Explaining Model Decisions and Fixing them via Focused Feedback

Ramprasaath R. Selvaraju

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

Plan for Today

- Visualizing the decision-making process of Deep networks
  - Approaches for Visual Explanations
  - Desirable properties of Explanations
  - Insights from explanations
- How can visual explanations help improve models?
  - Overcoming gender-bias
  - Transferring domain knowledge through language
  - Improving localization
  - Transferring knowledge across domains
  - Improving out of distribution generalization
  - Making models reason compositionally as humans
  - Debiasing Self-supervised representation learning
  - Reducing catastrophic forgetting

Approaches for visual explanations

Gradient-based methods

- Backpropagation [Simonyan *et al.*, 2013]
- Guided Backpropagation [Springenberg *et al.*, 2014]
- Layer-wise Relevance Propagation [Bach *et al.*, 2015]
- Integrated Gradients [Sundararajan *et al.*, 2017]
- Grad-CAM [Selvaraju *et al.*, 2017]

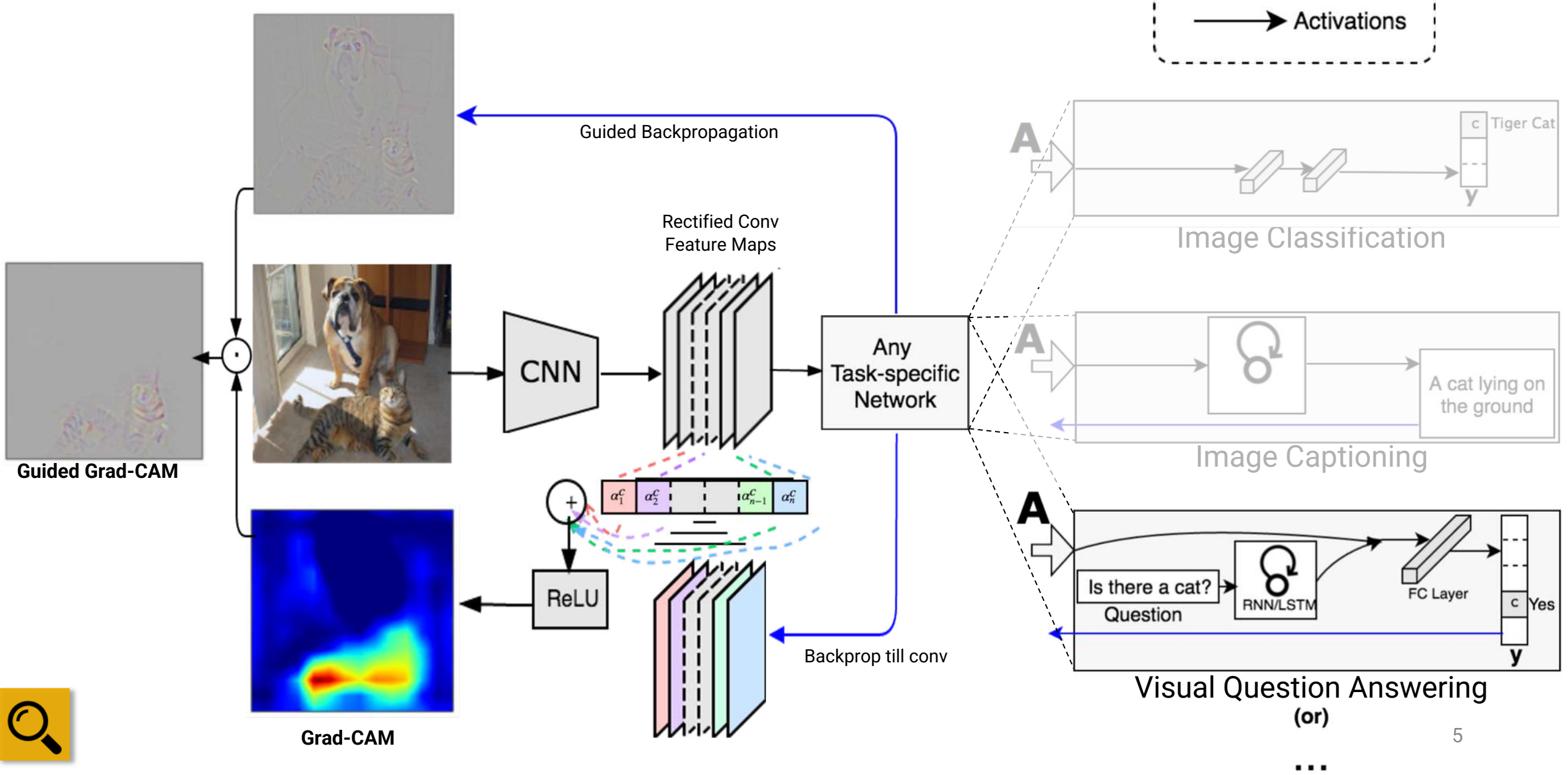
Simplifying model architectures

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

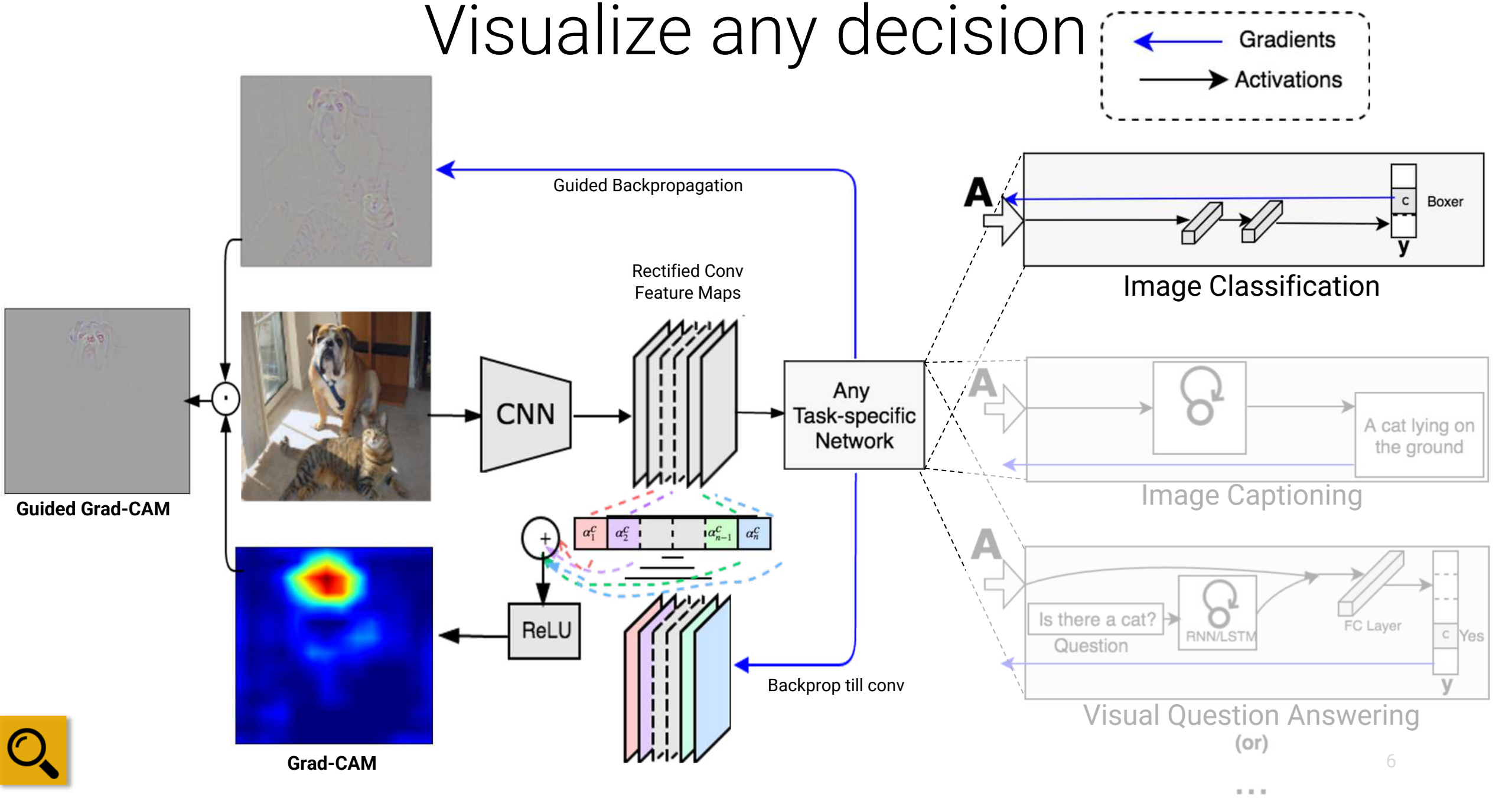
Black-box approaches

- LIME [Ribeiro *et al.*, 2016]
- SHAP [Lundberg *et al.*, 2017]

Guided Grad-CAM



Visualize any decision



Desirable properties of Visual Explanations

- Faithful
  - How faithful is the explanation to the underlying model?
- Interpretable
  - How easy/interpretable is the explanation to the human?
- Inexpensive
  - What is the cost of getting an explanation for a particular decision?
- Maintain model performance
  - Does the explanation require modifying the model?

Insights from Explanations

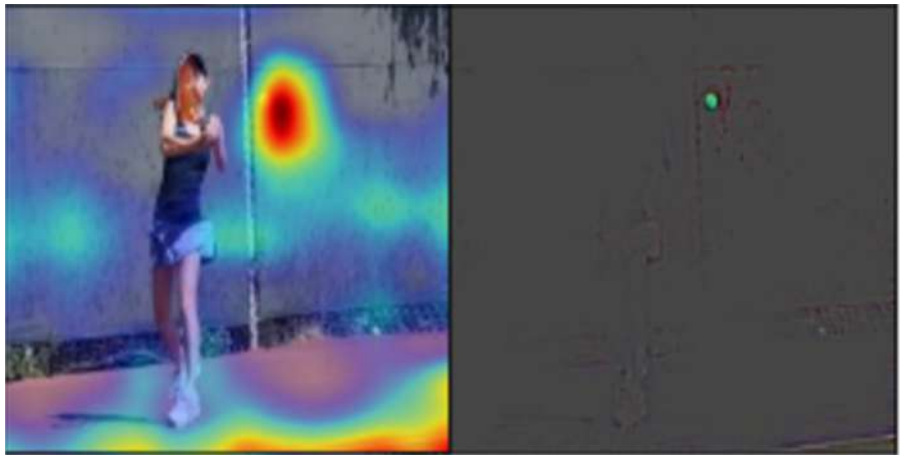
Visualizing Image Captioning models

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?



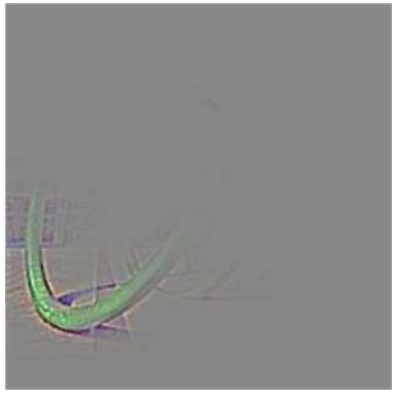
Even simple non-attention-based CNN+LSTM models attend to appropriate regions

