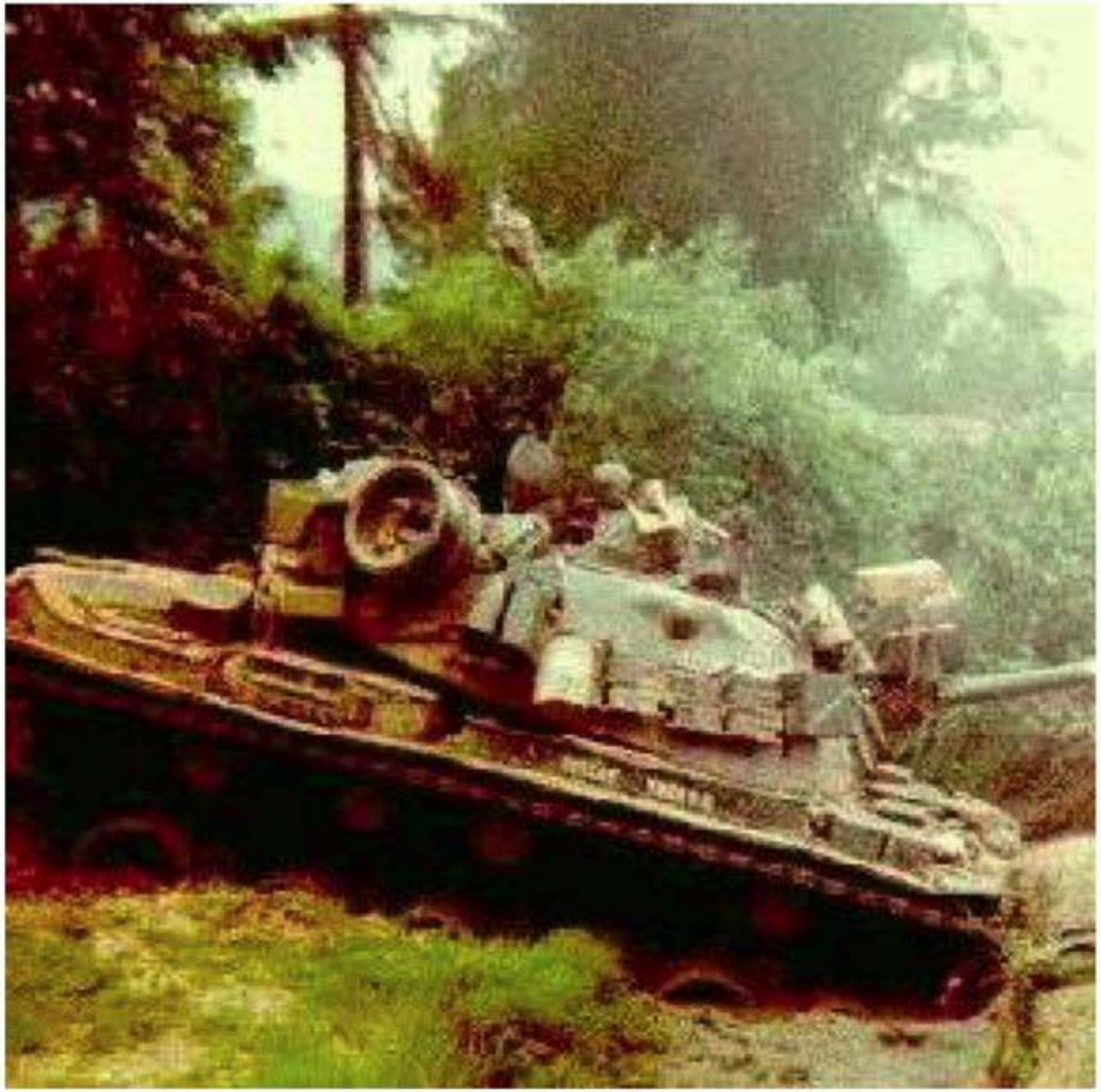
Explaining Model Decisions and Correcting them via Human Feedback

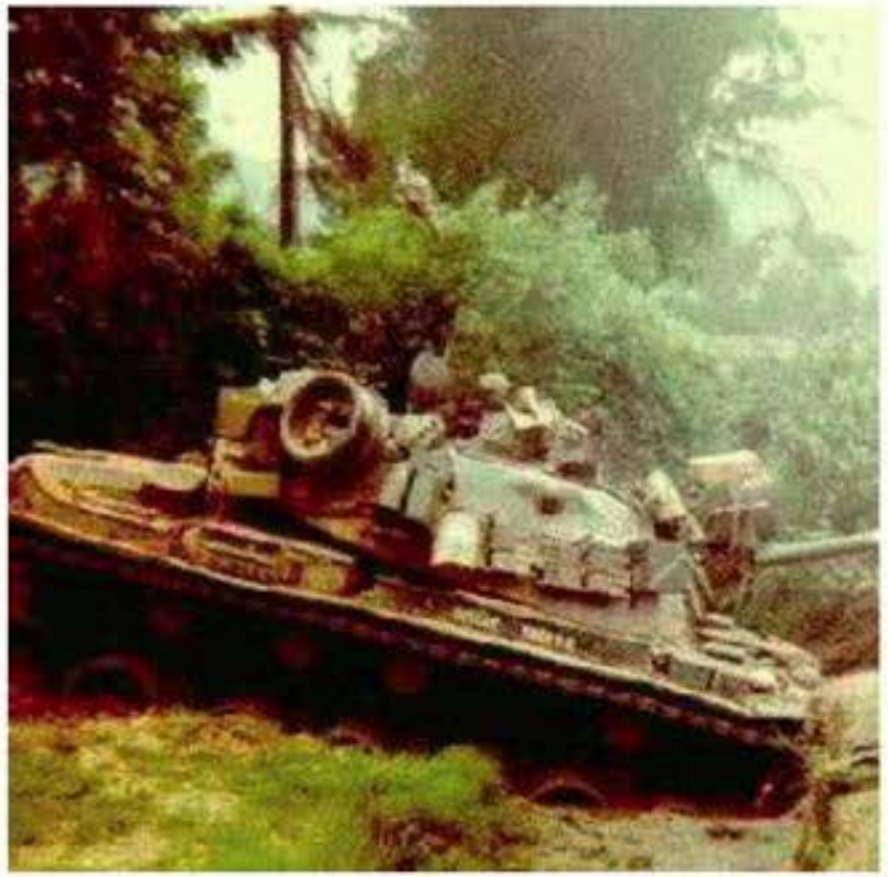
Ramprasaath R. Selvaraju

Georgia Institute of Technology



Mission: Detect Camouflaged Enemy Tanks

- Plan:
    - Collect training data
        - Positive: Tanks camouflaged by trees
        - Negative: Trees with no tanks
- Implementation:
    - Train a machine learning model
    - Achieved good performance on test set
- Right for wrong reasons
    - Issues
        - Images of forest taken on sunny day
        - Images with camouflaged tanks taken on a cloudy day



Gray skies for the US military!


lf-Driving
48
mph
Autonomous
Mode
NETWORK
DRIVER

Need for interpretability in life critical applications

towards
data science

Another Self-Driving Car Accident, Another AI Development Lesson

Photo from https://blogs.nvidia.com

Interpretability in Healthcare

EU introduces GDPR – Right to explanation

GDPR establishes that *a data subject has the right to "an explanation of the decision reached after [algorithmic] assessment."*

Talk outline

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV’17, IJCV’19)

Debias

Leveraging explanations to unbias models through HINT (ICCV’19)

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What future directions excite me?

Explain

Explain decisions from deep networks through Grad-CAM (ICCV’17, IJCV’19)

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, **“Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, ICCV’17, IJCV’19**

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

How can we explain decisions from deep models?

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, **"Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", ICCV'17, IJCV'19**

How can we explain decisions from deep models?

How can we explain decisions from deep models?

Visual Explanations

Where does an intelligent system “look” when making decisions?

How can we explain decisions from deep models?

Interpretability landscape in 2016

Earlier approaches for visual explanations

Earlier approaches for visual explanations

Gradient-based methods

- Backpropagation [Simonyan *et al.*, 2013]
- Deconvolution [Zeiler *et al.*, 2014]
- Guided Backpropagation [Springenberg *et al.*, 2014]
- Layer-wise Relevance Propagation [Bach *et al.*, 2015]

Noisy
Not class-discriminative

Earlier approaches for visual explanations

Gradient-based methods

- Backpropagation [Simonyan *et al.*, 2013]
- Deconvolution [Zeiler *et al.*, 2014]
- Guided Backpropagation [Springenberg *et al.*, 2014]
- Layer-wise Relevance Propagation [Bach *et al.*, 2015]

Noisy
Not class-discriminative

Simplifying model architectures

- Class Activation Mapping (CAM) [Zhou *et al.*, 2015]

Applicable only to limited architectures

Earlier approaches for visual explanations

Gradient-based methods

- Backpropagation [Simonyan *et al*., 2013]
- Deconvolution [Zeiler *et al*., 2014]
- Guided Backpropagation [Springenberg *et al*., 2014]
- Layer-wise Relevance Propagation [Bach *et al*., 2015]

Noisy
Not class-discriminative

Simplifying model architectures

- Class Activation Mapping (CAM) [Zhou *et al*., 2015]

Applicable only to limited architectures

Black-box approaches

- LIME [Ribeiro *et al*., 2016]

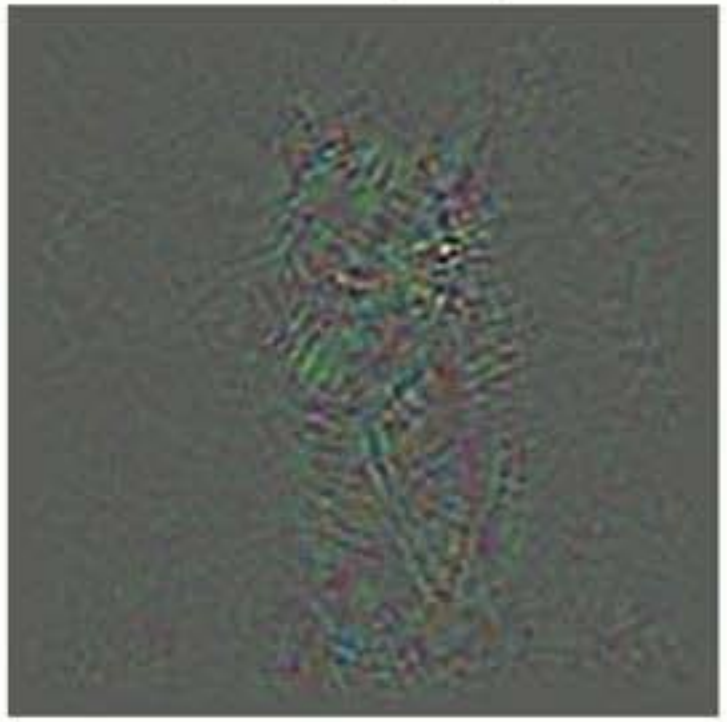
Model-agnostic

Problems with existing approaches

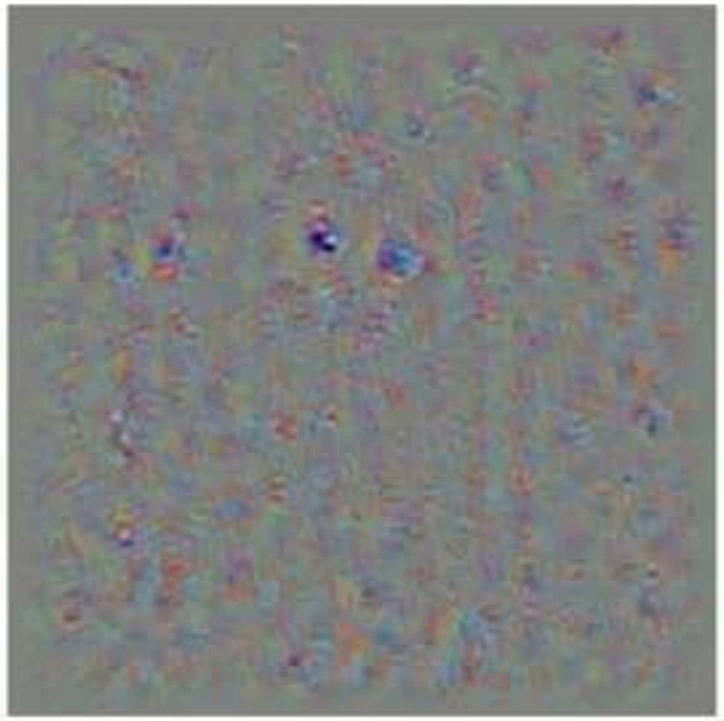
Problems with existing approaches

Input

Backprop



Deconv

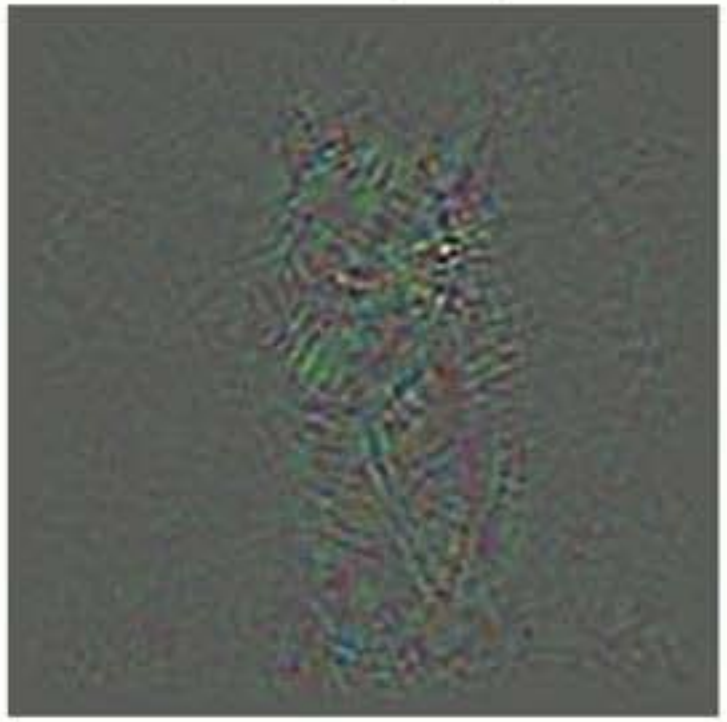


Problems with existing approaches

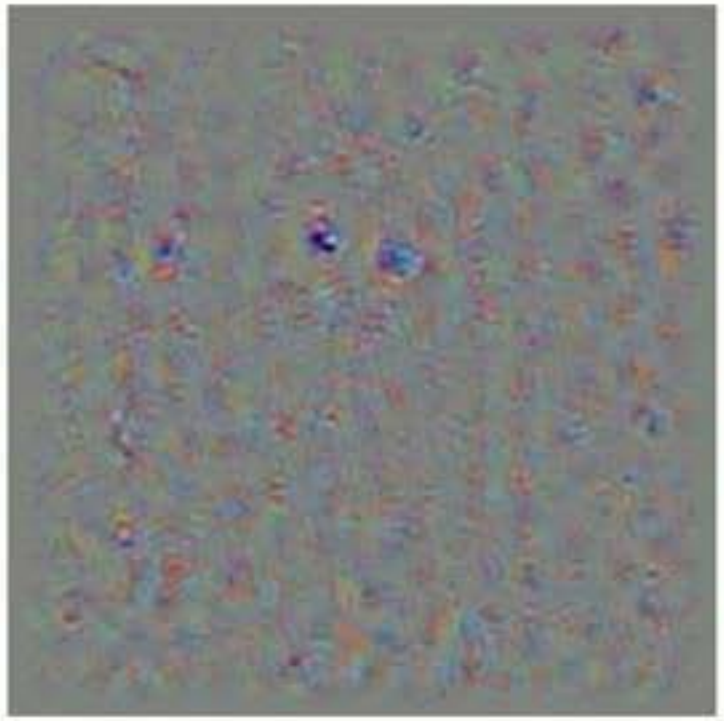
- Noisy

Input

Backprop



Deconv



Problems with existing approaches

- **Noisy**

Problems with existing approaches

- Noisy
- Not class-discriminative

Input

Problems with existing approaches

- Noisy
- Not class-discriminative

Guided Backprop for “*Cat*”

Input

Problems with existing approaches

- Noisy
- Not class-discriminative

Guided Backprop for *"Cat"*

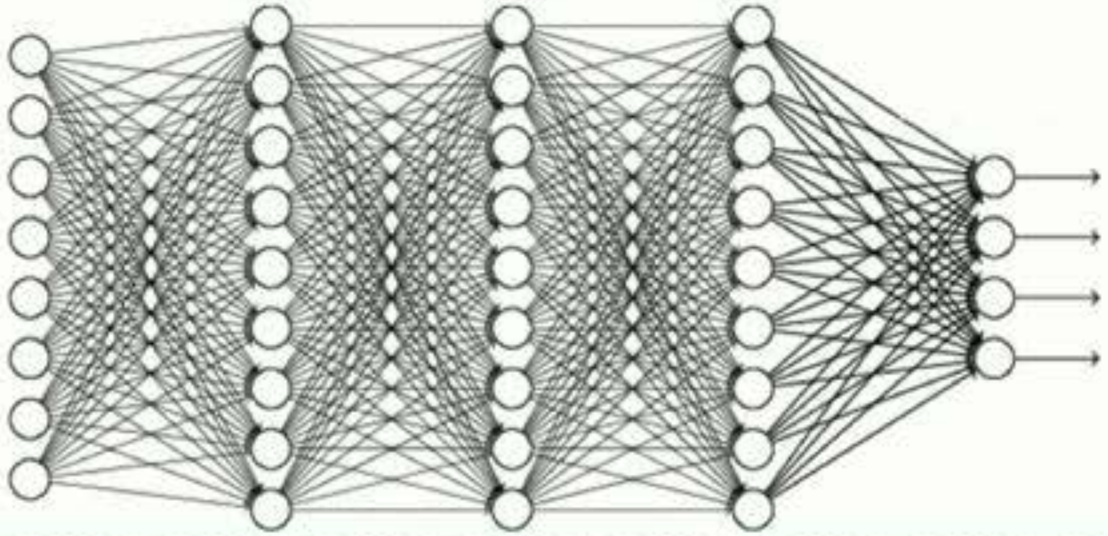
Input

Guided Backprop for *"Dog"*

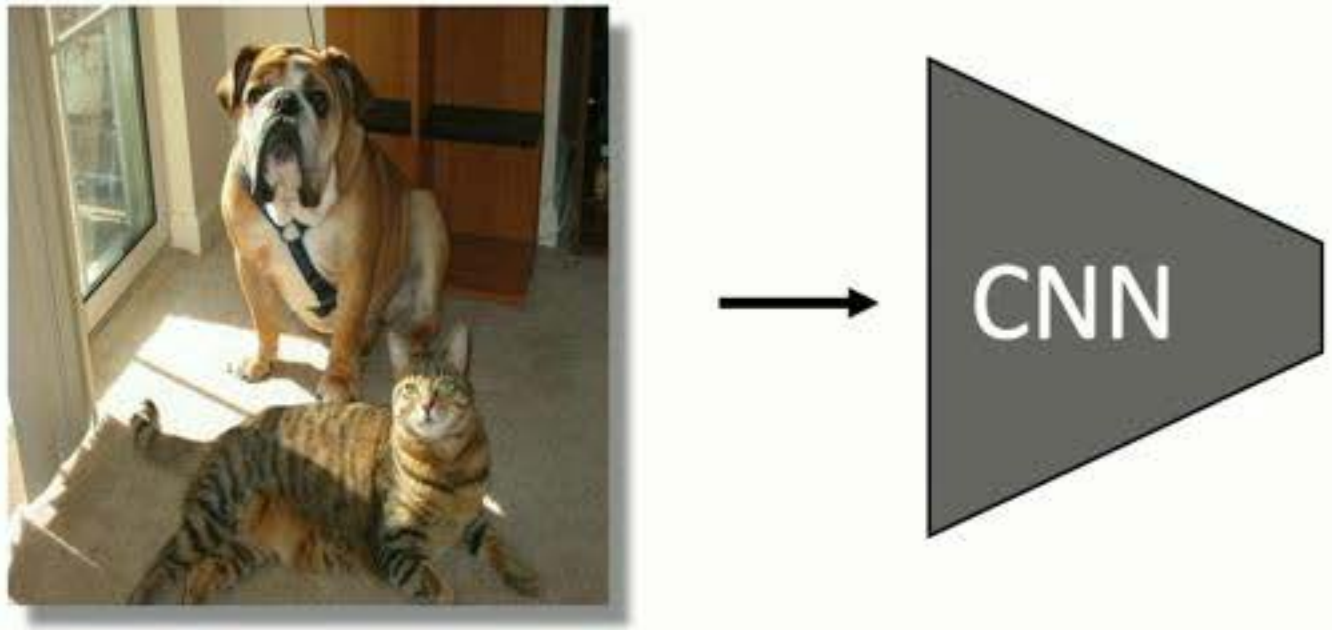
Why is explaining deep models particularly hard?

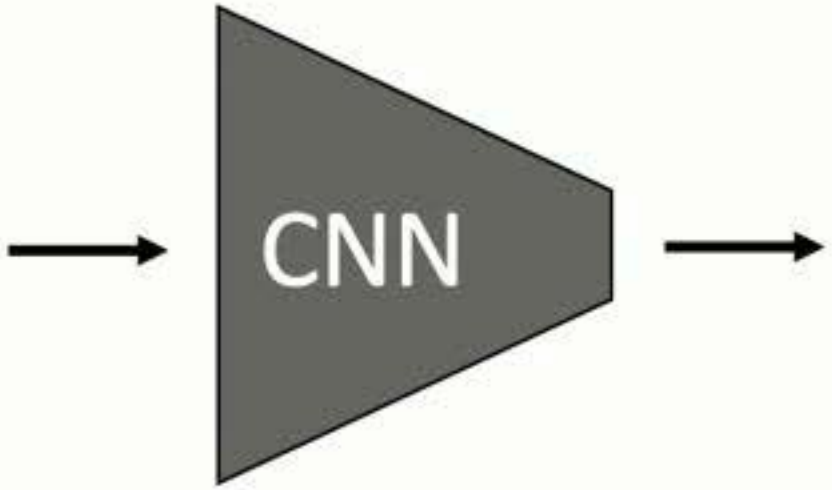
- Linear models with interpretable and normalized features are inherently interpretable
  - Weights of the features signify importance

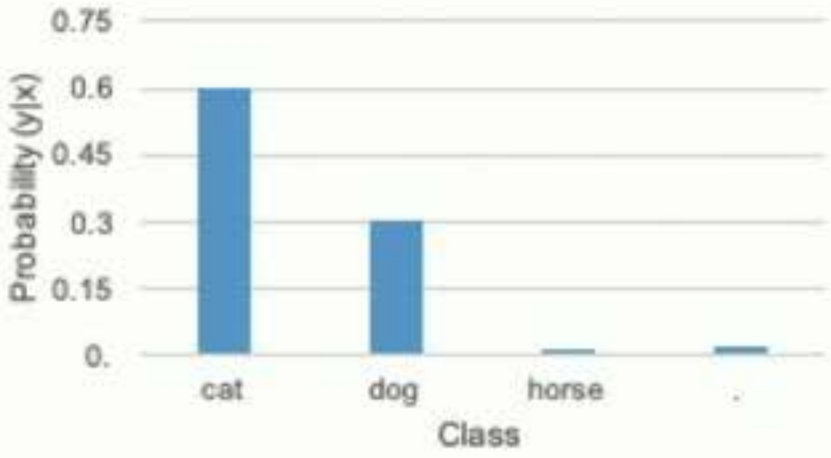
- Deep neural networks are highly nonlinear complex piece of functions

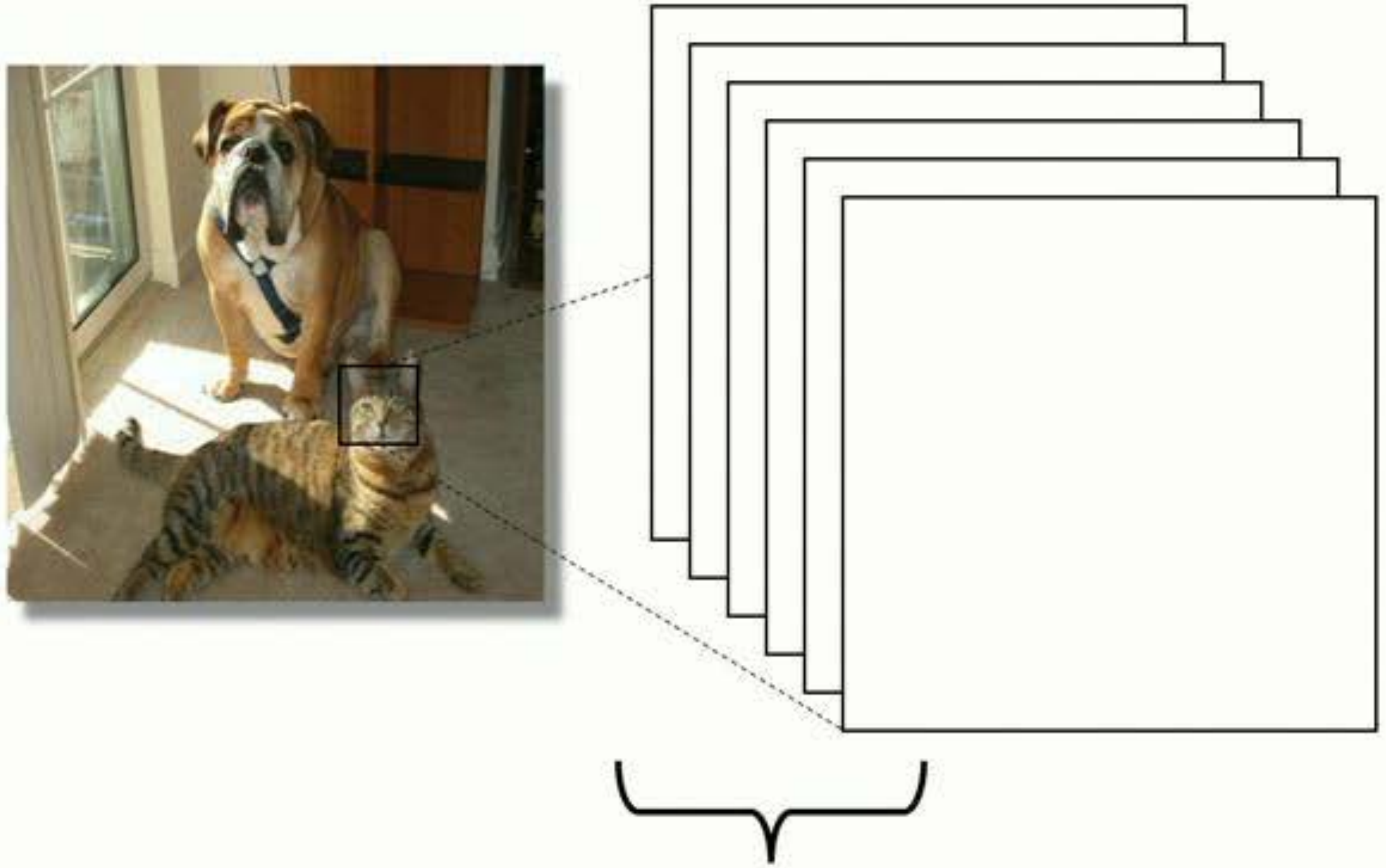


Estimating feature importance for deep networks is hard

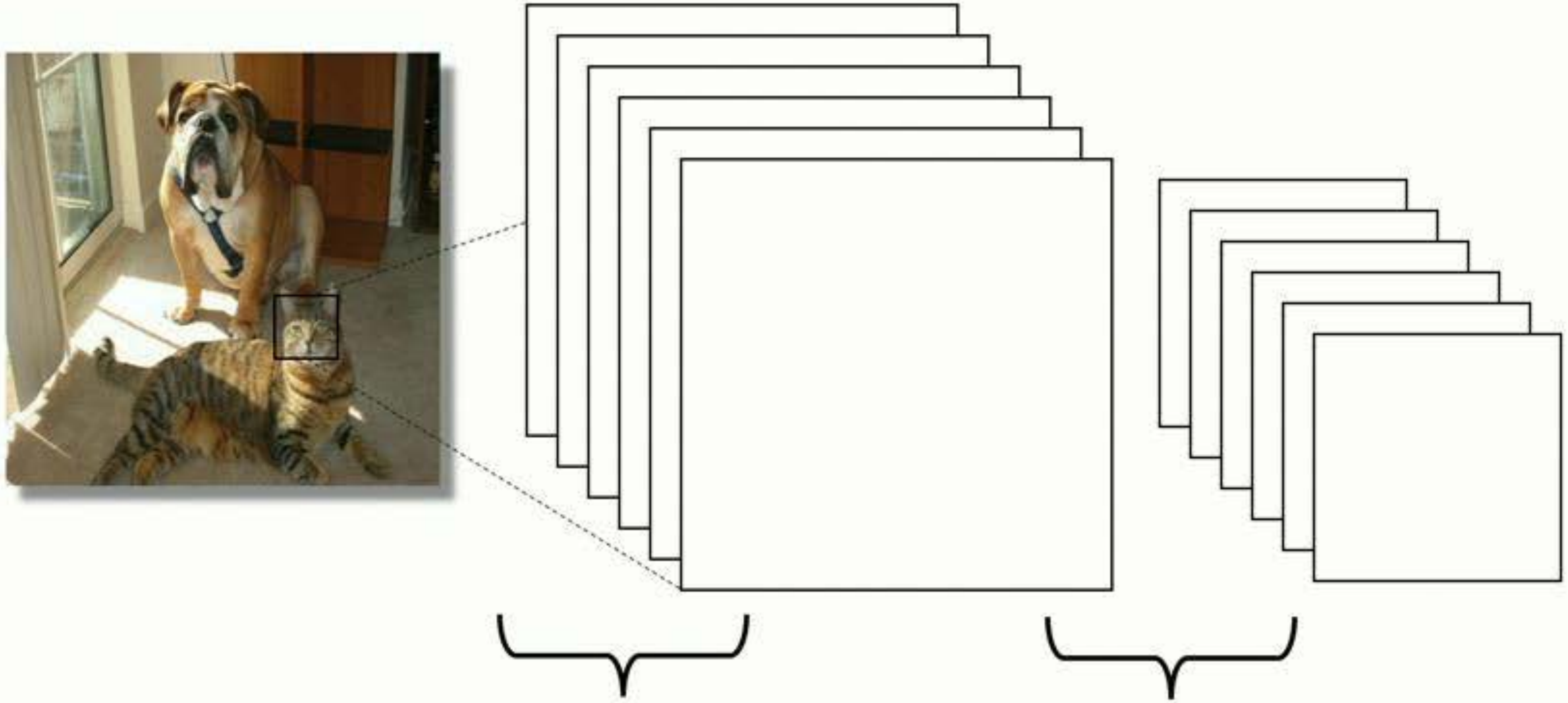

CNN


CNN


0.75
0.6
0.45
0.3
0.15
0.
Probability (y|x)
cat
dog
horse
.
Class

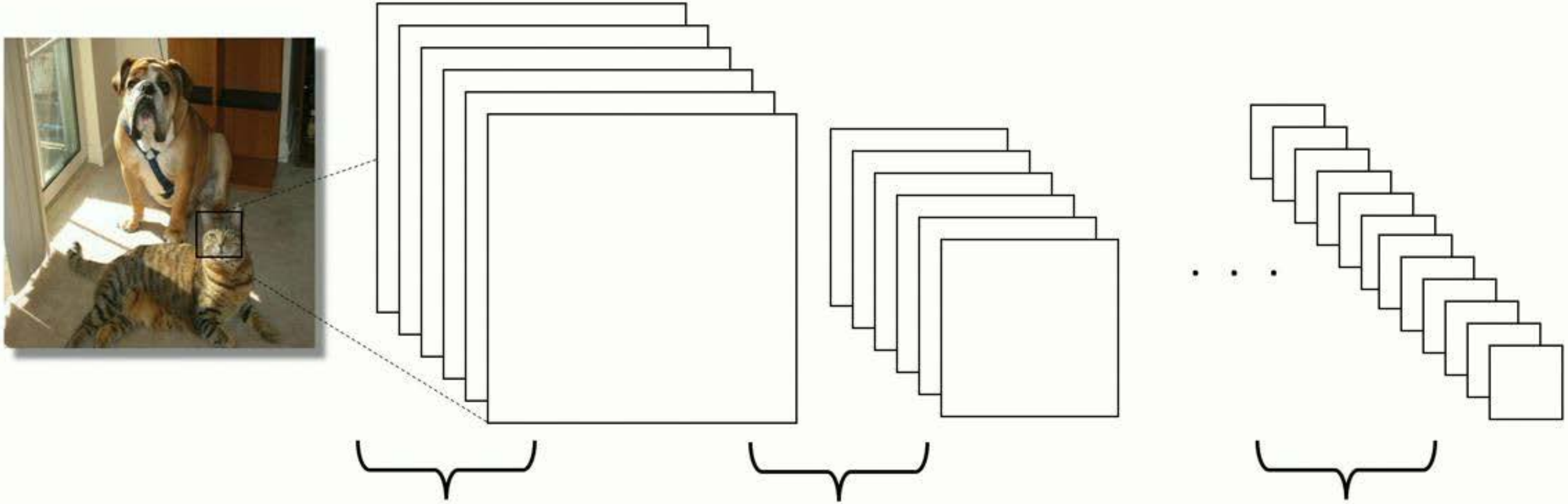


Convolution
+
ReLU



Convolution
+
ReLU

Pooling

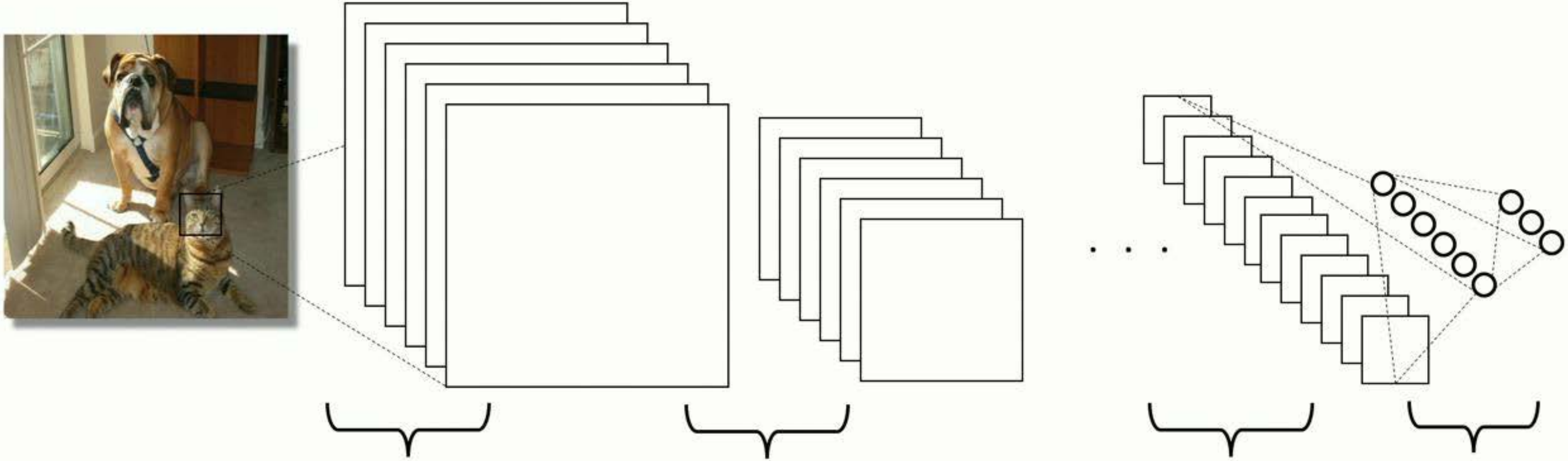


Convolution
+
ReLU

Pooling

. . .

Convolution
+
ReLU



Convolution
+
ReLU

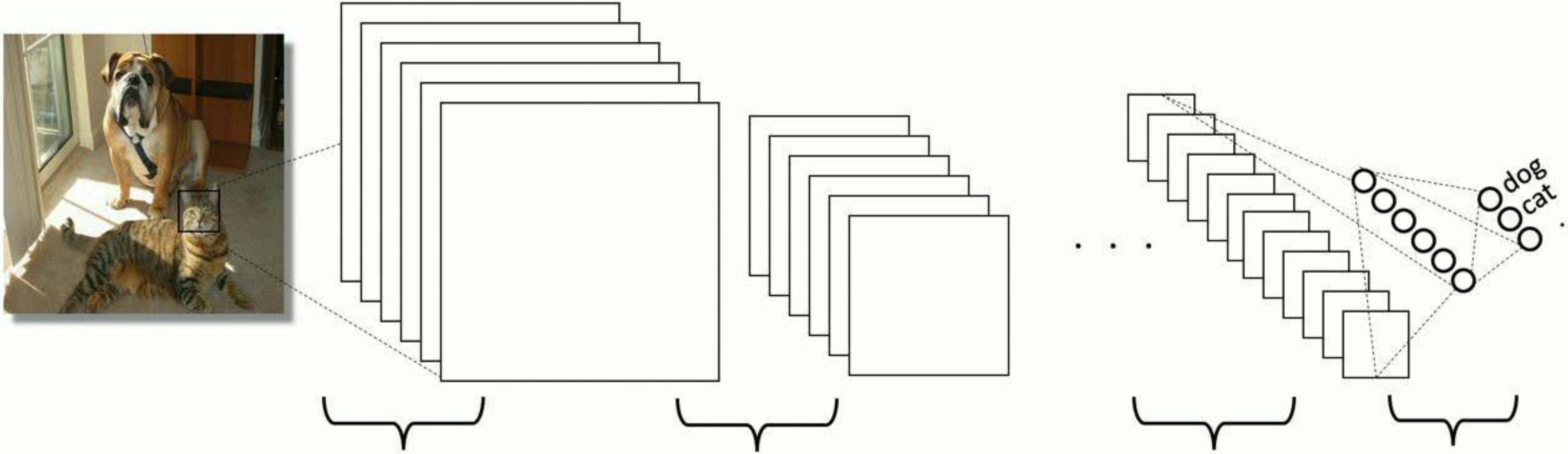
Pooling

. . .

Convolution
+
ReLU

Fully Connected
+
Softmax

What do individual layers in deep models learn?



