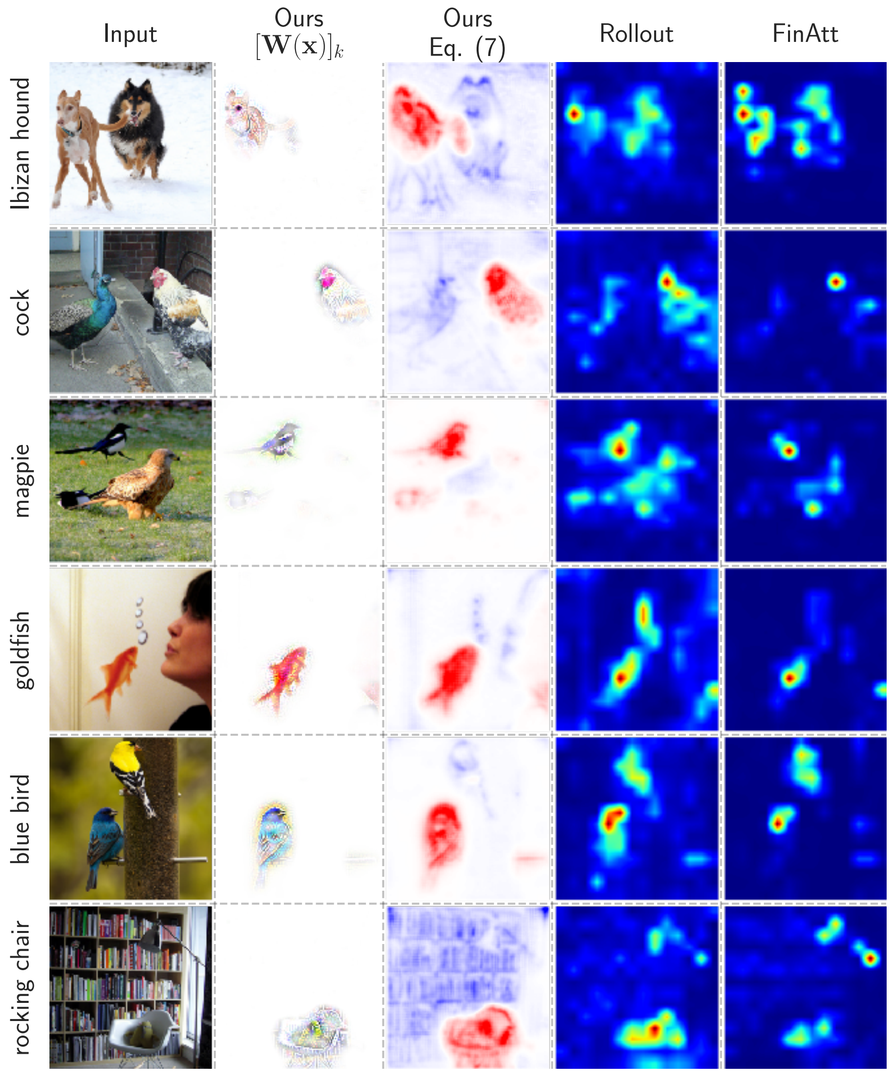




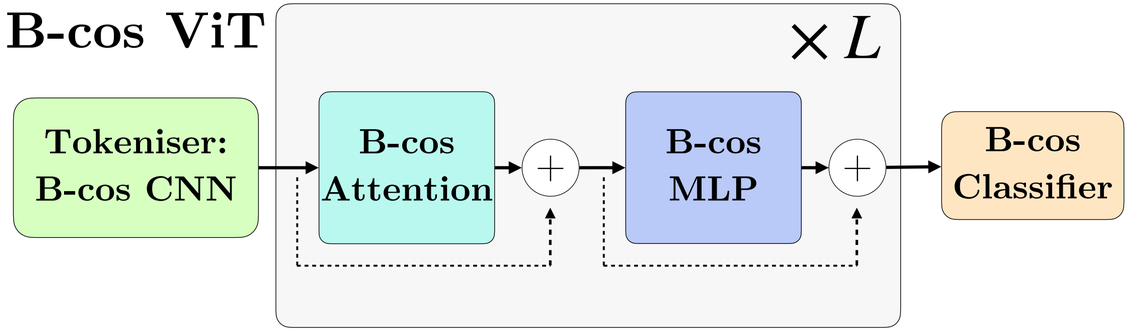


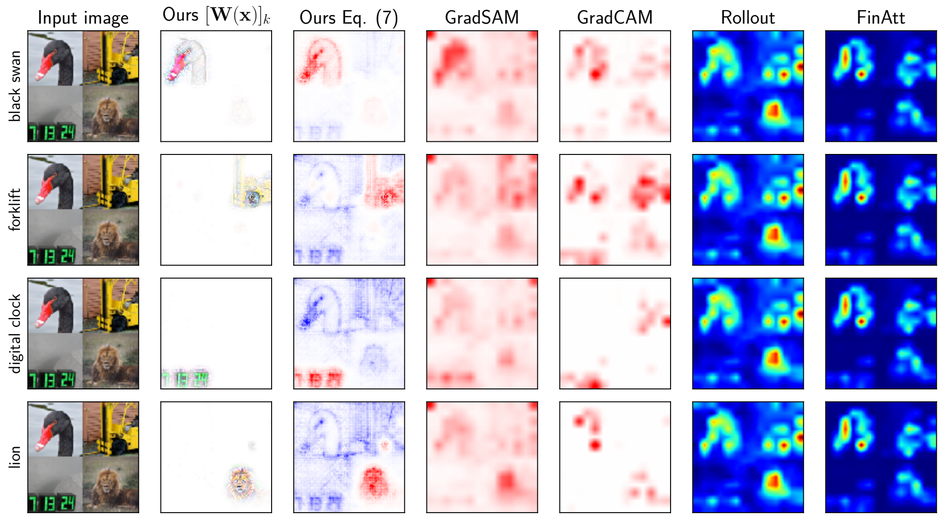
Attention is not All You Need (for XAI)

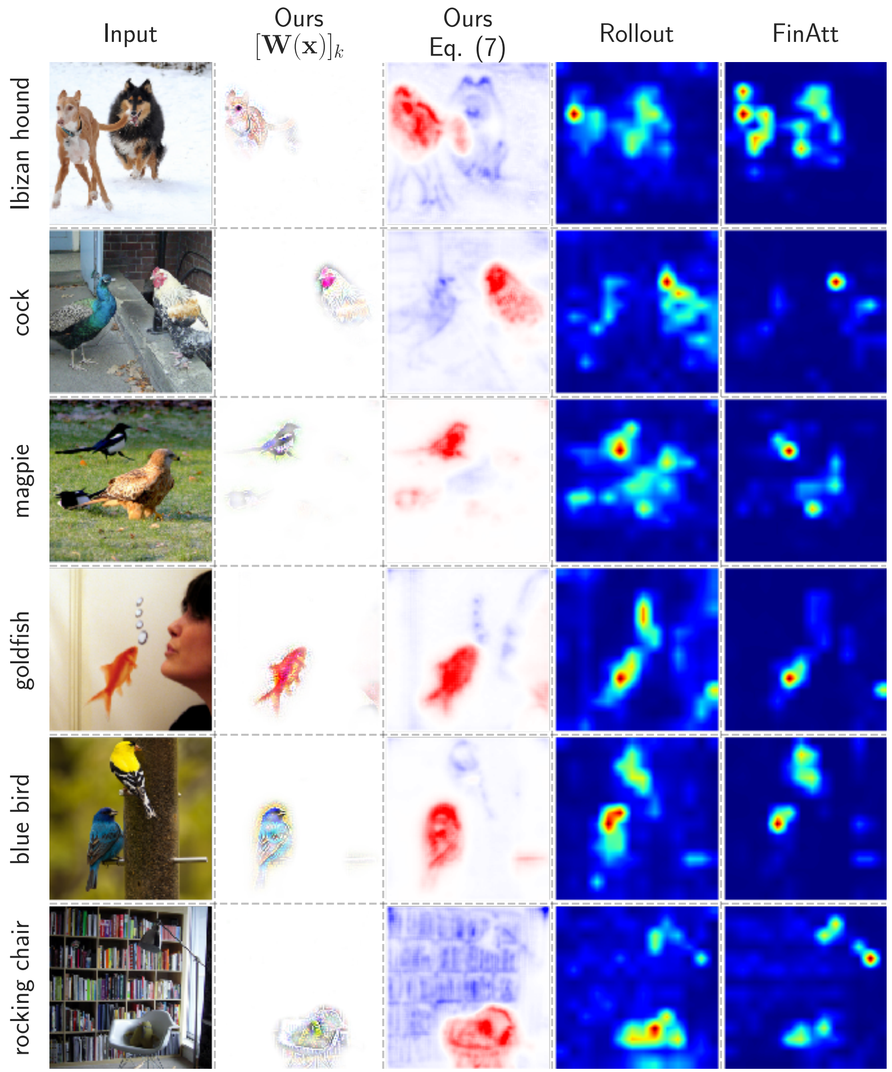




Attention is not All You Need (for XAI)

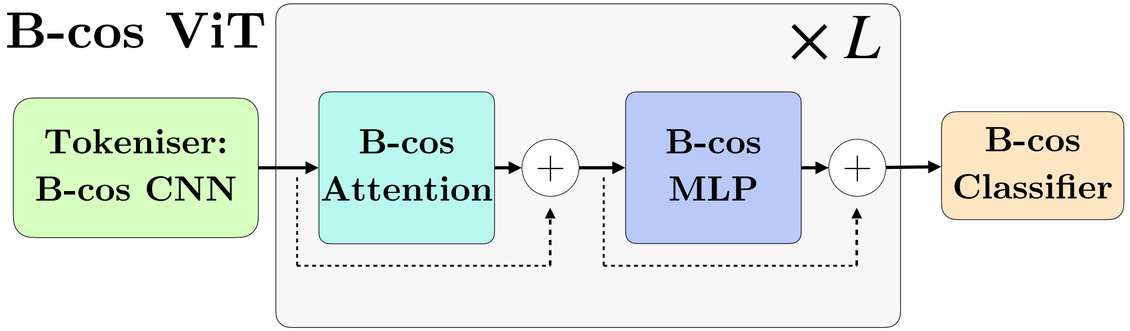




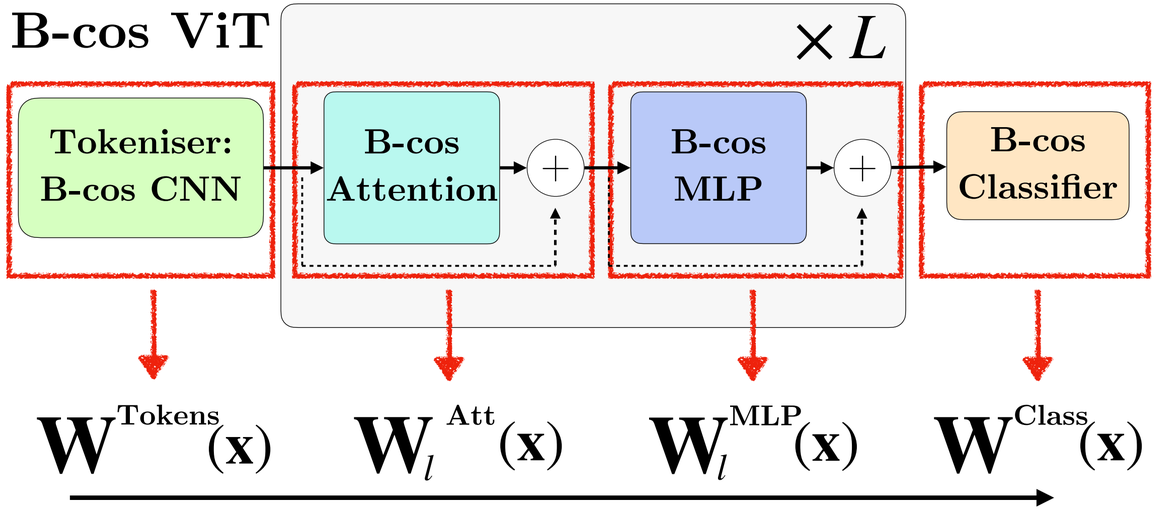


Attention Rollout: Abnar & Zuidema, ACL 2020; GradSAM: Barkan et al., CIKM 2021, GradCAM: Selvaraju et al., ICCV 2017

Attention is not All You Need (for XAI)



Attention is not All You Need (for XAI)