Analyzing Failure modes

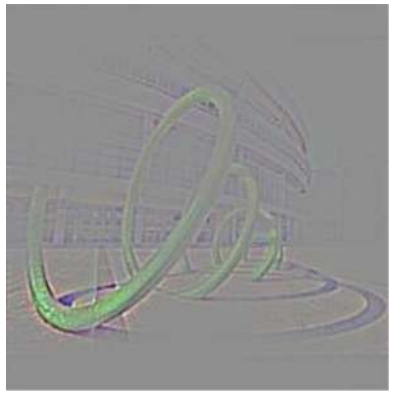
Predicted: *Car mirror*

Ground-truth: *Volcano*

Predicted: *Vine snake*



Ground-truth: *coil*

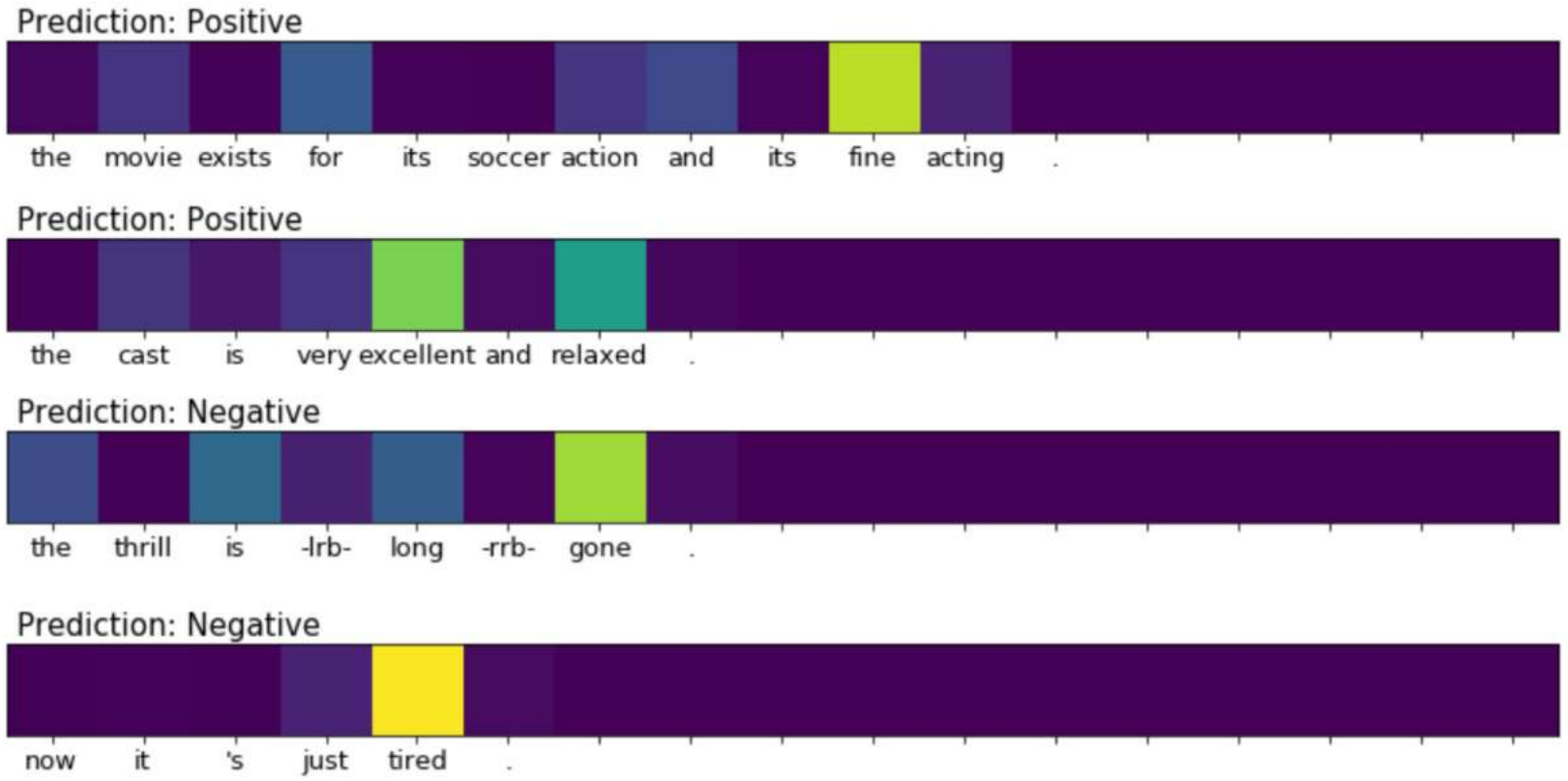


Even unreasonable predictions sometimes have reasonable explanations

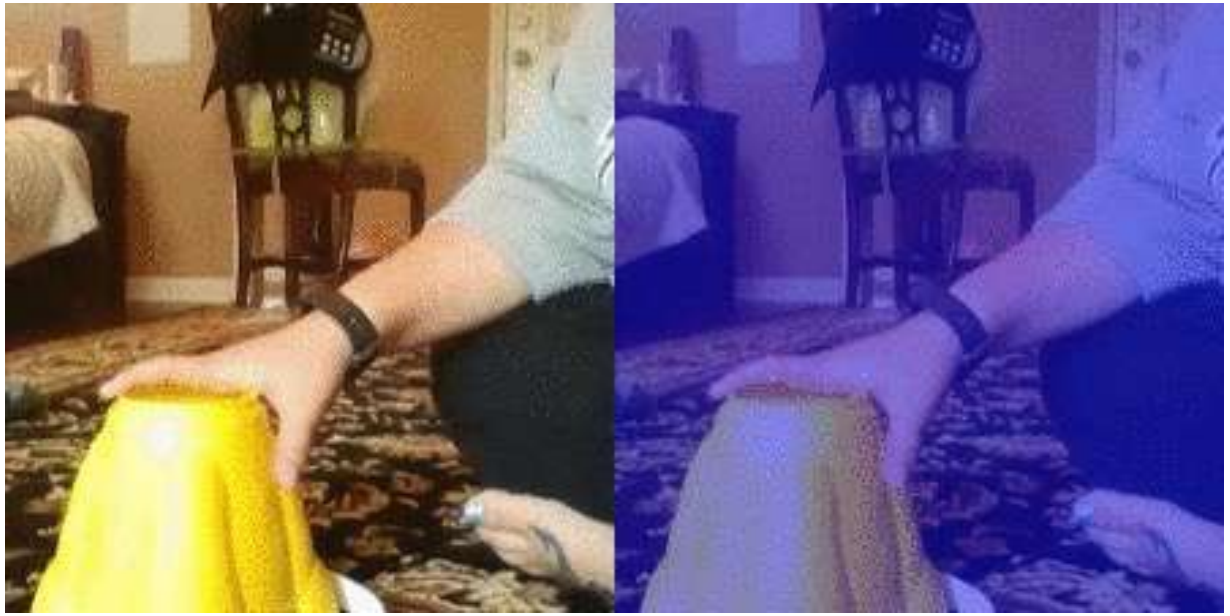
Demo

gradcam.cloudcv.org

Grad-CAM for Text



Grad-CAM for Videos

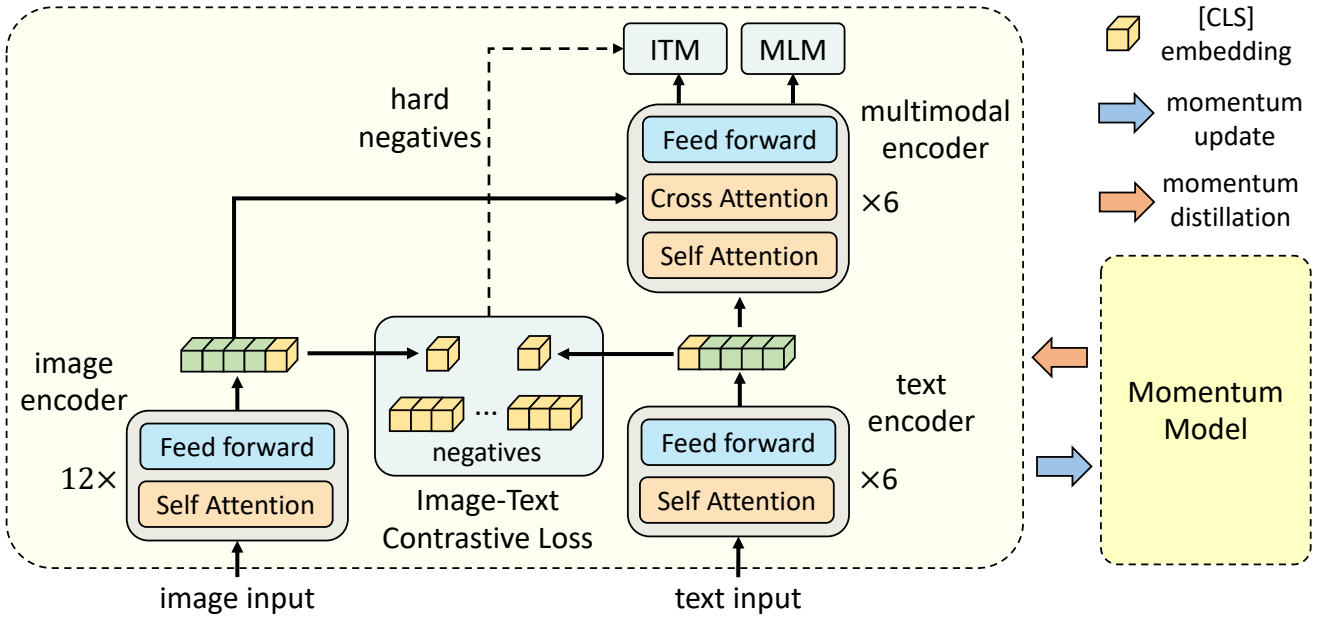


Unrolling [something]

Pushing [something]

Extensions to Multi-modal Transformer based Architectures

- Introducing a new vision and language pretraining framework ALBEF based on vision transformers and BERT Multimodal encoders.



Grad-CAM for multi-modal transformers

“a little girl holding a kitten next to a blue fence”


“girl”


“holding”


“kitten”


“next”


“blue”

Q: is this rice noodle soup?
A: yes

Q: what is to the right of the soup? A: chopsticks

Q: what is the man doing in the street? A: walking

Q: what does the truck on the left sell? A: ice cream

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 trust and confidence in users
- AI > Human
  - e.g., AlphaGo in the game of Go
  - Goal:
    - Machine teaching humans to make better decisions

How can explanations help improve models?

Using insights from explanations to overcome gender-bias

Women also Snowboard: Overcoming Bias in Image Captioning Models

ECCV 2018

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach

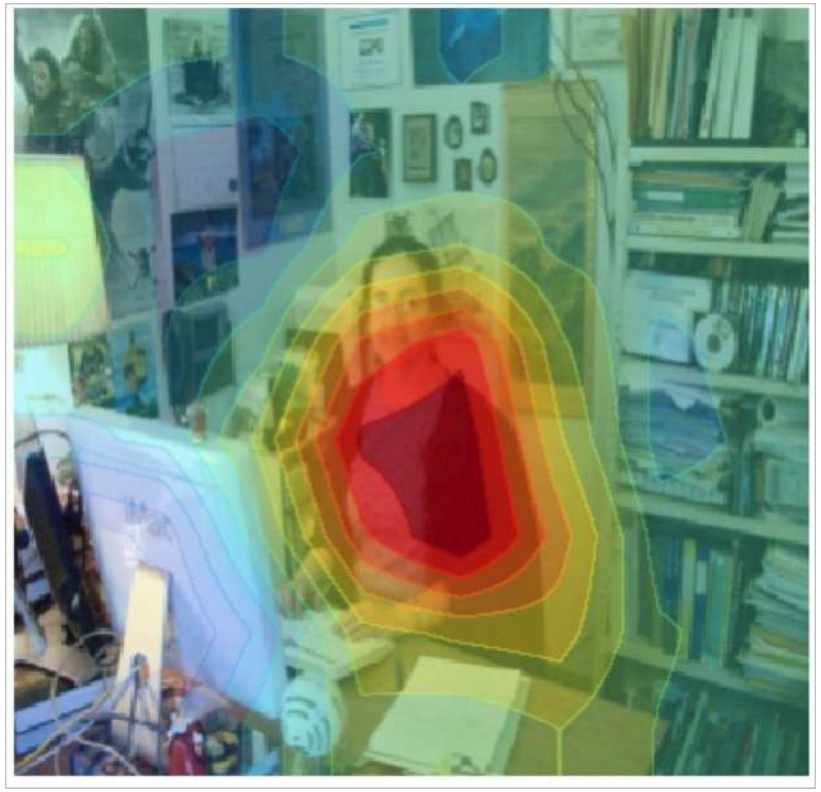
Wrong Prediction, Wrong Evidence

Baseline: *A **man** sitting at a desk with a laptop computer.*

Explanation from: Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. ICCV 2017.

Right Prediction, Right Evidence

Baseline: *A **man** sitting at a desk with a laptop computer.*



Correct Model: *A **woman** sitting in front of a laptop computer.*

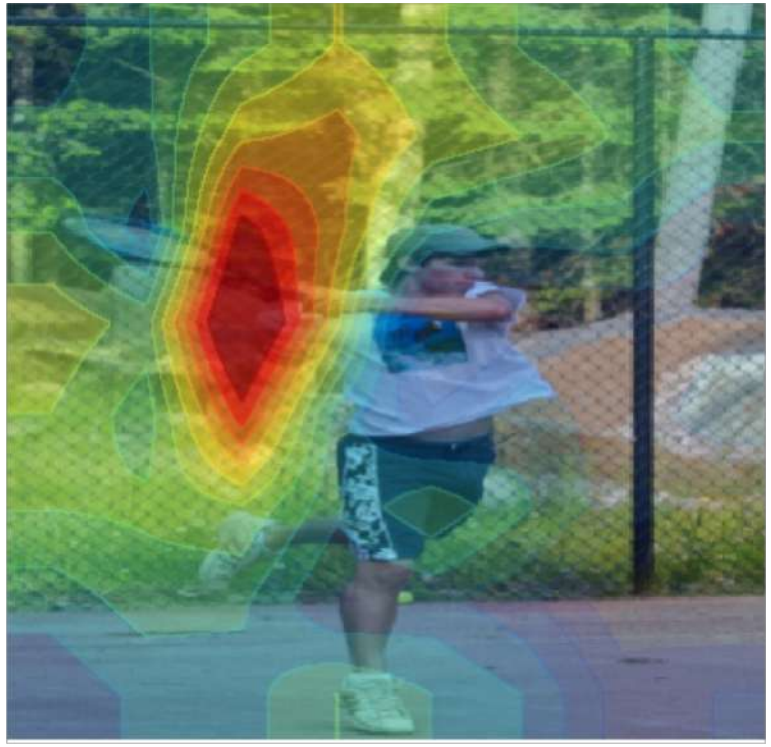
Right for the Right Reasons

Baseline: *A **man** sitting at a desk with a laptop computer.*

Right for the Right Reasons.

Correct Model: *A **woman** sitting in front of a laptop computer.*

Right for the Wrong Reasons



Baseline: *A **man** holding a tennis racquet on a tennis court.*

Right for the Wrong Reasons

Right for the Wrong Reasons.

Baseline: *A **man** holding a tennis racquet on a tennis court.*

Right for the Right Reasons

Right for the Wrong Reasons.

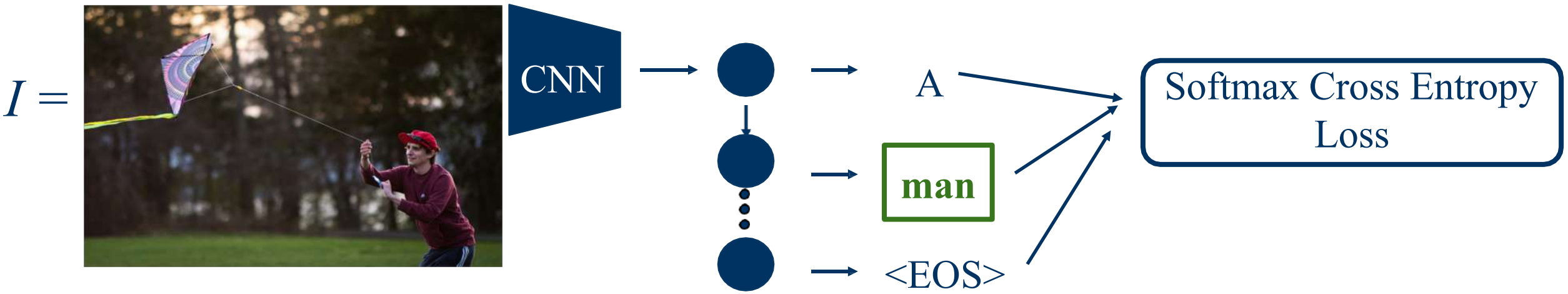
Baseline: *A **man** holding a tennis racquet on a tennis court.*

Right for the Right Reasons.

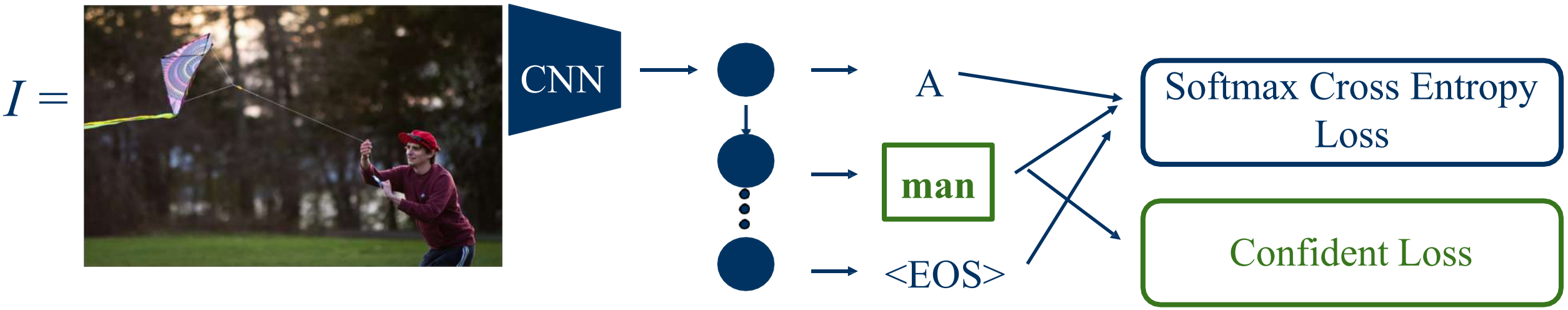
Correct Model: *A **man** holding a tennis racquet on a tennis court.*