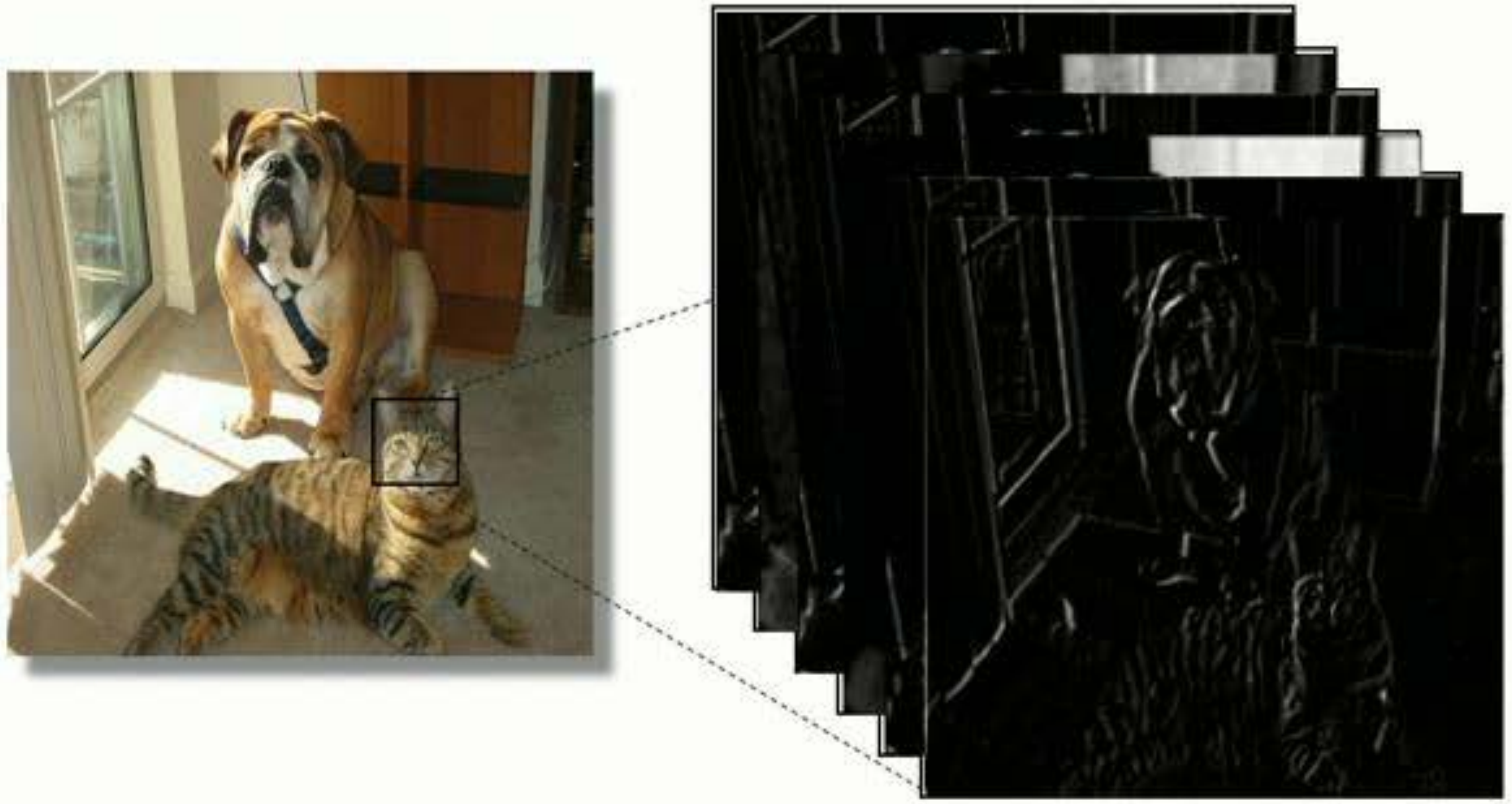
Convolution
+
ReLU

Pooling

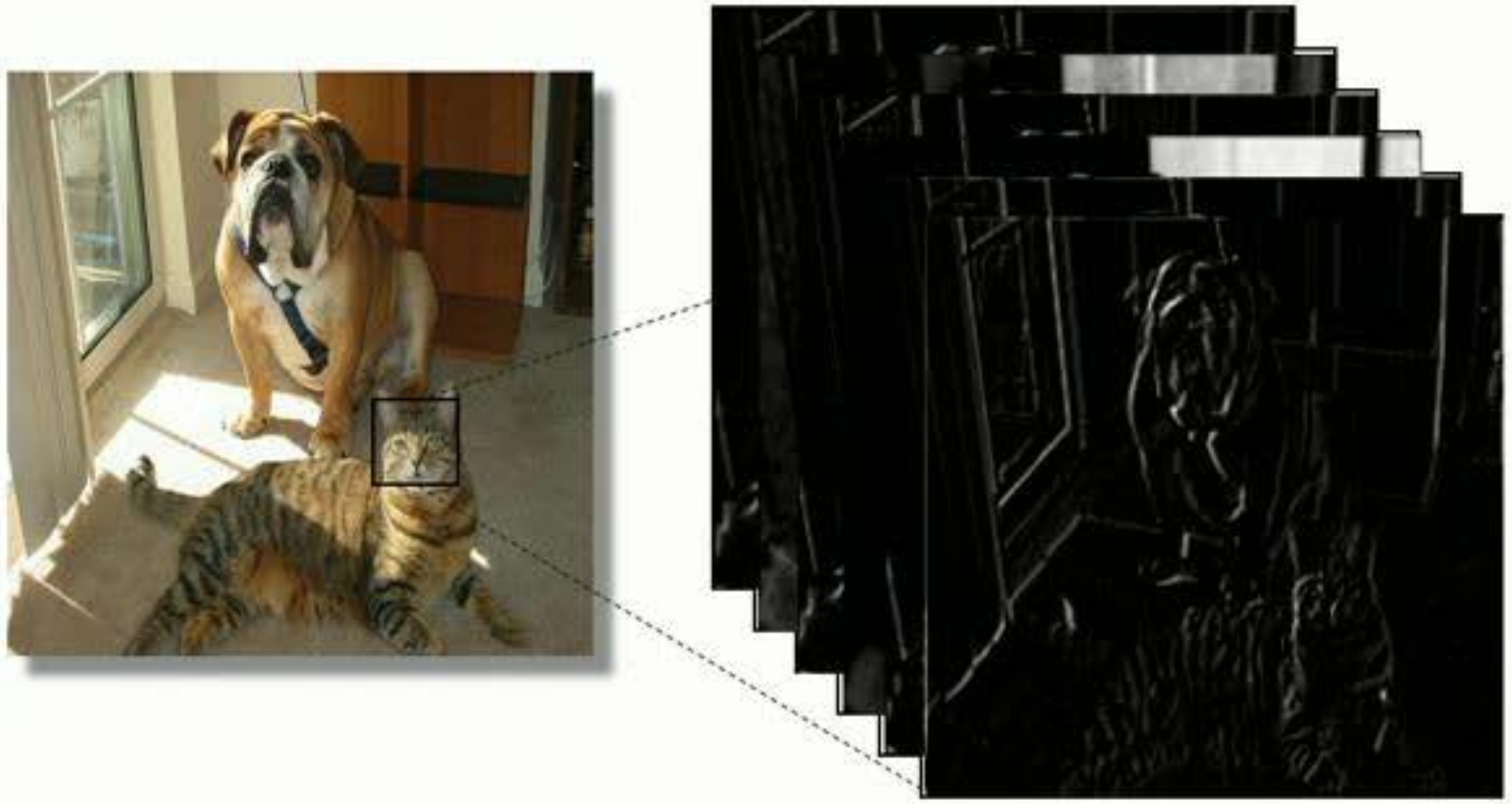
Convolution
+
ReLU

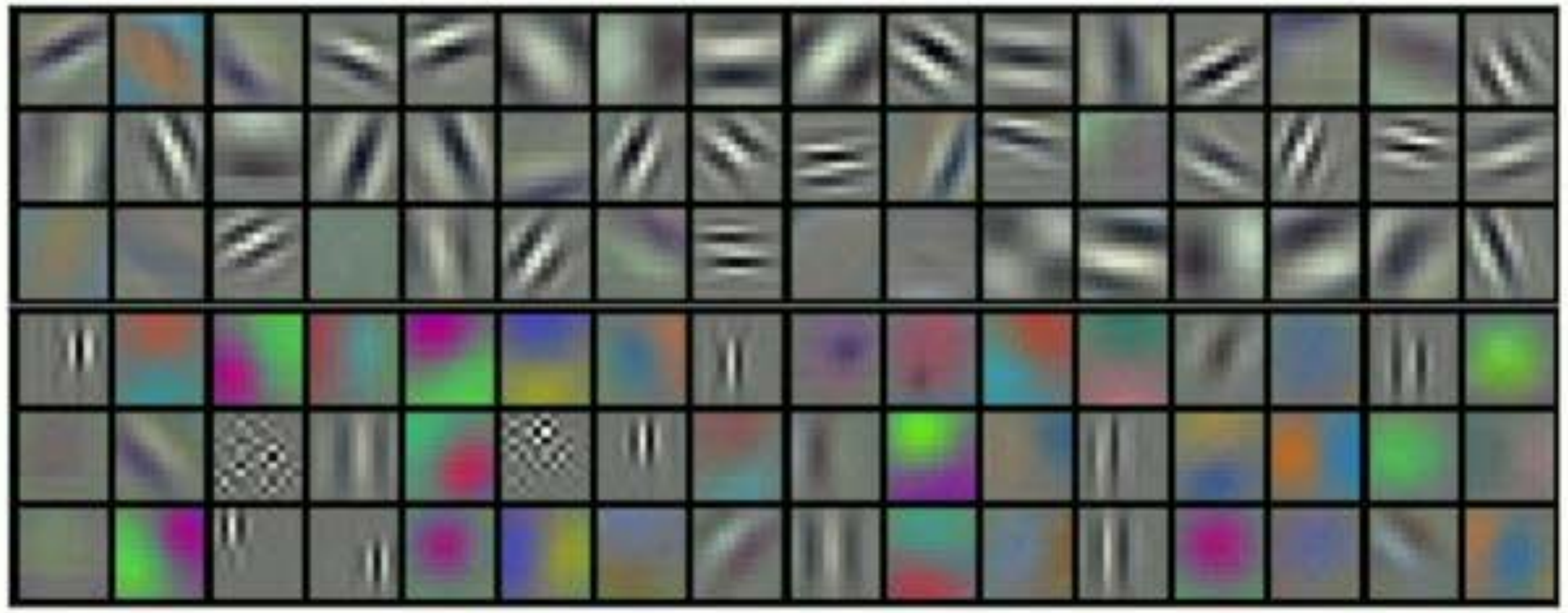
Fully Connected
+
Softmax

Lower layers look for edges/blobs

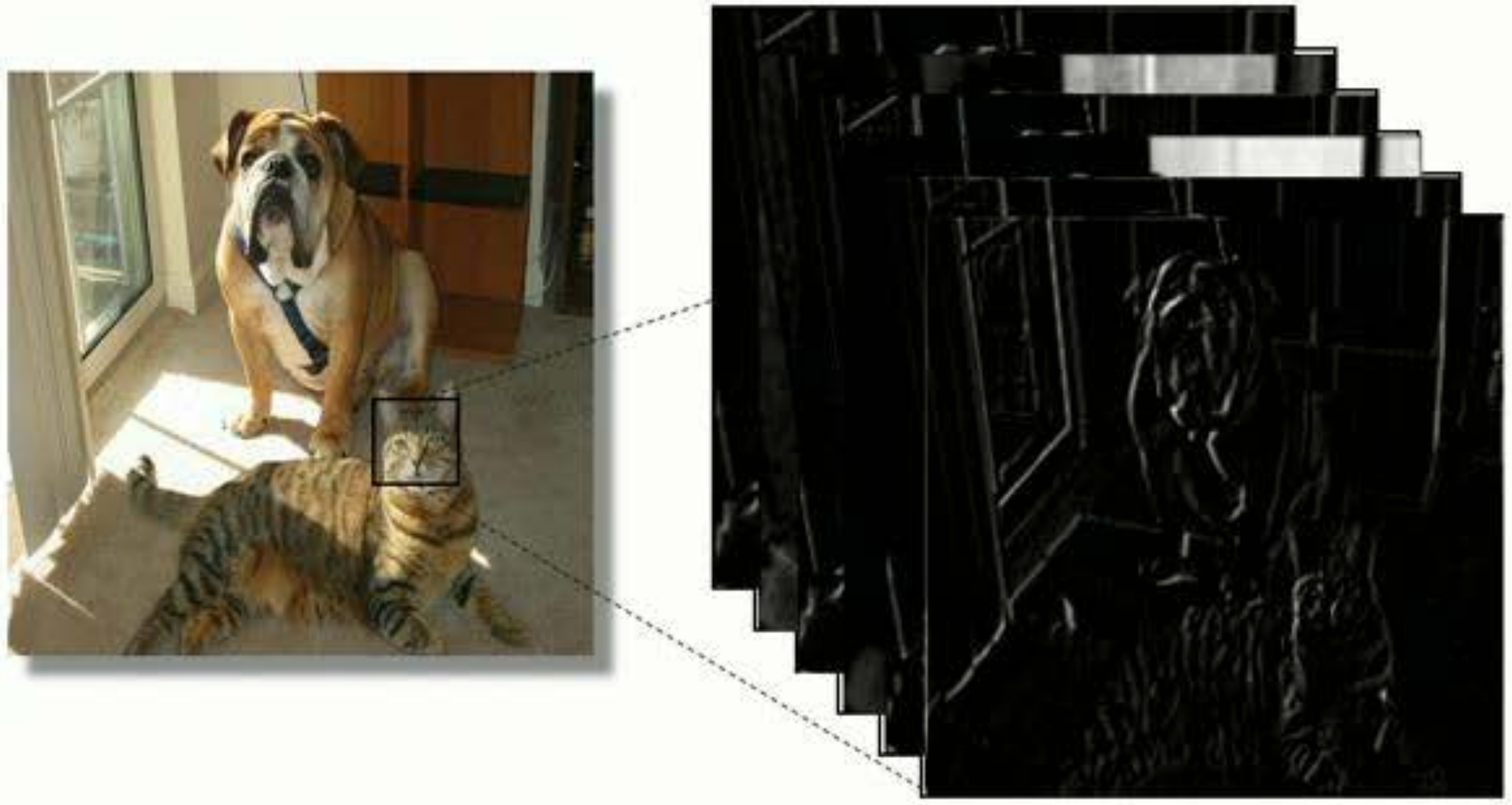


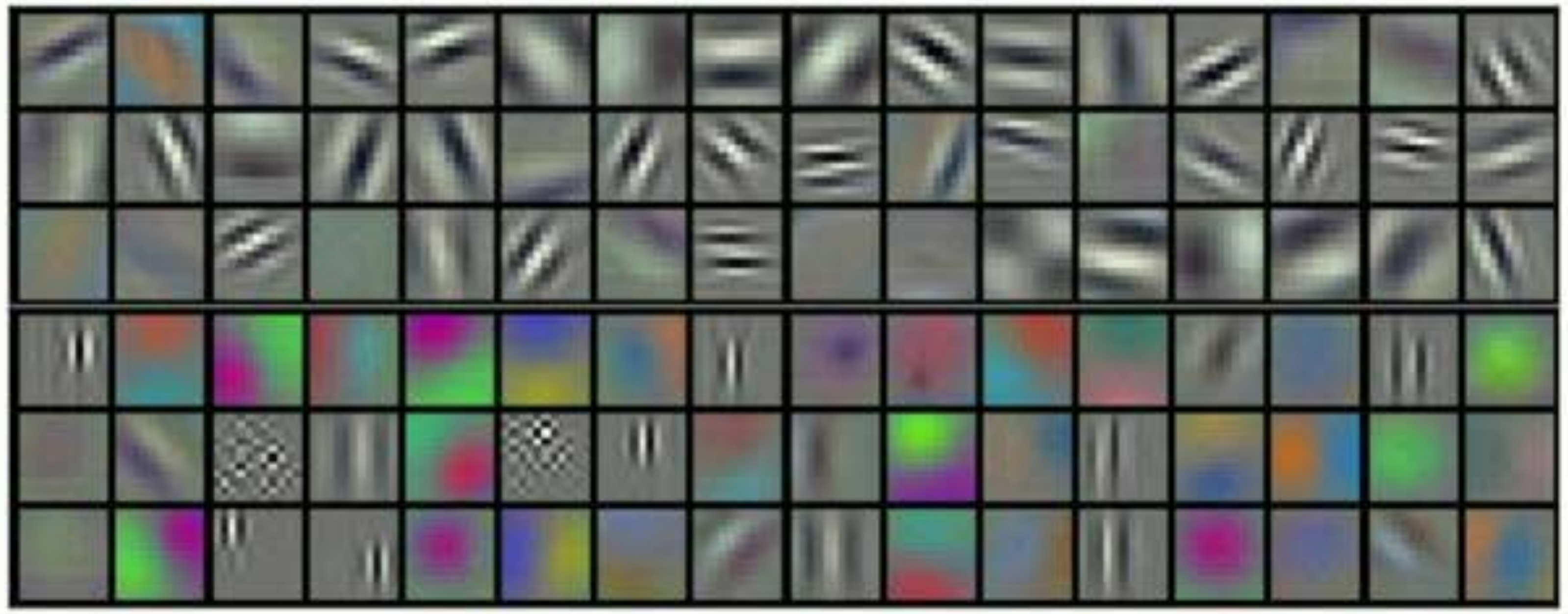
Lower layers look for edges/blobs





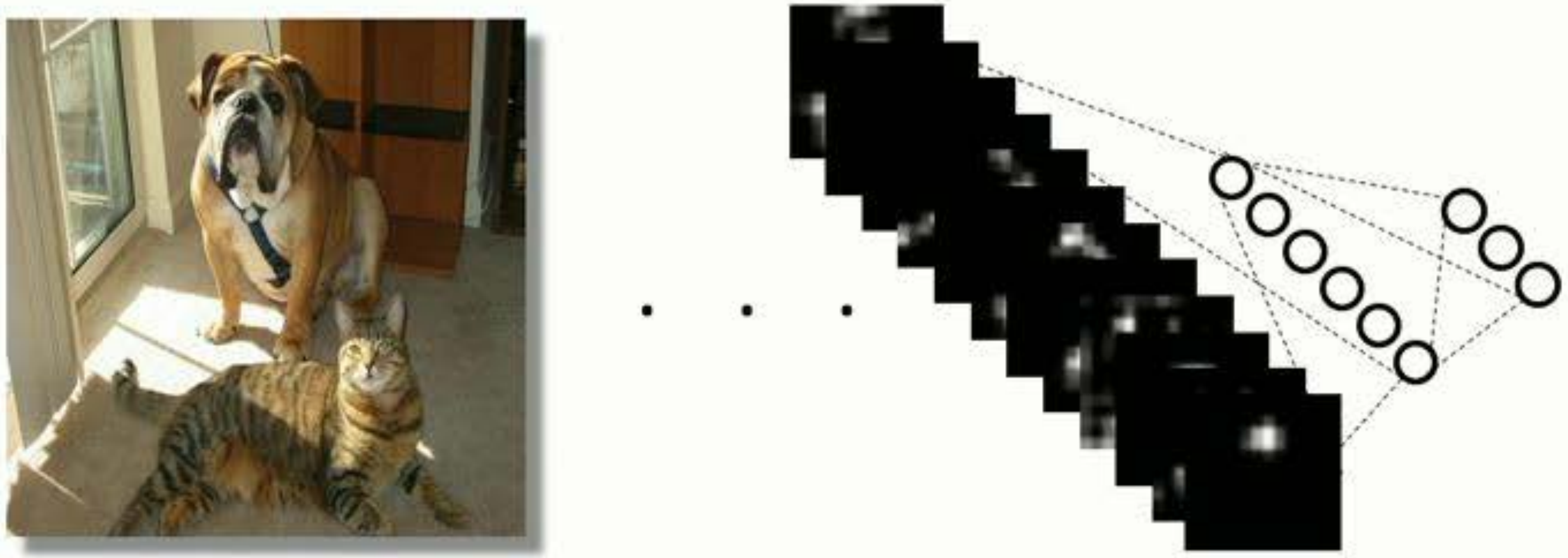
Lower layers look for edges/blobs



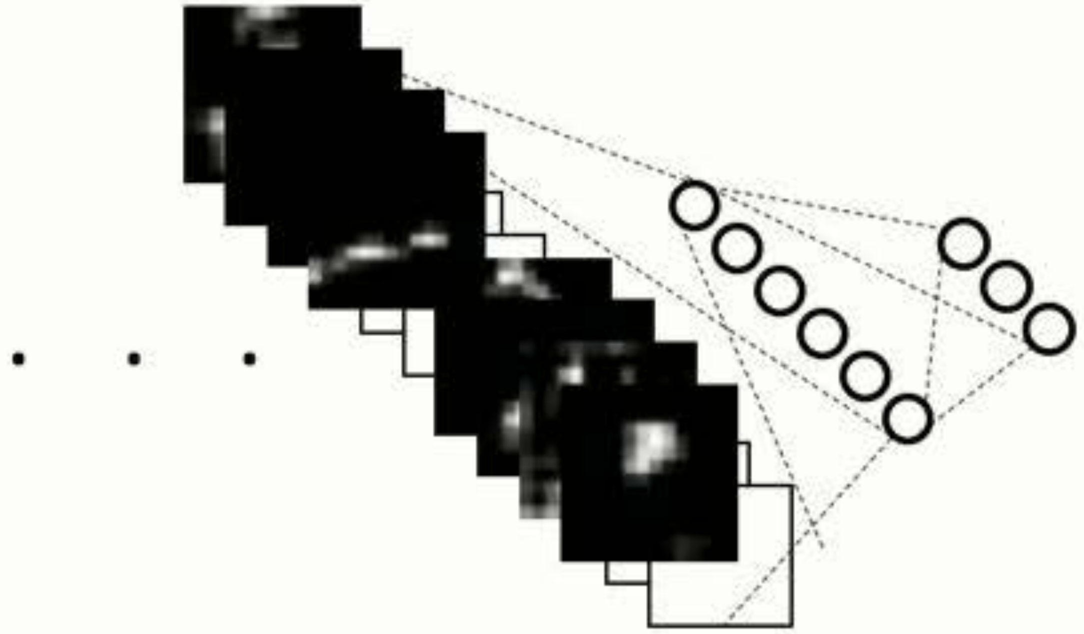


Importance in terms of first layer filters are not meaningful

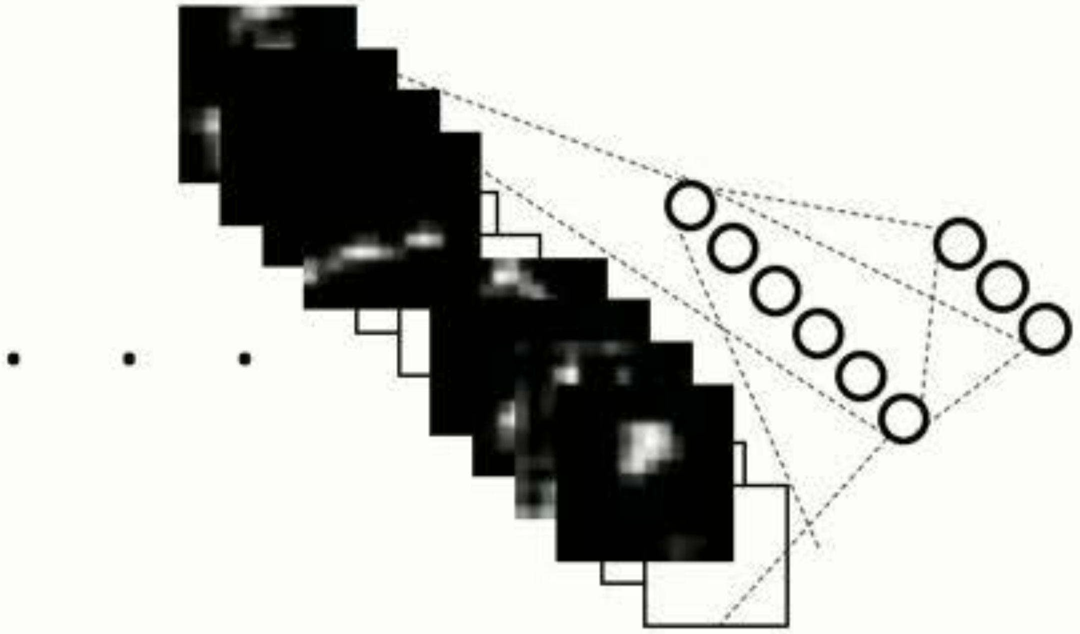
Higher layers look for semantic features



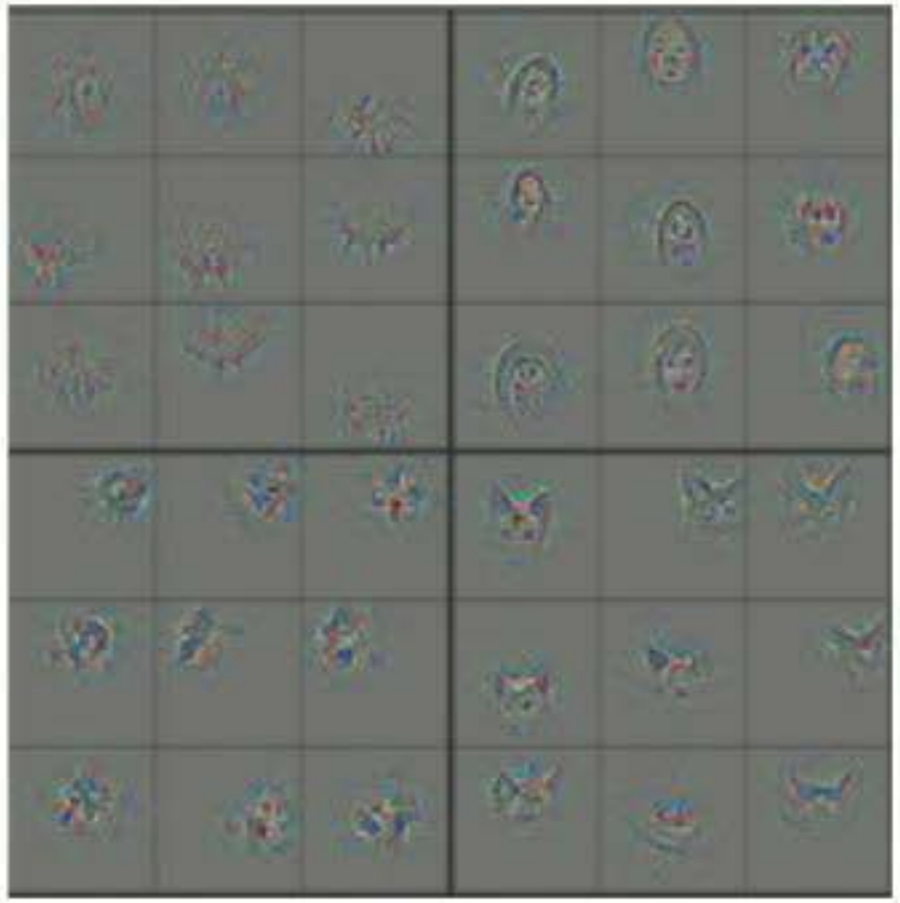
Higher layers look for semantic features



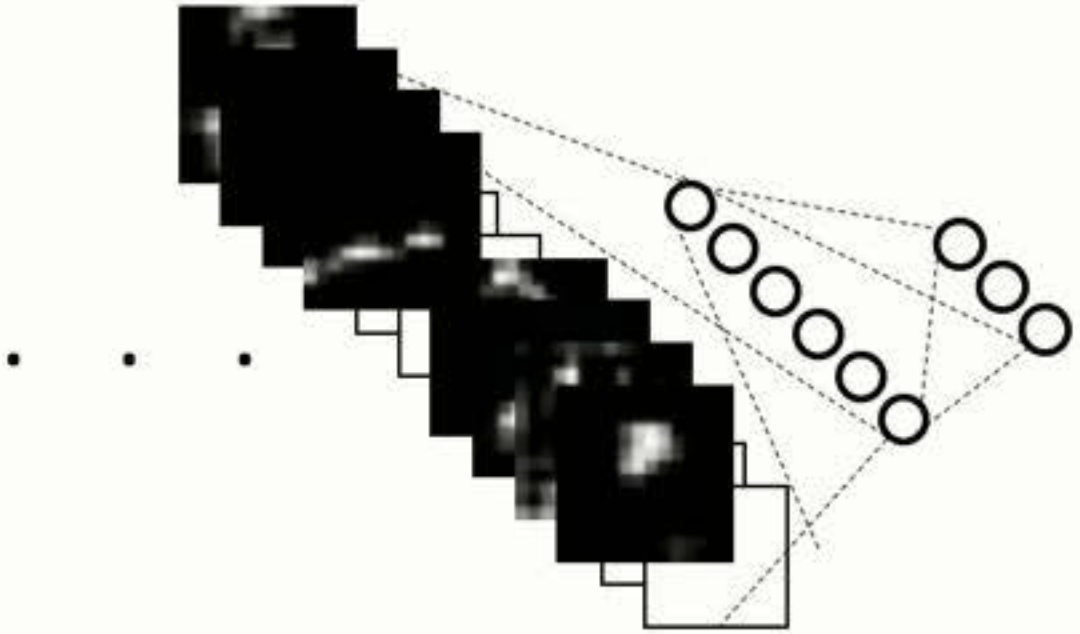
Higher layers look for semantic features



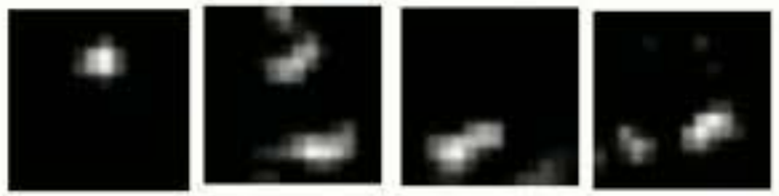
Conv 5

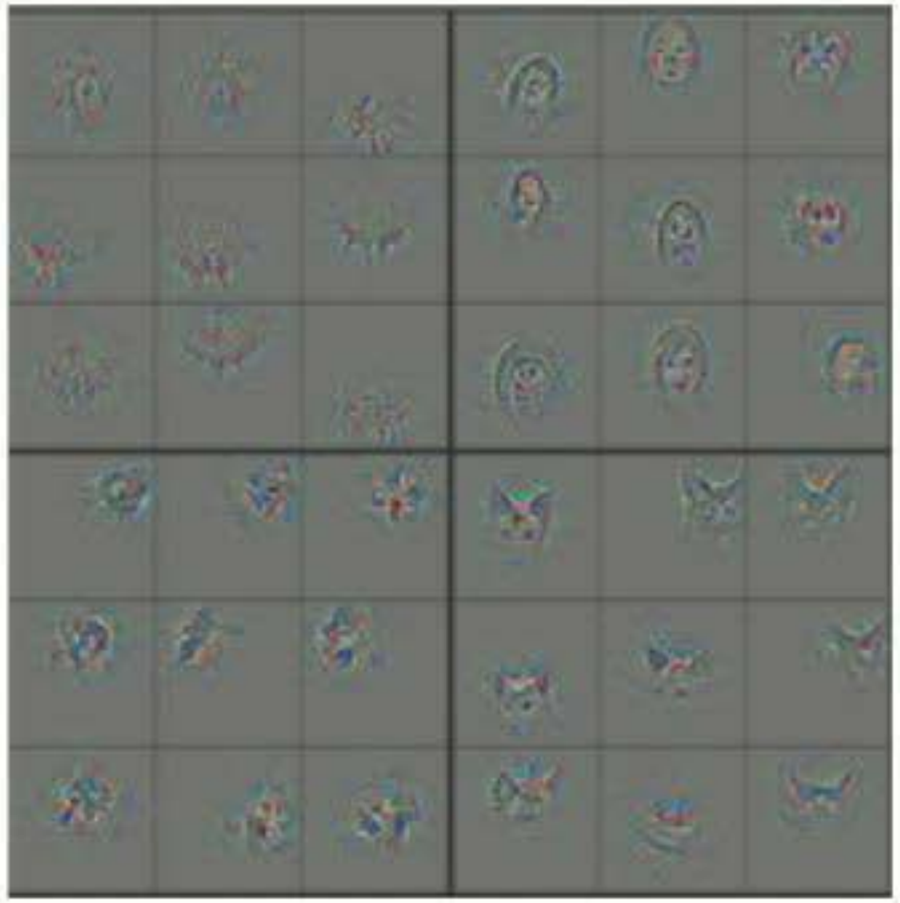


Higher layers look for semantic features



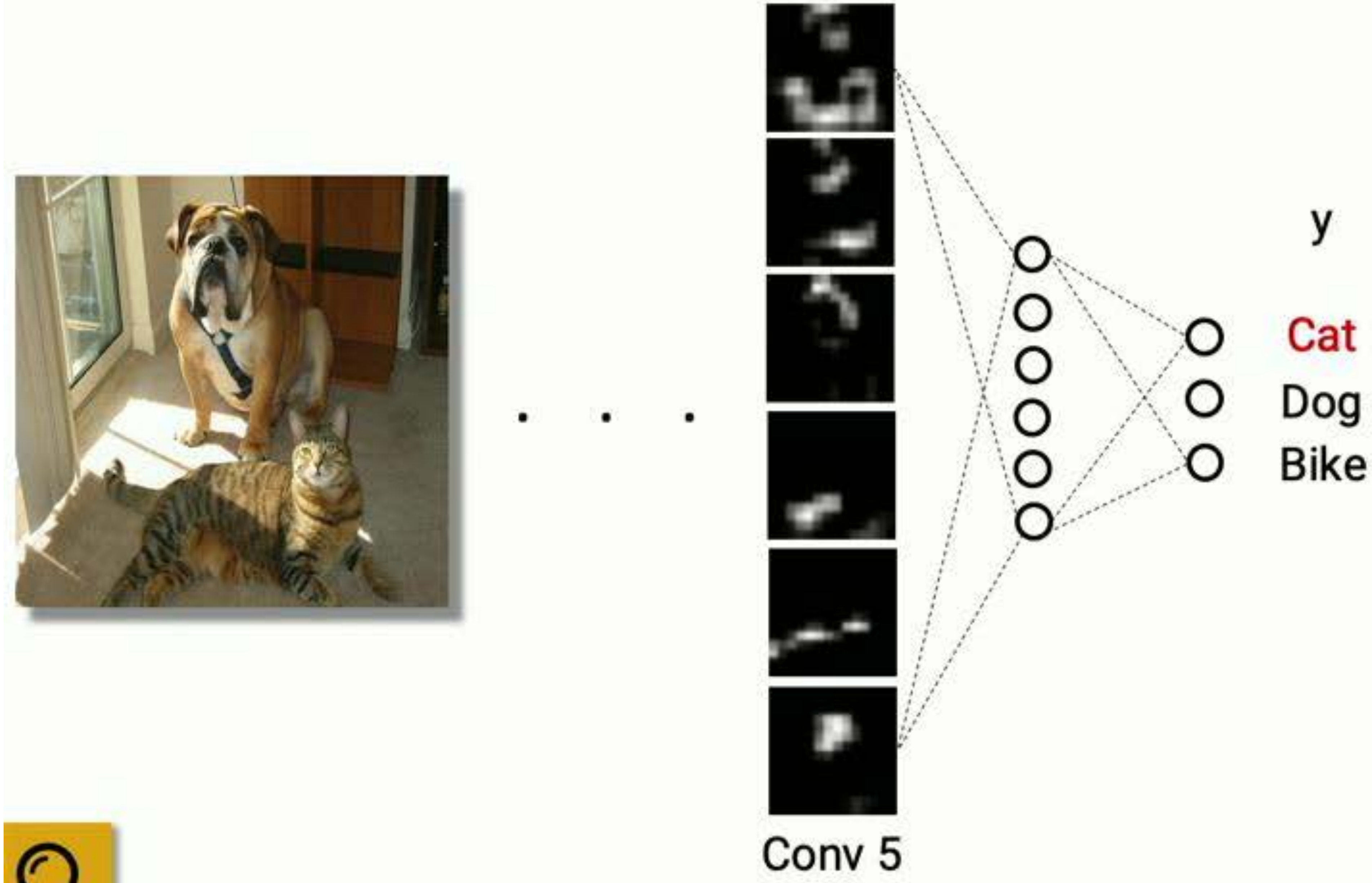
Conv 5





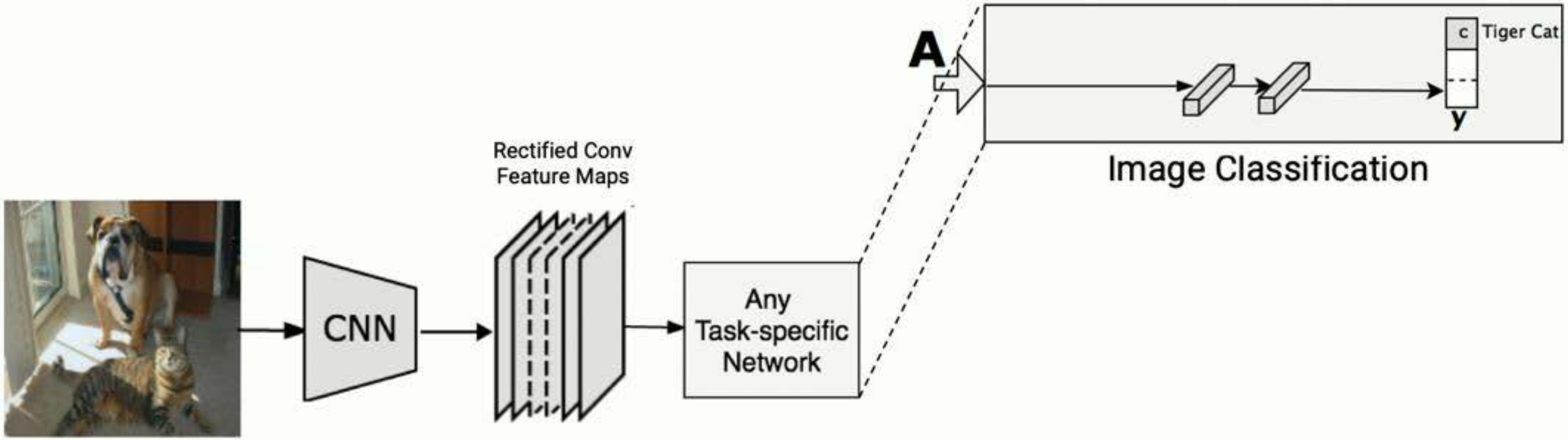
Importance in terms of later layer neurons make more sense

Neuron Importance of higher layers



Grad-CAM

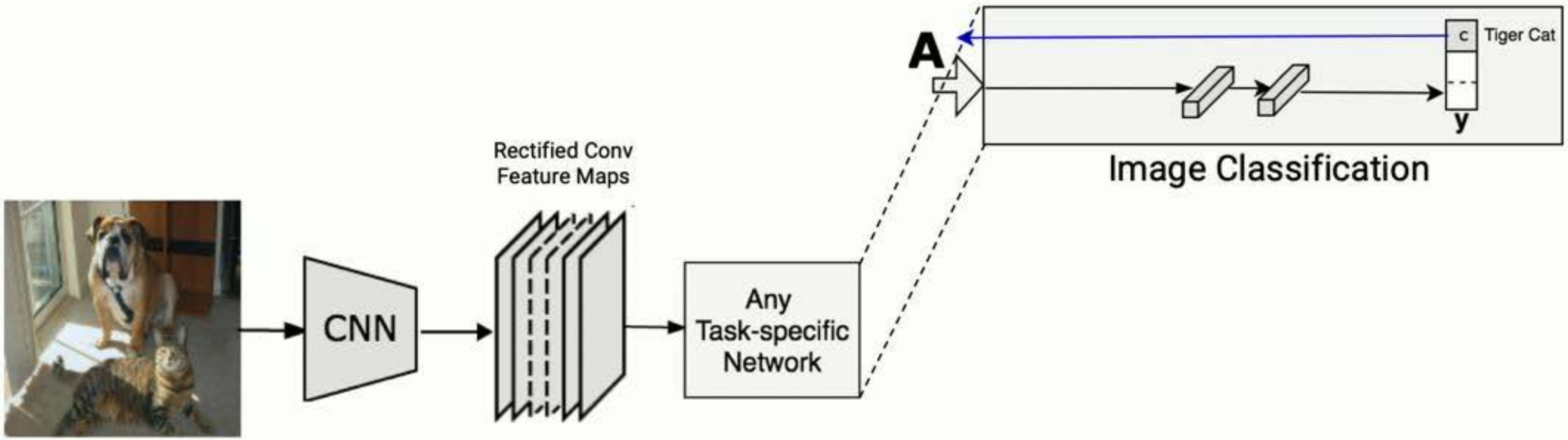
Grad-CAM


CNN
Rectified Conv
Feature Maps
Any
Task-specific
Network
A
c
Tiger Cat
y
Image Classification

Grad-CAM

← Gradients

→ Activations



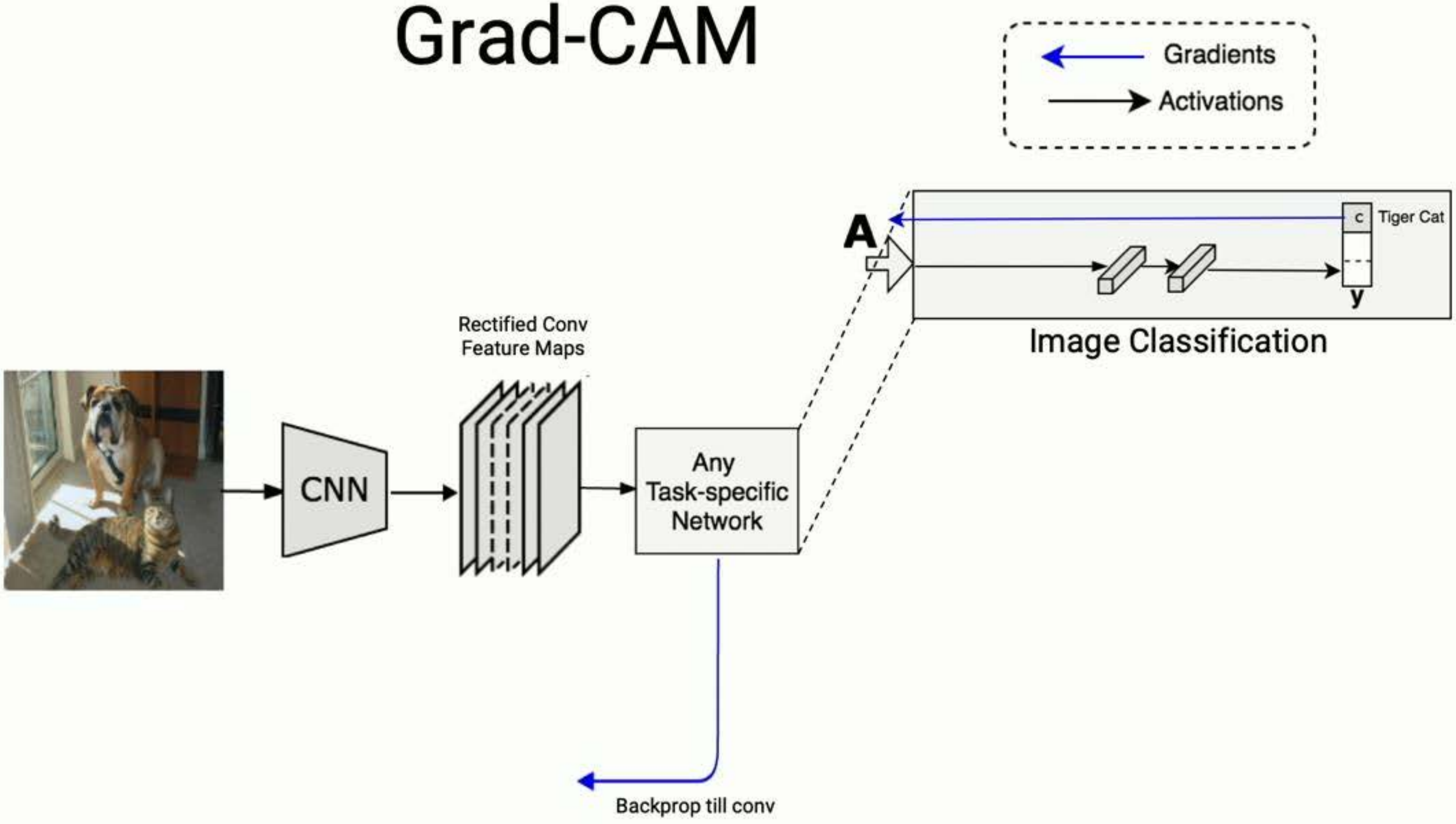
Rectified Conv Feature Maps

CNN

Any Task-specific Network

Image Classification

Grad-CAM



Gradients

Activations

Rectified Conv
Feature Maps

CNN

Any
Task-specific
Network

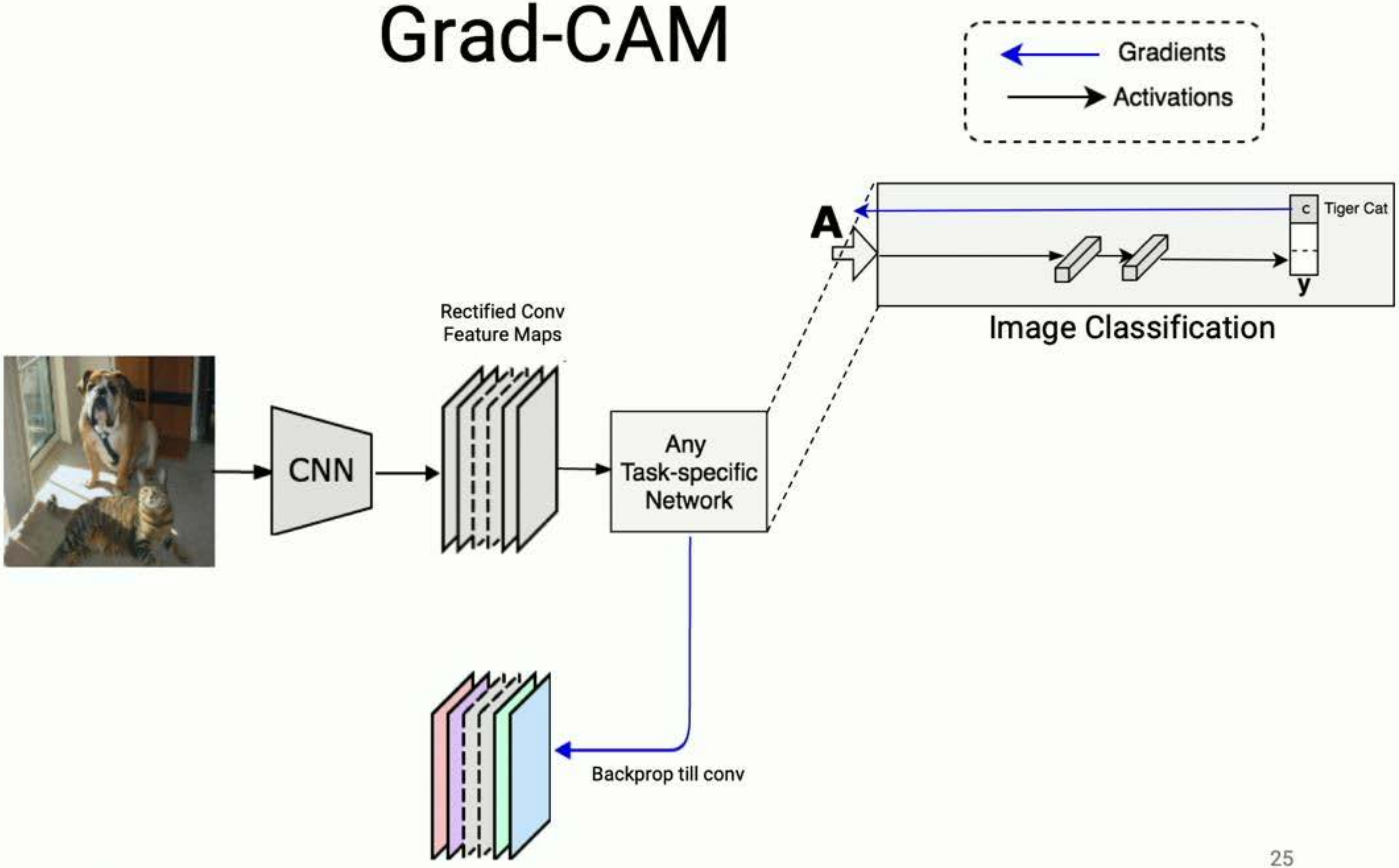
A

Tiger Cat

Image Classification

Backprop till conv

Grad-CAM



Gradients

Activations

A

c

Tiger Cat

y

Image Classification

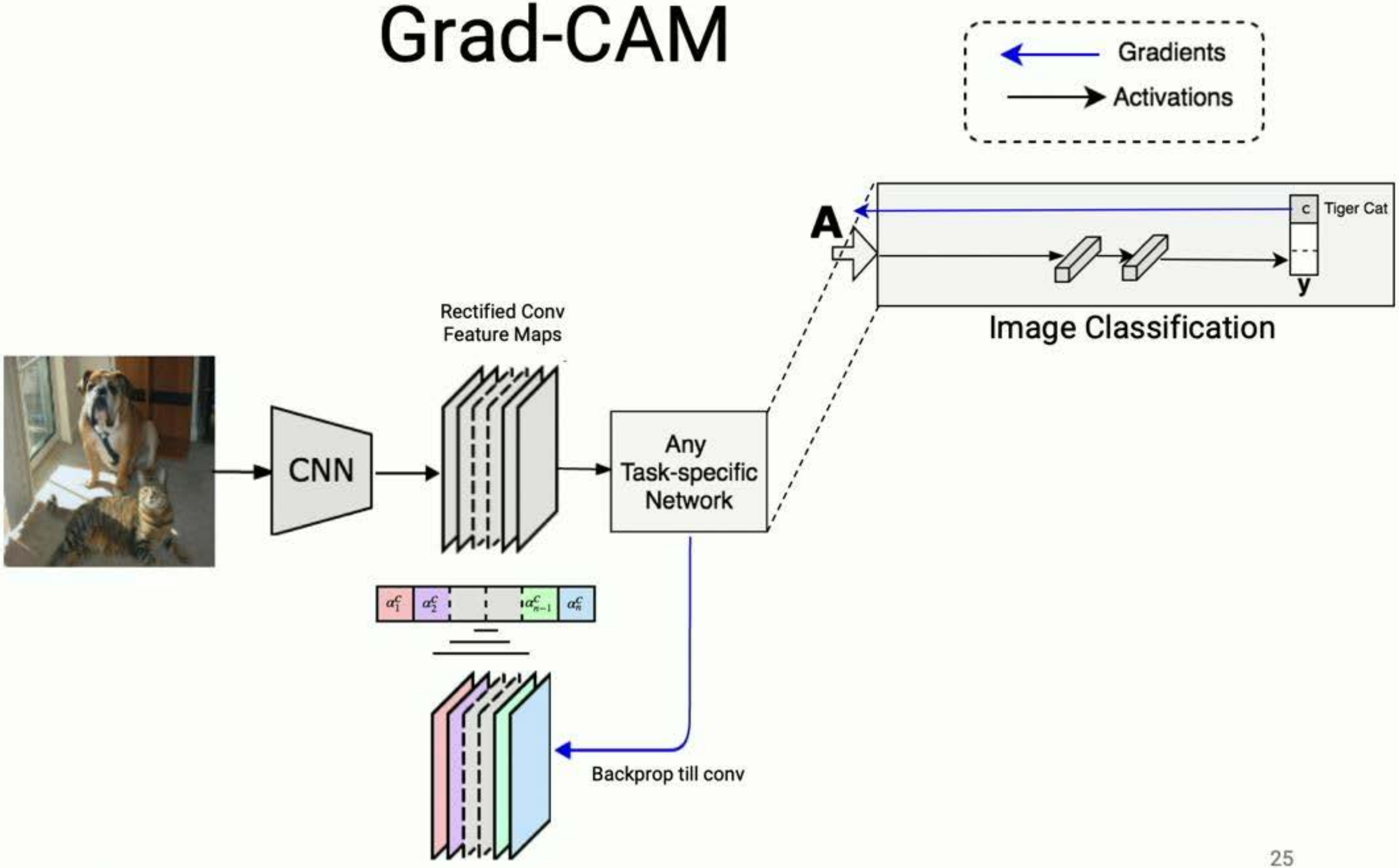
Rectified Conv Feature Maps

CNN

Any Task-specific Network

Backprop till conv

Grad-CAM



Gradients

Activations

A

c Tiger Cat

y

Image Classification

Rectified Conv Feature Maps

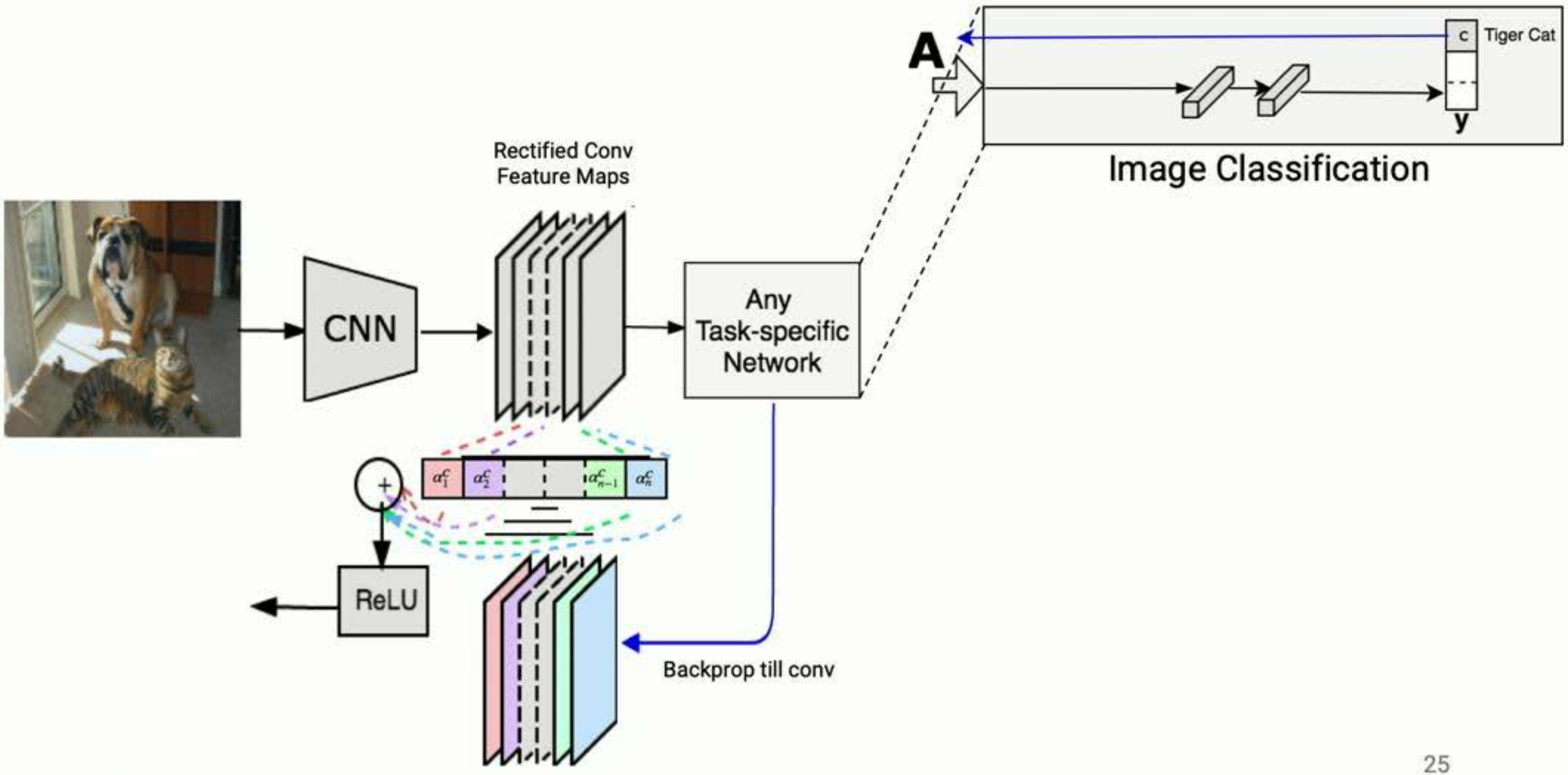
CNN

Any Task-specific Network

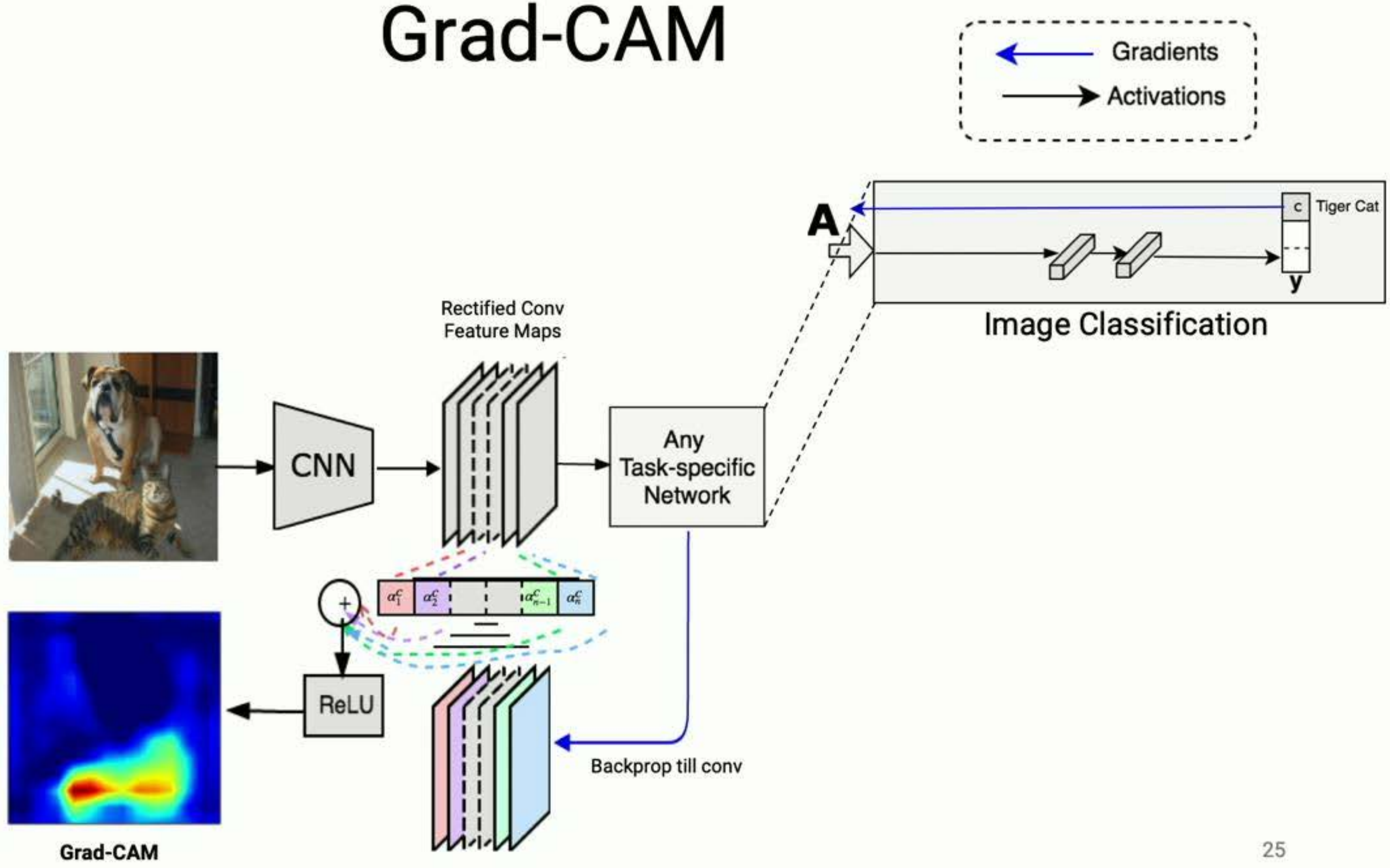
α_1^c α_2^c α_{n-1}^c α_n^c

Backprop till conv

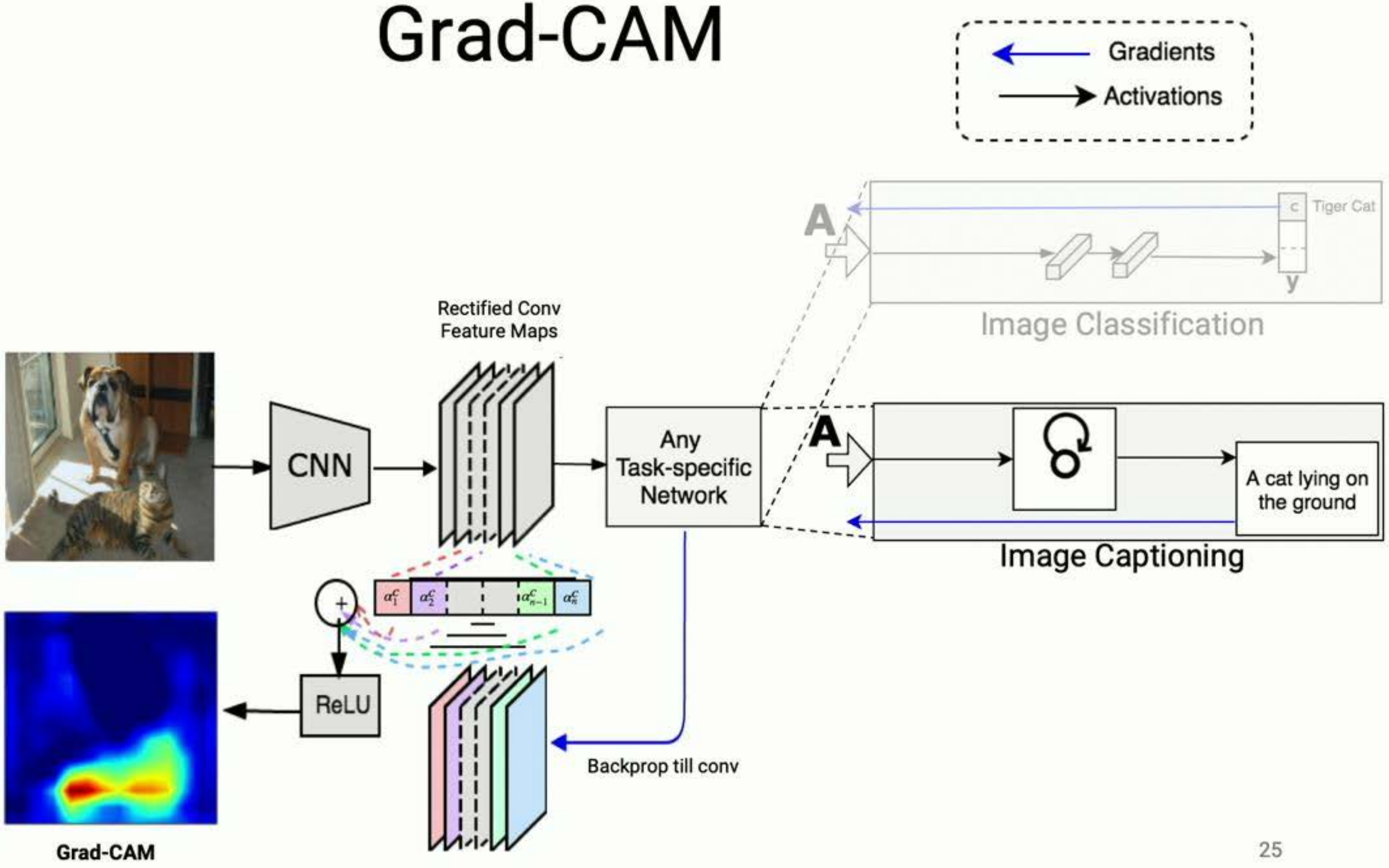
Grad-CAM


Gradients
Activations
A
c
Tiger Cat
y
Image Classification
Rectified Conv
Feature Maps
CNN
Any
Task-specific
Network
α_1^c
α_2^c
α_{n-1}^c
α_n^c
+
ReLU
Backprop till conv

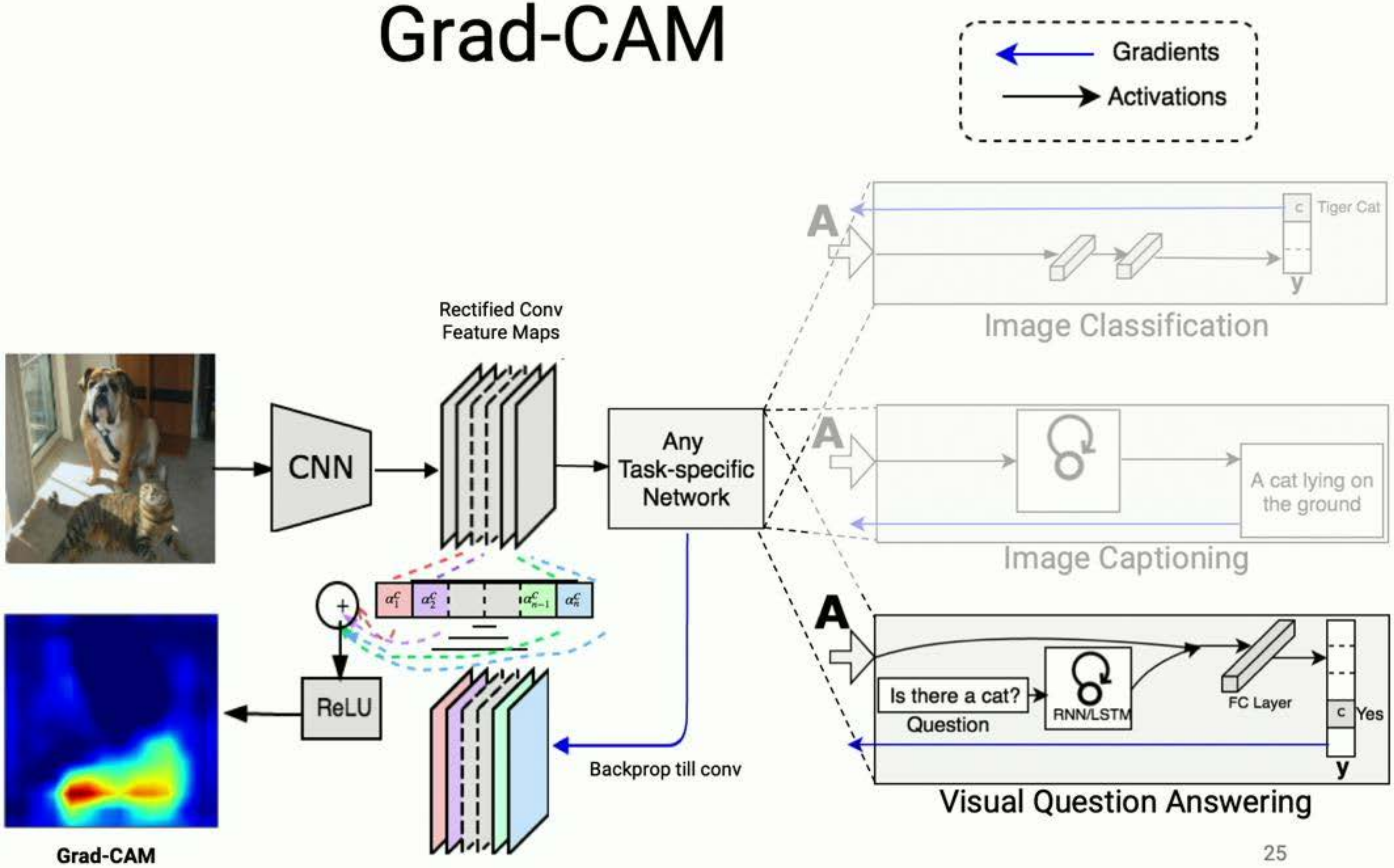
Grad-CAM


Gradients
Activations
A
c
Tiger Cat
y
Image Classification
Rectified Conv Feature Maps
CNN
Any Task-specific Network
α_1^c
α_2^c
α_{n-1}^c
α_n^c
+
ReLU
Backprop till conv
Grad-CAM