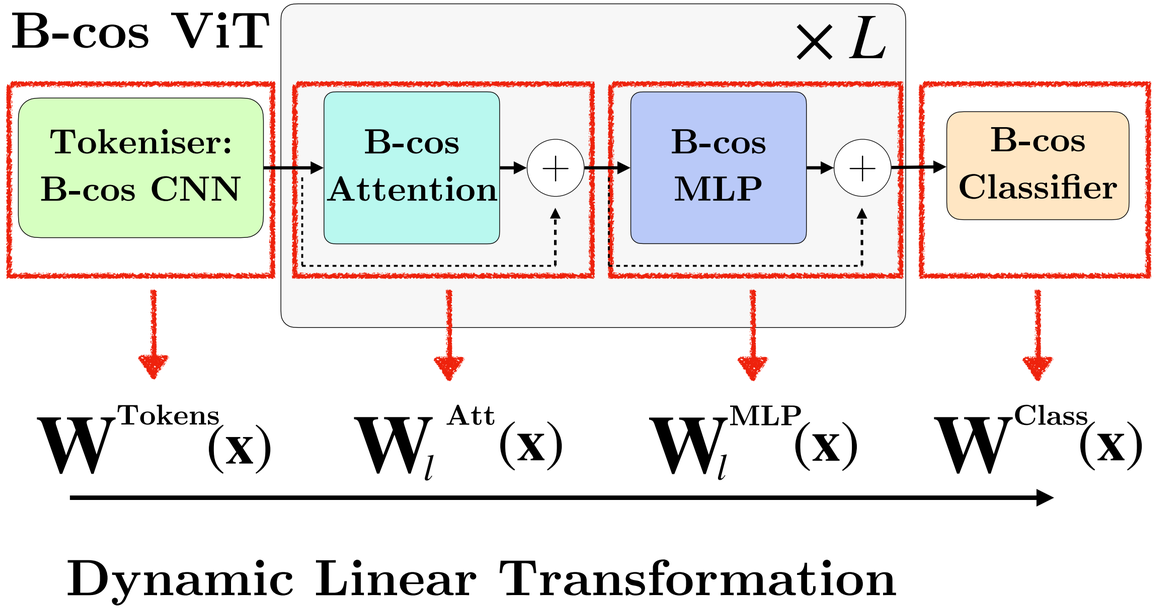


Dynamic Linear Transformation

$$\mathbf{W}(\mathbf{x}) = \mathbf{W}^{\text{Class}}(\mathbf{x}) \prod_{l=1}^{L} \left(\mathbf{W}_l^{\text{MLP}}(\mathbf{x}) \, \mathbf{W}_l^{\text{Att}}(\mathbf{x}) \right) \, \mathbf{W}^{\text{Tokens}}(\mathbf{x})$$

Attention is not All You Need (for XAI)

- Tokeniser + MLP + Classifier
  - interpret as CNNs,
    convert to B-cos CNNs

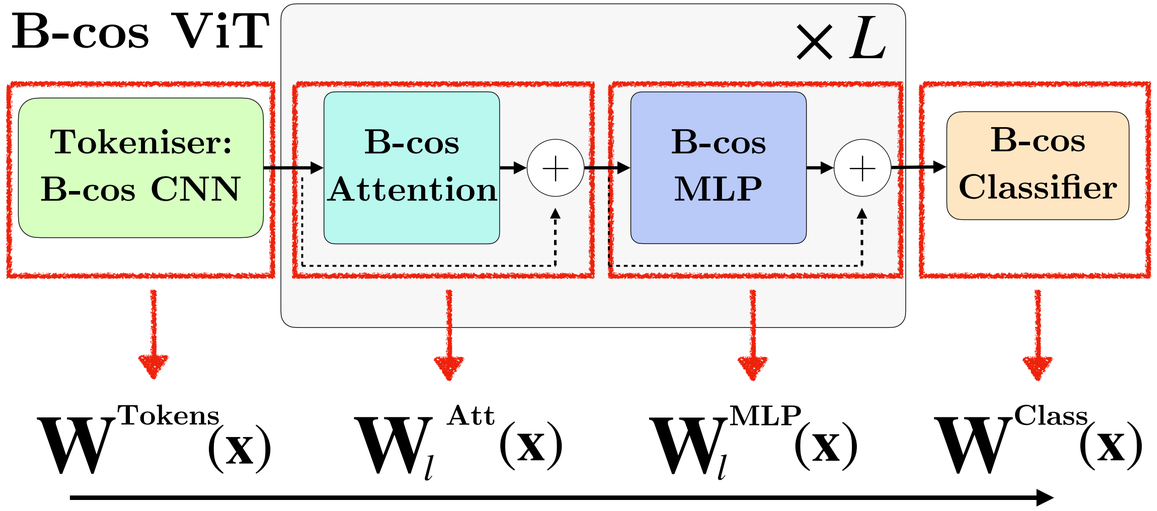


$$\mathbf{W}(\mathbf{x}) = \mathbf{W}^{\text{Class}}(\mathbf{x}) \prod_{l=1}^{L} \left(\mathbf{W}_l^{\text{MLP}}(\mathbf{x}) \, \mathbf{W}_l^{\text{Att}}(\mathbf{x}) \right) \, \mathbf{W}^{\text{Tokens}}(\mathbf{x})$$

Attention is not All You Need (for XAI)

- Tokeniser + MLP + Classifier
  - interpret as CNNs,
    convert to B-cos CNNs

- Self-Attention (SA) is dynamic linear
  - $$SA(\mathbf{X}) = \underbrace{\mathbf{A}(\mathbf{X})\,\mathbf{V}}_{\mathbf{W}(\mathbf{X})}\,\mathbf{X} = \mathbf{W}(\mathbf{X})\,\mathbf{X}$$

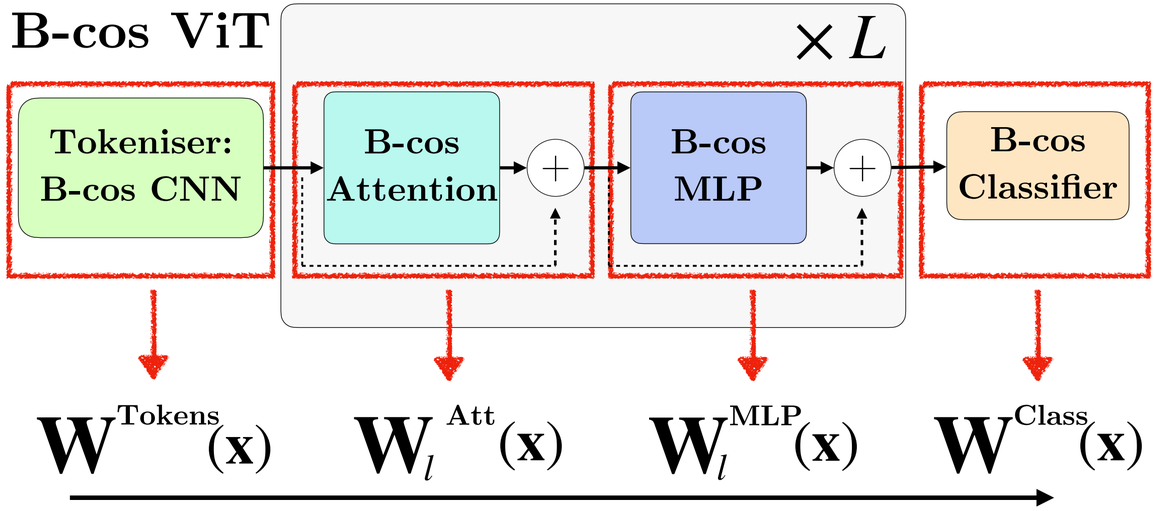


Dynamic Linear Transformation

$$\mathbf{W}(\mathbf{x}) = \mathbf{W}^{\text{Class}}(\mathbf{x}) \prod_{l=1}^{L} \left(\mathbf{W}_l^{\text{MLP}}(\mathbf{x})\, \mathbf{W}_l^{\text{Att}}(\mathbf{x}) \right) \mathbf{W}^{\text{Tokens}}(\mathbf{x})$$

Attention is not All You Need (for XAI)

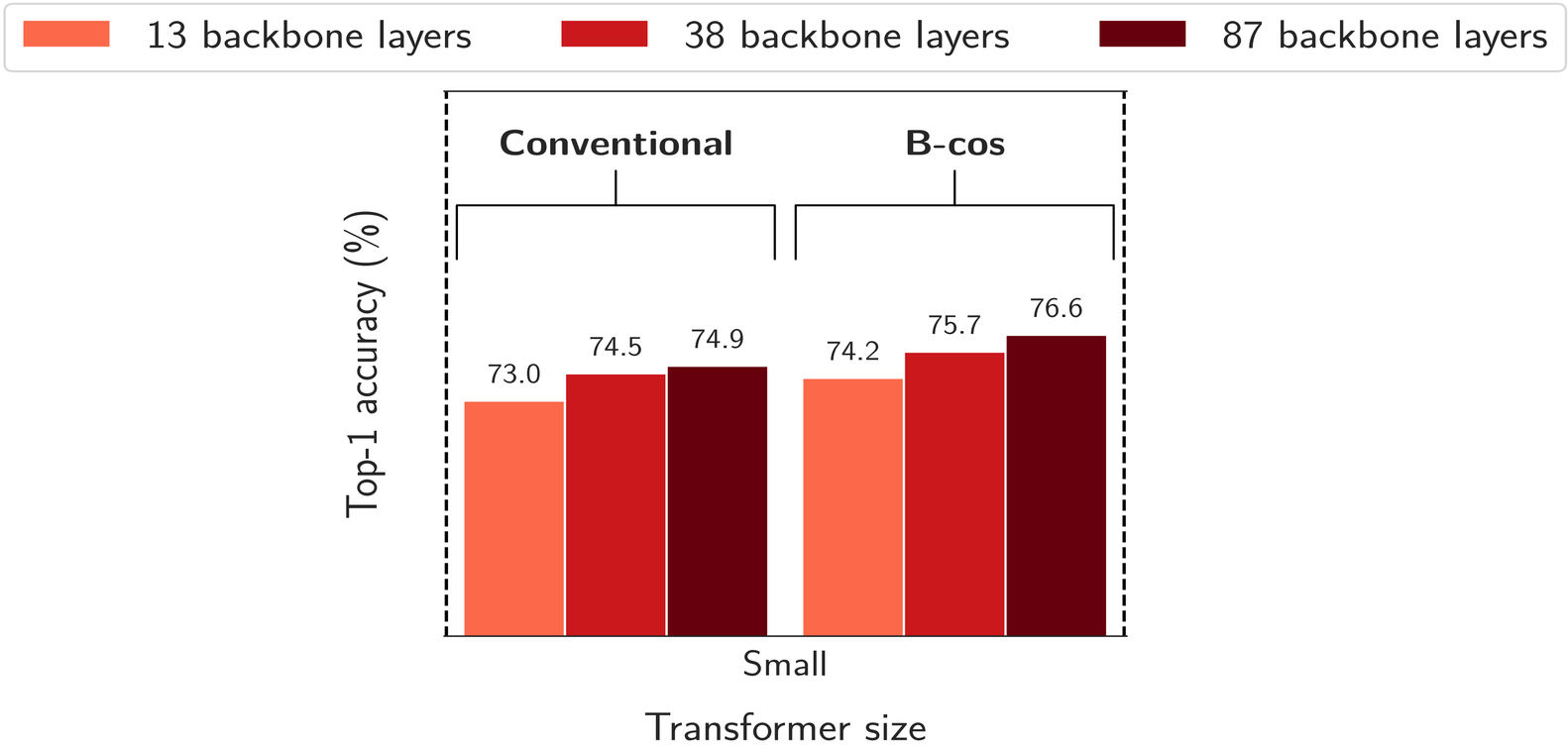
- Tokeniser + MLP + Classifier
  - interpret as CNNs,
    convert to B-cos CNNs
- Self-Attention (SA) is dynamic linear
  - $$SA(\mathbf{X}) = \underbrace{\mathbf{A}(\mathbf{X})\,\mathbf{V}}_{\mathbf{W}(\mathbf{X})}\,\mathbf{X} = \mathbf{W}(\mathbf{X})\,\mathbf{X}$$
- Side note:
  - for Tokenisation, use L layers of pretrained+frozen B-cos DenseNet-121



$$\mathbf{W}(\mathbf{x}) = \mathbf{W}^{\text{Class}}(\mathbf{x}) \prod_{l=1}^{L} \left(\mathbf{W}_l^{\text{MLP}}(\mathbf{x})\, \mathbf{W}_l^{\text{Att}}(\mathbf{x}) \right) \mathbf{W}^{\text{Tokens}}(\mathbf{x})$$

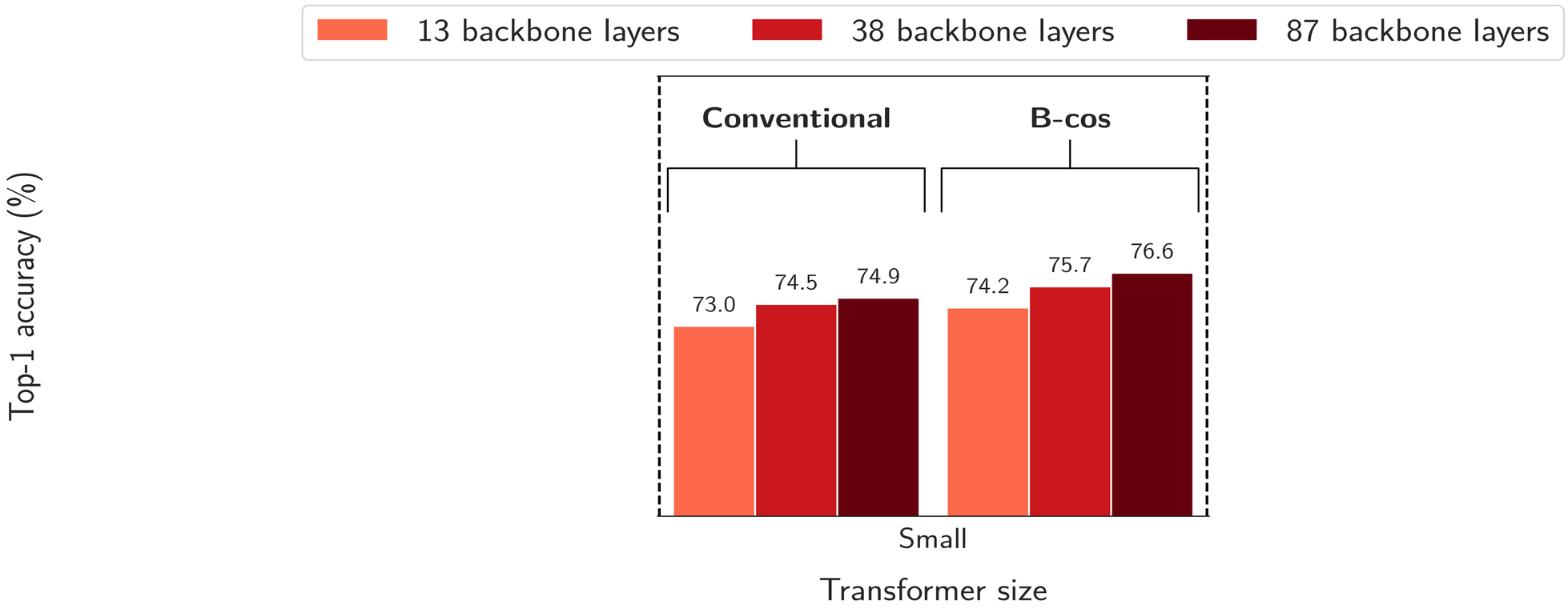
Attention is not All You Need (for XAI)