Equalizer

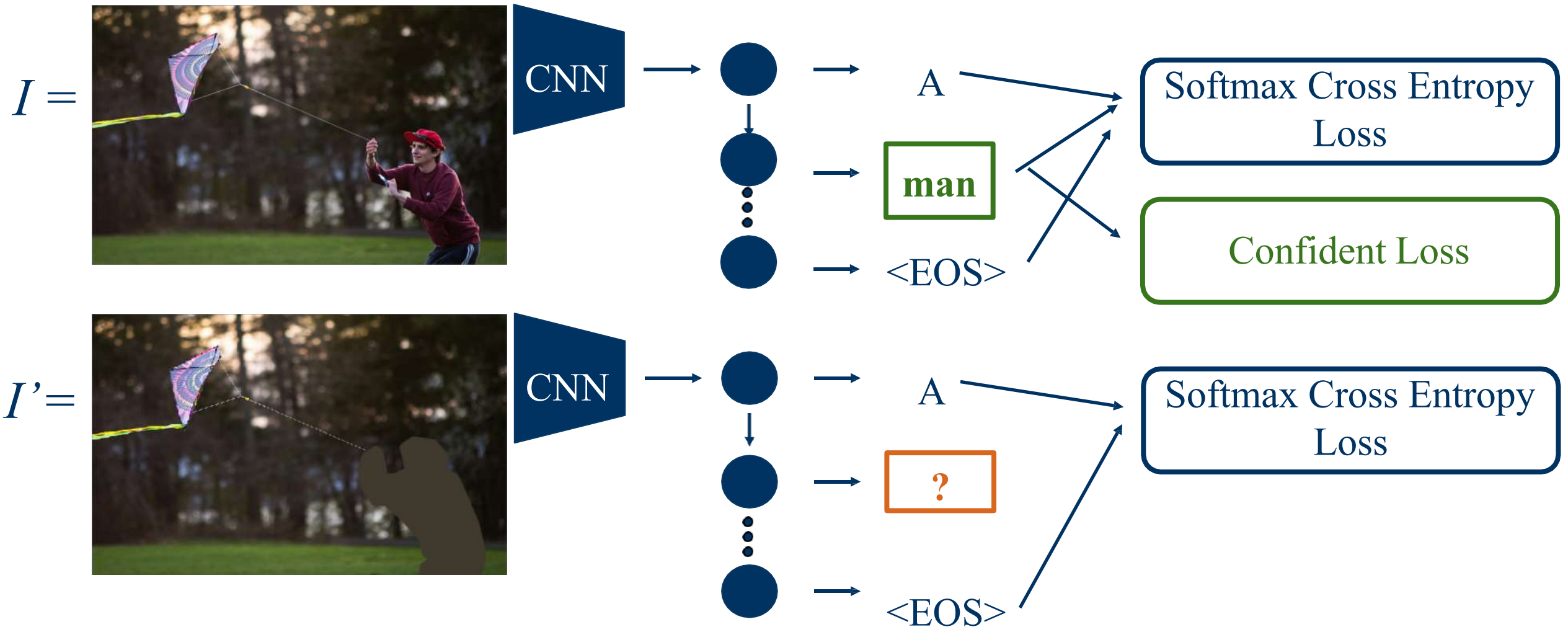
Equalizer



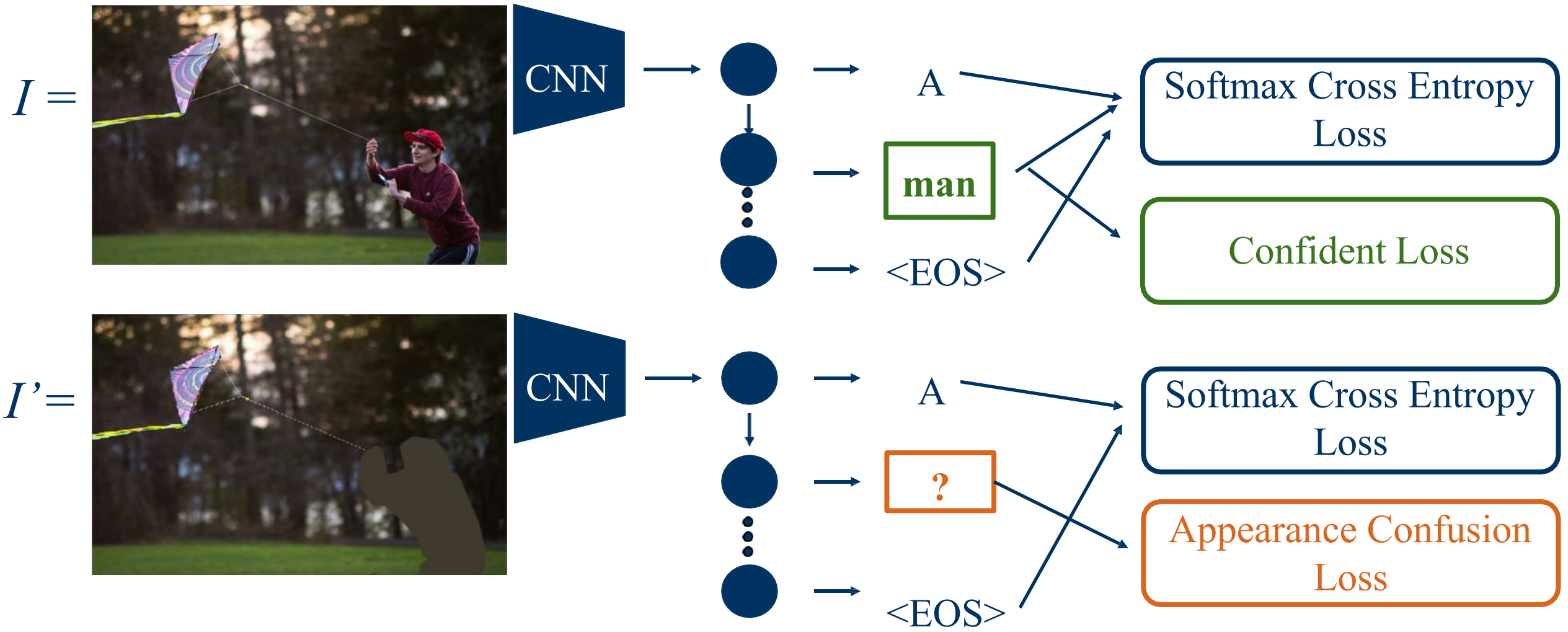
Equalizer



Equalizer


I =
CNN
A
man
<EOS>
Softmax Cross Entropy Loss
Confident Loss
I'=
CNN
A
?
<EOS>
Softmax Cross Entropy Loss

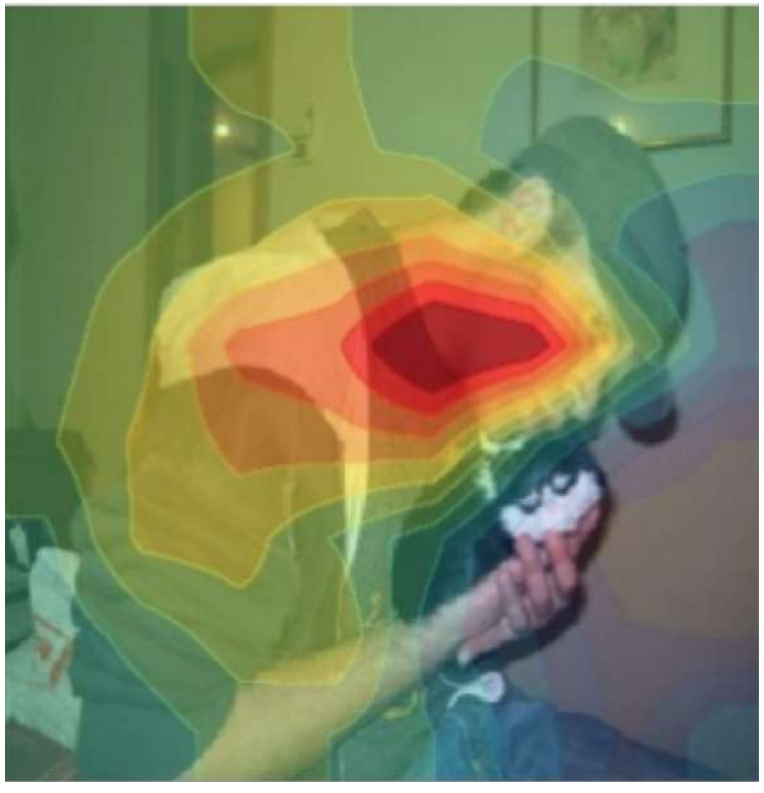
Equalizer


I =
CNN
A
man
<EOS>
Softmax Cross Entropy Loss
Confident Loss
I'=
CNN
A
?
<EOS>
Softmax Cross Entropy Loss
Appearance Confusion Loss

Baseline

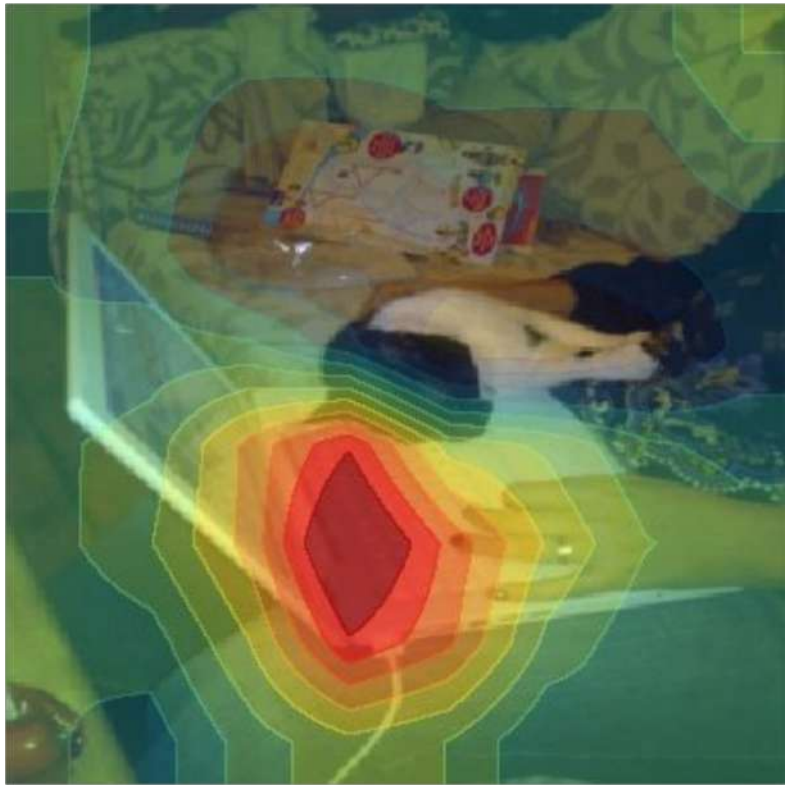
A ***woman*** holding a cat in her arms.

Equalizer



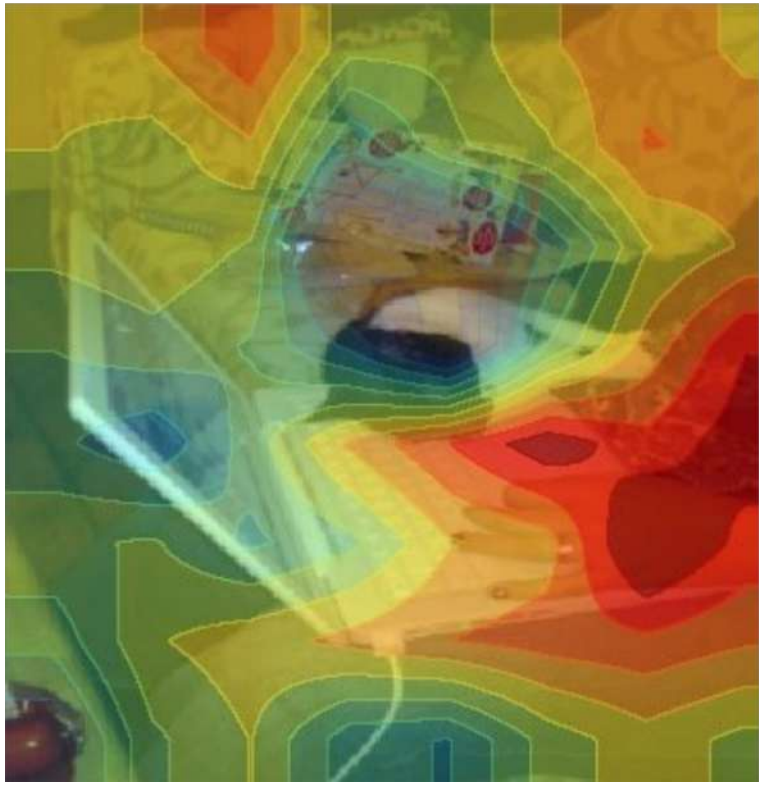
A ***man*** is holding a black and white cat.

Baseline



A ***man*** sitting on a couch with a laptop computer.

Equalizer



A ***person*** laying on a bed with a laptop.

Explanations to facilitate knowledge transfer between humans and AI

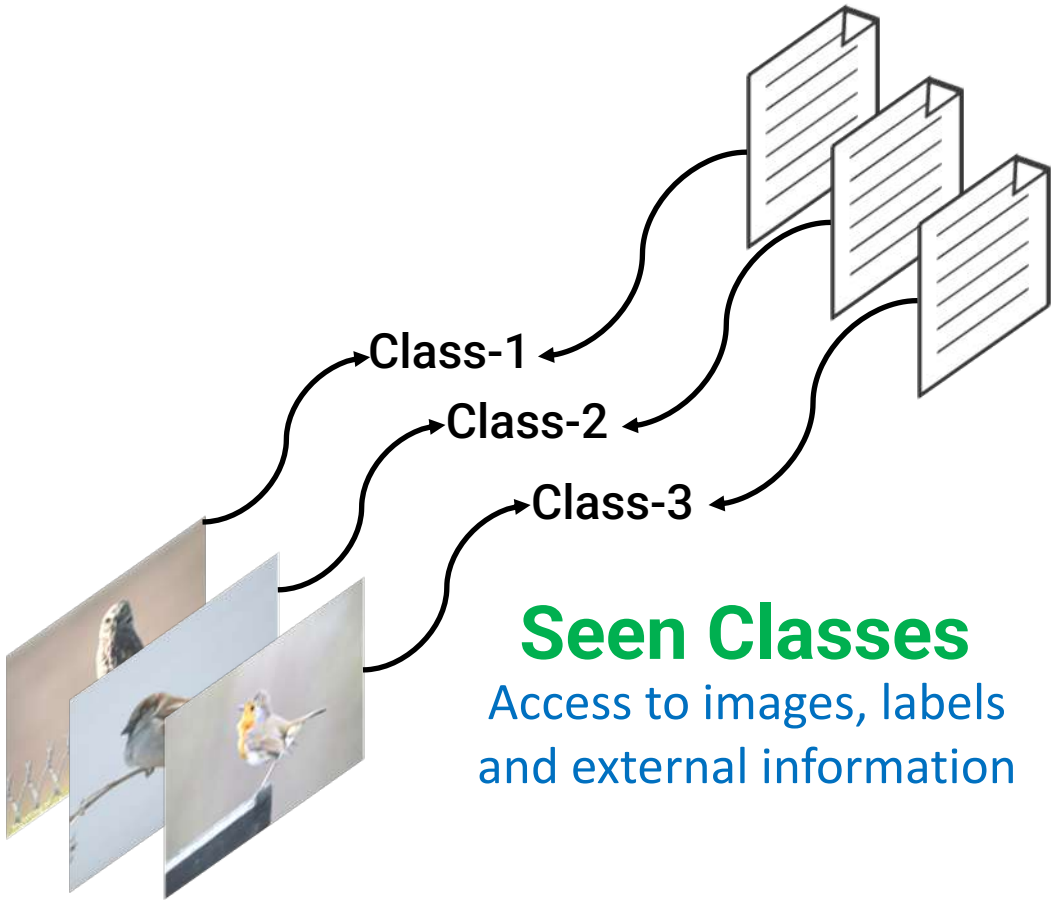
Choose Your Neuron: Incorporating Domain Knowledge through Neuron Importance

ECCV 2018

Ramprasaath R. Selvaraju, Prithvijit Chattopadhyay, Mohamed Elhoseiny, Tilak Sharma, Dhruv Batra, Devi Parikh, Stefan Lee.

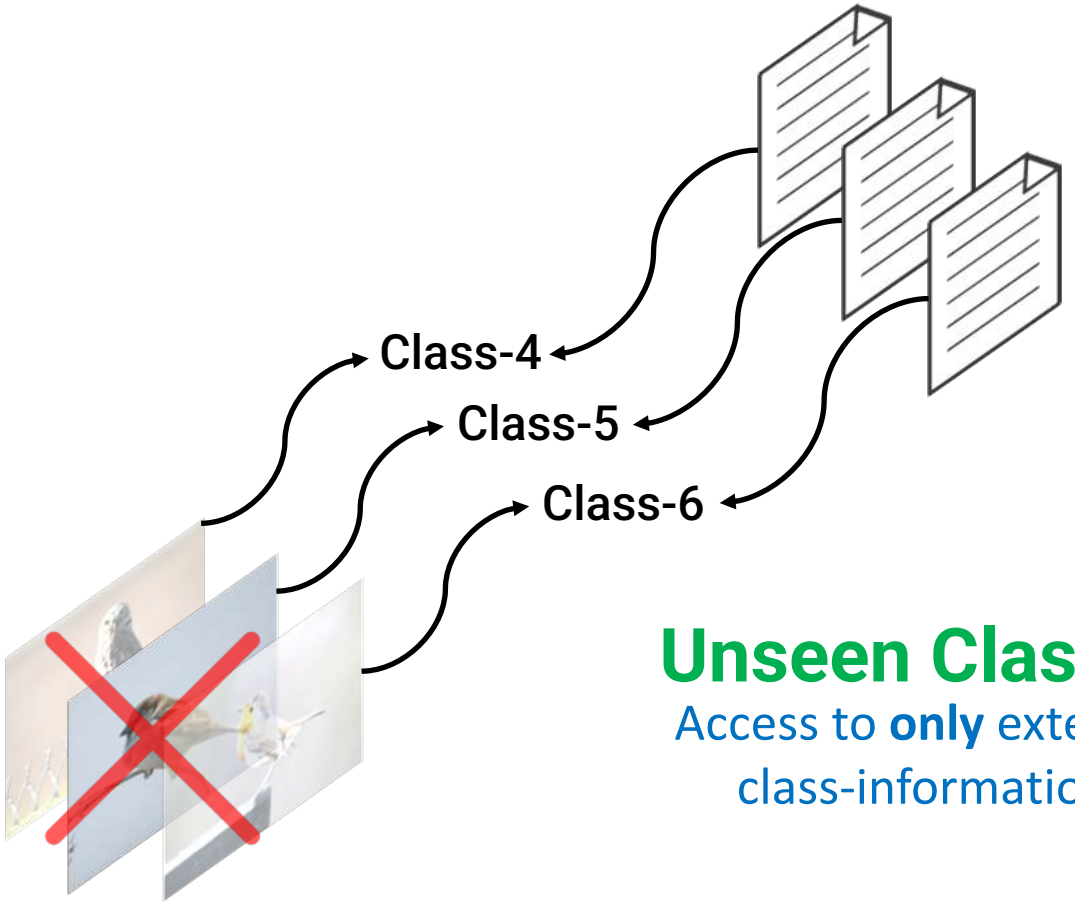
Can interpretability help us make networks generalize better?

