max planck institut informatik

SIC Saarland Informatics Campus

High Level Computer Vision

Some Recent Trends: Flamingo & B-Cos-Nets

@ July 5, 2023

Bernt Schiele

cms.sic.saarland/hlcvss23/

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

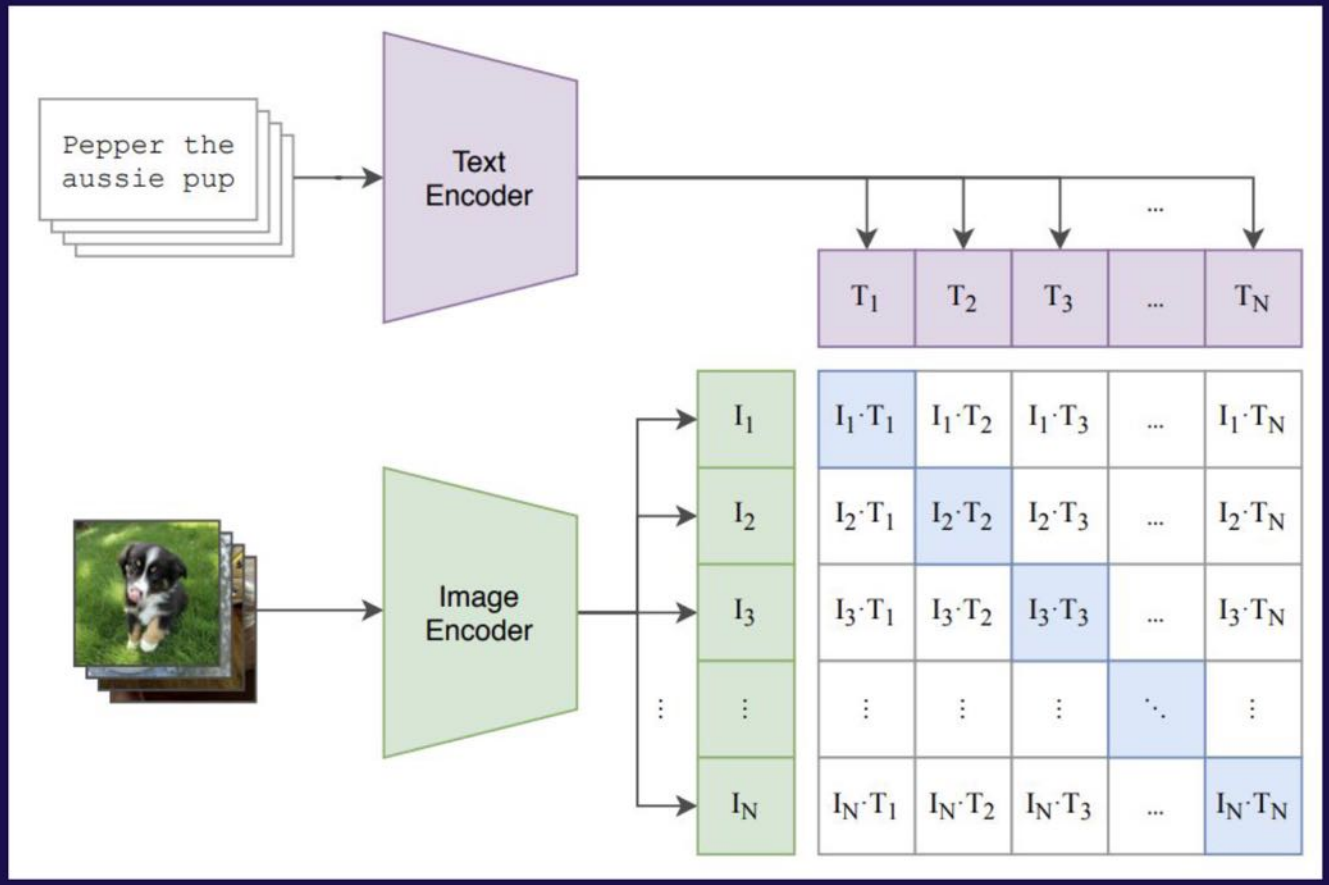
Overview of Today's Lecture

- Vision-Language Learning for Computer Vision
  - Recap: CLIP
  - Flamingo [neurips'22] - https://arxiv.org/abs/2204.14198

- Inherently Interpretable Neural Network
  - B-cos CNN [cvpr'22] - https://arxiv.org/abs/2205.10268
  - B-cos CNN & Vision transformers [arxiv'23] - https://arxiv.org/abs/2306.10898

Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

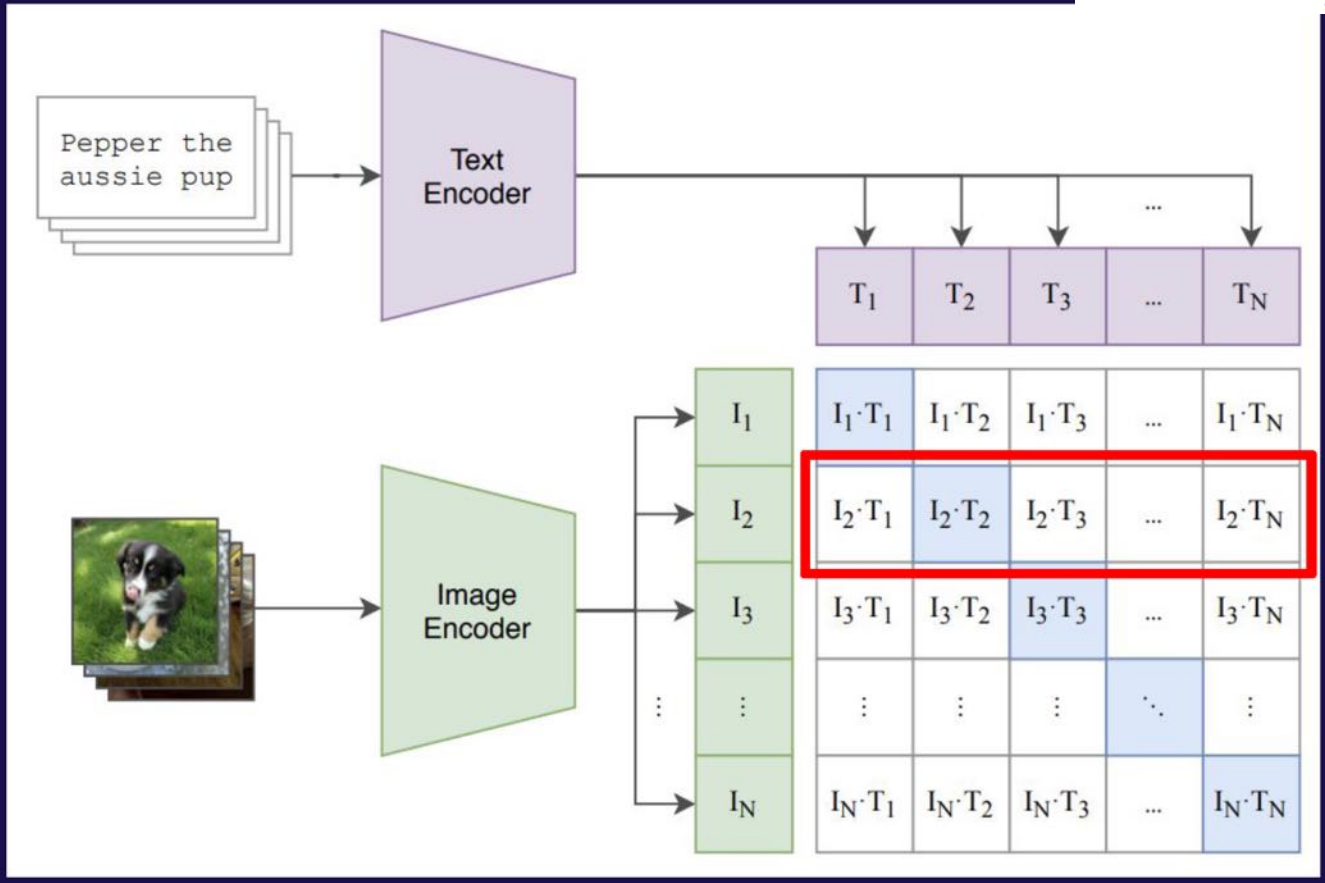
CLIP: Contrastive Language-Image Pre-training



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

CLIP: Contrastive Language-Image Pre-training

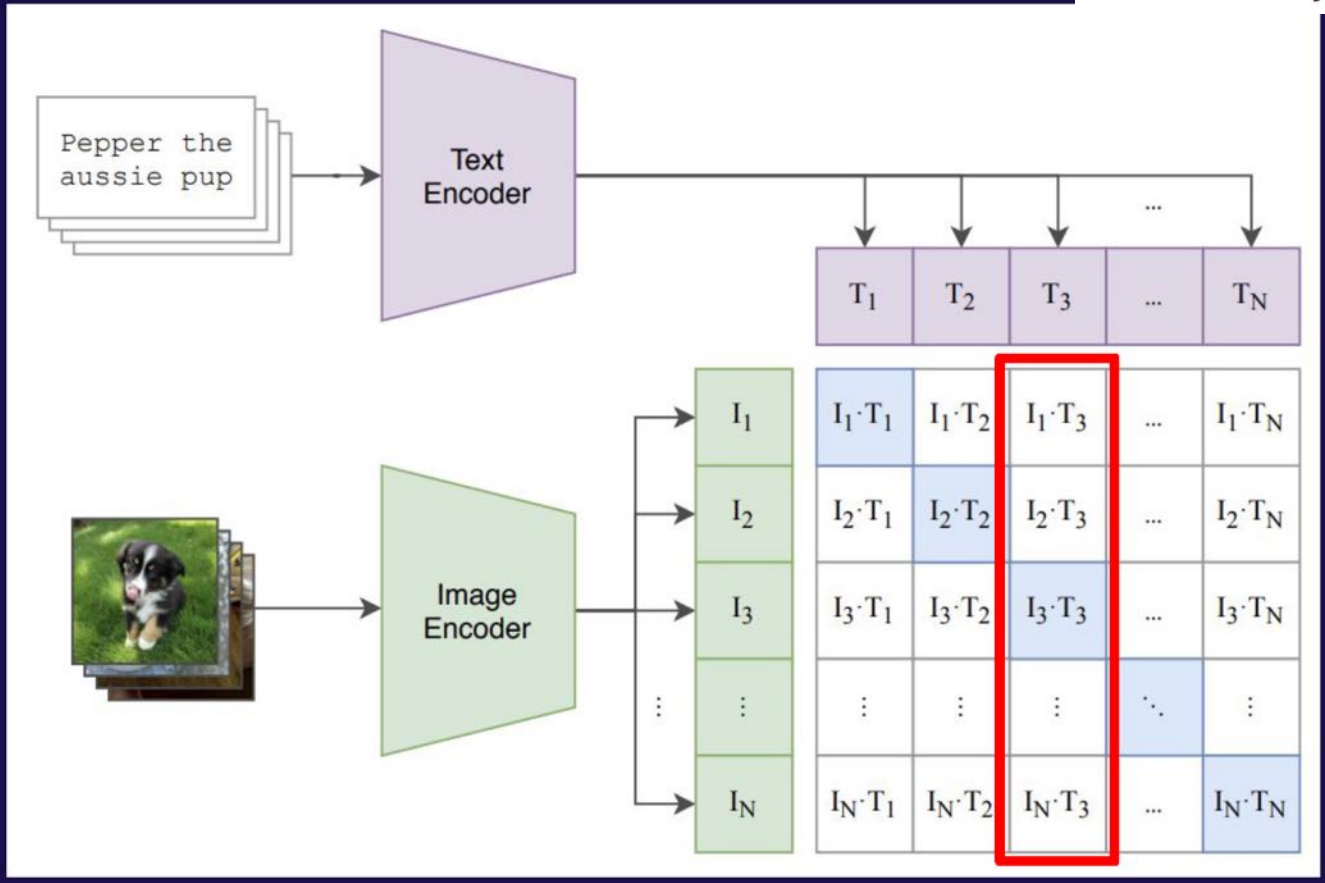
$$L_{i2t} = -\sum_{j} \log \frac{\exp I_j T_j^T}{\sum_k \exp I_j T_k^T}$$



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

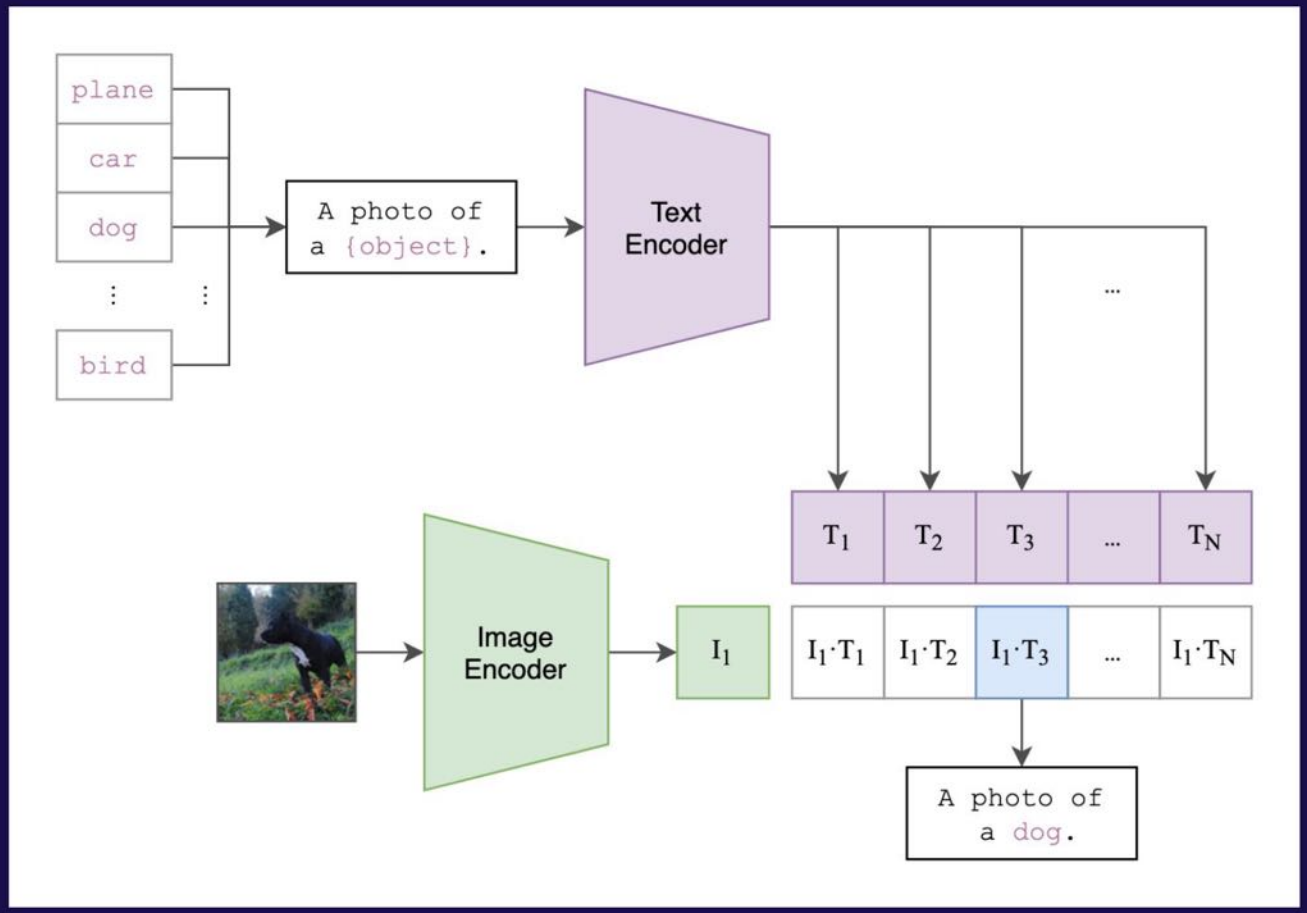
CLIP: Contrastive Language-Image Pre-training

$$L_{t2i} = -\sum_j \log \frac{\exp I_j T_j^T}{\sum_k \exp I_k T_j^T}$$



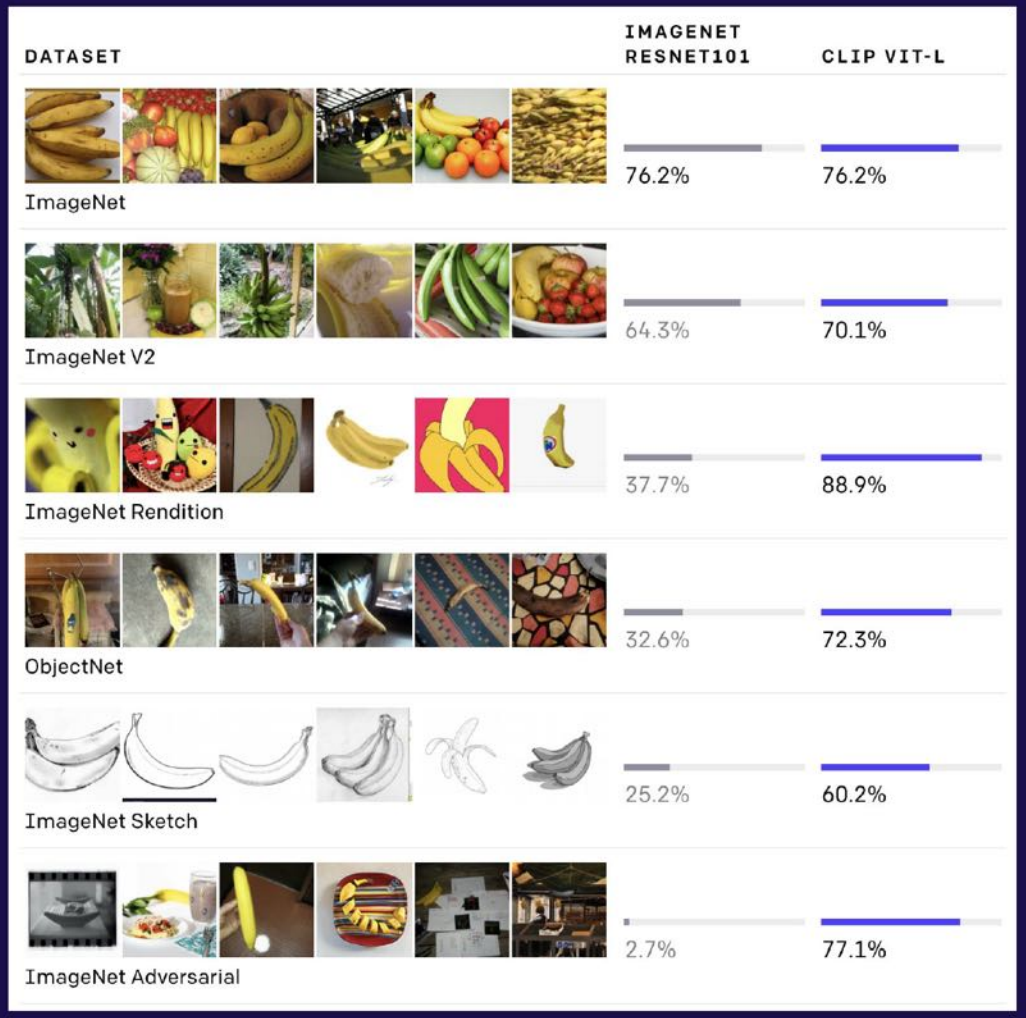
Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Zero-shot image classification



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

Zero-shot CLIP is much more robust



slide credit: Jianwei Yang

The Lesson from CLIP

- Image recognition can be formulated as an **image-text matching problem** instead of image-label mapping problem
- Image recognition does not require human-annotated image-label data but **huge amount of (noisy) image-text pairs**
- **Contrastive learning is a good learning objective** for multi-modal learning strategy compared with generative learning
- **Two-tower model without fusion** is sufficient to learn good and generic visual and language representations

Overview of Today's Lecture

- Continuation of Vision-Language Learning for Computer Vision
  - Recap: CLIP
  - Flamingo [neurips'22] - https://arxiv.org/abs/2204.14198

- Inherently Interpretable Neural Network
  - B-cos CNN [cvpr'22] - https://arxiv.org/abs/2205.10268
  - B-cos CNN & Vision transformers [arxiv'23] - https://arxiv.org/abs/2306.10898

slide credit: Jean-Babtiste Alayrac

DeepMind

Flamingo: a Visual Language Model for Few-Shot Learning

JB Alayrac

Authors: Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan

T4V: Transformers for Vision Workshop
June 19th 2022

slide credit: Jean-Babtiste Alayrac

What is Flamingo?

Output: Free-form text

A portrait of Salvador Dali with a robot head.

Flamingo Model

Input: Text and visual data interleaved

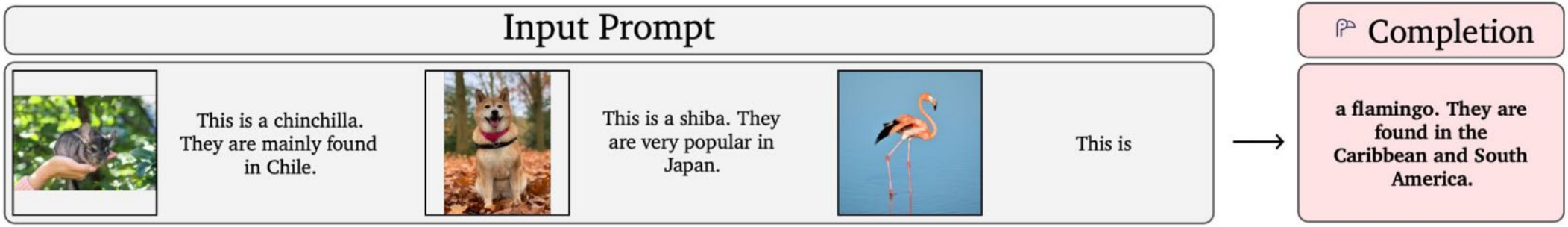
Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.

Output: A pink room with a flamingo pool float.

Output:

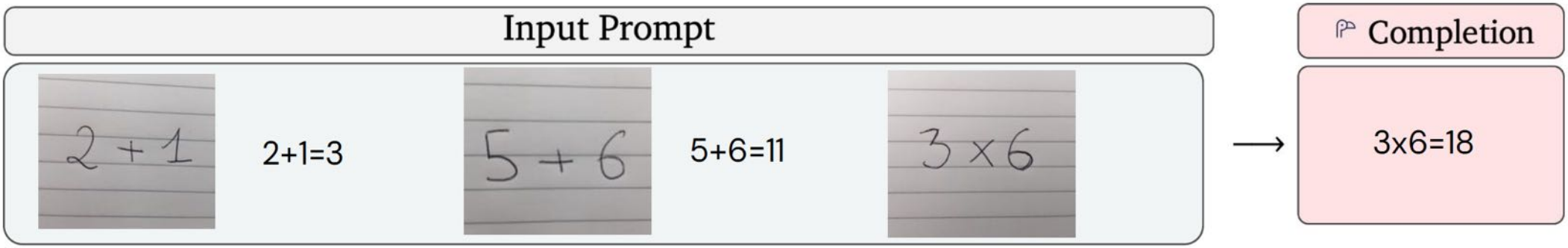
slide credit: Jean-Babtiste Alayrac

Flamingo comes in many colours depending on what it eats



slide credit: Jean-Babtiste Alayrac

Flamingo comes in many colours depending on what it eats



slide credit: Jean-Babtiste Alayrac

Flamingo comes in many colours depending on what it eats

Input Prompt

What happens to the man after hitting the ball? Answer:

→

Completion

he falls down.

slide credit: Jean-Babtiste Alayrac

Why did we build it?

Build a state-of-the-art, generalist Visual Language Model that can **be rapidly adapted to different multimodal tasks via few-shot learning**

- **Visual Language Model**: ingest visual data (images or videos) along with a language input, and produce language output.

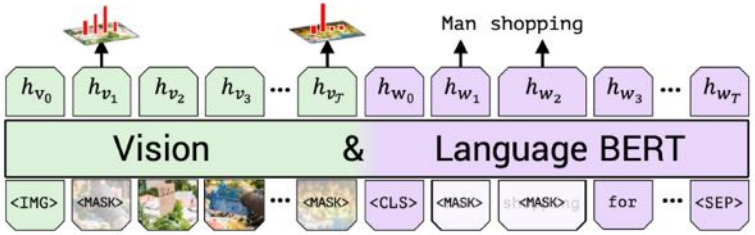
- **Generalist … rapidly adapted to different multimodal tasks:** one model can address multiple tasks (captioning, visual dialogue, classification) with the same weights and without any post-hoc training.