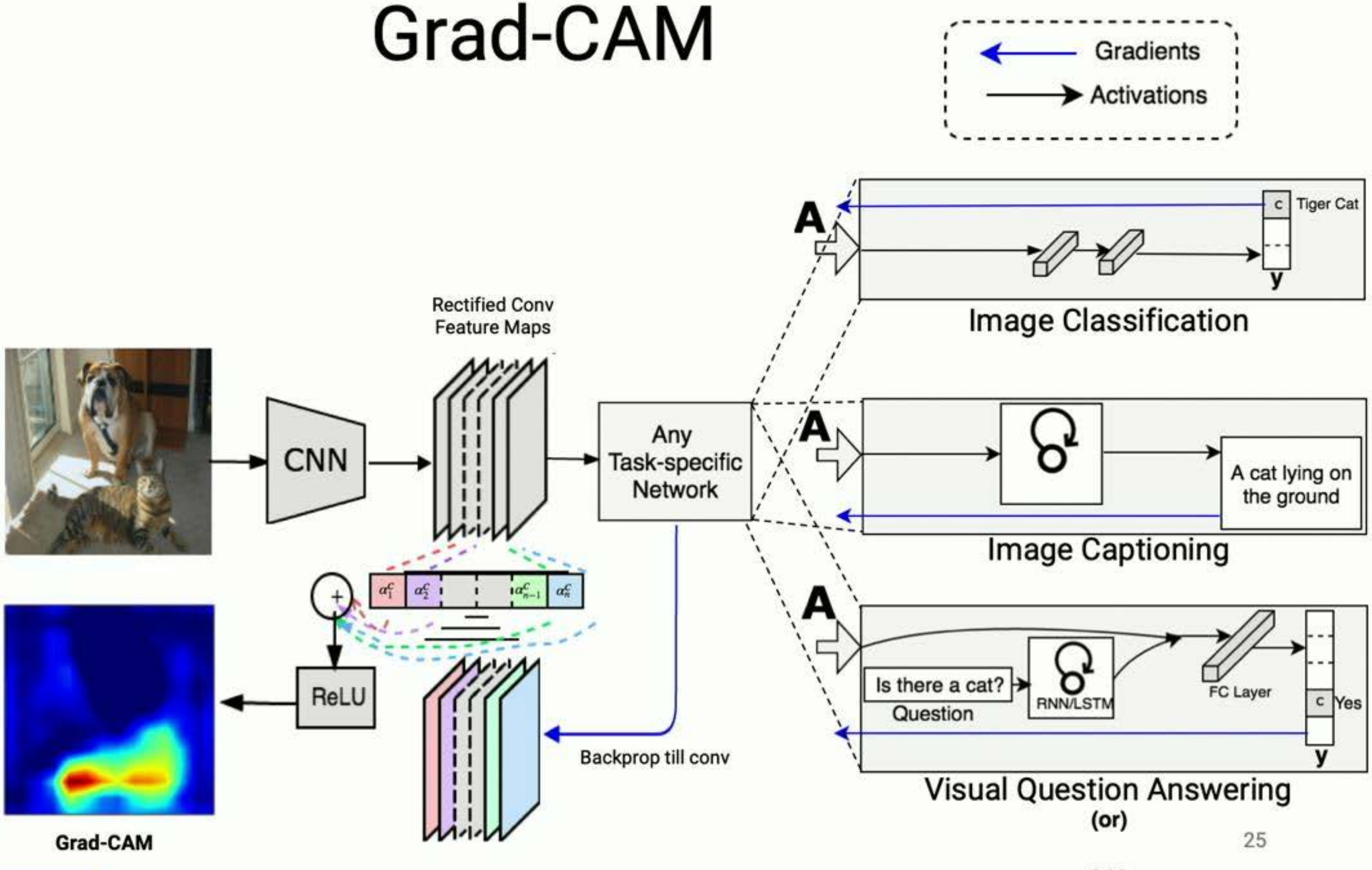
Grad-CAM


Gradients
Activations
A
c
Tiger Cat
y
Image Classification
Rectified Conv Feature Maps
CNN
Any Task-specific Network
A
A cat lying on the ground
Image Captioning
α_1^c
α_2^c
α_{n-1}^c
α_n^c
+
ReLU
Backprop till conv
Grad-CAM

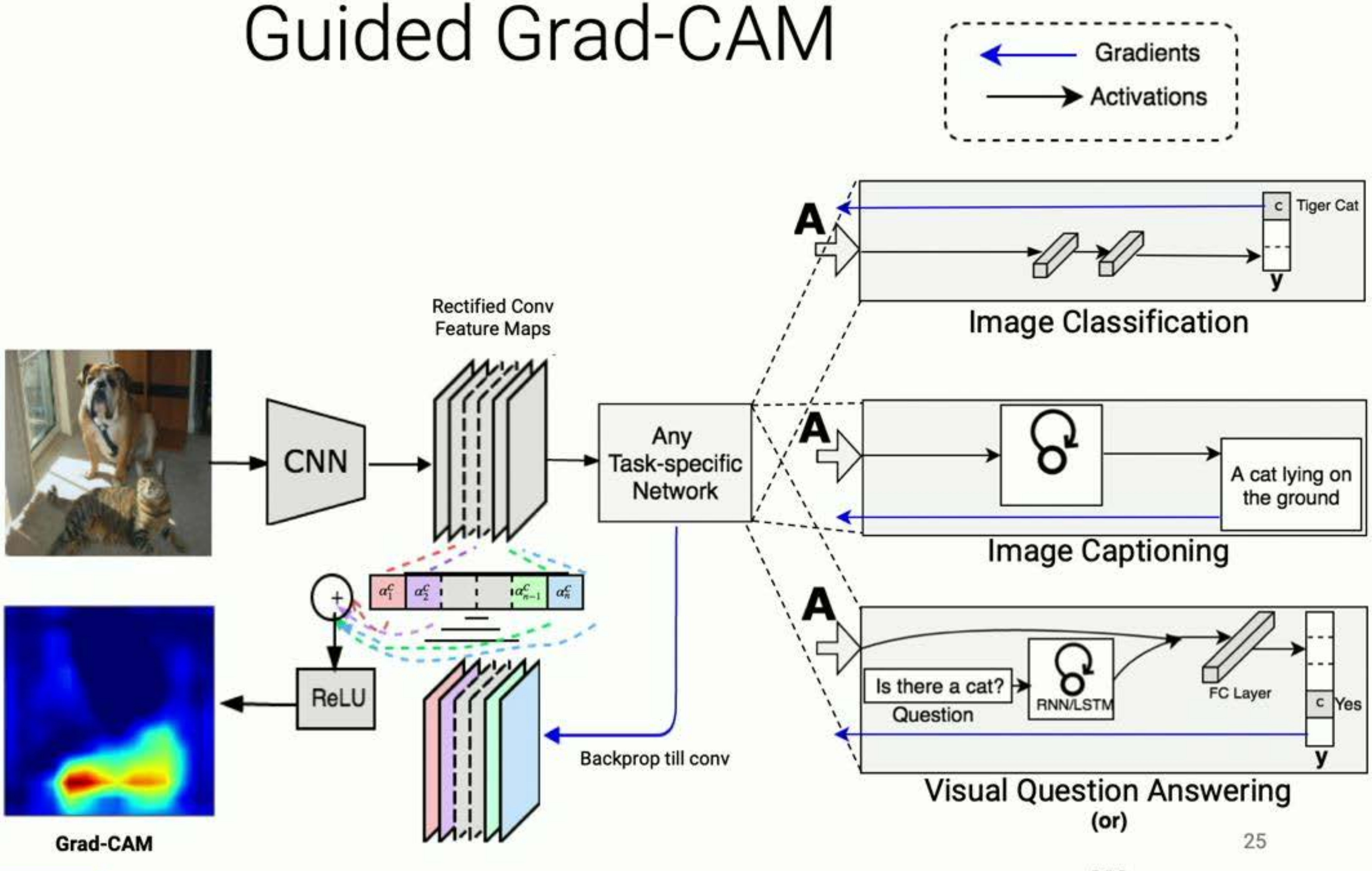
Grad-CAM

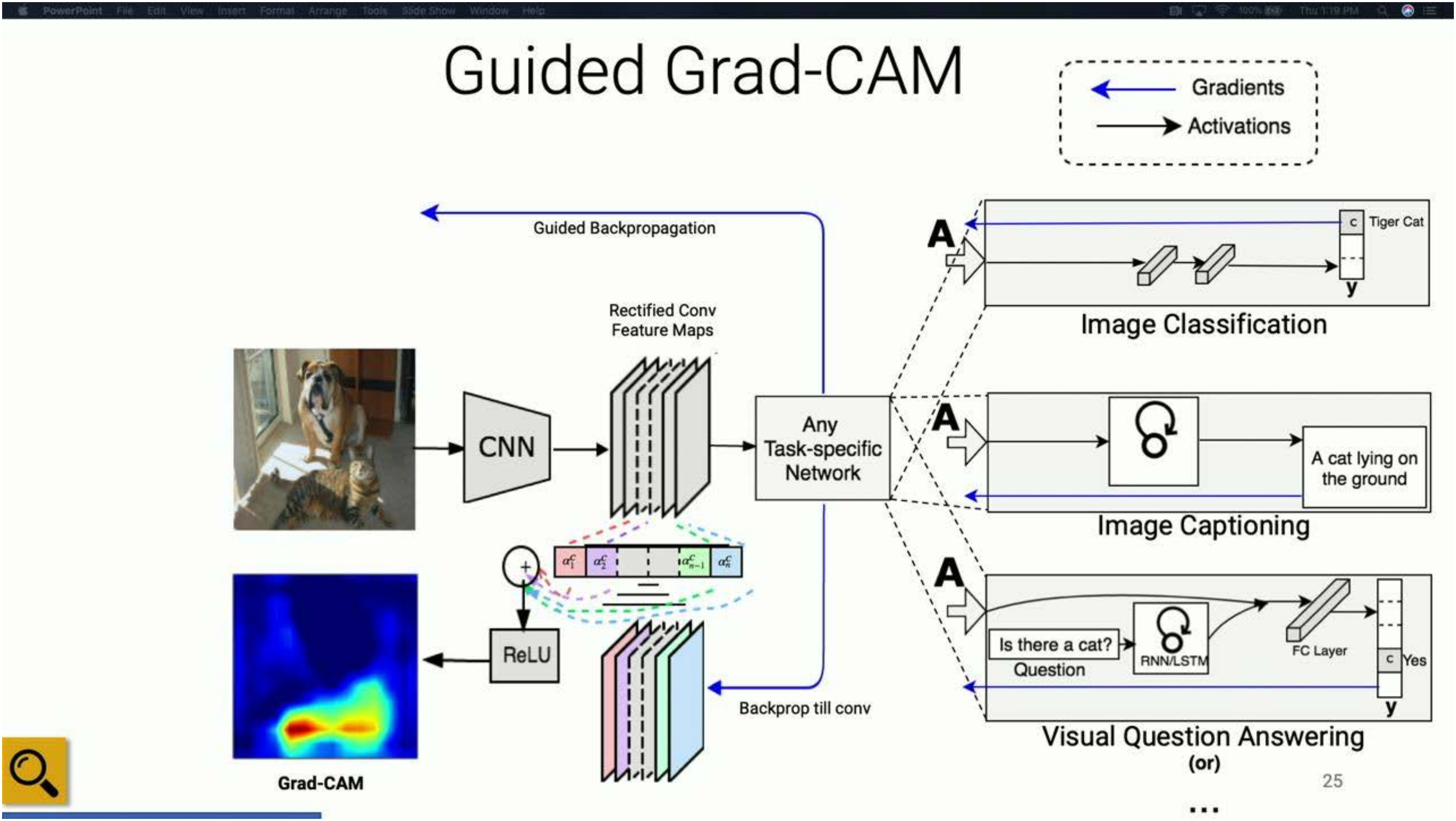

Gradients
Activations
Rectified Conv Feature Maps
CNN
Any Task-specific Network
α_1^c
α_2^c
α_{n-1}^c
α_n^c
+
ReLU
Backprop till conv
Grad-CAM
A
c
Tiger Cat
y
Image Classification
A
A cat lying on the ground
Image Captioning
A
Is there a cat?
Question
RNN/LSTM
FC Layer
c
Yes
y
Visual Question Answering

Grad-CAM

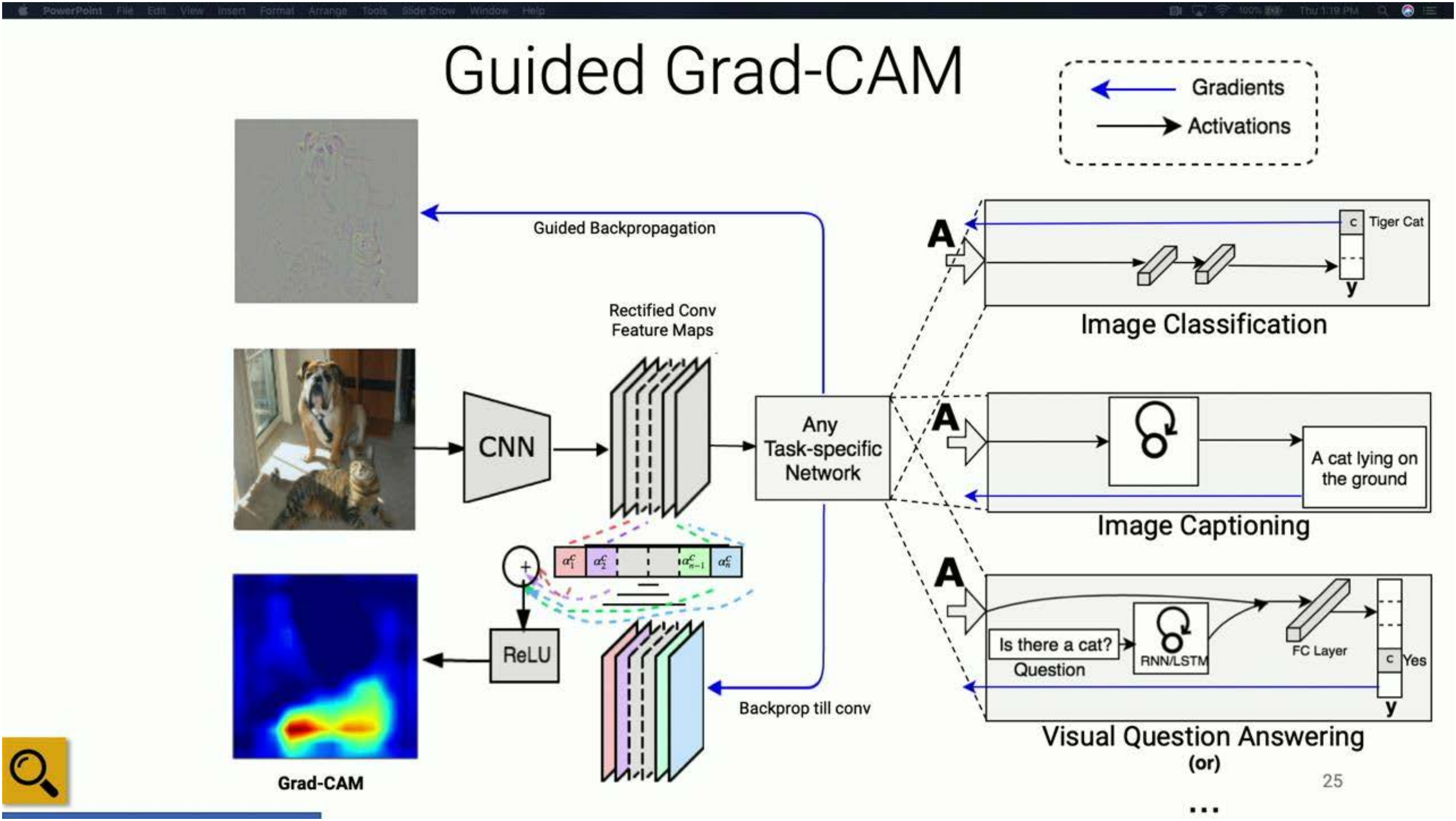

Gradients
Activations
Rectified Conv Feature Maps
CNN
Any Task-specific Network
A
c
Tiger Cat
y
Image Classification
A
A cat lying on the ground
Image Captioning
A
Is there a cat?
Question
RNN/LSTM
FC Layer
c
Yes
y
Visual Question Answering
(or)
...
α_1^c
α_2^c
α_{n-1}^c
α_n^c
+
ReLU
Backprop till conv
Grad-CAM

Guided Grad-CAM





Guided Grad-CAM



Guided Grad-CAM

Gradients

Activations

Guided Backpropagation

Rectified Conv Feature Maps

CNN

Any Task-specific Network

α_1^c α_2^c α_{n-1}^c α_n^c

ReLU

Backprop till conv

Grad-CAM

A

c Tiger Cat

y

Image Classification

A

A cat lying on the ground

Image Captioning

A

Is there a cat?

Question

RNN/LSTM

FC Layer

c Yes

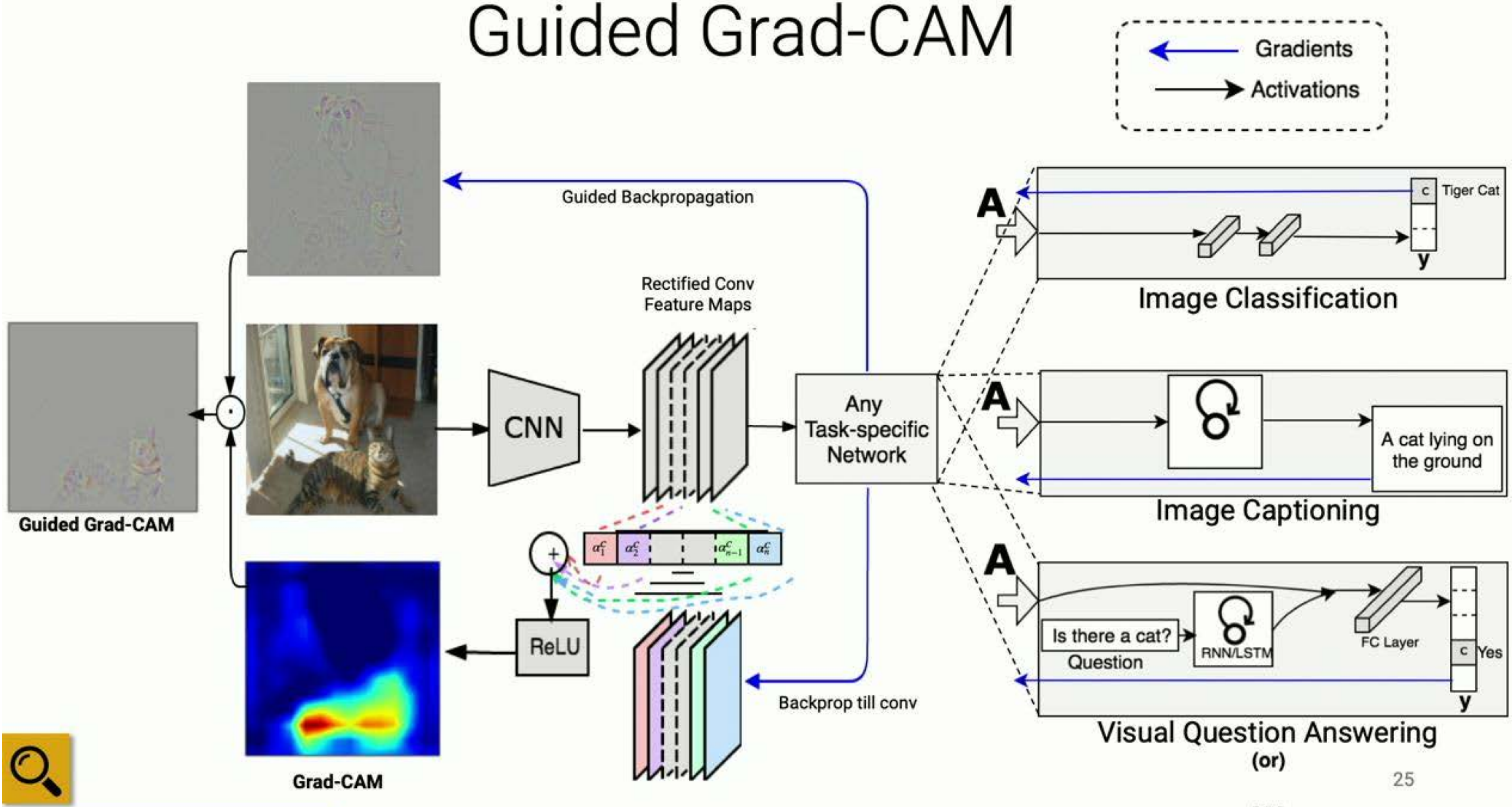
y

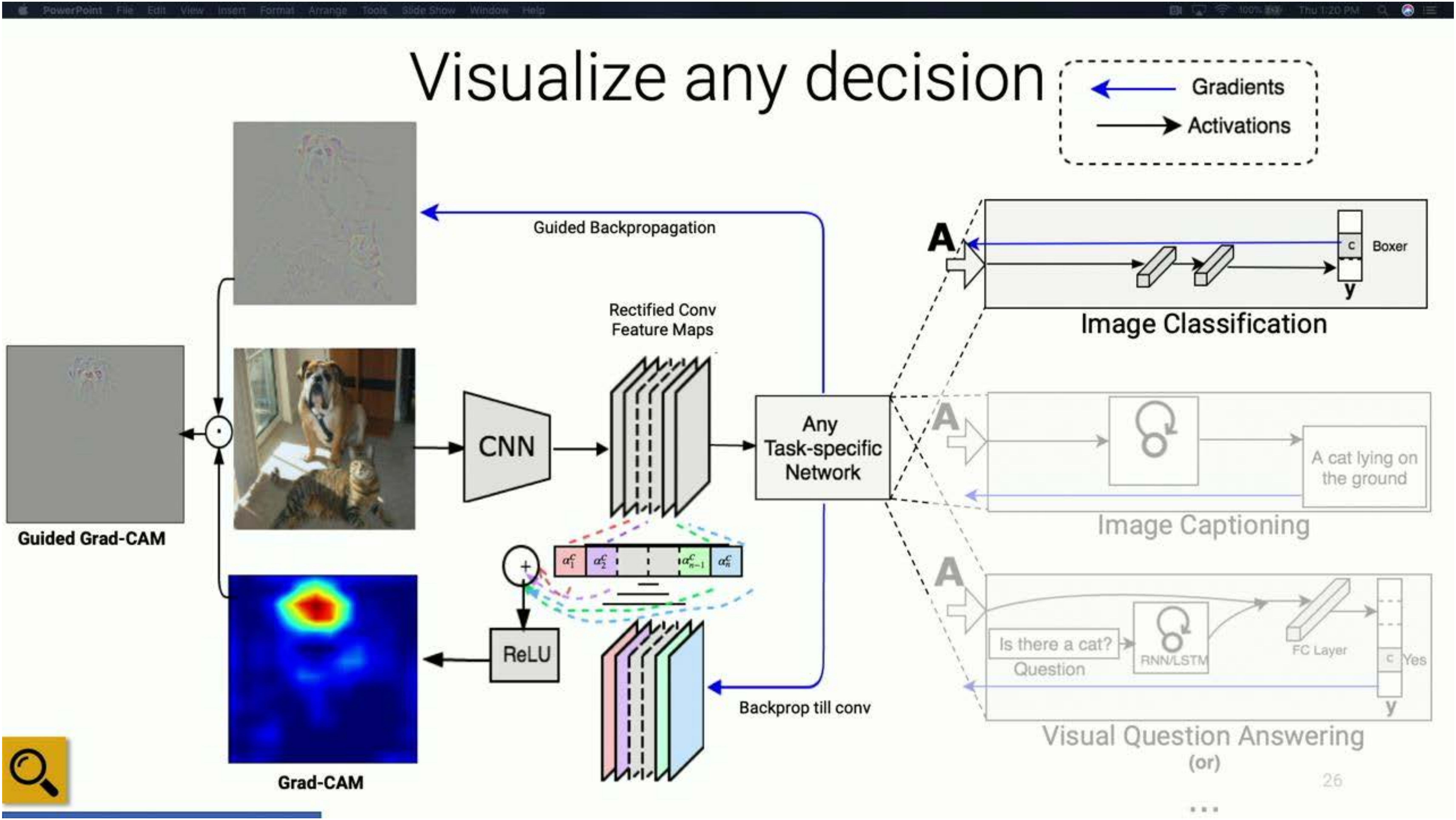
Visual Question Answering

(or)

...

Guided Grad-CAM





Visualize any decision

Gradients

Activations

Guided Backpropagation

Rectified Conv Feature Maps

Guided Grad-CAM

CNN

Any Task-specific Network

α_1^c α_2^c α_{n-1}^c α_n^c

ReLU

Grad-CAM

Backprop till conv

A

c

Boxer

y

Image Classification

A

A cat lying on the ground

Image Captioning

A

Is there a cat?

Question

RNN/LSTM

FC Layer

c

Yes

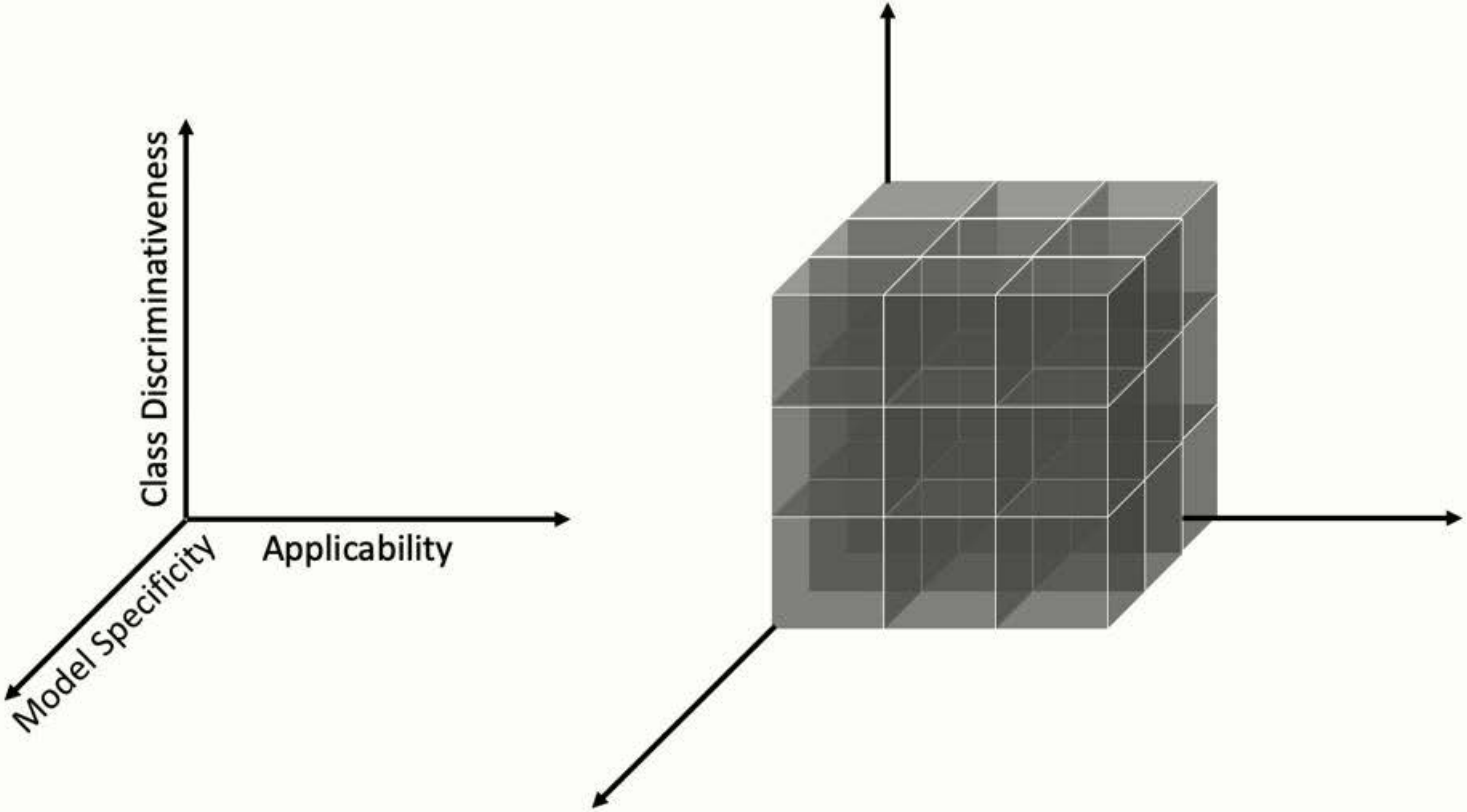
y

Visual Question Answering

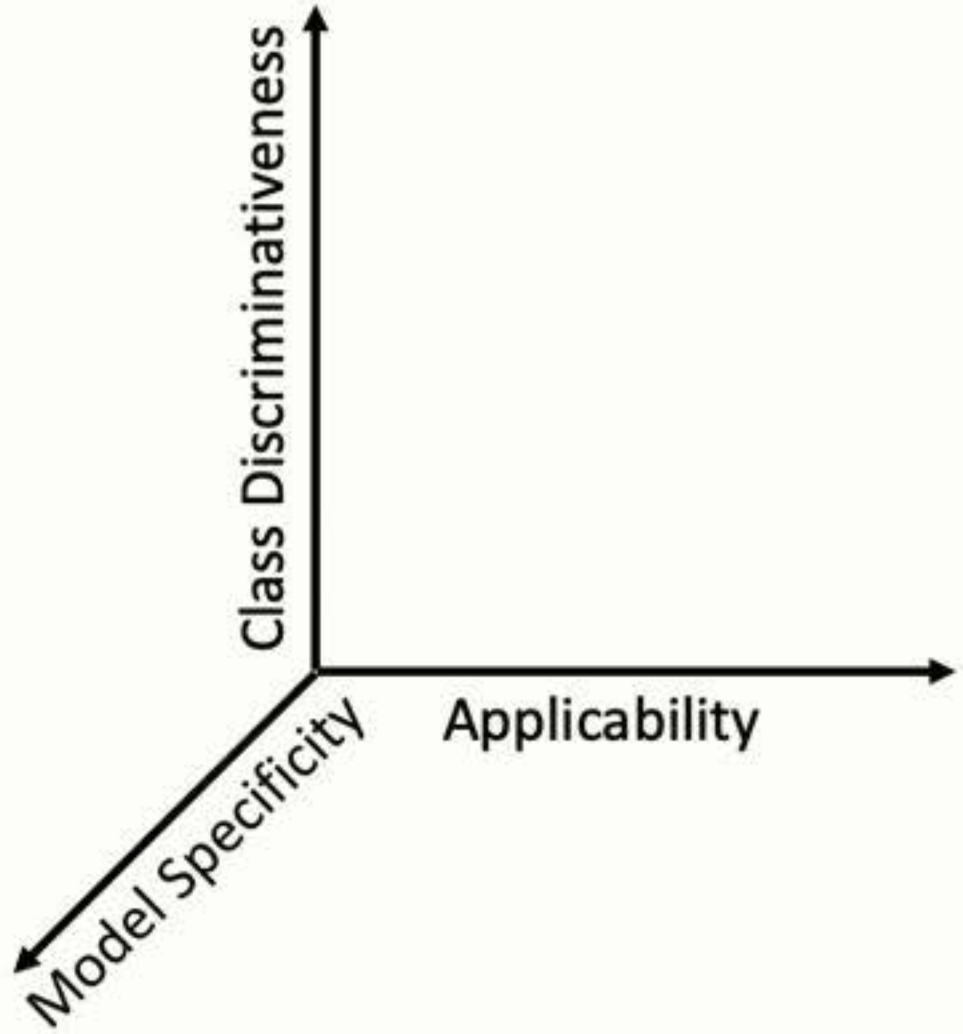
(or)

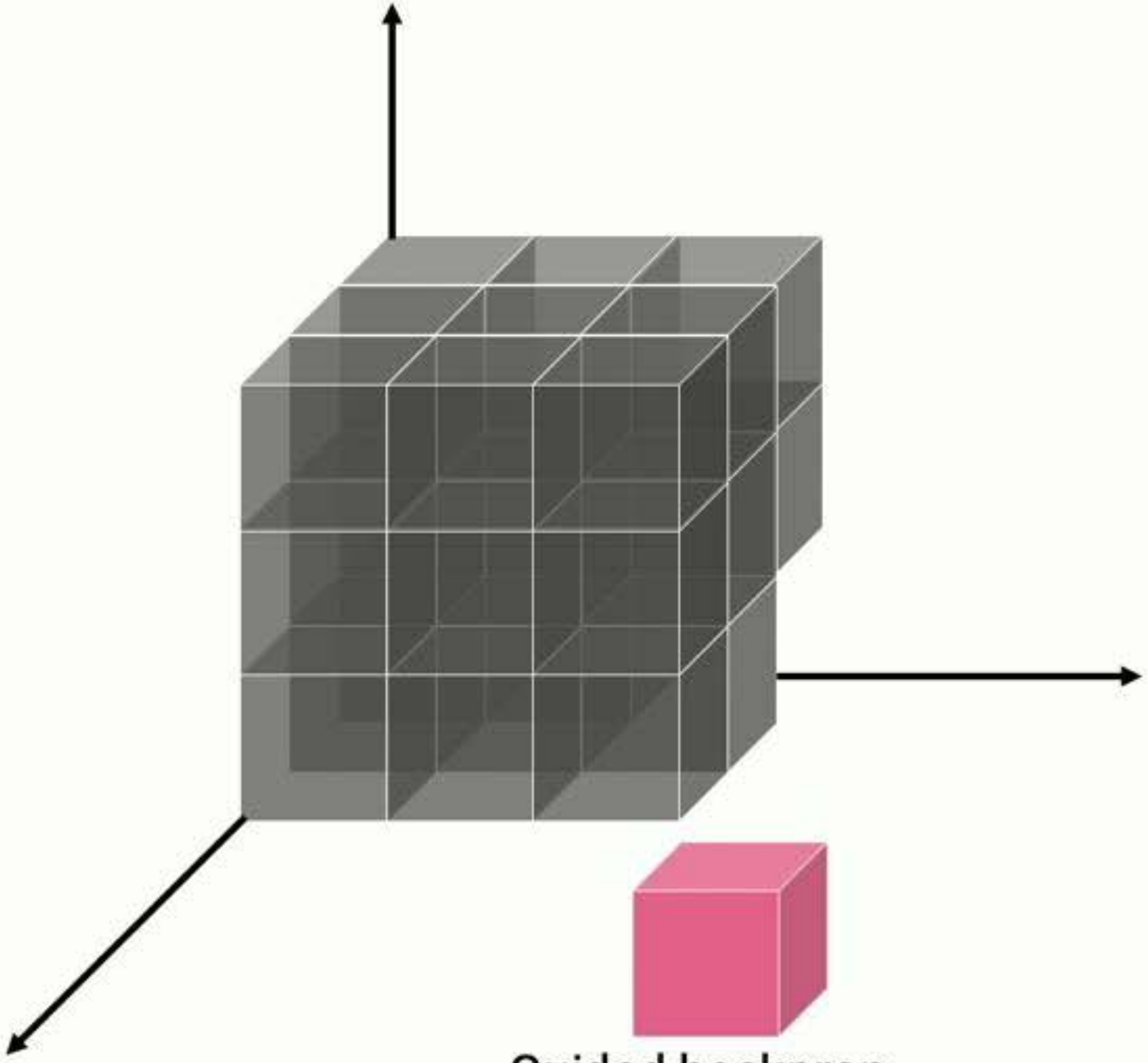
...

Advantages of Grad-CAM



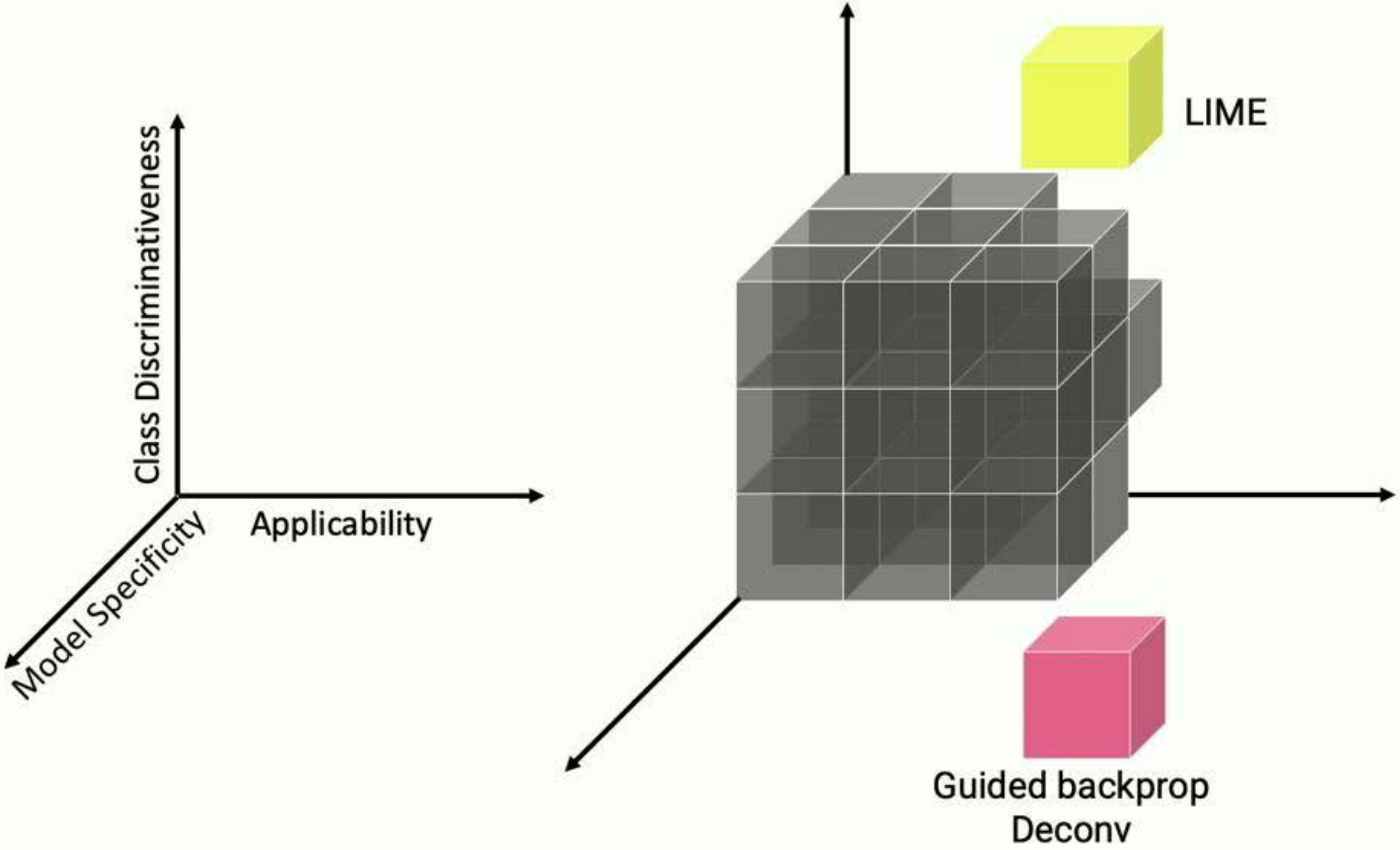
Advantages of Grad-CAM



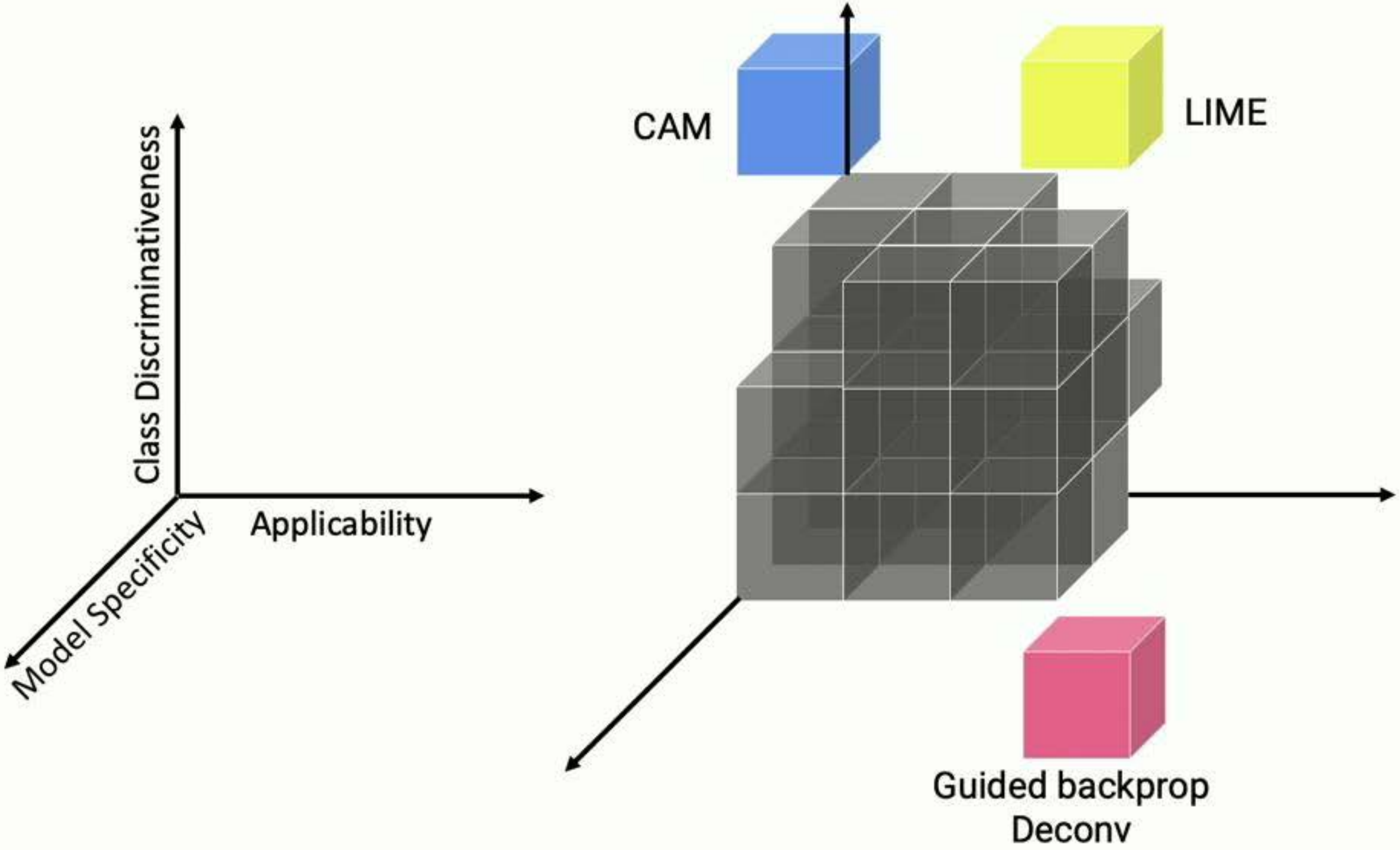


Guided backprop
Deconv

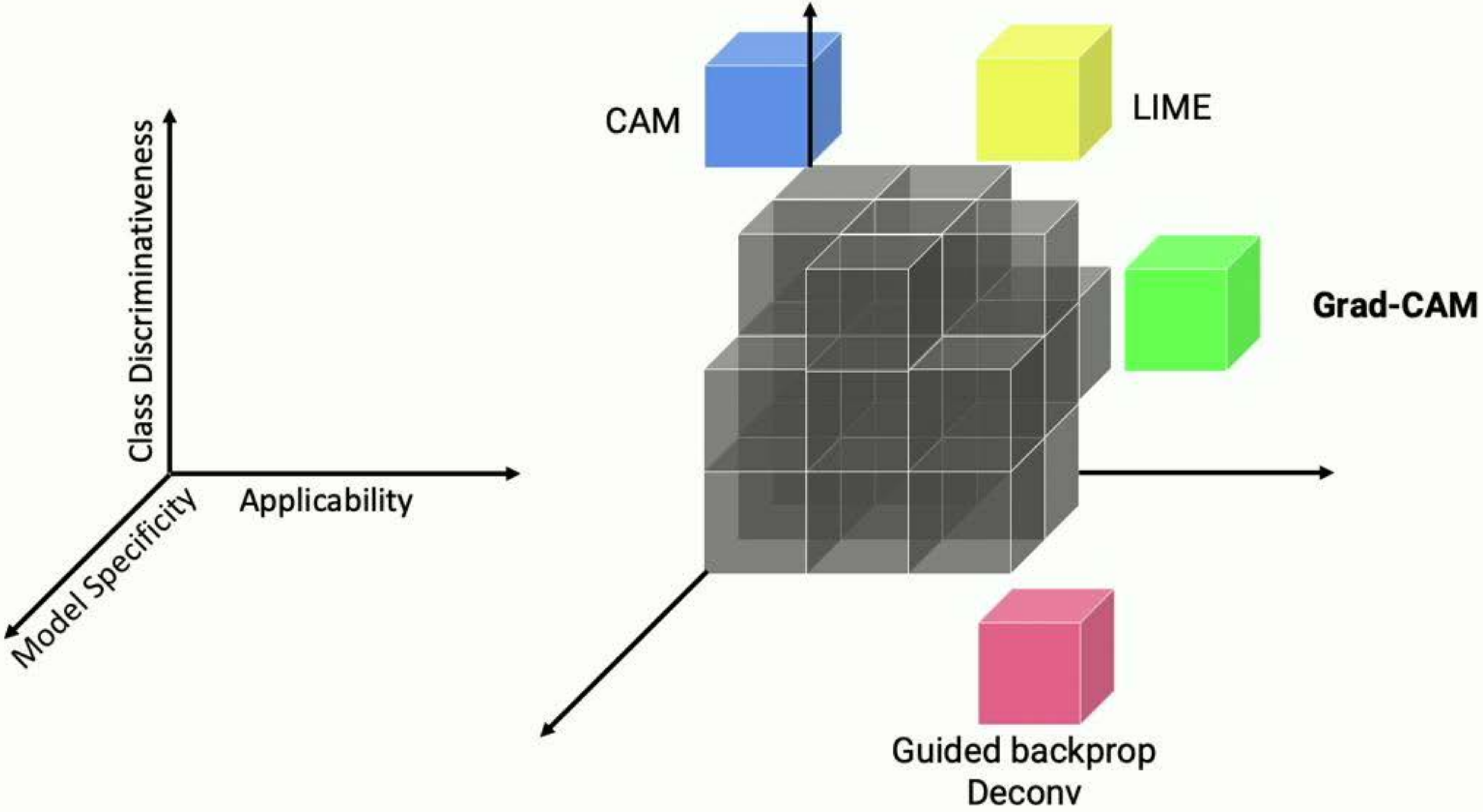
Advantages of Grad-CAM



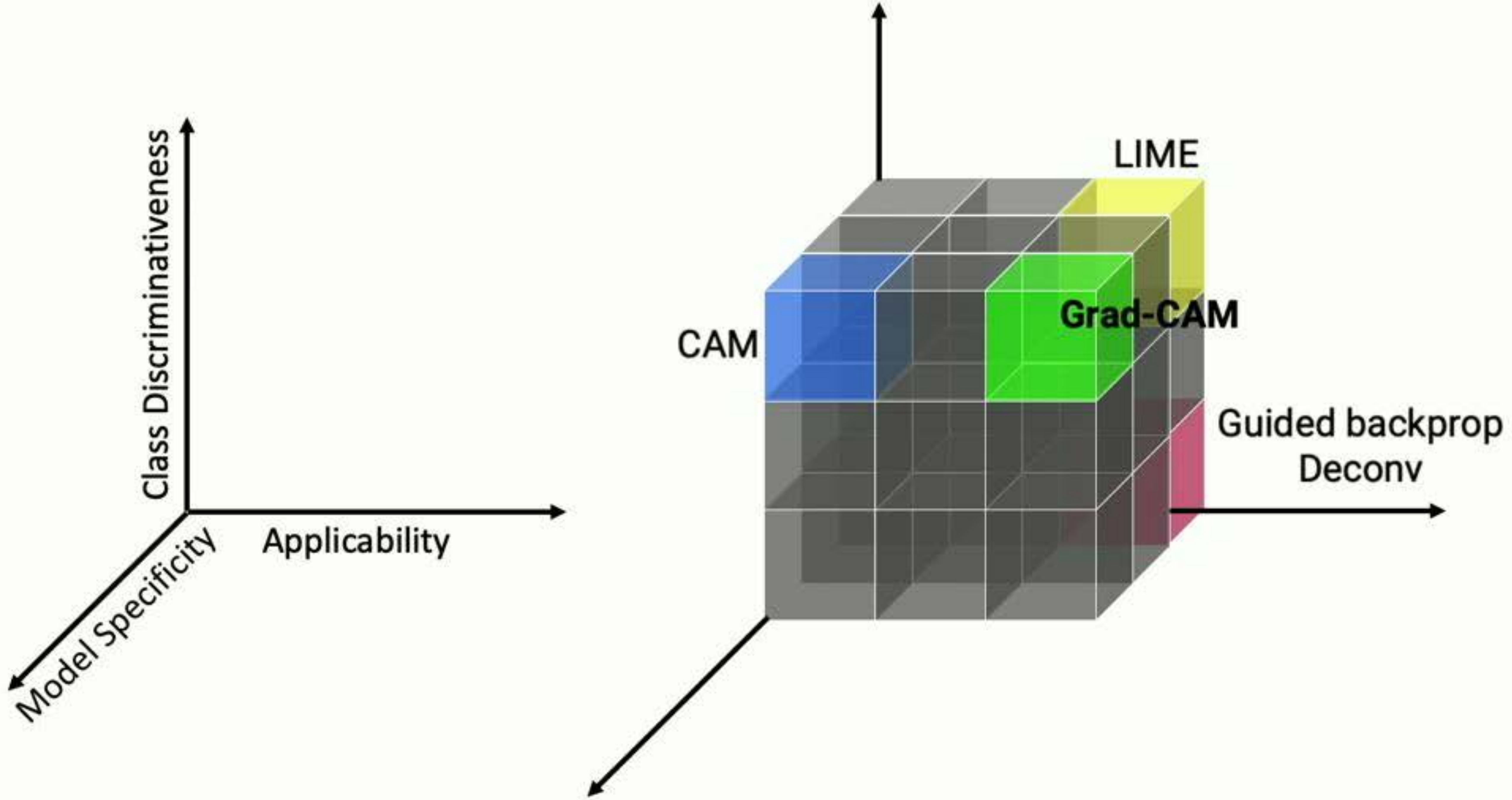
Advantages of Grad-CAM



Advantages of Grad-CAM



Advantages of Grad-CAM



Evaluating Explanations

Evaluating Explanations

- Interpretability
  - How interpretable are explanations to humans?
- Faithfulness
  - How faithful are the explanations to the underlying model?
- Trustworthiness
  - Can visualizations help establish user trust?

Evaluating Interpretability

- Can visualizations tell users which class is being visualized?

What do you see?

○ Horse
○ Person

Evaluating Interpretability

- Can visualizations tell users which class is being visualized?

What do you see?

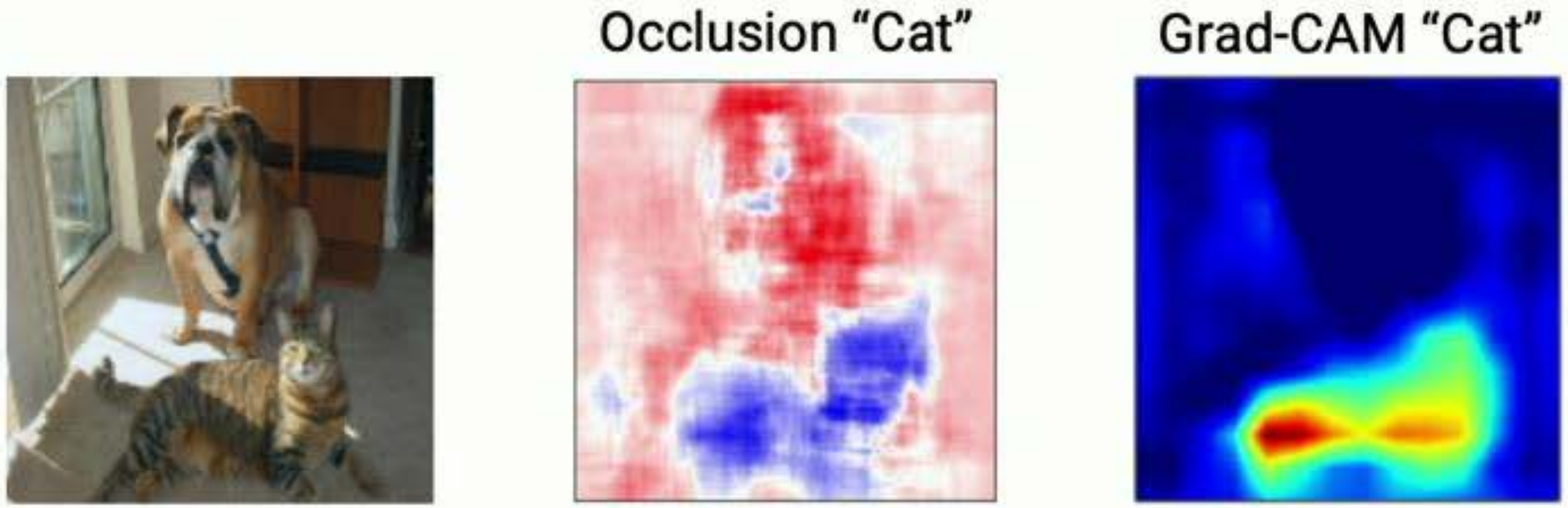
○ Horse
○ Person

Method	**Human Classification accuracy**
Guided Backpropagation	44.44
Guided Grad-CAM	61.23

Grad-CAM is Class-discriminative

Evaluating Faithfulness

- Comparison: Occlude patches of input image and see how it affects decision


Occlusion “Cat”

Grad-CAM “Cat”

Method	Rank Correlation with Occlusion
Guided Backpropagation	0.168
Grad-CAM	0.254
Guided Grad-CAM	**0.261**

Grad-CAM portrays the model more accurately

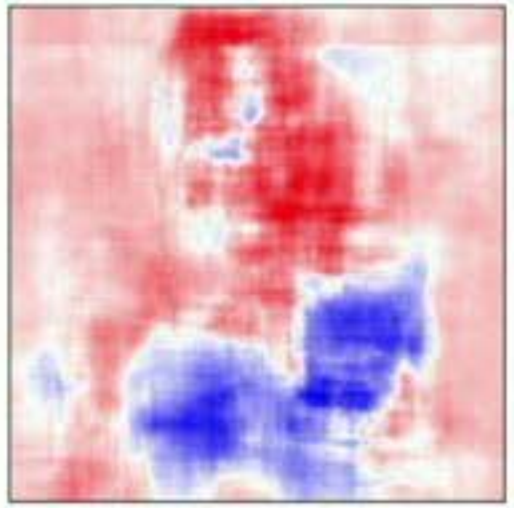
Evaluating Trustworthiness

- Can visualizations help establish user trust?