Test-Bed: Zero-Shot Learning



Seen Classes

Access to images, labels and external information



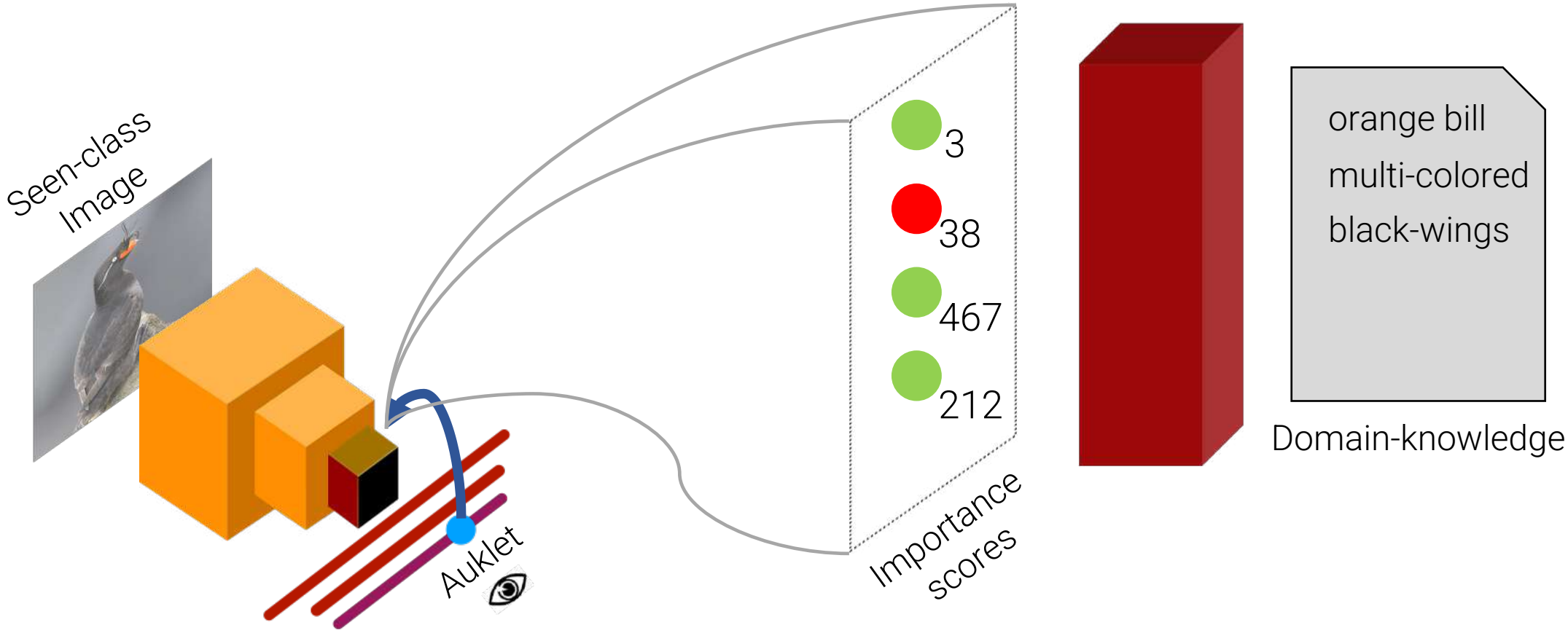
Unseen Classes

Access to **only** external class-information

Task: Generalized classification across seen and unseen classes

Approach

Neuron-Importance Aware Weight Transfer (NIWT)

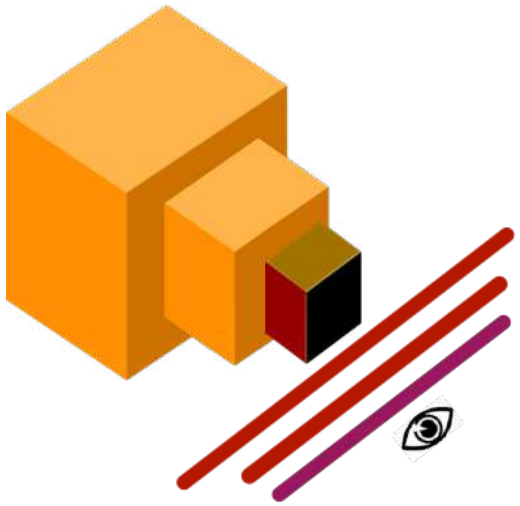


Compute neuron-importance scores for seen-classes

Learn to map neuron-importance scores to domain-knowledge

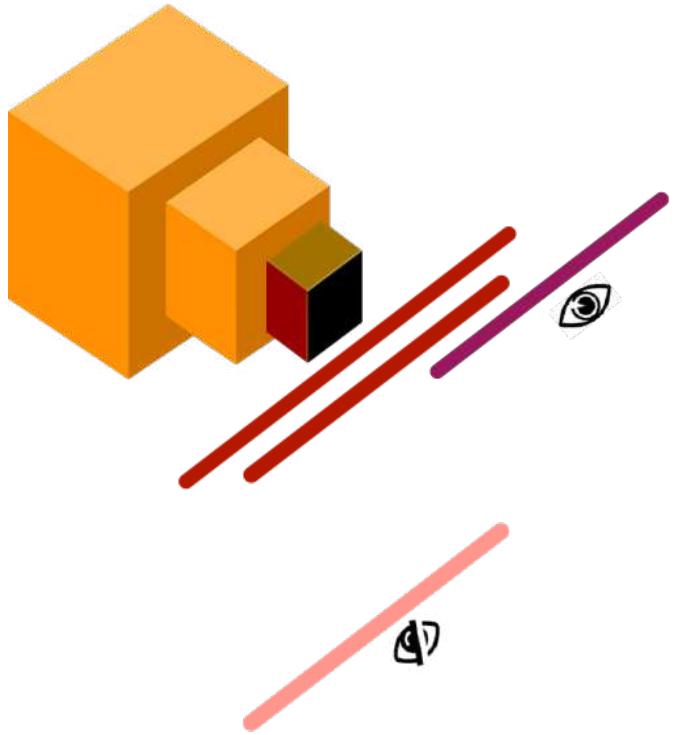
Approach

Neuron-Importance Aware Weight Transfer (NIWT)



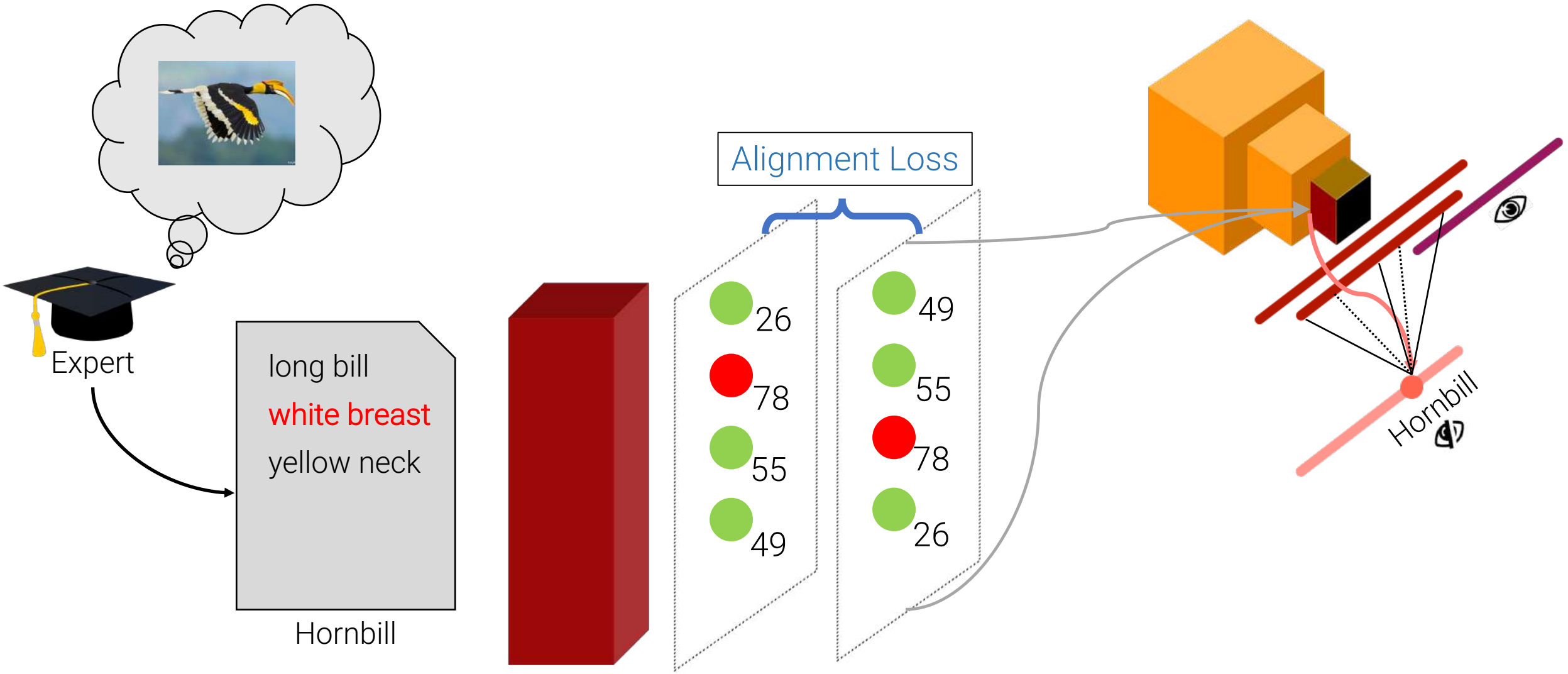
Approach

Neuron-Importance Aware Weight Transfer (NIWT)



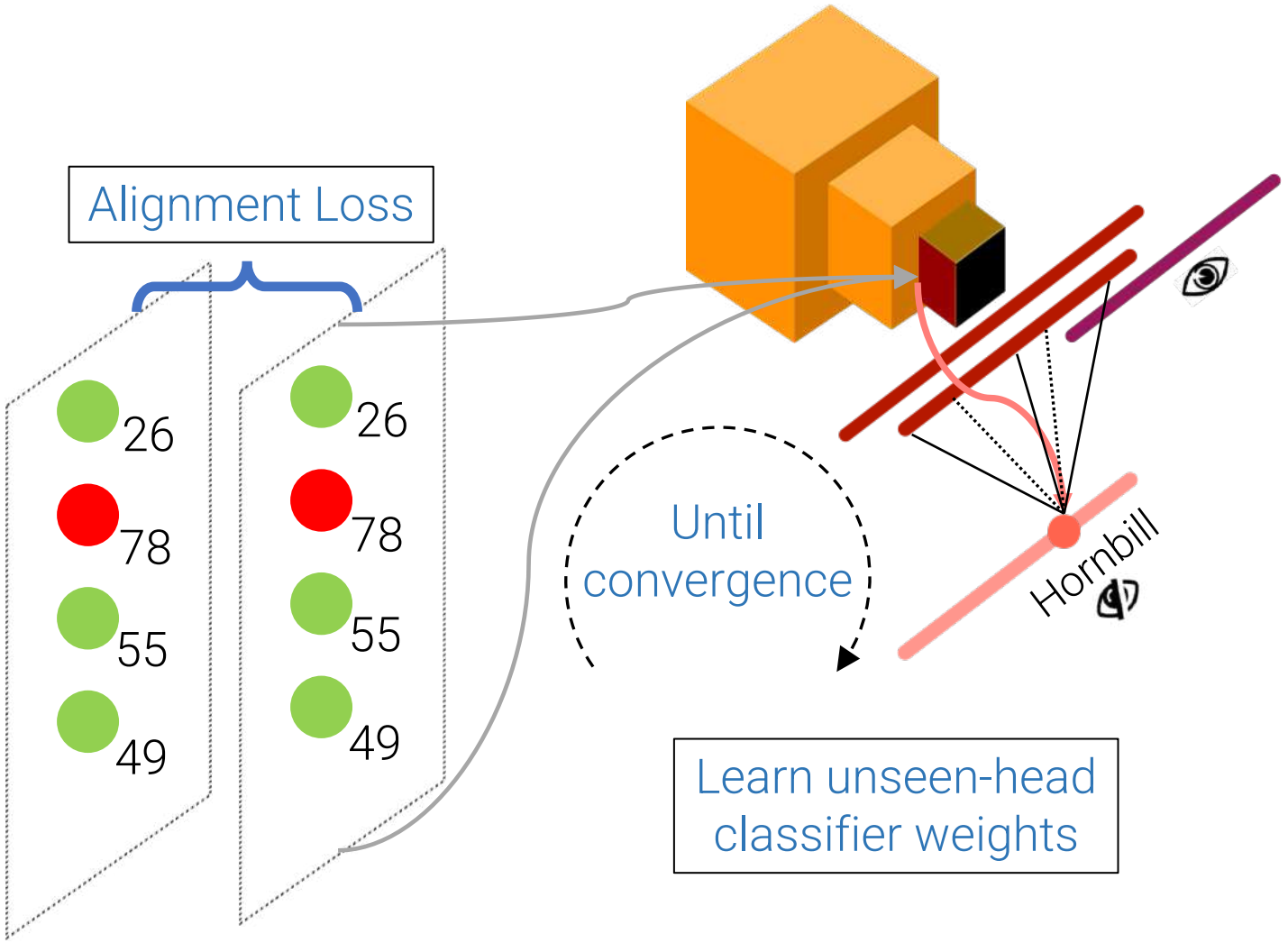
Approach

Neuron-Importance Aware Weight Transfer (NIWT)



Approach

Neuron-Importance Aware Weight Transfer (NIWT)



Qualitative Examples

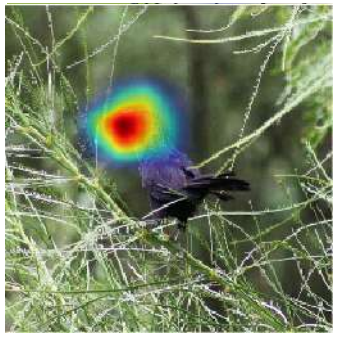
GT Class

Visual + Textual Explanations

Top-3 important neurons with focus

Groove-Billed Ani

ID = 131
has_throat_color = *black*



ID = 116
has_breast_color = *black*

ID = 50
has_underparts_color = *black*

has_throat_color = *black*,
has_breast_color = *black*,
has_nape_color = *black*,
has_primary_color = *black*,
has_forehead_color = *black*