- **Few-shot learning:** condition the model to solve various tasks with only a few input-output examples (32 examples are used)

slide credit: Jean-Babtiste Alayrac

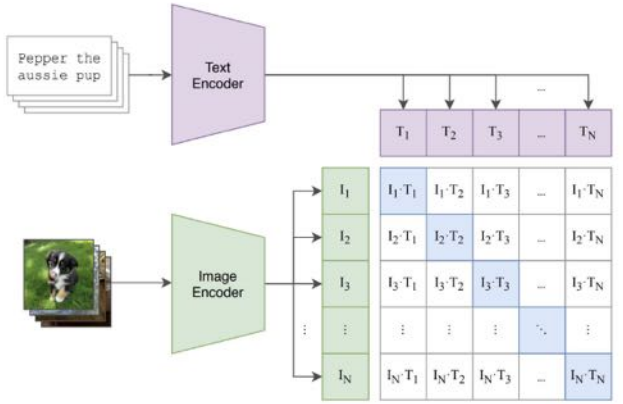
Vision and Language related work

BERT-based models



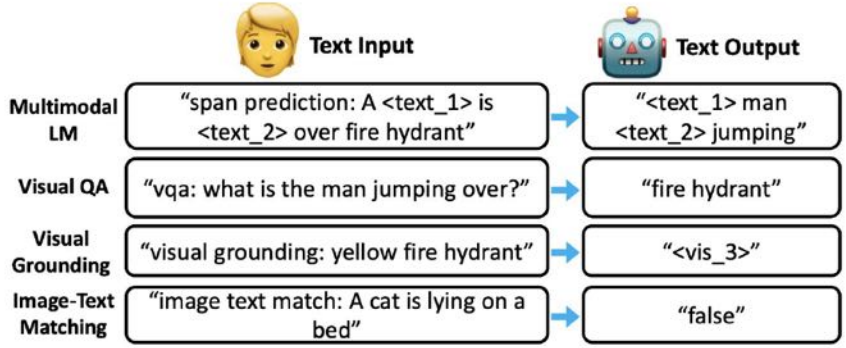
- VilBert
- VisualBERT
- VL-BERT
- UNITER
- OSCAR
- VideoBERT
- ActBERT
- Unicoder-VL
- LXMERT
- MERLOT
- HERO
- ALBEF
- **Many more...**

Dual-encoder contrastive models



- CLIP
- ALIGN
- CoCa
- Florence
- MIL-NCE
- BASIC
- LiT
- FILIP
- MMV

Visual language models



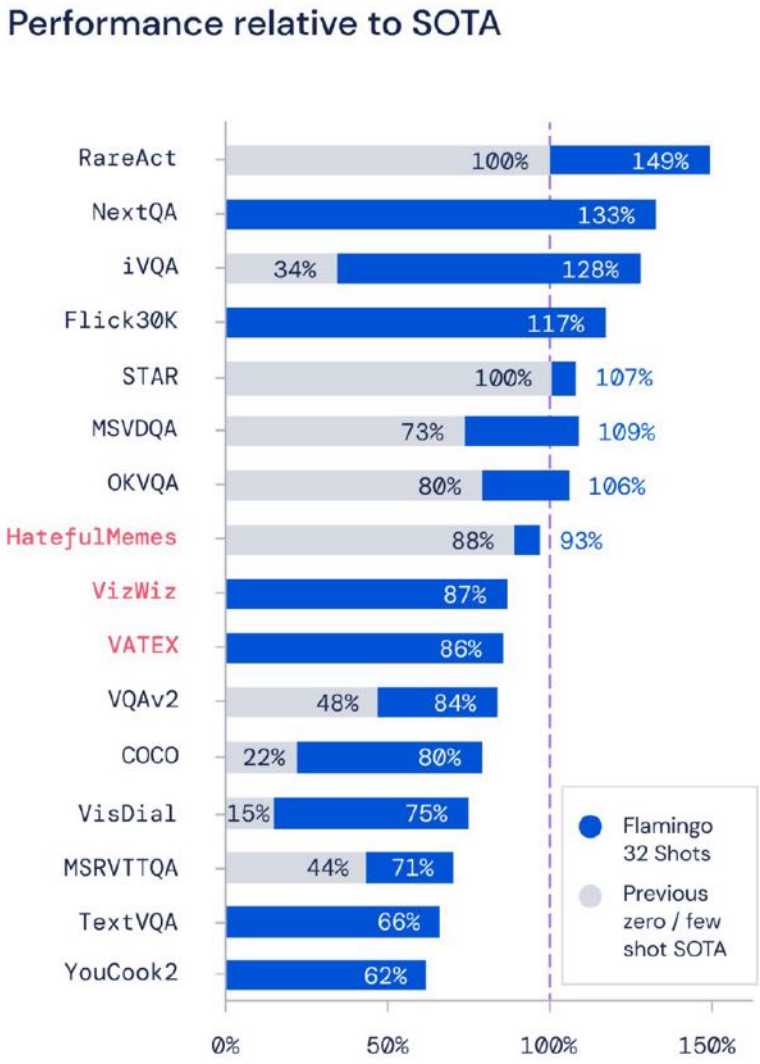
- SimVLM
- Virtex
- MAGMA
- Frozen
- VisualGPT
- ClipClap
- VC-GPT
- CM3
- BLIP
- Uni-Perceiver
- VL-BART
- VL-T5
- VLM

Flamingo sits here

Main result

Flamingo sets a new state of the art in few-shot learning on a wide range of open-ended vision and language tasks.

On the 16 tasks we consider, Flamingo outperform the fine-tuned state-of-art in 7 of the cases despite using orders of magnitude less task-specific training data.

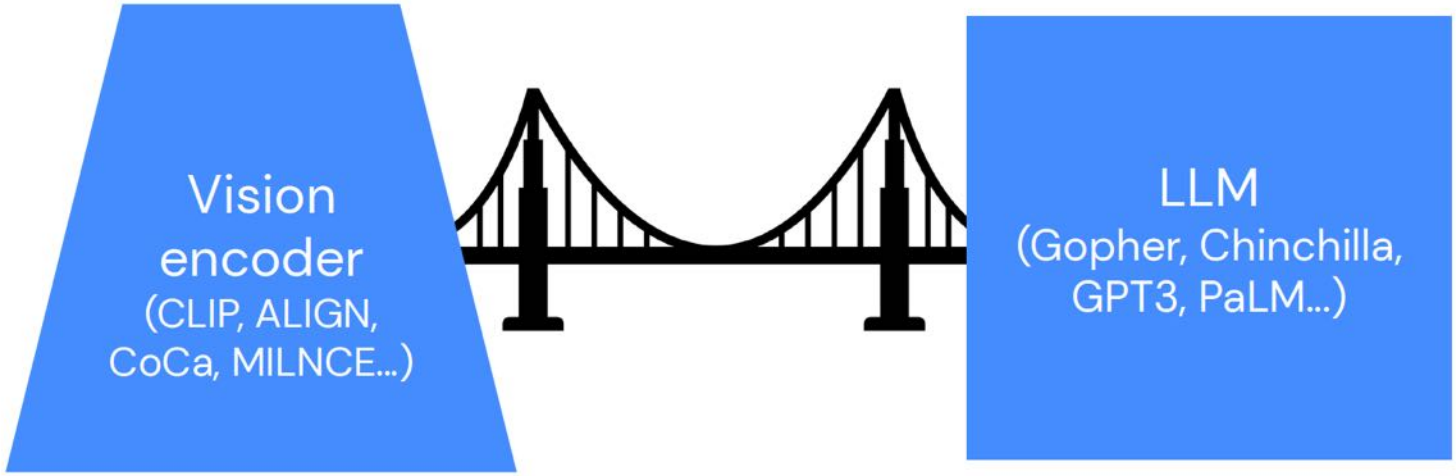


slide credit: Jean-Babtiste Alayrac

How does it work?

Model overview

Pretrained parts of the model are frozen:
the Vision Encoder and the LLM.



👀

The perception

🧠

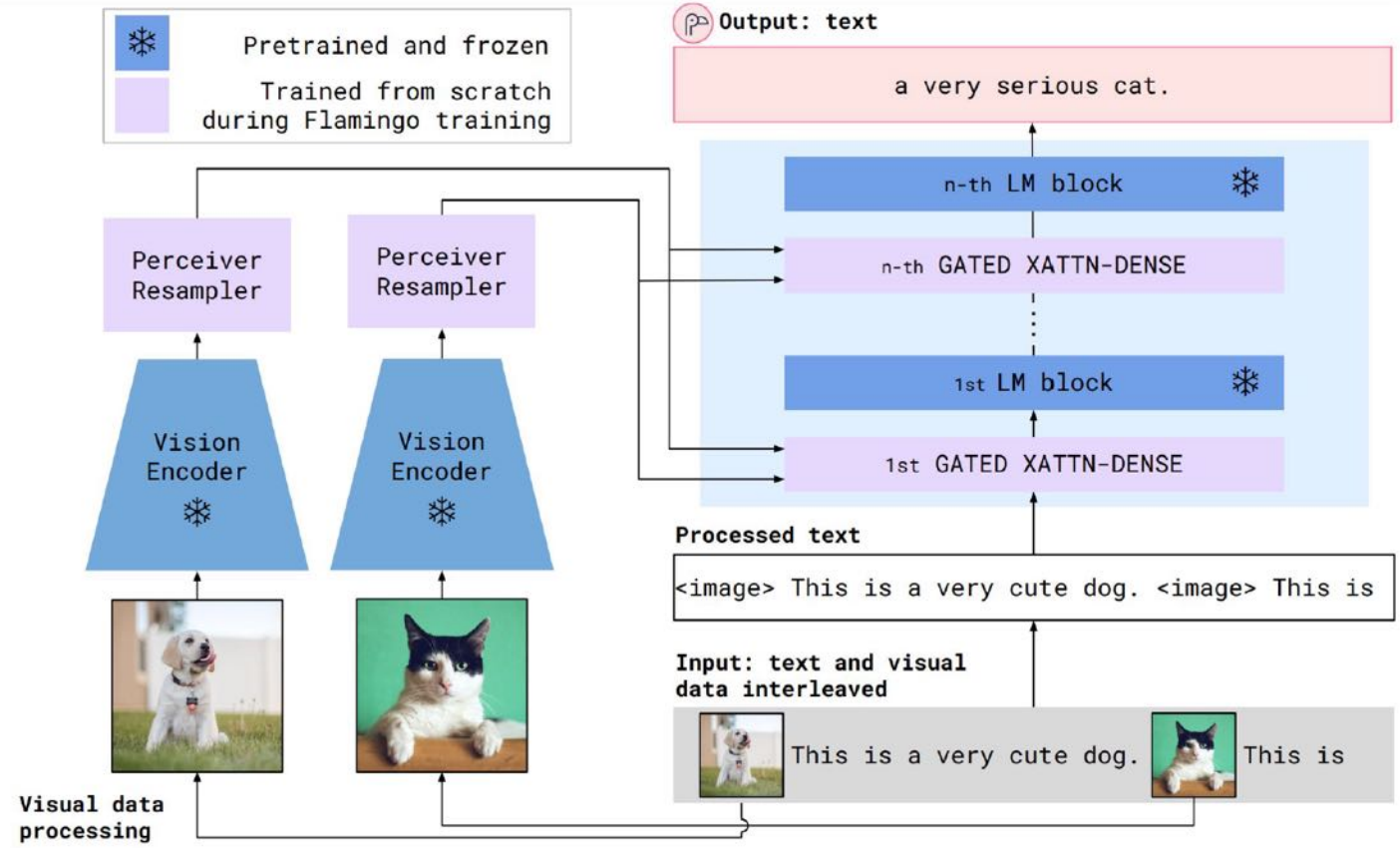
The "reasoning part" and "knowledge source"

slide credit: Jean-Babtiste Alayrac

How does it work?

Model overview

Pretrained parts of the model are frozen:
the Vision Encoder and the LLM.



Training datasets

Image-Text Pairs dataset
[N=1, T=1, H, W, C]

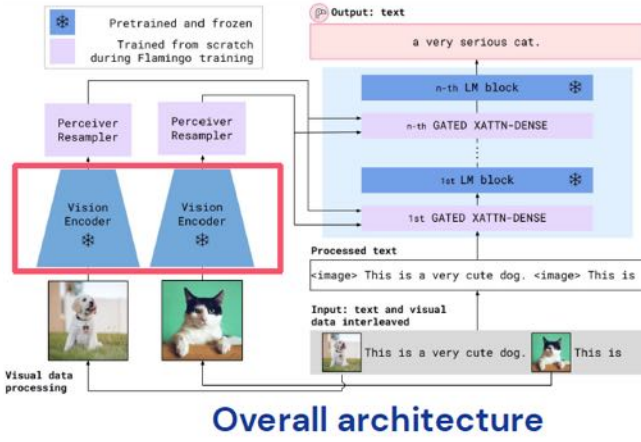
Video-Text Pairs dataset
[N=1, T>1, H, W, C]

Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

Visual processing

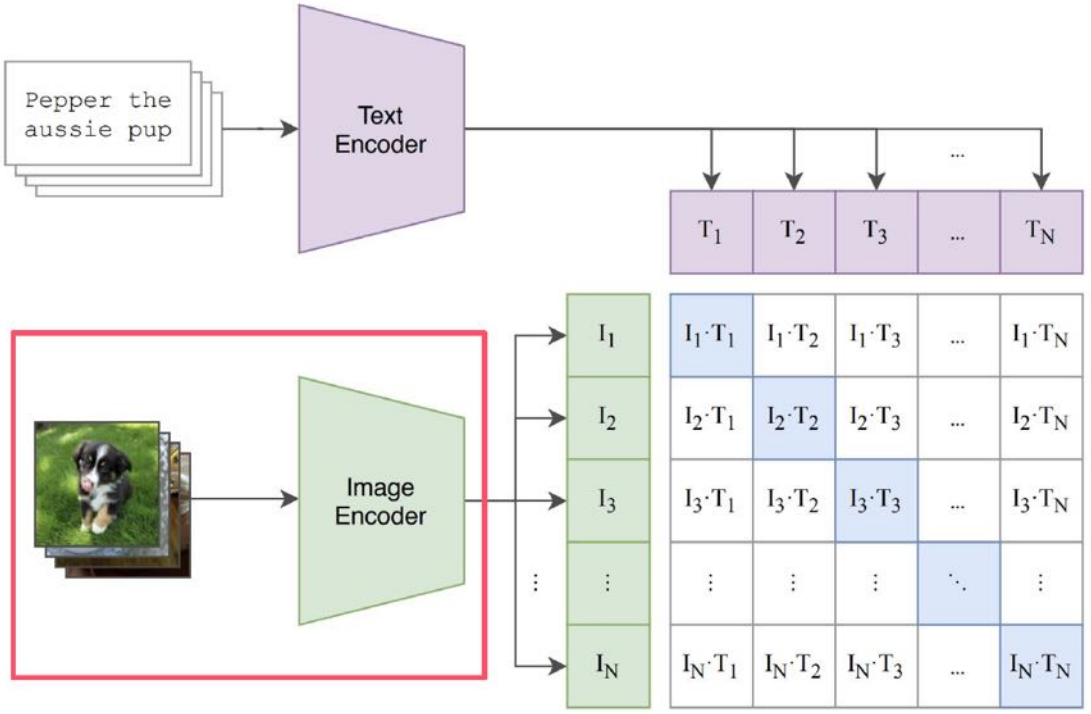
Vision Encoder:

Pretrained with image-text contrastive training (CLIP-like) and kept frozen during Flamingo training.



Overall architecture

We only keep the vision encoder and discard the text encoder.

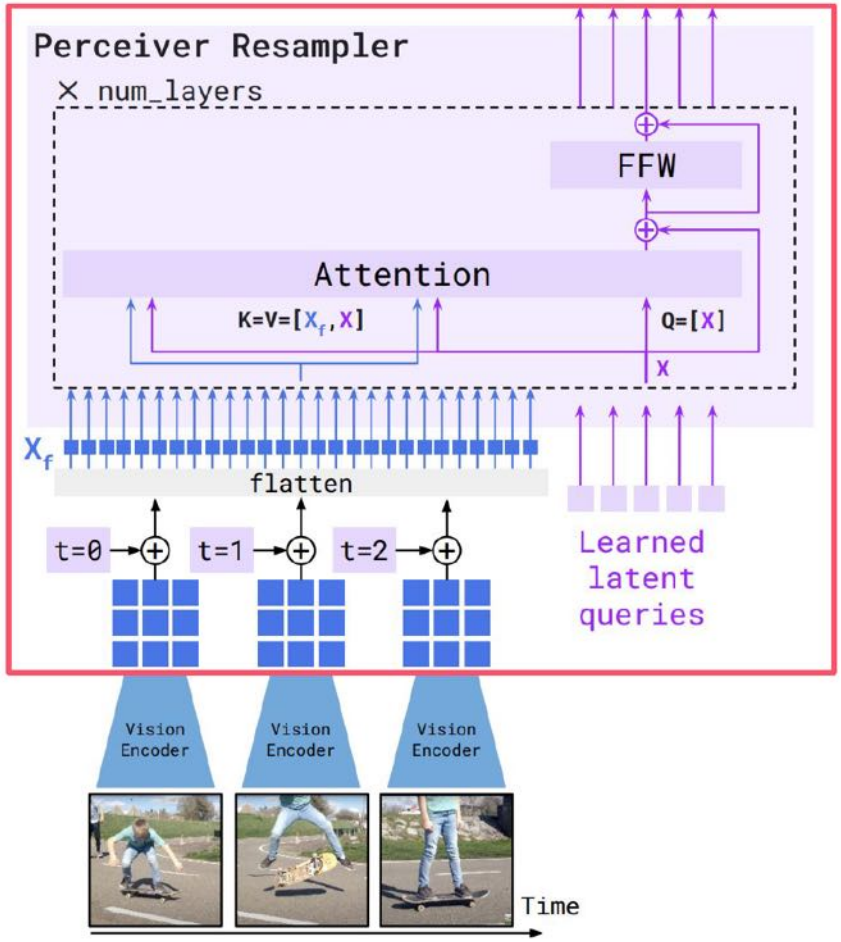


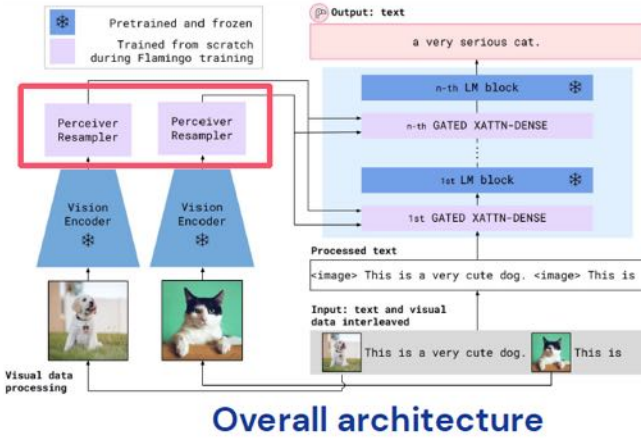
slide credit: Jean-Babtiste Alayrac

Visual processing

Perceiver Resampler:

Takes as *input a variable number of features* (image or videos) and *outputs a fixed number* of "visual tokens".