Results — classification accuracy

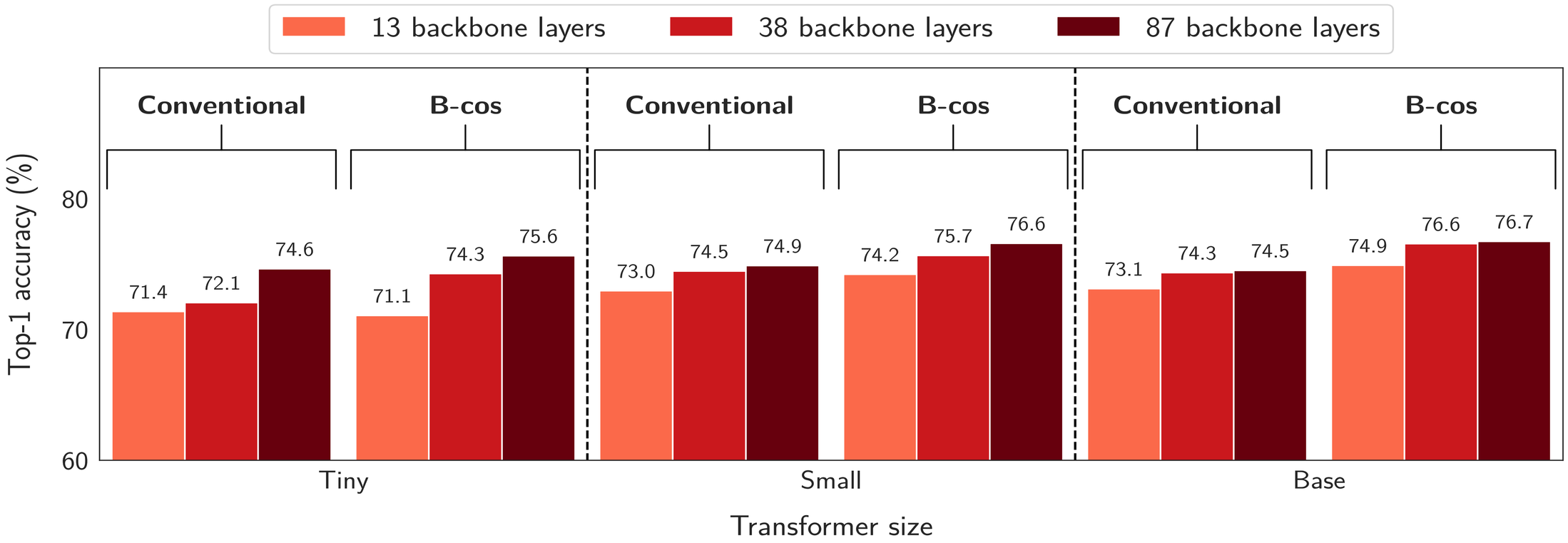


Attention is not All You Need (for XAI)

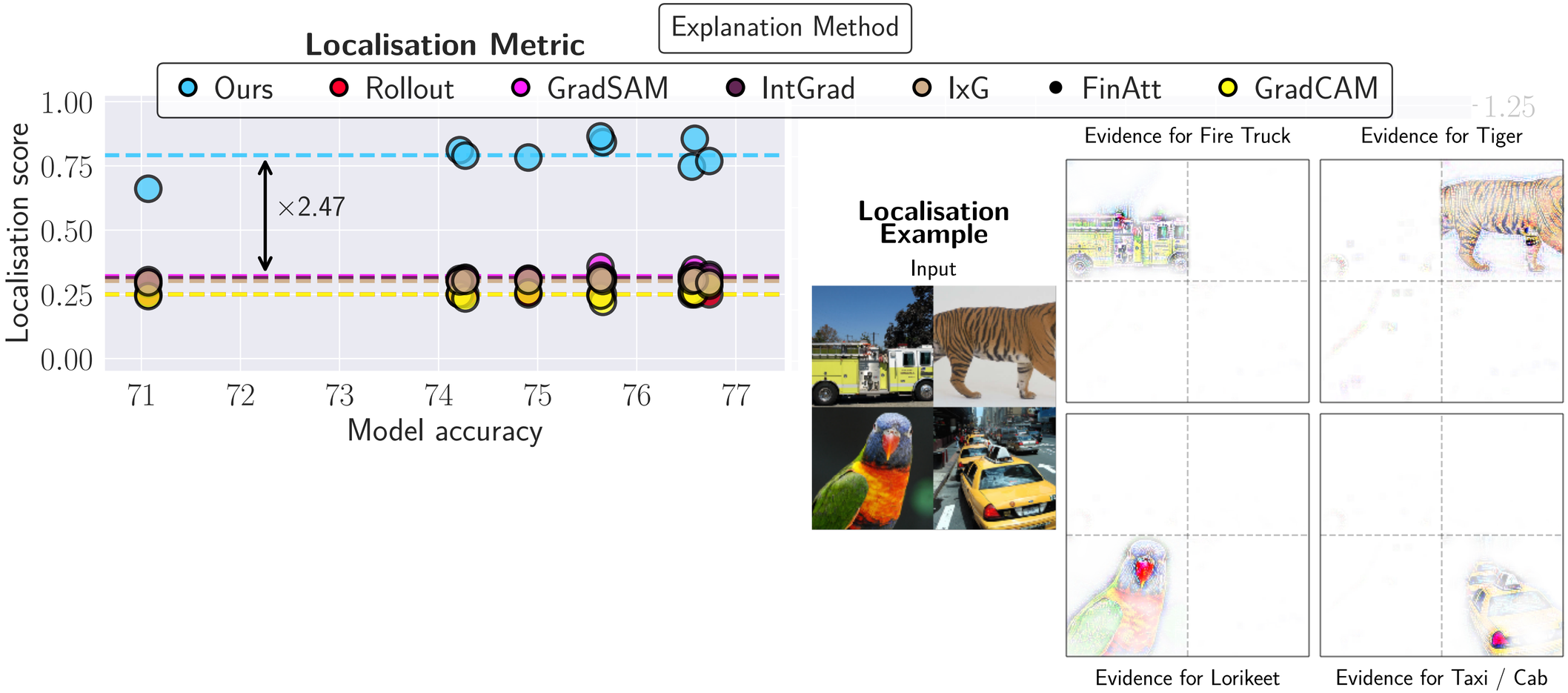
Results — classification accuracy



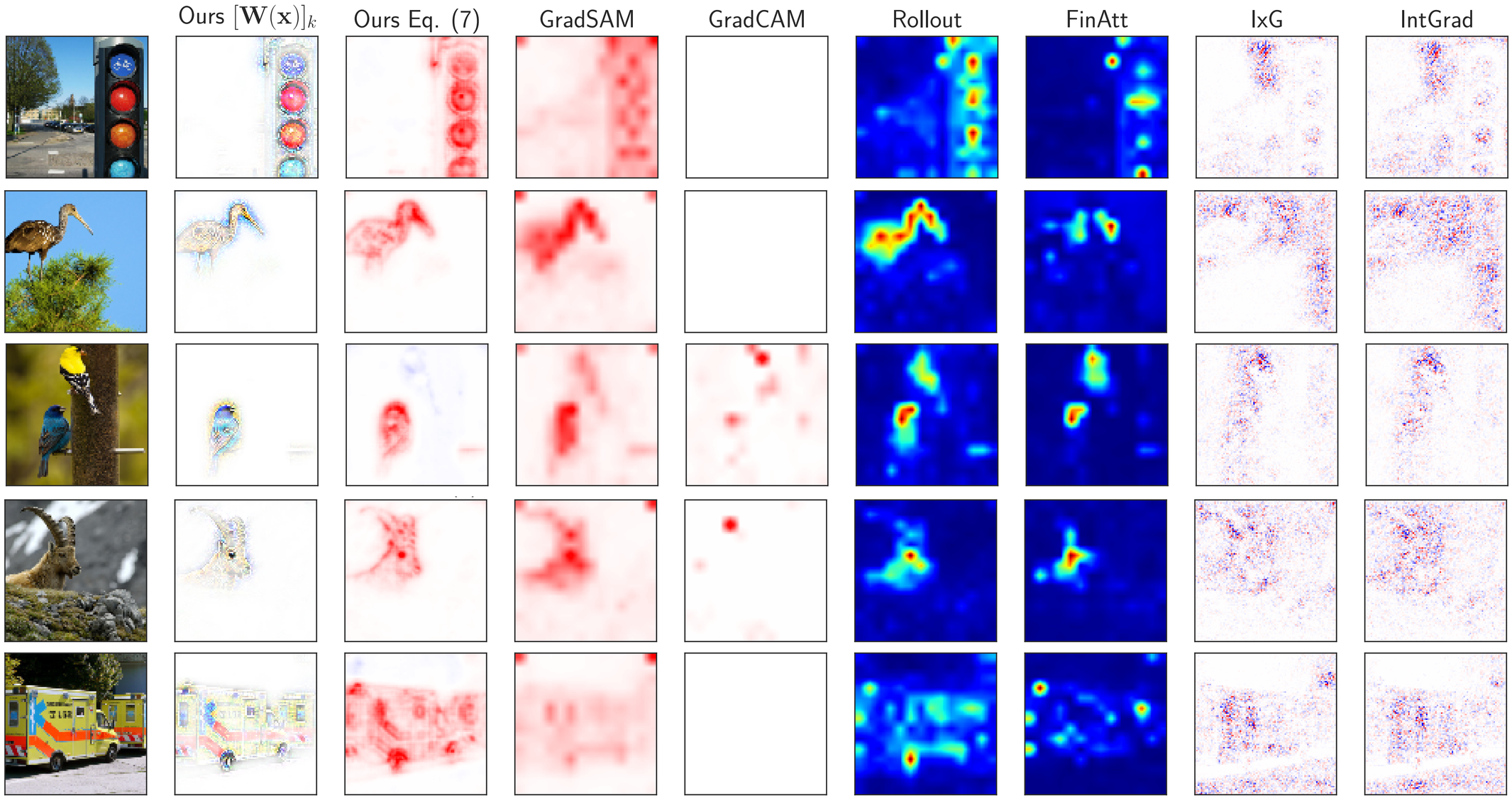
Results — classification accuracy



Results — interpretability metrics (grid pointing game)



Qualitative Results



Attention is not All You Need (for XAI) - Summary

- B-cos framework generally compatible with ViTs
  - Attention already dynamic linear $\mathrm{SA}(\mathbf{X}) = \mathbf{W}(\mathbf{X})\mathbf{X}$
  - remaining modules → B-cos CNNs

Attention is not All You Need (for XAI) - Summary

- B-cos framework generally compatible with ViTs
  - Attention already dynamic linear $\mathrm{SA}(\mathbf{X}) = \mathbf{W}(\mathbf{X})\mathbf{X}$
  - remaining modules → B-cos CNNs
- B-cos ViTs can be highly performant
  - similar results as with standard ViTs in comparable setting

Attention is not All You Need (for XAI) - Summary

- B-cos framework generally compatible with ViTs
  - Attention already dynamic linear $\mathrm{SA}(\mathbf{X}) = \mathbf{W}(\mathbf{X})\mathbf{X}$
  - remaining modules → B-cos CNNs
- B-cos ViTs can be highly performant
  - similar results as with standard ViTs in comparable setting
- B-cos ViTs highly interpretable
  - similar interpretability as B-cos CNNs

Overview of Today's Lecture

- Inherently Interpretable Neural Network
  - B-cos CNN [cvpr'22] - https://arxiv.org/abs/2205.10268
  - B-cos CNN & Vision transformers [pami'24] - https://arxiv.org/abs/2306.10898

- Segment Anything Model (SAM)
  - [iccv'23] - https://arxiv.org/abs/2304.02643
  - https://segment-anything.com/

Segment Anything

slides credit:

Alexander Kirillov @ Meta AI

slide credit: Alexander Kirillov

Research Team

Alexander Kirillov

Eric Mintun

Nikhila Ravi

Hanzi Mao

Chloe Rolland

Laura Gustafson

Tete Xiao

Spencer Whitehead

Alex Berg

Wan-Yen Lo

Piotr Dollar

Ross Girshick

Project Contributors (alphabetical):