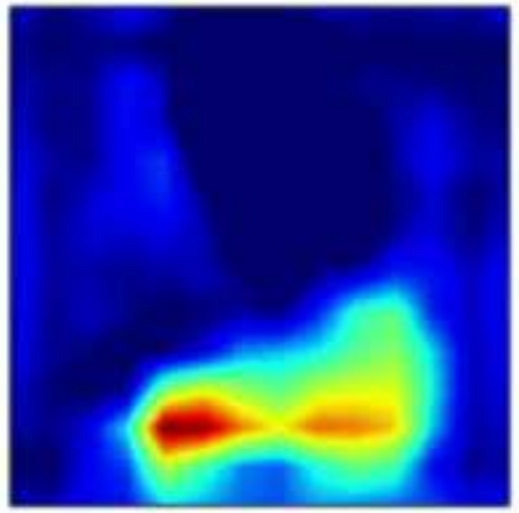
Evaluating Faithfulness

- Comparison: Occlude patches of input image and see how it affects decision

Occlusion “Cat”



Grad-CAM “Cat”



Method	Rank Correlation with Occlusion
Guided Backpropagation	0.168
Grad-CAM	0.254
Guided Grad-CAM	**0.261**

Grad-CAM portrays the model more accurately

Evaluating Trustworthiness

- Can visualizations help establish user trust?

Both robots predicted: Person

Robot A based its decision on

Robot B based its decision on

Which robot is more reasonable?

Evaluating Trustworthiness

- Can visualizations help establish user trust?

Both robots predicted: Person

Robot A based its decision on

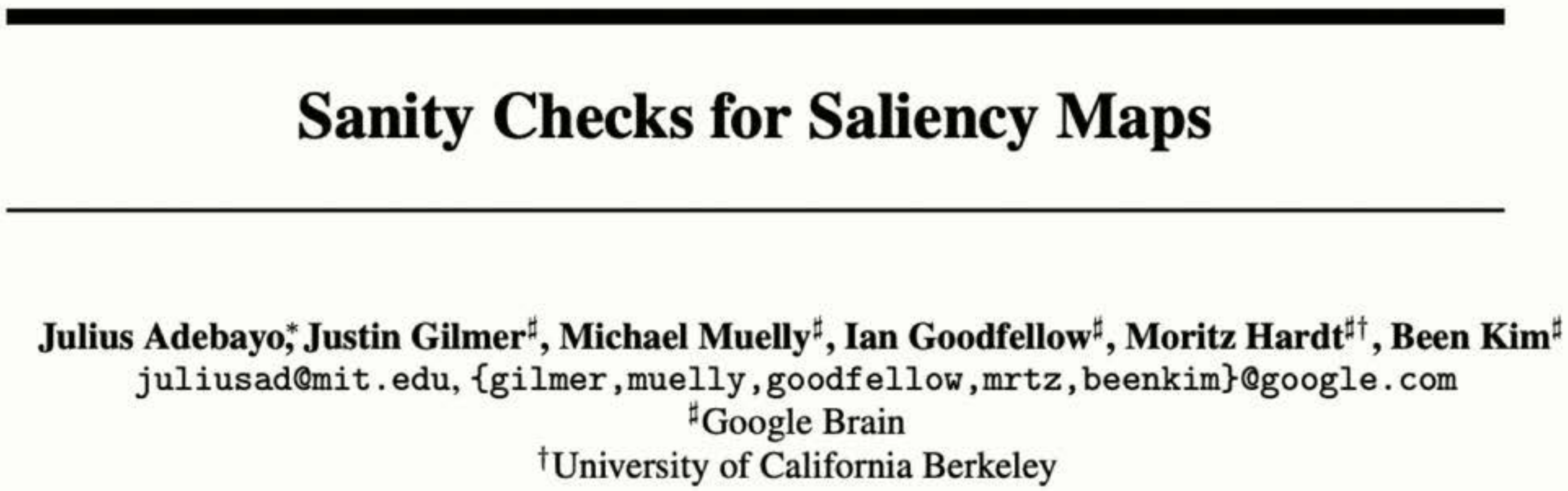
Robot B based its decision on

Which robot is more reasonable?

Method	Relative Reliability
Guided Backpropagation	+1.00
Guided Grad-CAM	+1.27

Grad-CAM helps users place higher trust in a model that generalizes better

Evaluations conducted by other papers

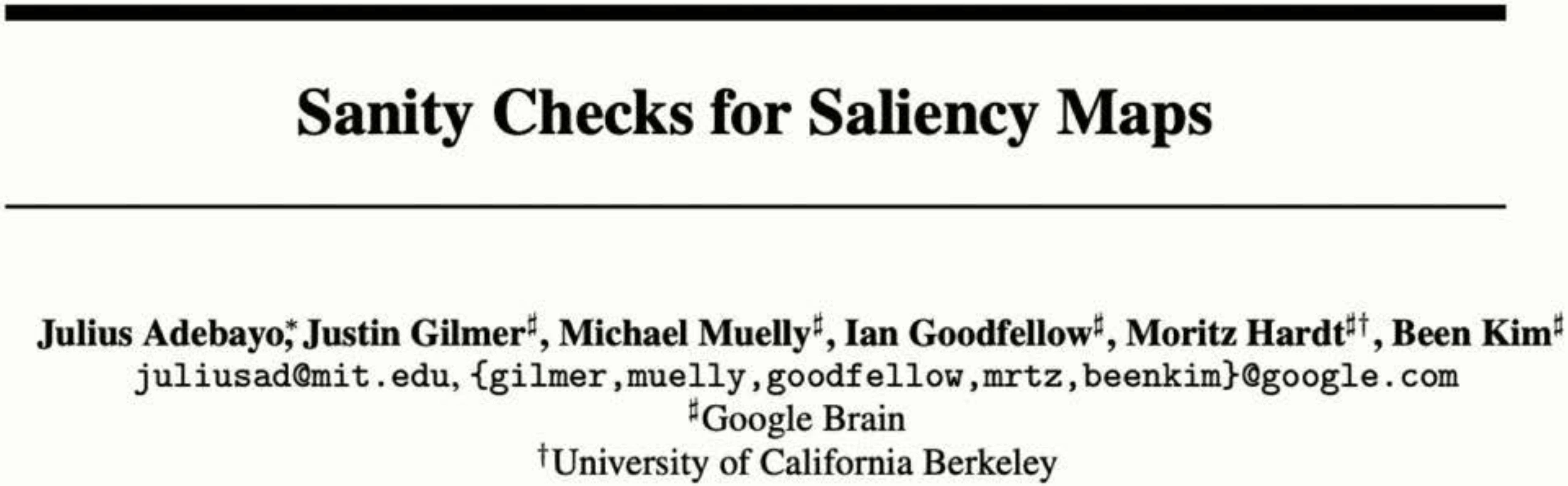


Sanity Checks for Saliency Maps

Julius Adebayo[*], **Justin Gilmer**[♯], **Michael Muelly**[♯], **Ian Goodfellow**[♯], **Moritz Hardt**[♯†], **Been Kim**[♯]
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
[♯]Google Brain
[†]University of California Berkeley

Sanity check for saliency maps, Adebayo et al.

Evaluations conducted by other papers



Sanity Checks for Saliency Maps

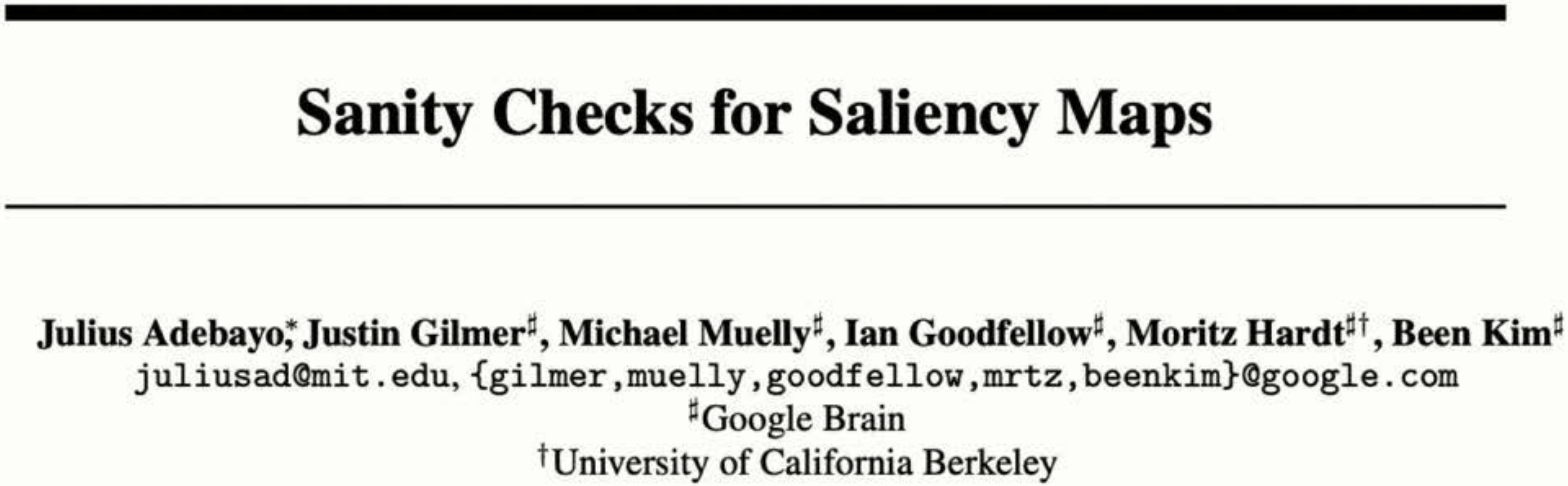
Julius Adebayo[*], Justin Gilmer[♯], Michael Muelly[♯], Ian Goodfellow[♯], Moritz Hardt[♯†], Been Kim[♯]
`juliusad@mit.edu`, `{gilmer,muelly,goodfellow,mrtz,beenkim}@google.com`
[♯]Google Brain
[†]University of California Berkeley

Parameter randomization test

Data randomization test

Sanity check for saliency maps, Adebayo et al.

Evaluations conducted by other papers



Sanity Checks for Saliency Maps

Julius Adebayo[*], **Justin Gilmer**[♯], **Michael Muelly**[♯], **Ian Goodfellow**[♯], **Moritz Hardt**[♯†], **Been Kim**[♯]
`juliusad@mit.edu`, `{gilmer,muelly,goodfellow,mrtz,beenkim}@google.com`
[♯]Google Brain
[†]University of California Berkeley

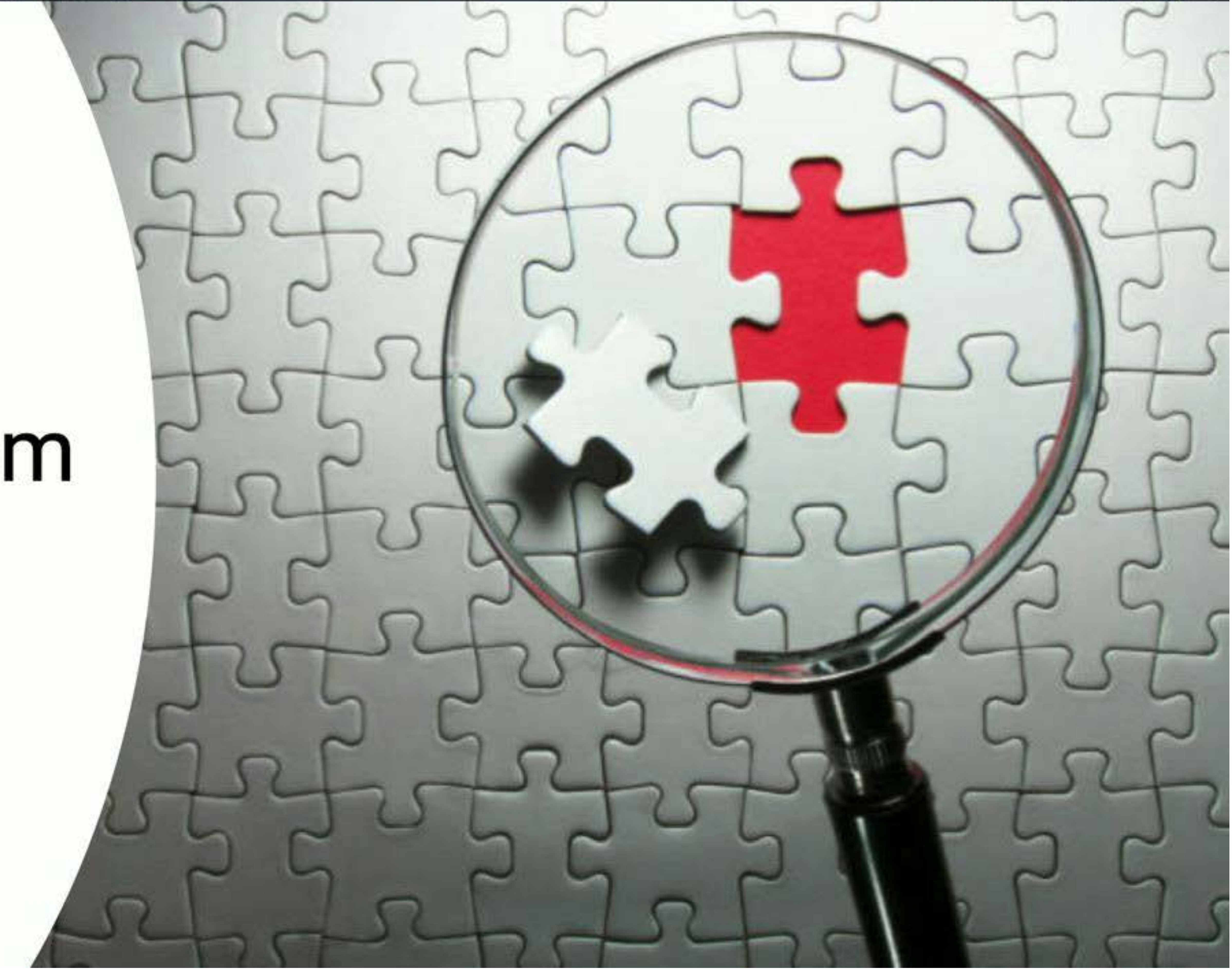
Parameter randomization test

Data randomization test

Only Grad-CAM and Backprop satisfied all the sanity checks

Sanity check for saliency maps, Adebayo et al.

Insights from Grad-CAM



Visualizing Image Captioning models

A group of people flying kites on a beach

Visualizing Image Captioning models

A group of people flying kites on a beach

Visualizing Image Captioning models

A group of people flying kites on a beach

A man is sitting at a table with a pizza

PowerPoint File Edit View Insert Format Arrange Tools Slide Show Window Help 100% Thu 1:25 PM

Visualizing Image Captioning models

A group of people flying kites on a beach

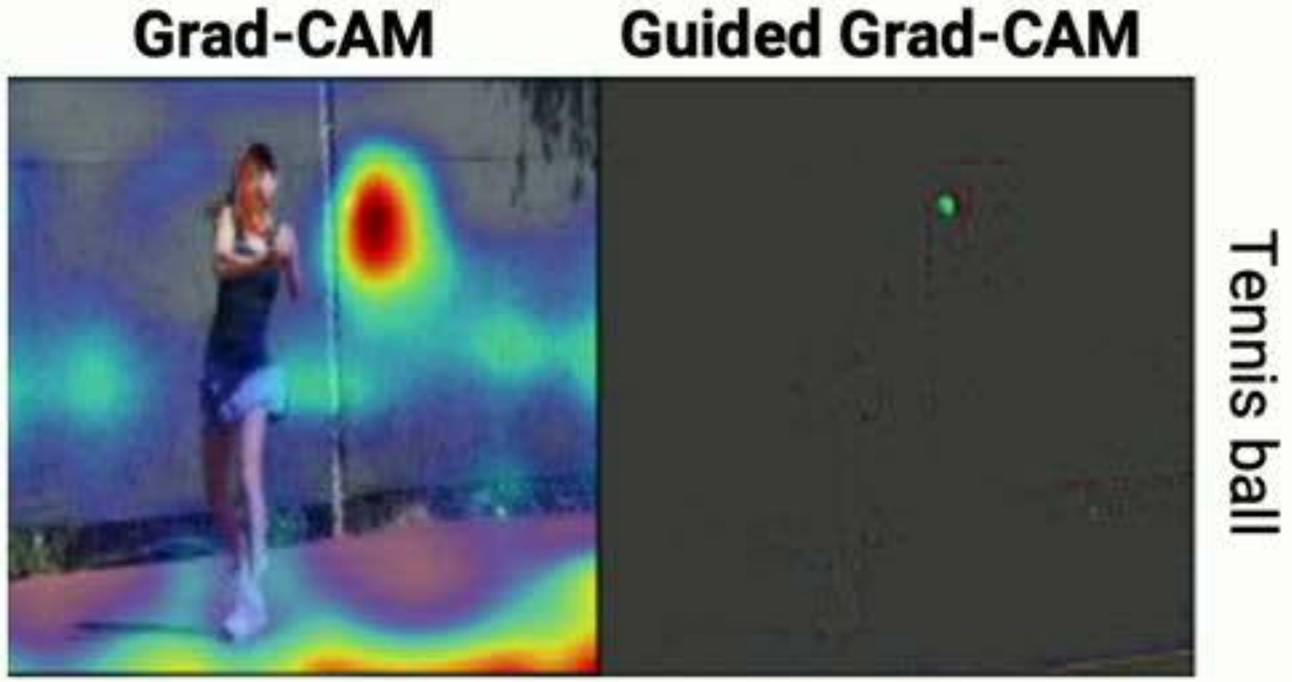
A man is sitting at a table with a pizza

Visualizing Visual Question Answering models

What is the person hitting?

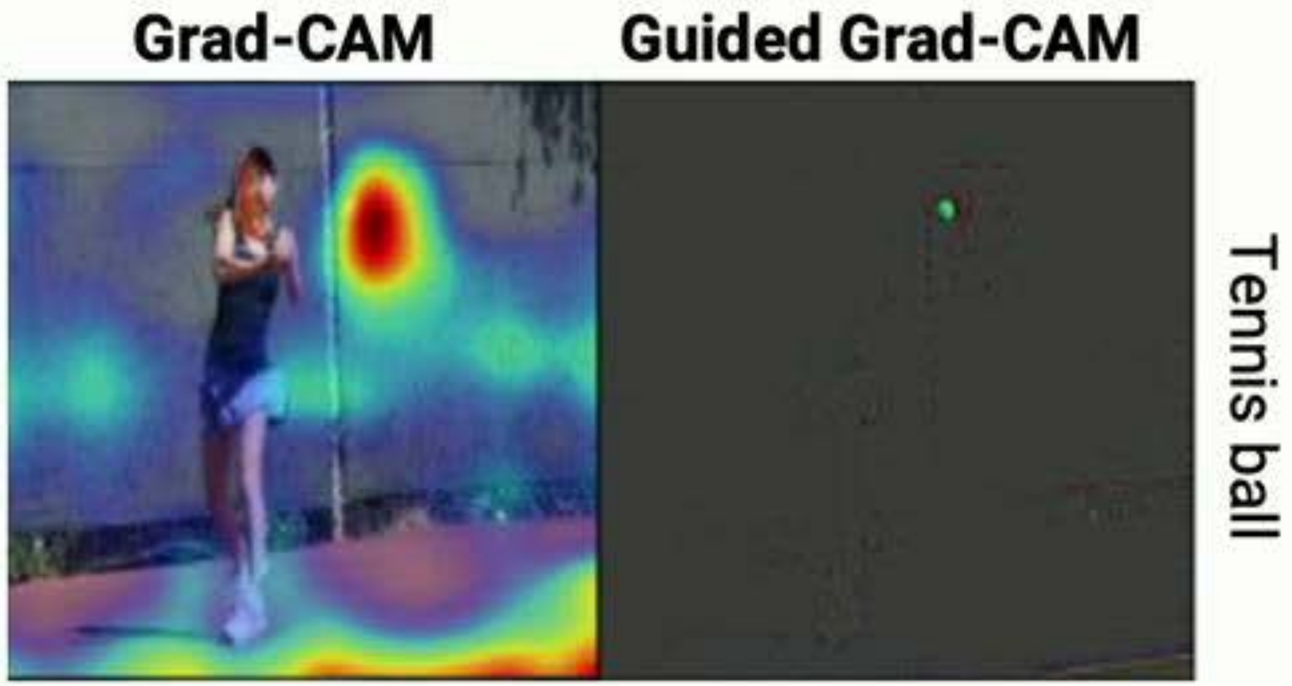
Visualizing Visual Question Answering models

What is the person hitting?



Visualizing Visual Question Answering models

What is the person hitting?



Even simple non-attention-based CNN+LSTM models attend to appropriate regions

Analyzing Failure modes

Predicted: *Car mirror*

Analyzing Failure modes

Predicted: *Car mirror*

Ground-truth: *Volcano*

Analyzing Failure modes

Predicted: *Car mirror*

Ground-truth: *Volcano*

Analyzing Failure modes

Predicted: *Car mirror*

Ground-truth: *Volcano*

Predicted: *Vine snake*

Analyzing Failure modes

Predicted: *Car mirror*

Ground-truth: *Volcano*

Predicted: *Vine snake*

Analyzing Failure modes

Predicted: *Car mirror*

Ground-truth: *Volcano*

Predicted: *Vine snake*

Ground-truth: *coil*

Analyzing Failure modes

Predicted: *Car mirror*

Ground-truth: *Volcano*

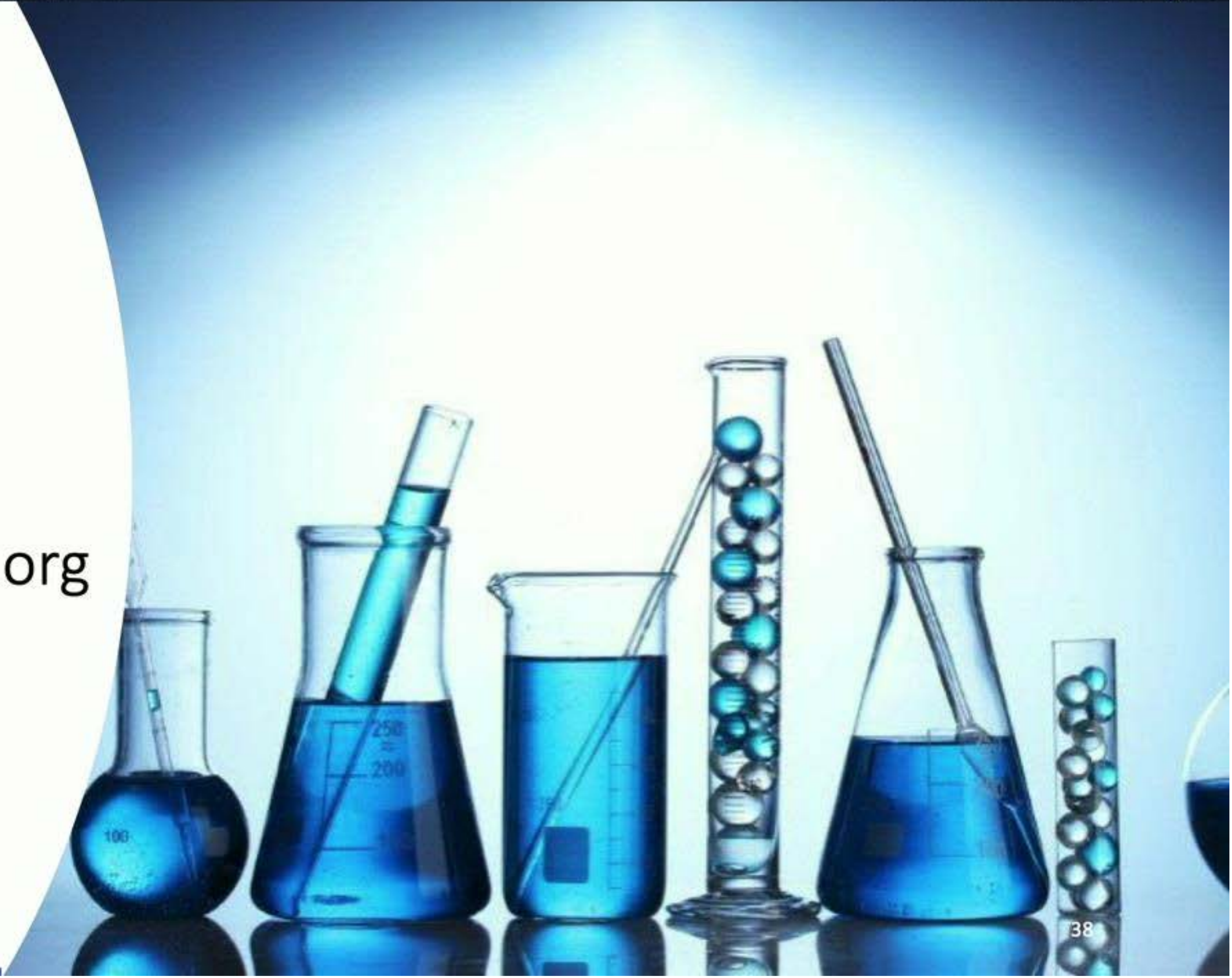
Predicted: *Vine snake*

Ground-truth: *coil*

Even unreasonable predictions sometimes have reasonable explanations

Demo

gradcam.cloudcv.org



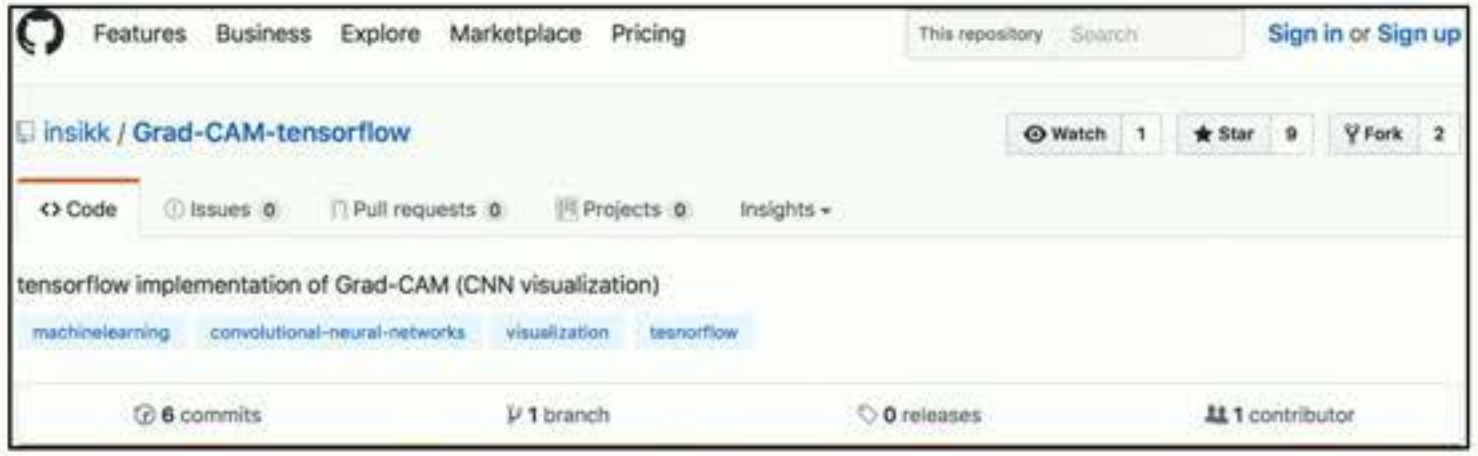
Grad-CAM re-implementations

jacobgil / keras-grad-cam

An implementation of Grad-CAM with keras

keras grad-cam visualization deep-learning

10 commits · 1 branch · 0 releases · 2 contributors · MIT

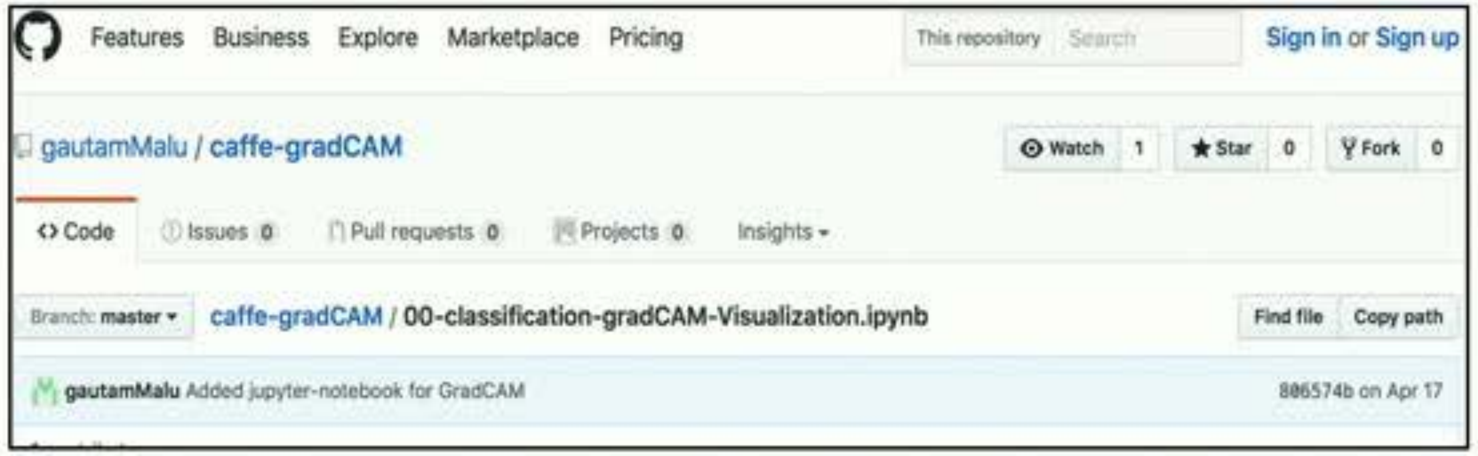


insikk / Grad-CAM-tensorflow

tensorflow implementation of Grad-CAM (CNN visualization)

machinelearning convolutional-neural-networks visualization tesnorflow

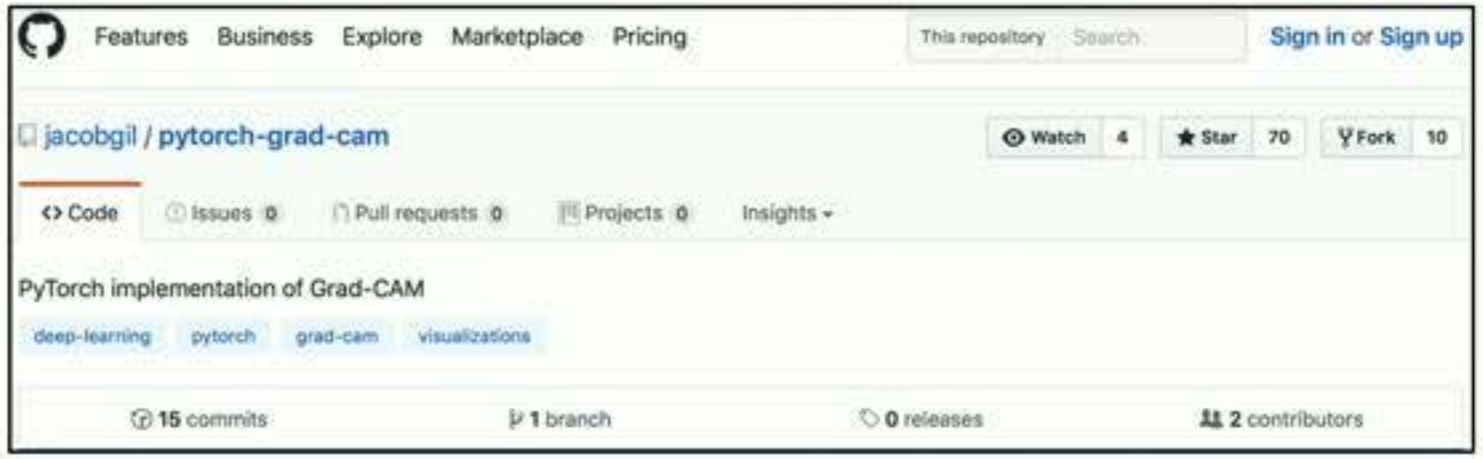
6 commits · 1 branch · 0 releases · 1 contributor



gautamMalu / caffe-gradCAM

caffe-gradCAM / 00-classification-gradCAM-Visualization.ipynb

gautamMalu Added jupyter-notebook for GradCAM · 886574b on Apr 17



jacobgil / pytorch-grad-cam

PyTorch implementation of Grad-CAM

deep-learning pytorch grad-cam visualizations

15 commits · 1 branch · 0 releases · 2 contributors

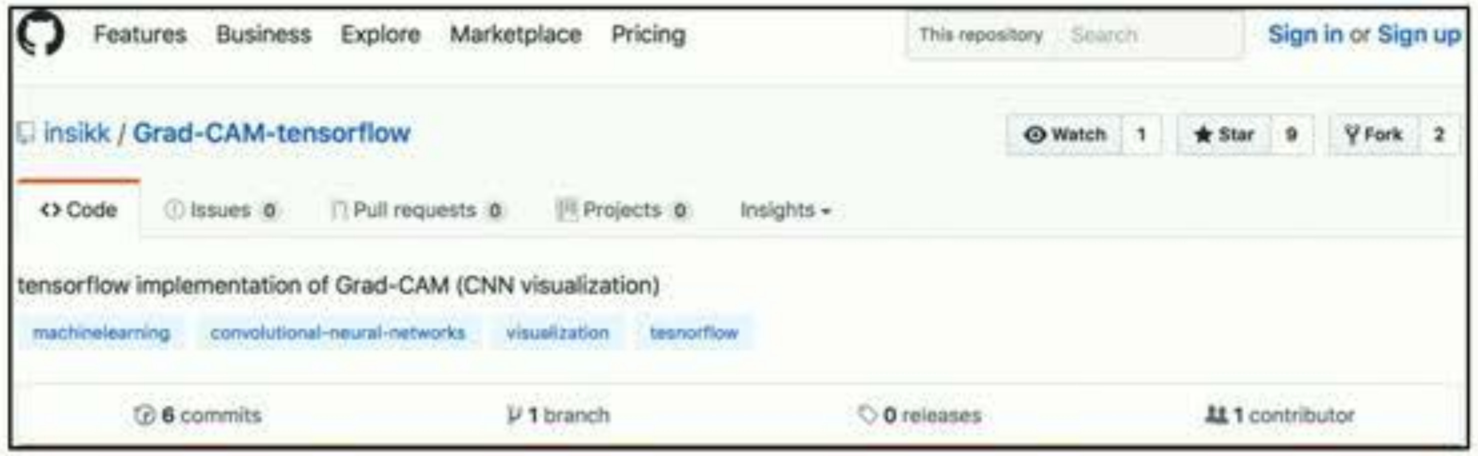
Grad-CAM re-implementations

jacobgil / keras-grad-cam

An implementation of Grad-CAM with keras

keras grad-cam visualization deep-learning

10 commits · 1 branch · 0 releases · 2 contributors · MIT

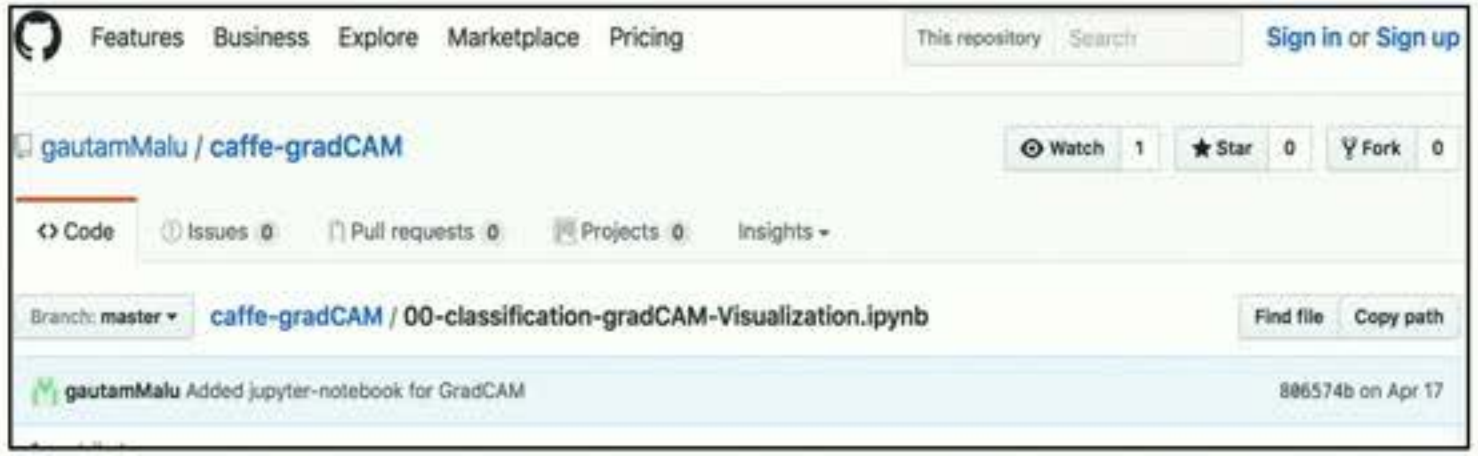


insikk / Grad-CAM-tensorflow

tensorflow implementation of Grad-CAM (CNN visualization)

machinelearning convolutional-neural-networks visualization tesnorflow

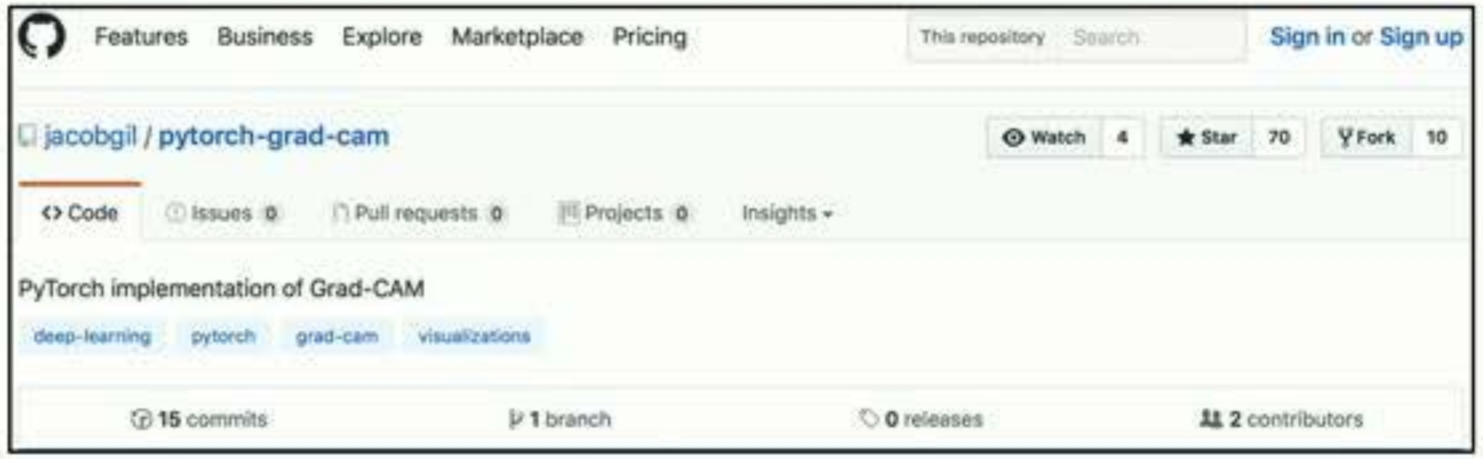
6 commits · 1 branch · 0 releases · 1 contributor



gautamMalu / caffe-gradCAM

caffe-gradCAM / 00-classification-gradCAM-Visualization.ipynb

gautamMalu Added jupyter-notebook for GradCAM 886574b on Apr 17



jacobgil / pytorch-grad-cam

PyTorch implementation of Grad-CAM

deep-learning pytorch grad-cam visualizations

15 commits · 1 branch · 0 releases · 2 contributors

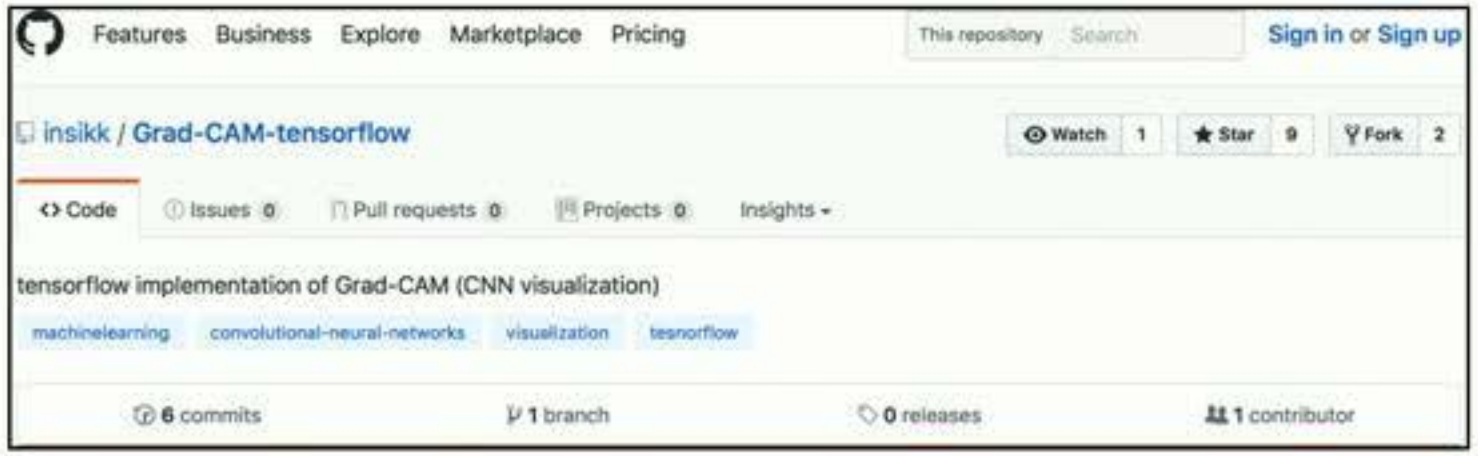
Captum

Model Interpretability for PyTorch

Grad-CAM re-implementations

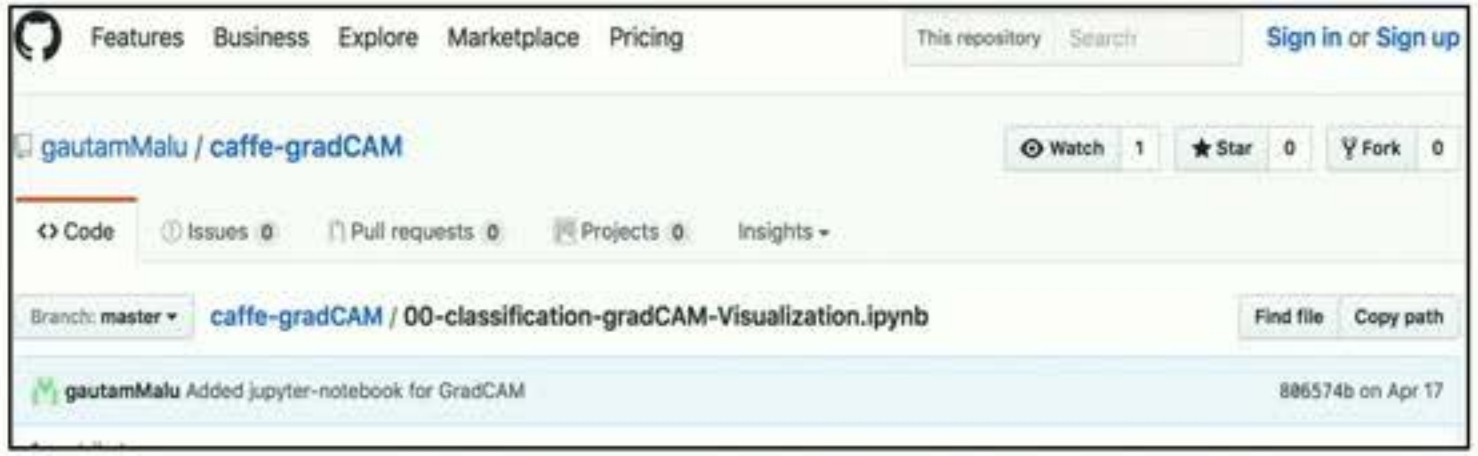
jacobgil / keras-grad-cam

An implementation of Grad-CAM with keras



insikk / Grad-CAM-tensorflow

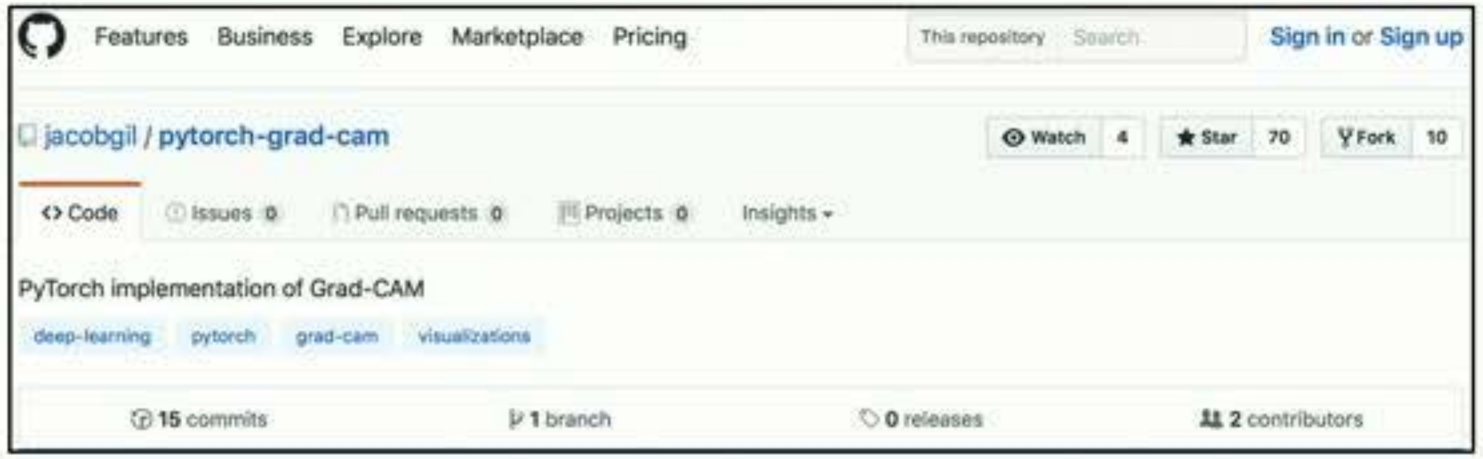
tensorflow implementation of Grad-CAM (CNN visualization)



gautamMalu / caffe-gradCAM

caffe-gradCAM / 00-classification-gradCAM-Visualization.ipynb

gautamMalu Added jupyter-notebook for GradCAM 886574b on Apr 17

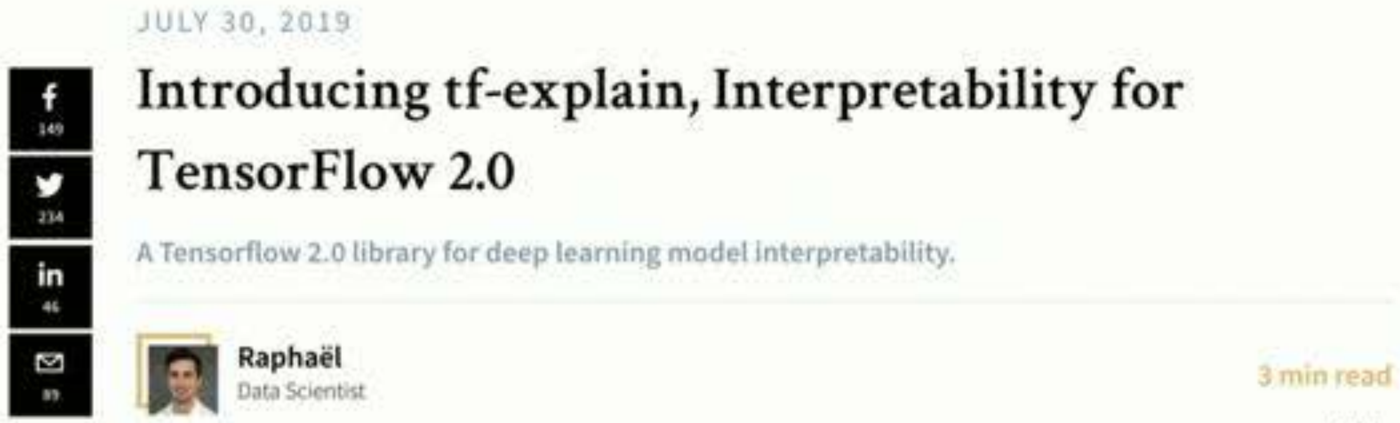


jacobgil / pytorch-grad-cam

PyTorch implementation of Grad-CAM

Captum

Model Interpretability for PyTorch



JULY 30, 2019

Introducing tf-explain, Interpretability for TensorFlow 2.0

A Tensorflow 2.0 library for deep learning model interpretability.

Raphaël
Data Scientist

What have others used Grad-CAM for?

- 1600+ citations

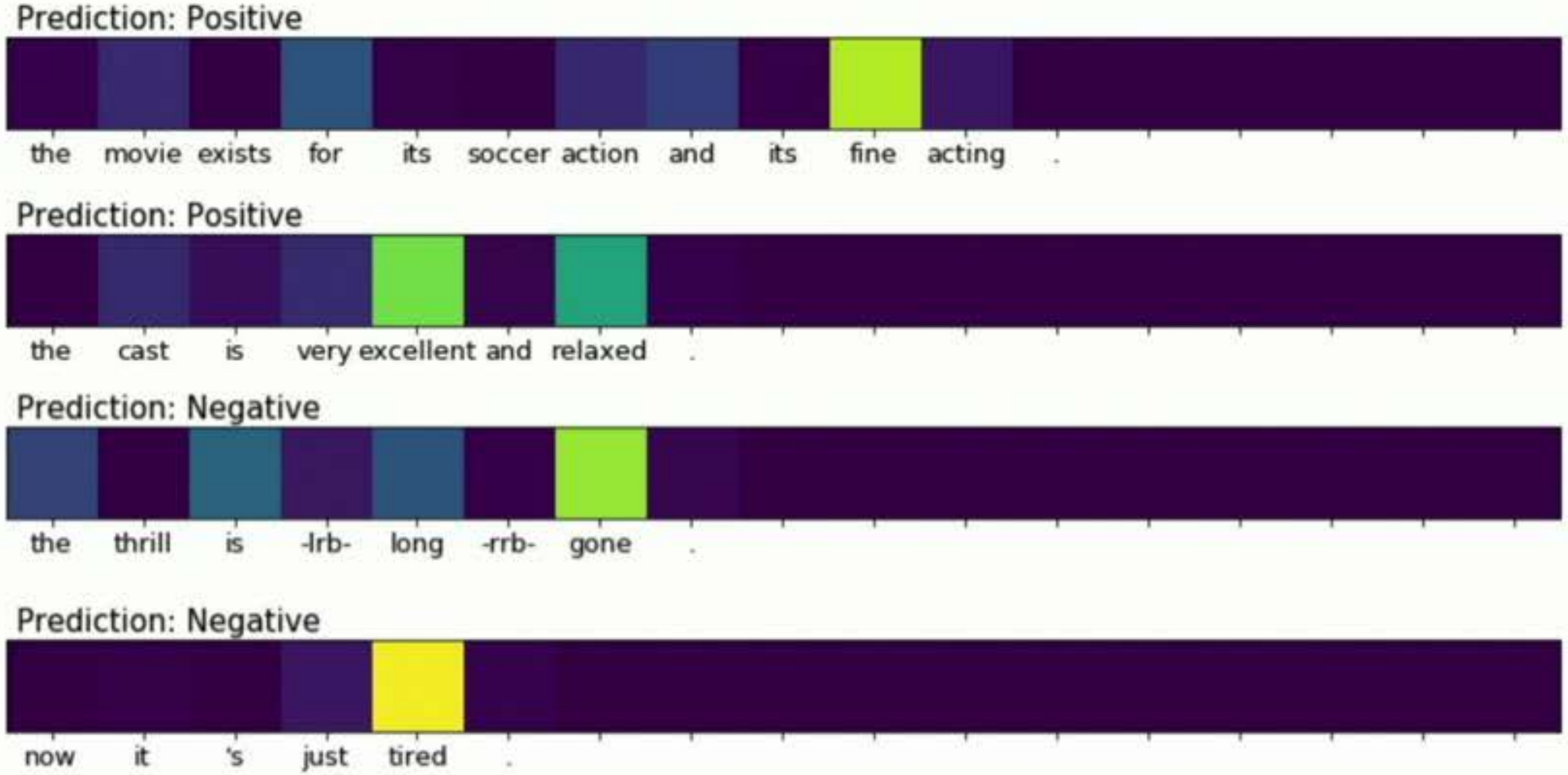
What have others used Grad-CAM for?