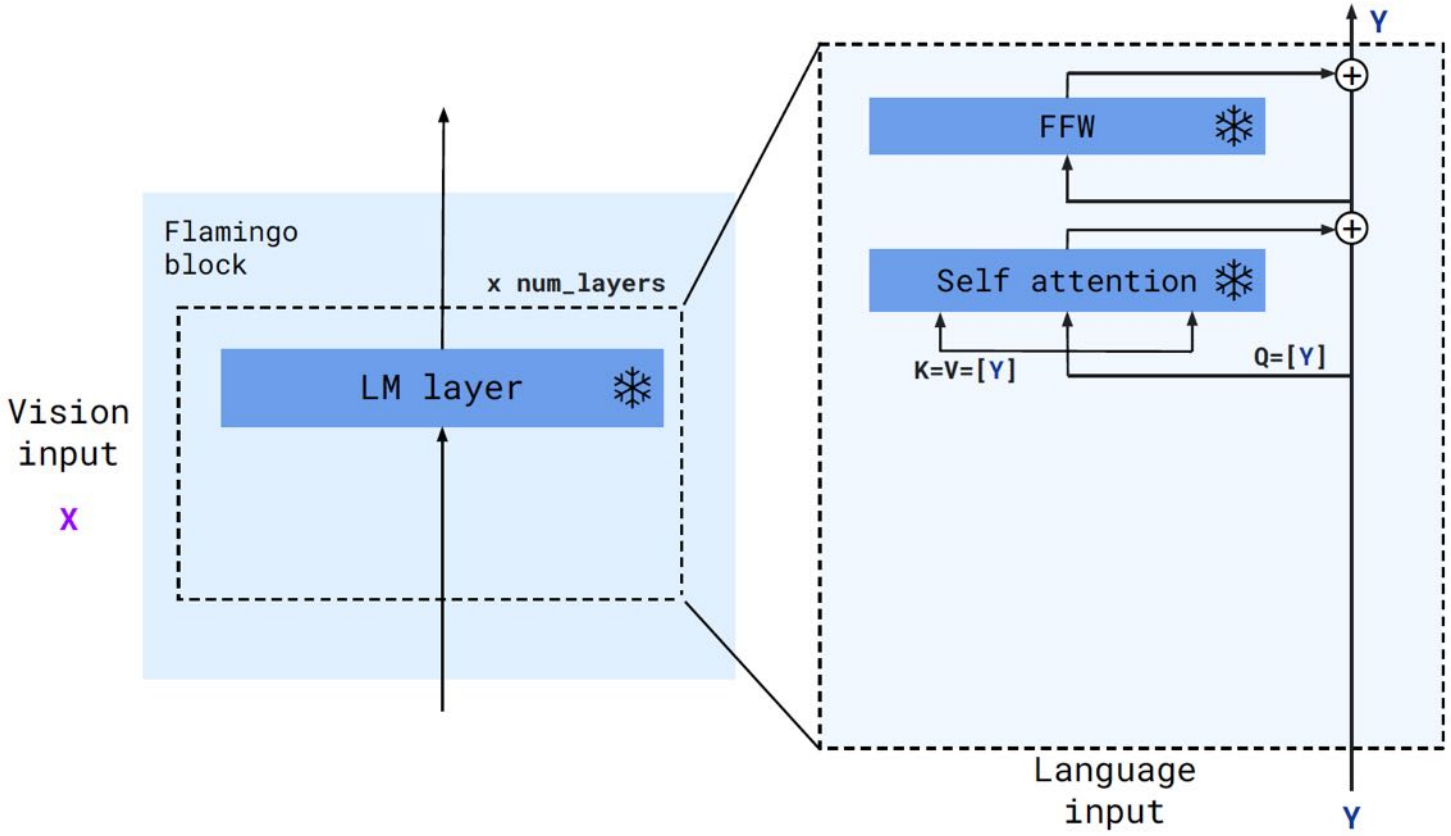


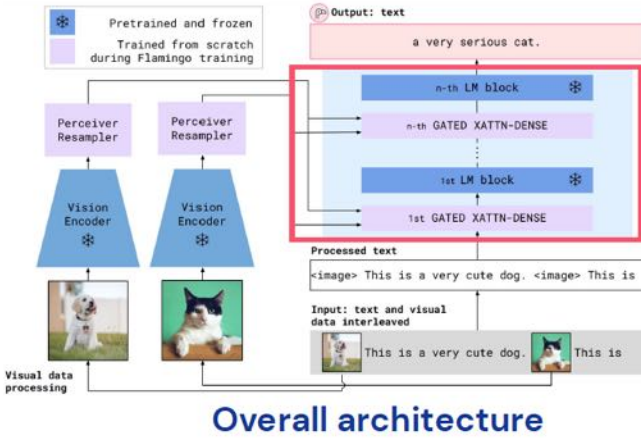


Overall architecture

slide credit: Jean-Babtiste Alayrac

Leveraging an existing language model

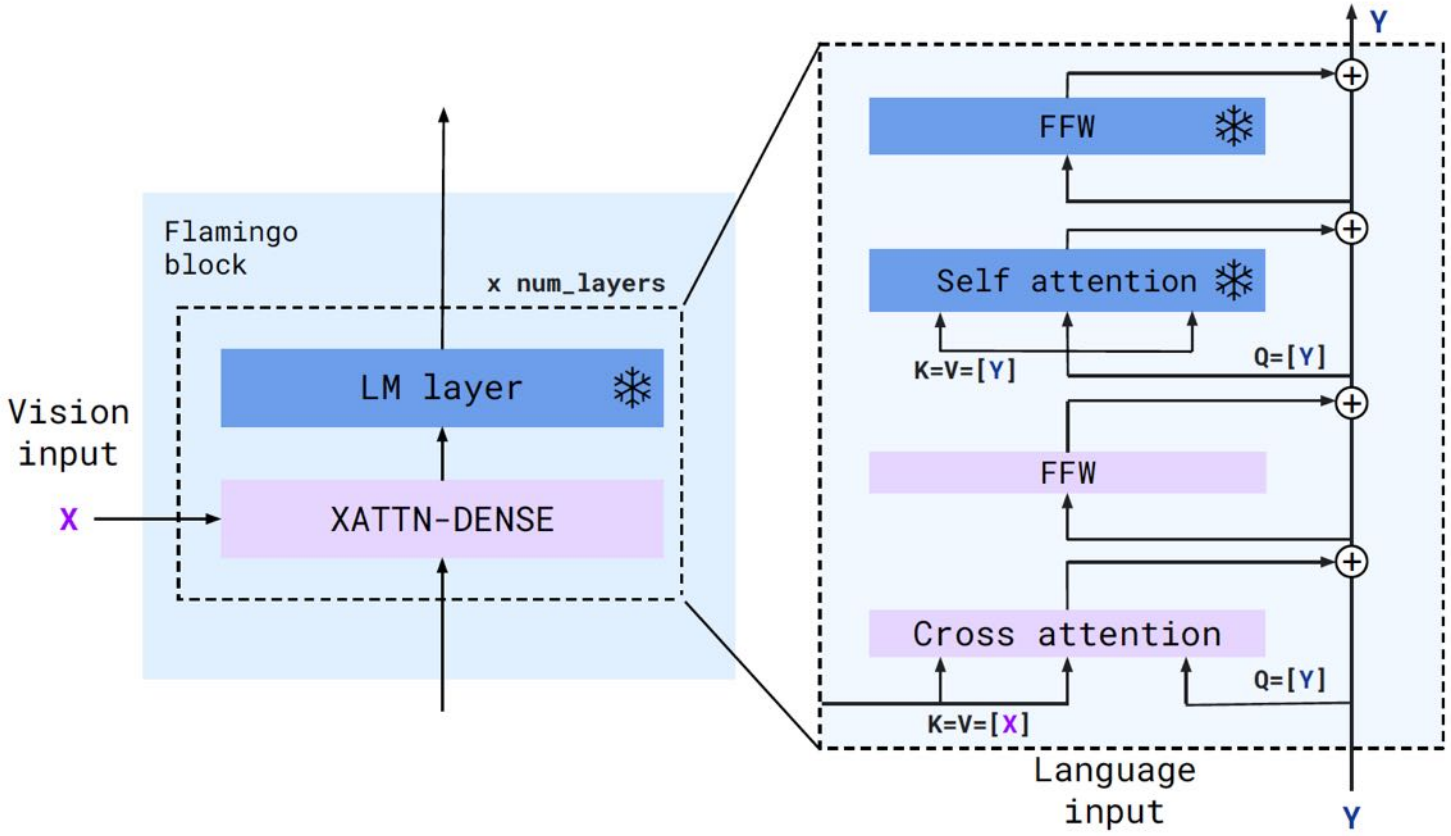


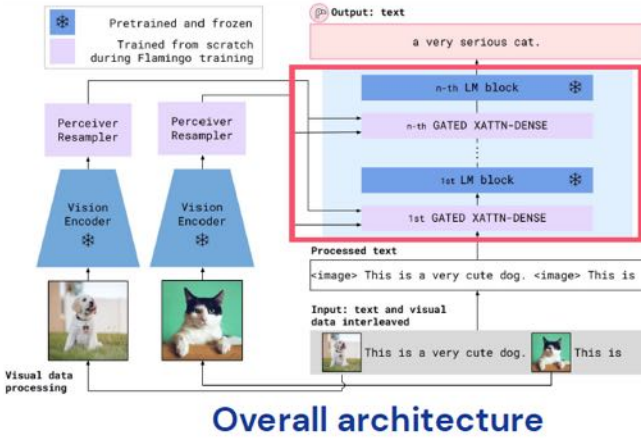


Overall architecture

slide credit: Jean-Babtiste Alayrac

Leveraging an existing language model

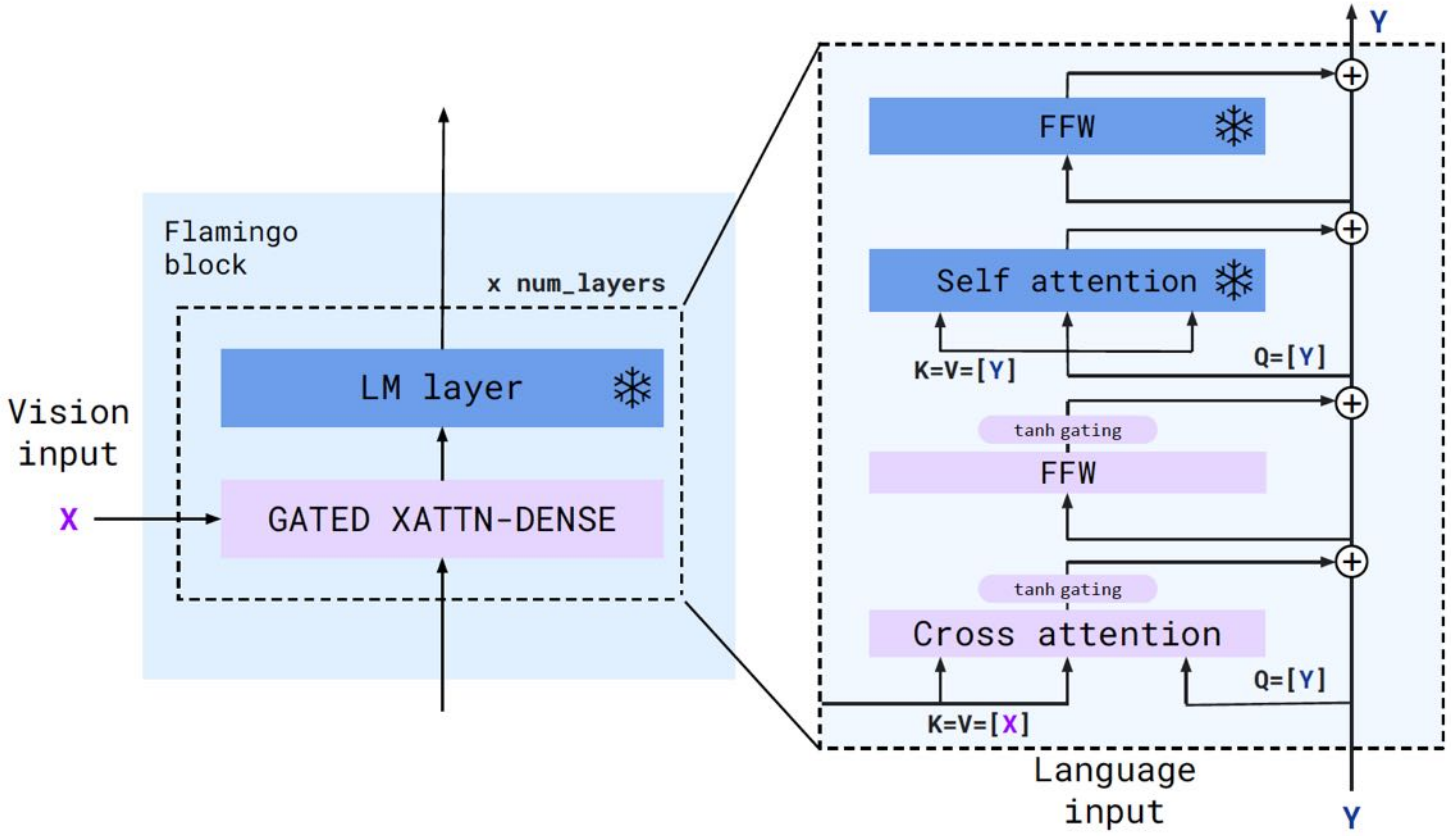


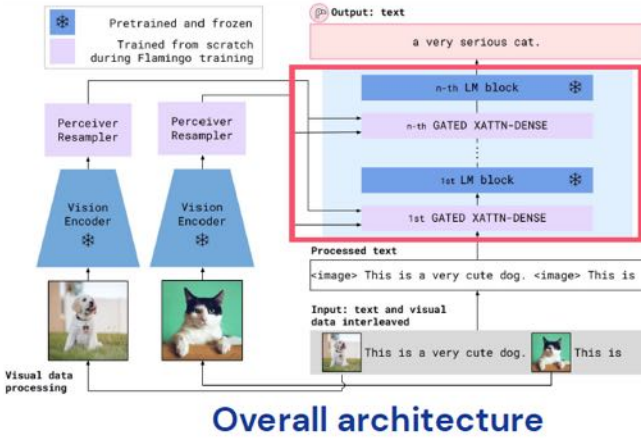


Overall architecture

slide credit: Jean-Babtiste Alayrac

Leveraging an existing language model

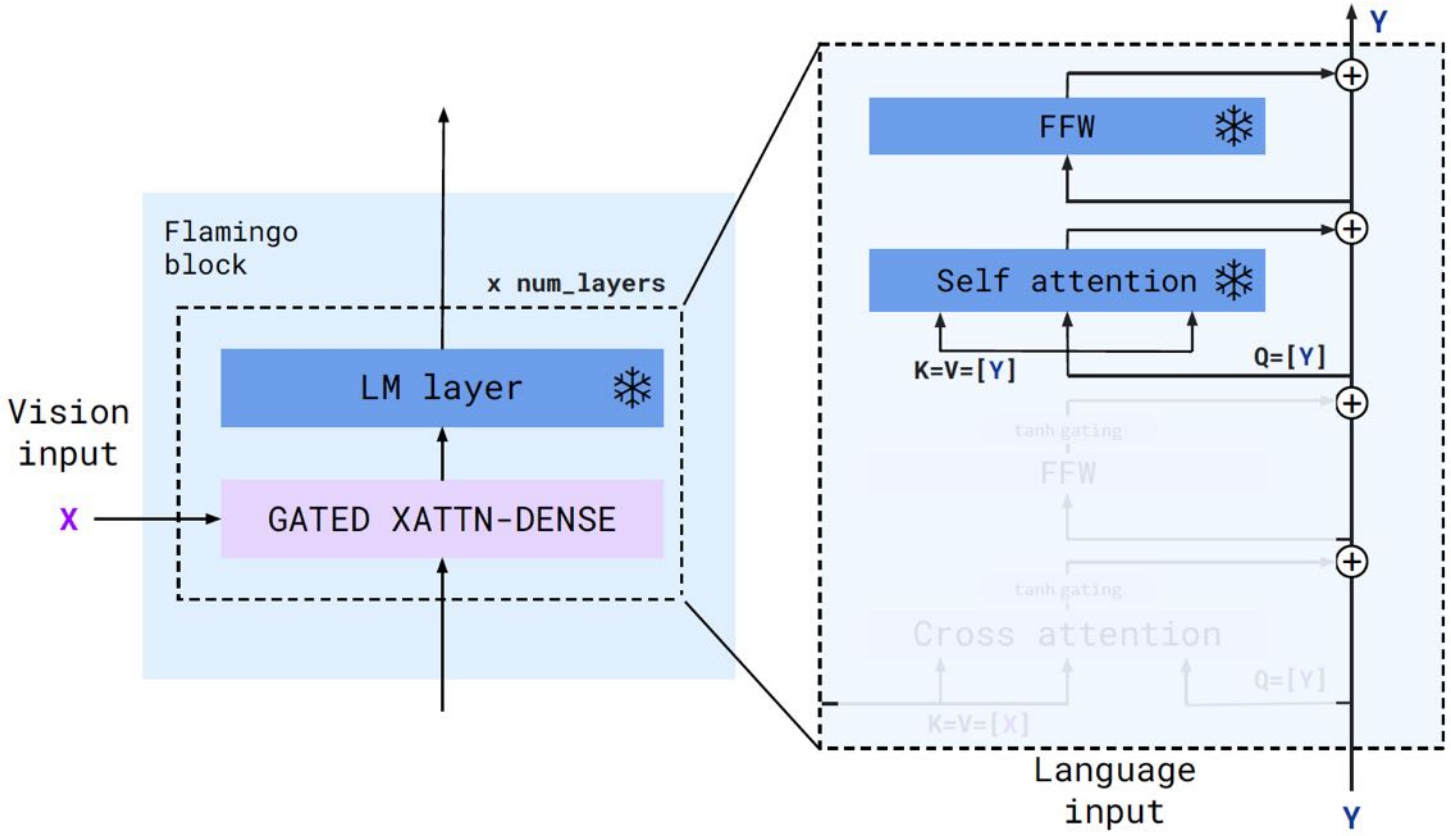


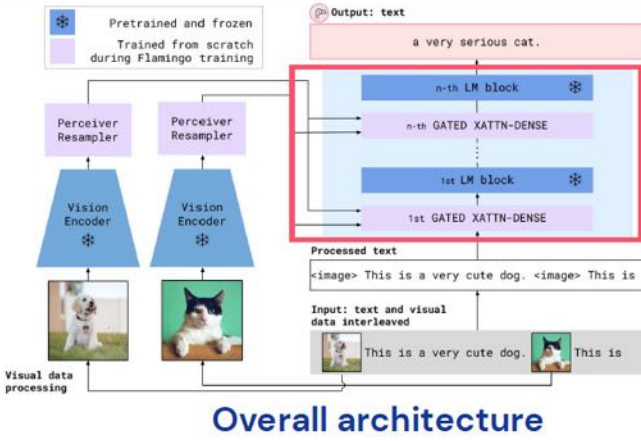


Overall architecture

slide credit: Jean-Babtiste Alayrac

Leveraging an existing language model



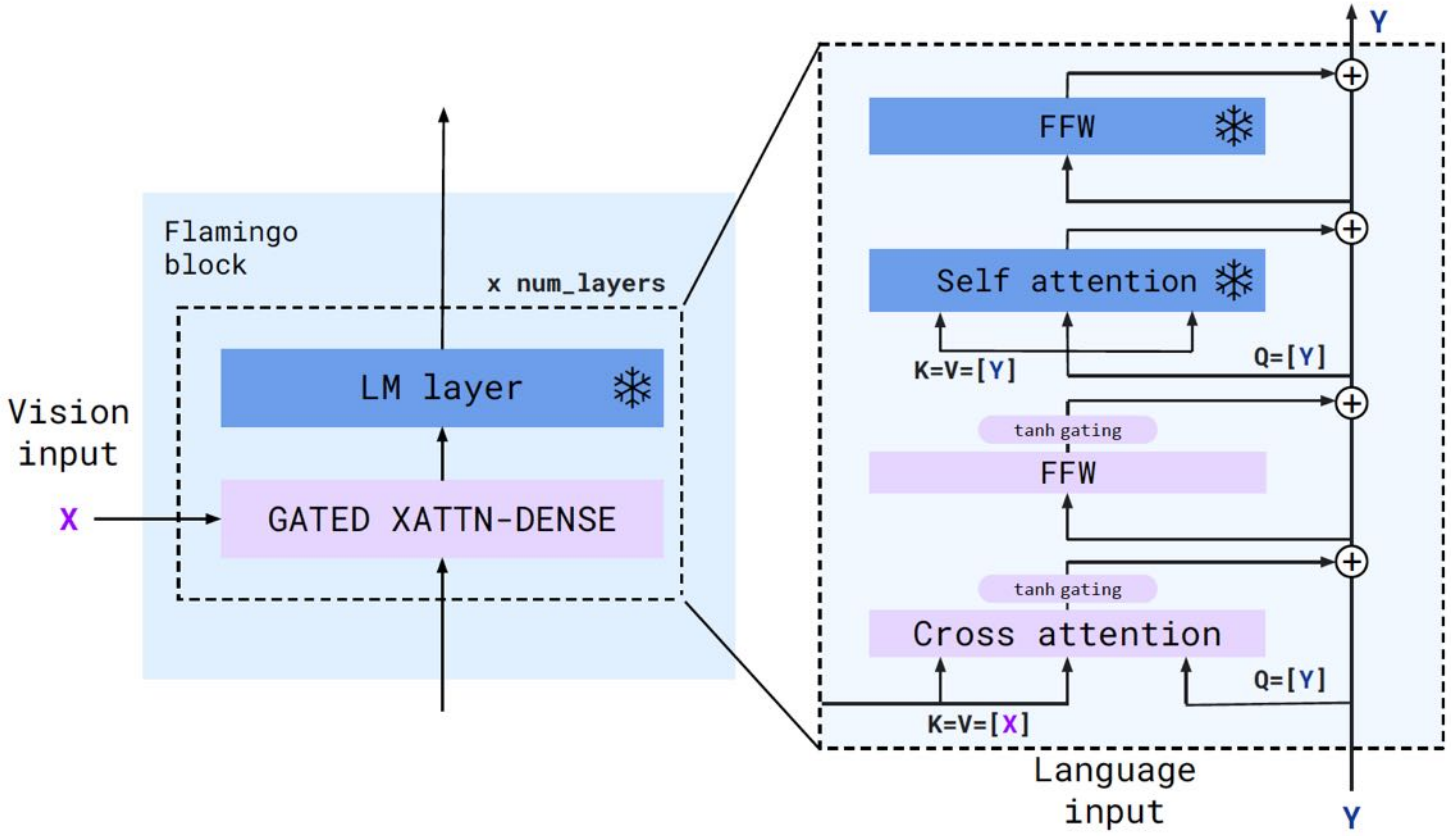


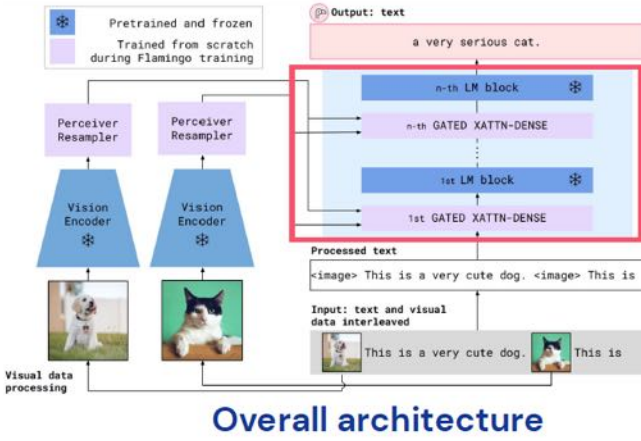
Overall architecture

At initialisation, tanh gates are all 0.

slide credit: Jean-Babtiste Alayrac

Leveraging an existing language model





Overall architecture

At initialisation, tanh gates are all 0.

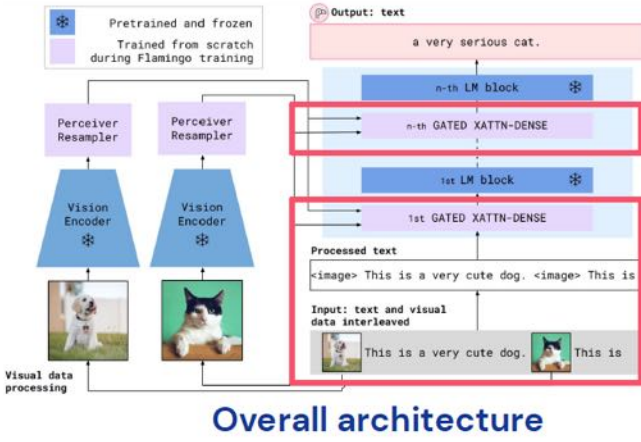
They slowly open as training progresses.

slide credit: Jean-Babtiste Alayrac

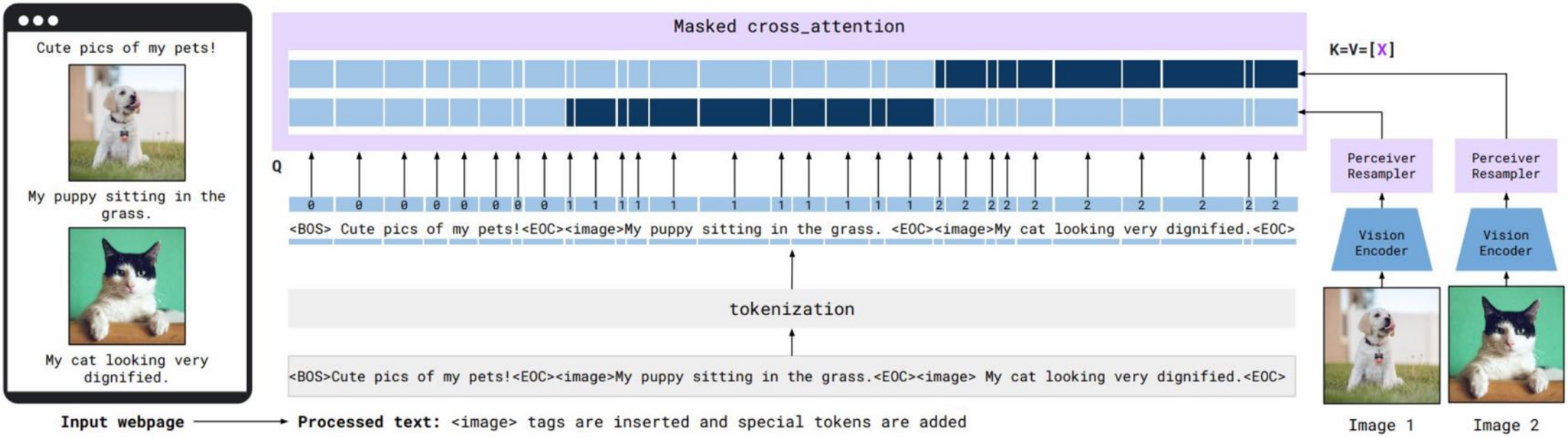
Deal with interleaved visual and text sequence

Each text token cross-attend to the image that precedes it in the interleaved sequence.

$$p(y|x) = \prod_{\ell=1}^{L} p(y_\ell | y_{<\ell}, x_{\leq \ell})$$



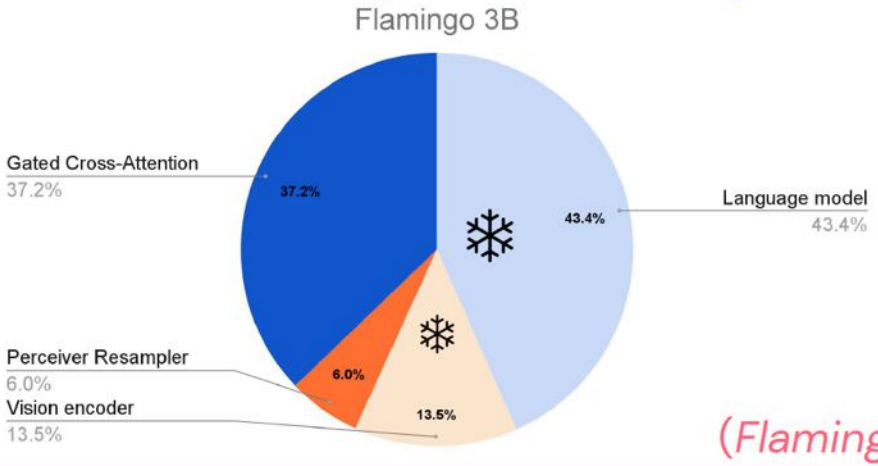
Overall architecture

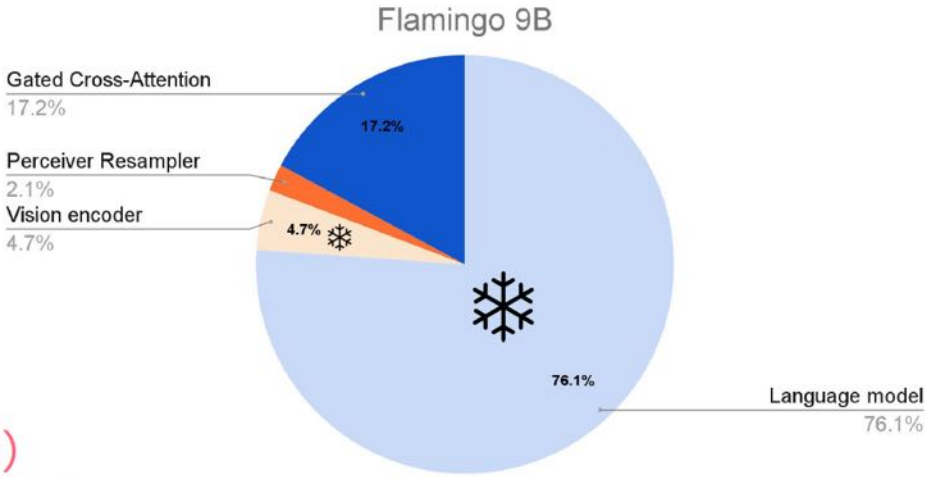


slide credit: Jean-Babtiste Alayrac

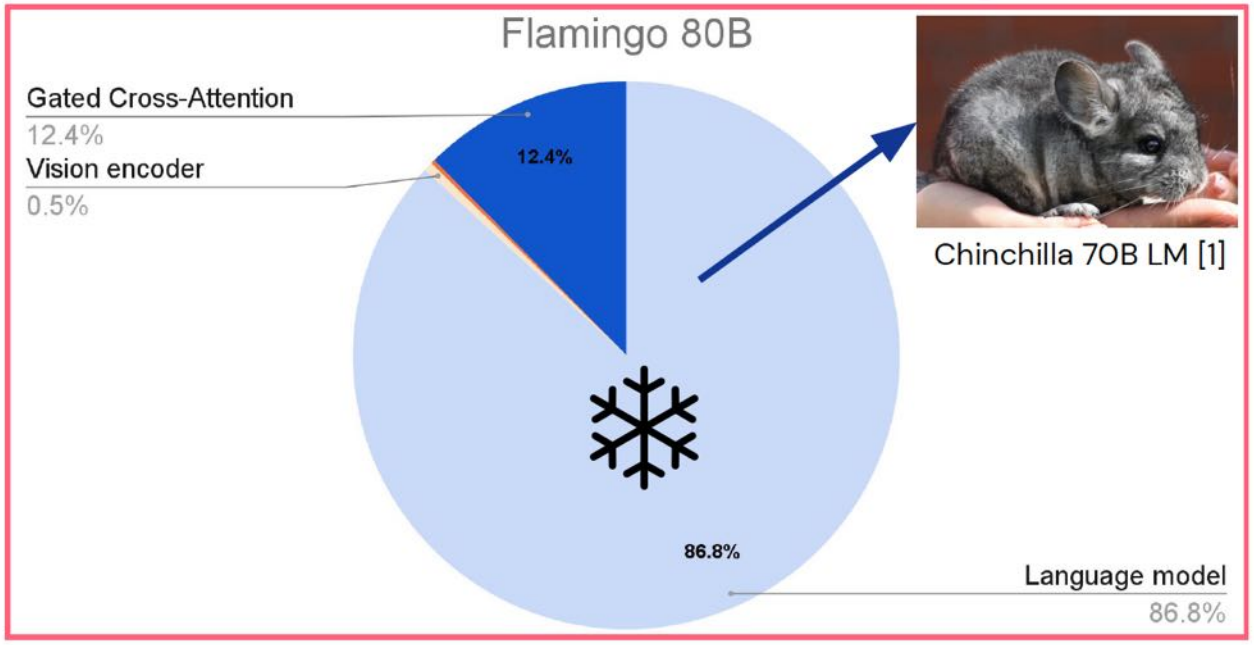
slide credit: Jean-Babtiste Alayrac

The Flamingo family





(*Flamingo* in short)