Explanations to improve localization

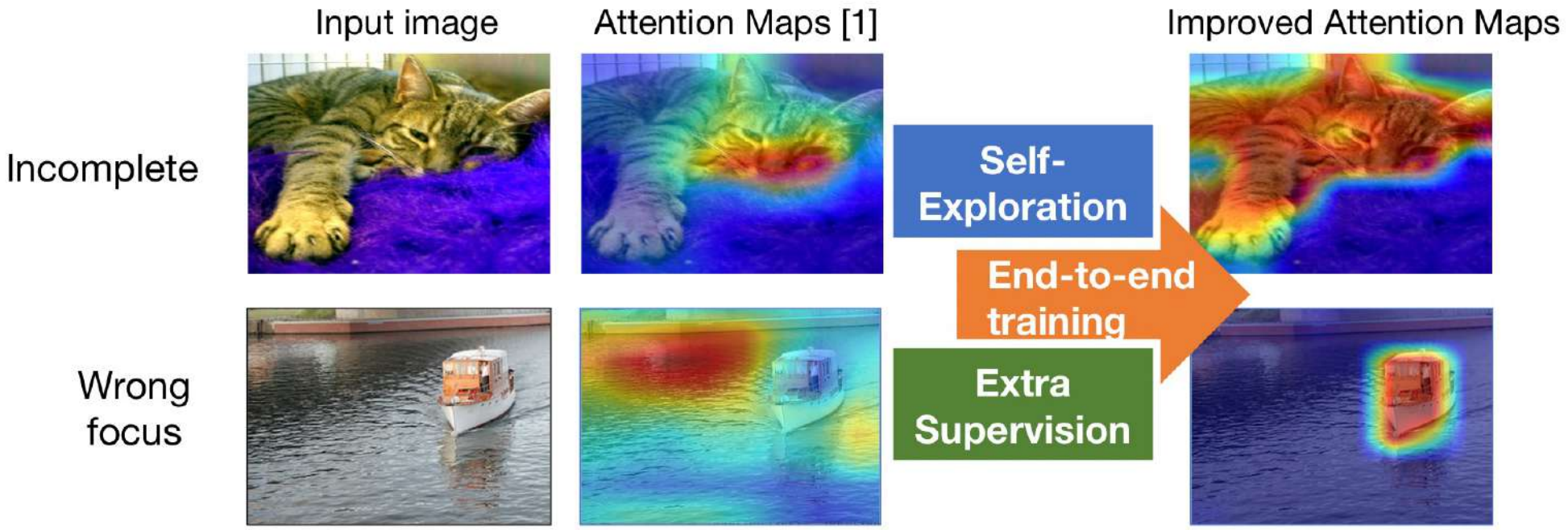
Tell me where to look: Guided attention inference network

CVPR’18 Spotlight

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, Yun Fu.

Issues with Attention maps

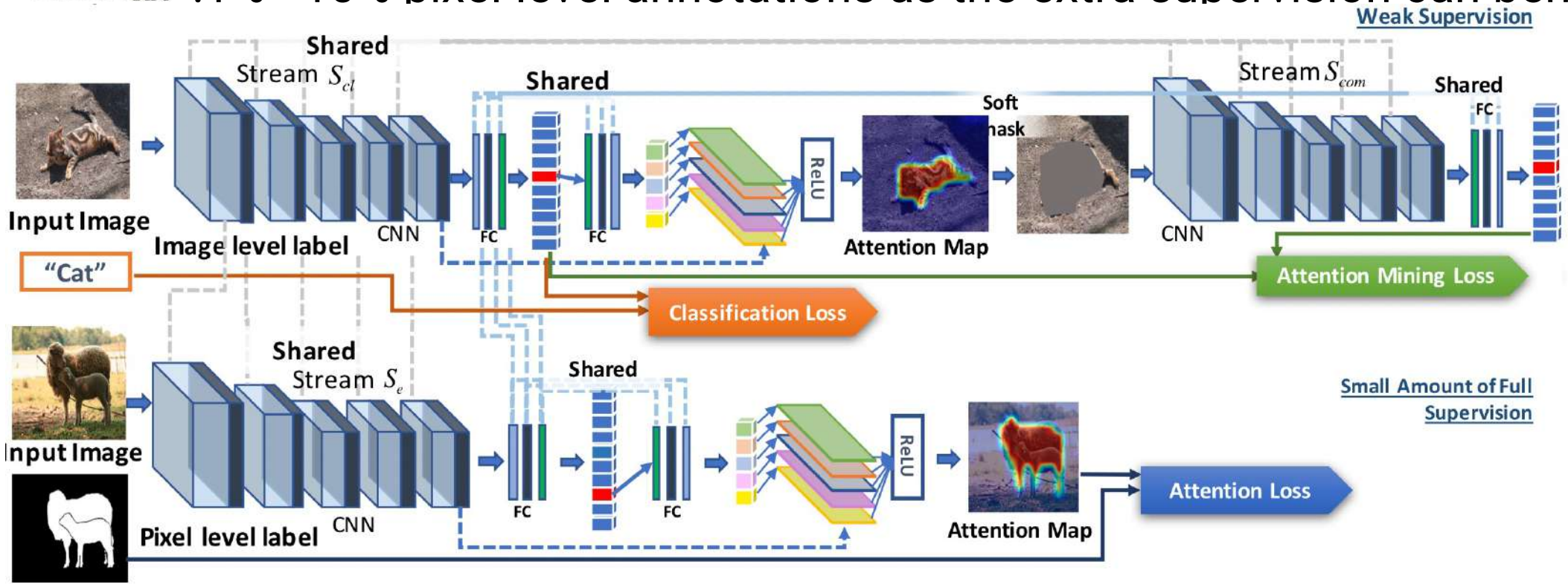
- Attention maps only cover small and most discriminative regions of object of interest
- Network may focus on wrong regions due to the dataset bias



Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, Yun Fu. Tell me where to look: Guided attention inference network, CVPR 2018

Guided Attention Inference Networks

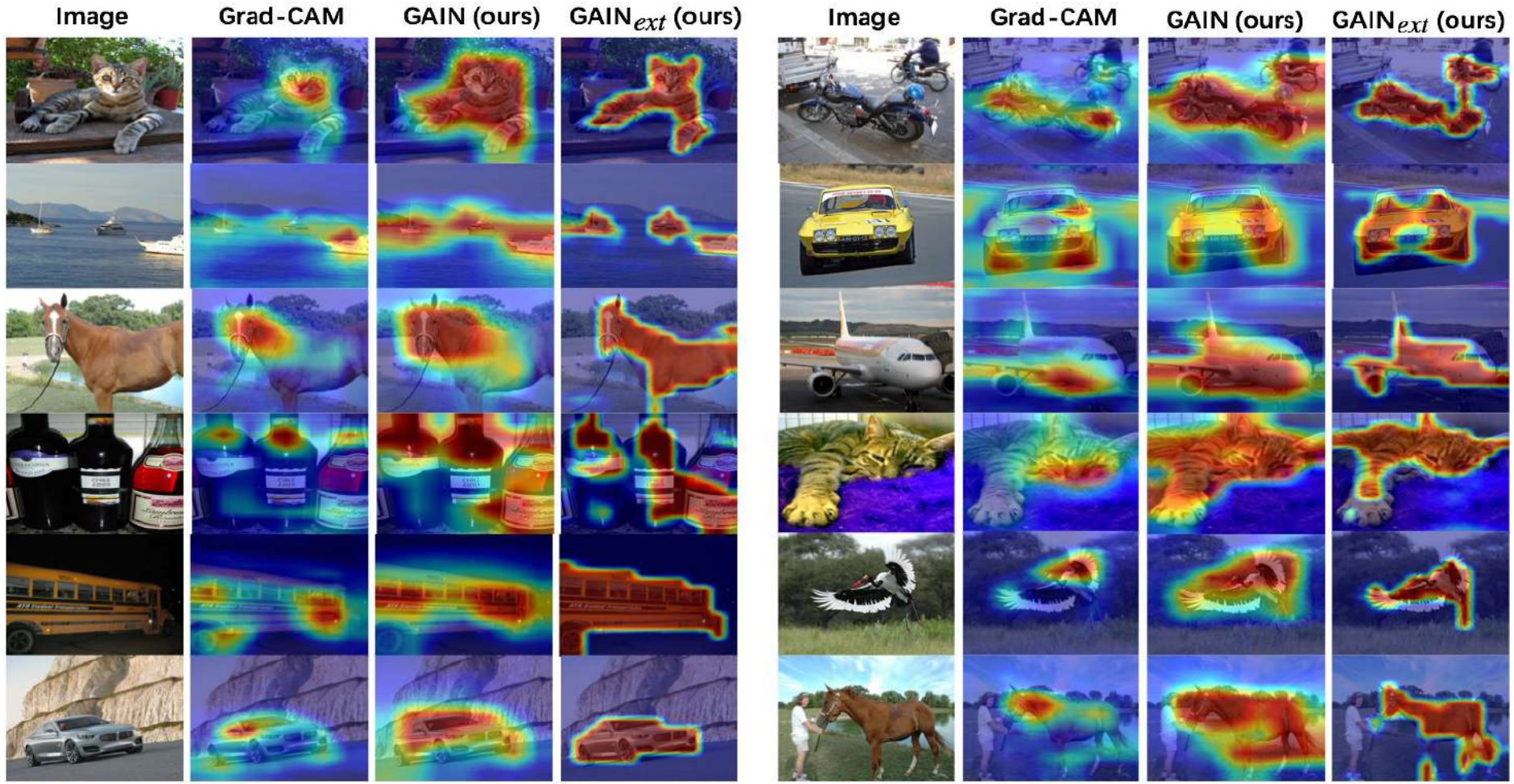
- GAIN builds guidance directly on the attention map during the training process
- $\mathbf{GAIN}_{ext}$:1% - 10% pixel-level annotations as the extra supervision can benefit



Attention Calculation: $\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backp}}$ $L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$

Qualitative Results for Attention Maps

- Our GAIN and $\mathbf{GAIN}_{ext}$ cover more areas belonging to the class of interest



Explanations to transfer knowledge across domains

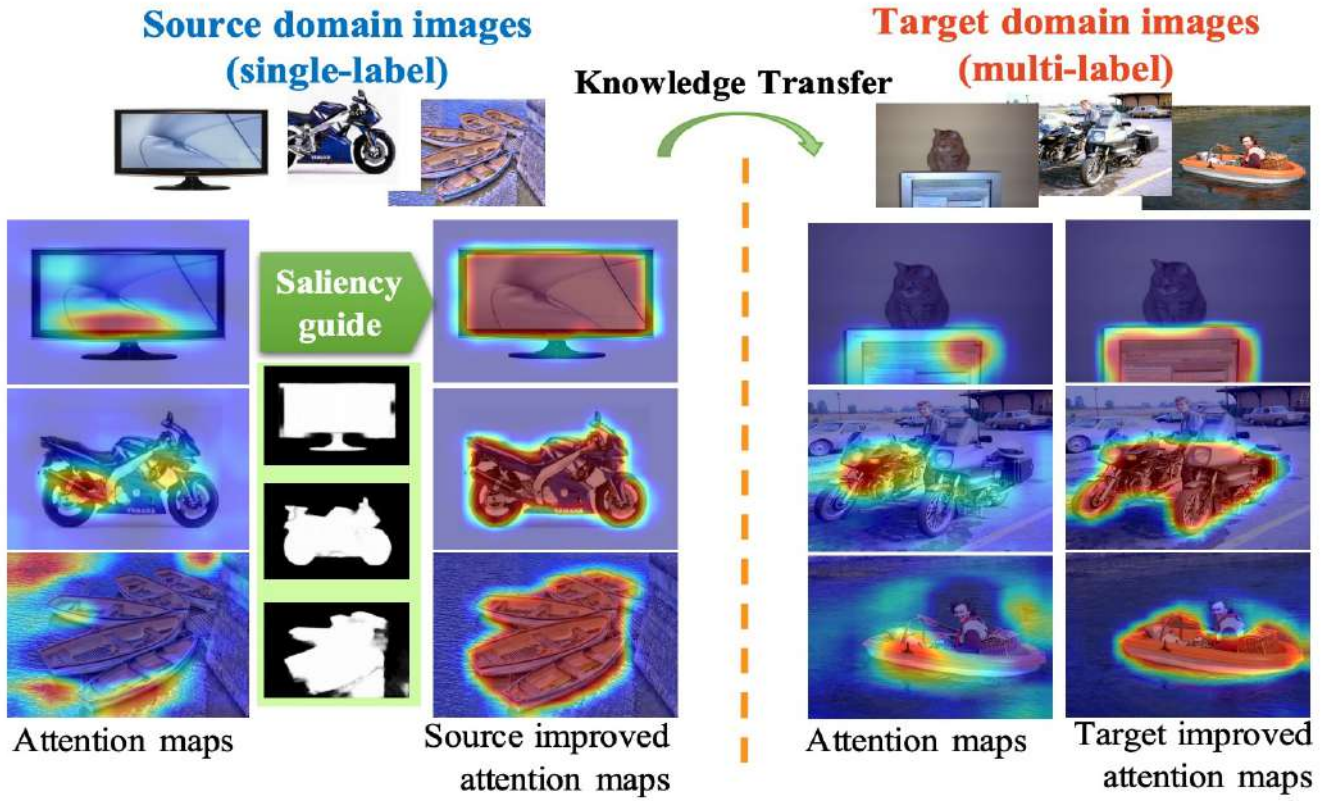
Attention bridging network for knowledge transfer

ICCV 2019

Kunpeng Li, Yulun Zhang, Yuanyuan Li, Yun Fu.

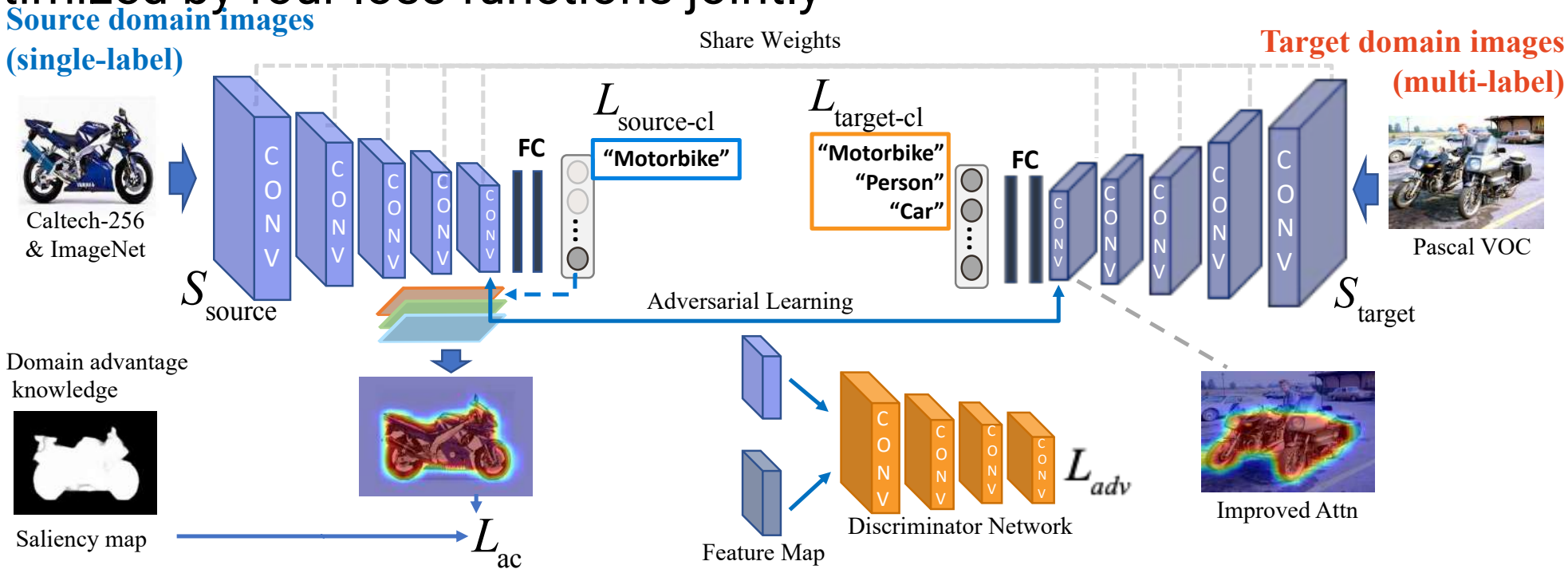
Attention Bridging Network (AttnBN)