Grad-CAM for Videos

Uncovering [something]

What have others used Grad-CAM for?

Grad-CAM for Text

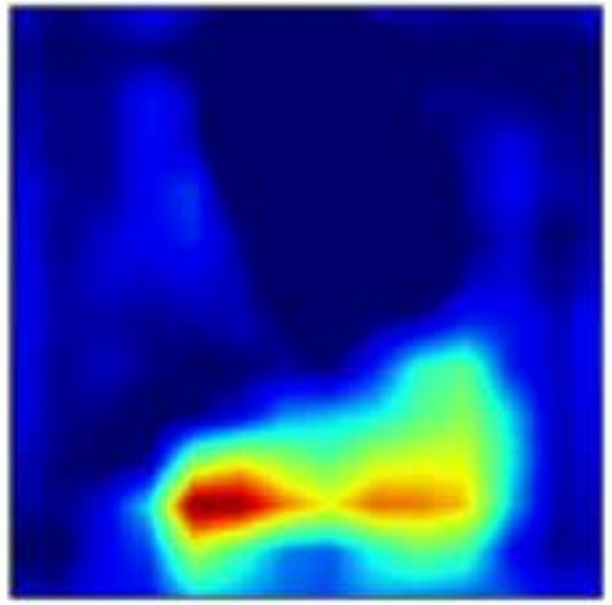


Industry Impact

facebook

Microsoft

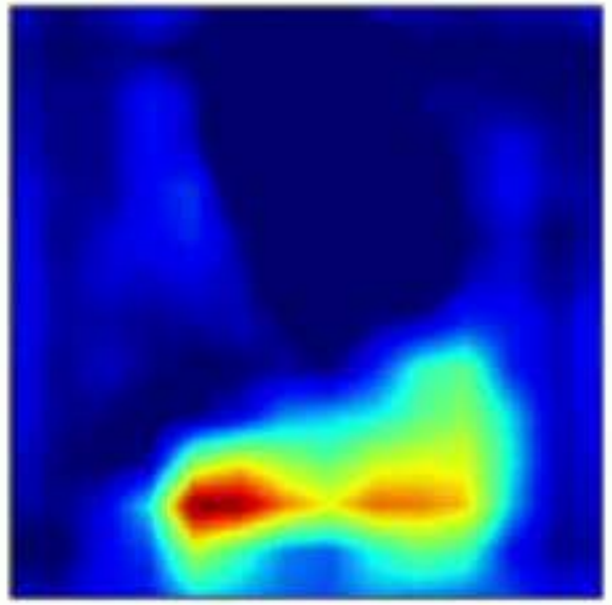
Grad-CAM Limitations



Grad-CAM for 'cat'

Guided-Grad-CAM for 'cat'

Grad-CAM Limitations



Grad-CAM for ‘cat’

Guided-Grad-CAM for ‘cat’

Unclear what concept is represented by the pixels below

Summary

Explain

Explain decisions from deep networks through Grad-CAM (ICCV’17, IJCV’19)

- Introduced a generic technique to interpret decisions from any CNN-based deep network
- Interesting findings with Grad-CAM
- Impact at various places
- Evaluation

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What future directions excite me?

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What future directions excite me?

Here is a riddle

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims “I can't operate on this boy, he's my son!”

How could this be?

Boston University study

Here is a riddle

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims “I can't operate on this boy, he's my son!”

How could this be?

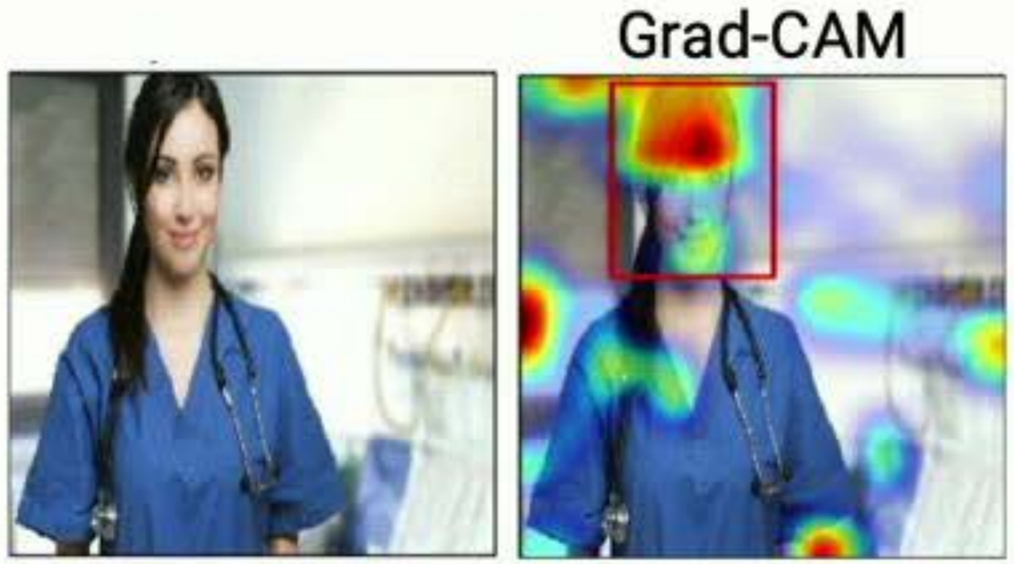
Biases in real world datasets

"Female Doctor"

"Doctor"

AI Models capture the same bias

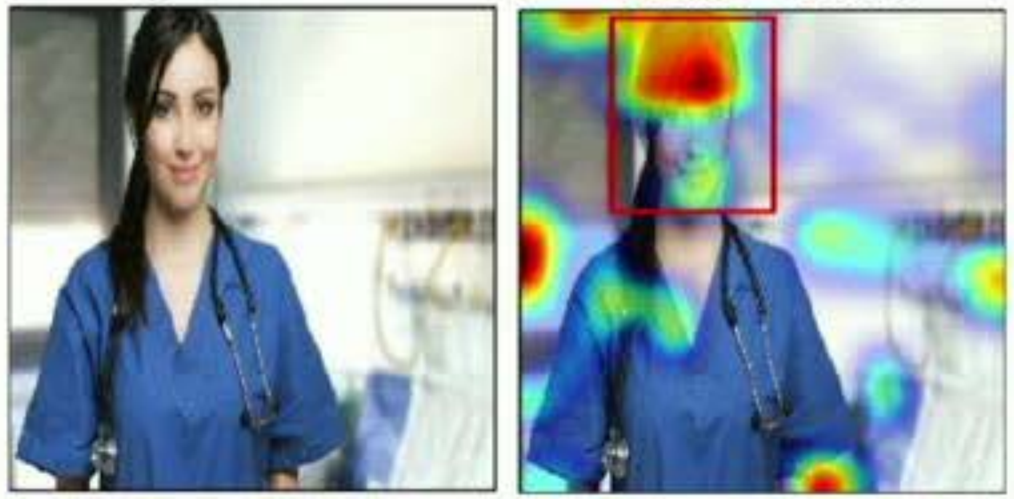
AI Models capture the same bias



Ground-Truth:
Nurse

Predicted:
Nurse

AI Models capture the same bias


Grad-CAM

Ground-Truth: Nurse

Predicted: **Nurse**


Grad-CAM

Ground-Truth: Doctor

Predicted: **Nurse**

AI Models capture the same bias

Ground-Truth: Nurse

Predicted: **Nurse**

Ground-Truth: Doctor

Predicted: **Nurse**

Models learns a gender stereotype

Fixing bias by balancing training set fixes model

- Balancing training data
  - Doctors and Nurses:
    - 50% male 50% female

Ground-Truth: Nurse
Predicted: **Nurse**

Ground-Truth: Doctor
Predicted: **Doctor**

Fixing bias by balancing training set fixes model

- **Balancing training data**
  - Doctors and Nurses:
    - 50% male 50% female

Ground-Truth: Nurse

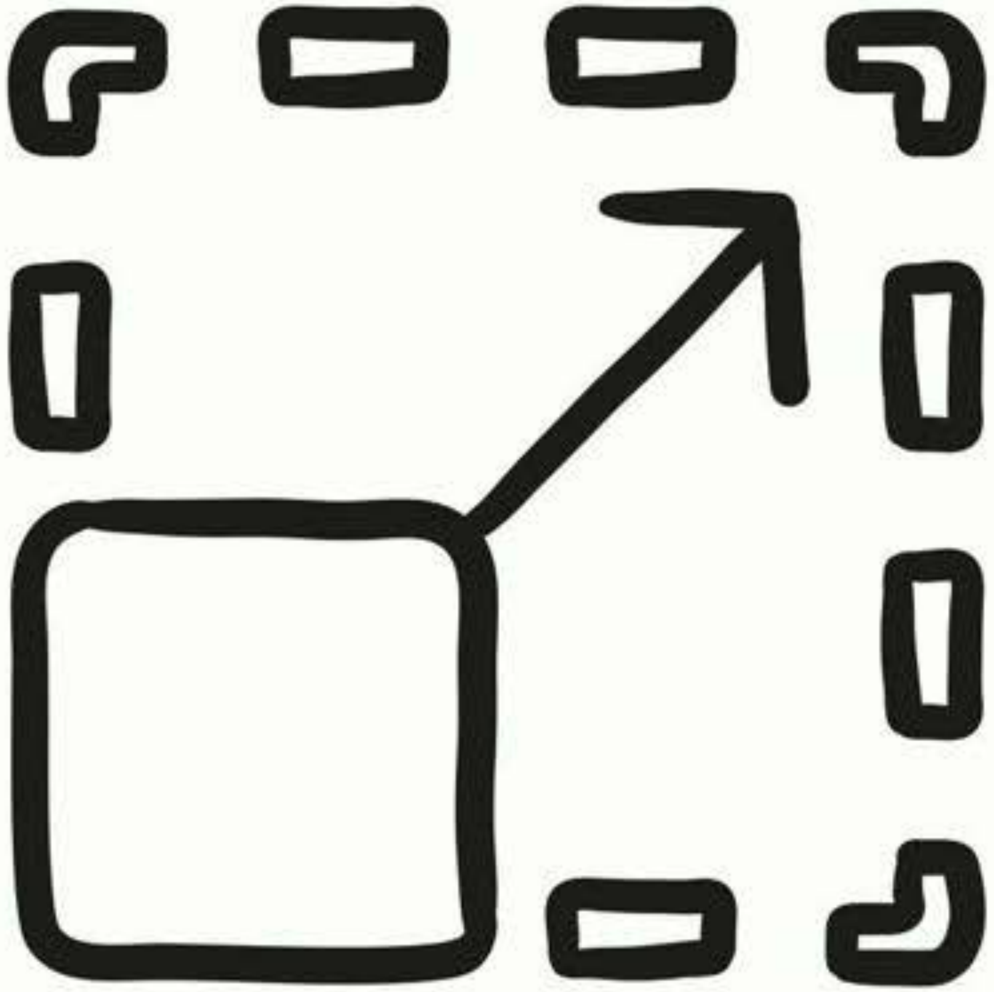
Predicted: **Nurse**

Ground-Truth: Doctor

Predicted: **Doctor**

Model not only makes correct predictions but also looks at appropriate regions

Is balancing datasets always scalable?



Biases in Vision and Language models

Giraffe standing next to a tree

Biases in Vision and Language models

Giraffe standing next to a tree

COCO images

Biases in Vision and Language models

What color are the bananas?

Yellow

Biases in Vision and Language models

What color are the bananas?

Yellow

COCO training dataset images

Biases in Vision and Language models

What color are the bananas?
Yellow

COCO training dataset images

Problematic when distributions change

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, Devi Parikh, **"Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded"**, ICCV'19

How can explanations help debias AI models?

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, Devi Parikh, **"Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded"**, ICCV'19

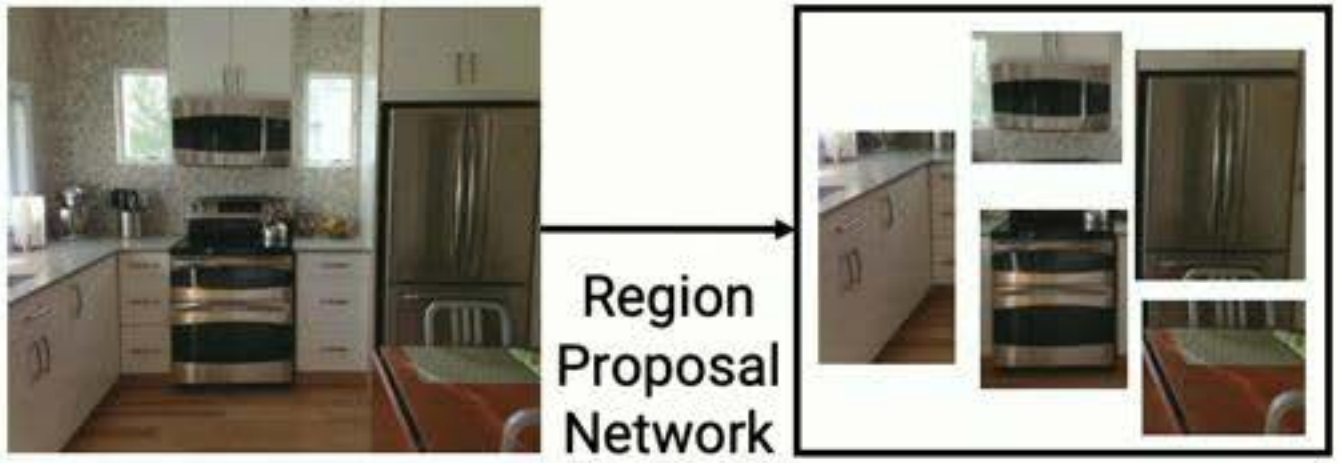
Bottom-up Top-down (UpDn) architecture

Bottom-up Top-down (UpDn) architecture

What room is this?

Bottom-up Top-down (UpDn) architecture

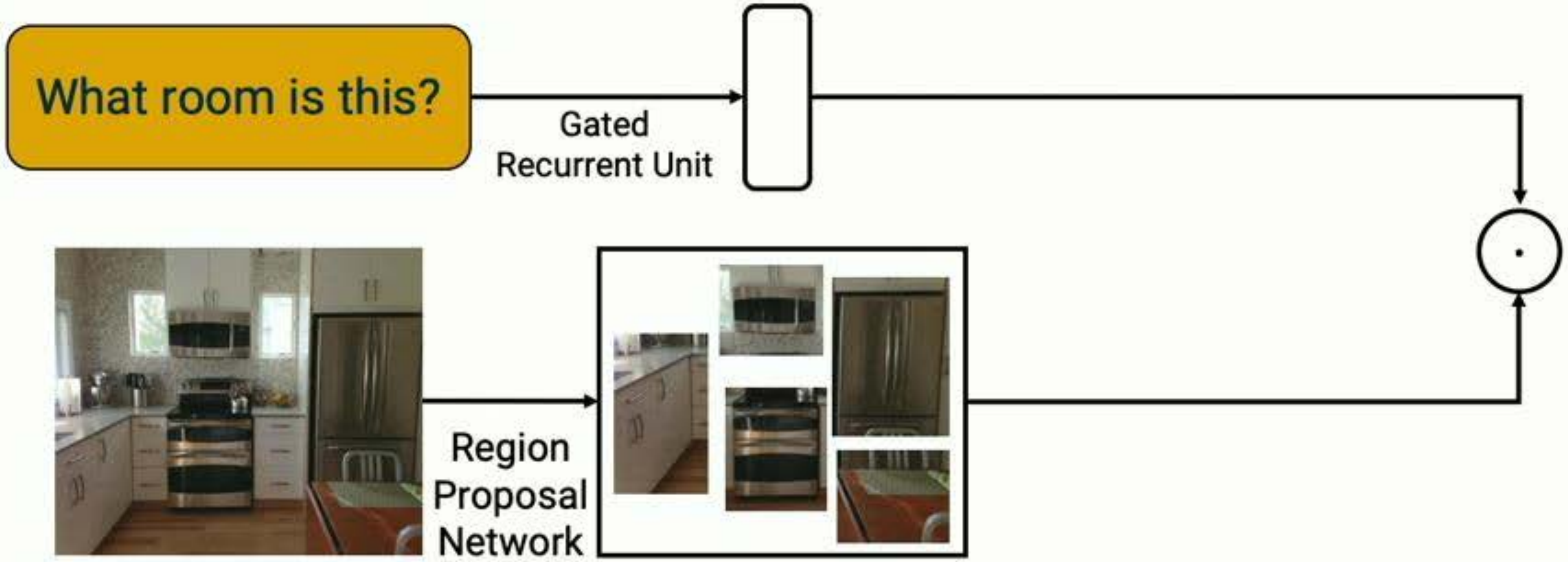
What room is this?



Region Proposal Network

PowerPoint File Edit View Insert Format Arrange Tools Slide Show Window Help 100% Thu 1:35 PM

Bottom-up Top-down (UpDn) architecture

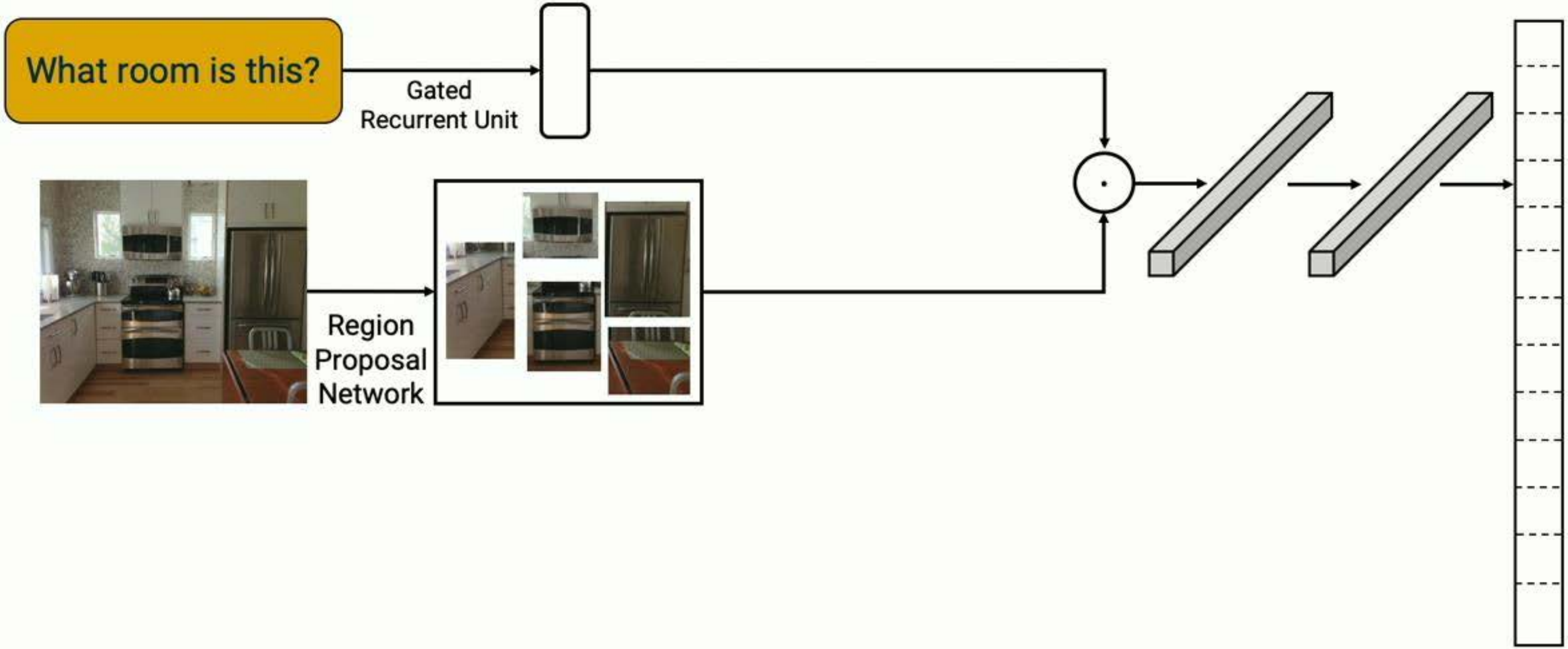


What room is this?

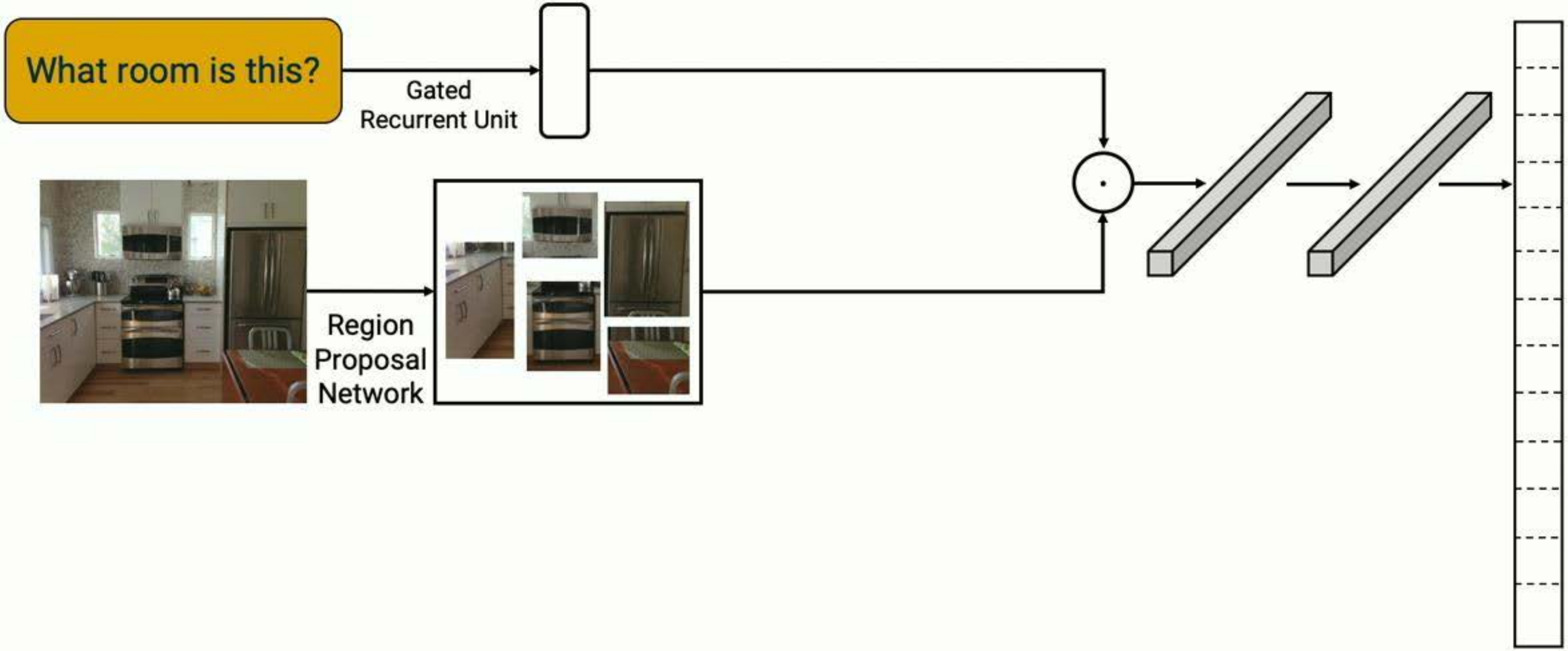
Gated Recurrent Unit

Region Proposal Network

Bottom-up Top-down (UpDn) architecture



UpDn Network Importance

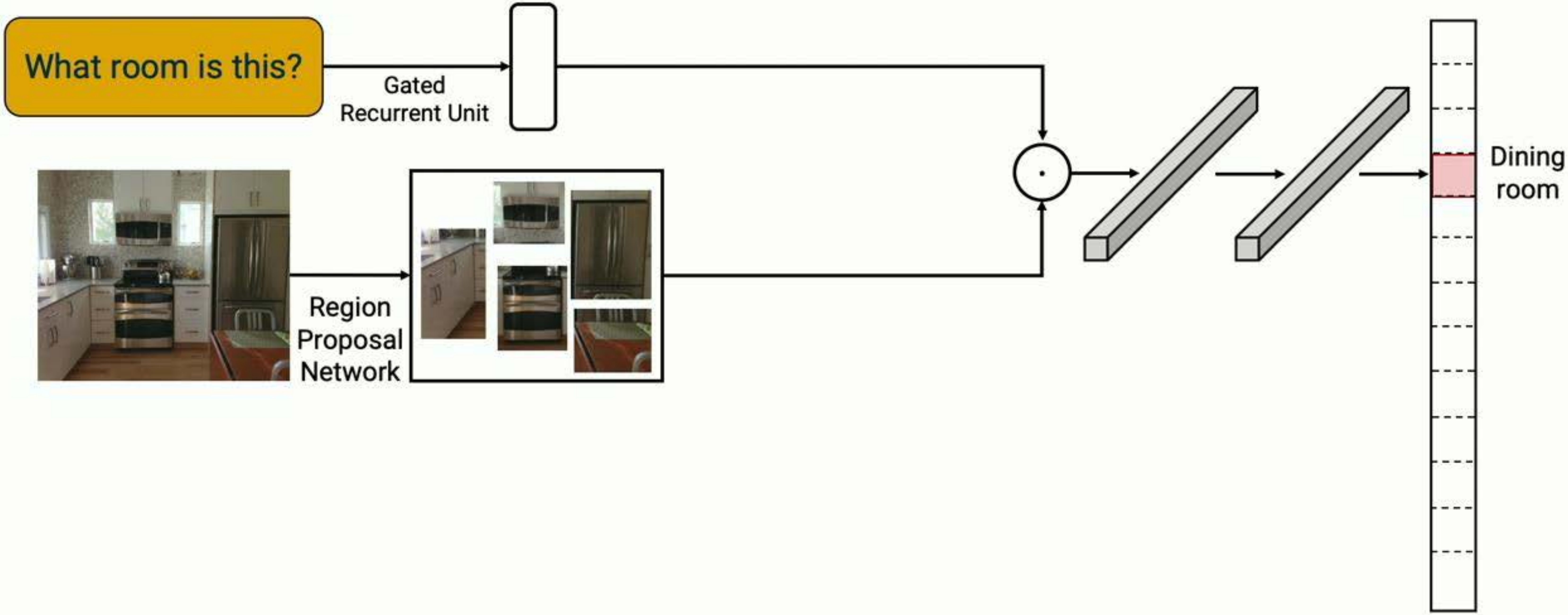


What room is this?

Gated Recurrent Unit

Region Proposal Network

UpDn Network Importance



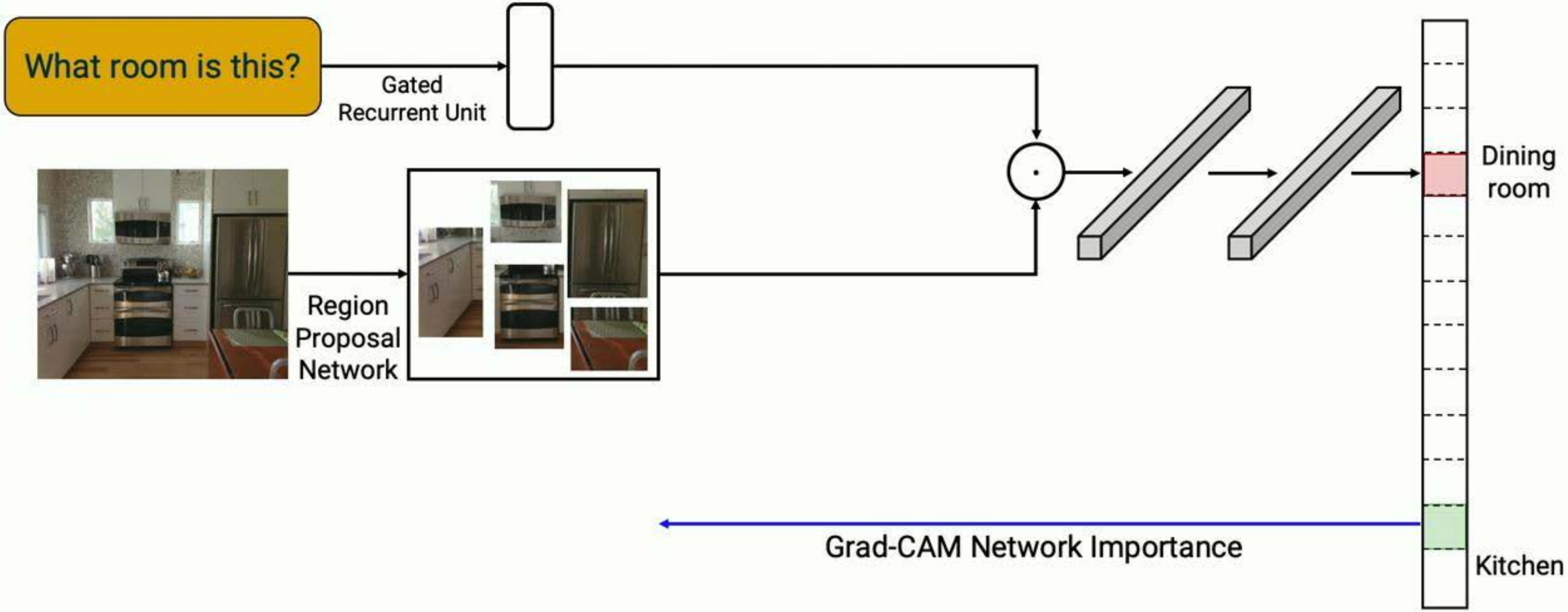
What room is this?

Gated Recurrent Unit

Region Proposal Network

Dining room

UpDn Network Importance



What room is this?

Gated Recurrent Unit

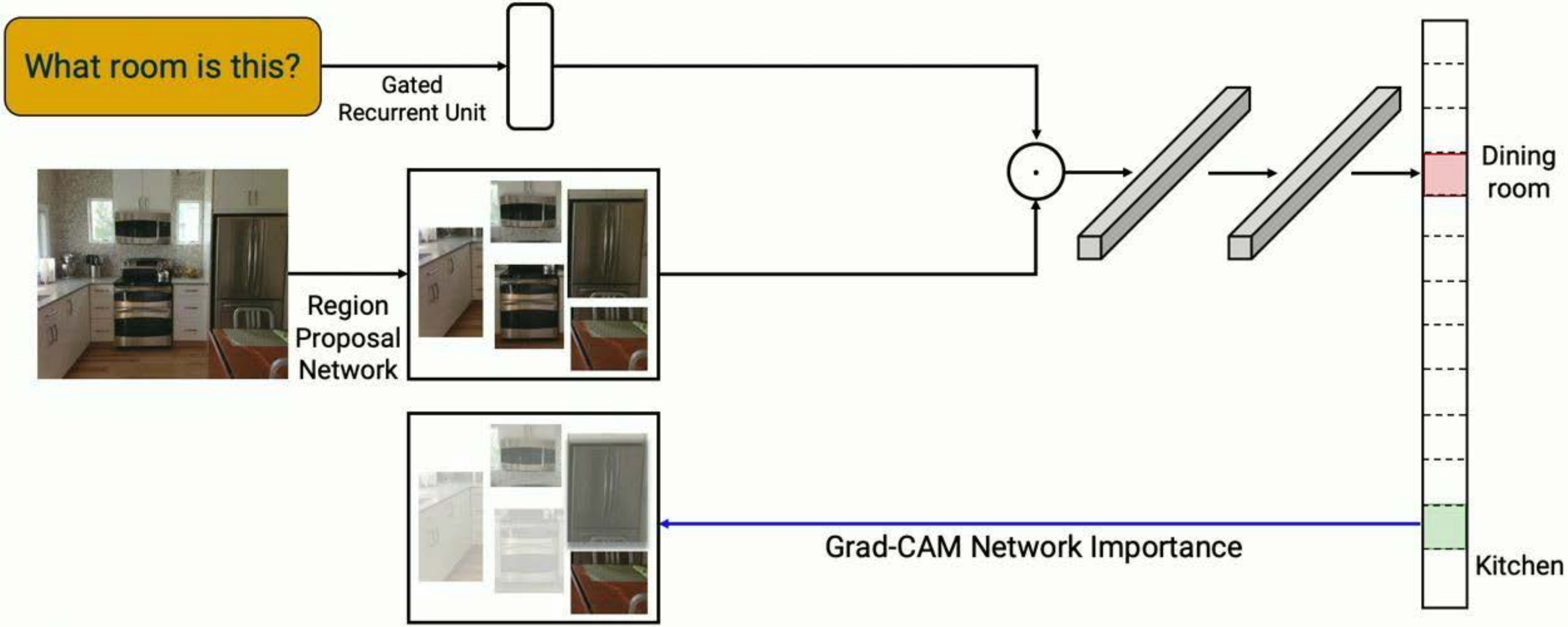
Region Proposal Network

Dining room

Kitchen

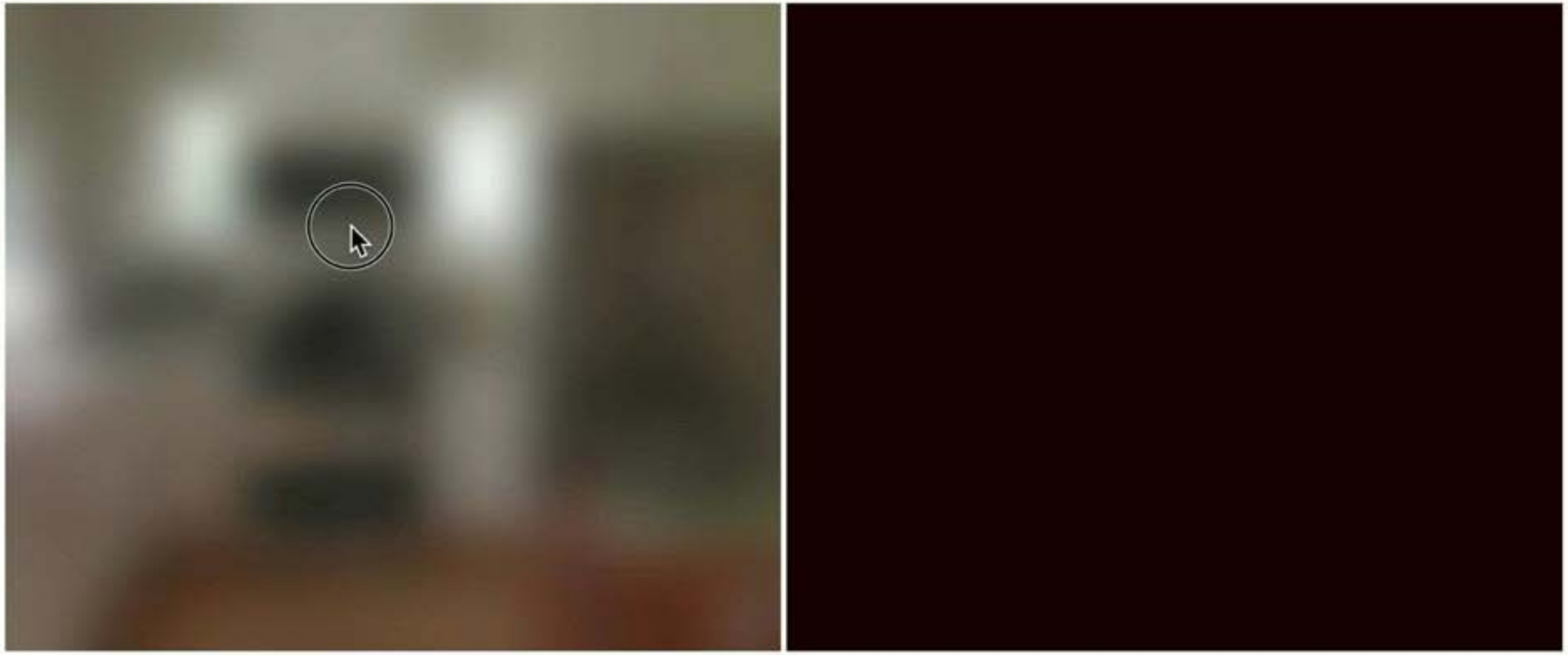
Grad-CAM Network Importance

UpDn Network Importance



Where do humans look when making decisions?

Question: What room is this?



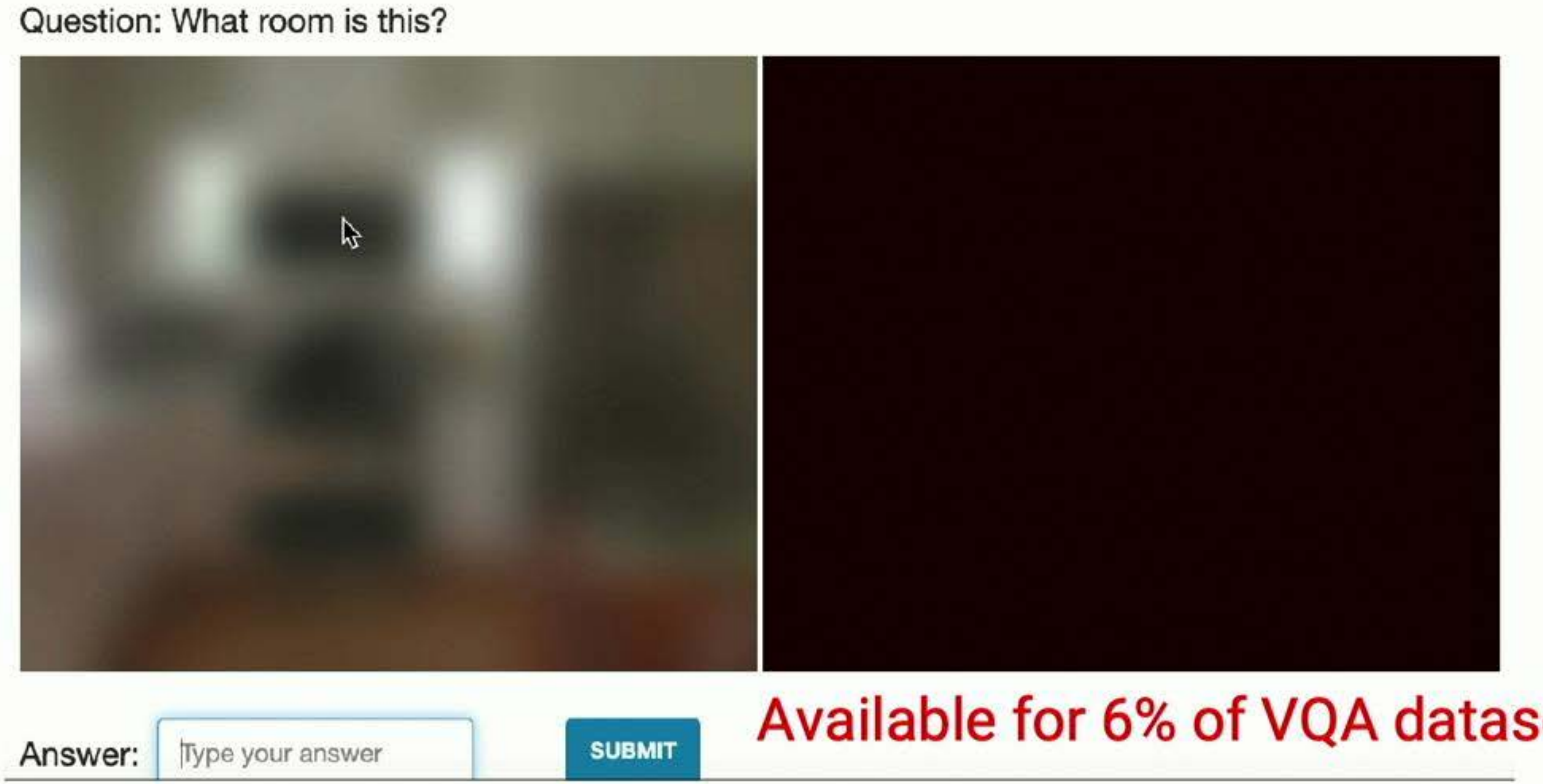
Answer: Type your answer

SUBMIT

Available for 6% of VQA dataset

Das & Agrawal et al.

Where do humans look when making decisions?



Question: What room is this?

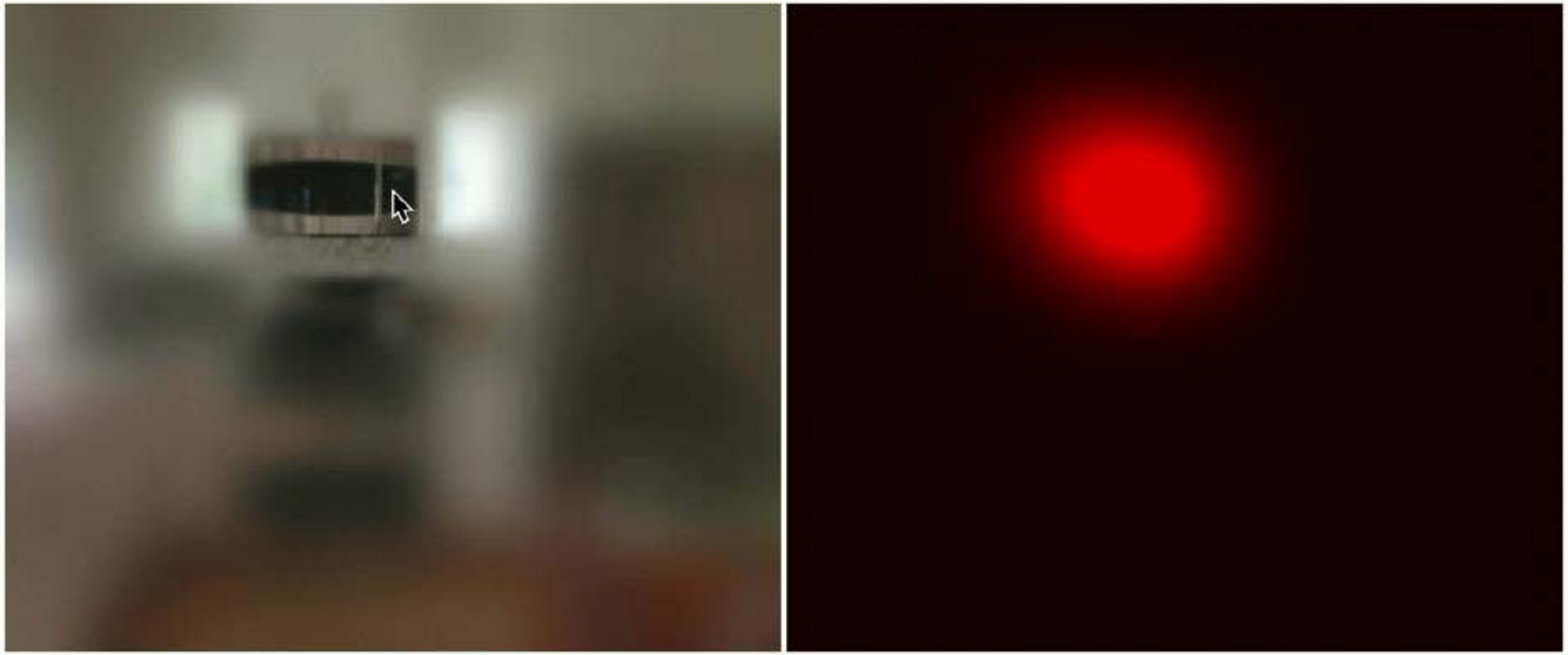
Answer: Type your answer

Available for 6% of VQA dataset

Das & Agrawal et al.

Where do humans look when making decisions?

Question: What room is this?

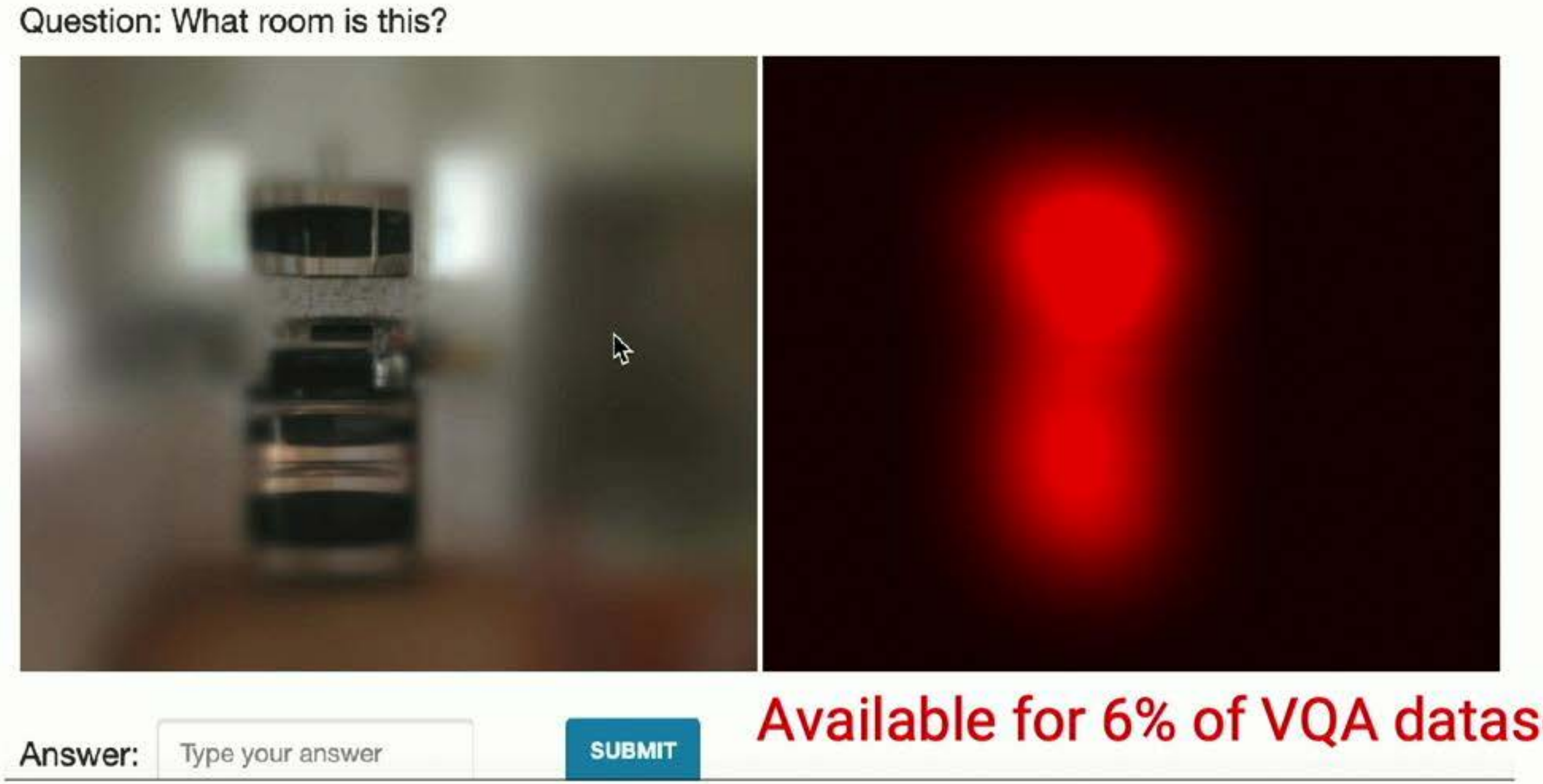


Answer: Type your answer

Available for 6% of VQA dataset

Das & Agrawal et al.

Where do humans look when making decisions?

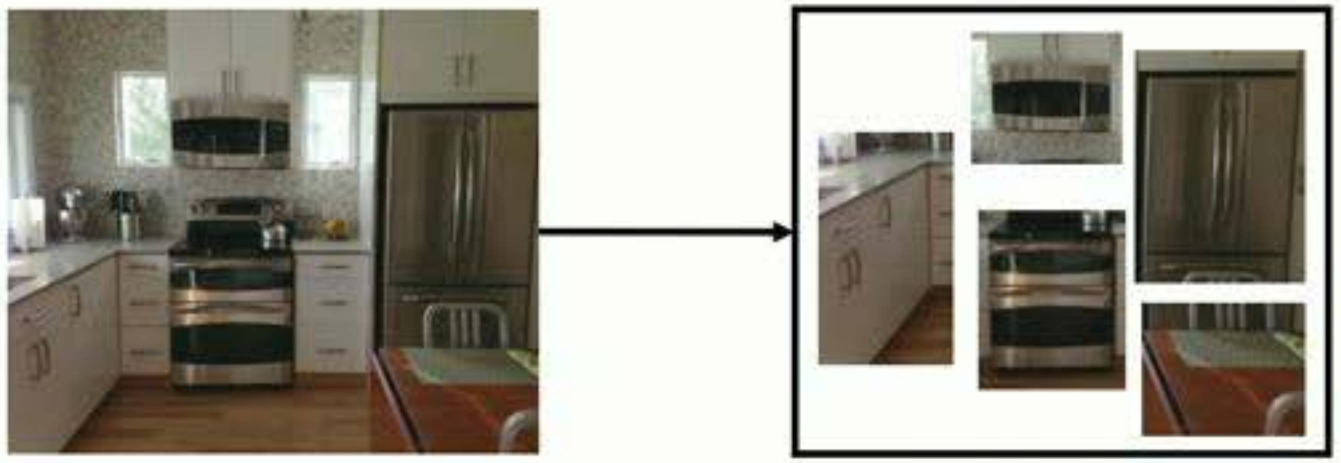


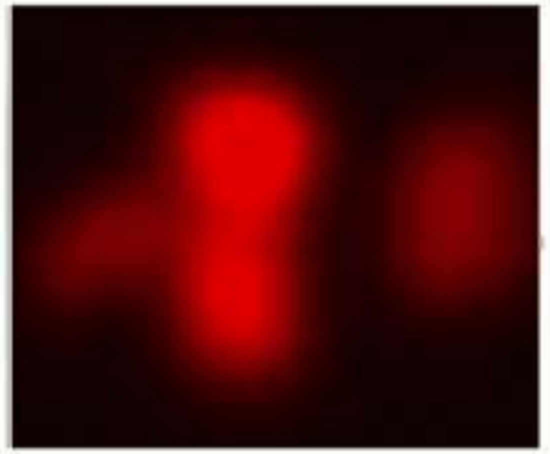
Question: What room is this?

Available for 6% of VQA dataset

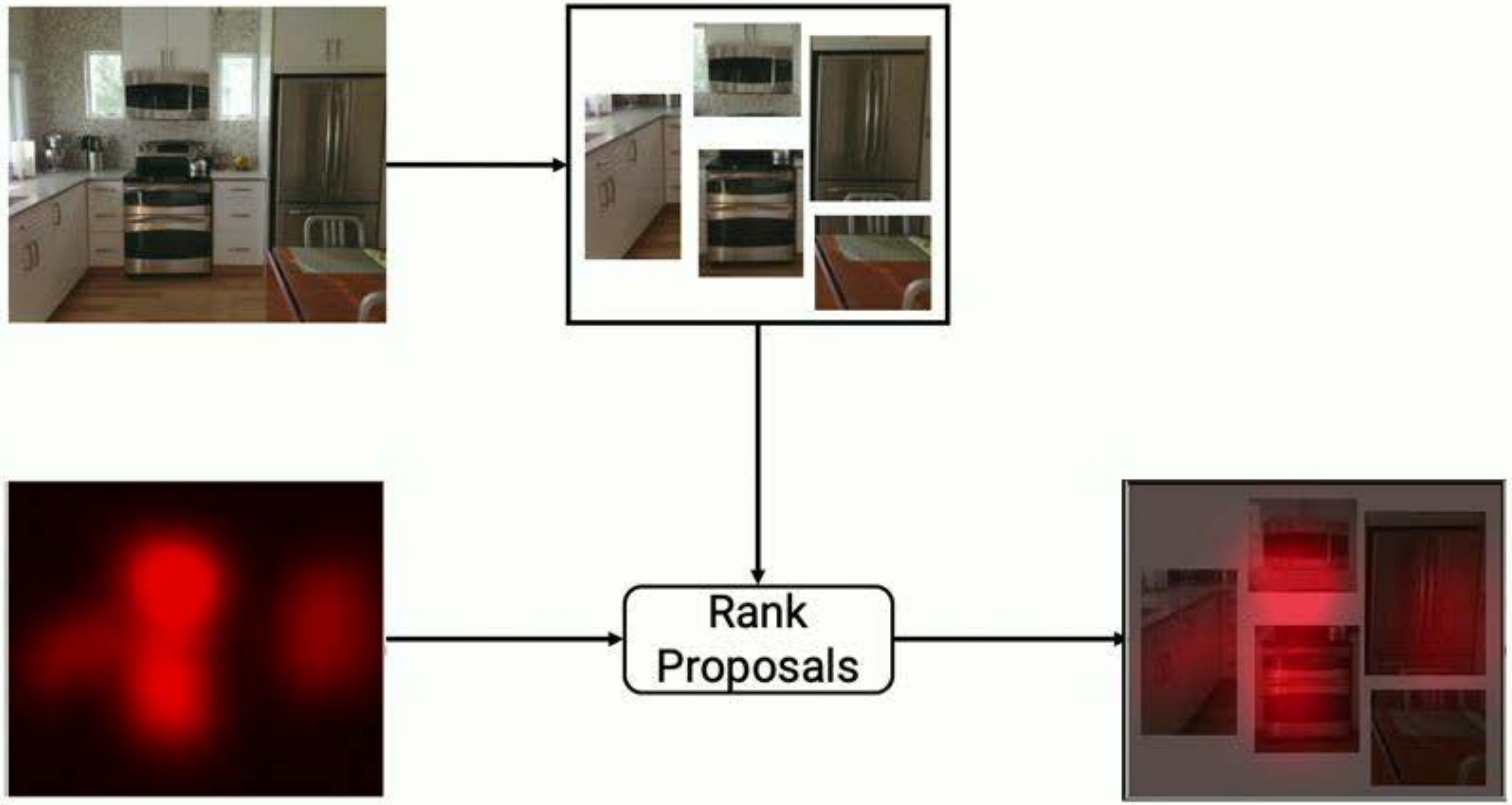
Das & Agrawal et al.