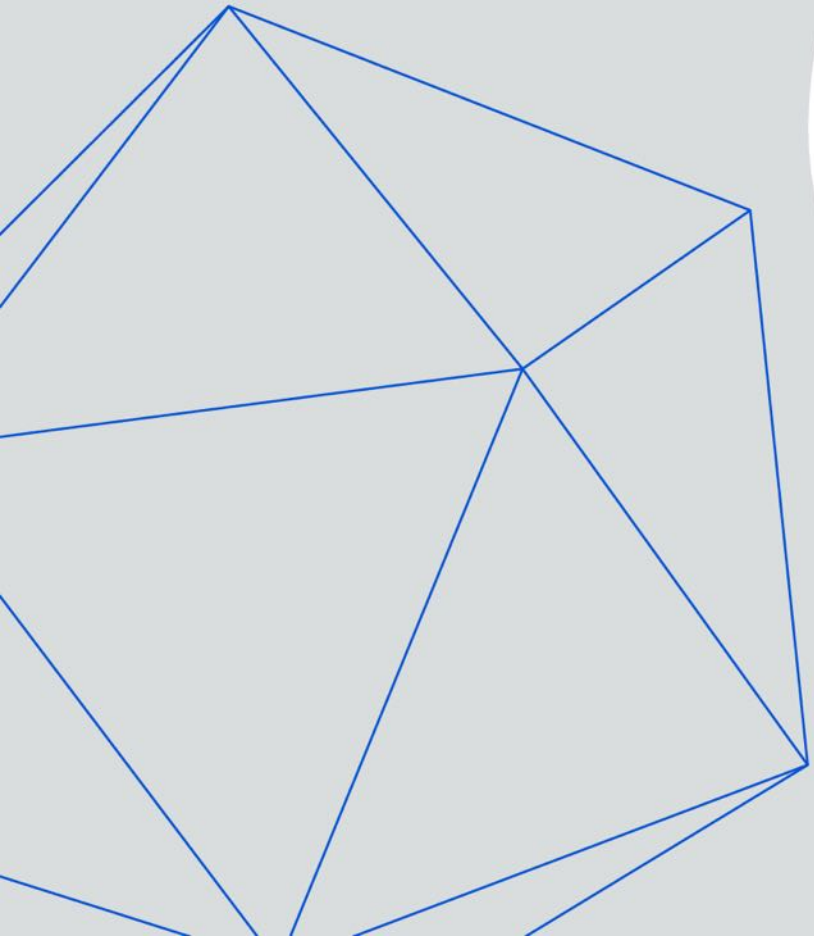
- Vision encoder (NFNet-F6) size fixed.
- Resampler size fixed.
- Focus on scaling the frozen language model.

[1] Hoffmann et al., Training Compute-Optimal Large Language Models, 2022

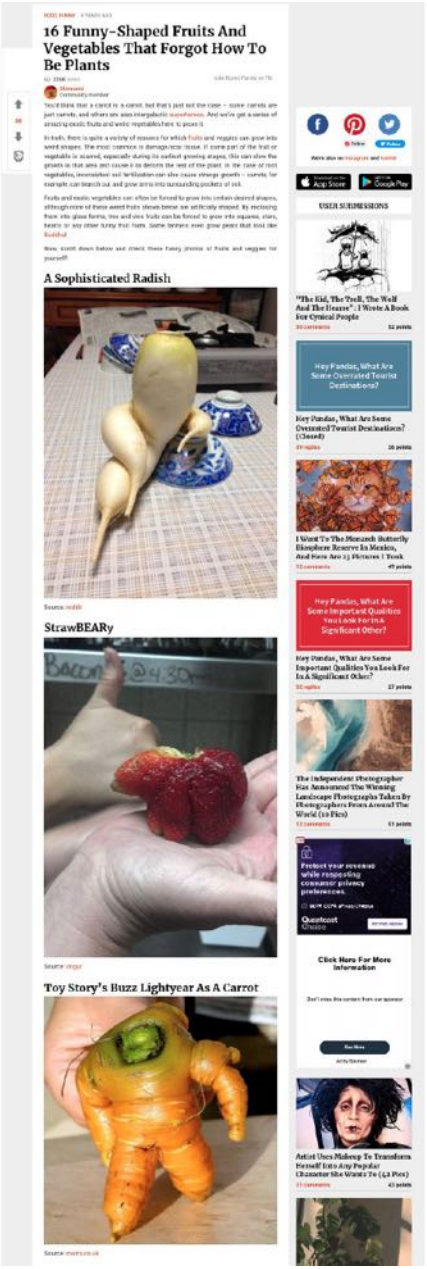
slide credit: Jean-Babtiste Alayrac

DeepMind

Training data



Flamingo training data



M3W: Massive MultiModal Web Dataset

slide credit: Jean-Babtiste Alayrac

44M scraped webpages with interleaved text and images.

180M images in total. (4 on average per webpage)

Processing

```
16 Funny-Shaped Fruits And Vegetables That
Forgot How To Be Plants

You'd think that a carrot is a carrot, but
that's just not the case - some carrots are
just carrots, and others are also
intergalactic superheroes. [...]. Some
farmers even grow pears that look like
Buddha!

Now, scroll down below and check these
funny photos of fruits and veggies for
yourself!

A Sophisticated Radish

<IMAGE PLACEHOLDER 1>

StrawBEARy

<IMAGE PLACEHOLDER 2>

Toy Story's Buzz Lightyear As A Carrot

<IMAGE PLACEHOLDER 3>
```

<IMAGE PLACEHOLDER 1> →

<IMAGE PLACEHOLDER 2> →

<IMAGE PLACEHOLDER 3> →

slide credit: Jean-Babtiste Alayrac

Image / Video paired with Captions

An English bulldog standing on a skateboard.

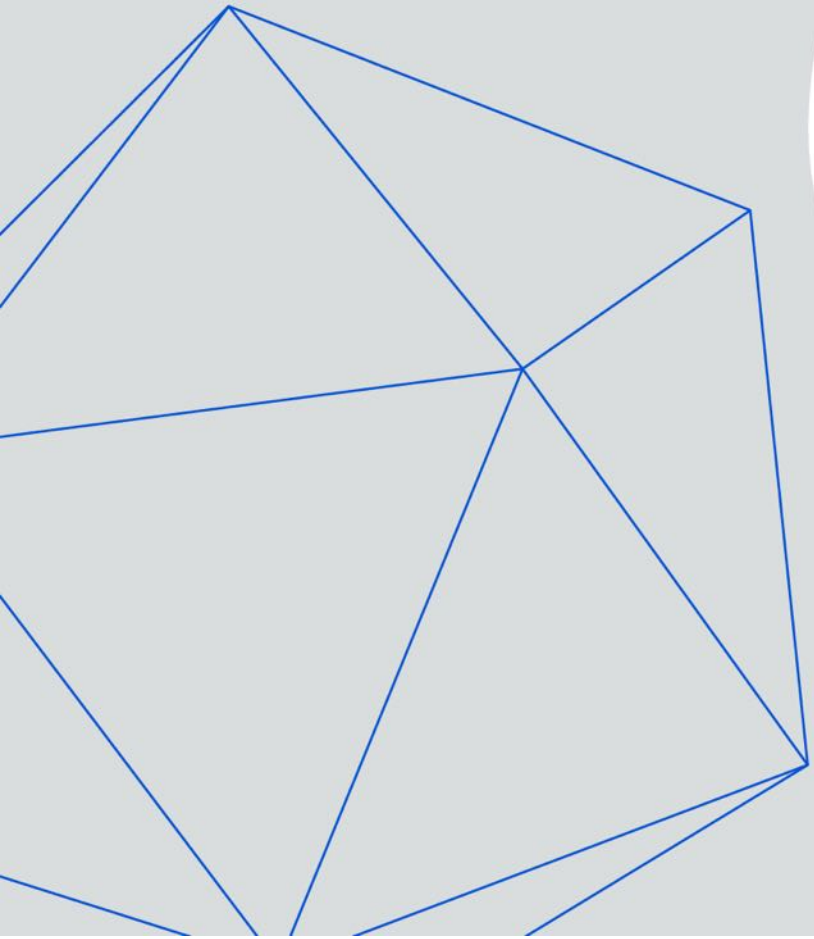
Image-Text pairs (2B examples)

A little girl playing with a flashlight.

Video-Text pairs (27M examples)

slide credit: Jean-Babtiste Alayrac

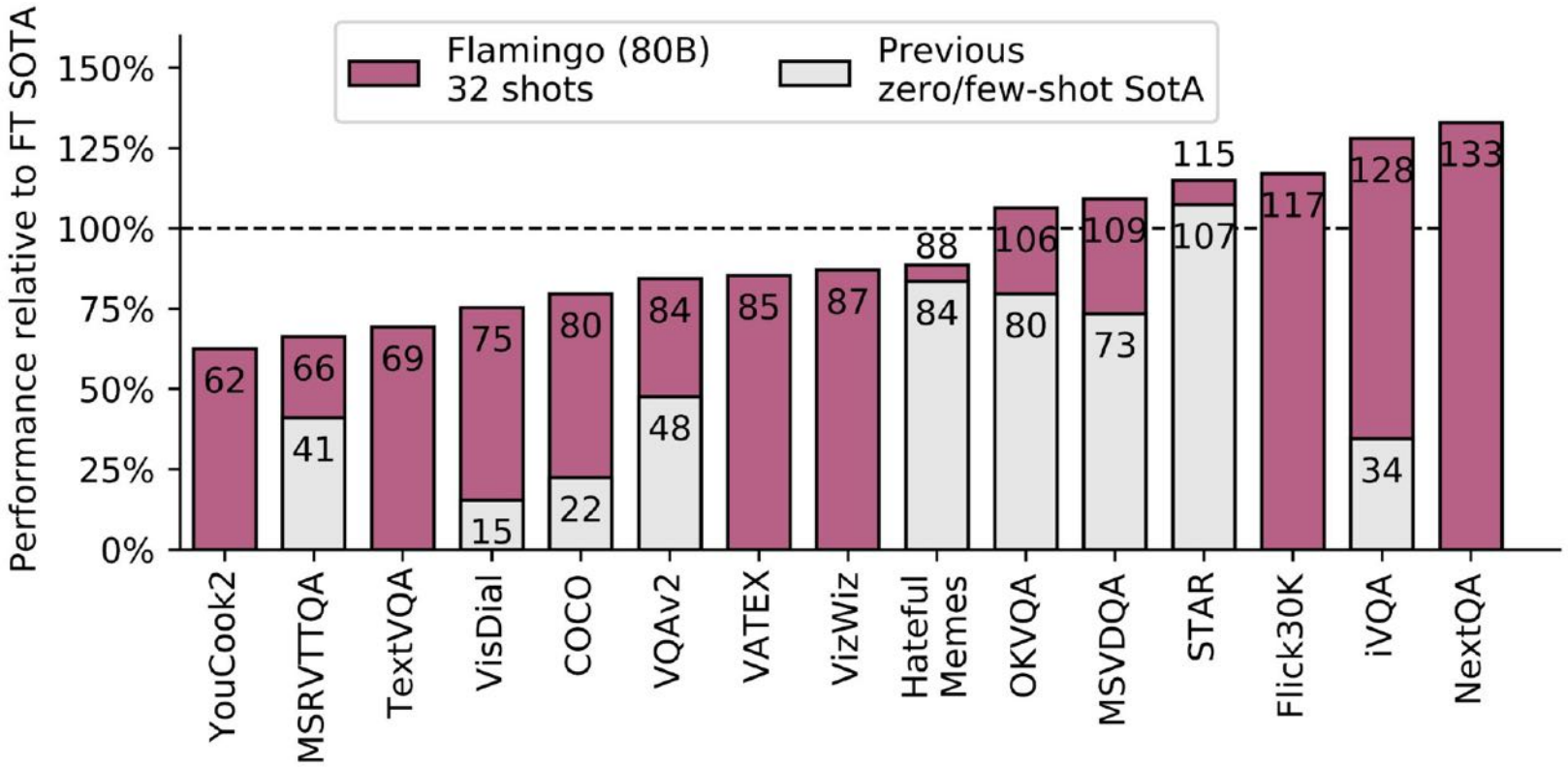
DeepMind



Experiments Results

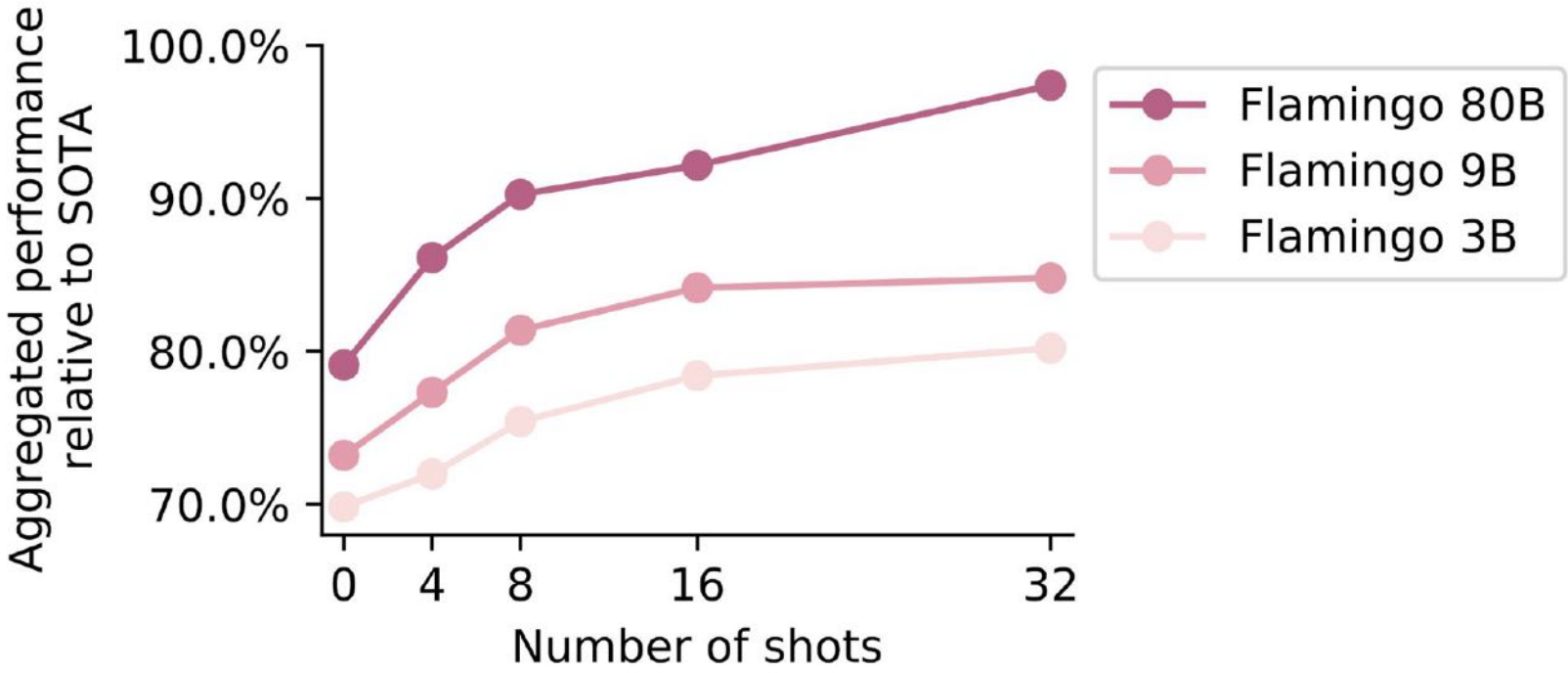
slide credit: Jean-Babtiste Alayrac

Main results: Few-shot results



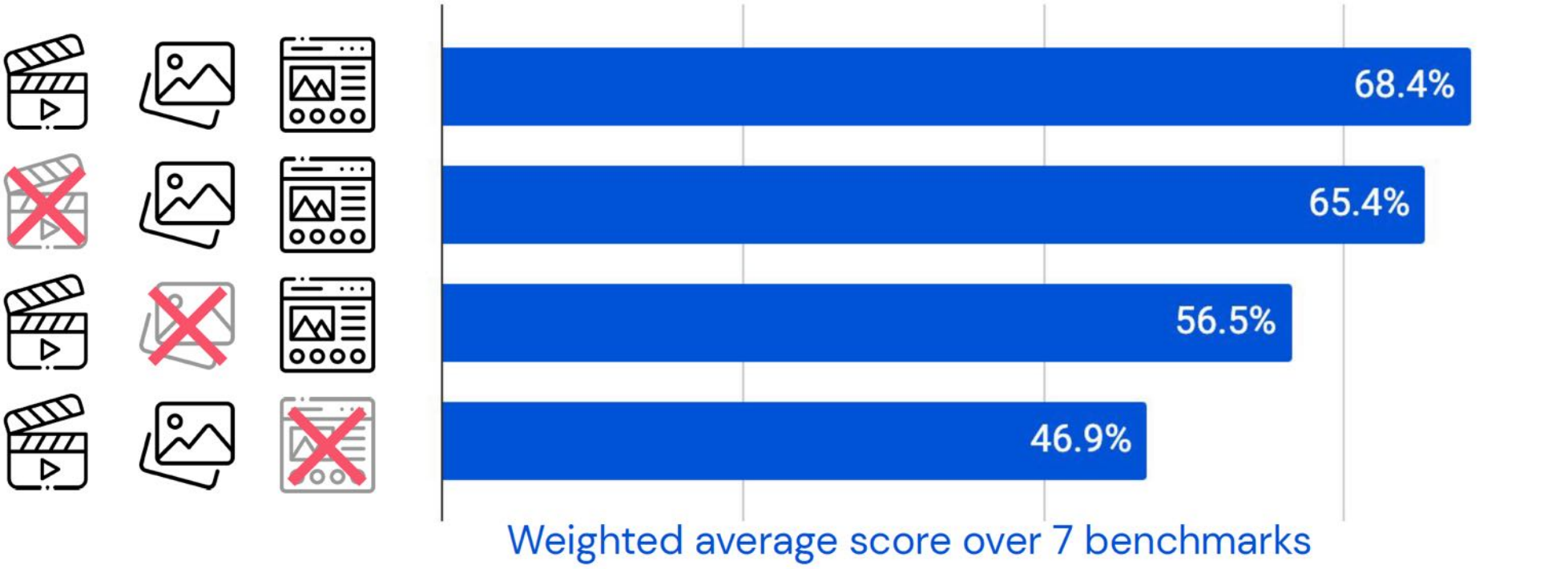
slide credit: Jean-Babtiste Alayrac

Scaling model size / number of shot



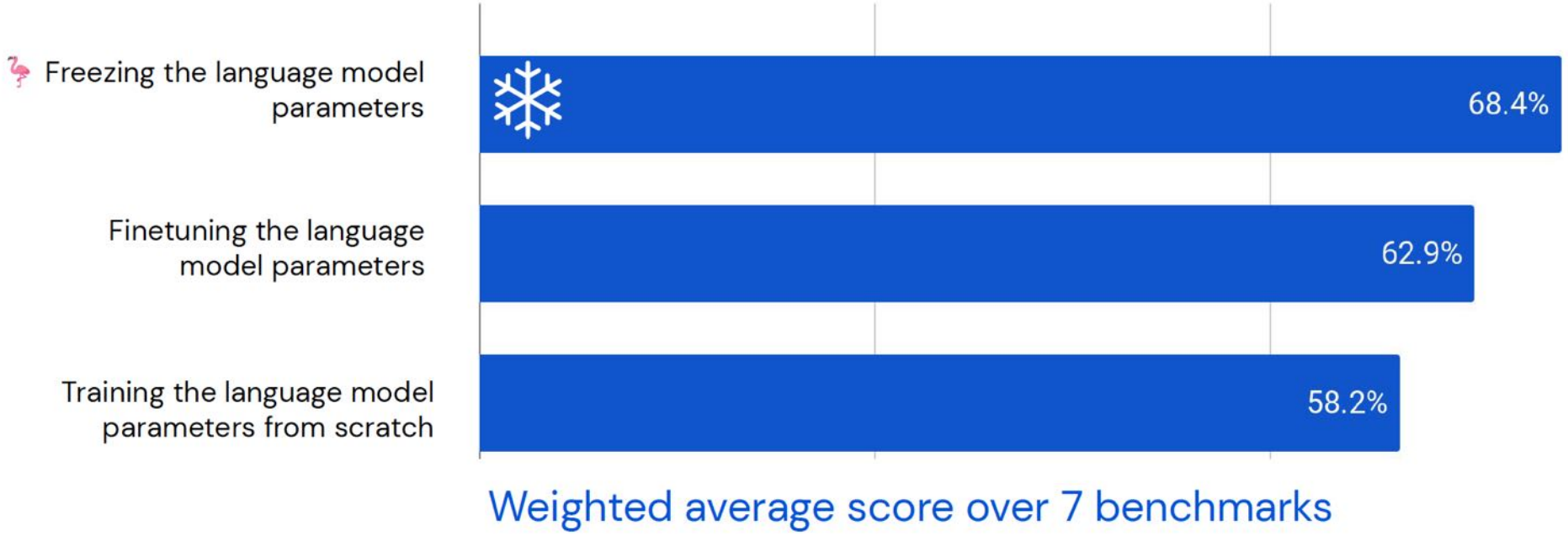
slide credit: Jean-Babtiste Alayrac

On the importance of the training data



slide credit: Jean-Babtiste Alayrac

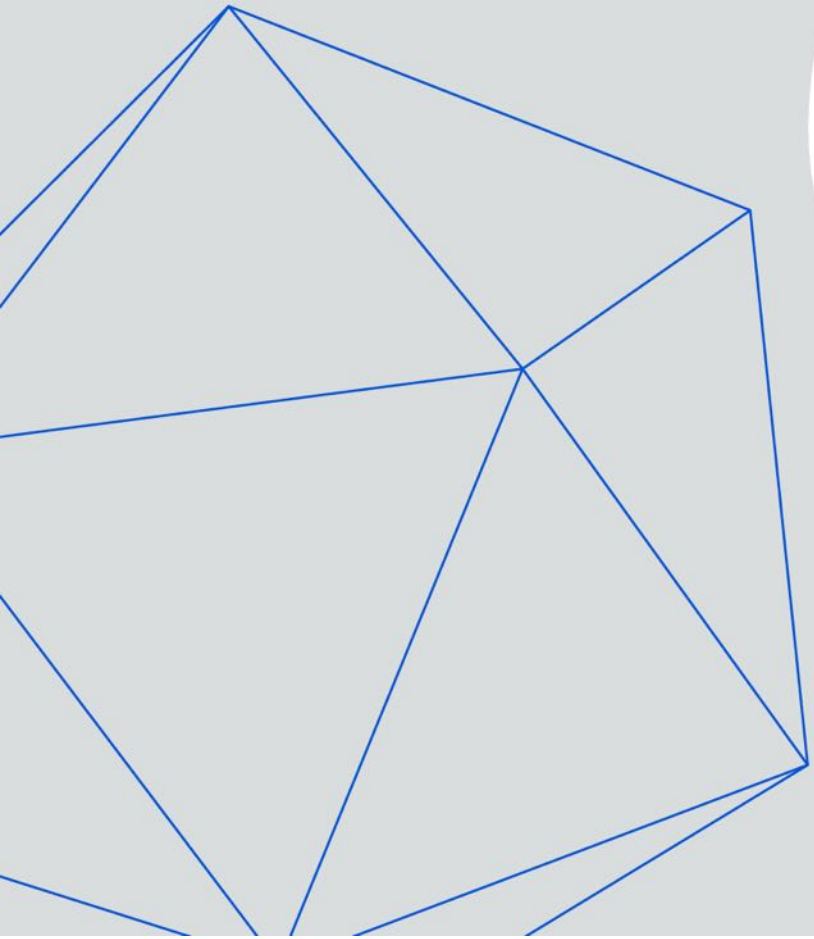
On the importance of freezing the language model



Freezing the language model lead to the best results while also being cheaper to train.

slide credit: Jean-Babtiste Alayrac

DeepMind



Qualitative examples

slide credit: Jean-Babtiste Alayrac

Input Prompt	 Question: What do you think the capacities of these are? Answer:	 Question: What is odd about this image? Explain why it is unusual. Answer:	 Question: What country is this? Why do you think so? Answer:
Completion	**The floppy disk is 1.44MB and the CD is 700MB.**	**The image is odd because the elephant is in the back of the truck. It is unusual because elephants are not usually transported in the back of a truck.**	**It is Canada. I think so because the flag is the Canadian flag.**
Input Prompt	 Question: What nutrients is the fruit in the image rich in? Answer:	 Question: What is the purpose of this? Answer:	 Question: What season is this? Explain why. Answer:
Completion	**Vitamin C, potassium, and fiber.**	**To keep the mosquitoes out of the bed.**	**This is winter season. The Christmas tree and the presents are the main reasons why it is winter.**

slide credit: Jean-Babtiste Alayrac

Which city is this?