Aaron Adcock, Vaibhav Aggarwal, Morteza Behrooz, Cheng-Yang Fu, Ashley Gabriel, Ahuva Goldstand, Allen Goodman, Sumanth Gurram, Jiabo Hu, Somya Jain, Devansh Kukreja, Robert Kuo, Joshua Lane, Yanghao Li, Lilian Luong, Jitendra Malik, Mallika Malhotra, William Ngan, Omkar Parkhi, Nikhil Raina, Dirk Rowe, Neil Seejoor, Vanessa Stark, Bala Varadarajan, Bram Wasti, Zachary Winstrom

slide credit: Alexander Kirillov

Deep learning paradigm shift

[2012 – 2022]

- **Apply DL to task X**
- End-to-end
- Step change on **specific tasks** and **specific data** distributions
- Adaptation via **fine-tuning only**

slide credit: Alexander Kirillov

Deep learning paradigm shift

[2012 – 2022] → [2022 – present]

- **Apply DL to task X**
- End-to-end
- Step change on **specific tasks** and **specific data** distributions
- Adaptation via **fine-tuning only**

- **Combine foundation/base models for X, Y, Z to do tasks Q, R, S**
- Powerful **components** (LLMs, CLIP …)
- Step change on **broad tasks** and **broad data** distributions
- Adaptation without parameter updates (**prompting**)

slide credit: Alexander Kirillov

Broadly applicable solution for segmentation

- The Segment Anything project / Segment Anything Model (SAM)
  - A component (SAM) that broadly solves segmentation
  - Useful to both humans and machines (i.e., a component in a system)
  - Broad generalization to new domains and use cases / tasks
  - Adaptation primarily via prompting

slide credit: Alexander Kirillov

Segment Anything Project Goal

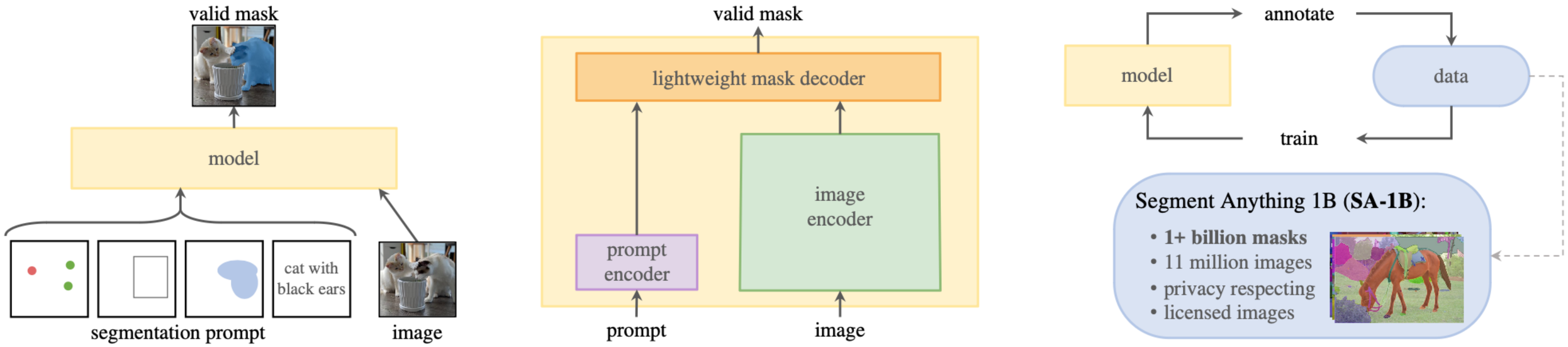
- Develop a **promptable model** and pre-train it on a broad **dataset** using a **task** that enables powerful generalization
  - Model: *Doesn't exist!*
  - Dataset: *Doesn't exist!*
  - Task: *Doesn't exist!*

- Model/data/task are highly coupled

- We need a comprehensive solution

slide credit: Alexander Kirillov

Segment Anything: Task, Model, Data



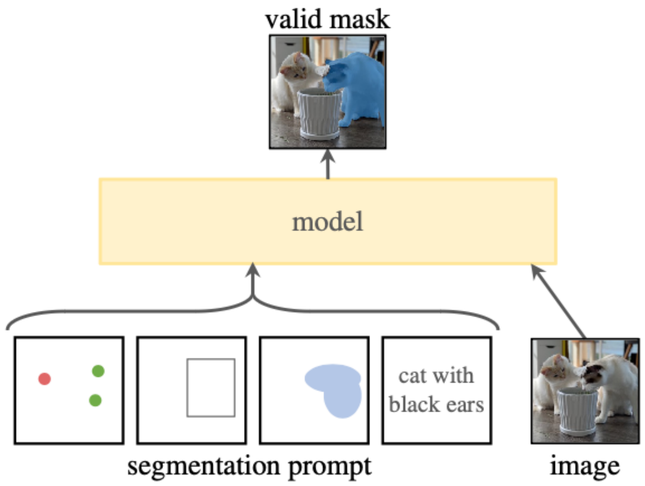
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

slide credit: Alexander Kirillov

Task: Promptable Segmentation

- Translate “prompt” from NLP to segmentation
  - Segmentation prompt: anything that specifies *what* to segment
  - E.g., foreground/background points, boxes, masks, text

- The task: Return a *valid* segmentation mask given *any prompt*



(a) **Task**: promptable segmentation

slide credit: Alexander Kirillov

Example: AR gaze-to-mask

slide credit: Alexander Kirillov

Example: AR gaze-to-mask

slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

Why Promptable Segmentation?

- Composable with other models.

- Prompts easy to simulate at training.

- Enables human-in-loop annotation.

fast annotation

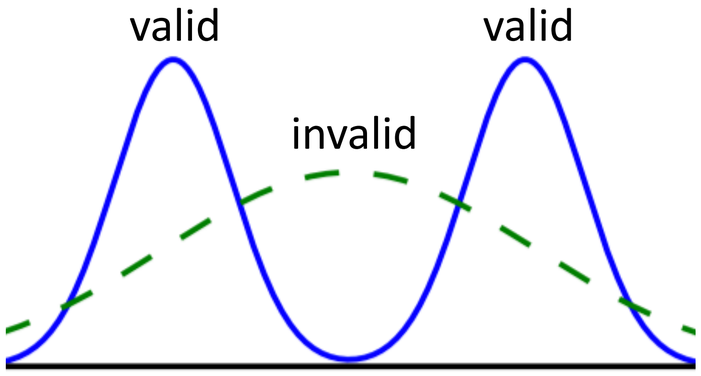
1 click 2 clicks

slide credit: Alexander Kirillov

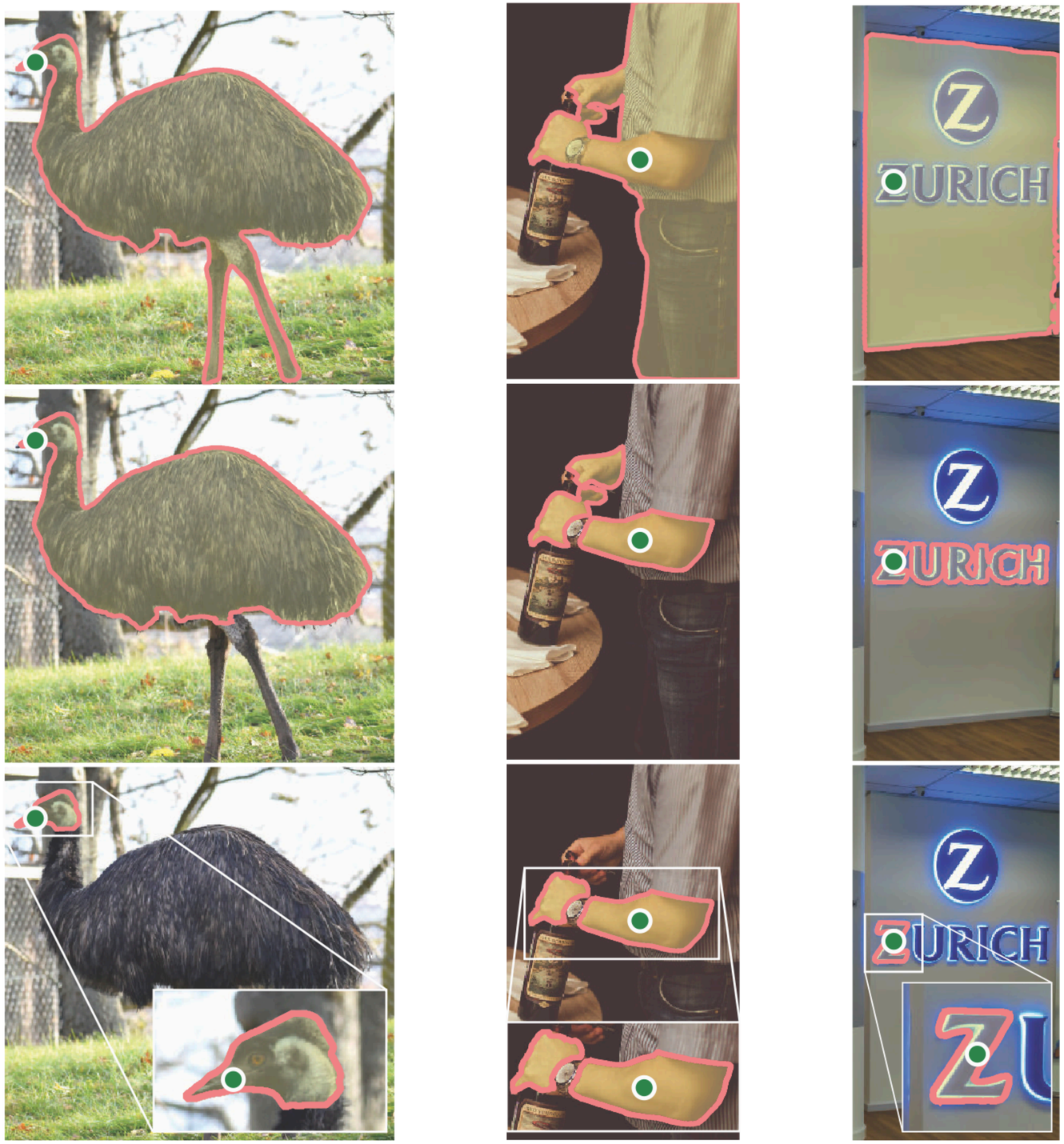
A “Valid” Mask?

- Prompts can be *ambiguous*
- Segmentation is *ambiguous*
- The task should train a model that handles ambiguity

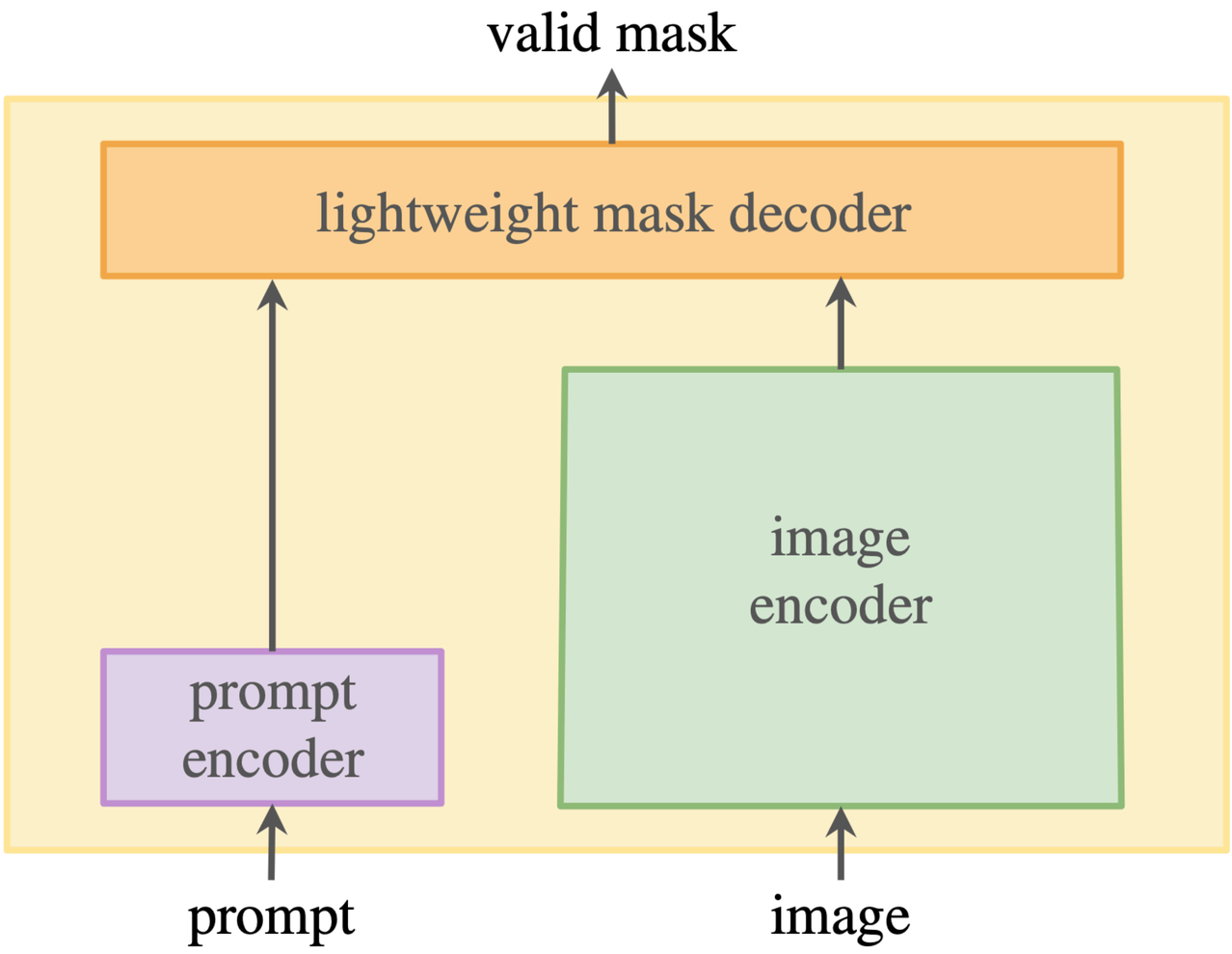
- Avoid “mode averaging”