- Take network gradient-based attention as a bridge across domains
- Transfer knowledge of integral objects from a single-label dataset (source domain) to another multi-label dataset (target domain)



AttnBN Framework

- AttnBN includes one discriminator network and two streams of classification networks. The attention map is online trainable and optimized by four loss functions jointly



Results

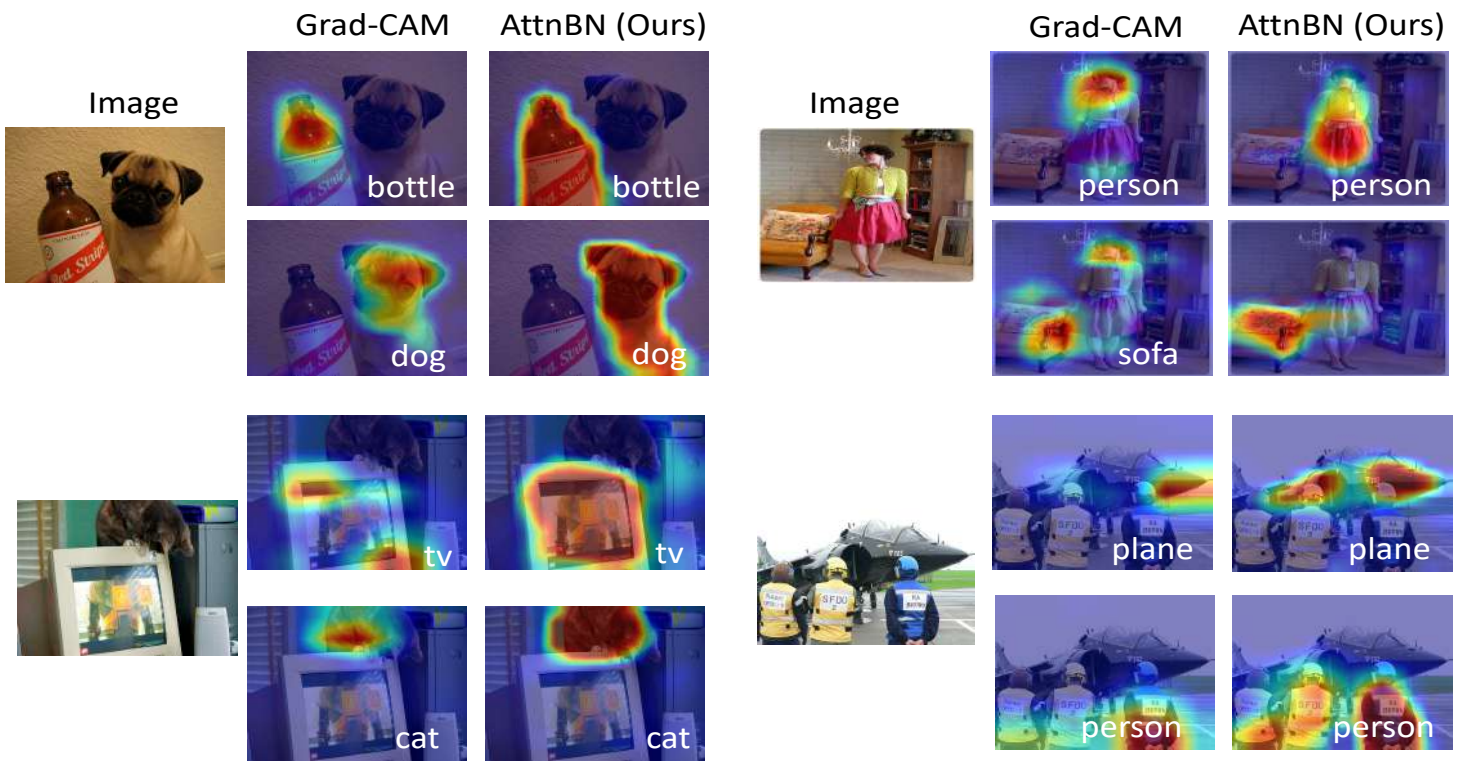


Image
Grad-CAM
AttnBN (Ours)
bottle
bottle
dog
dog
tv
tv
cat
cat
Image
Grad-CAM
AttnBN (Ours)
person
person
sofa
plane
plane
person
person

Explanations to tackle distribution shift

Taking a HINT: Leveraging Explanations to Improve Grounding in Vision and Language

ICCV 2019

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, Devi Parikh

Biases in Vision and Language models

Giraffe standing next to a tree

COCO images

Biases in Vision and Language models

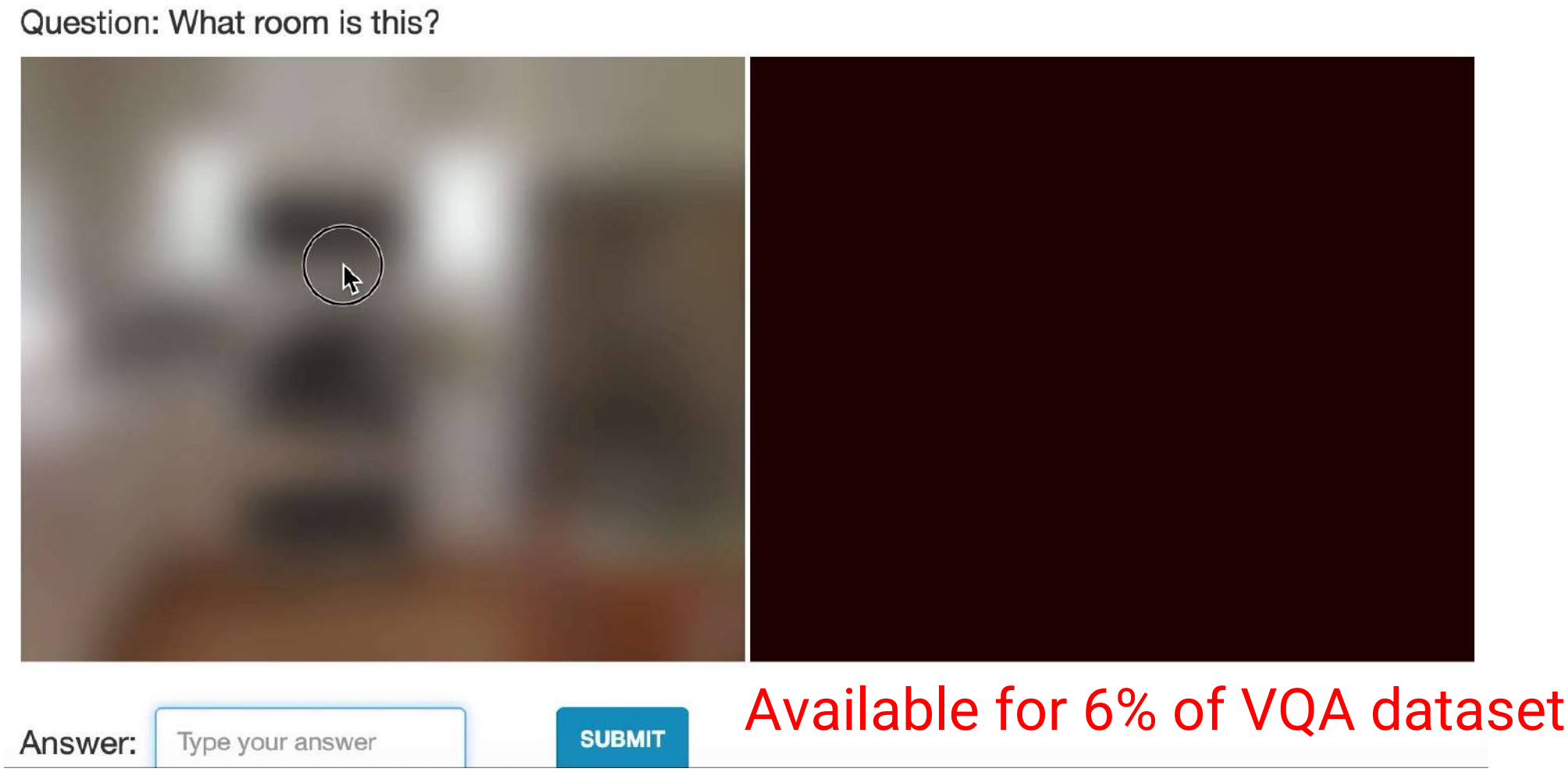
What color are the bananas?

Yellow

COCO training dataset images

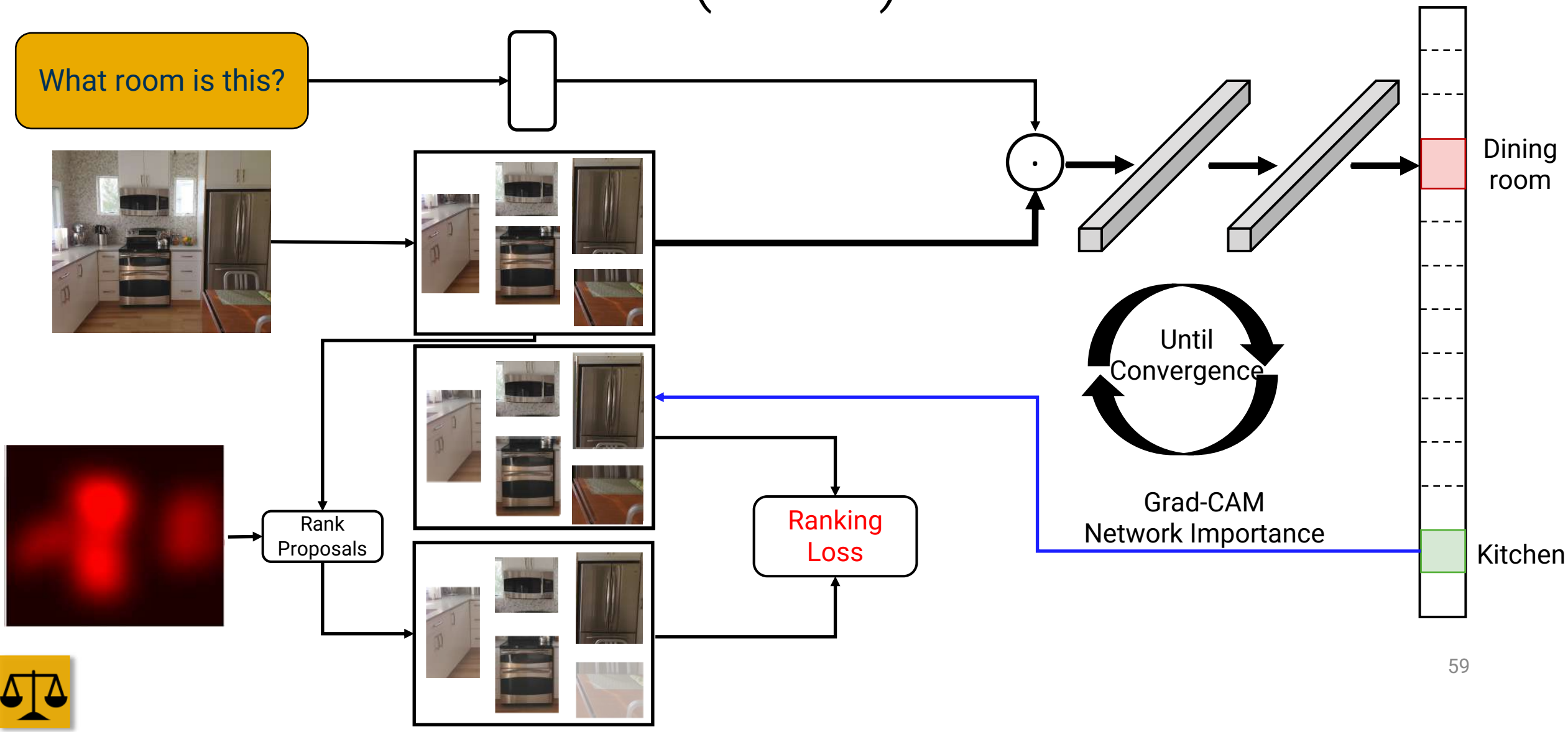
Problematic when distributions change

Where do humans look when making decisions?