This is a street sign in New York City.

Which street corner is it?

It is the intersection of Madison Avenue and East 42nd Street.

Which subway station is nearby?

The closest subway station is Grand Central Station.

How about the nearest park?

The nearest park is Bryant Park.

There is another sign. What does it say?

It says "One way".

What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

when you are fully grown but you can't resist using the playground equipment

What is funny about this meme?

This is a hamster in a slide. It is funny because the hamster is too big to fit in the slide.

Is the hamster enjoying that?

I don't think so.

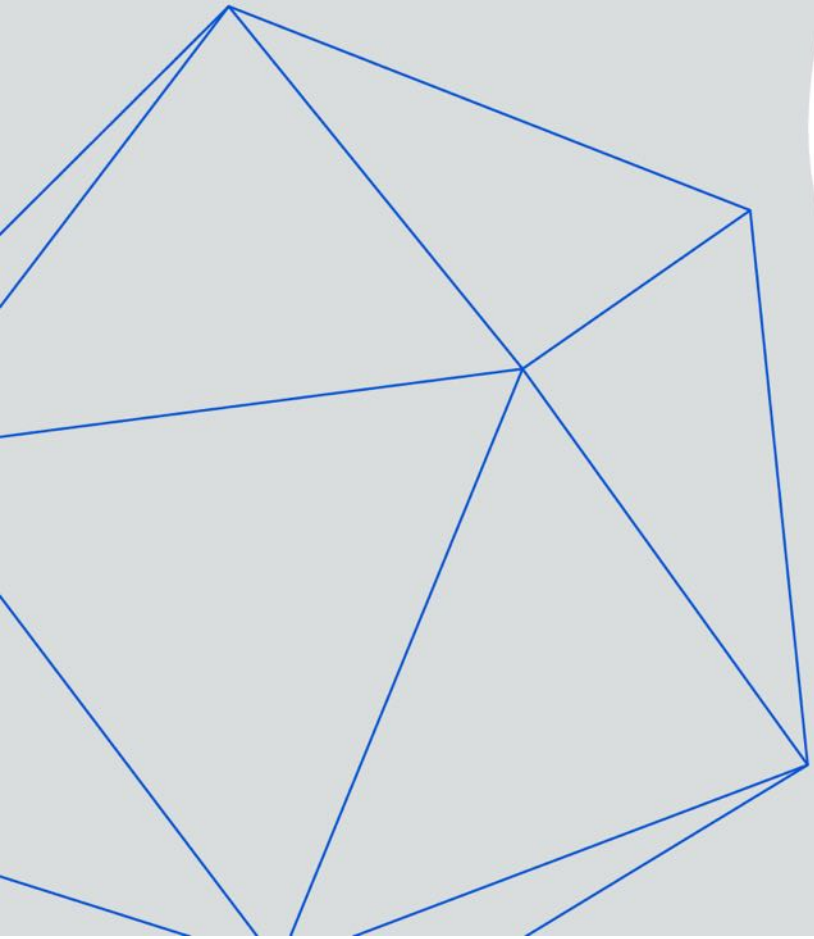
How is the hamster feeling?

The hamster is feeling frustrated.

slide credit: Jean-Babtiste Alayrac

DeepMind

Discussion



slide credit: Jean-Babtiste Alayrac

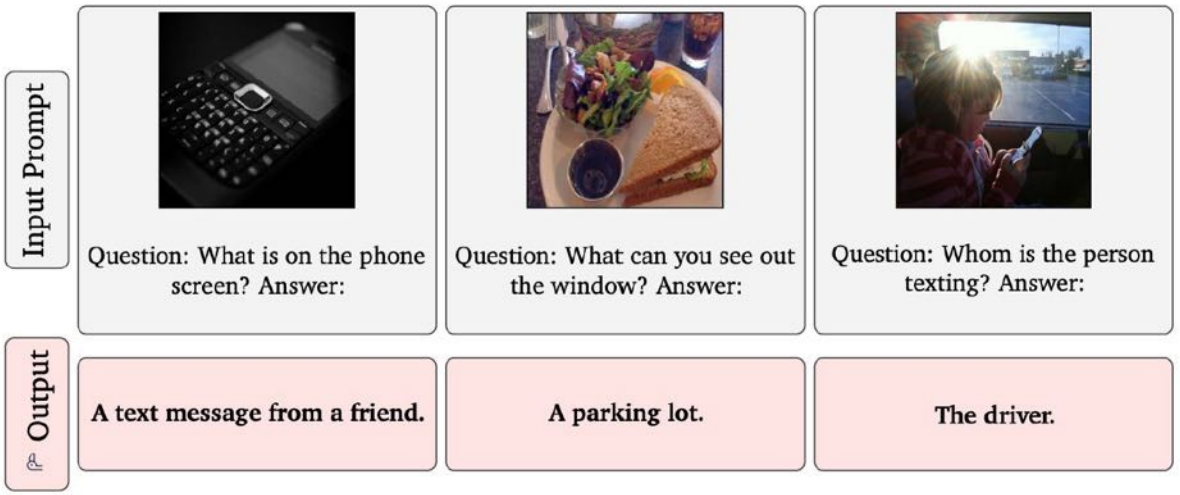
Limitations

- Performance on classification

slide credit: Jean-Babtiste Alayrac

Limitations

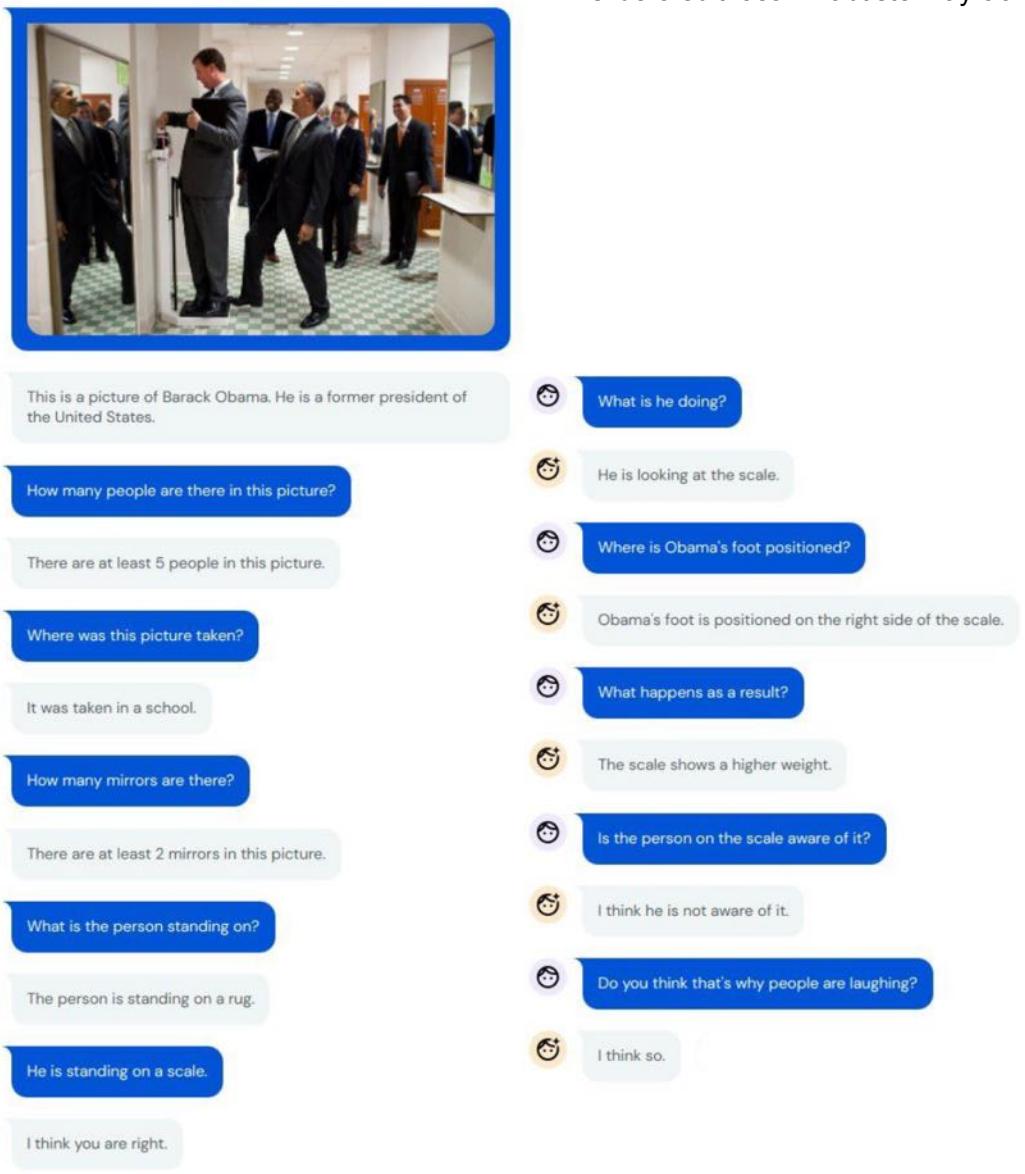
- Performance on classification
- Hallucinations



slide credit: Jean-Babtiste Alayrac

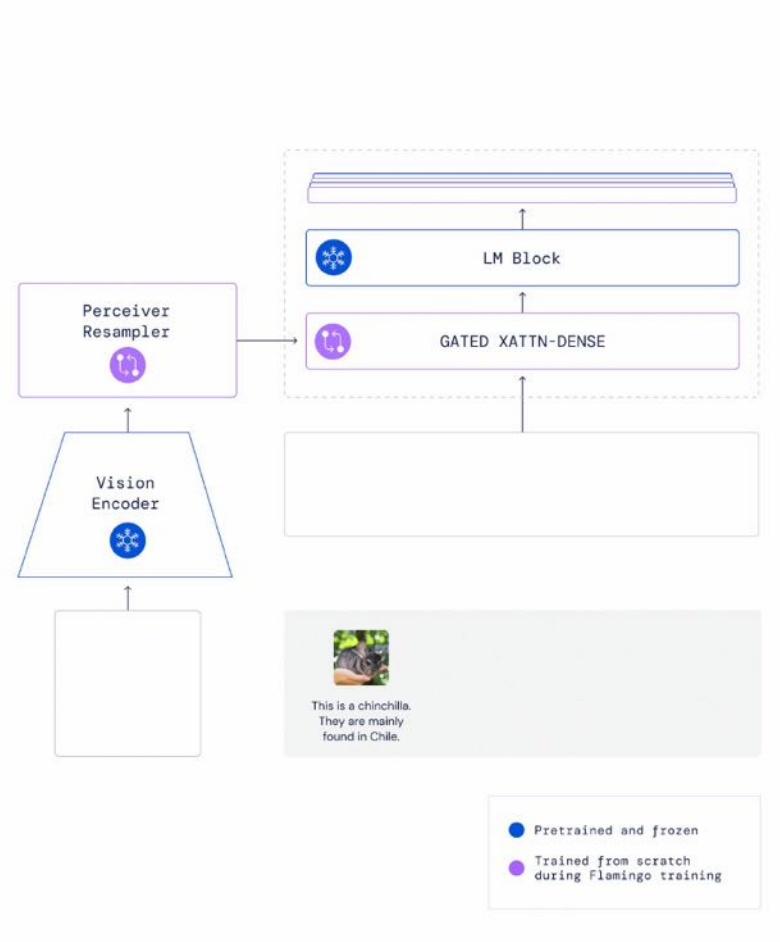
Limitations

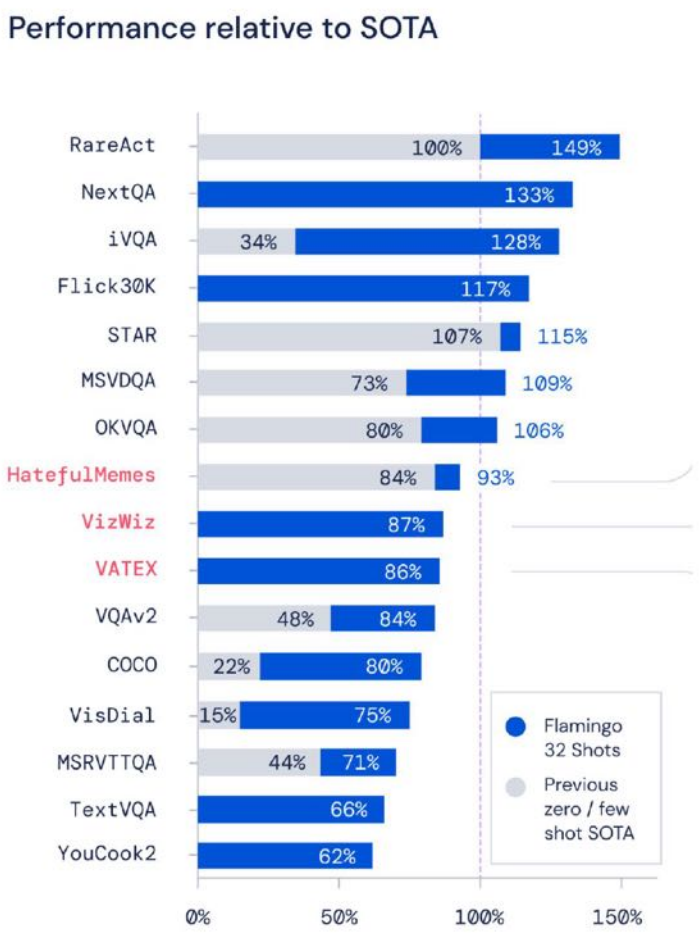
- Performance on classification
- Hallucinations
- Coarse spatial understanding
- Needs hinting to understand complex scenes



slide credit: Jean-Babtiste Alayrac

Summary and conclusion





Overview of Today's Lecture

- Continuation of Vision-Language Learning for Computer Vision
  - Recap: CLIP
  - Flamingo [neurips'22] - https://arxiv.org/abs/2204.14198

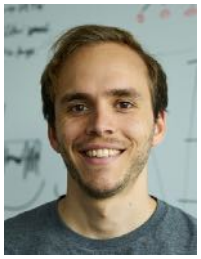
- Inherently Interpretable Neural Network
  - B-cos CNN [cvpr'22] - https://arxiv.org/abs/2205.10268
  - B-cos CNN & Vision transformers [arxiv'23] - https://arxiv.org/abs/2306.10898

max planck institut informatik

SIC Saarland Informatics Campus

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

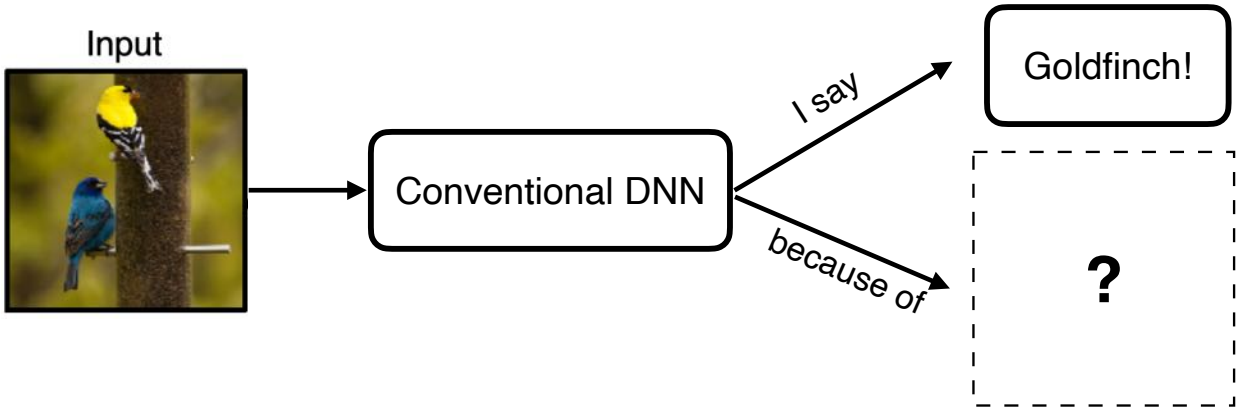
Moritz Böhle[1], Mario Fritz[2] and Bernt Schiele[1]
[1]Max Planck Institute for Informatics, [2]CISPA Helmholtz Center for Information Security



Motivation

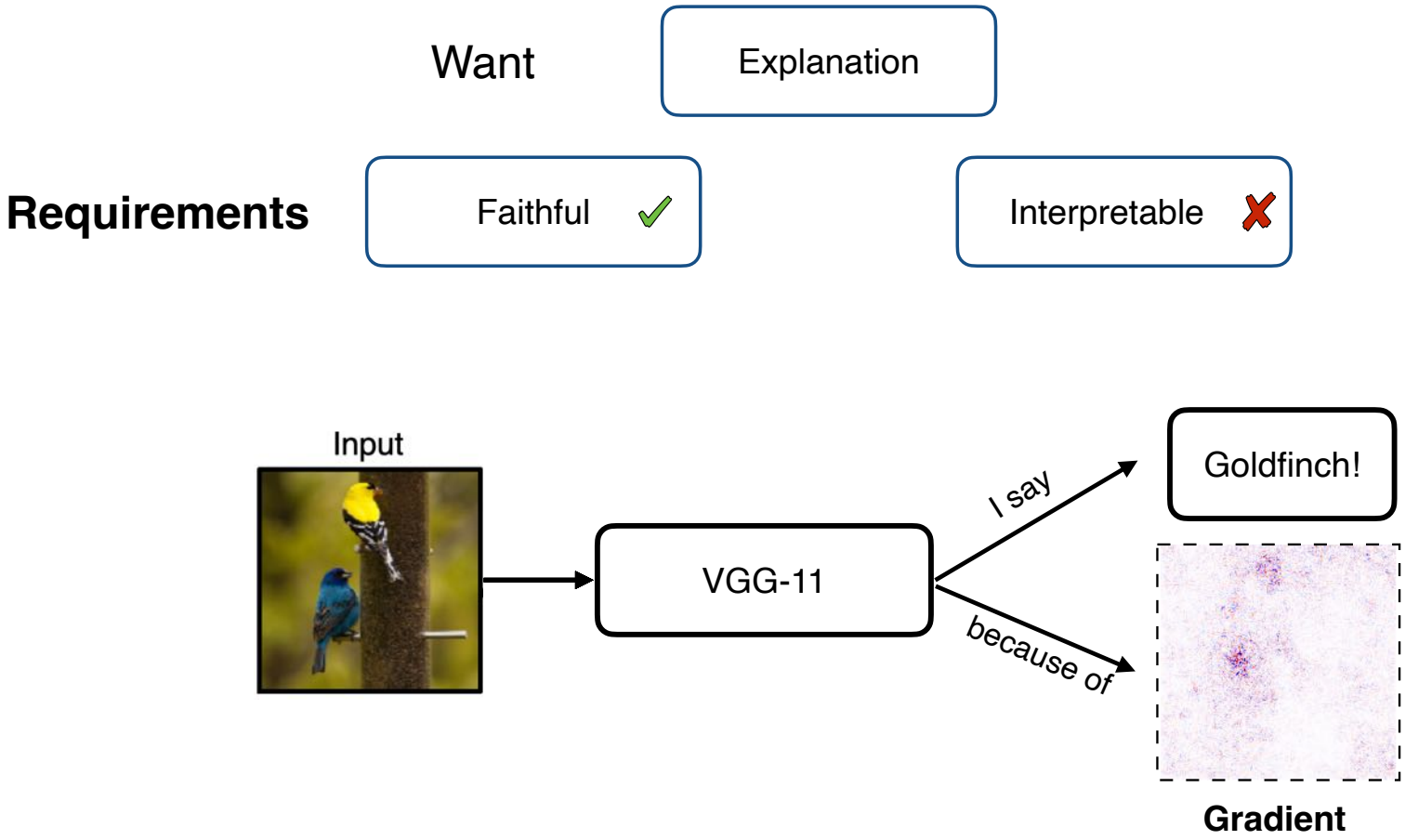
Want Explanation to understand better

Requirements Faithful Interpretable



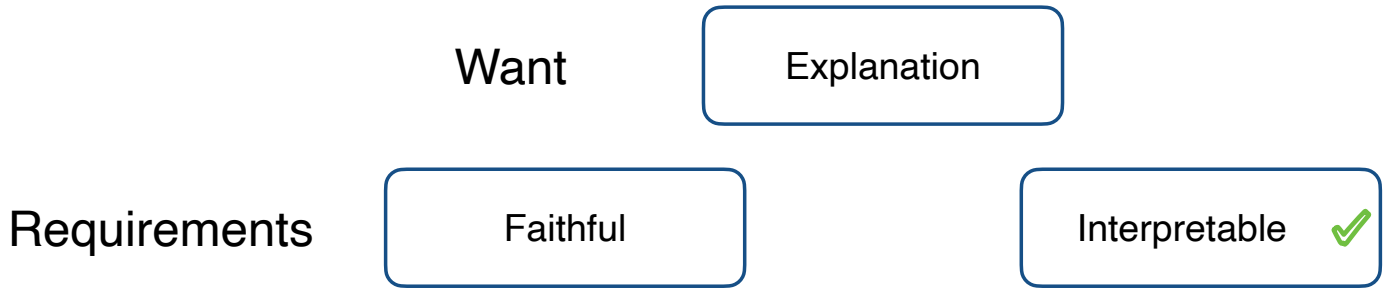
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