Human importance



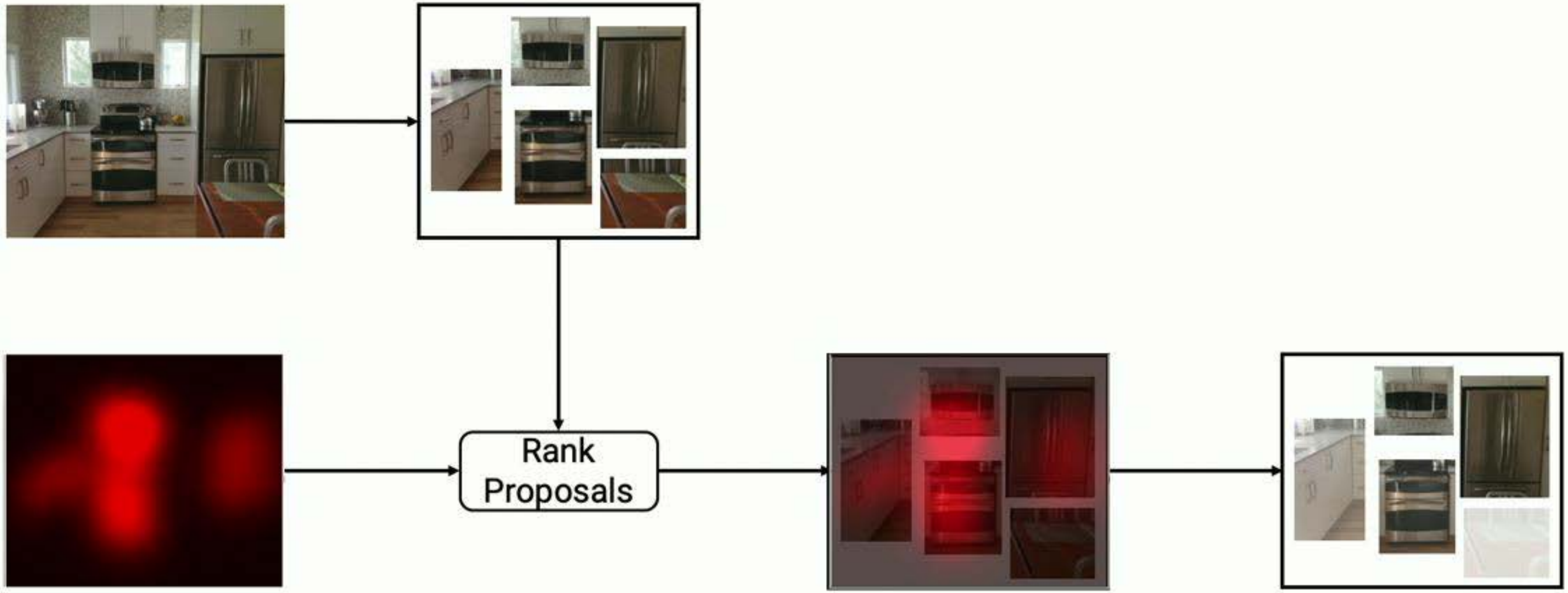


Human importance



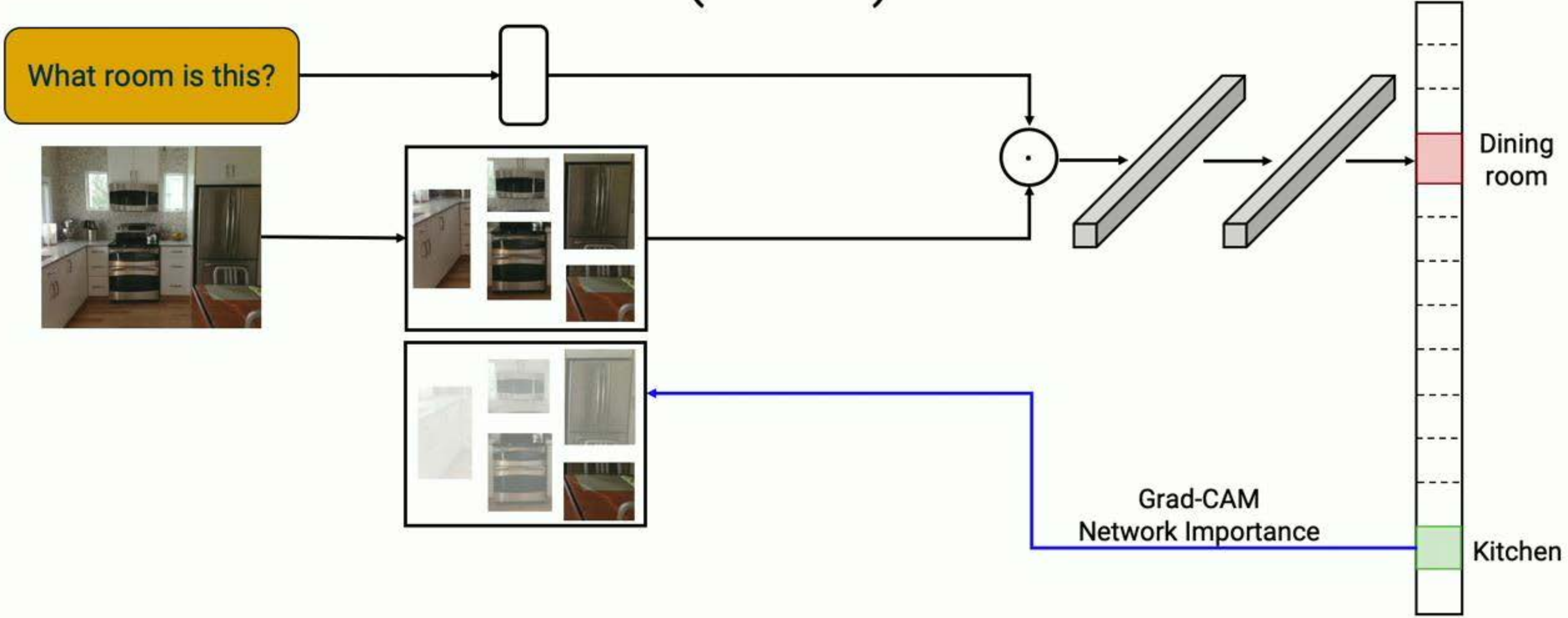
Rank Proposals

Human importance



Human Importance-aware Network Tuning (HINT)

Human Importance-aware Network Tuning (HINT)



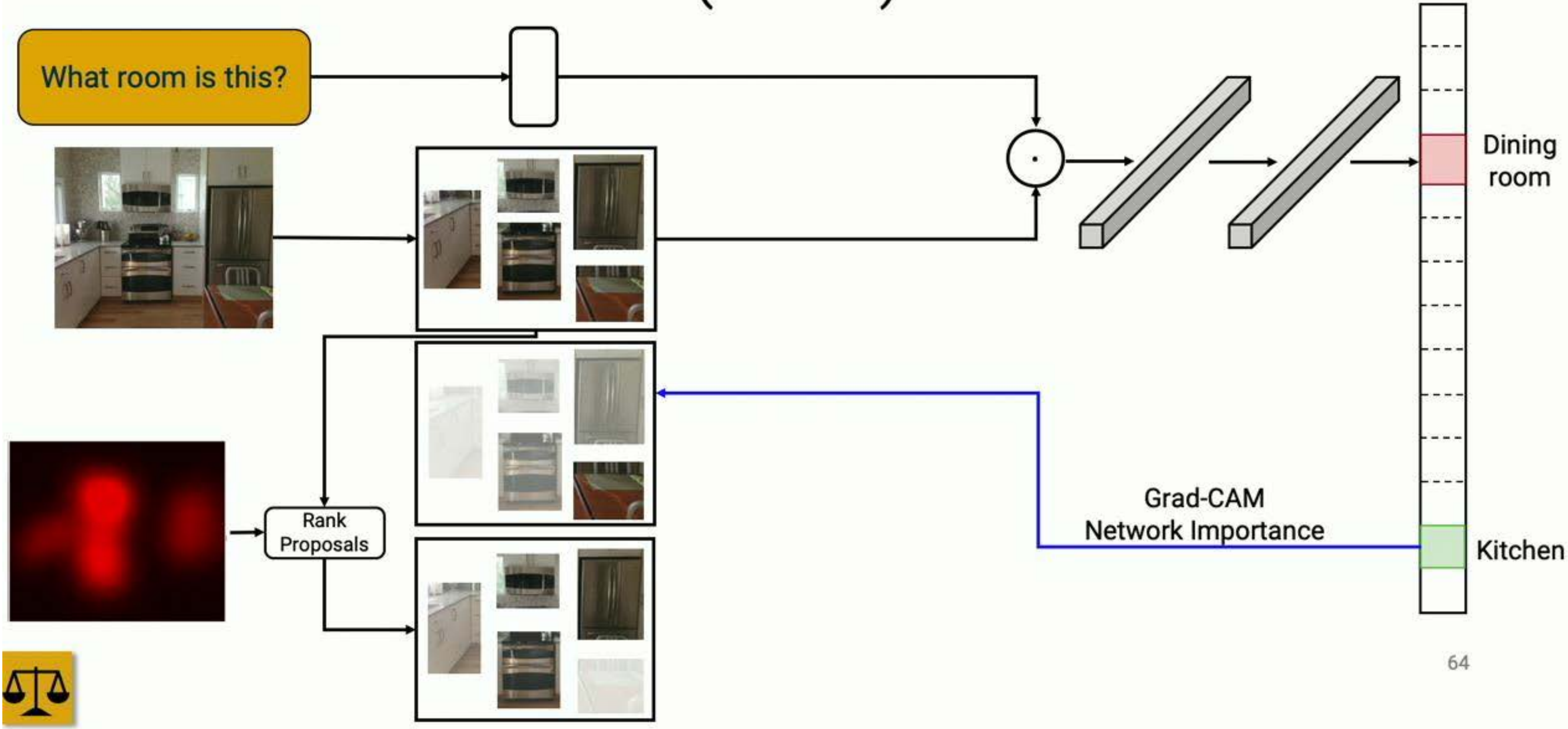
What room is this?

Dining room

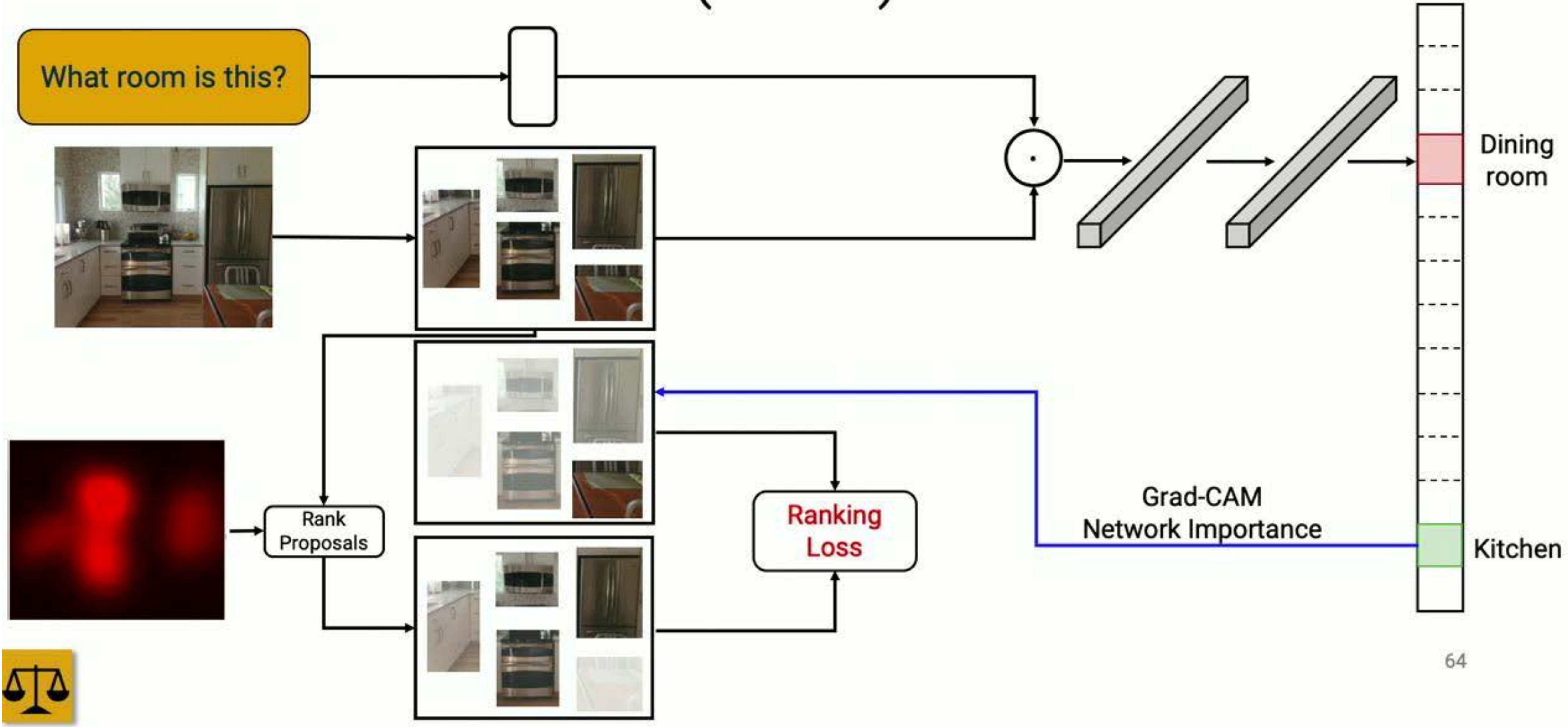
Grad-CAM
Network Importance

Kitchen

Human Importance-aware Network Tuning (HINT)



Human Importance-aware Network Tuning (HINT)



What room is this?

Dining room

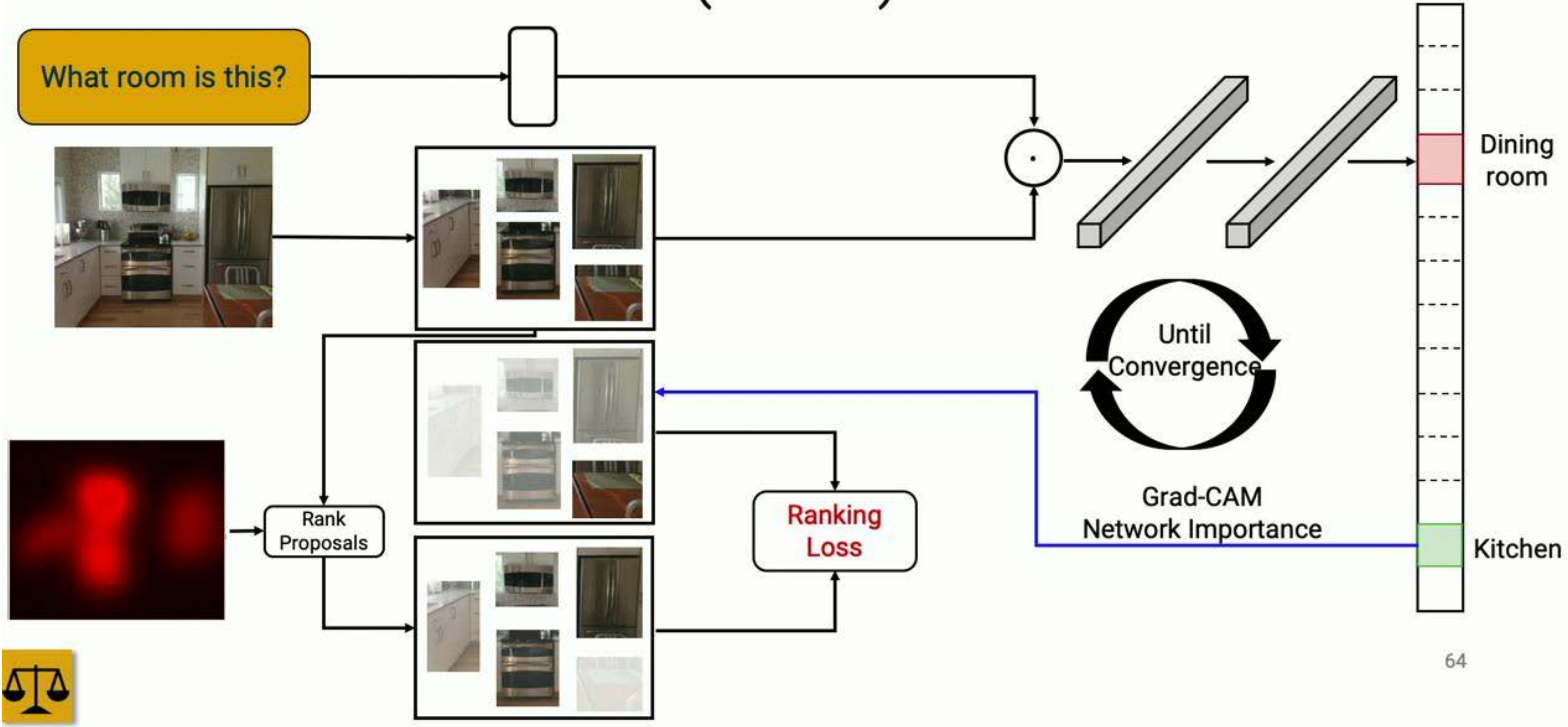
Rank Proposals

Ranking Loss

Grad-CAM Network Importance

Kitchen

Human Importance-aware Network Tuning (HINT)



What room is this?

Rank Proposals

Ranking Loss

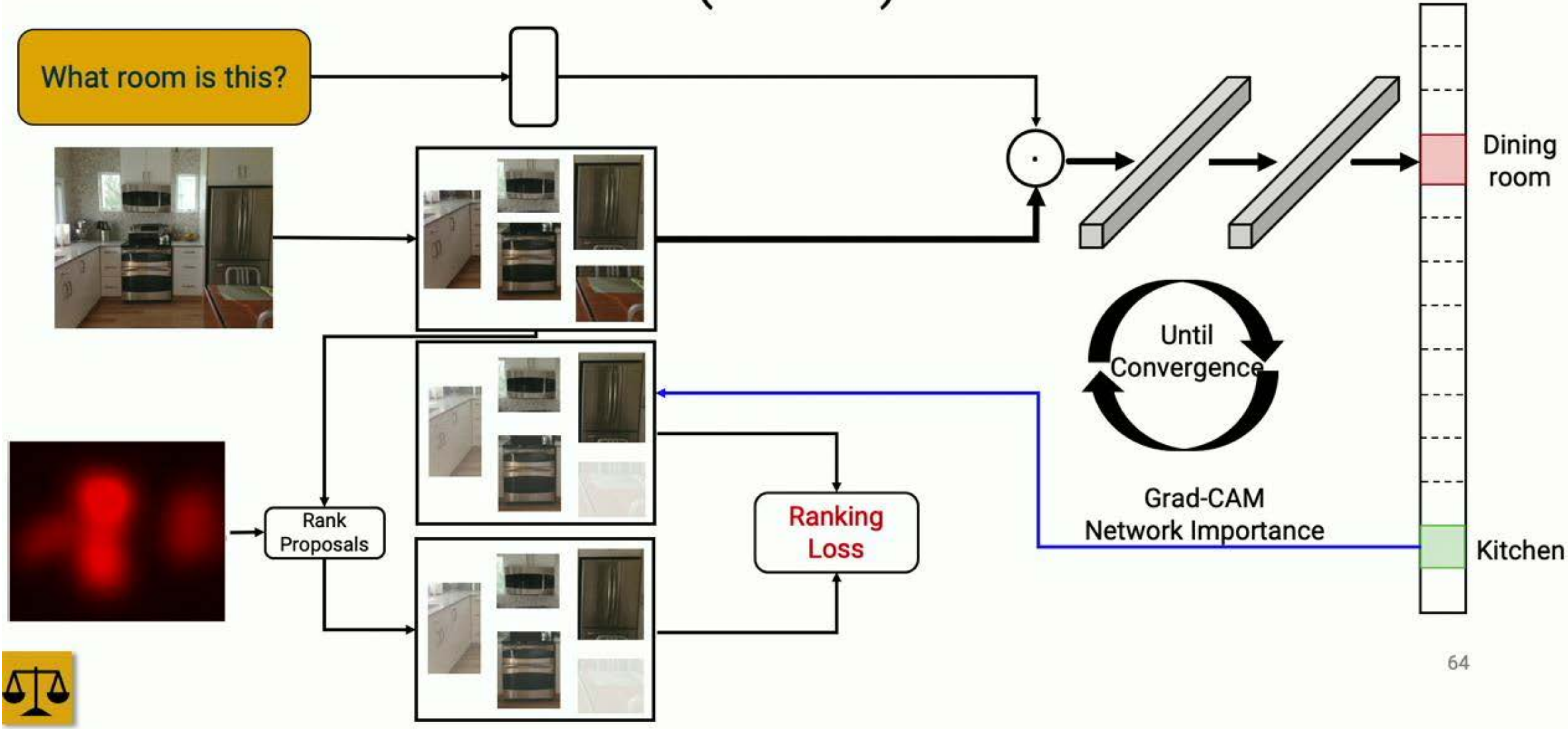
Until Convergence

Grad-CAM Network Importance

Dining room

Kitchen

Human Importance-aware Network Tuning (HINT)



What room is this?

Dining room

Until Convergence

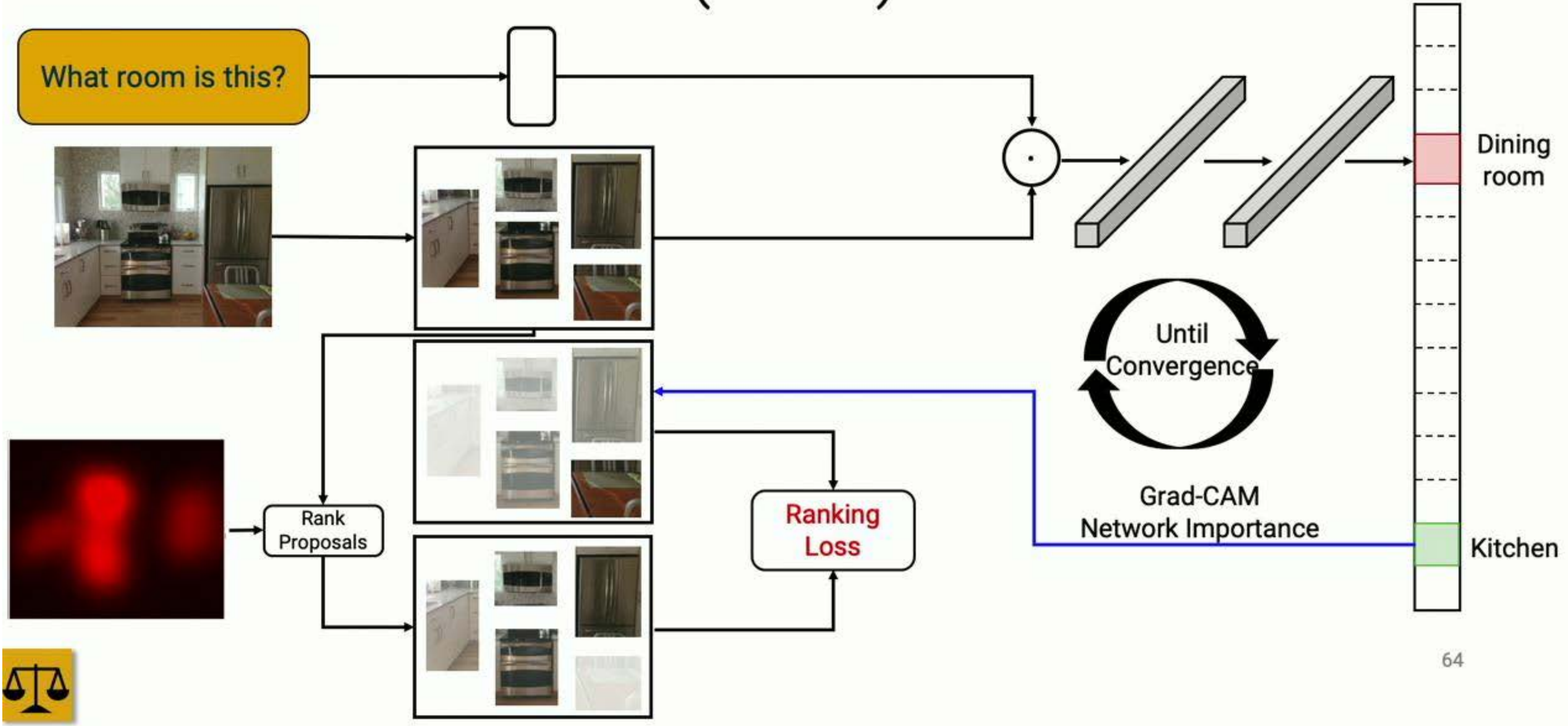
Rank Proposals

Ranking Loss

Grad-CAM Network Importance

Kitchen

Human Importance-aware Network Tuning (HINT)



What room is this?

Dining room

Until Convergence

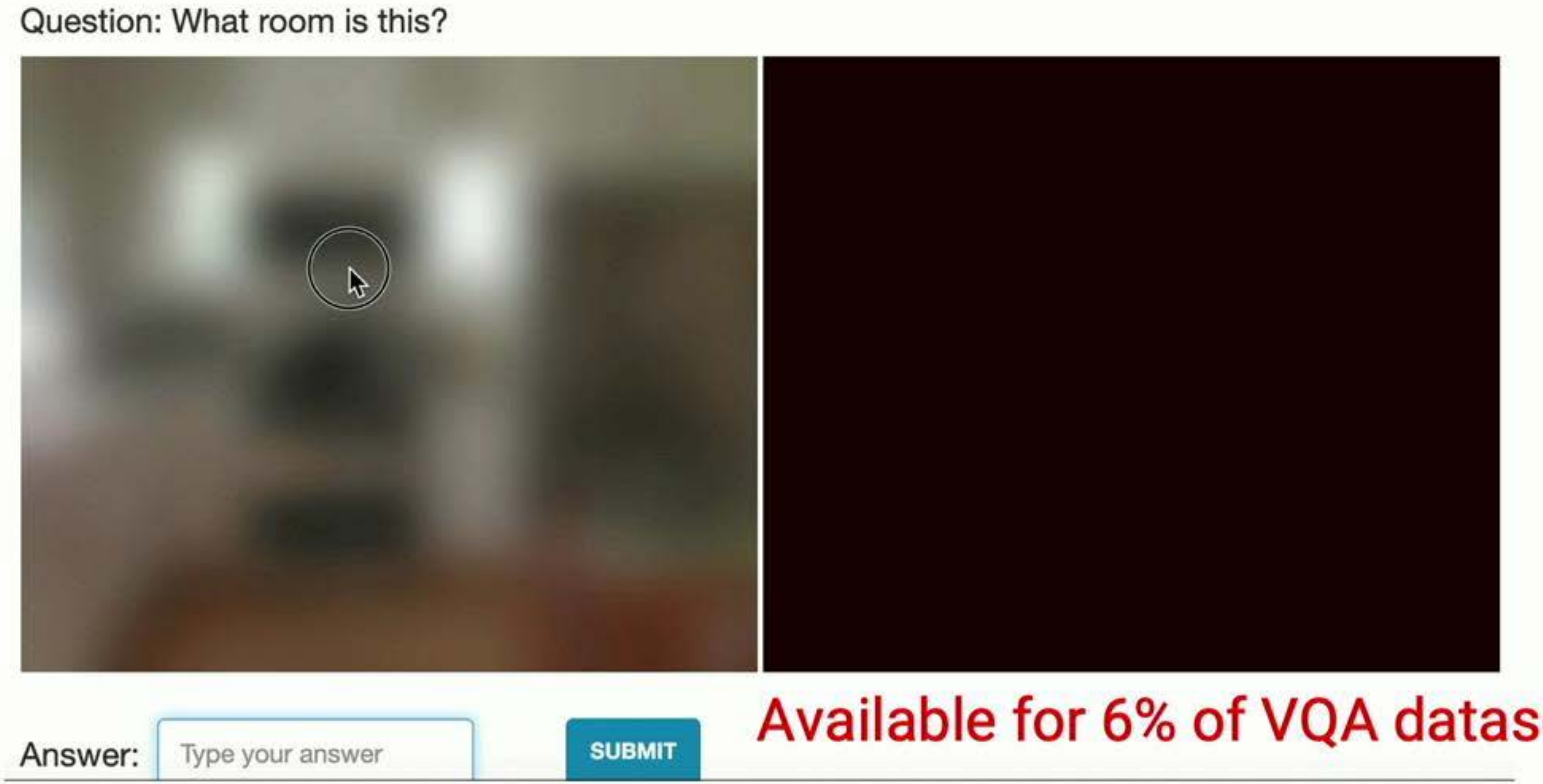
Rank Proposals

Ranking Loss

Grad-CAM Network Importance

Kitchen

Where do humans look when making decisions?



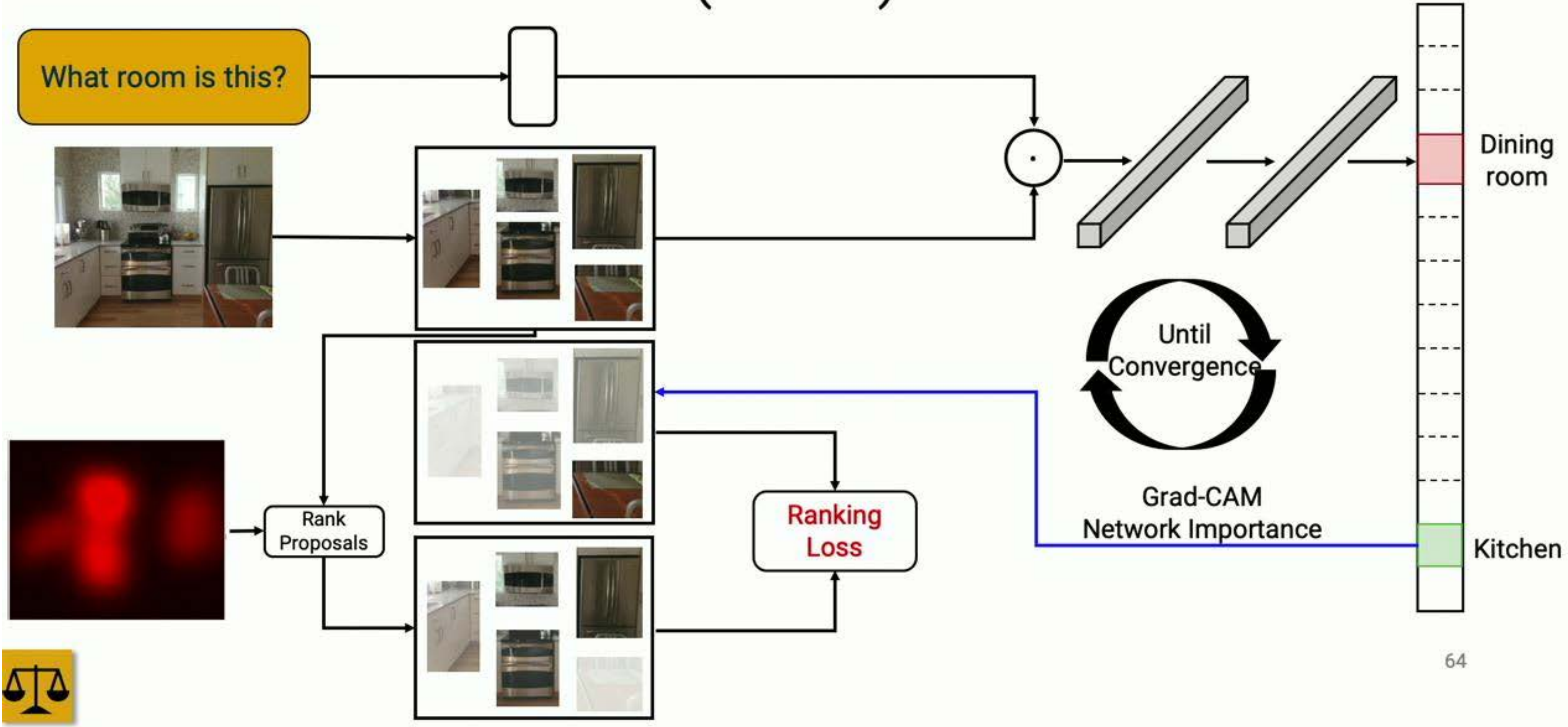
Question: What room is this?

Available for 6% of VQA dataset

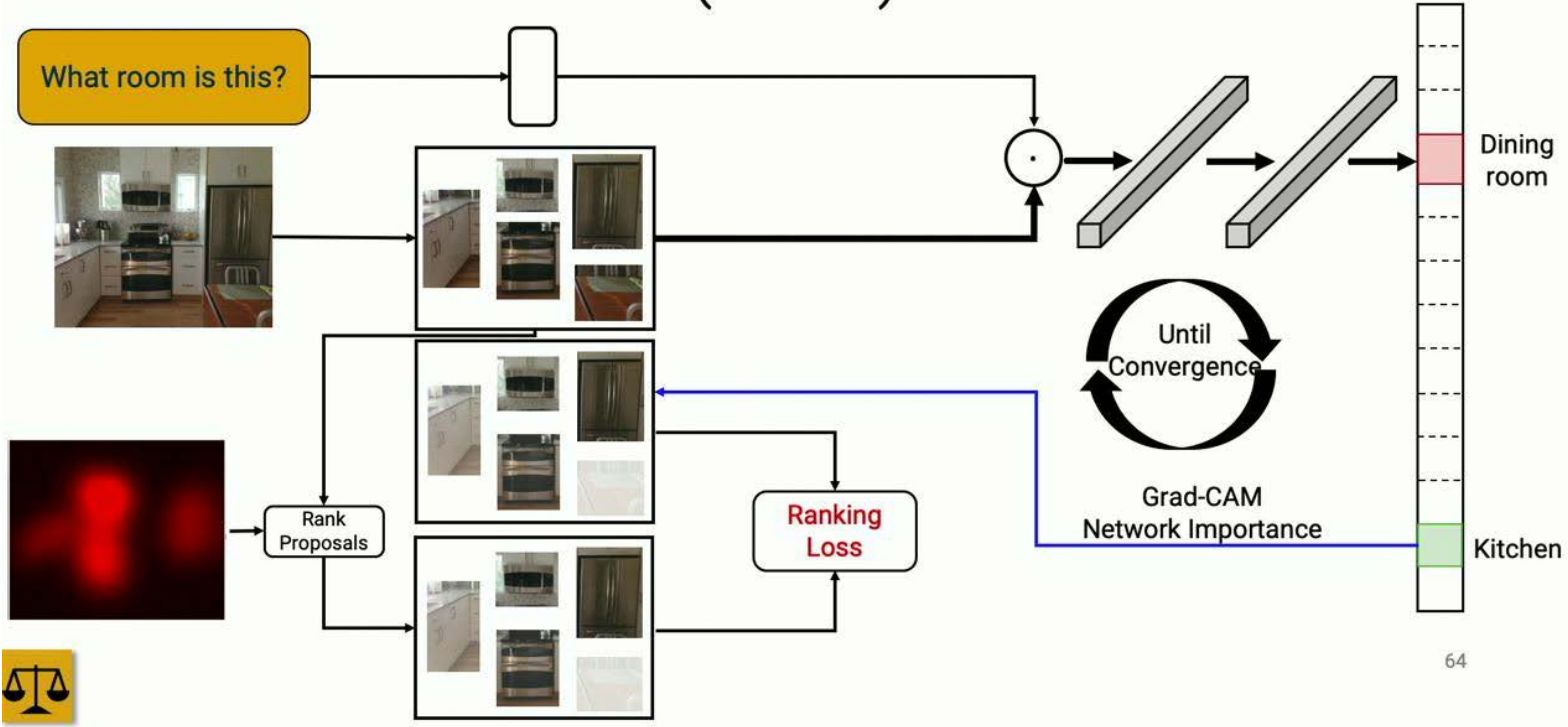
Das & Agrawal et al.

Human importance

Human Importance-aware Network Tuning (HINT)



Human Importance-aware Network Tuning (HINT)



What room is this?

Dining room

Until Convergence

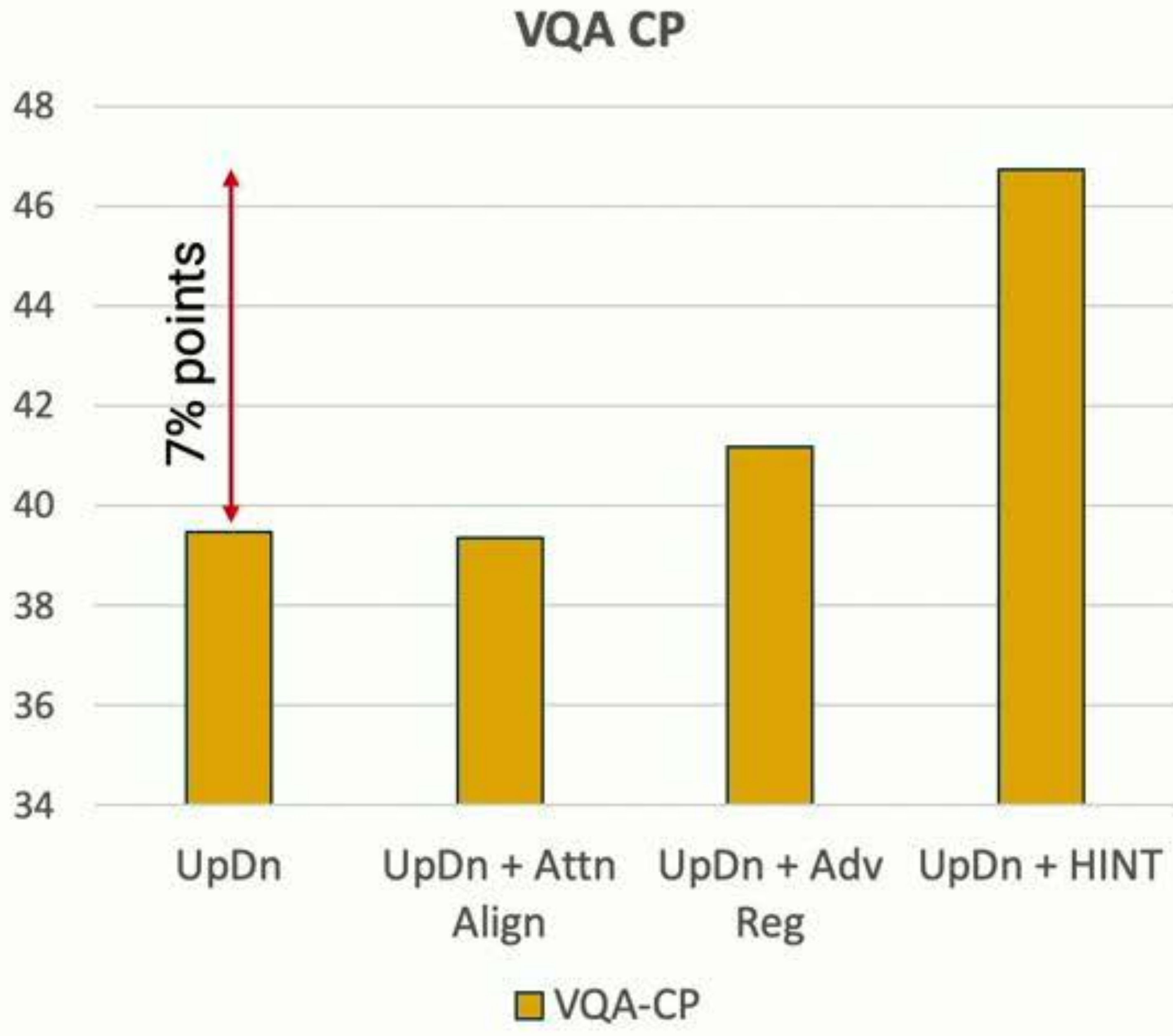
Grad-CAM Network Importance

Kitchen

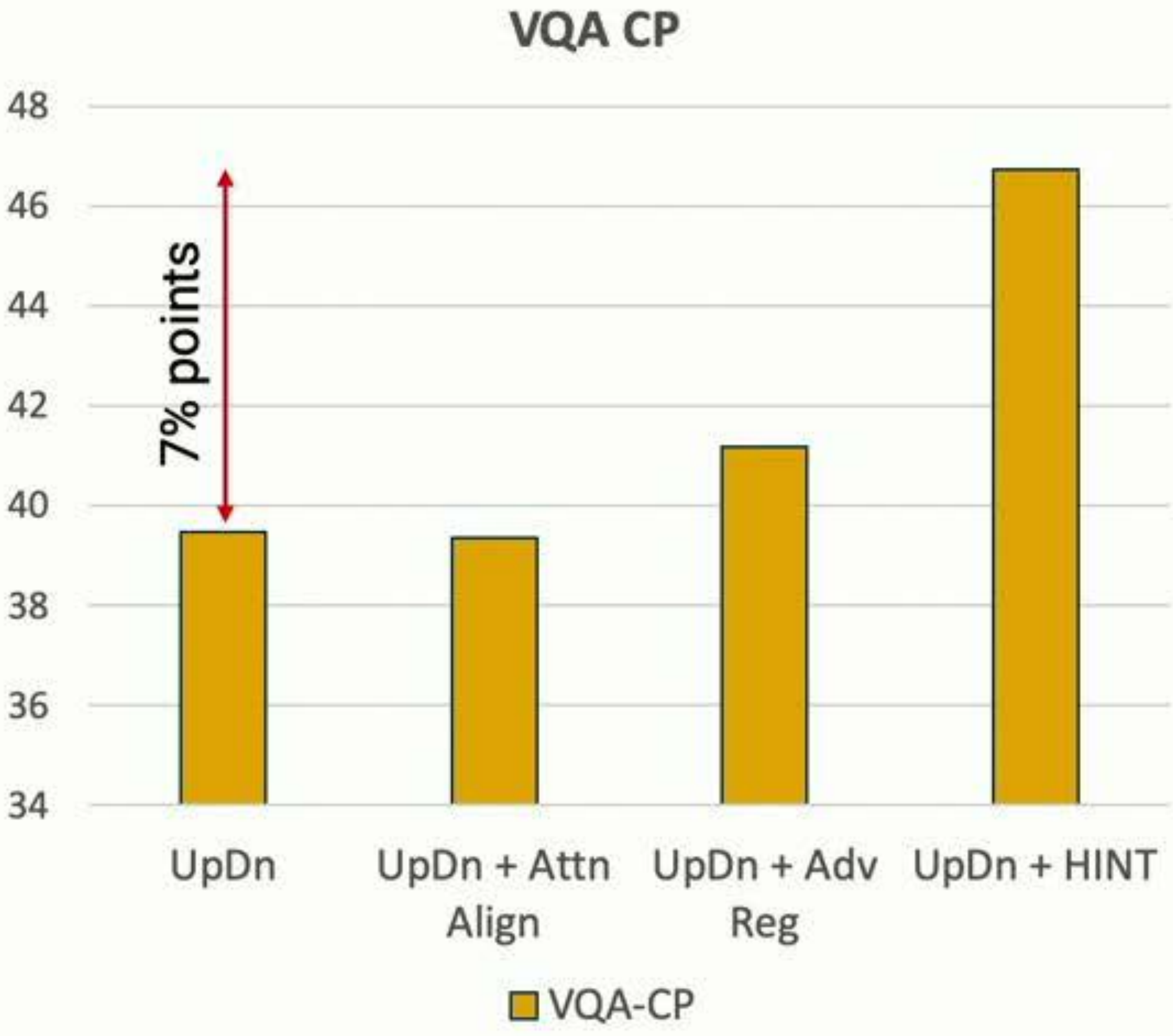
Rank Proposals

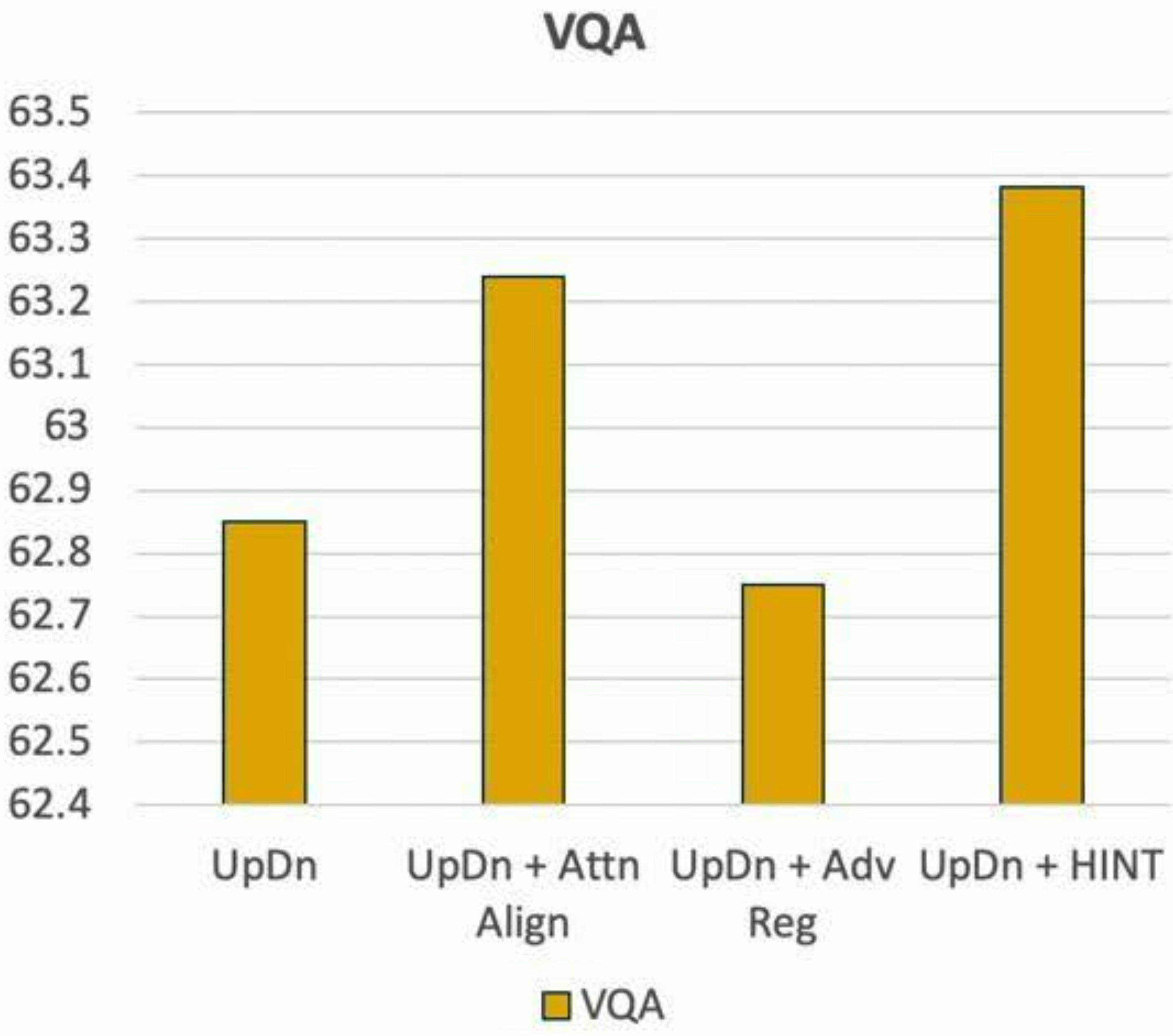
Ranking Loss

Results

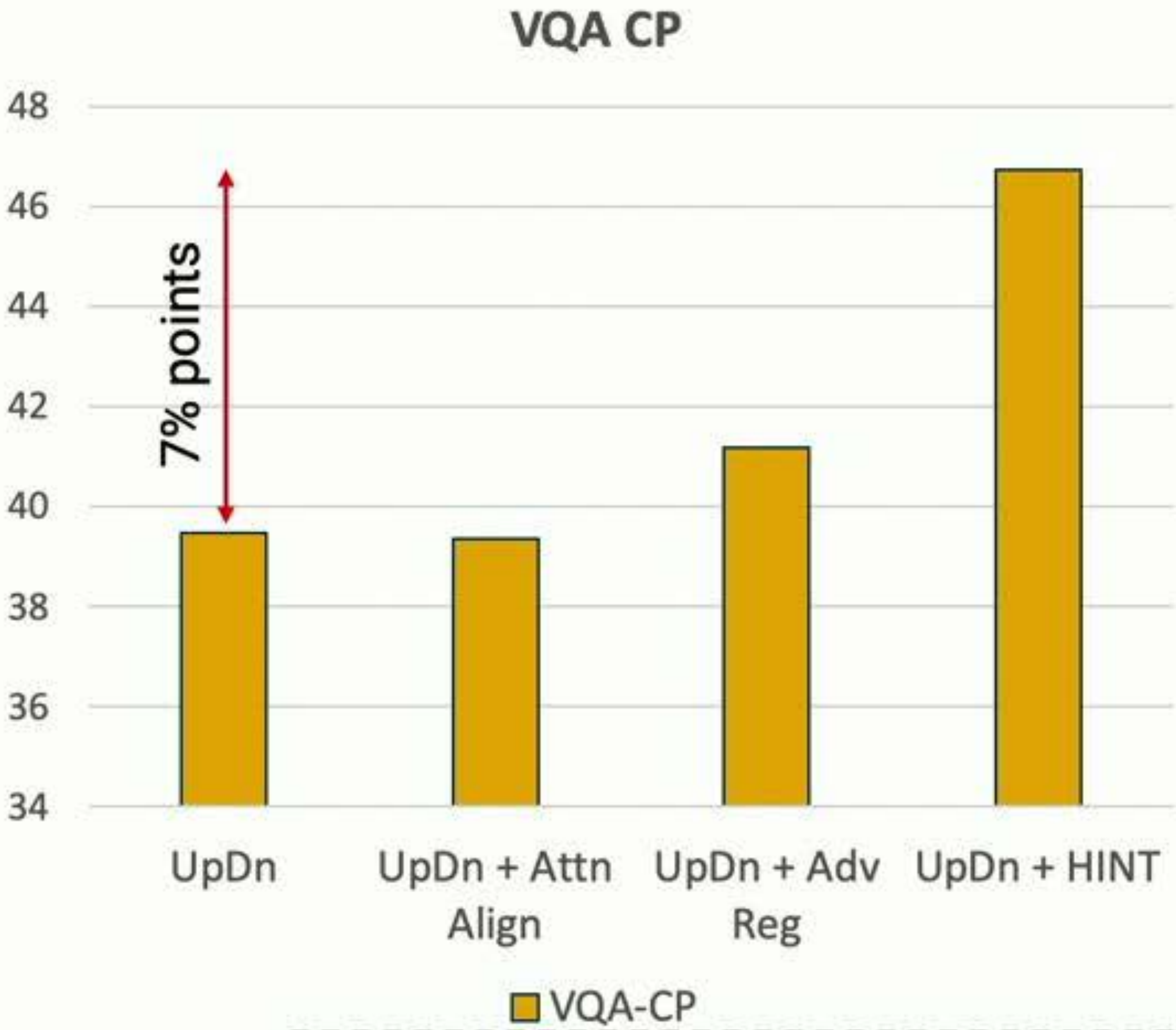

VQA CP
48
46
44
42
40
38
36
34
7% points
UpDn
UpDn + Attn Align
UpDn + Adv Reg
UpDn + HINT
VQA-CP

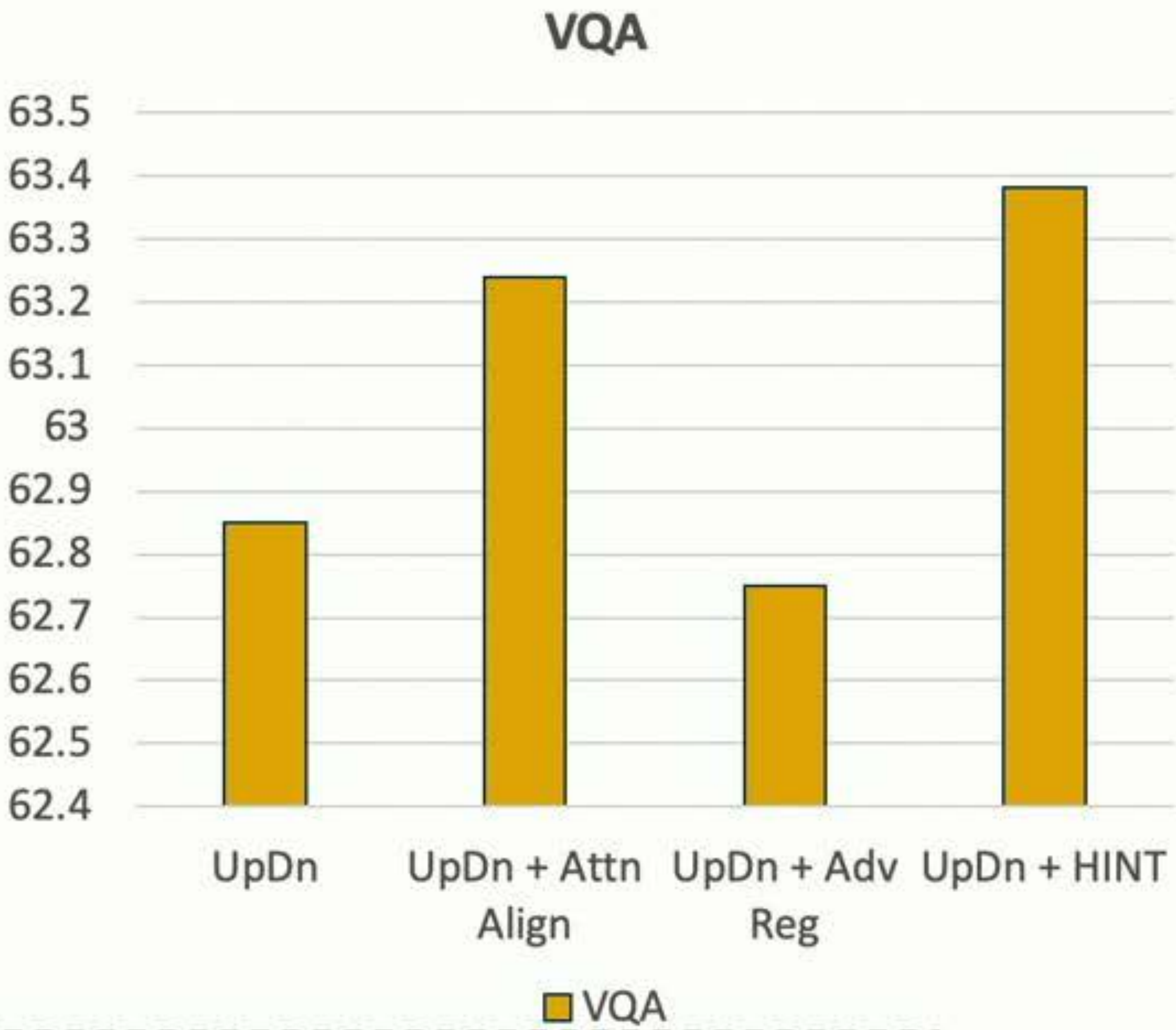
Results


VQA CP
48
46
44
42
40
38
36
34
7% points
UpDn
UpDn + Attn Align
UpDn + Adv Reg
UpDn + HINT
VQA-CP


VQA
63.5
63.4
63.3
63.2
63.1
63
62.9
62.8
62.7
62.6
62.5
62.4
UpDn
UpDn + Attn Align
UpDn + Adv Reg
UpDn + HINT
VQA

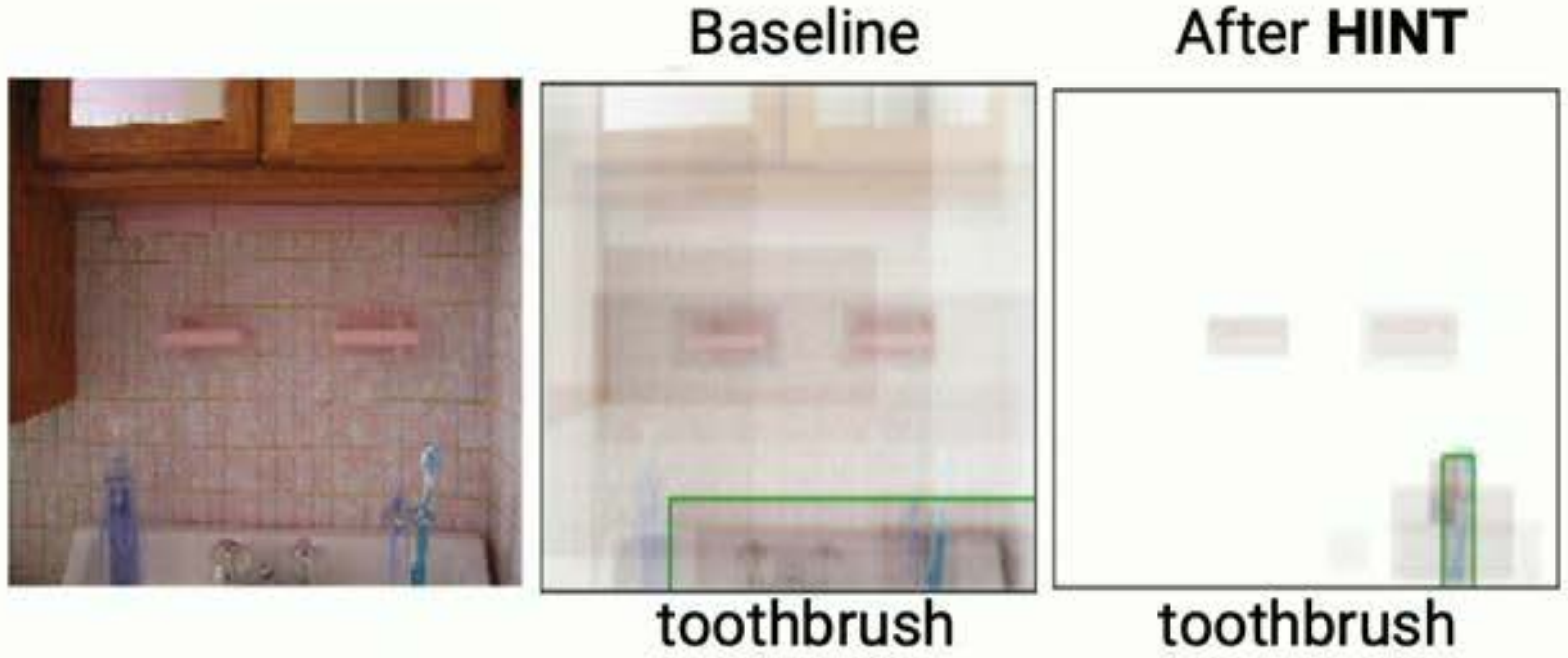
Results





Making machines look at regions like humans makes them generalize to arbitrary distributions better

Image Captioning – Do HINTed models look at right regions?