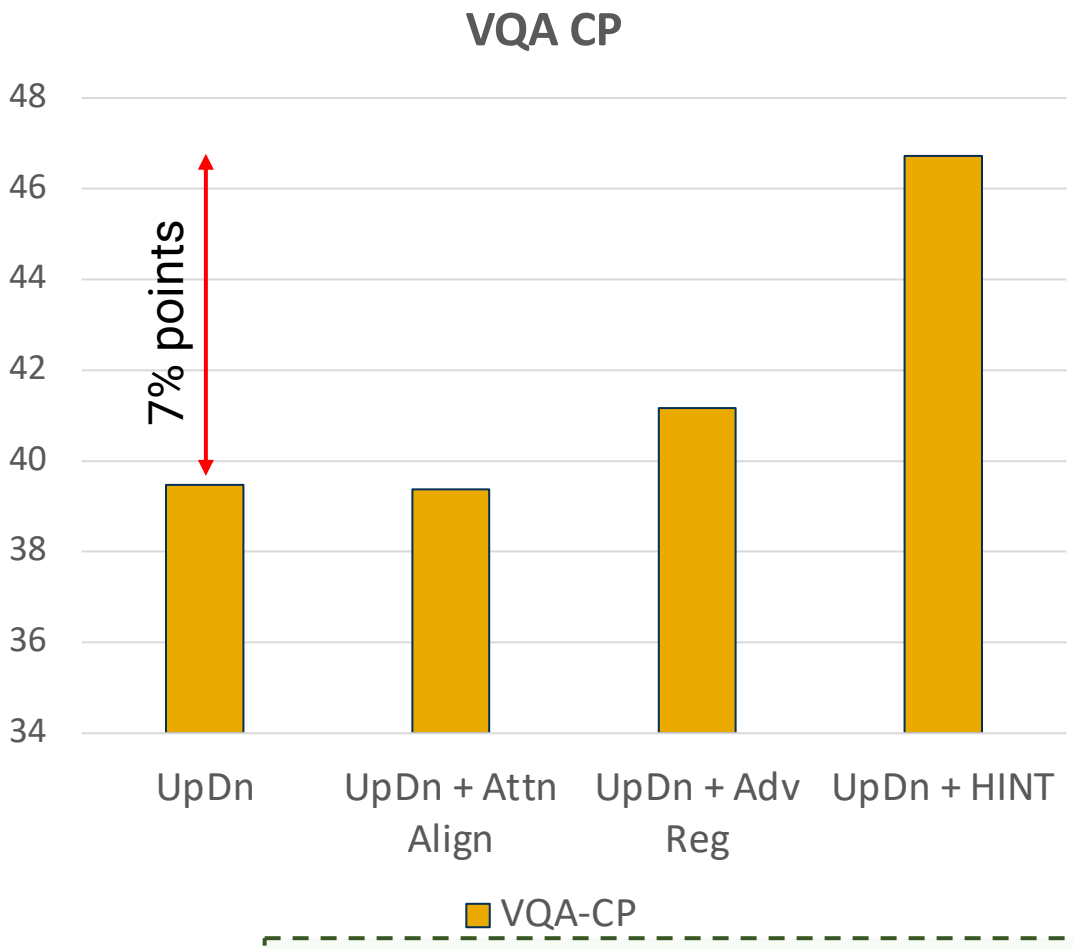


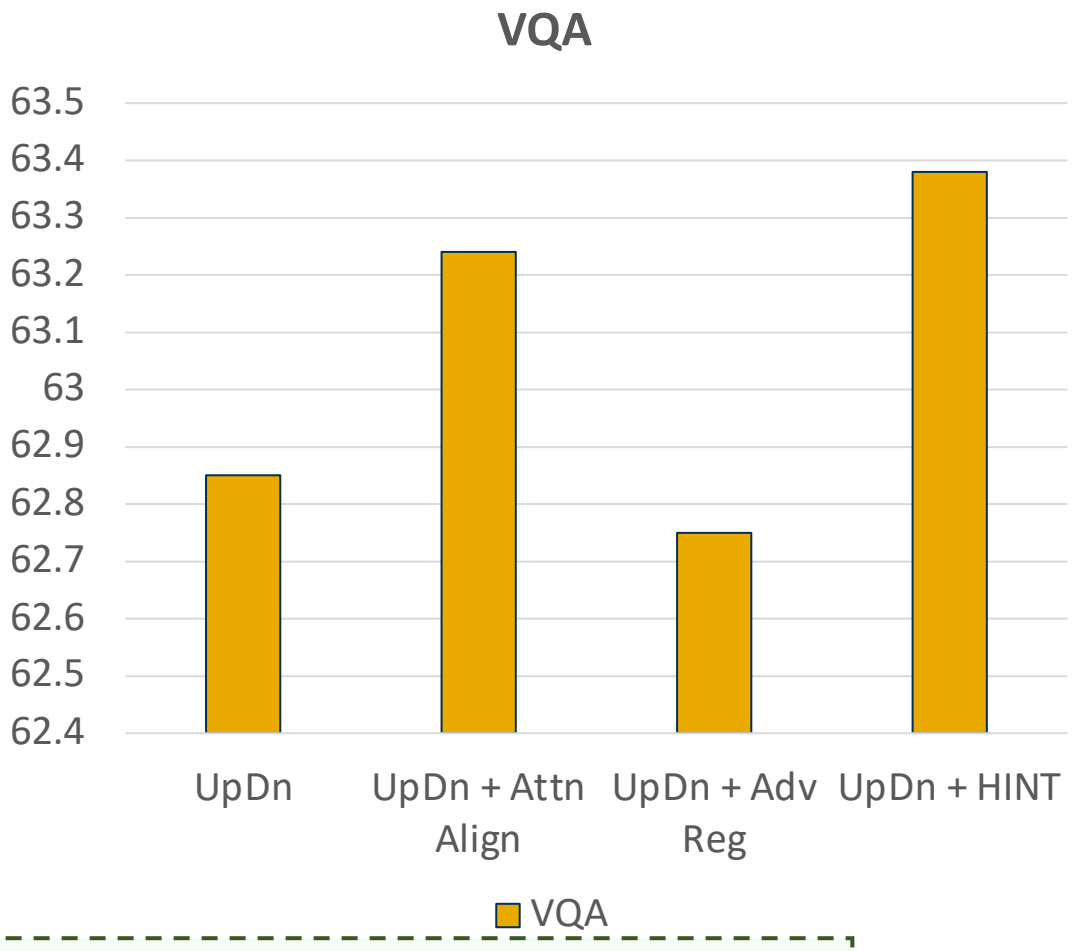
Available for 6% of VQA dataset

Human Importance-aware Network Tuning (HINT)



Results





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

Image Captioning – Do HINTed models look at right regions?

Baseline

toothbrush

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

Baseline

After **HINT**

cat

cat

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

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

Explanations to make models reason compositionally

SQuINTing at VQA Models: Introspecting VQA Models with Sub-Questions

CVPR'20 Oral

Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, Ece Kamar

VQA for visually impaired users

Is the banana ripe enough to eat?

Yes

Is the banana mostly green or yellow?

Green

VQA-Introspect 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”

Do current models reason compositionally?

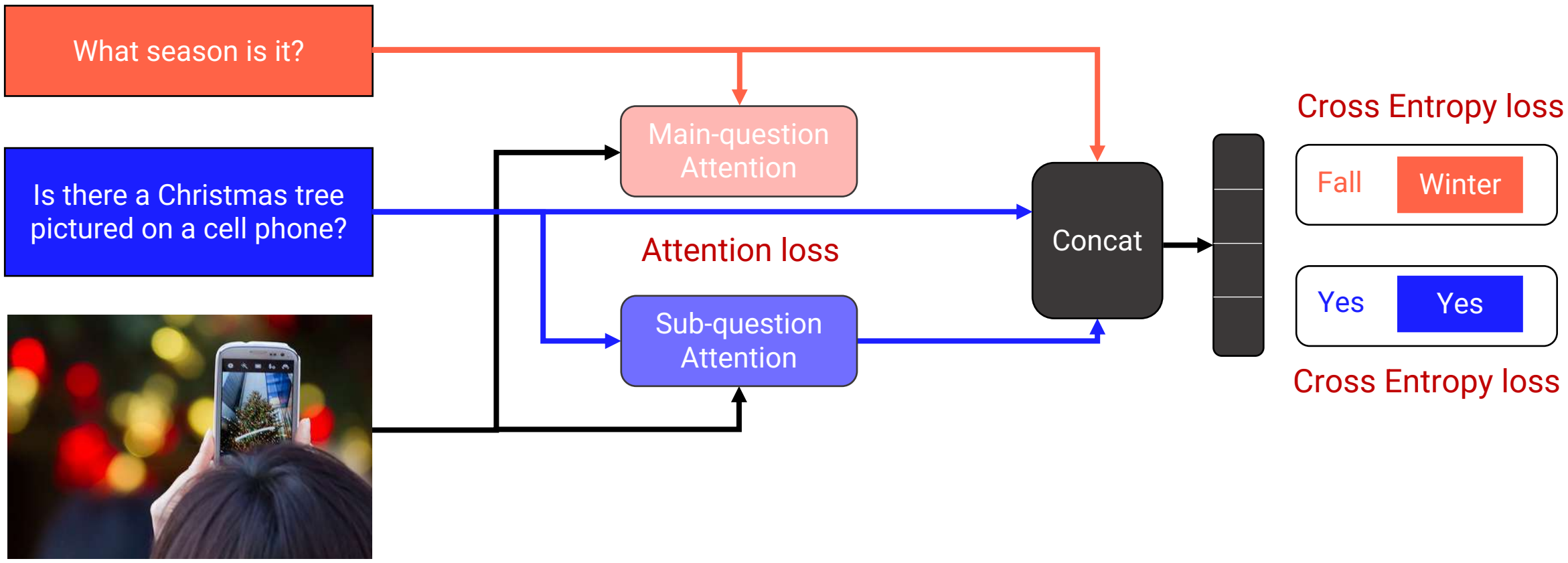
- How consistent are SOTA approaches?

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

Sub-Question Importance-aware Network Tuning (SQuINT)



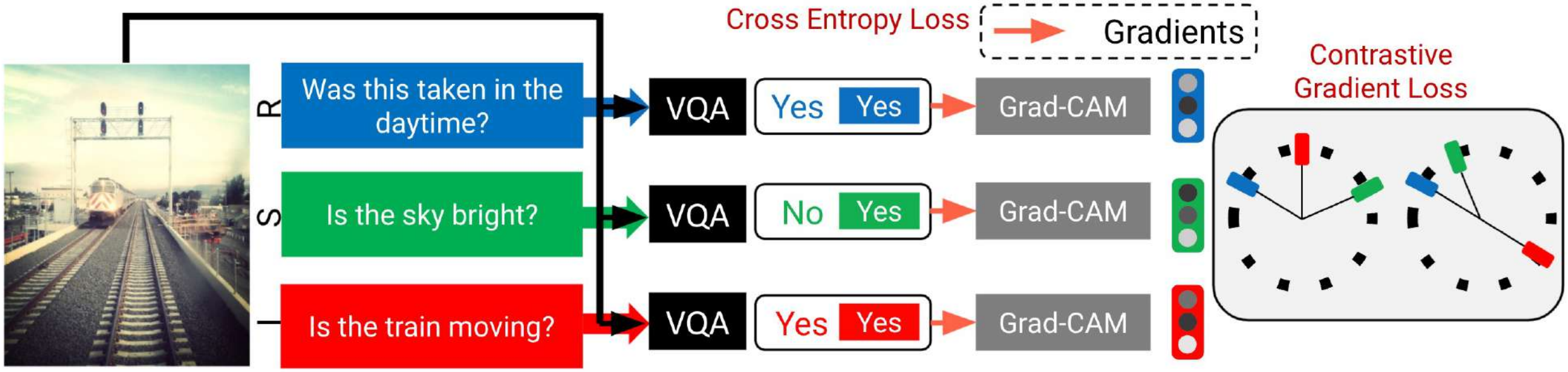
Results

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

Making VQA models consistent

- Idea: Language-based interpretability technique to enforce models to rely on relevant sub-questions more than irrelevant ones



R = Reasoning Question

S = Sub-Question

I = Irrelevant Question

Explanations to improve Self-Supervised representation learning

CASTing Your Model: Learning to Localize Improves Self-supervised Representations

CVPR'21

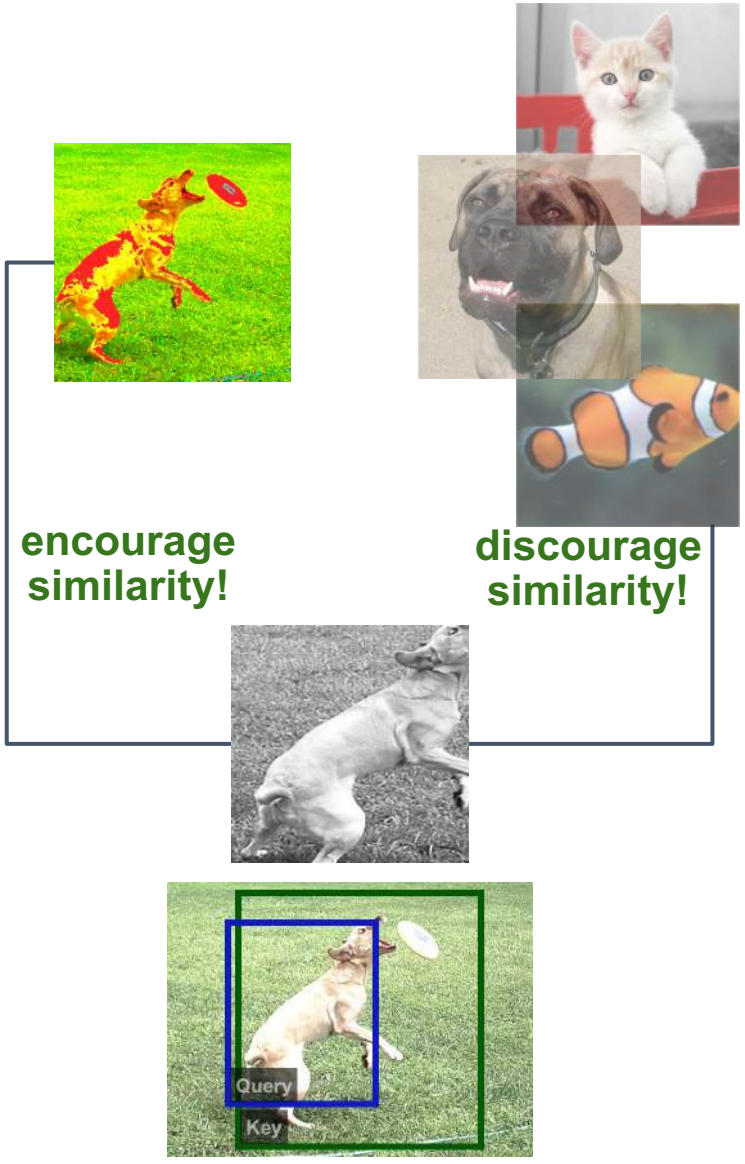
Ramprasaath R. Selvaraju, Karan Desai, Justin Johnson, Nikhil Naik

Contrastive Self-Supervised Learning (SSL)

- Learn visual representations from unlabeled images through instance discrimination
- Recent methods outperform fully supervised Imagenet pretraining.

Shortcomings:

- Success largely confined to iconic images.
- Direct application to uncurated web or scene-level images show marginal gains

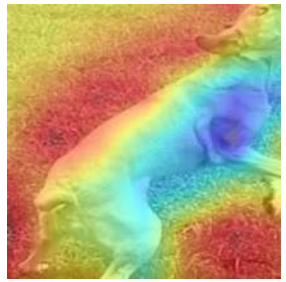


Why SSL methods fail to generalize to arbitrary images?

- Poor visual grounding ability

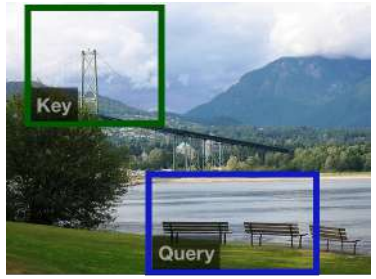

Query


Key


Grad-CAM on Query

Models exploit low level visual cues or spurious background correlations

- Receive imperfect supervisory signal when augmented views contain different visual concepts




Query


Key


Grad-CAM on Query

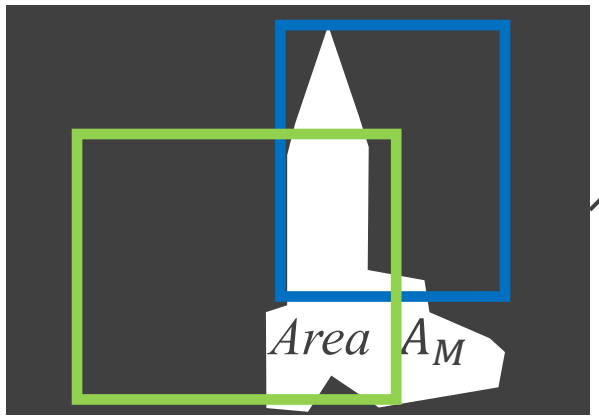
May discourage semantic understanding

Leads to diminishing gains on web-scale images and performance regression on complex scene images