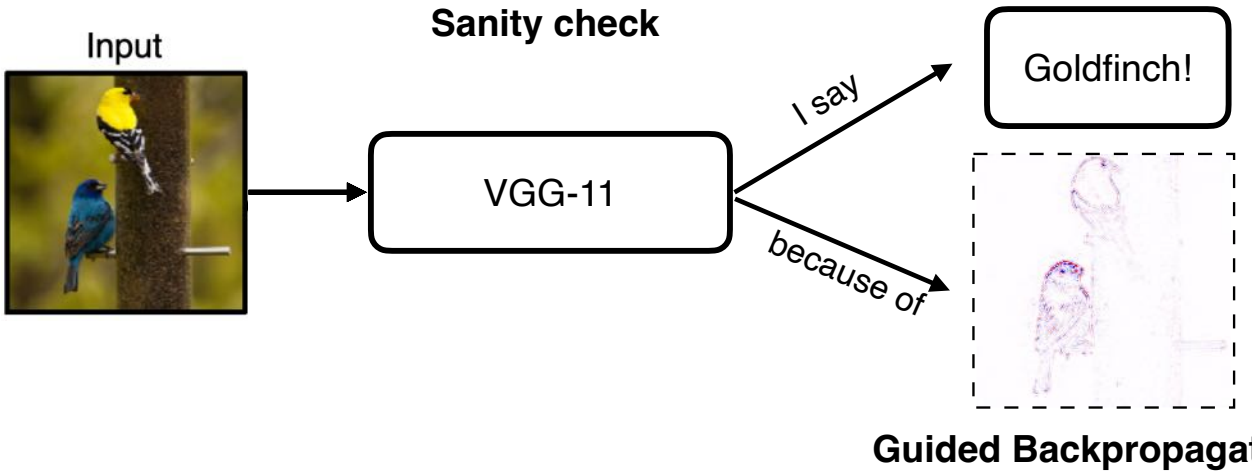
Motivation



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)

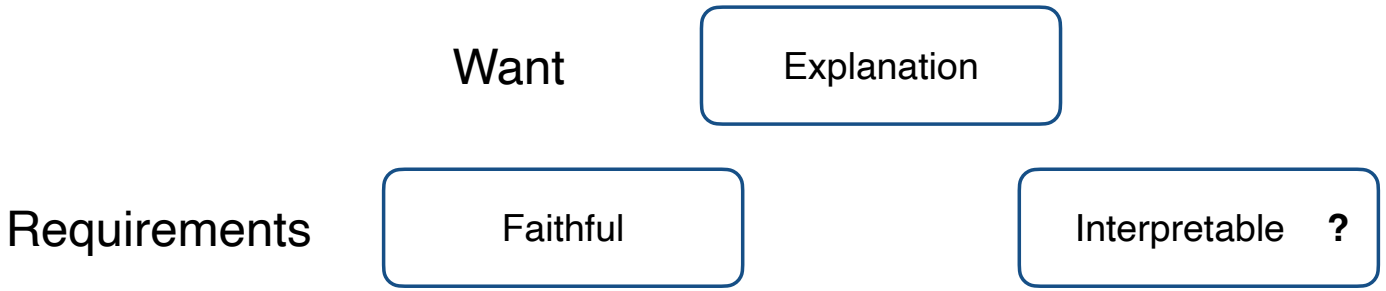
Motivation

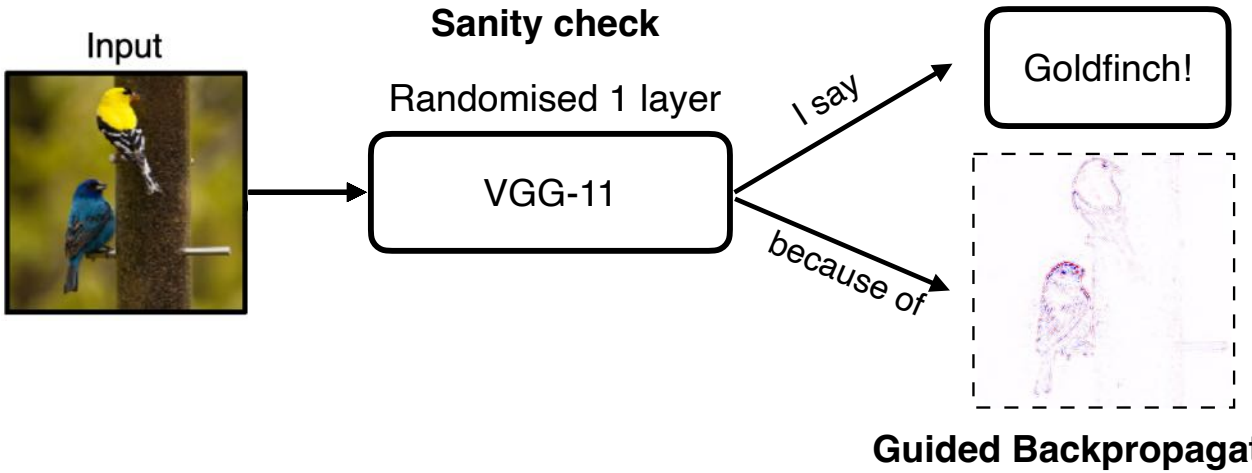




References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

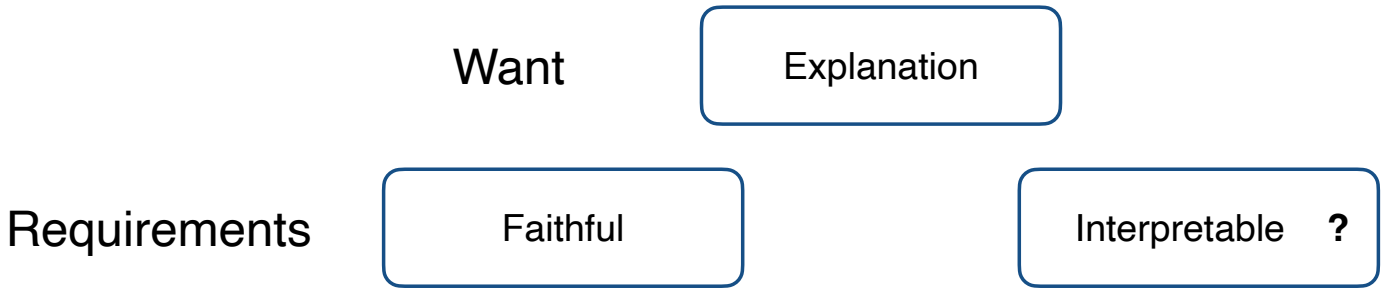
Motivation

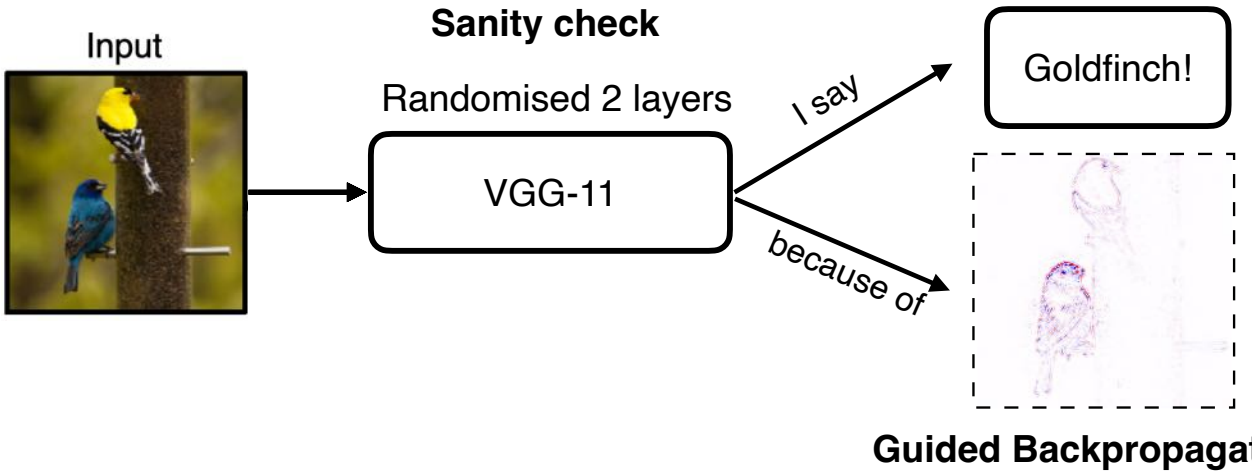




References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

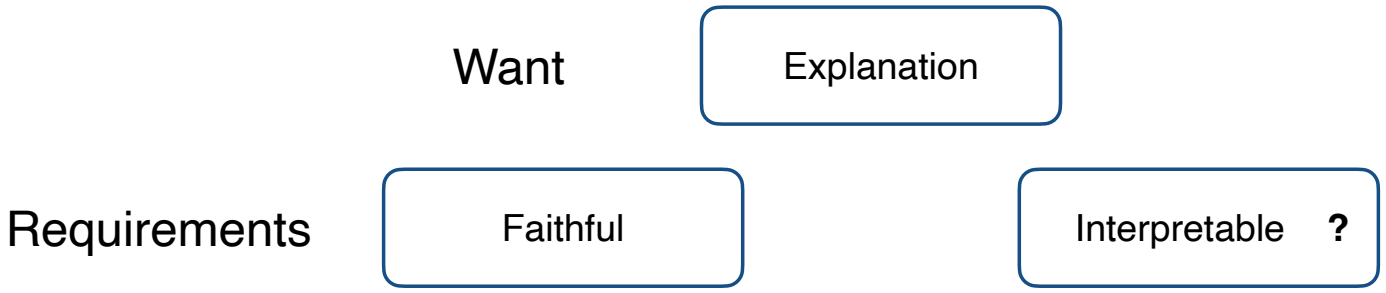
Motivation

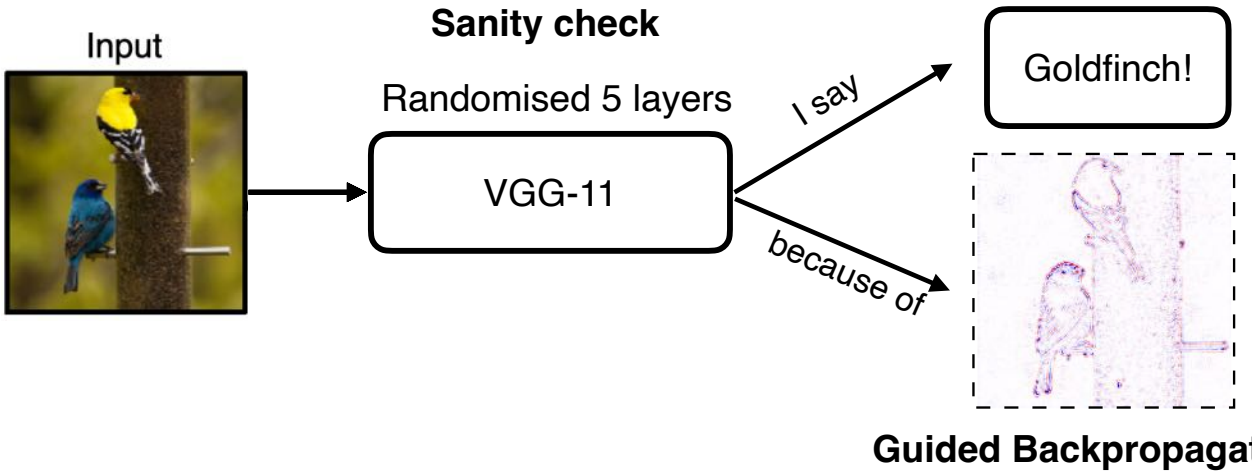




References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

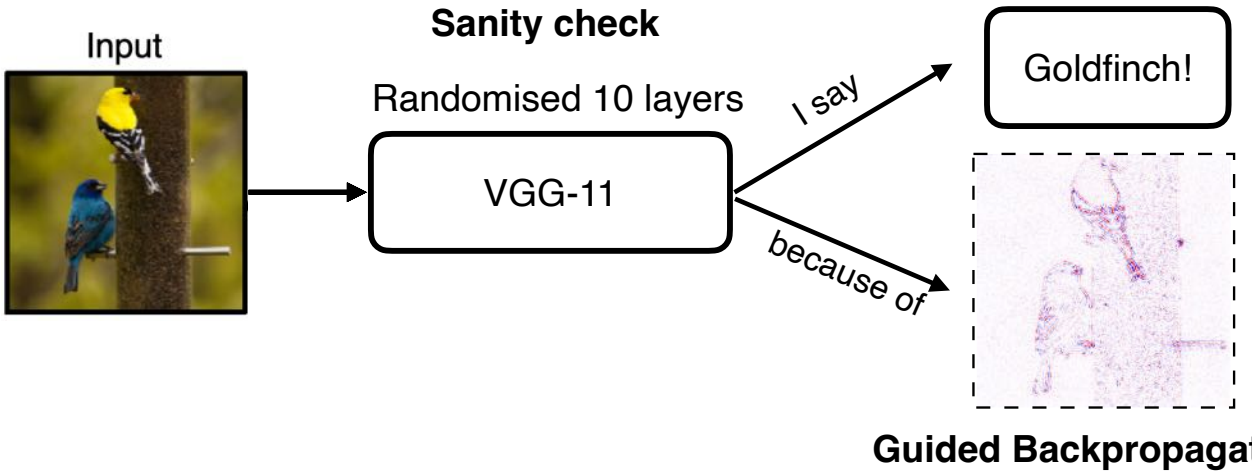
Motivation





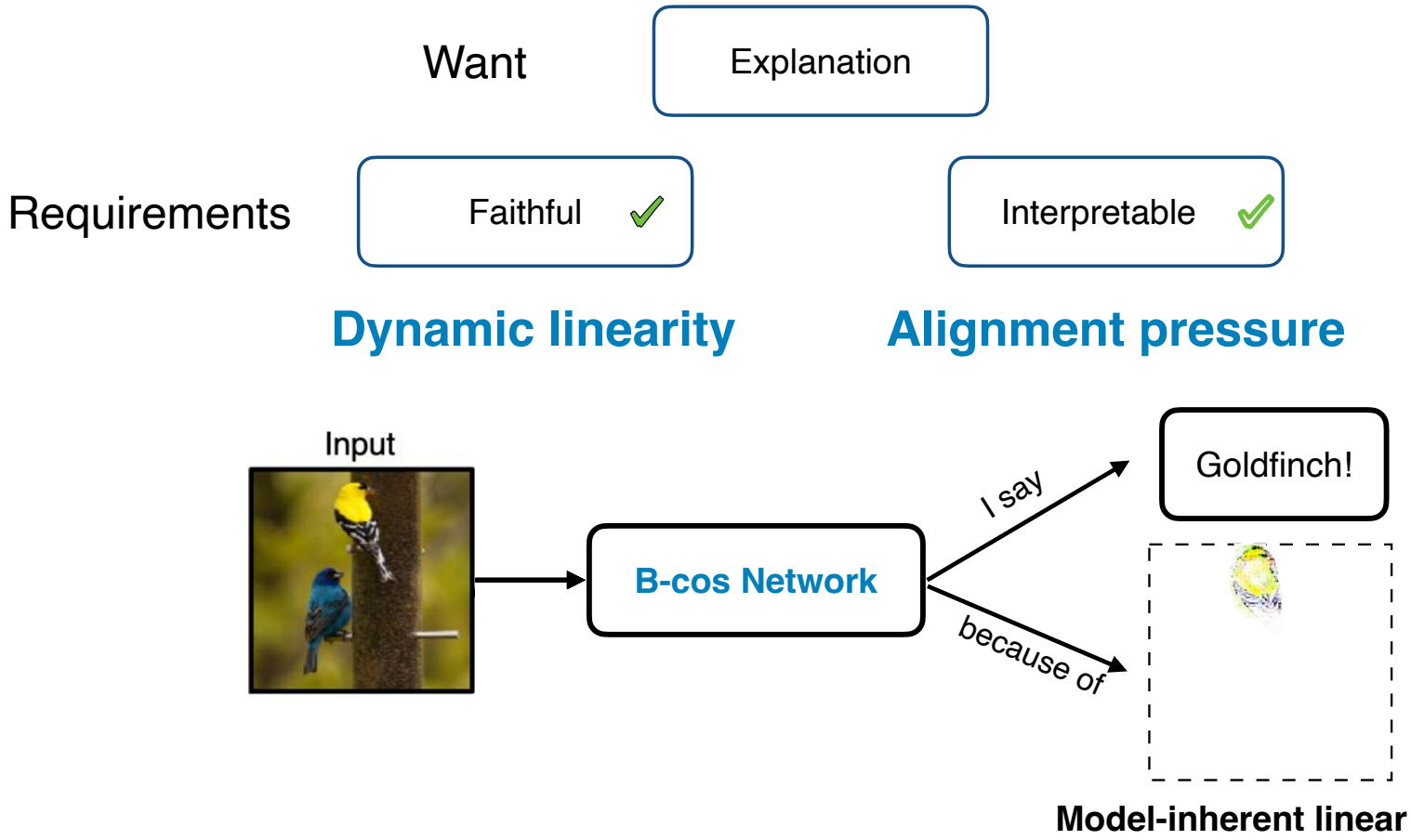
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation



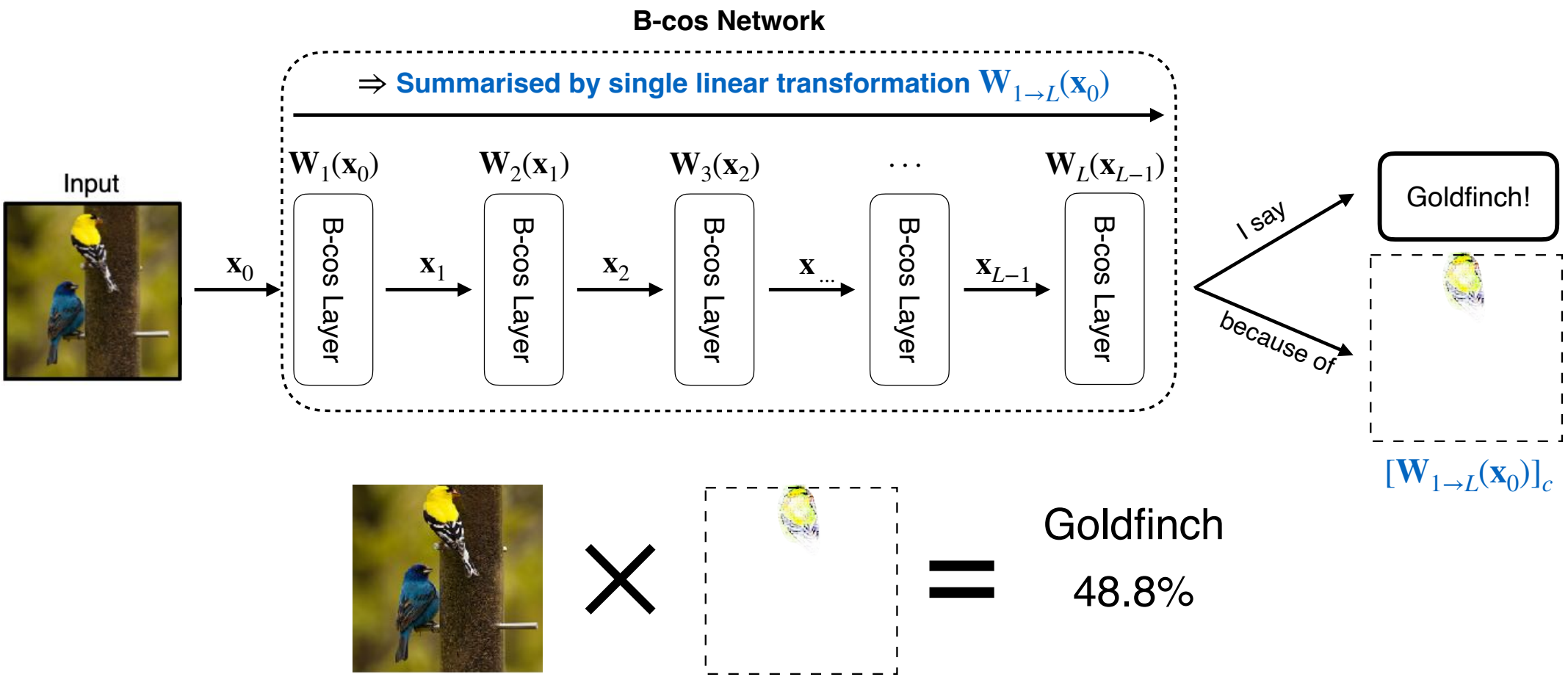
References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

Motivation: we aim for Inherent Interpretability

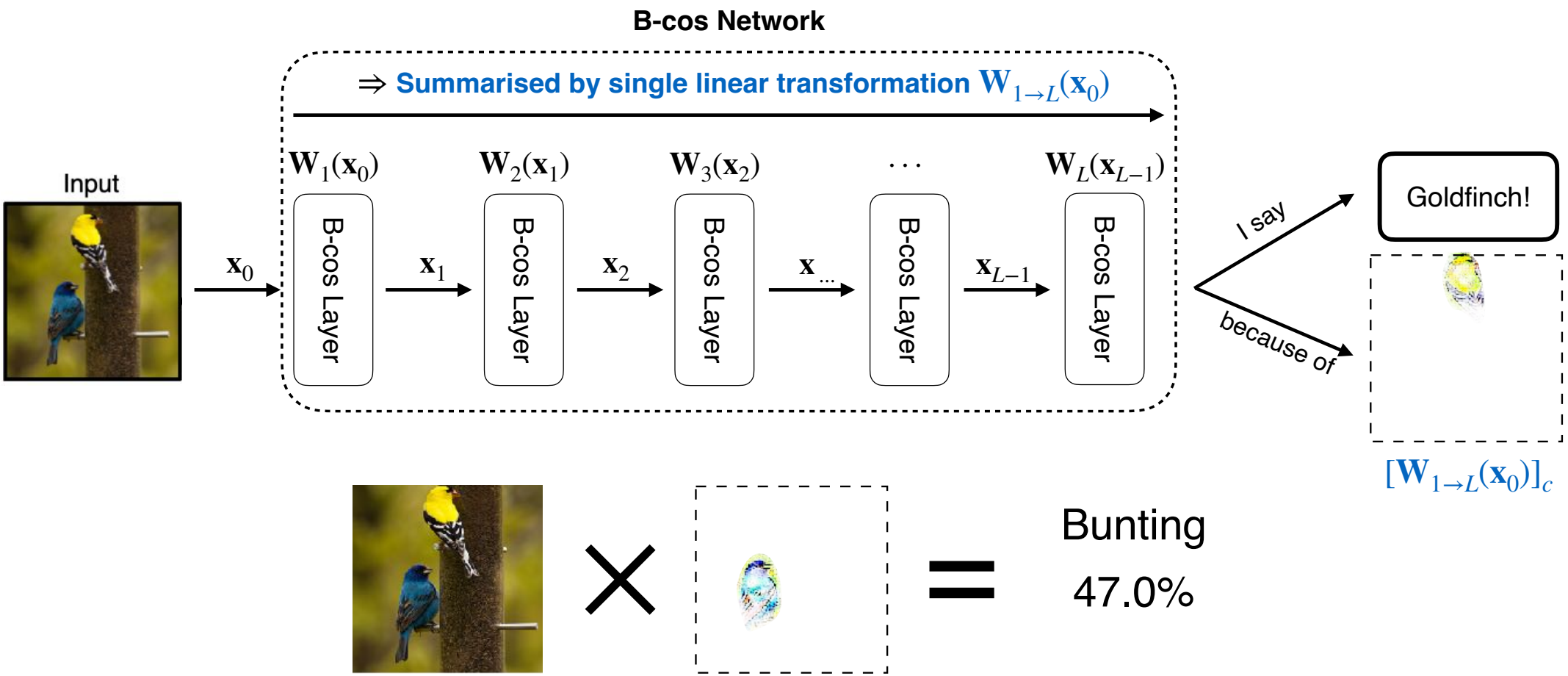


References: 'Requirements' (Gilpin et al., 2018), VGG-11 (Simonyan et al., 2014), Grad (Baehrens et al., 2010), Guided Backpropagation (Springenberg et al., 2014), Sanity check (Adebayo et al., 2018)