Adapted from NIPS 2016 Tutorial: “Generative Adversarial Networks" by Ian Goodfellow

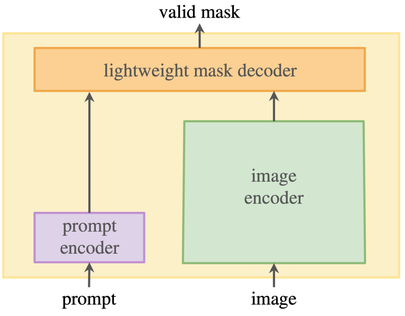


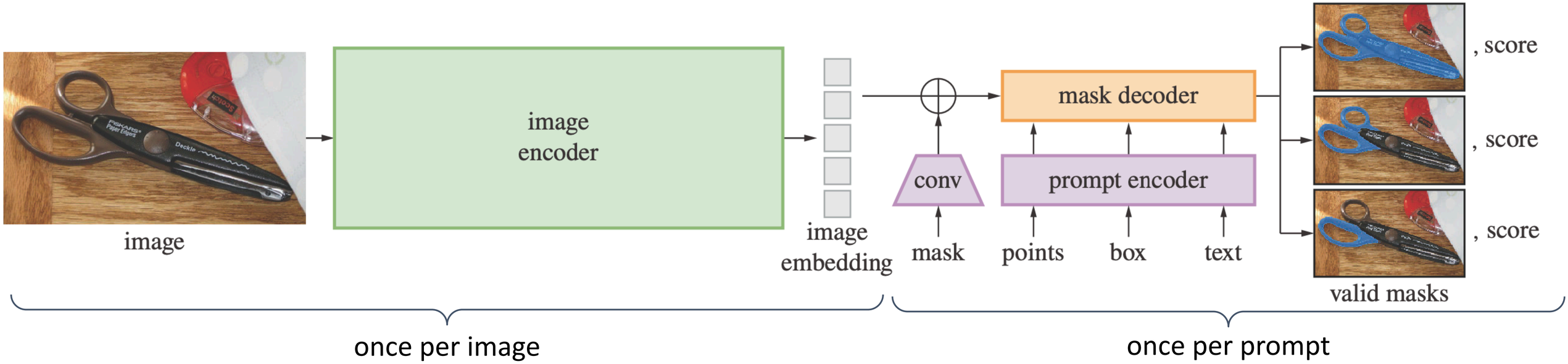
SAM: Model Overview



slide credit: Alexander Kirillov

SAM in More Detail

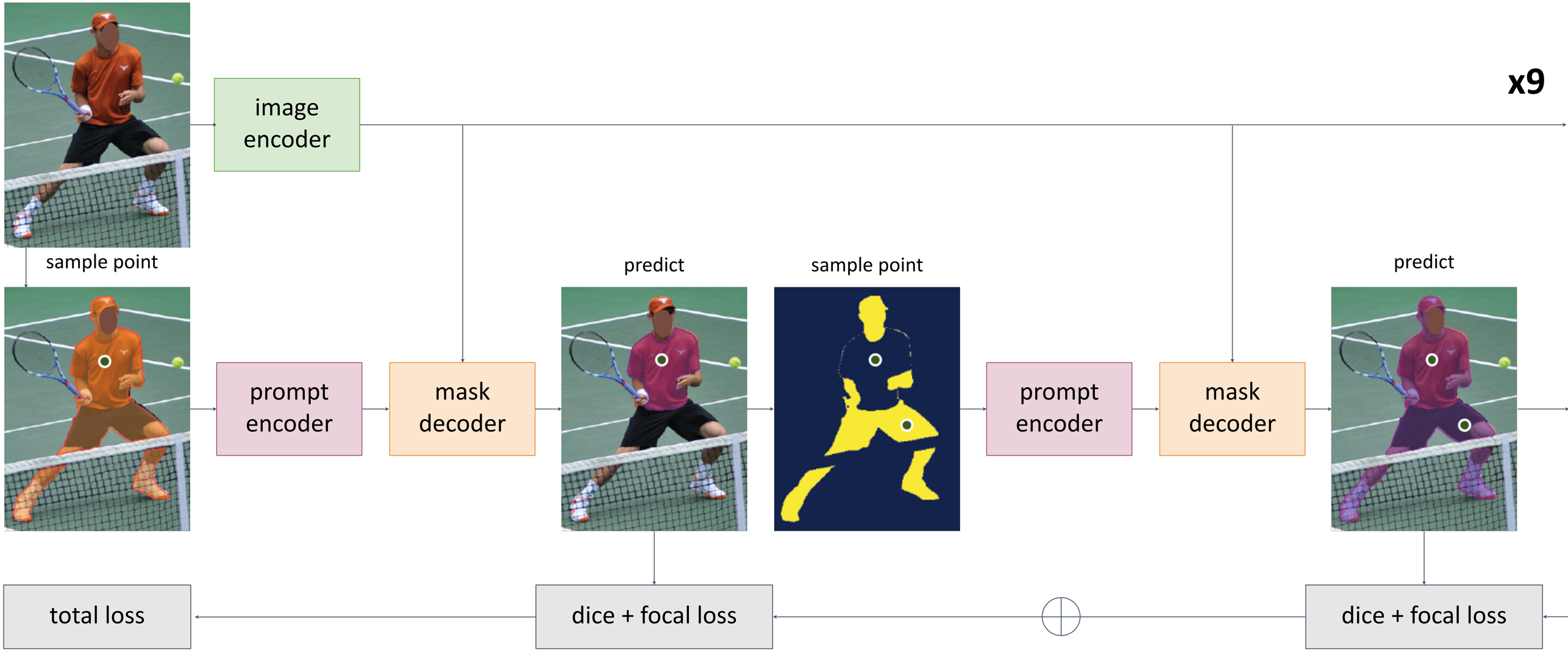




- **Image encoder:** ViT-H adapted for 1024x1024 input resolution
- **Prompt encoder:** Positional and label embeddings for points/boxes; CLIP for text
- **Mask decoder:** Lightweight transformer (~50ms on CPU)
- **Output:** 1 or 3 masks + model's prediction of mask quality

slide credit: Alexander Kirillov

SAM training



slide credit: Alexander Kirillov

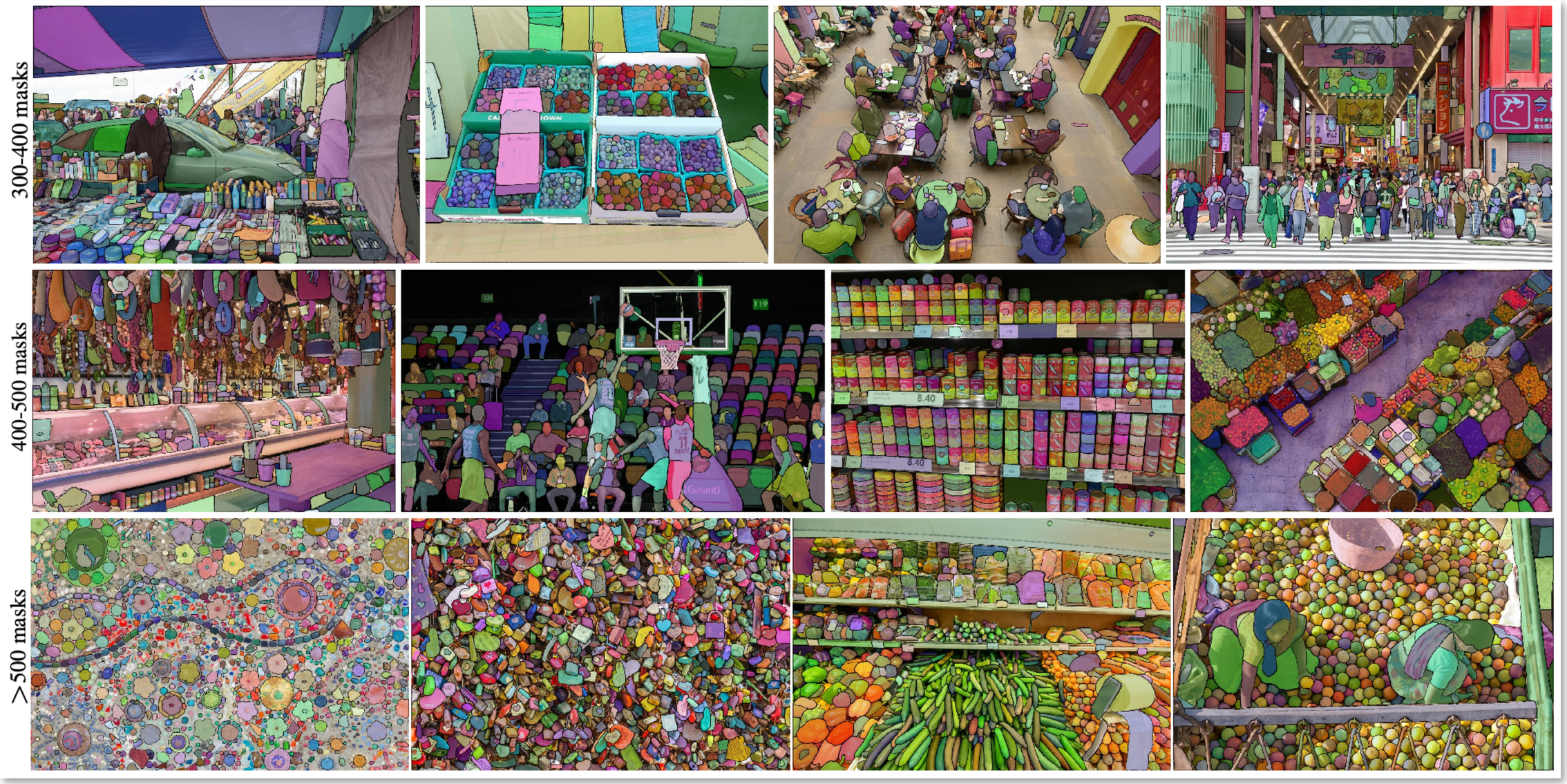
Transformer details

- Image encoder is ViT-H modified for large resolution (ViTDet):
  - initialized with MAE weights
  - Patch size 16
  - Size 14 windowed attention in all but three layers
  - Final 1x1 conv mapping ViT-H channel dim. (1280) to mask decoder dim. (256)
  - For 1024x1024 input image, output embeddings is 256x64x64
- Mask decoder
  - Large internal MLP dimension of 2048
  - Down-sampled q/k/v dimension of 128 in cross attention
  - Entire prompt token added as positional encoding to q/k in attention every layer (DETR style)
  - After transformer, embeddings upscaled to 256x256 by 2 conv transposes of dim 64 and 32

slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

Dataset Construction

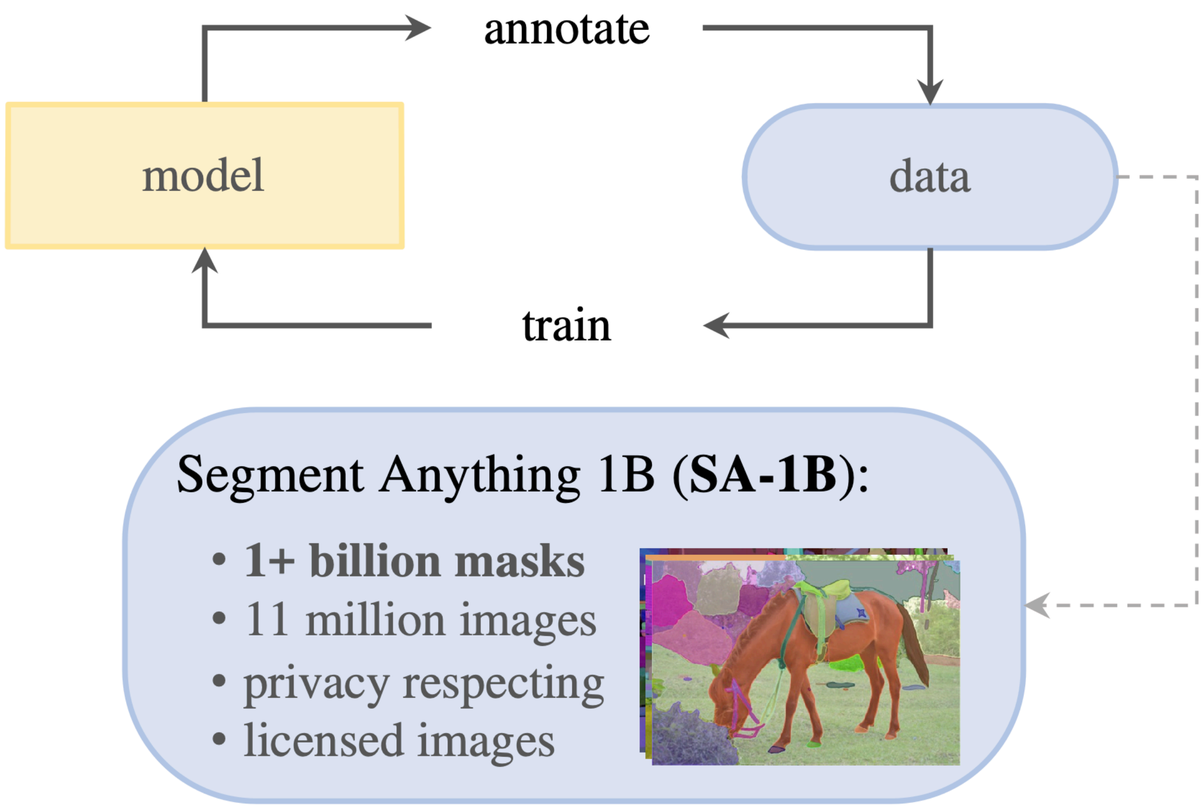


Data Engine

- Model-in-the-loop dataset construction

- Requires SAM to run in (amortized) real-time on CPU in a browser

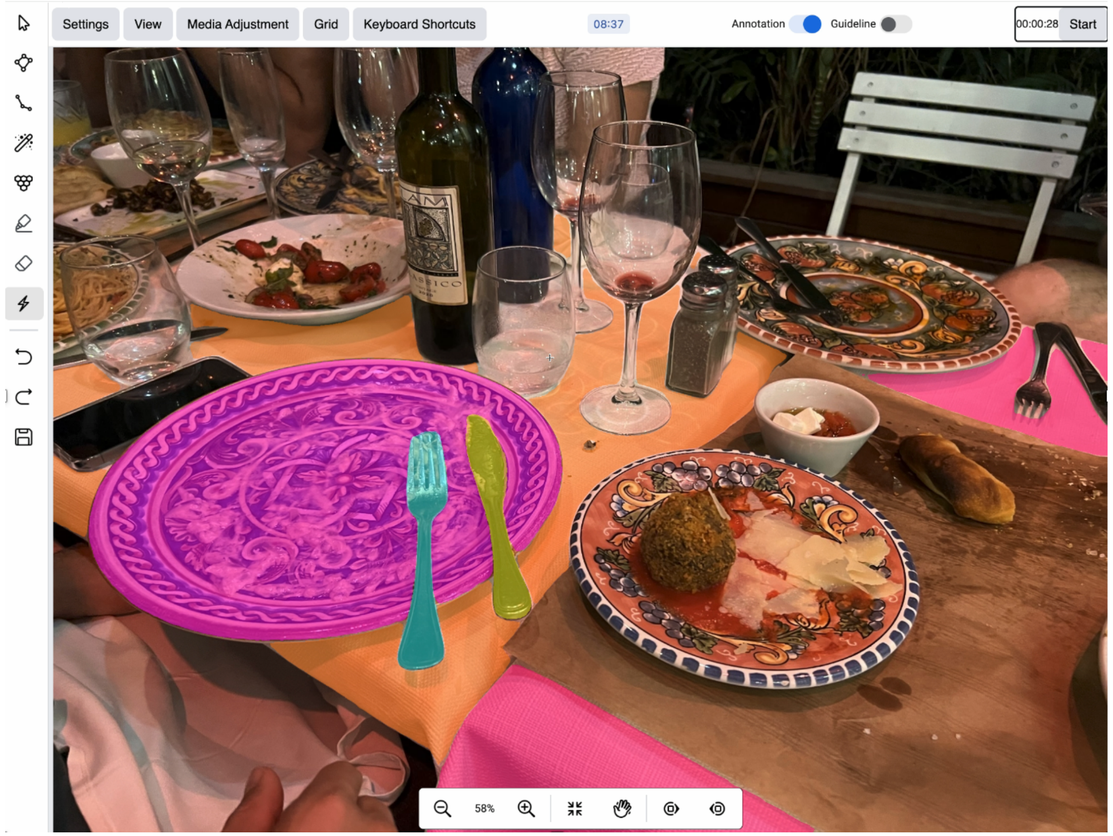
- Three stages
  - *Assisted manual,*
  - *semi-manual,*
  - *fully automatic*



slide credit: Alexander Kirillov

Data Engine: Assisted Manual Stage

- SAM in a web-browser
  - Interactive segmentation tool
- 4.3 million masks (120k images)
- 6.5x faster than COCO masks
  - 2x slower than box annotation
- Retrained SAM 6 times
  - Scaled model size
  - Other improvements



slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

Data Engine: Semi-Manual Stage

- Goal: increase diversity beyond most prominent objects
- Automatically segment many objects using detector + SAM
- Added 5.9 million more masks (180k more images)
- Retrained SAM 5 more times

Automatic masks + Manually annotated on top = Automatic + Manual

Data Engine: Fully Automatic Stage

- We ran SAM in “segment everything” mode on 11 million images
- Generated 1.1 billion masks

slide credit: Alexander Kirillov

SA-1B Dataset

- 11 million images
- 1.1 **billion** masks
- All automatically generated by SAM
- 400x larger than previous largest segmentation dataset
- Quality evaluation
  - Estimated 94% have at least 90% IoU with professional masks
  - Prior work estimates human inter-annotator IoU at 85-91%

slide credit: Alexander Kirillov

SA-1B Examples (Automatically Labelled)

slide credit: Alexander Kirillov

SA-1B Examples (Automatically Labelled)

slide credit: Alexander Kirillov

SA-1B Examples (Automatically Labelled)

300-400 masks

400-500 masks

slide credit: Alexander Kirillov

>500 masks

slide credit: Alexander Kirillov

Zero-Shot Transfer Tasks

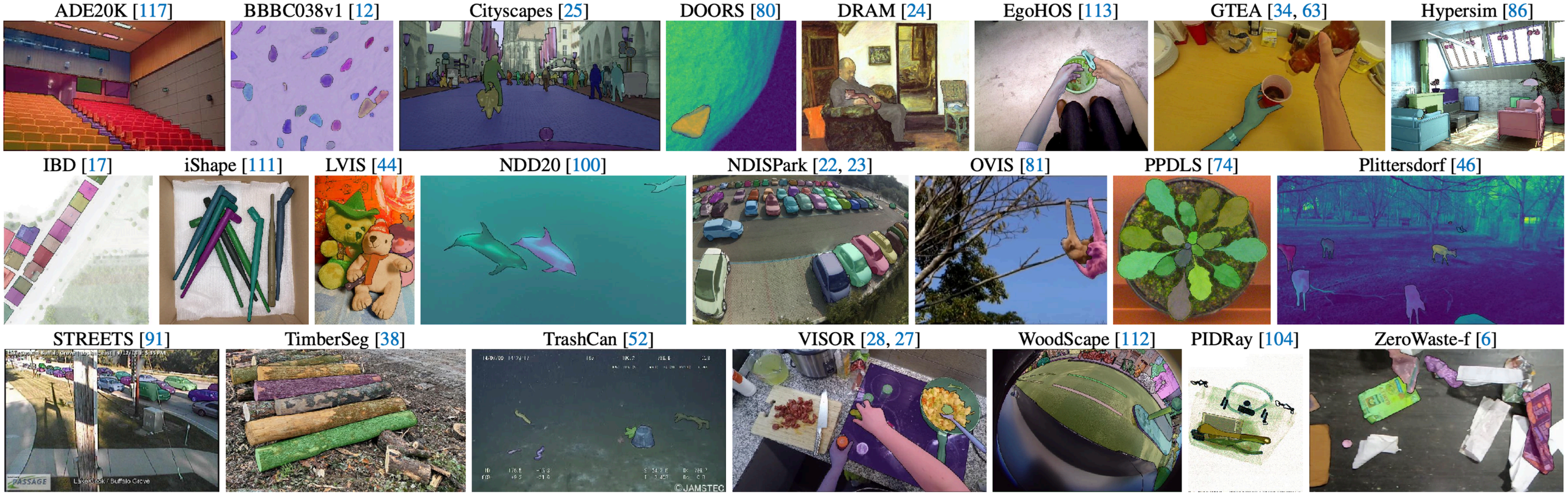
1. Single point valid mask evaluation
2. Edge detection
3. (Object proposals)
4. (Instance segmentation)
5. Text-to-mask

slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

Single Point Valid Mask Evaluation

- SA-1B: mostly “scene” images (photographers walking in the world)
- 23 diverse dataset suite – mostly out-of-distribution vis-à-vis SA-1B



Single Point Valid Mask Evaluation Protocol

slide credit: Alexander Kirillov

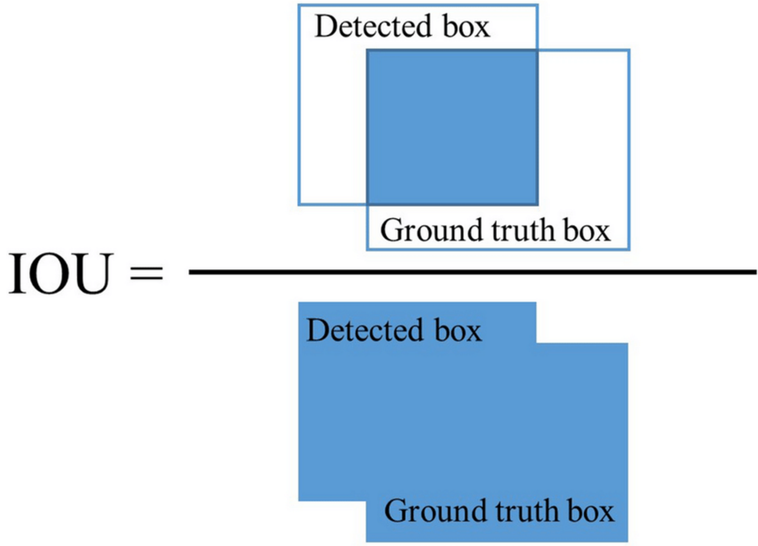
Single Point Valid Mask Evaluation Protocol

- Select g.t. mask from dataset
  - Sample point from mask
  - Prompt model with point, model predicts mask

slide credit: Alexander Kirillov

Single Point Valid Mask Evaluation Protocol

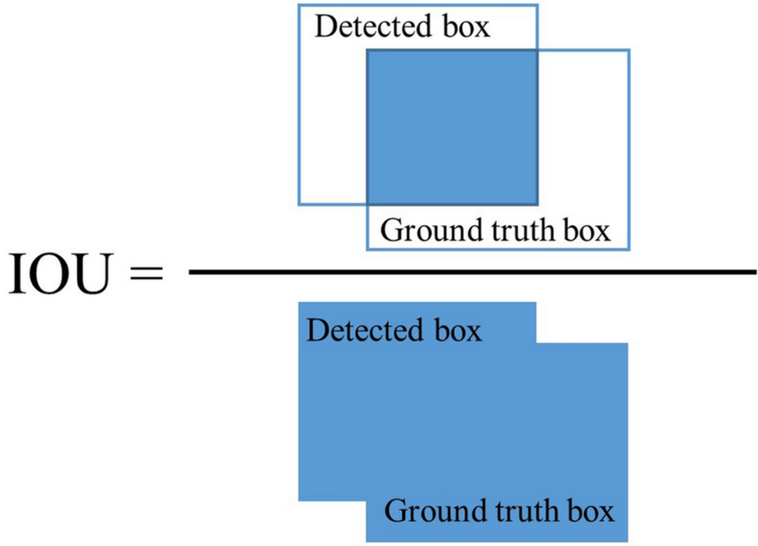
- Select g.t. mask from dataset
  - Sample point from mask
  - Prompt model with point, model predicts mask
- Compute IoU (prediction, g.t.)
  - Metric: mean IoU (mIoU) over mask



slide credit: Alexander Kirillov

Single Point Valid Mask Evaluation Protocol

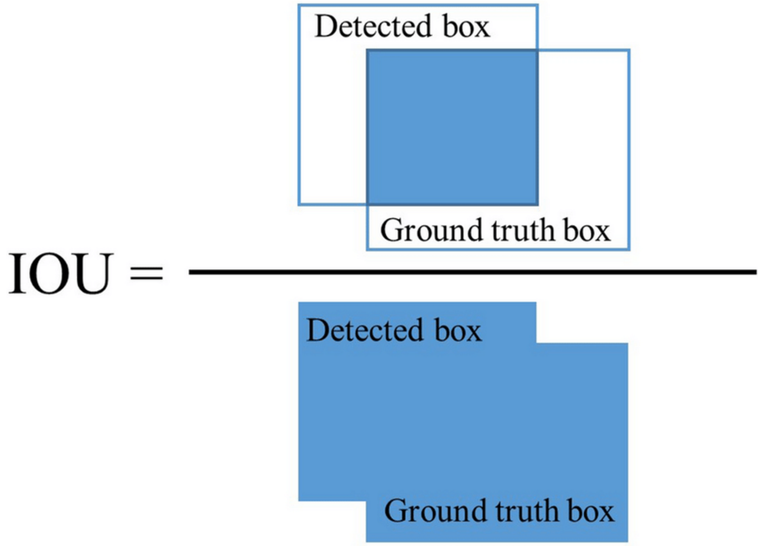
- Select g.t. mask from dataset
  - Sample point from mask
  - Prompt model with point, model predicts mask
- Compute IoU (prediction, g.t.)
  - Metric: mean IoU (mIoU) over mask
- Oracle evaluation
  - Recall SAM can predict 3 masks to handle ambiguity