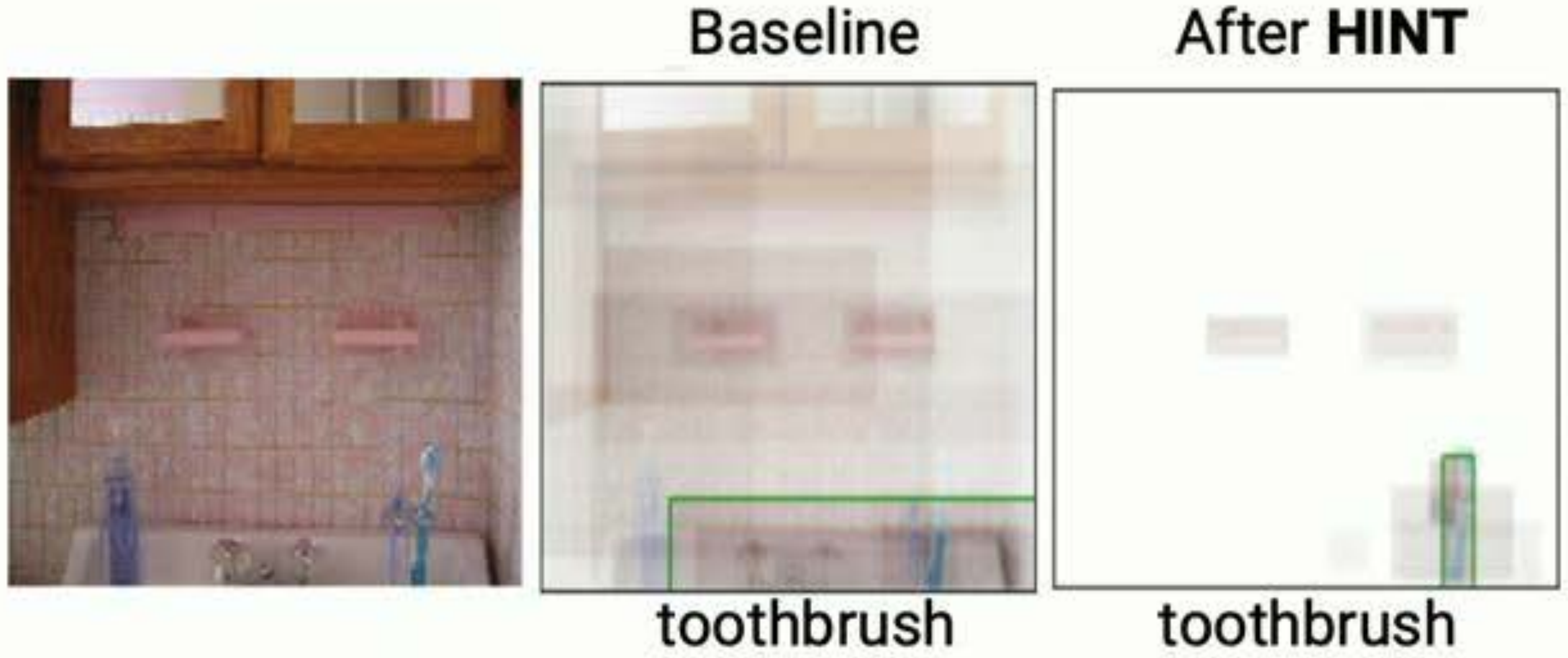


A bathroom sink with a **toothbrush**, soap dispenser and mirror

Image Captioning – Do HINTed models look at right regions?

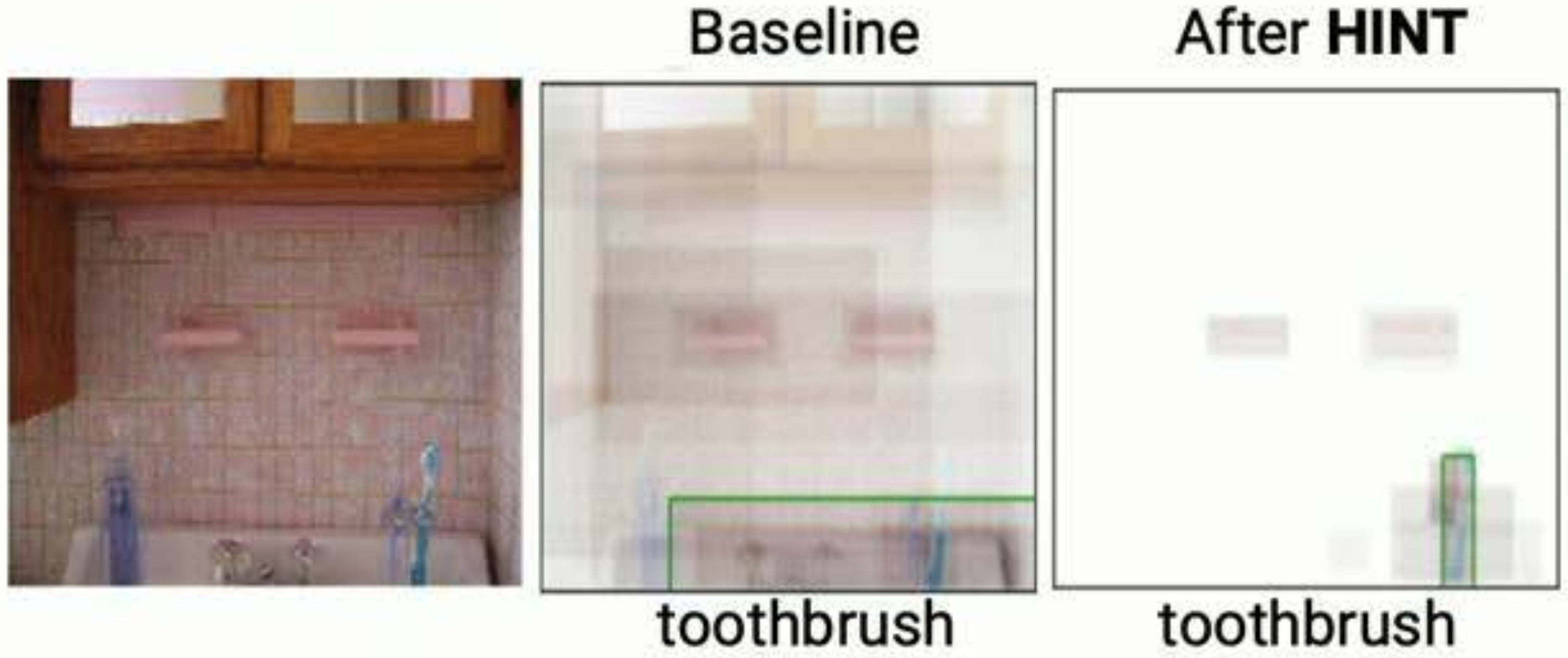
A bathroom sink with a **toothbrush**, soap dispenser and mirror

Image Captioning – Do HINTed models look at right regions?

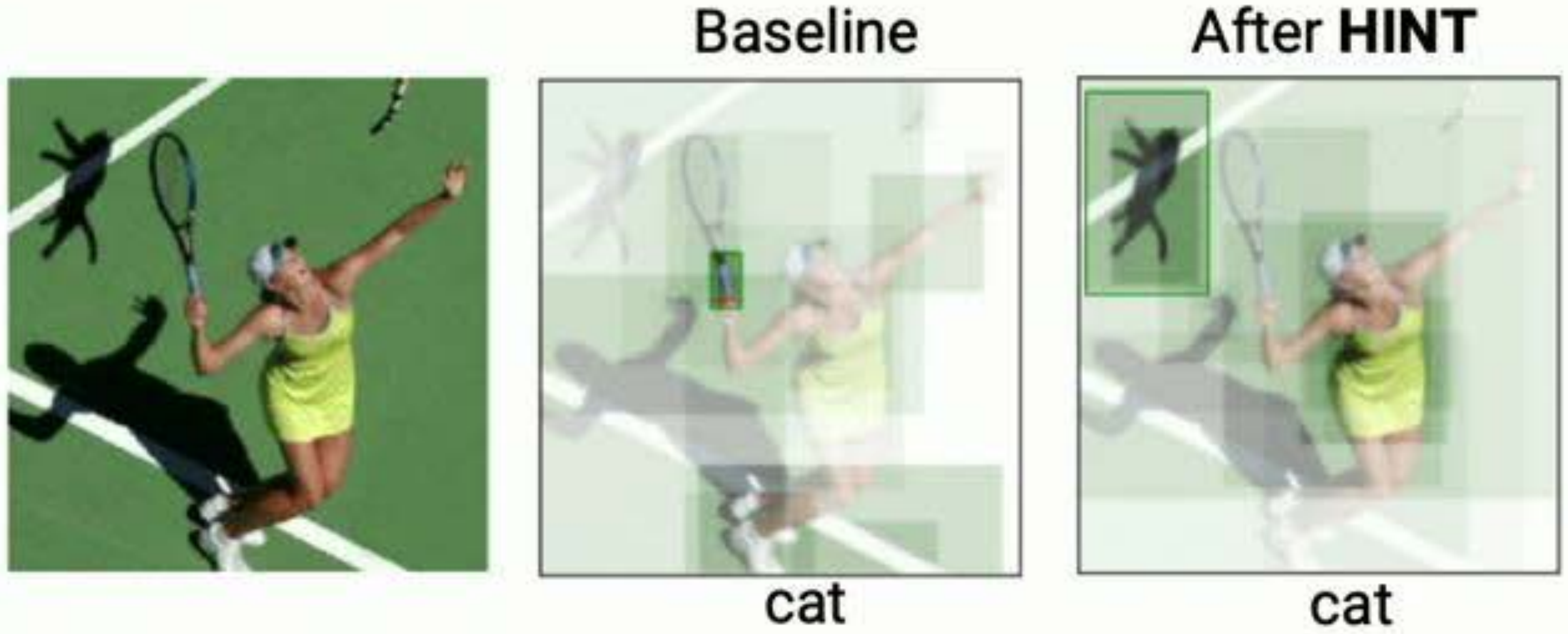


A bathroom sink with a **toothbrush**, soap dispenser and mirror

Image Captioning – Do HINTed models look at right regions?



A bathroom sink with a **toothbrush**, soap dispenser and mirror



A woman with a tennis racket with a **cat** in the air

HINT- Limitations

- Human attention can be misleading

HINT- Limitations

- **Human attention can be misleading**

What color is the sky?

Gray

HINT- Limitations

- Human attention can be misleading

What color is the sky?

Gray

Human Attention

Gray

HINT- Limitations

- Human attention can be misleading

What color is the sky?

Gray

Human Attention

Gray

Forcing machines to look at such regions might confuse them

HINT- Limitations

- In some cases it is not clear what region is even important

HINT- Limitations

- In some cases it is not clear what region is even important

Are the man and woman together?

Human Attention

No

Need for deeper understanding beyond visual context

Summary

- Models are biased
  - Tend to make decisions based on statistical correlations in the training data

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Summary

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

- Models are biased
  - Tend to make decisions based on statistical correlations in the training data
- Introduce **HINT**: Making machines look at regions like humans makes them generalize to arbitrary distributions

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What's future directions excite me?

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What's future directions excite me?

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, Ece Kamar, **" SQuINTing at VQA Models: Interrogating VQA Models with Sub-Questions",** Under Review

Can models benefit from human-like compositional reasoning?

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, Ece Kamar, " **SQuINTing at VQA Models: Interrogating VQA Models with Sub-Questions",** Under Review

Collaborators at Microsoft

Ece Kamar

Besmira Nushi

Marco Tulio Ribeiro

Eric Horvitz

VQA for visually impaired users

VQA for visually impaired users

Is the banana ripe enough to eat?

VQA for visually impaired users

Is the banana ripe enough to eat?

Yes

VQA for visually impaired users

Is the banana ripe enough to eat?

Yes

Is the banana mostly green or yellow?

VQA for visually impaired users

Is the banana ripe enough to eat?

Yes

Is the banana mostly green or yellow?

Green

VQA for situationally blind

VQA for situationally blind

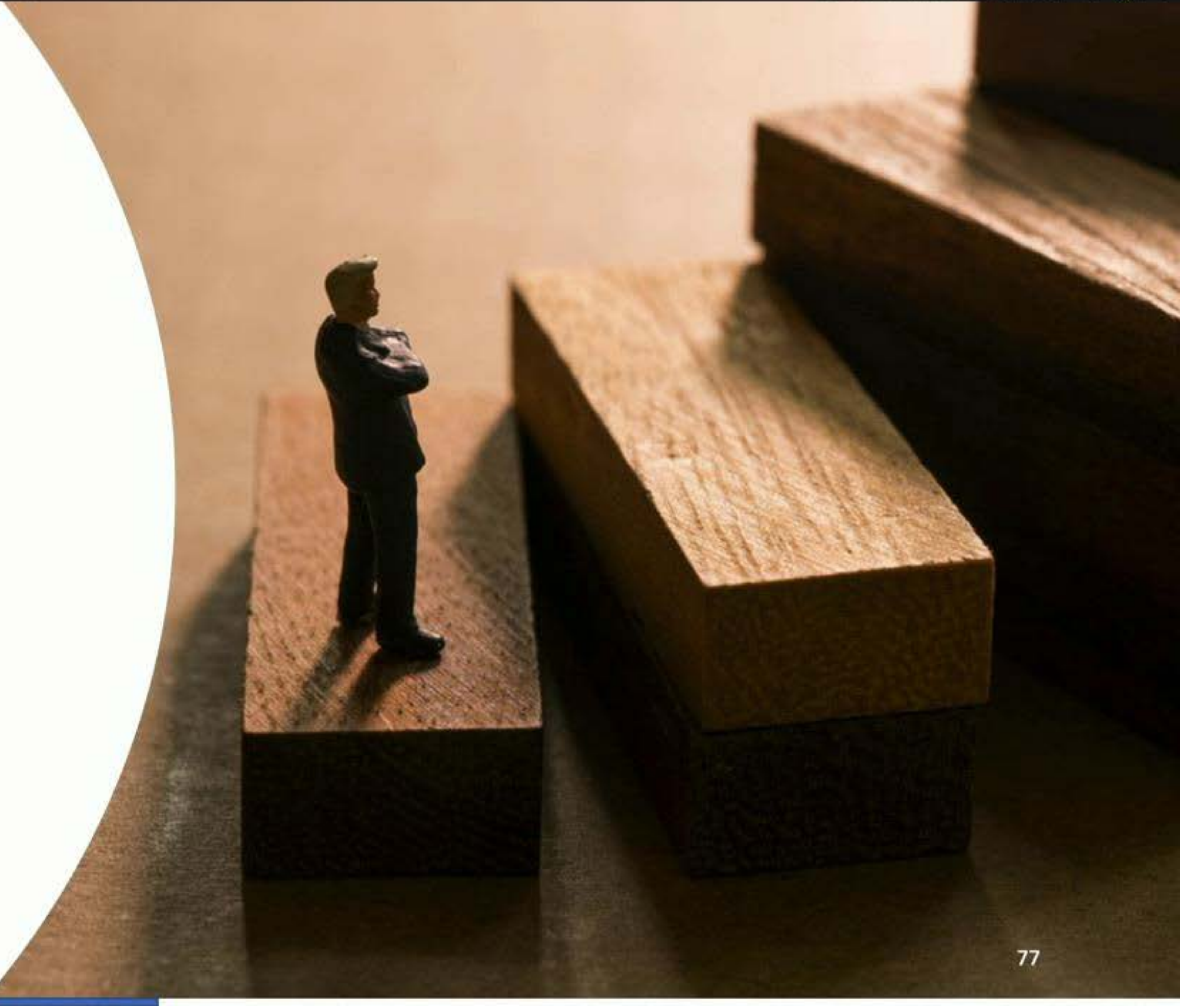
Is there an emergency?
No

Is the room on fire?
Yes

Is there a lot of smoke in the room?
Yes

Are there people?
Yes

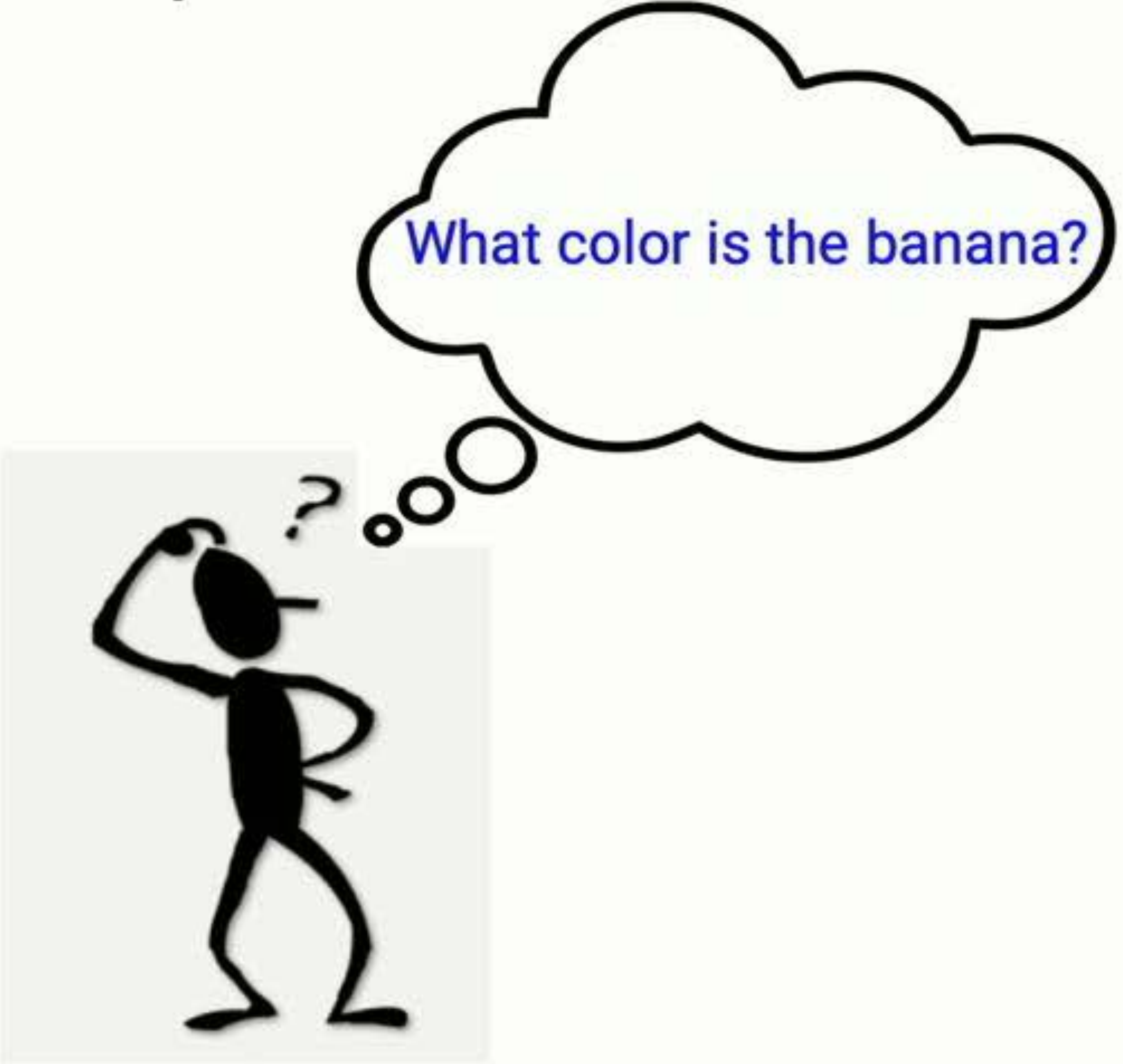
How do we reason?



Human compositional reasoning

Is the banana ripe enough to eat?

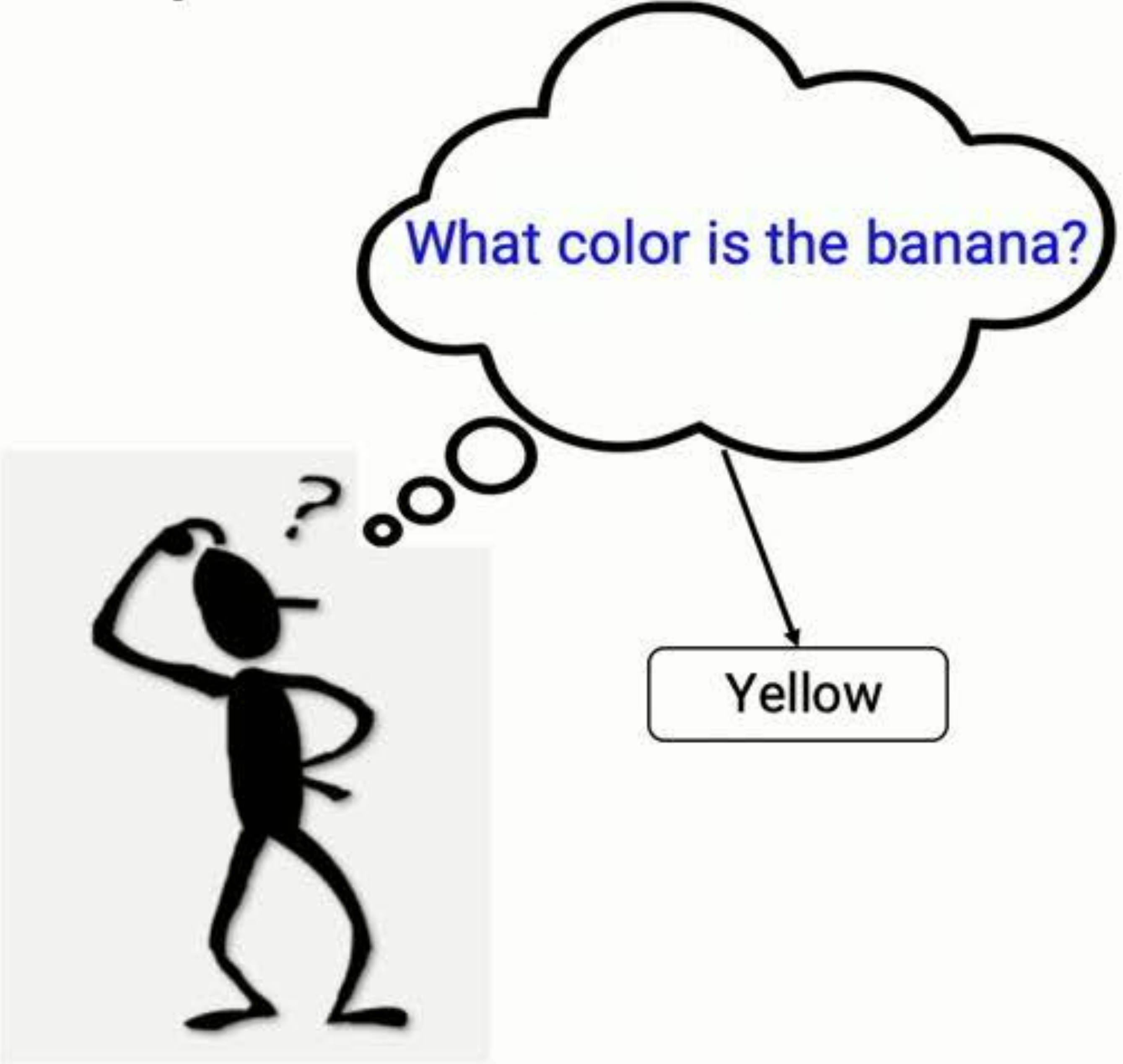
Human compositional reasoning



What color is the banana?

Is the banana ripe enough to eat?

Human compositional reasoning

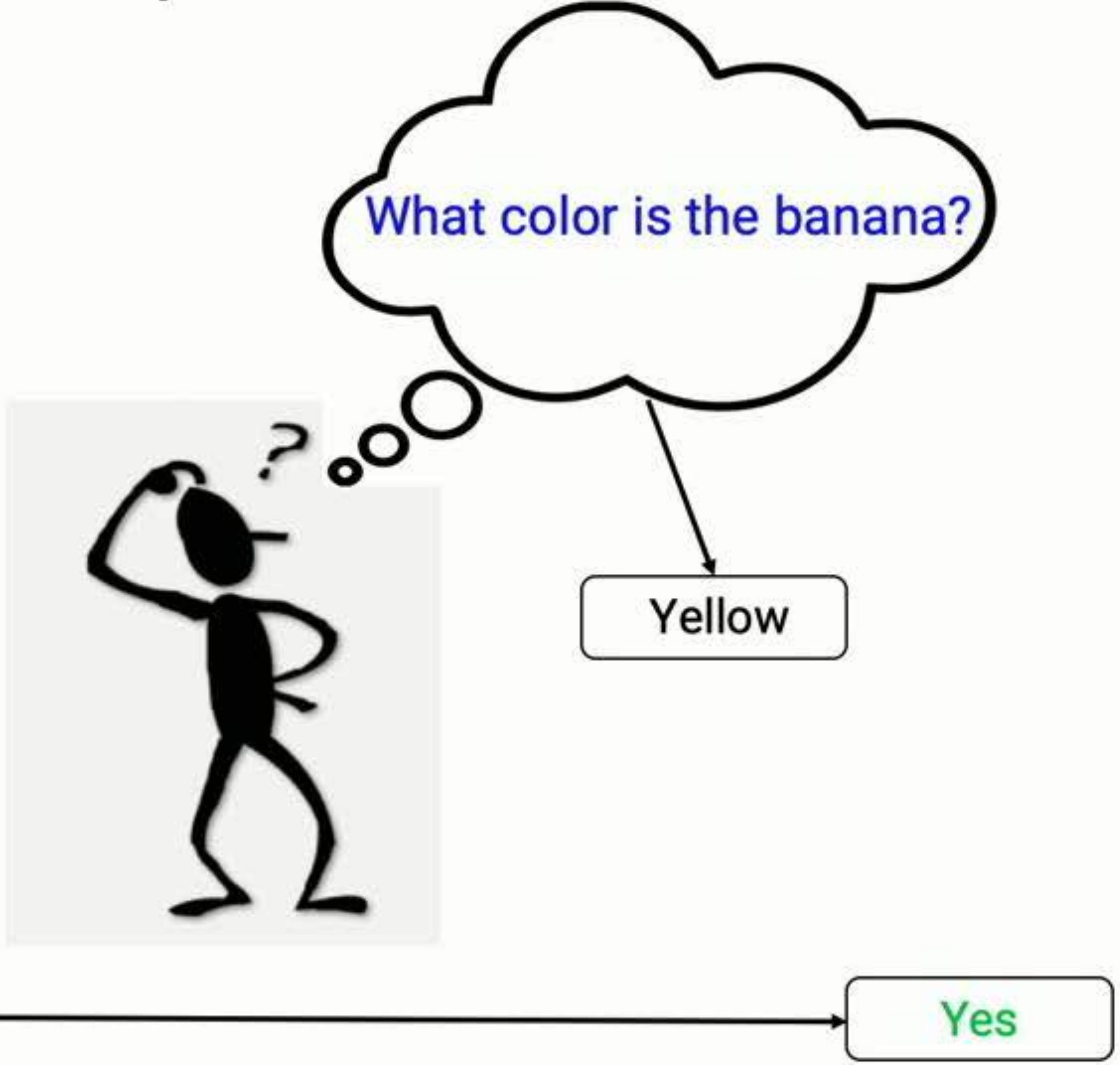


What color is the banana?

Yellow

Is the banana ripe enough to eat?

Human compositional reasoning



Is the banana ripe enough to eat?

How are current VQA models trained/evaluated?

- All questions are treated equally

How are current VQA models trained/evaluated?

- All questions are treated equally
    - Feed all kinds of questions randomly

How are current VQA models trained/evaluated?

- All questions are treated equally
    - Feed all kinds of questions randomly
    - No regard to complexity or commonsense requirement

How are current VQA models trained/evaluated?

- All questions are treated equally
  - Feed all kinds of questions randomly
  - No regard to complexity or commonsense requirement

What color is the man's shirt?

Is this a good idea for a rainy day?

How are current VQA models trained/evaluated?

- All questions are treated equally
  - Feed all kinds of questions randomly
  - No regard to complexity or commonsense requirement
- Expect models to learn compositionality

Need to classify questions based on complexity

Need to classify questions based on complexity

- Physical properties of objects/entities:
    - What color is the couch? Red

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Do they look relaxed? Yes

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Do they look relaxed? Yes

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Do they look relaxed? Yes

Need to classify questions based on complexity

- Physical properties of objects/entities:
    - What color is the couch? Red
- Existence:
    - Is there a fork? Yes
- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
  - How many bears are there? 2
- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
  - How many bears are there? 2
- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
  - How many bears are there? 2

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
  - How many bears are there? 2
- Spatial relationship:
  - What is to the right of the plate? glass

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes

Need to classify questions based on complexity

- Physical properties of objects/entities:
    - What color is the couch? Red
- Existence:
    - Is there a fork? Yes
- Counts:
    - How many bears are there? 2
- Spatial relationship:
    - What is to the right of the plate? glass

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes
- Is this breakfast, lunch or dinner? Breakfast

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
  - How many bears are there? 2
- Spatial relationship:
  - What is to the right of the plate? glass
- Text/Symbol recognition:
  - What does the sign say? Stop

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes
- Is this breakfast, lunch or dinner? Breakfast

Need to classify questions based on complexity

- Physical properties of objects/entities:
    - What color is the couch? Red
- Existence:
    - Is there a fork? Yes
- Counts:
    - How many bears are there? 2
- Spatial relationship:
    - What is to the right of the plate? glass
- Text/Symbol recognition:
    - What does the sign say? Stop

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes
- Is this breakfast, lunch or dinner? Breakfast
- Is it going to rain here? No

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
  - How many bears are there? 2
- Spatial relationship:
  - What is to the right of the plate? glass
- Text/Symbol recognition:
  - What does the sign say? Stop

Perception

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes
- Is this breakfast, lunch or dinner? Breakfast
- Is it going to rain here? No

Need to classify questions based on complexity

- Physical properties of objects/entities:
  - What color is the couch? Red
- Existence:
  - Is there a fork? Yes
- Counts:
  - How many bears are there? 2
- Spatial relationship:
  - What is to the right of the plate? glass
- Text/Symbol recognition:
  - What does the sign say? Stop

- Do they look relaxed? Yes
- Could you pick up this pizza to eat it? No
- Was this picture taken in Australia? Yes
- Is this breakfast, lunch or dinner? Breakfast
- Is it going to rain here? No

Perception

Reasoning

Sub-VQA Dataset

- Collect sub-questions to obtain perceptual evidence for Reasoning questions

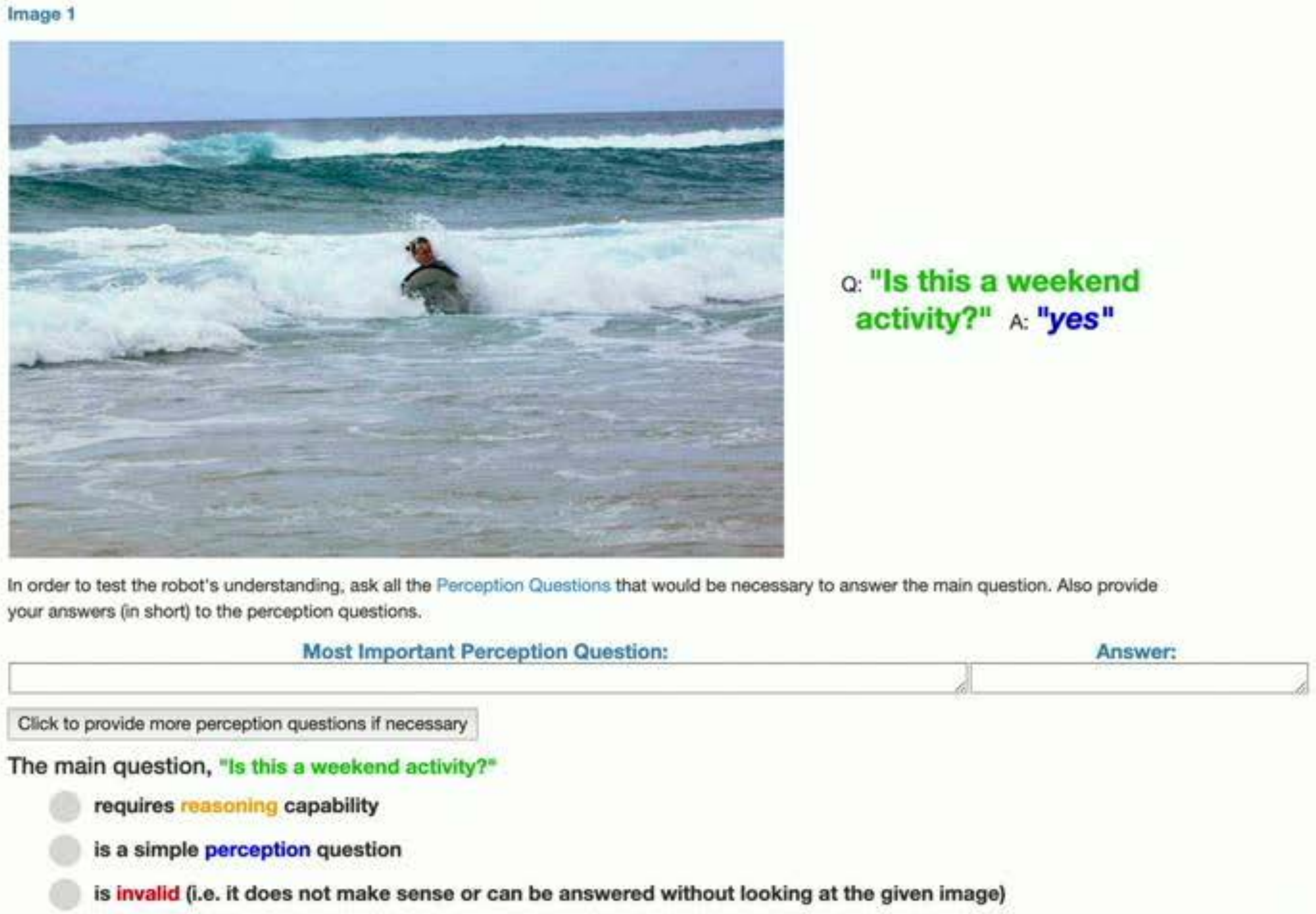


Image 1

Q: "Is this a weekend activity?" A: "yes"

In order to test the robot's understanding, ask all the Perception Questions that would be necessary to answer the main question. Also provide your answers (in short) to the perception questions.

Most Important Perception Question:	Answer:

The main question, "Is this a weekend activity?"

- ○ requires reasoning capability
- ○ is a simple perception question
- ○ is invalid (i.e. it does not make sense or can be answered without looking at the given image)

Sub-VQA Dataset

Main Reasoning Question:

- Is this a keepsake photo? "Yes"

Perception Sub-questions:

- Is this a black and white photo? "Yes"
- Is the woman wearing a white veil and holding flowers? "Yes"
- Is the woman wearing a veil? "Yes"
- What is the woman next to the man wearing? "Gown"

Sub-VQA Dataset

Main Reasoning Question:
- Is this a keepsake photo? "Yes"

Perception Sub-questions:
- Is this a black and white photo? "Yes"
- Is the woman wearing a white veil and holding flowers? "Yes"
- Is the woman wearing a veil? "Yes"
- What is the woman next to the man wearing? "Gown"

Main Reasoning Question:
- Is this giraffe at the zoo? "Yes"

Perception Sub-questions:
- Is the giraffe fenced in? "Yes"
- Is the grass shorter than 3 inches? "Yes"
- Is there a fence? "Yes"
- Is a fence around the giraffe? "Yes"

Sub-VQA Dataset

Main Reasoning Question:

- Is this a keepsake photo? "Yes"

Perception Sub-questions:

- Is this a black and white photo? "Yes"
- Is the woman wearing a white veil and holding flowers? "Yes"
- Is the woman wearing a veil? "Yes"
- What is the woman next to the man wearing? "Gown"

Main Reasoning Question:

- Is this giraffe at the zoo? "Yes"

Perception Sub-questions:

- Is the giraffe fenced in? "Yes"
- Is the grass shorter than 3 inches? "Yes"
- Is there a fence? "Yes"
- Is a fence around the giraffe? "Yes"

Main Reasoning Question:

- Does this appear to be an emergency? "Yes"

Perception Sub-questions:

- Are there a lot of ambulances? "Yes"
- Are people standing in the middle of the street? "Yes"
- Is there a firetruck? "Yes"
- Does the white vehicle say "ambulance"? "Yes"
- Does the red truck say "fire department"? "Yes"

Sub-VQA Dataset

Main Reasoning Question:

- Is this a keepsake photo? "Yes"

Perception Sub-questions:

- Is this a black and white photo? "Yes"
- Is the woman wearing a white veil and holding flowers? "Yes"
- Is the woman wearing a veil? "Yes"
- What is the woman next to the man wearing? "Gown"

Main Reasoning Question:

- Is this giraffe at the zoo? "Yes"

Perception Sub-questions:

- Is the giraffe fenced in? "Yes"
- Is the grass shorter than 3 inches? "Yes"
- Is there a fence? "Yes"
- Is a fence around the giraffe? "Yes"

Main Reasoning Question:

- Does this appear to be an emergency? "Yes"

Perception Sub-questions:

- Are there a lot of ambulances? "Yes"
- Are people standing in the middle of the street? "Yes"
- Is there a firetruck? "Yes"
- Does the white vehicle say "ambulance"? "Yes"
- Does the red truck say "fire department"? "Yes"

Main Reasoning Question:

- Is this a good idea for a rainy day? "No"

Perception Sub-questions:

- Is there a roof on the bus? "No"
- Does the vehicle have a roof? "No"

How can Sub-VQA help?

How can Sub-VQA help?

- Evaluation:
  - Do current models reason compositionally?
    - How consistent are SOTA VQA models?
- Improving models:
  - Does human-like compositional reasoning help current models reason better?

Do current models reason compositionally?

- How consistent are SOTA approaches?

Perception and Reasoning Success

Perception Failure

Reasoning Failure

Perception and Reasoning Failure

Do current models reason compositionally?

- How consistent are SOTA approaches?

Overall : 60.26%

Perception and Reasoning Success	Perception Failure
Reasoning Failure	Perception and Reasoning Failure

Do current models reason compositionally?

- How consistent are SOTA approaches?

Overall : 60.26%

Perception and Reasoning Success 47.42%	Perception Failure 18.57%
Reasoning Failure 20.70%	Perception and Reasoning Failure 13.31%

28% of the times model is right for the wrong reasons

Does human-like compositional reasoning help current models?

- Does making models use the right perception concepts make models reason better?

Does human-like compositional reasoning help current models?

- Does making models use the right perception concepts make models reason better?

Is this a good idea for a rainy day?

Does the bus have a roof?

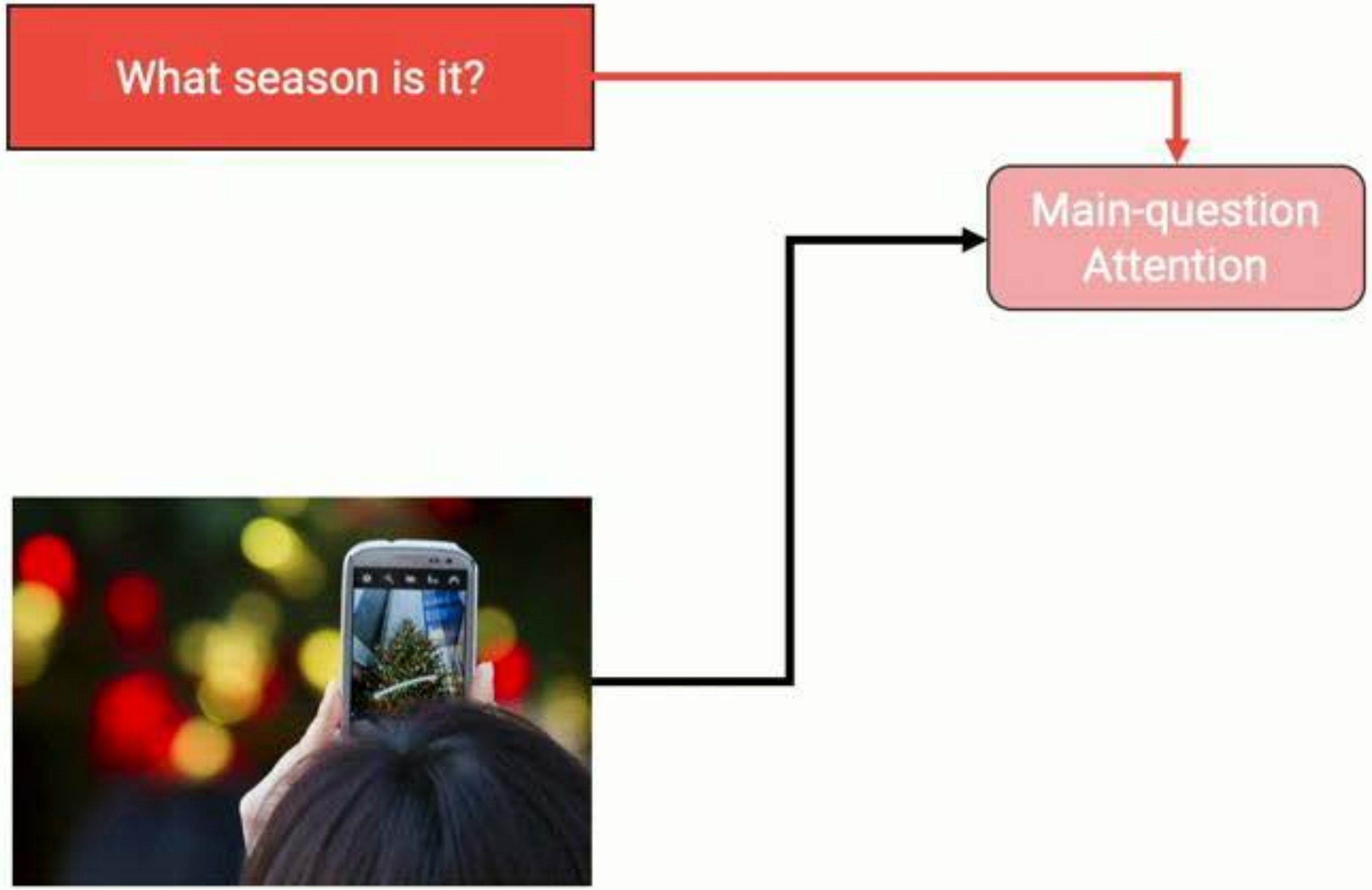
SQuINT

Sub-**Qu**estion **I**mportance-aware **N**etwork **T**uning

Sub-Question Importance-aware Network Tuning (SQuINT)

What season is it?

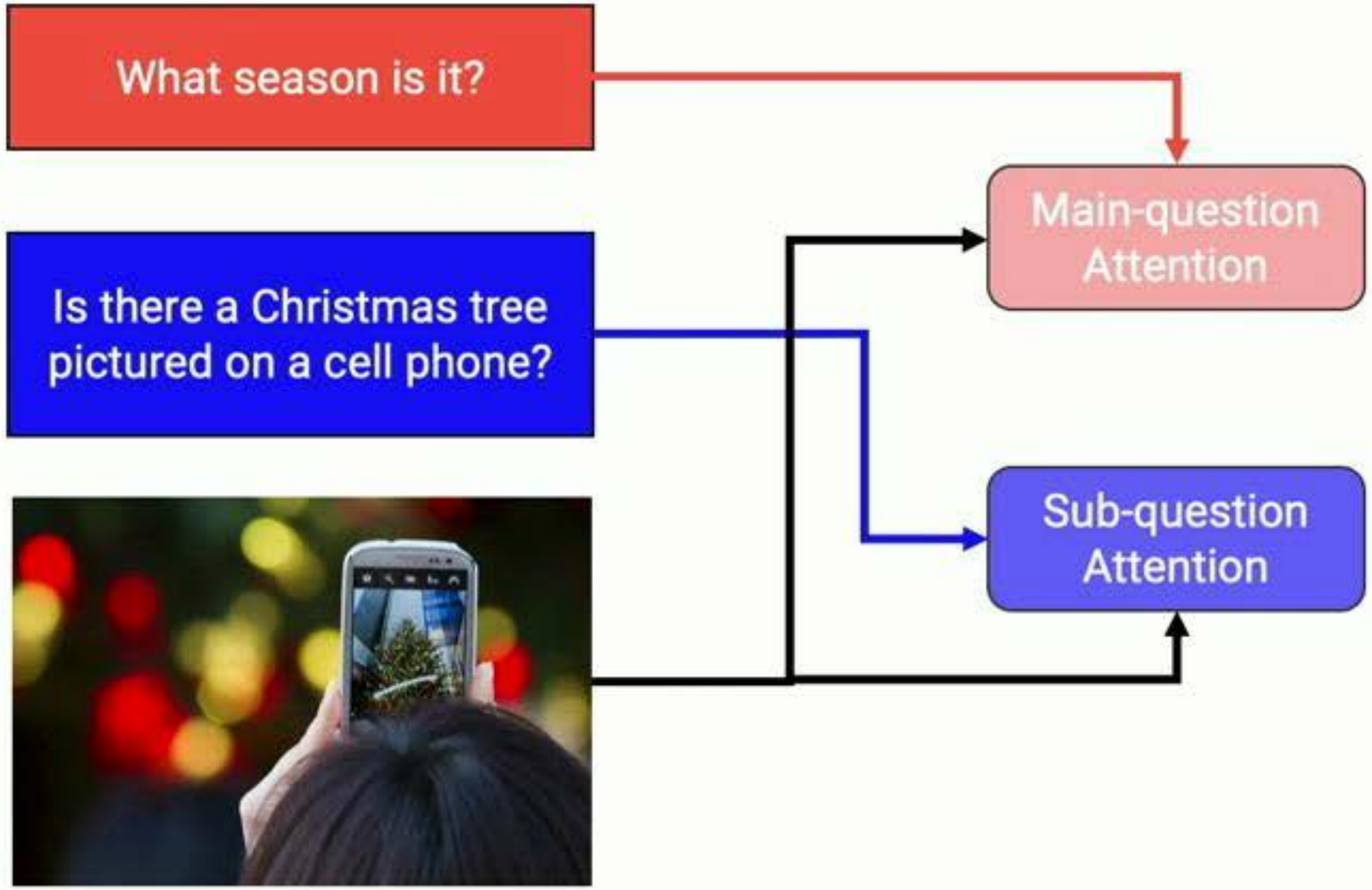
Sub-Question Importance-aware Network Tuning (SQuINT)



What season is it?

Main-question Attention

Sub-Question Importance-aware Network Tuning (SQuINT)



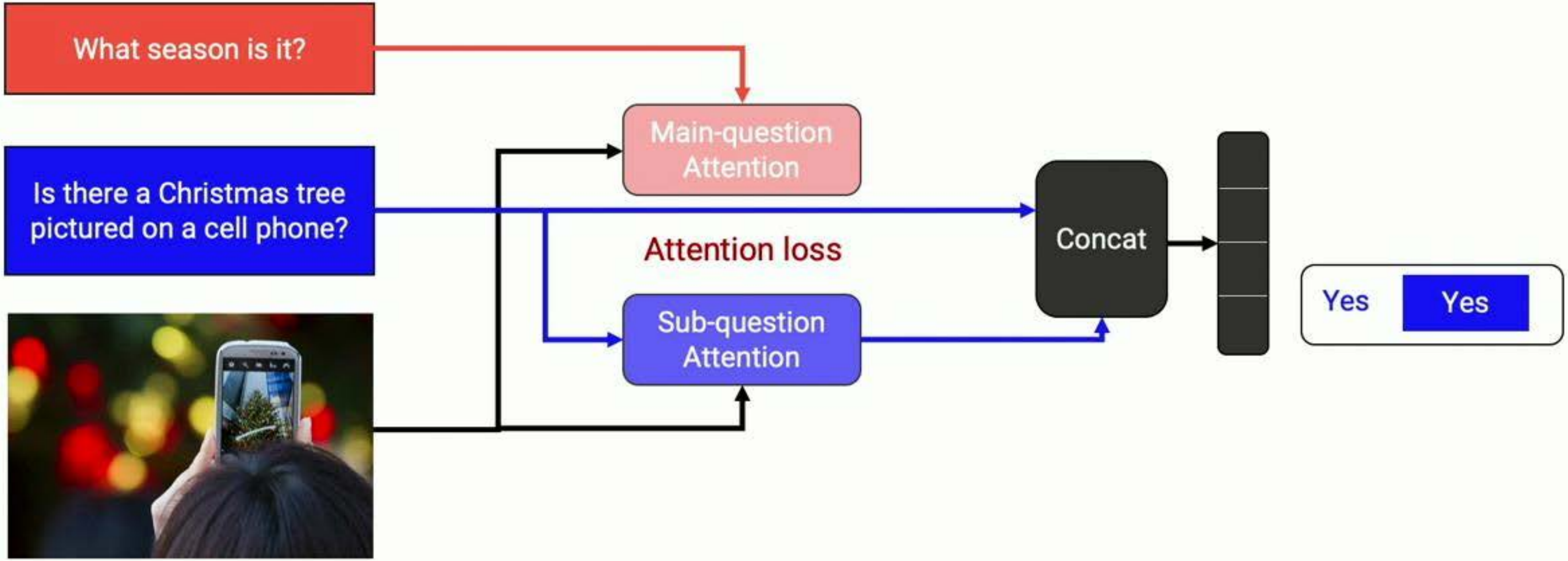
What season is it?