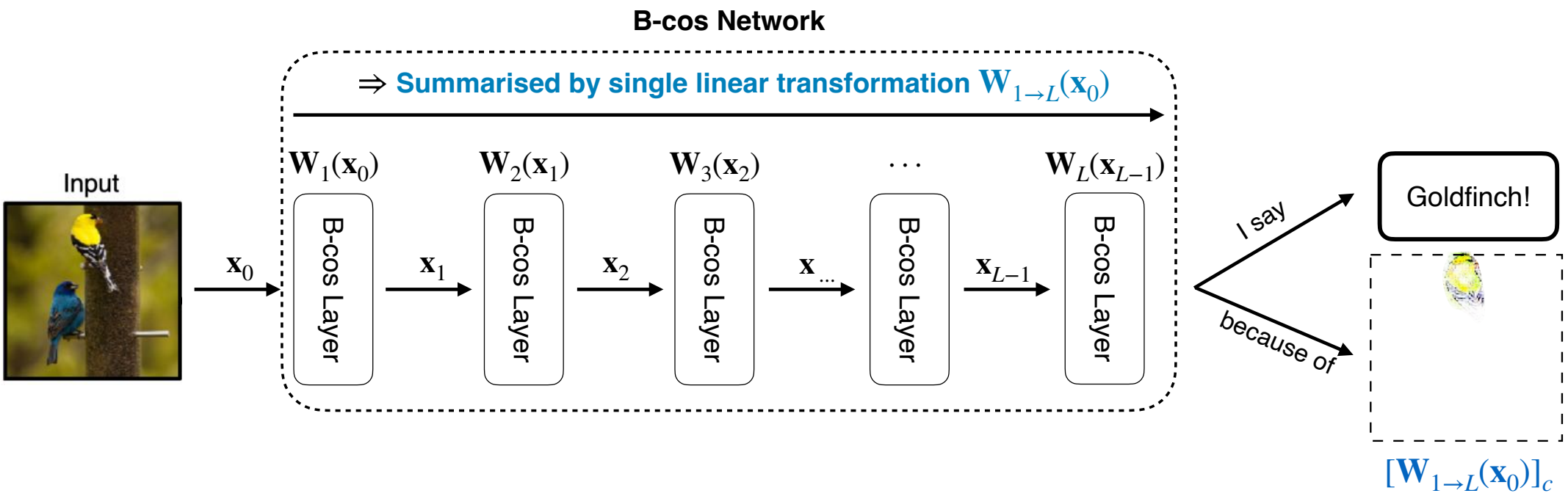
B-cos Networks: Dynamic Linearity



B-cos Networks: Dynamic Linearity



B-cos Networks: Dynamic Linearity



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

slide credit: Moritz Böhle

B-cos transformation vs. linear transformation

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

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

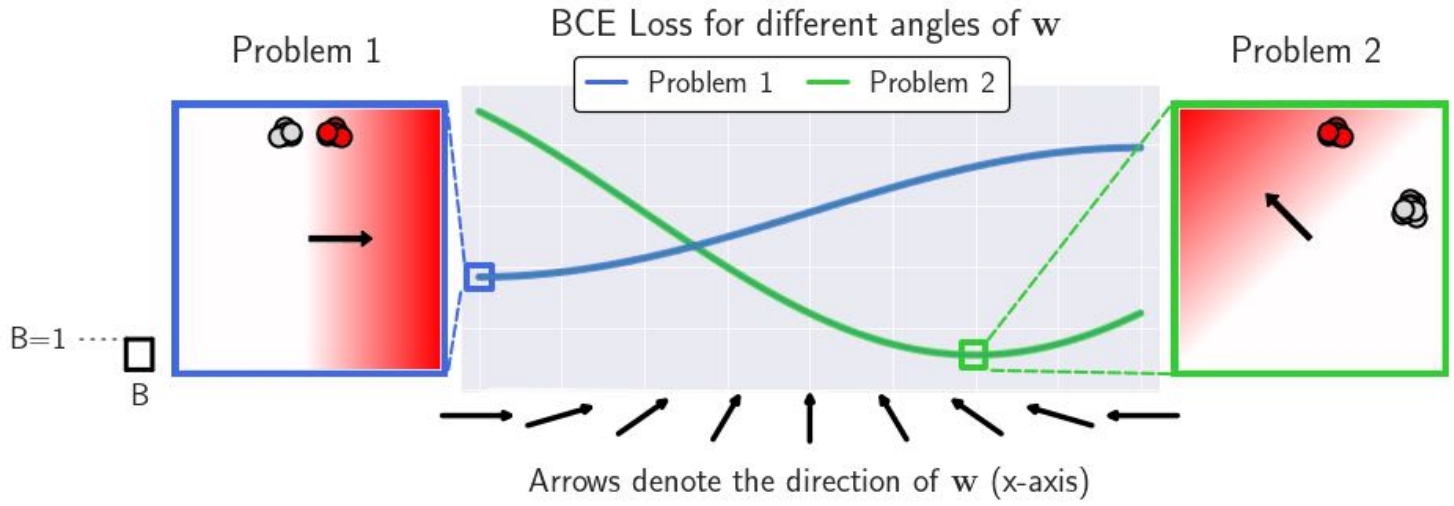
slide credit: Moritz Böhle

B-cos transformation vs. linear transformation

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

$$= \left(\ |\cos(\mathbf{x}, \mathbf{w})|^{B-1} \times \widehat{\mathbf{w}}\ \right)^{T} \mathbf{x} = \mathbf{w}^{T}(\mathbf{x})\,\mathbf{x}$$

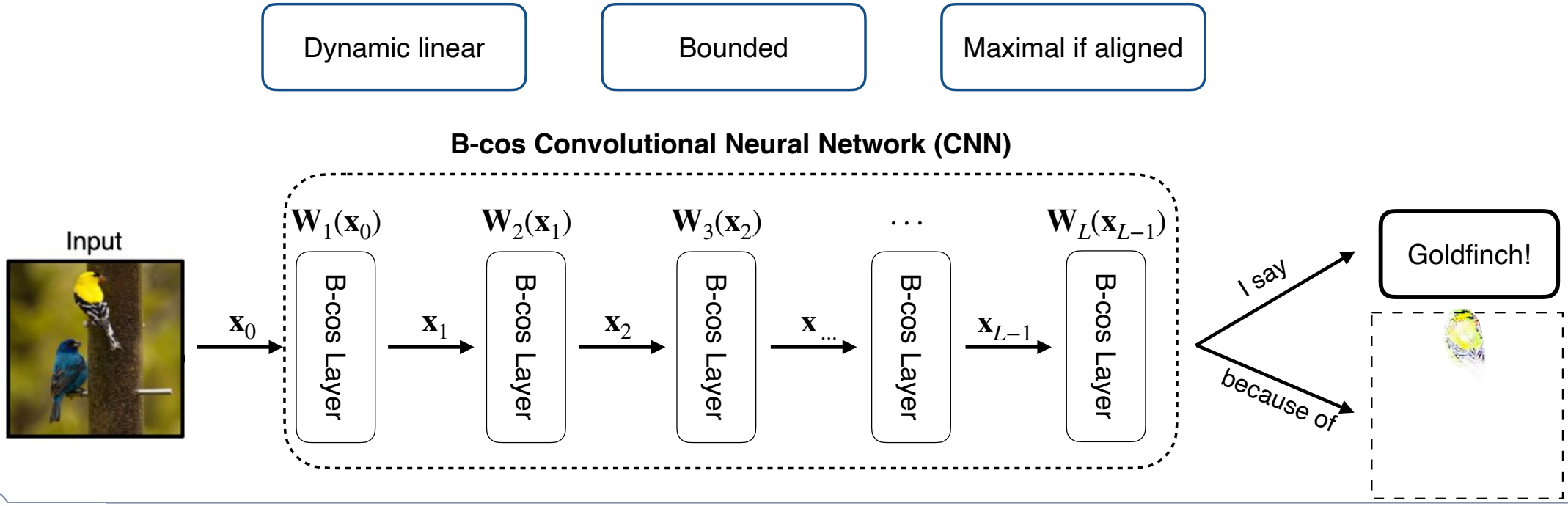
- Dynamic linear
- Bounded
- Maximal if aligned



slide credit: Moritz Böhle

Alignment pressure: B-cos networks

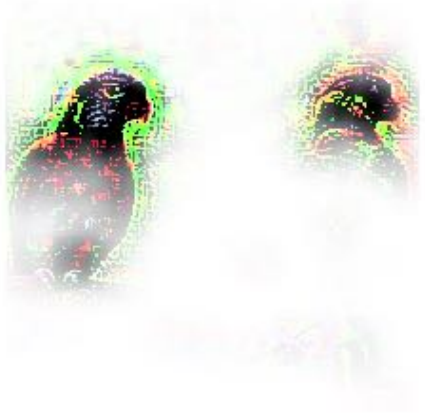
- **Properties inherited by B-cos network**
  - ⇒ Output maximisation induces alignment pressure



slide credit: Moritz Böhle

Results

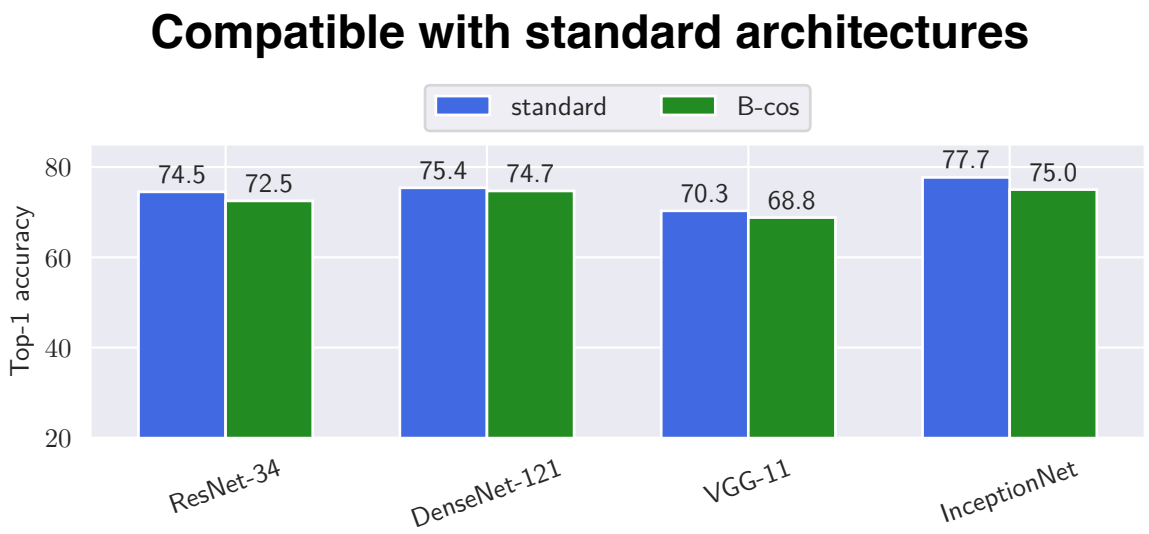
- Quantitative results on ImageNet
  - Classification accuracy
  - Interpretability scores
- Some qualitative results on ImageNet
  - Explaining class logits
  - explaining intermediate neurons



slide credit: Moritz Böhle

ImageNet results

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

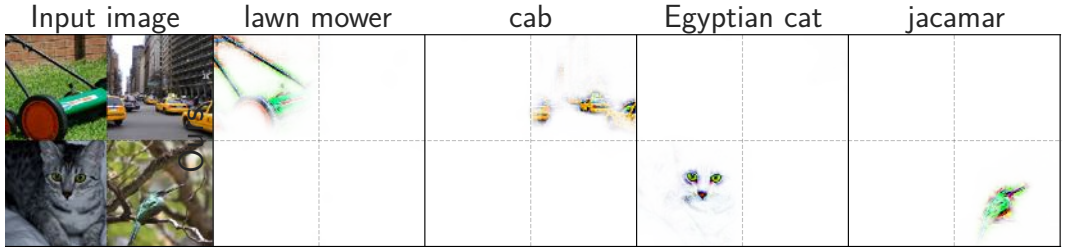


Measuring Interpretability via Grid Pointing Game

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

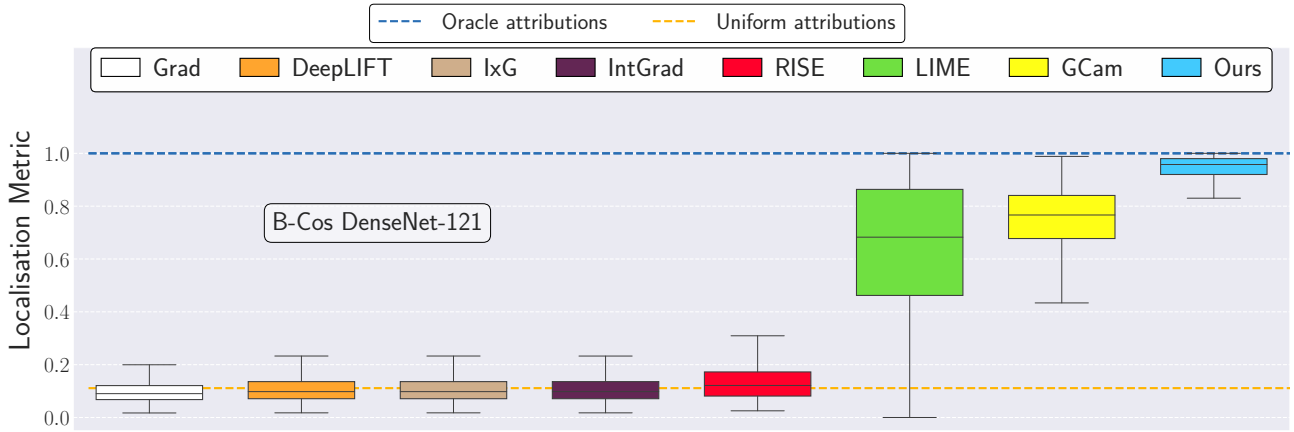
- In particular:
  - evaluate models on synthetic image grid
  - measure how well an explanation *localises* the correct image grid

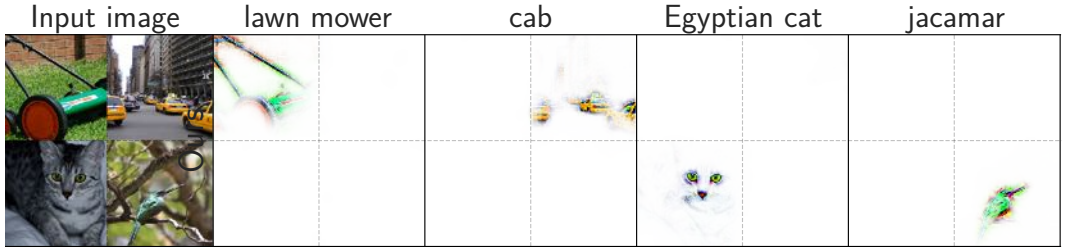
  (score $s = \frac{A_i^+}{\sum_j A_j^+}$ with A_i^+ the positive attribution to subimage i)



ImageNet results

High interpretability

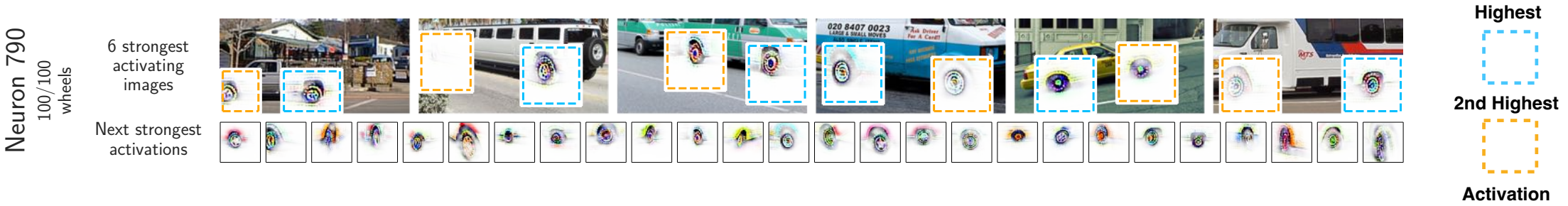




Gradient (Baehrens (2010)), DeepLIFT (Shrikumar (2017)), Input x Gradient (cf. Adebayo (2018)), IntGrad (Sundararajan (2017)), RISE (Petsiuk, 2018), LIME (Ribeiro, 2016), GradCam (Ramprasaath et al. (2017))

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

Visualisations: intermediate neurons



Interim Summary

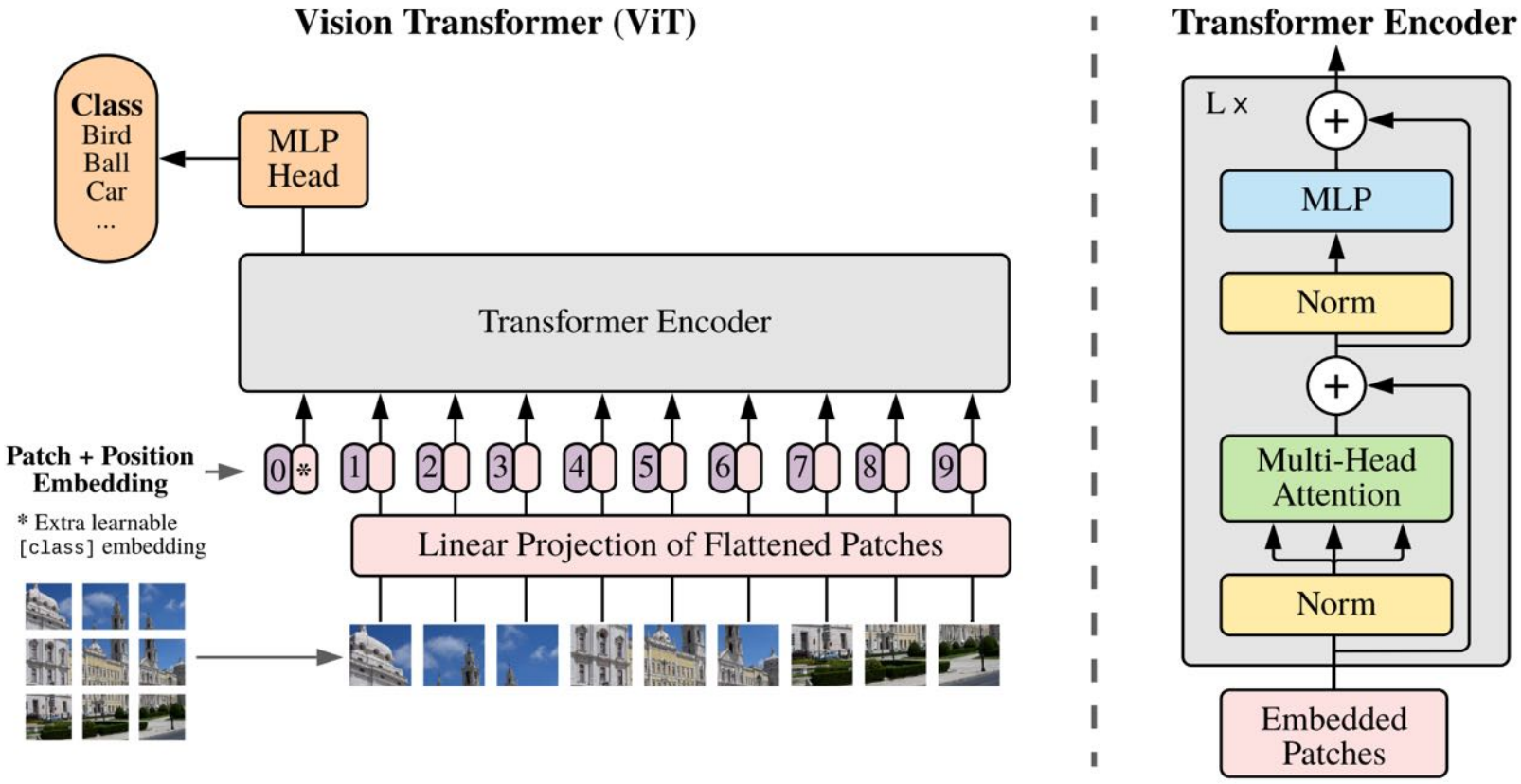
- Deep Neural Network explanations need to be **faithful** & **interpretable**
  - for faithfulness: B-cos is designed to be **dynamic linear**
  - for interpretability: B-cos induces **alignment pressure**
- The resulting networks are **competitive classifiers**...
- ... and **provide interpretable explanations** for their decisions

Overview of Today's Lecture

- Continuation of Vision-Language Learning for Computer Vision
  - Recap: CLIP
  - Flamingo [neurips'22] - https://arxiv.org/abs/2204.14198

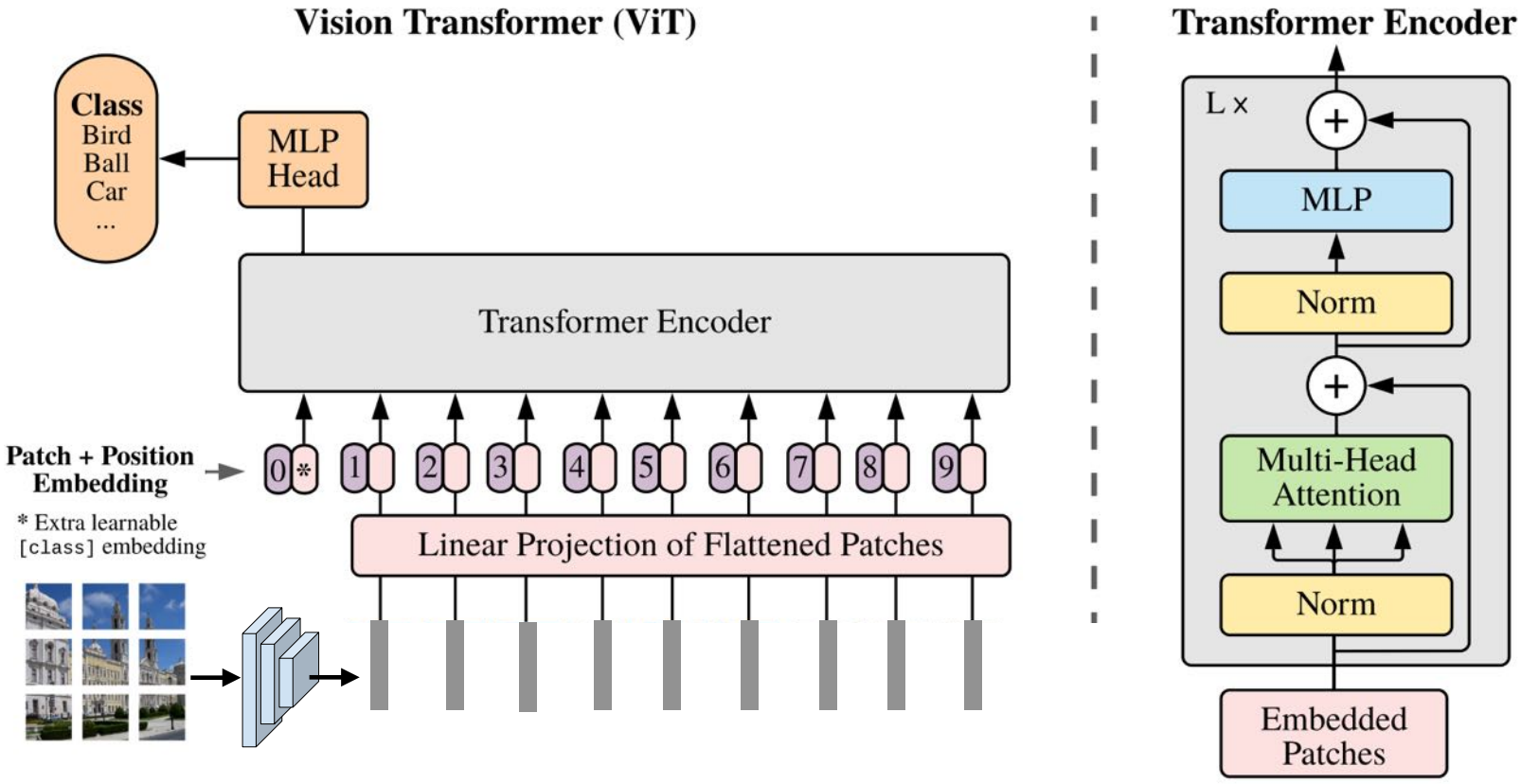
- Inherently Interpretable Neural Network
  - B-cos CNN [cvpr'22] - https://arxiv.org/abs/2205.10268
  - B-cos CNN & Vision transformers [arxiv'23] - https://arxiv.org/abs/2306.10898

Vision Transformer (ViT)