slide credit: Alexander Kirillov

Single Point Valid Mask Evaluation Protocol

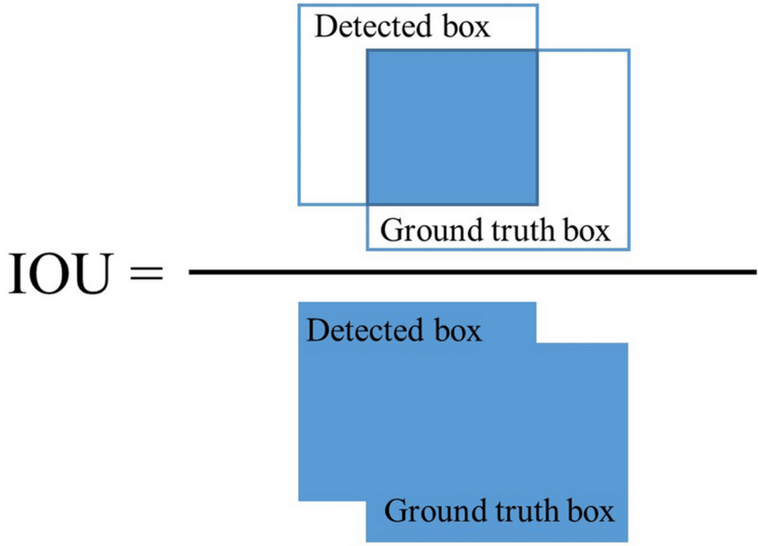
- Select g.t. mask from dataset
  - Sample point from mask
  - Prompt model with point, model predicts mask
- Compute IoU (prediction, g.t.)
  - Metric: mean IoU (mIoU) over mask
- Oracle evaluation
  - Recall SAM can predict 3 masks to handle ambiguity
  - Oracle evaluation: mask with best g.t. IoU is selected



slide credit: Alexander Kirillov

Single Point Valid Mask Evaluation Protocol

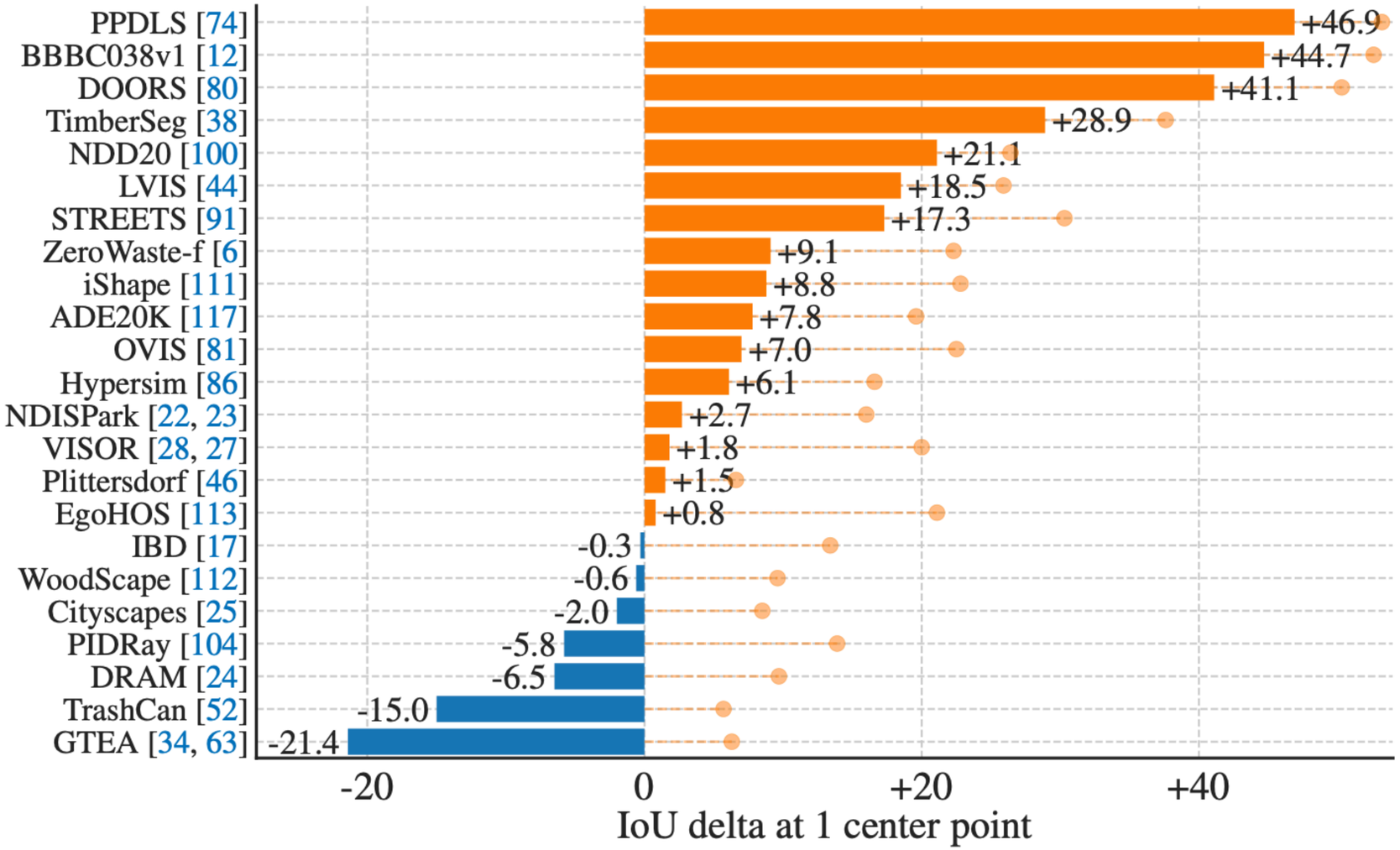
- Select g.t. mask from dataset
  - Sample point from mask
  - Prompt model with point, model predicts mask
- Compute IoU (prediction, g.t.)
  - Metric: mean IoU (mIoU) over mask
- Oracle evaluation
  - Recall SAM can predict 3 masks to handle ambiguity
  - Oracle evaluation: mask with best g.t. IoU is selected
  - Regular evaluation: model puts forth most confident mask



slide credit: Alexander Kirillov

Single Point Valid Mask Evaluation

- SAM vs. RITM
- Circle = oracle



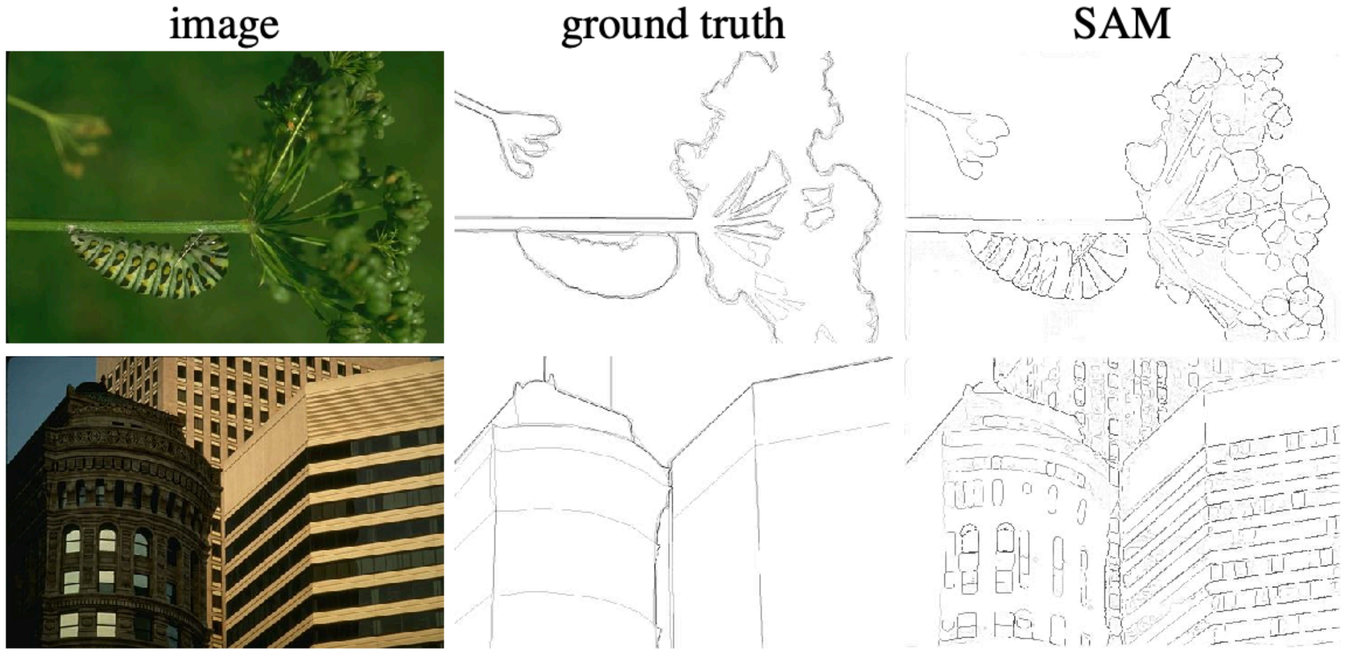
slide credit: Alexander Kirillov

RITM: Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. ICIP, 2022

Edge Detection

⇨

- BSDS500
- Segment everything (regular grid of foreground points)
  → filter edges from masks → edge NMS



method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

slide credit: Alexander Kirillov

Text-to-Mask

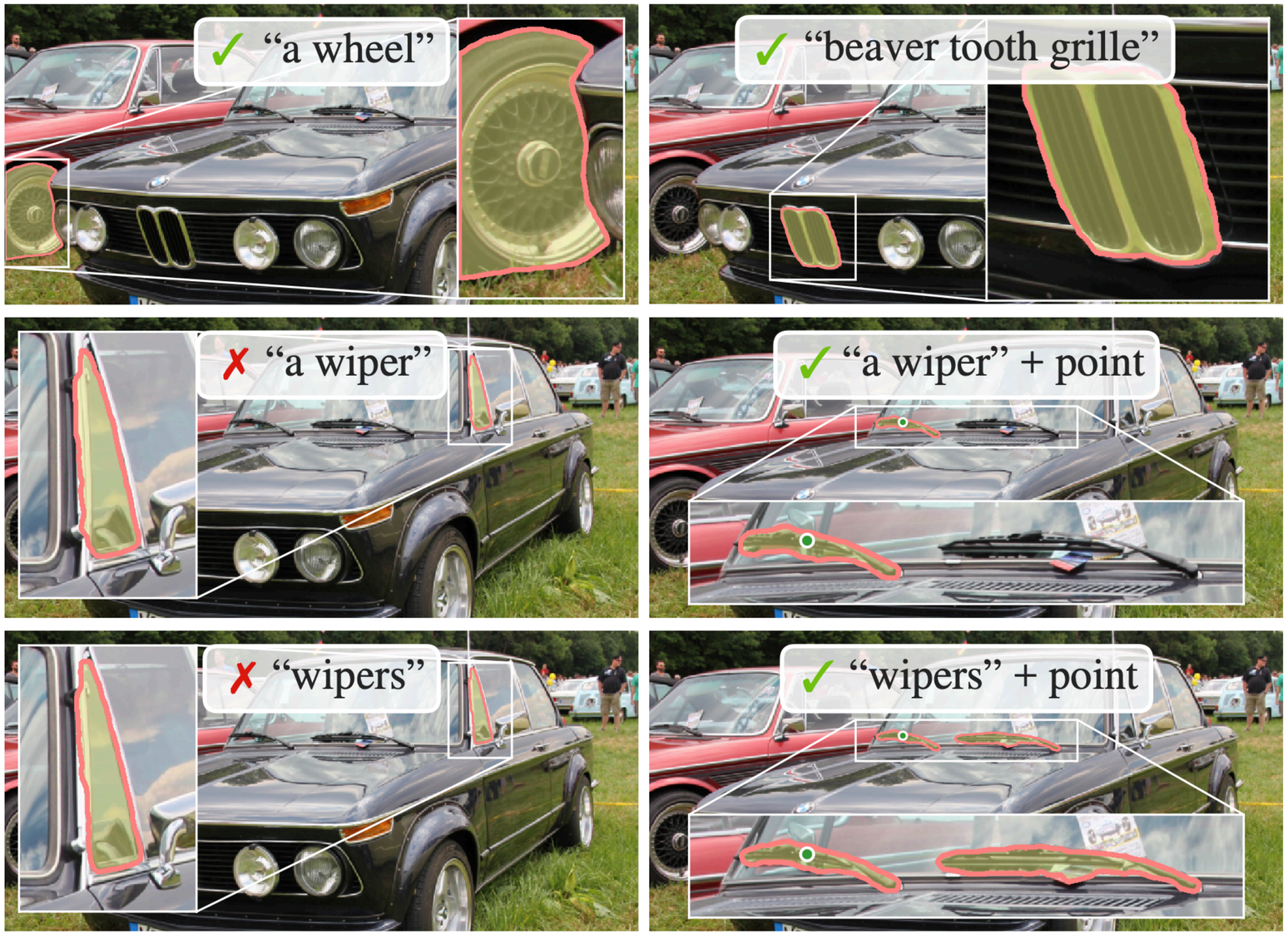
- Training:
  - For each mask (larger than 100x100), compute the CLIP image encoder embedding
  - Prompt with embedding during simulated interactions