Is there a Christmas tree pictured on a cell phone?

Main-question Attention

Sub-question Attention

Sub-Question Importance-aware Network Tuning (SQuINT)



What season is it?

Is there a Christmas tree pictured on a cell phone?

Main-question Attention

Attention loss

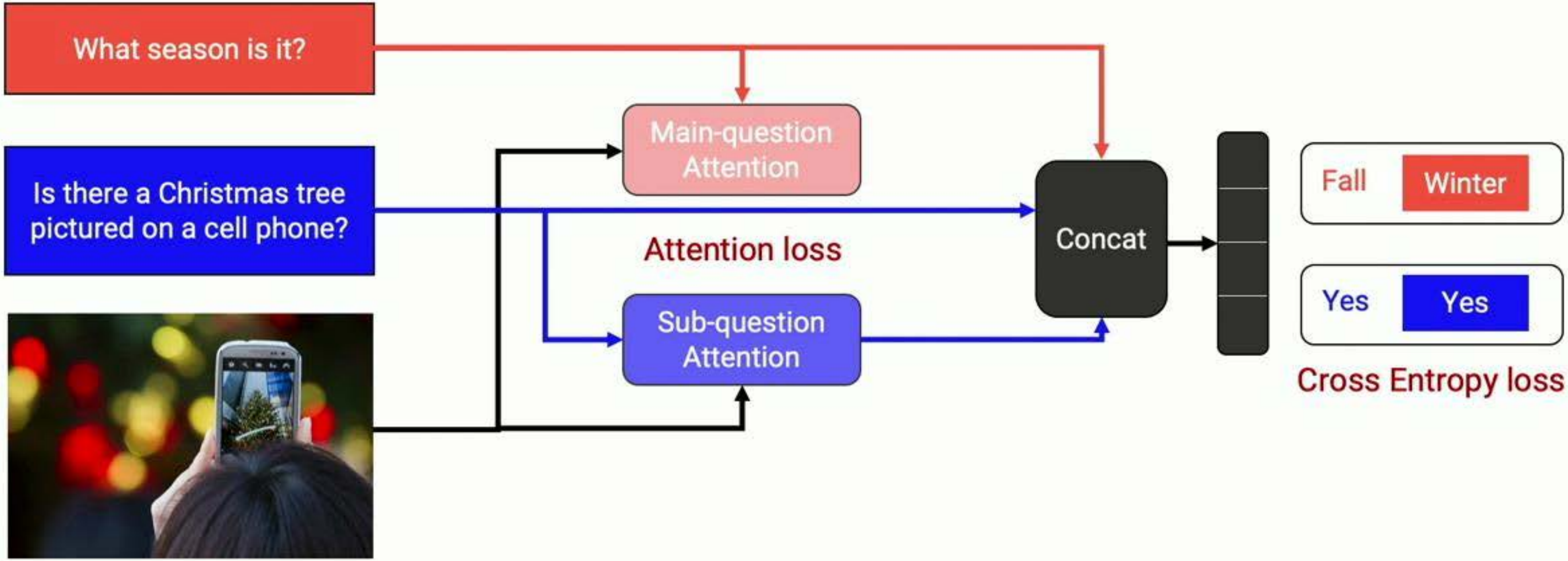
Sub-question Attention

Concat

Yes

Yes

Sub-Question Importance-aware Network Tuning (SQuINT)



What season is it?

Is there a Christmas tree pictured on a cell phone?

Main-question Attention

Attention loss

Sub-question Attention

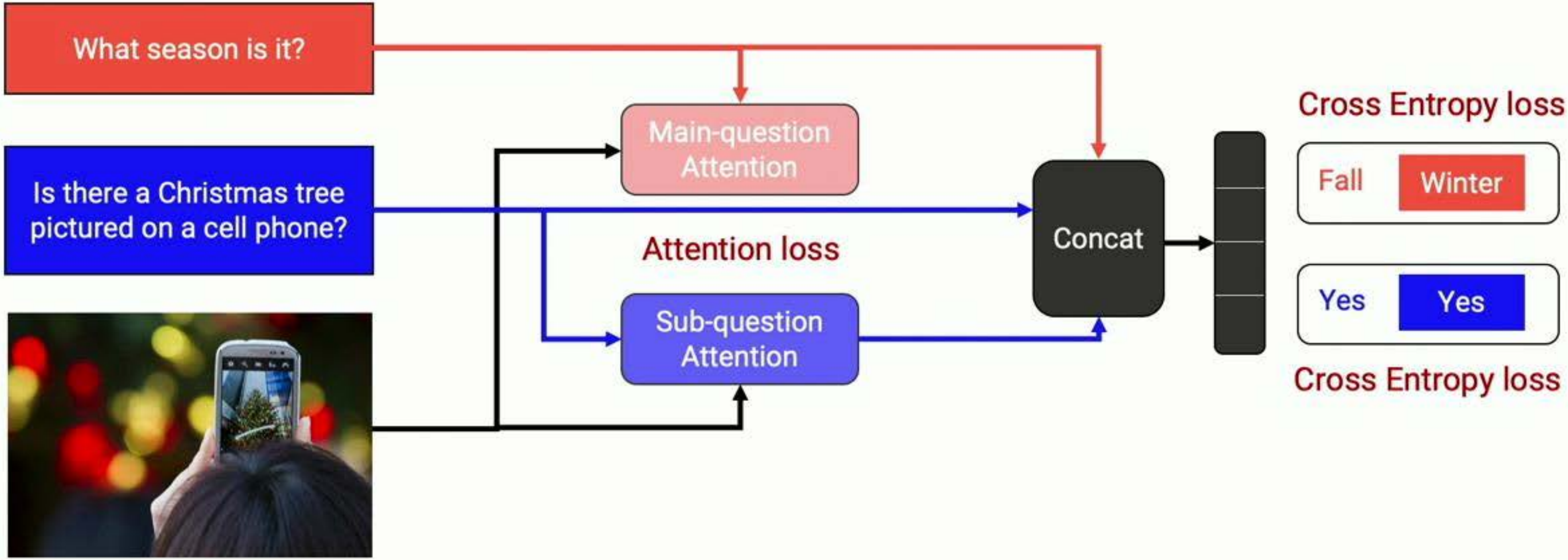
Concat

Fall Winter

Yes Yes

Cross Entropy loss

Sub-Question Importance-aware Network Tuning (SQuINT)



What season is it?

Is there a Christmas tree pictured on a cell phone?

Main-question Attention

Attention loss

Sub-question Attention

Concat

Cross Entropy loss

Fall

Winter

Yes

Yes

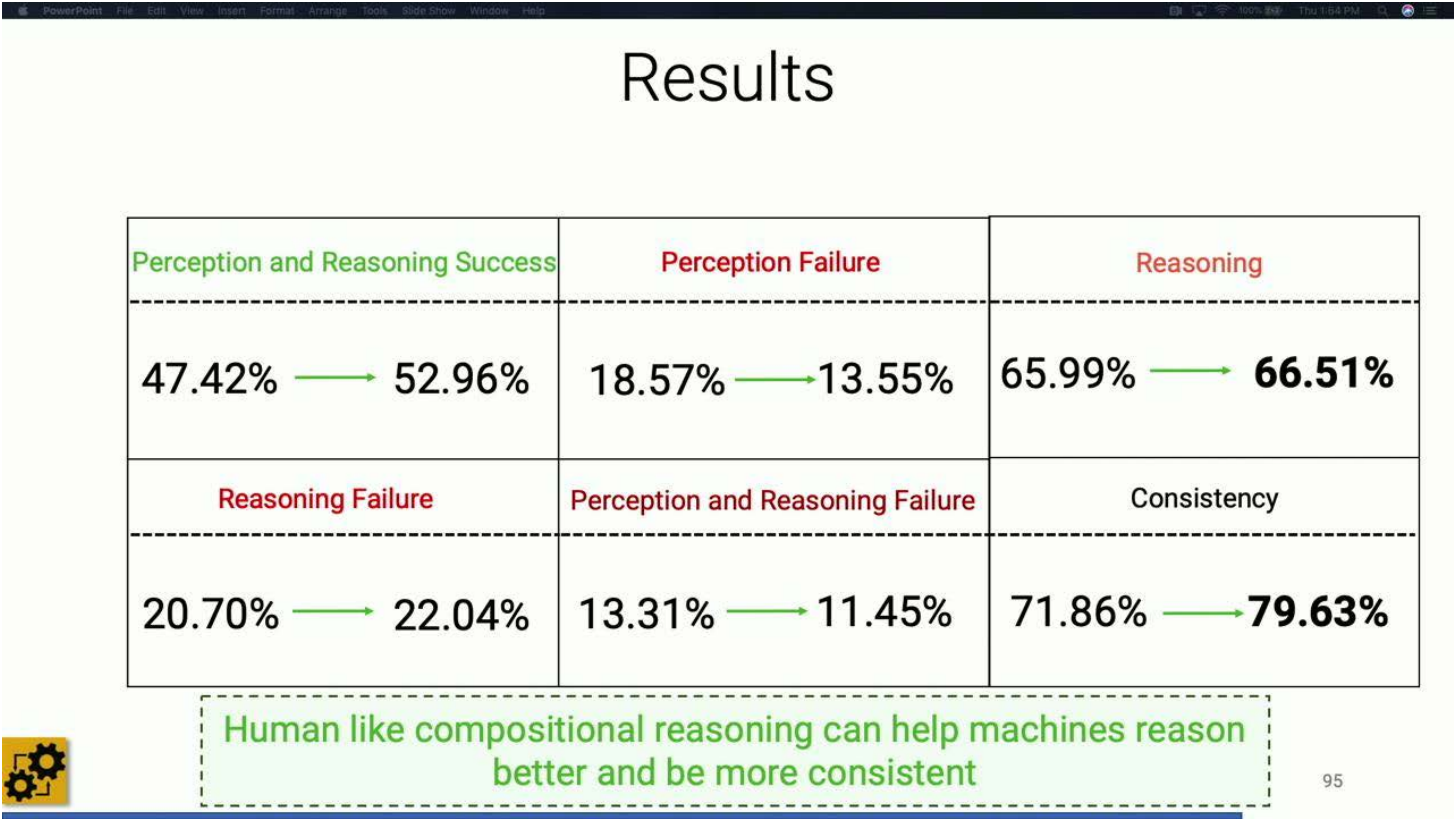
Cross Entropy loss

Results

Perception and Reasoning Success	Perception Failure	Reasoning
Reasoning Failure	Perception and Reasoning Failure	Consistency

Results

Perception and Reasoning Success	Perception Failure	Reasoning
47.42%	18.57%	65.99%
Reasoning Failure	Perception and Reasoning Failure	Consistency
20.70%	13.31%	71.86%



Results

Perception and Reasoning Success	Perception Failure	Reasoning
47.42% → 52.96%	18.57% → 13.55%	65.99% → **66.51%**
Reasoning Failure	Perception and Reasoning Failure	Consistency
20.70% → 22.04%	13.31% → 11.45%	71.86% → **79.63%**

Human like compositional reasoning can help machines reason better and be more consistent

Do SQuINTed models look at right regions?

Do SQuINTed models look at right regions?

Main Question

Is this clock in America? Yes

Do SQuINTed models look at right regions?

Main Question

Is this clock in America? Yes

Baseline

No

Do SQuINTed models look at right regions?

Main Question

Is this clock in America? Yes

Sub Question

Is there an American flag? Yes

Baseline

No

Do SQuINTed models look at right regions?

Main Question

Is this clock in America? Yes

Sub Question

Is there an American flag? Yes

Do SQuINTed models look at right regions?

Main Question

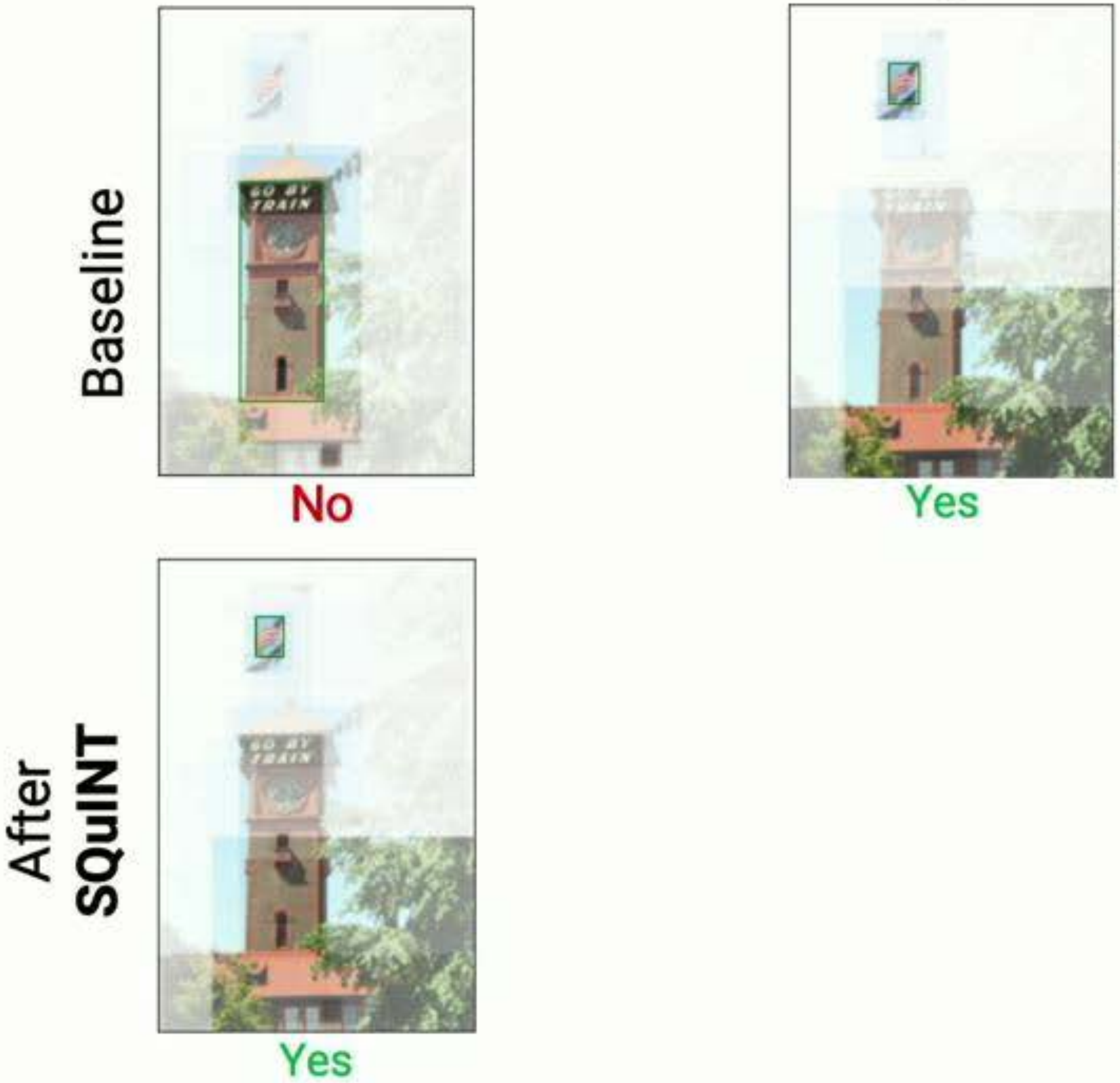
Is this clock in America? Yes

Sub Question

Is there an American flag? Yes

Reasoning Failure

Do SQuINTed models look at right regions?



Reasoning Failure

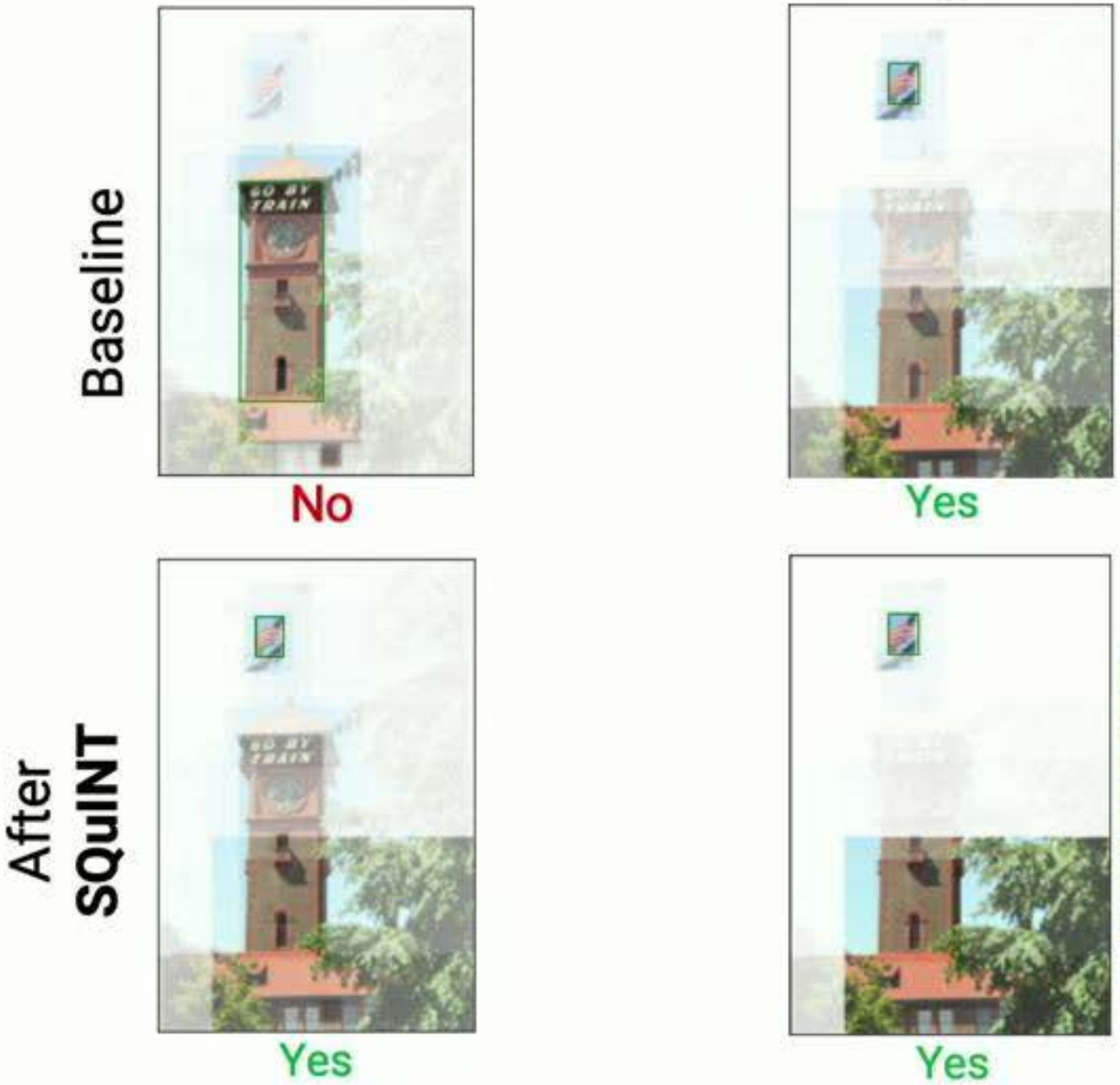
Do SQuINTed models look at right regions?

Main Question

Is this clock in America? Yes

Sub Question

Is there an American flag? Yes



Reasoning Failure

Correcting Reasoning failure through SQuINT

Summary

- New split of VQA dataset (Perception vs Reasoning)

Reason

Enabling human-like compositional reasoning in models through SQuINT

Do SQuINTed models look at right regions?

Main Question

Is this clock in America? Yes

Sub Question

Is there an American flag? Yes

Reasoning Failure

Correcting Reasoning failure through SQuINT

Summary

Reason

Enabling human-like compositional reasoning in models through SQuINT

- New split of VQA dataset (Perception vs Reasoning)
- Introduced a new Sub-VQA dataset
  - to evaluate and enforce compositionality
- SQuINT as a first step towards how human-like compositional reasoning can help improve VQA performance on complex questions

Talk outline

Explain

Explain decisions from deep networks through Grad-CAM (ICCV'17, IJCV'19)

Debias

Leveraging explanations to make models human-like through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What future directions excite me?

Future Work

What future directions excite me?

What are my immediate next steps?

Modality-specific Explanations

- VQA:
  - What questions does the model use when arriving at an answer?

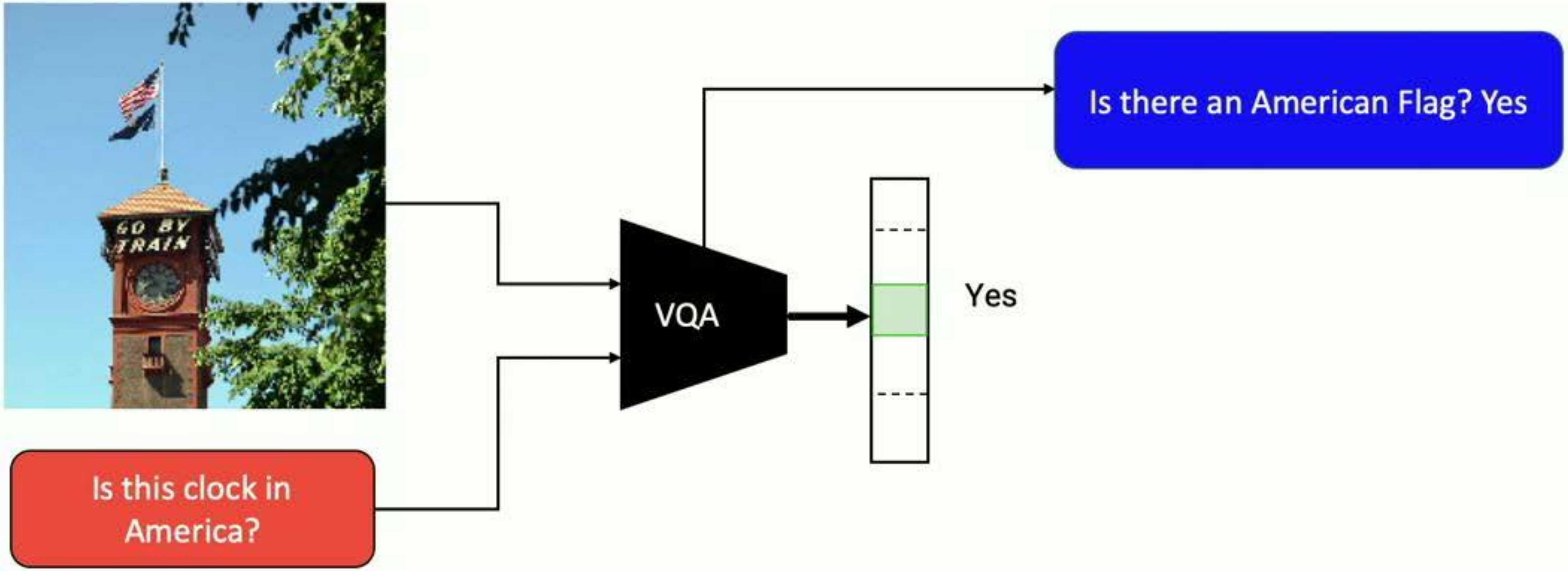
Modality-specific Explanations

- VQA:
  - What questions does the model use when arriving at an answer?

Is this clock in America?

Modality-specific Explanations

- VQA:
  - What questions does the model use when arriving at an answer?



Is there an American Flag? Yes

VQA

Yes

Is this clock in America?

Provide intuitive ways to fix models

Provide intuitive ways to fix models

- VQA:
    - Can fixing the answer to the generated sub-question fix the model?

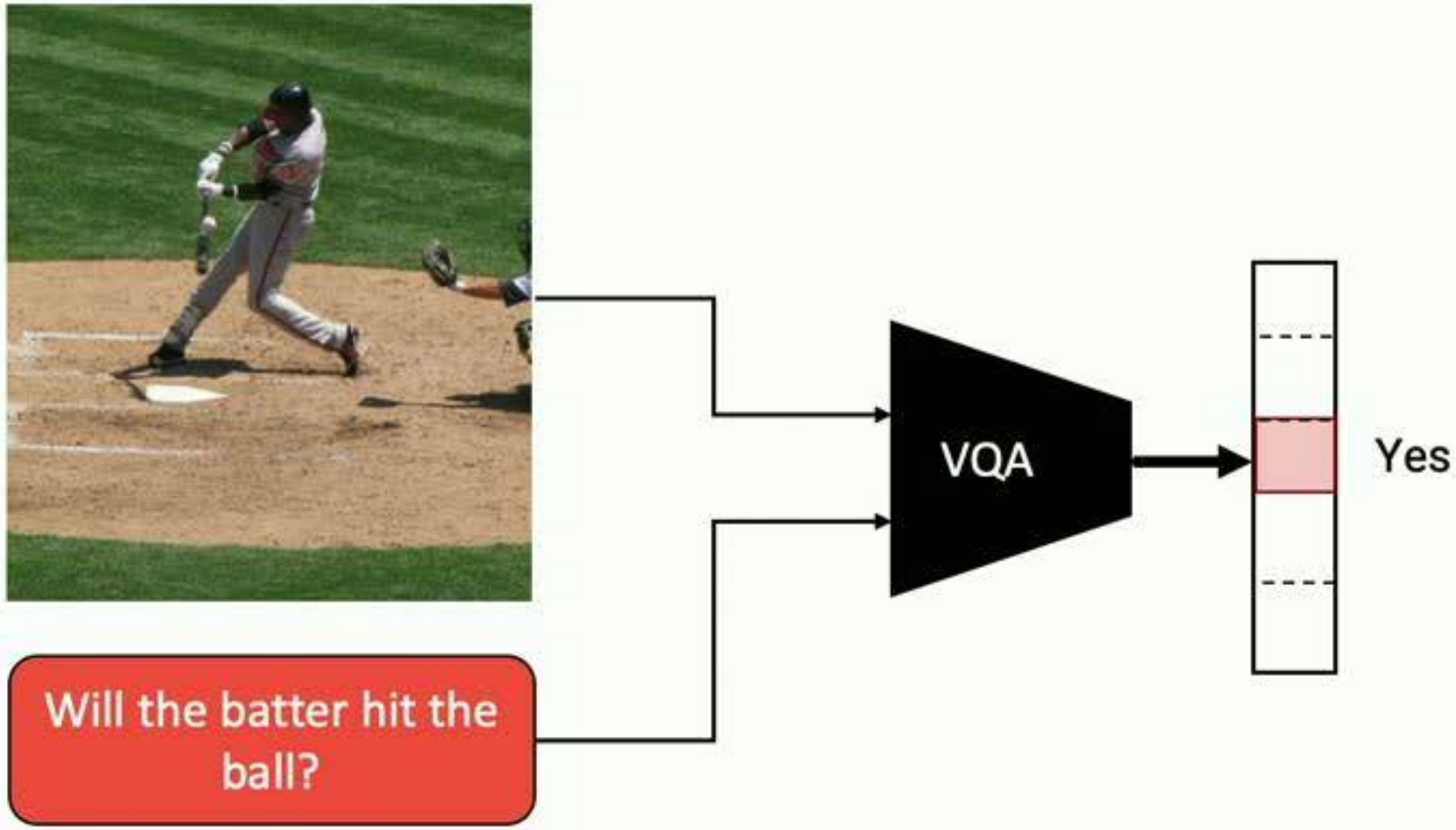
Provide intuitive ways to fix models

- VQA:
  - Can fixing the answer to the generated sub-question fix the model?

Will the batter hit the ball?

Provide intuitive ways to fix models

- VQA:
  - Can fixing the answer to the generated sub-question fix the model?



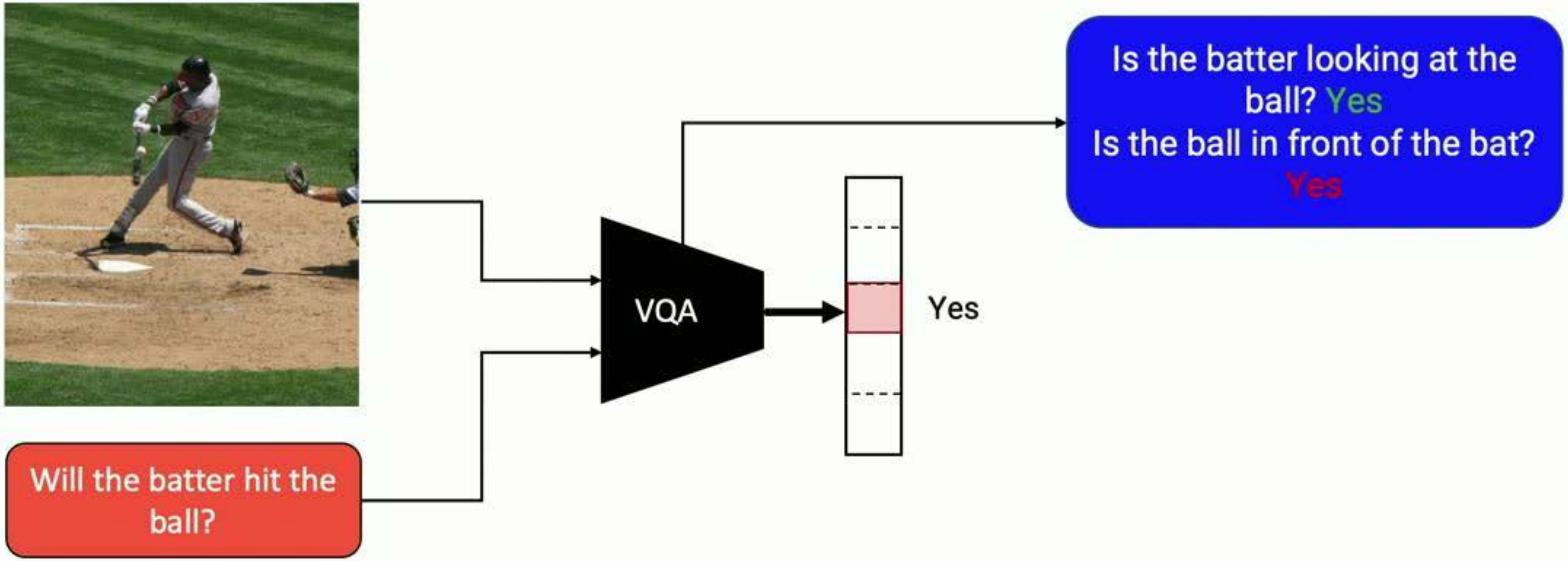
Will the batter hit the ball?

VQA

Yes

Provide intuitive ways to fix models

- VQA:
  - Can fixing the answer to the generated sub-question fix the model?



Will the batter hit the ball?

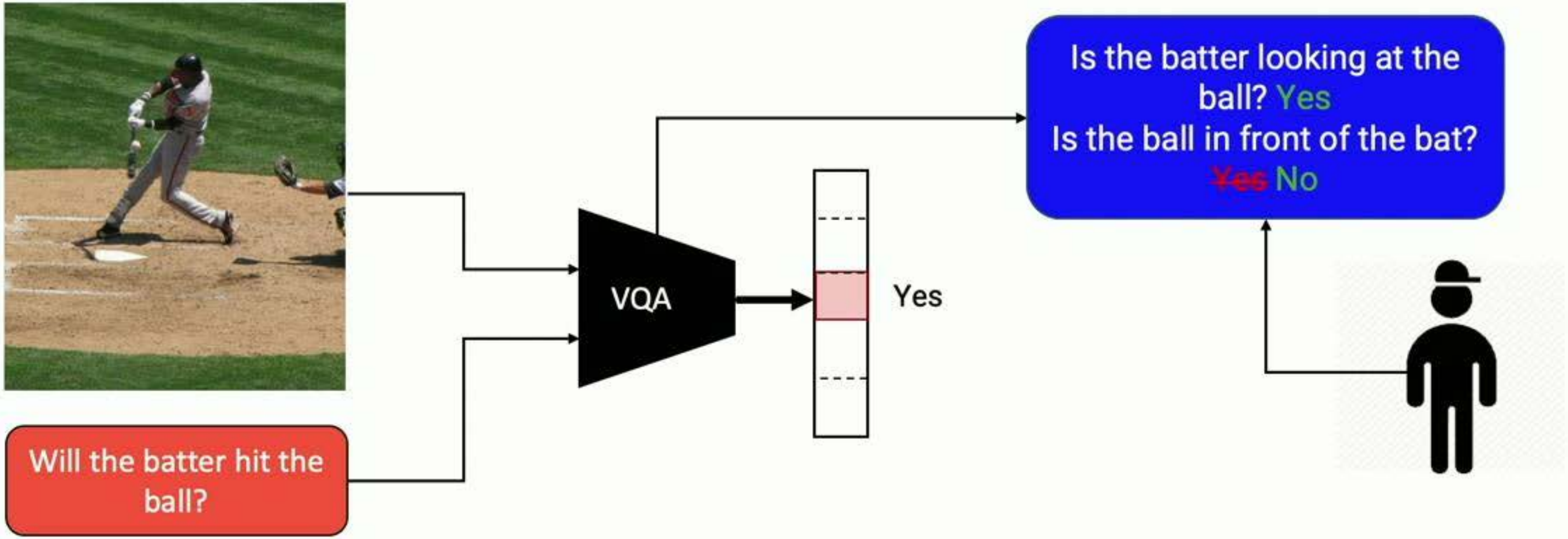
VQA

Yes

Is the batter looking at the ball? Yes
Is the ball in front of the bat? Yes

Provide intuitive ways to fix models

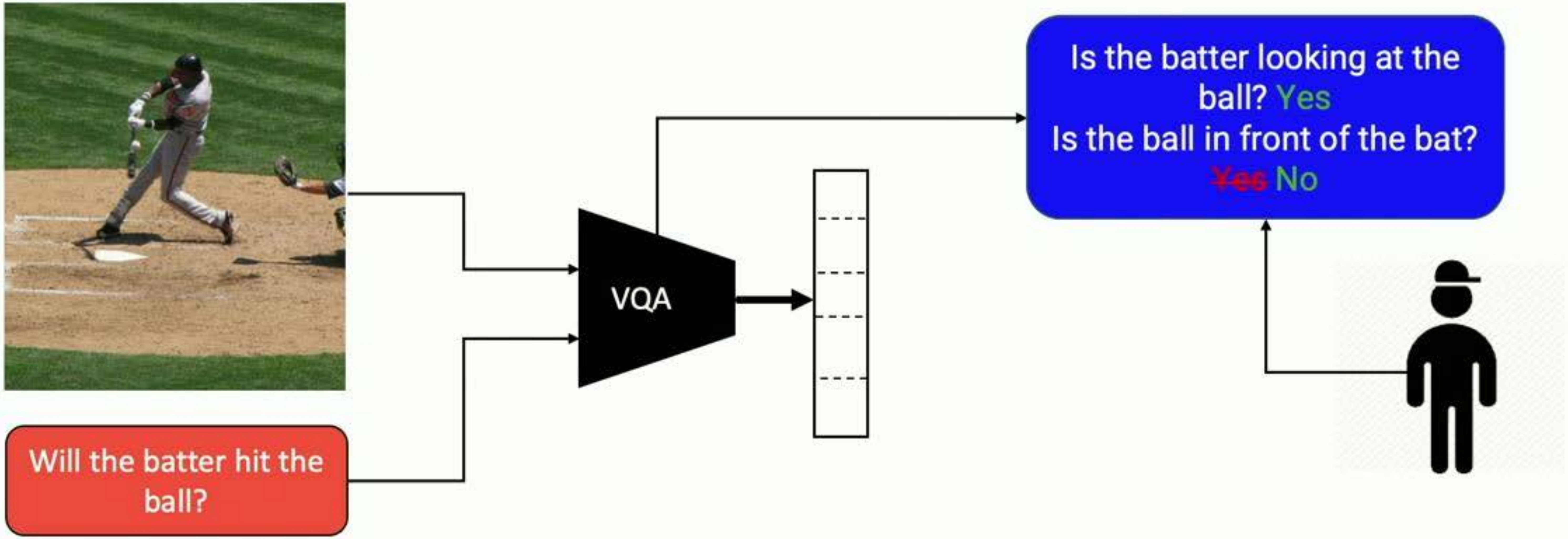
- VQA:
  - Can fixing the answer to the generated sub-question fix the model?



Is the batter looking at the ball? Yes
Is the ball in front of the bat? ~~Yes~~ No

Provide intuitive ways to fix models

- VQA:
  - Can fixing the answer to the generated sub-question fix the model?

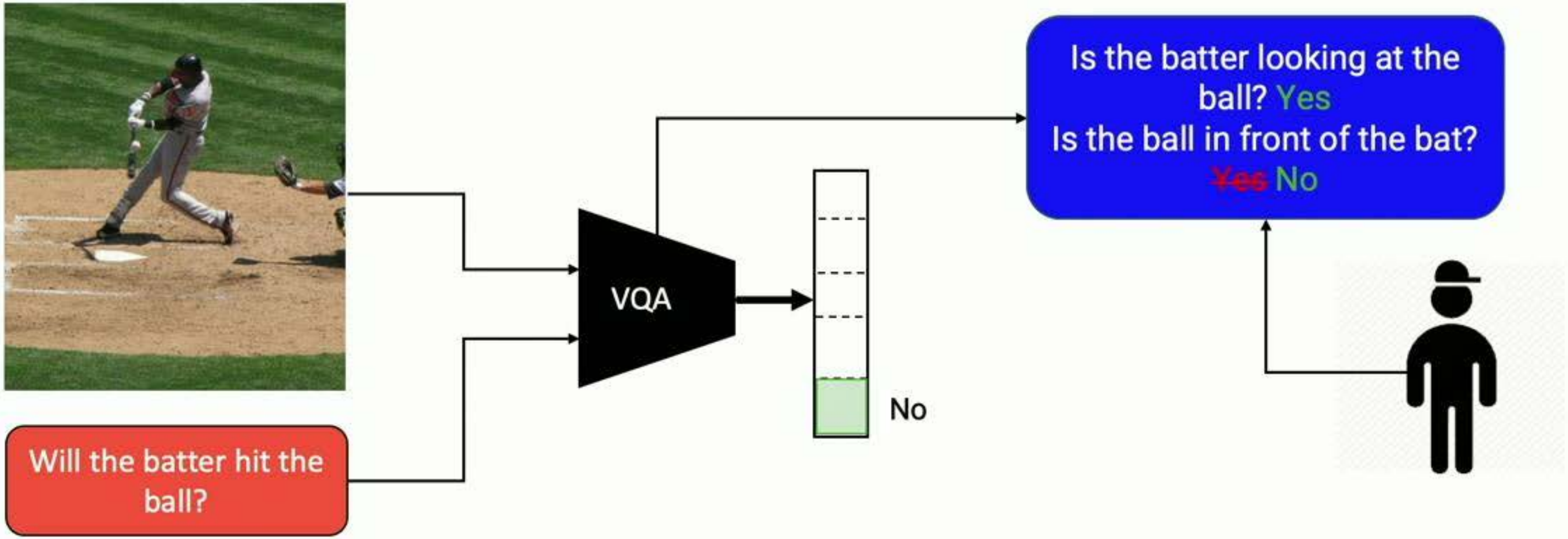


Will the batter hit the ball?

Is the batter looking at the ball? Yes
Is the ball in front of the bat? ~~Yes~~ No

Provide intuitive ways to fix models

- VQA:
  - Can fixing the answer to the generated sub-question fix the model?



Will the batter hit the ball?

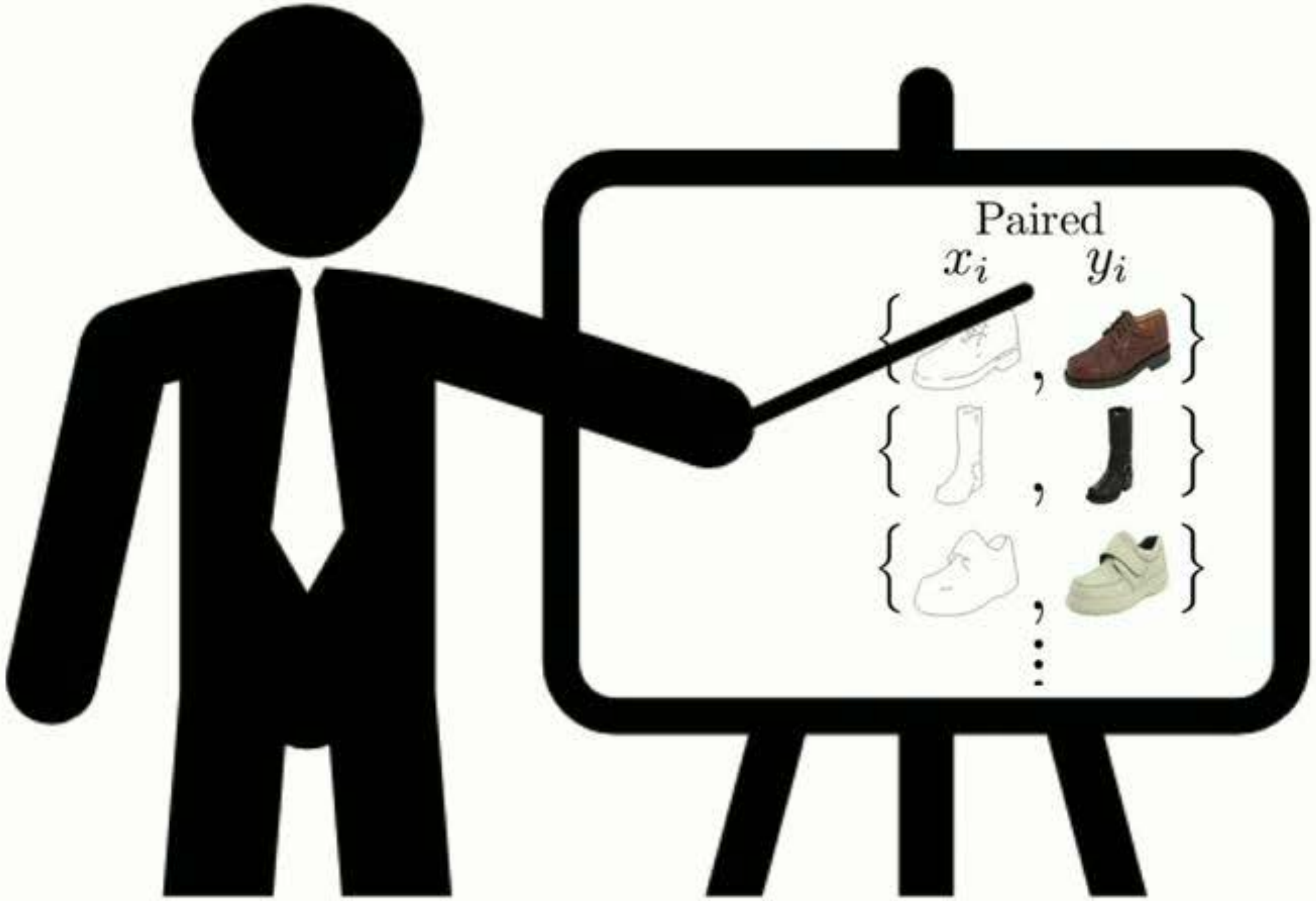
VQA

No

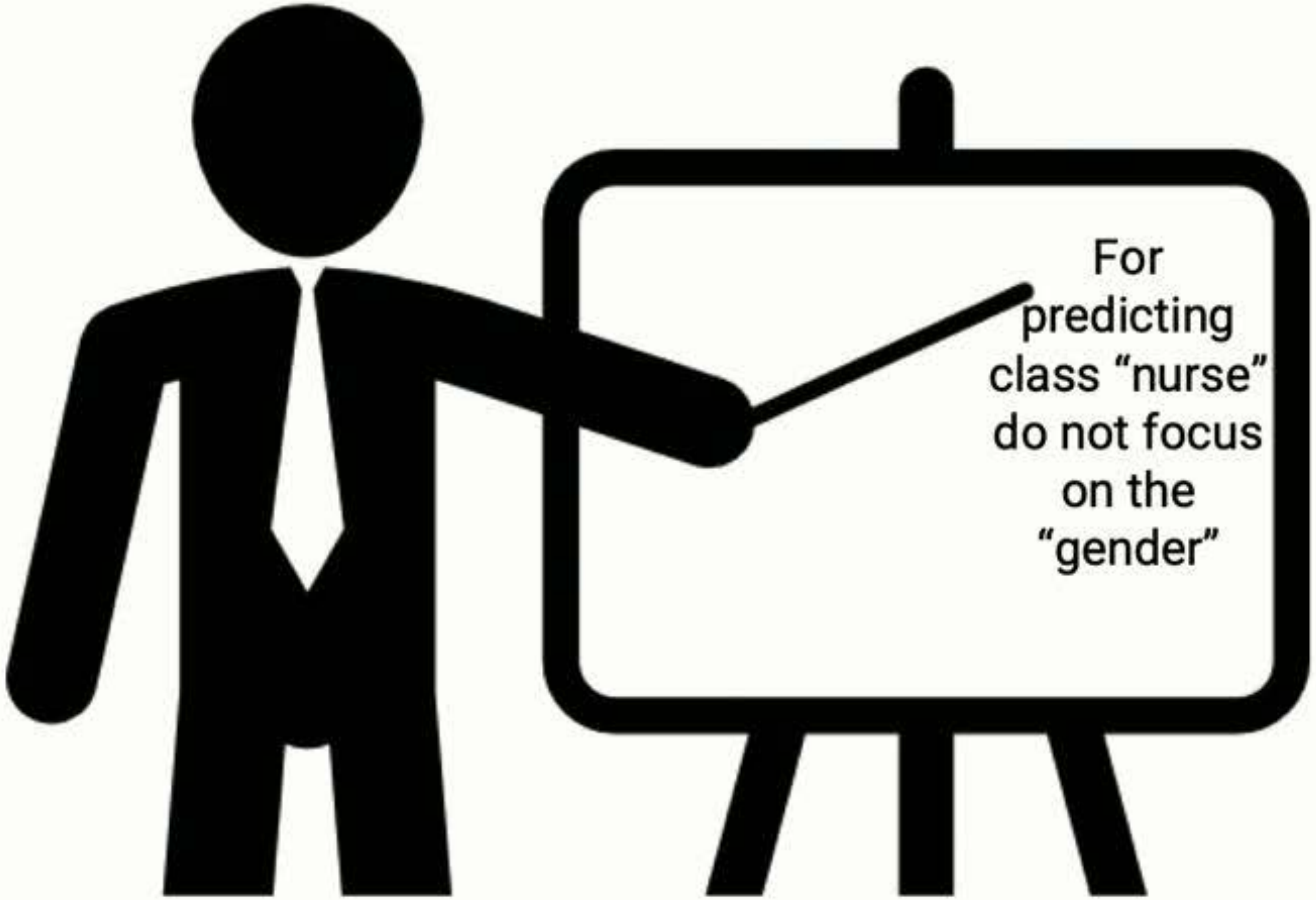
Is the batter looking at the ball? Yes
Is the ball in front of the bat? ~~Yes~~ No

How to incorporate human domain knowledge or rules into deep networks?

Feeding paired data is often an indirect way to teach AI



Convey domain knowledge in natural form



For predicting class “nurse” do not focus on the “gender”

Choose your Neuron: Incorporating Domain Knowledge into Deep Networks through Neuron Importance

Red bellied Woodpecker

A Red Bellied Woodpecker is a small, round bird with a white breast, red crown, and spotted wings

Ramprasaath R. Selvaraju*, Prithvijit Chattopadhyay*, Mohamed Elhoseiny, Tilak Sharma, Dhruv Batra, Devi Parikh, Stefan Lee, " **Choose your Neuron: Incorporating Domain Knowledge into Deep Networks through Neuron Importance**", ECCV'18

Choose your Neuron: Incorporating Domain Knowledge into Deep Networks through Neuron Importance

Red bellied Woodpecker

A Red Bellied Woodpecker is a small, round bird with a white breast, red crown, and spotted wings

Use Grad-CAM as a medium to incorporate human domain knowledge to extend a classifier to detect new classes

Ramprasaath R. Selvaraju*, Prithvijit Chattopadhyay*, Mohamed Elhoseiny, Tilak Sharma, Dhruv Batra, Devi Parikh, Stefan Lee, " **Choose your Neuron: Incorporating Domain Knowledge into Deep Networks through Neuron Importance",** ECCV'18

How will interpretability play a role in the future of AI?

Future Work

What future directions excite me?

Interpretability in different stages of AI evolution

- AI < Human
  - e.g. VQA
  - Goal:
    - Identify failure modes
    - Help researchers focus their efforts on specific modules
- AI ~ Human (ready to be deployed)
  - e.g. Image classification trained on sufficient data
  - Goal:
    - Help establish appropriate trust and confidence in users
- AI > Human
  - e.g. AlphaGo in the game of Go
  - Goal:
    - Machine teaching a human about how to make better decisions

Future where humans can specify what models should be doing

Explaining Model Decisions and Correcting them via Human Feedback

Explain

Explain decisions from deep networks through Grad-CAM (ICCV' 17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

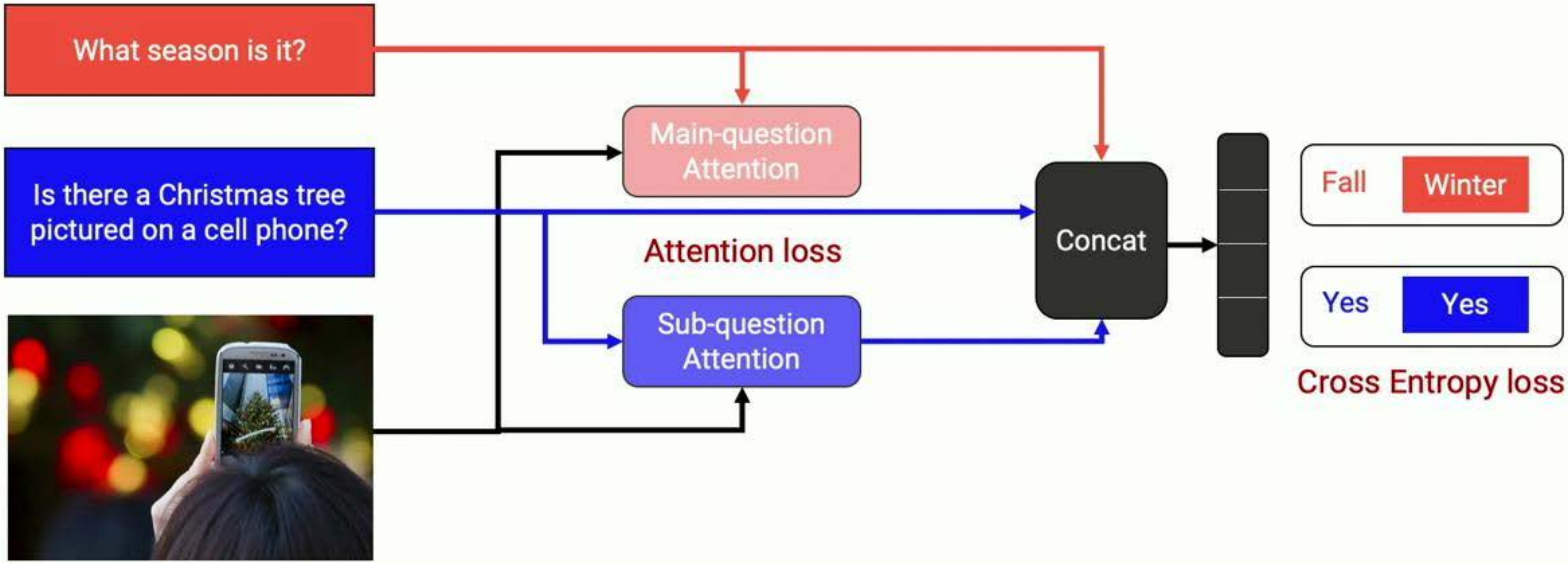
Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What future directions excite me?

Thank you

Sub-Question Importance-aware Network Tuning (SQuINT)



What season is it?

Is there a Christmas tree pictured on a cell phone?

Main-question Attention

Attention loss

Sub-question Attention

Concat

Fall

Winter

Yes

Yes

Cross Entropy loss

Explaining Model Decisions and Correcting them via Human Feedback

Explain

Explain decisions from deep networks through Grad-CAM (ICCV' 17, IJCV'19)

Debias

Leveraging explanations to unbias models through HINT (ICCV'19)

Reason

Enabling human-like compositional reasoning in models through SQuINT (Under Review)

Future Work

What future directions excite me?

Thank you