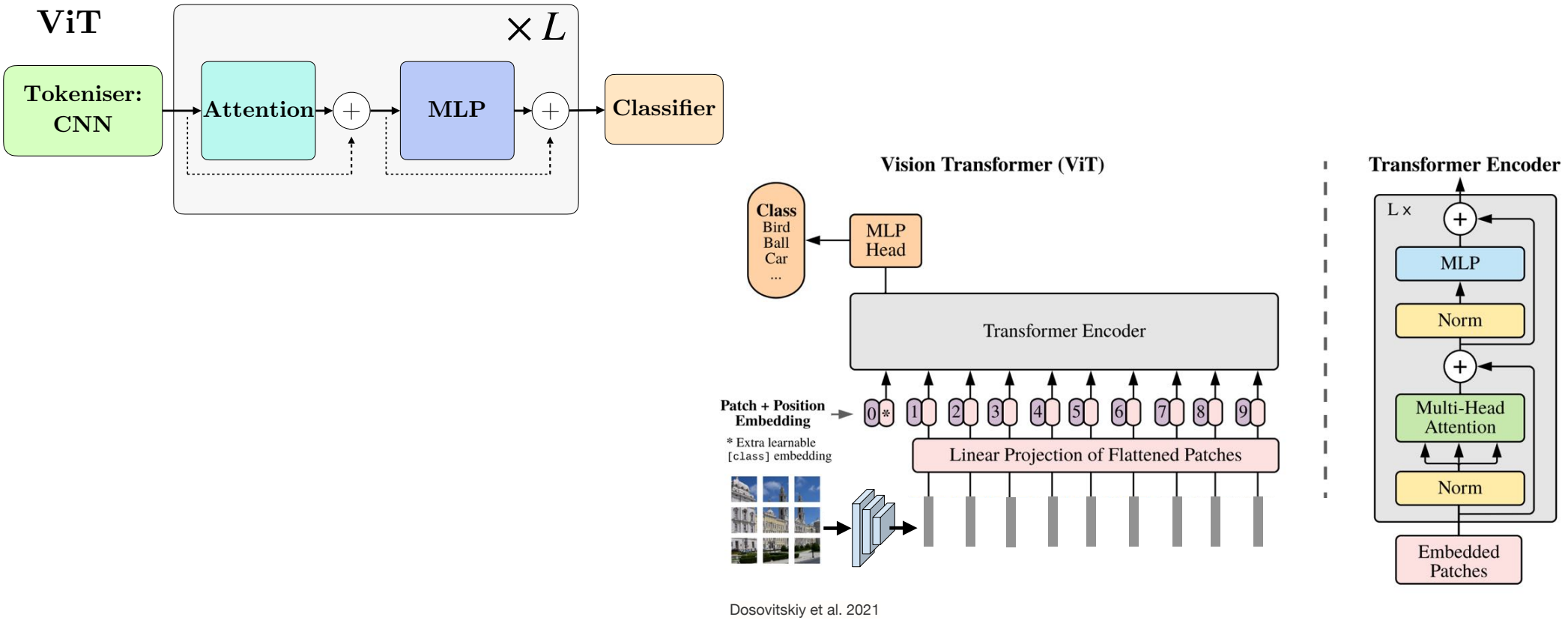
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. ICLR 2021

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



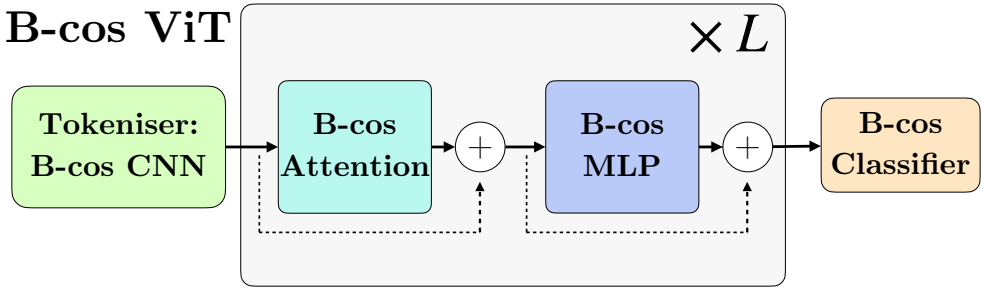
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. ICLR 2021

Attention is not All You Need (for XAI)

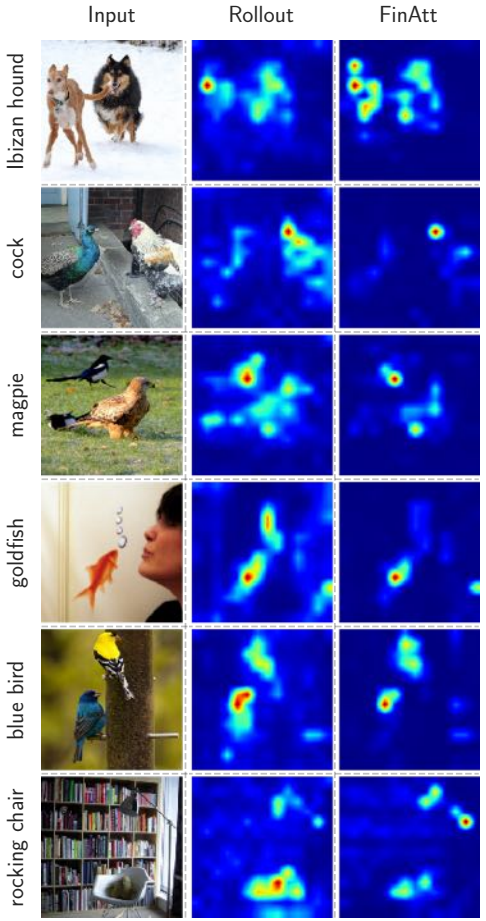


Dosovitskiy et al. 2021

Attention is not All You Need (for XAI)

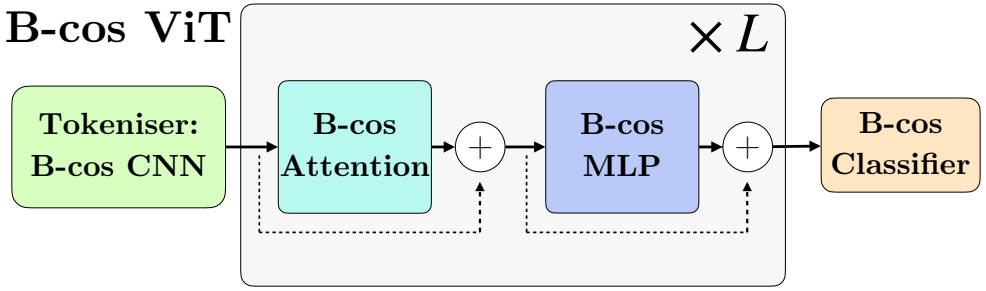


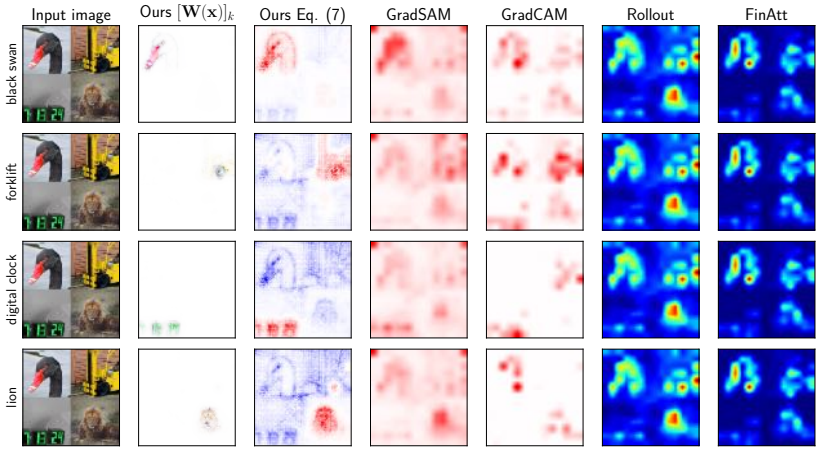
- High resolution
- Holistic explanation

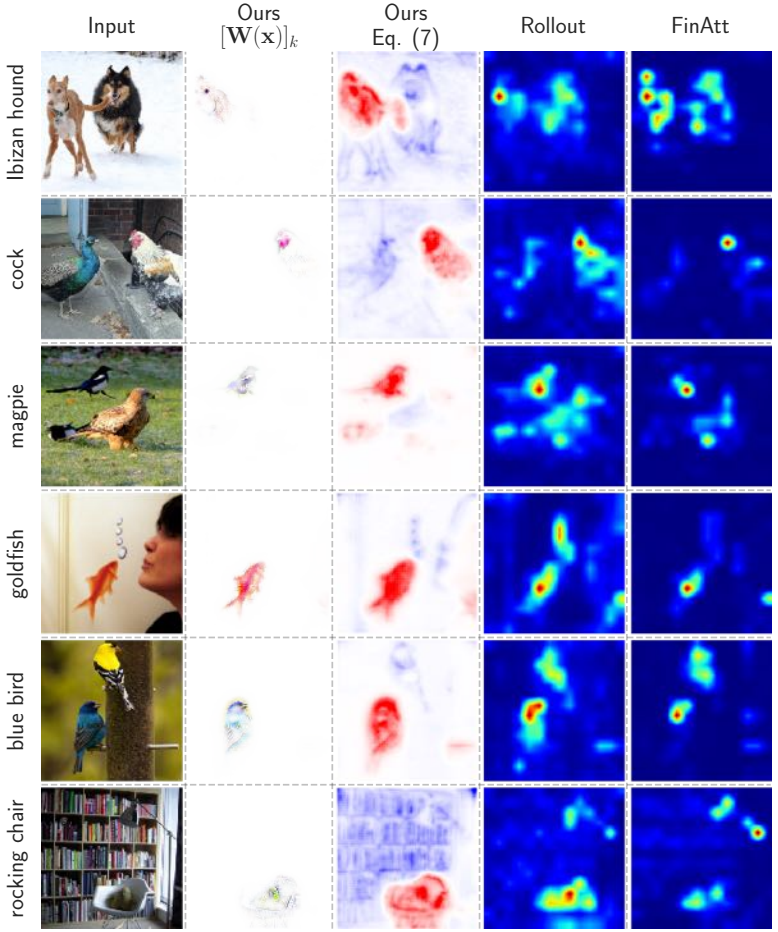


Attention Rollout: Abnar & Zuidema, ACL 2020

Attention is not All You Need (for XAI)



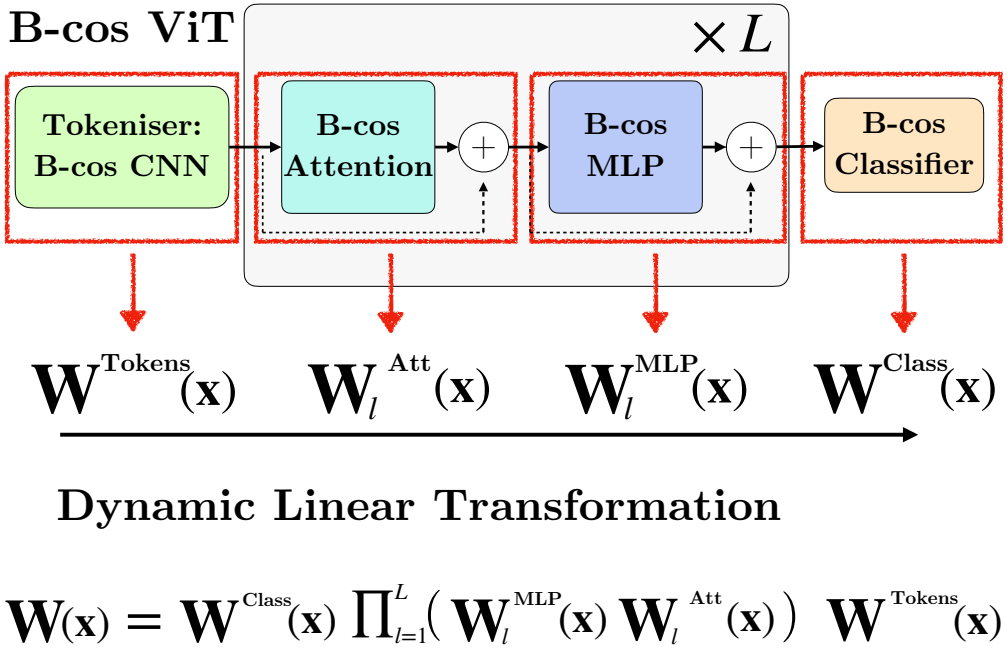




Attention Rollout: Abnar & Zuidema, ACL 2020; GradSAM: Barkan et al., CIKM 2021, GradCAM: Selvaraju et al., ICCV 2017

Attention is not All You Need (for XAI)

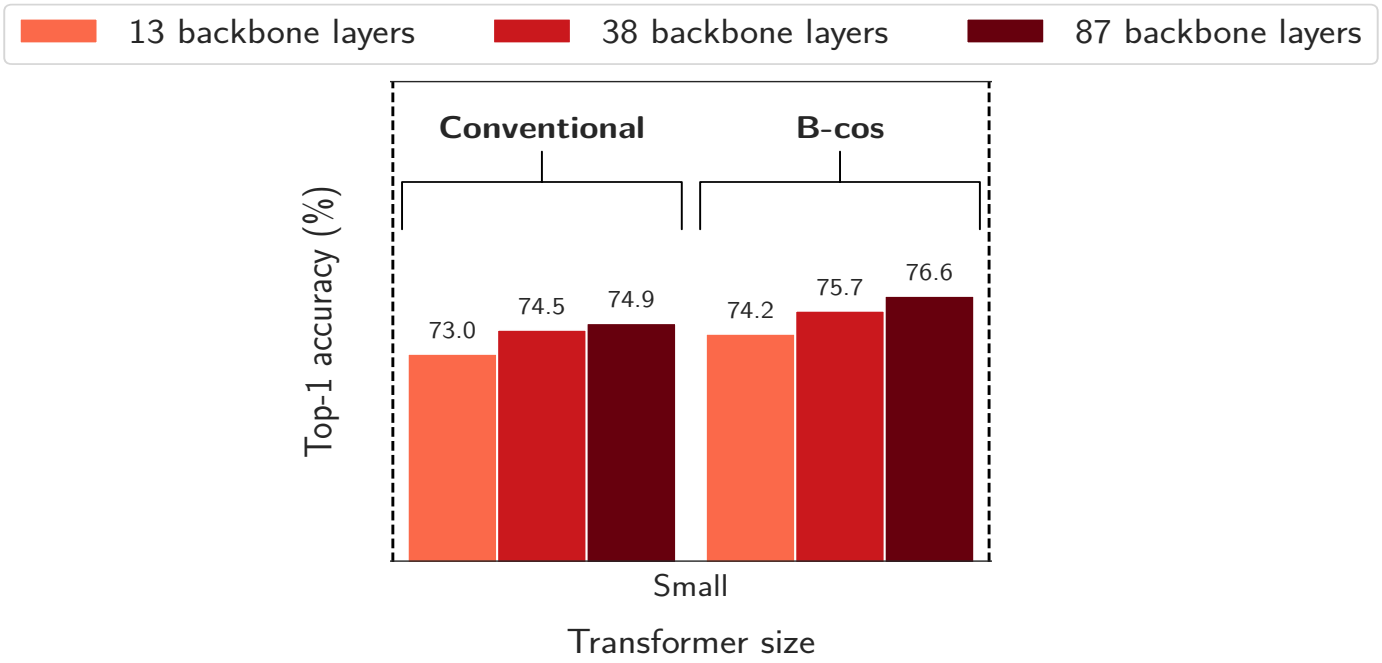
- Tokeniser + MLP + Classifier
  - interpret as CNNs,
    convert to B-cos CNNs
- Self-Attention (SA) is dynamic linear
  - $$SA(\mathbf{X}) = \underbrace{\mathbf{A}(\mathbf{X})\,\mathbf{V}}_{\mathbf{W}(\mathbf{X})}\,\mathbf{X} = \mathbf{W}(\mathbf{X})\,\mathbf{X}$$
- Side note:
  - for Tokenisation,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



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

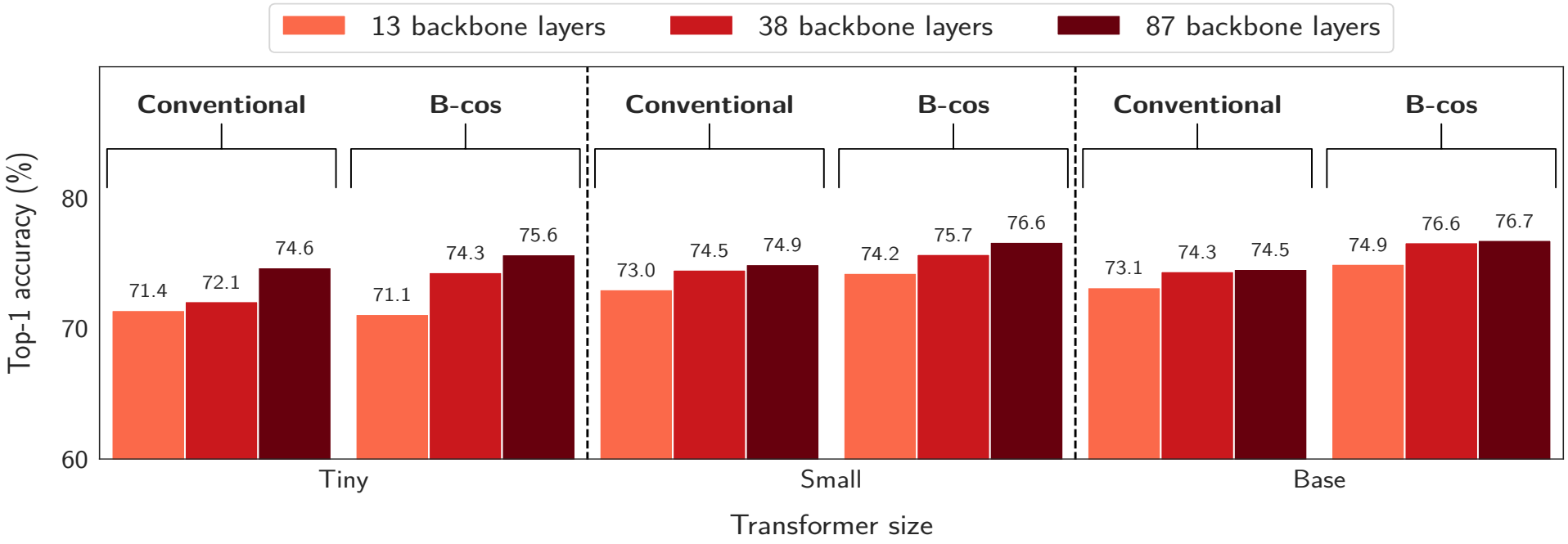
Attention is not All You Need (for XAI)

Results — classification accuracy



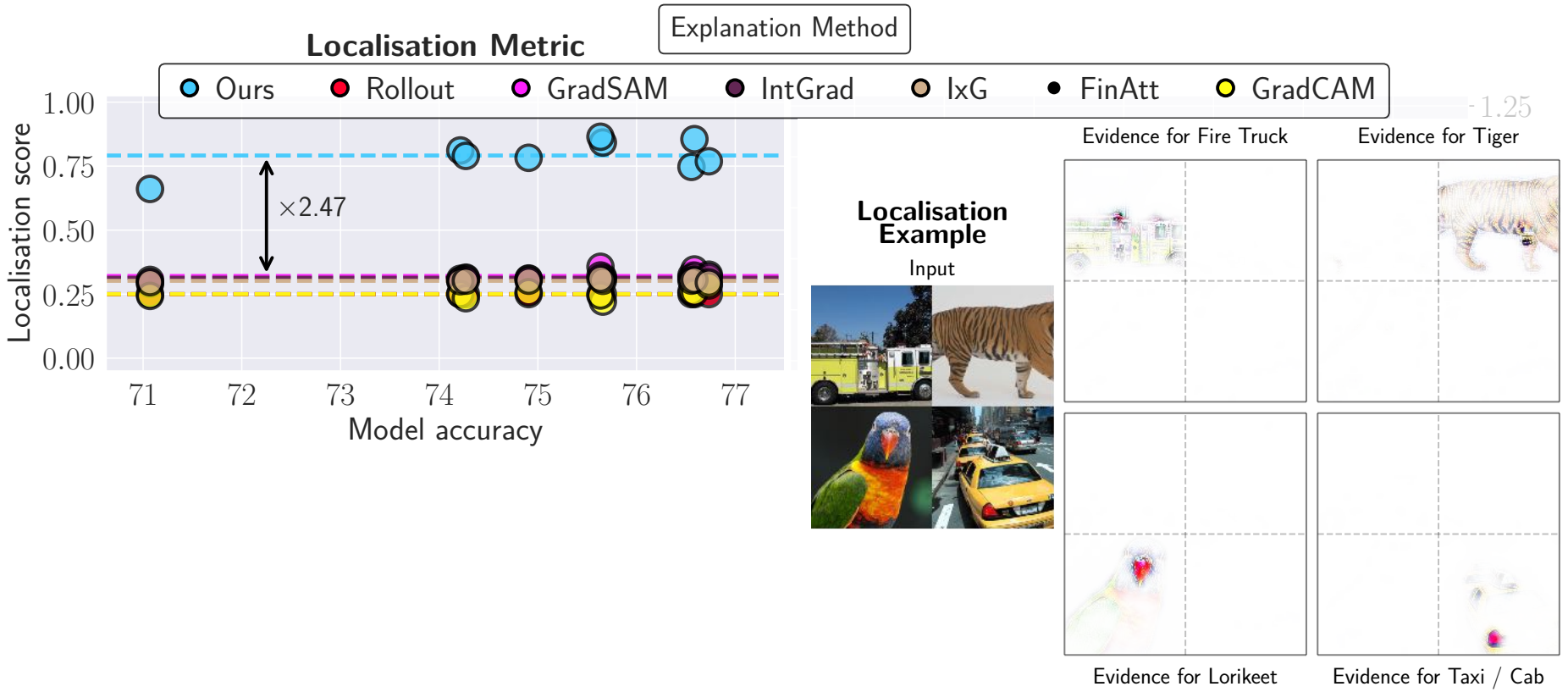
Attention is not All You Need (for XAI)

Results — classification accuracy

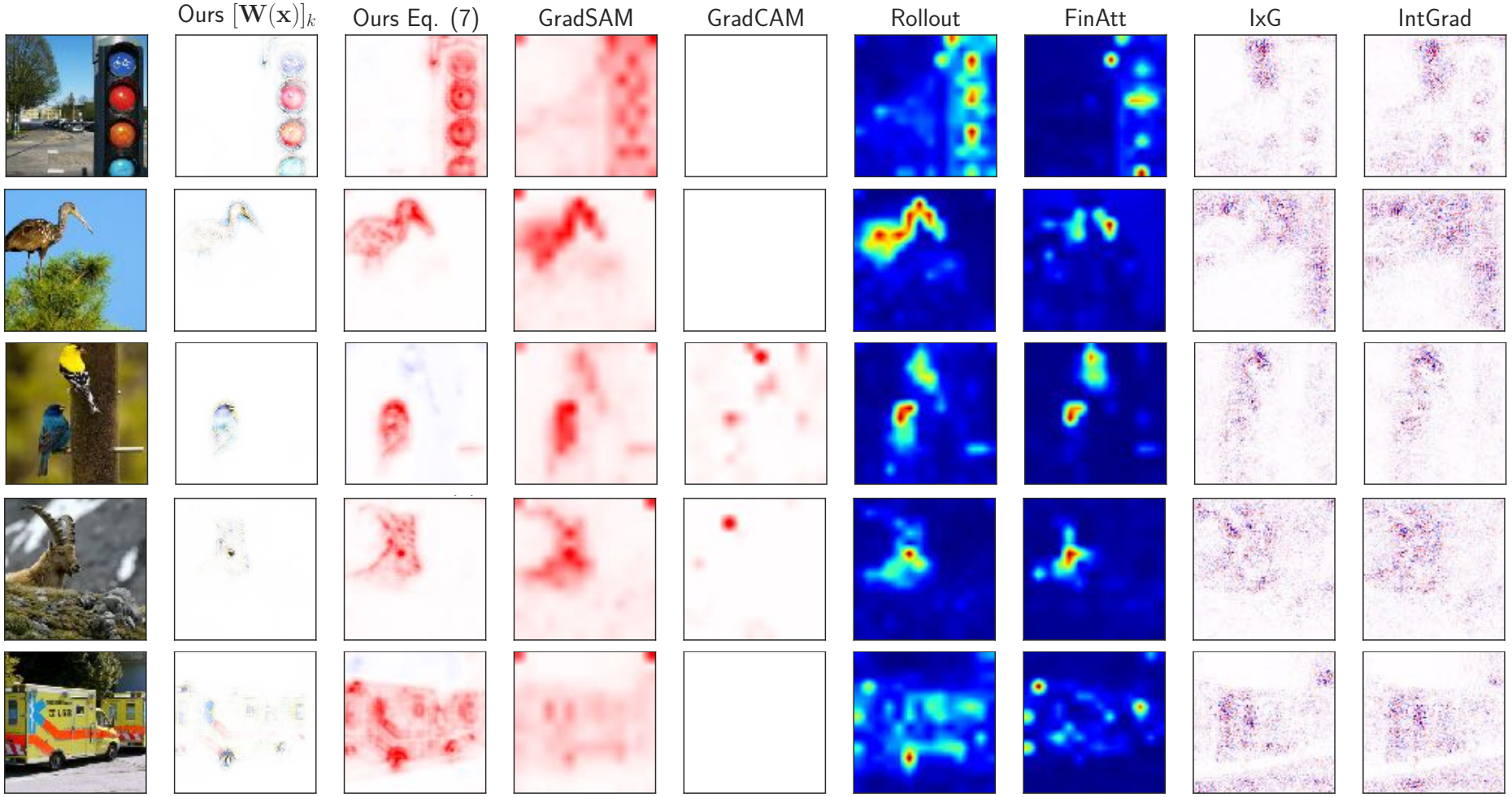


Attention is not All You Need (for XAI)

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 $\text{SA}(\mathbf{X}) = \mathbf{W}(\mathbf{X})\mathbf{X}$
  - 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