- Recall in CLIP image and text embeddings are *aligned*

- During inference: prompt with text embedding

slide credit: Alexander Kirillov

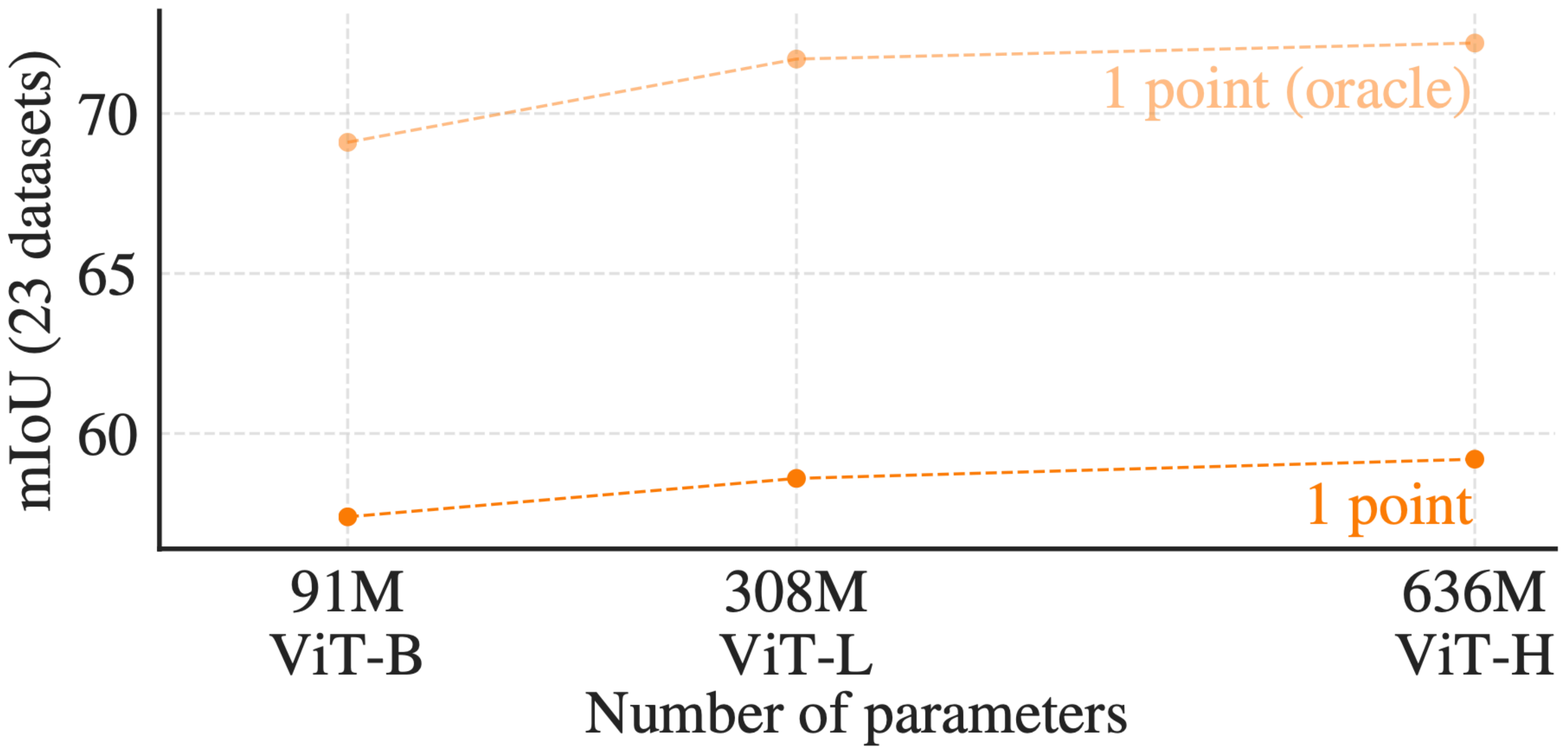
Text-to-Mask Qualitative Examples



slide credit: Alexander Kirillov

Ablations

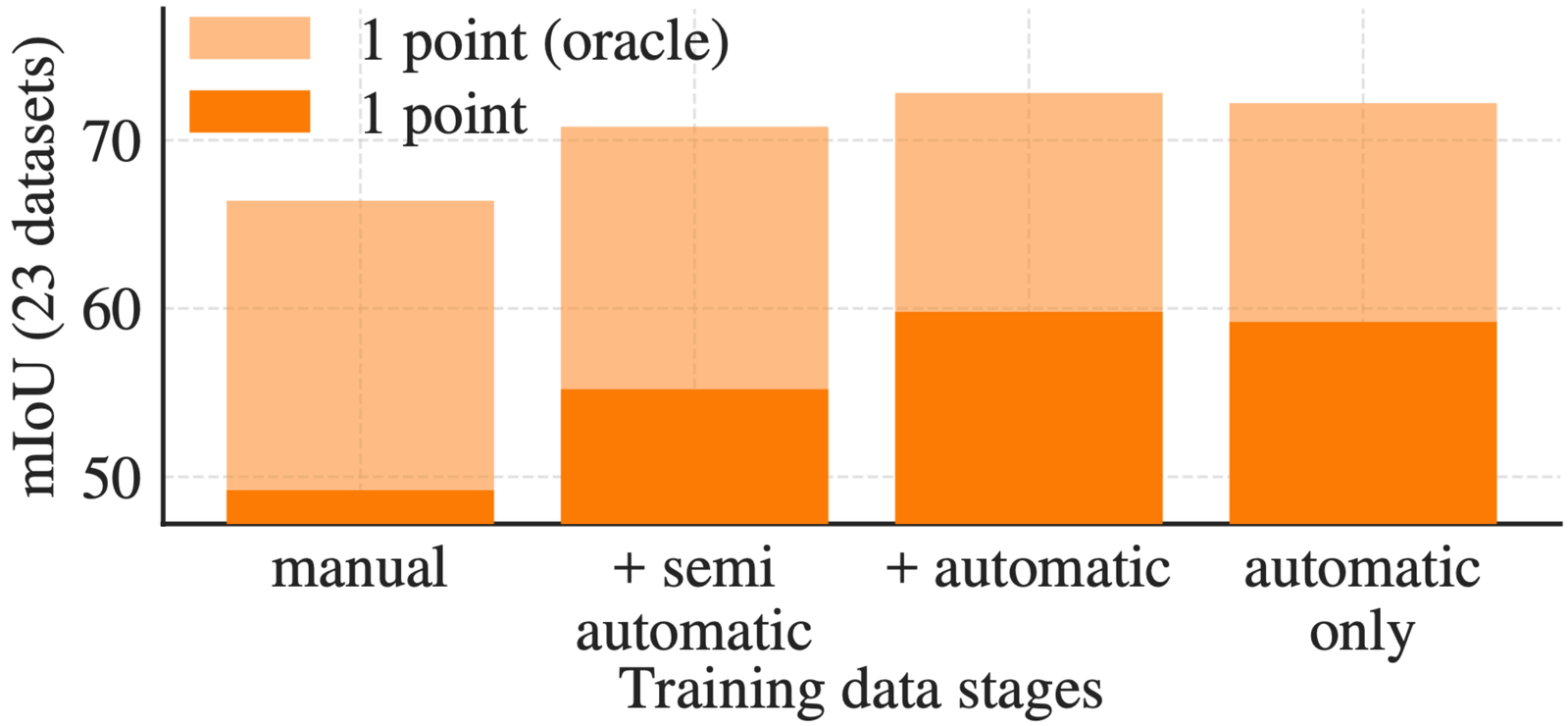
- Scaling saturation



slide credit: Alexander Kirillov

Ablations

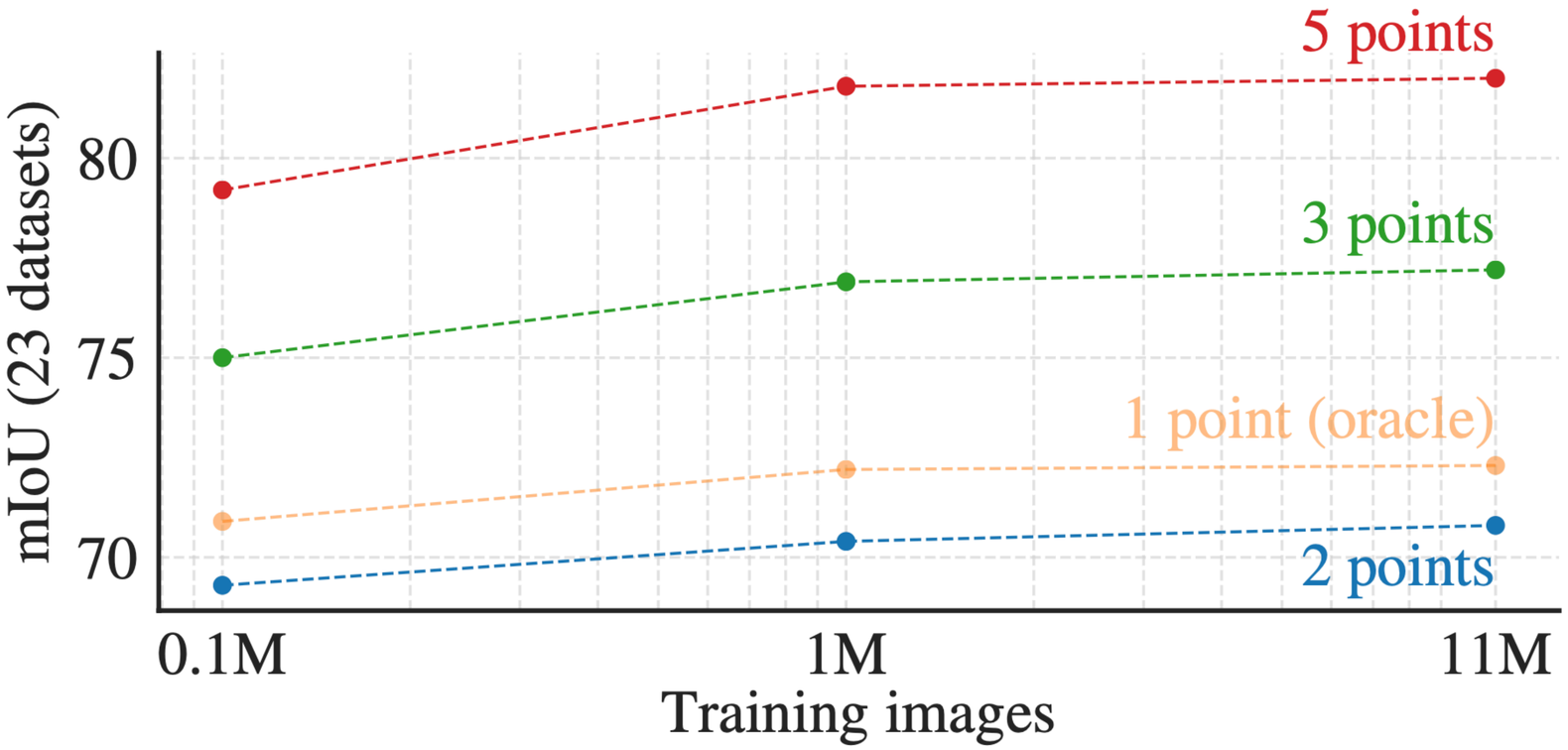
- Automatic data is good enough



slide credit: Alexander Kirillov

Ablations

- Training with 10% of the data is good



slide credit: Alexander Kirillov

Limitations

slide credit: Alexander Kirillov

Limitations

- Can miss fine structures

slide credit: Alexander Kirillov

Limitations

- Can miss fine structures
- Small “sprinkles” and holes (fixed with post-processing)

slide credit: Alexander Kirillov

Limitations

- Can miss fine structures
- Small “sprinkles” and holes (fixed with post-processing)
- Boundaries not as crisp as “zoom-in” approaches

slide credit: Alexander Kirillov

Limitations

- Can miss fine structures
- Small “sprinkles” and holes (fixed with post-processing)
- Boundaries not as crisp as “zoom-in” approaches
- Interactive-only models may work better with many points

slide credit: Alexander Kirillov

Limitations

- Can miss fine structures
- Small “sprinkles” and holes (fixed with post-processing)
- Boundaries not as crisp as “zoom-in” approaches
- Interactive-only models may work better with many points
- Amortized real-time, but not real-time (with ViT-H anyway)

slide credit: Alexander Kirillov

Limitations

- Can miss fine structures
- Small “sprinkles” and holes (fixed with post-processing)
- Boundaries not as crisp as “zoom-in” approaches
- Interactive-only models may work better with many points
- Amortized real-time, but not real-time (with ViT-H anyway)
- Text-to-mask exploration is preliminary

slide credit: Alexander Kirillov

Limitations

- Can miss fine structures
- Small “sprinkles” and holes (fixed with post-processing)
- Boundaries not as crisp as “zoom-in” approaches
- Interactive-only models may work better with many points
- Amortized real-time, but not real-time (with ViT-H anyway)
- Text-to-mask exploration is preliminary
- Domain-specialized tools may work better

slide credit: Alexander Kirillov

slide credit: Alexander Kirillov

Conclusion

- Main contributions
  - New task (yields pre-training and inference)
  - New model (SAM)
  - New dataset (SA-1B – 1 billion masks!)
- The SA Project: attempt at a base/foundation model for segmentation
  - Up to community use and evaluation to see if that is the case

- NLP:LLMs :: CV:???
  - Is there something as simple, general, and powerful for CV?

Links

- Project website: https://segment-anything.com
- Interactive demo: https://segment-anything.com/demo
- Dataset explorer: https://segment-anything.com/dataset/index.html
- Code (inference only for now): https://github.com/facebookresearch/segment-anything

slide credit: Alexander Kirillov