Saliency Constrained Random Cropping

Crops share a common salient region between them

Image



Unsupervised Saliency Map

Saliency Constrained Random Crop

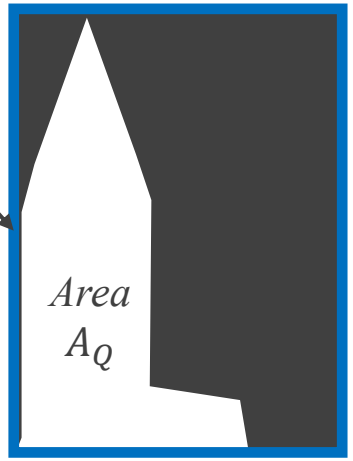
$A_Q, A_K \geq \varphi A_M$

$\varphi \in [0, 1)$

Constraints

Query

Key



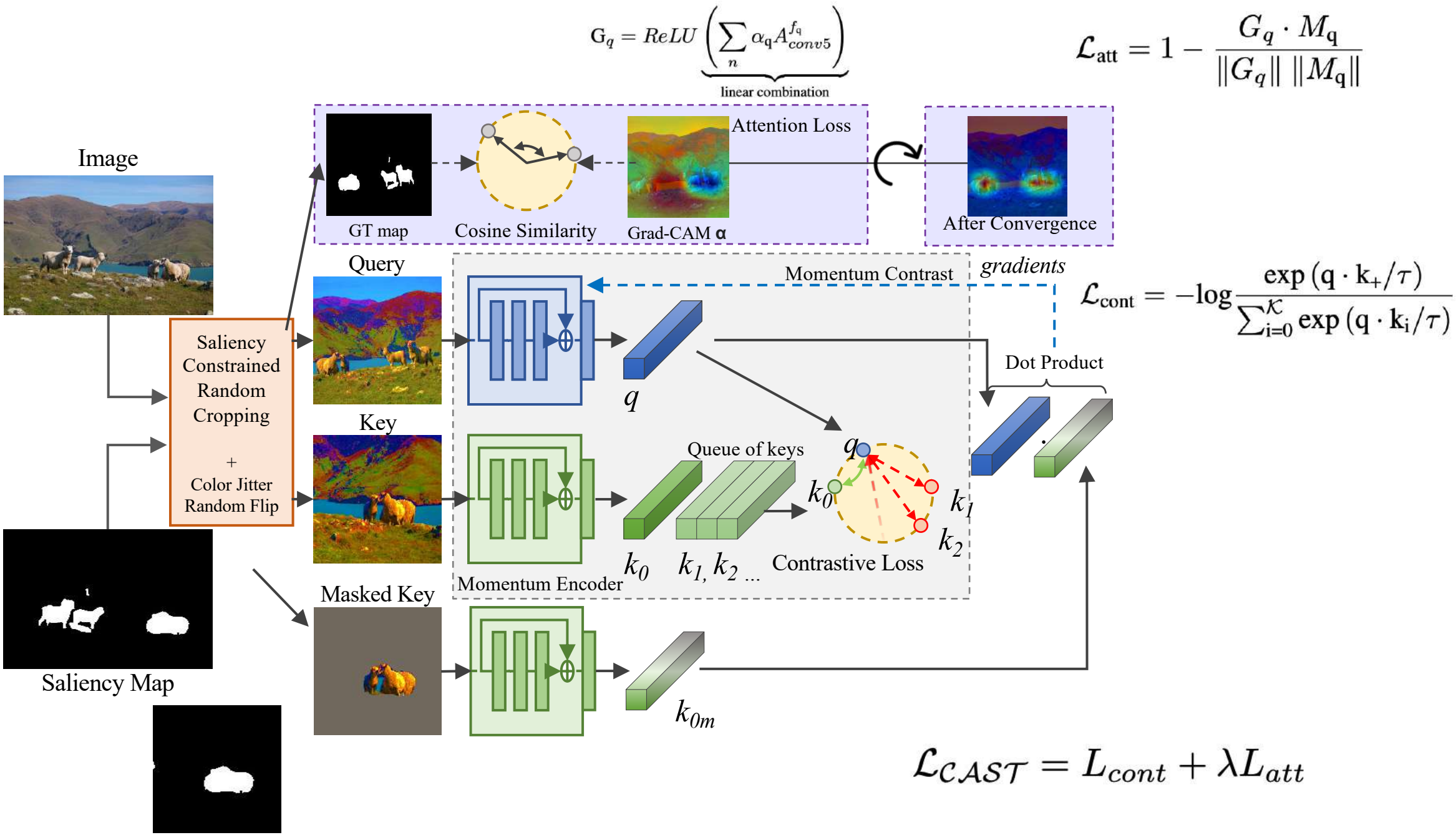
Query Saliency

Key Saliency

Crop oriented saliency maps which can be used as supervision

CAST: Contrastive Attention-Supervised Tuning

Idea: Encourage models to rely on the appropriate salient regions during contrastive pretraining

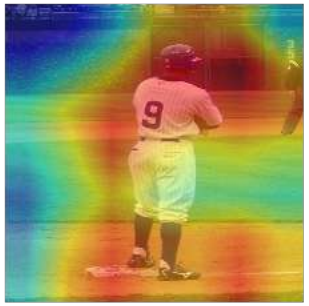

$G_q = ReLU\left(\sum_n \alpha_q A^{f_q}_{conv5}\right)$
linear combination
$\mathcal{L}_{att} = 1 - \frac{G_q \cdot M_q}{\|G_q\| \, \|M_q\|}$
Image
GT map
Cosine Similarity
Grad-CAM α
Attention Loss
After Convergence
Query
Key
Masked Key
Saliency Constrained Random Cropping
+
Color Jitter
Random Flip
Saliency Map
Momentum Contrast
gradients
q
Momentum Encoder
Queue of keys
k_0
k_1, k_2 ...
Contrastive Loss
k_1
k_2
Dot Product
k_{0m}
$\mathcal{L}_{cont} = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)}$
$\mathcal{L}_{CAST} = L_{cont} + \lambda L_{att}$

Transfer Learning on Downstream Tasks

Method	VOC07 clf.	IN-1k clf.	PASCAL VOC Detection			COCO Instance Segmentation					
	mAP	Top-1 acc.	AP^{bbox}_{all}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{bbox}_{all}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}_{all}	AP^{mask}_{50}	AP^{mask}_{75}
1) Random Init	–	–	33.8	60.2	33.1	36.7	56.7	40.0	33.7	53.8	35.9
2) ImageNet Fully Sup	–	–	53.5	81.3	59.1	38.9	59.6	42.7	35.4	56.5	38.1
3) COCO Fully Sup	86.2	46.4	50.9	79.2	54.7	40.3	61.3	43.7	36.5	58.1	39.1
4) MoCo-COCO	67.5	46.5	47.5	75.4	51.5	38.3	58.7	41.5	34.9	55.7	37.2
5) + Constrained Crop	71.1 +3.6	46.0 −0.5	49.0 +1.5	77.4 +2.0	52.4 +0.9	38.3 +0.0	58.7 +0.0	41.6 +0.1	34.8 −0.1	55.7 +0.0	37.2 +0.0
6) + CAST	74.0 +6.5	48.7 +2.1	54.2 +6.7	80.1 +4.7	59.9 +8.4	39.4 +1.1	60.0 +1.3	42.8 +1.3	35.8 +0.9	57.1 +1.4	38.6 +1.4

CAST outperforms all baselines by a huge margin on all downstream tasks

Does improved SSL grounding transfer to downstream tasks?

Original	Fully-supervised Imagenet Grad-CAM	MoCo-COCO Grad-CAM	MoCO-COCO + CAST Grad-CAM
			
GT: Ball Player	Pred: Ball Player	Pred: Ball Player	Pred: Ball Player
			
GT: Street Sign	Pred: Street Sign	Pred: Street Sign	Pred: Street Sign

CAST models rely less on correlations

CAST models attend to whole extent of objects

CAST makes models more robust

Backgrounds Challenge

MoCo-COCO Performance	Original	Mixed-Same	Mixed-Rand	Only FG
Default	72.62	45.75	30.44	30.42
+ Constrained Cropping	74.79	52.64	39.14	33.73
+ CAST	77.33	54.42	39.93	43.26

13% improvement

CAST makes models resilient to background changes

Original

GT Class	Bird
Pred MoCo	Bird
Pred MoCo + CAST	Bird

Mixed Rand

GT Class	Bird
Pred MoCo	Fish
Pred MoCo + CAST	Bird

Original

GT Class	Reptile
Pred MoCo	Reptile
Pred MoCo + CAST	Reptile

Mixed Rand

GT Class	Reptile
Pred MoCo	Instrument
Pred MoCo + CAST	Reptile

Explanations to reduce catastrophic forgetting

Remembering for the Right Reasons: Explanations Reduce Catastrophic Forgetting

ICLR’21

Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E. Gonzalez, Marcus Rohrbach, Trevor Darrell

Continual Learning

Definition:

learning a sequence of tasks without *catastrophic forgetting*

Continual Learning

Definition:

learning a sequence of tasks without *catastrophic forgetting*

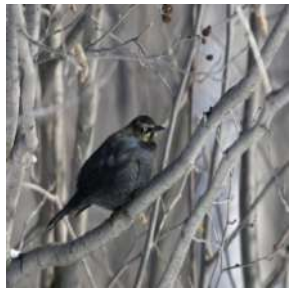
Hypothesis:

Catastrophic forgetting is due in part to forgetting the original reasoning for a previous prediction.

eXplainable AI for Continual Learning

Hypothesis:

Catastrophic forgetting is due in part to forgetting the original reasoning for a previous prediction.

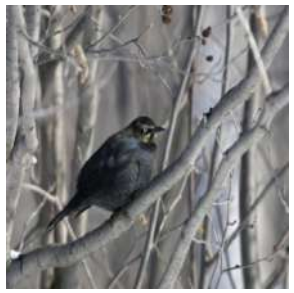


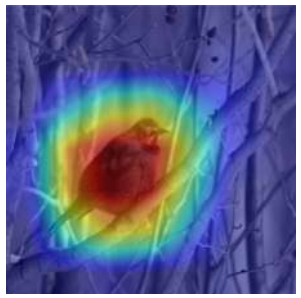
Task t

eXplainable AI for Continual Learning

Hypothesis:

Catastrophic forgetting is due in part to not being able to rely on the **same reasoning** as was used for a previously seen observation.





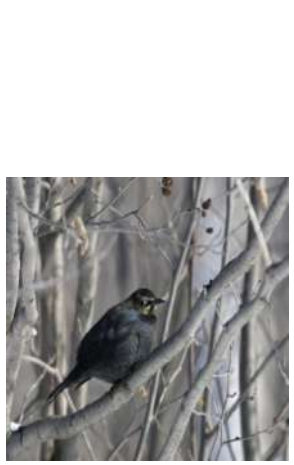
Initial explanation/saliency map

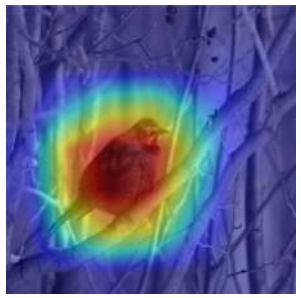
Task t

eXplainable AI for Continual Learning

Hypothesis:

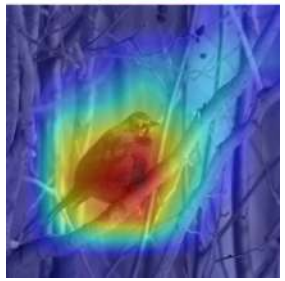
Catastrophic forgetting is due in part to not being able to rely on the **same reasoning** as was used for a previously seen observation.





Initial explanation/saliency map

Task t



Task $t+1$

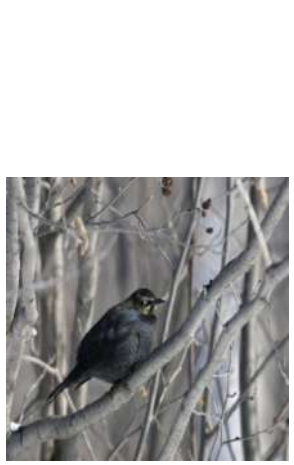
Final Task

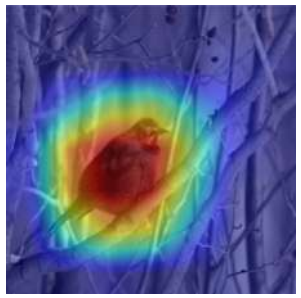
Model *remembers the initial reason*

eXplainable AI for Continual Learning

Hypothesis:

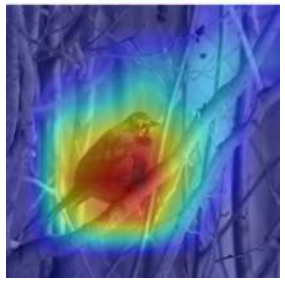
Catastrophic forgetting is due in part to not being able to rely on the **same reasoning** as was used for a previously seen observation.





Initial explanation/saliency map

Task t



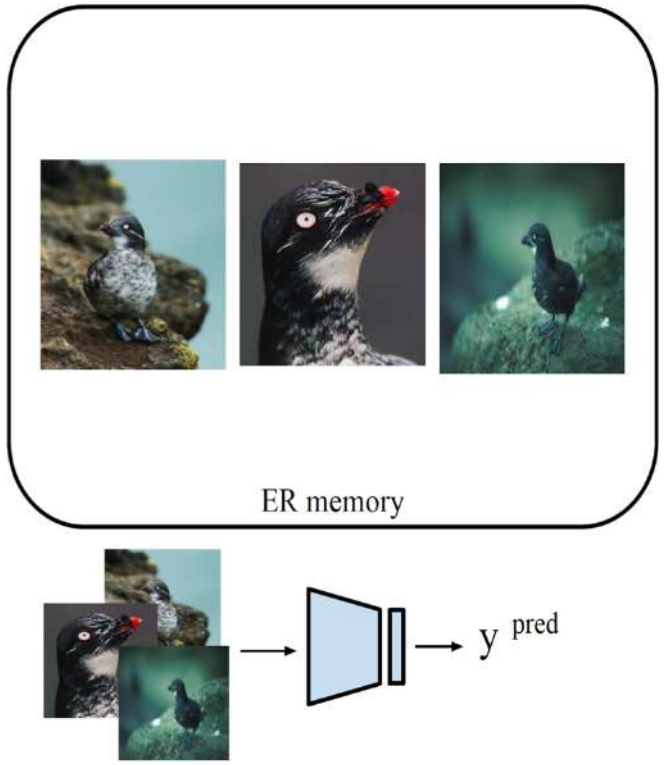
Task $t+1$

Final Task

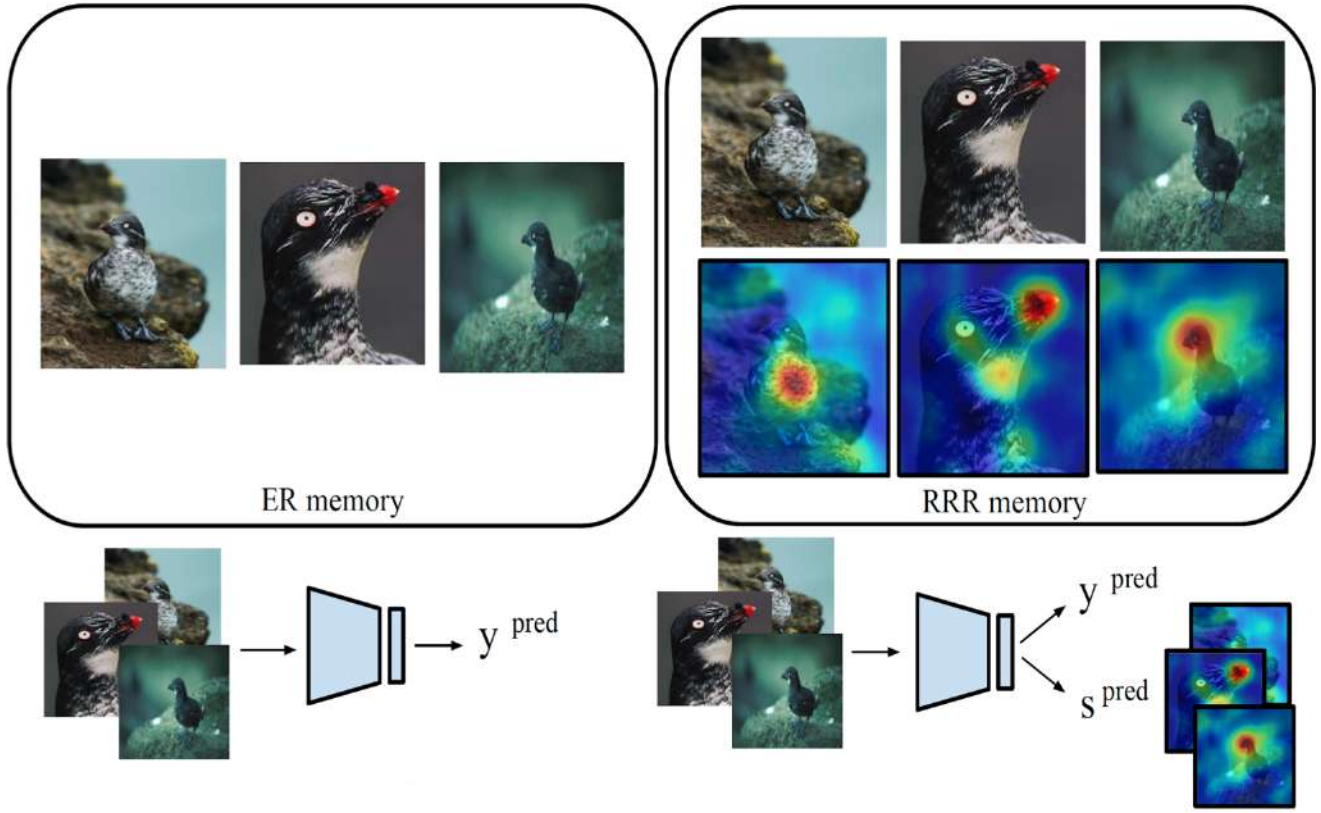
Model *remembers the initial reason*

Model *forgets the initial reason*

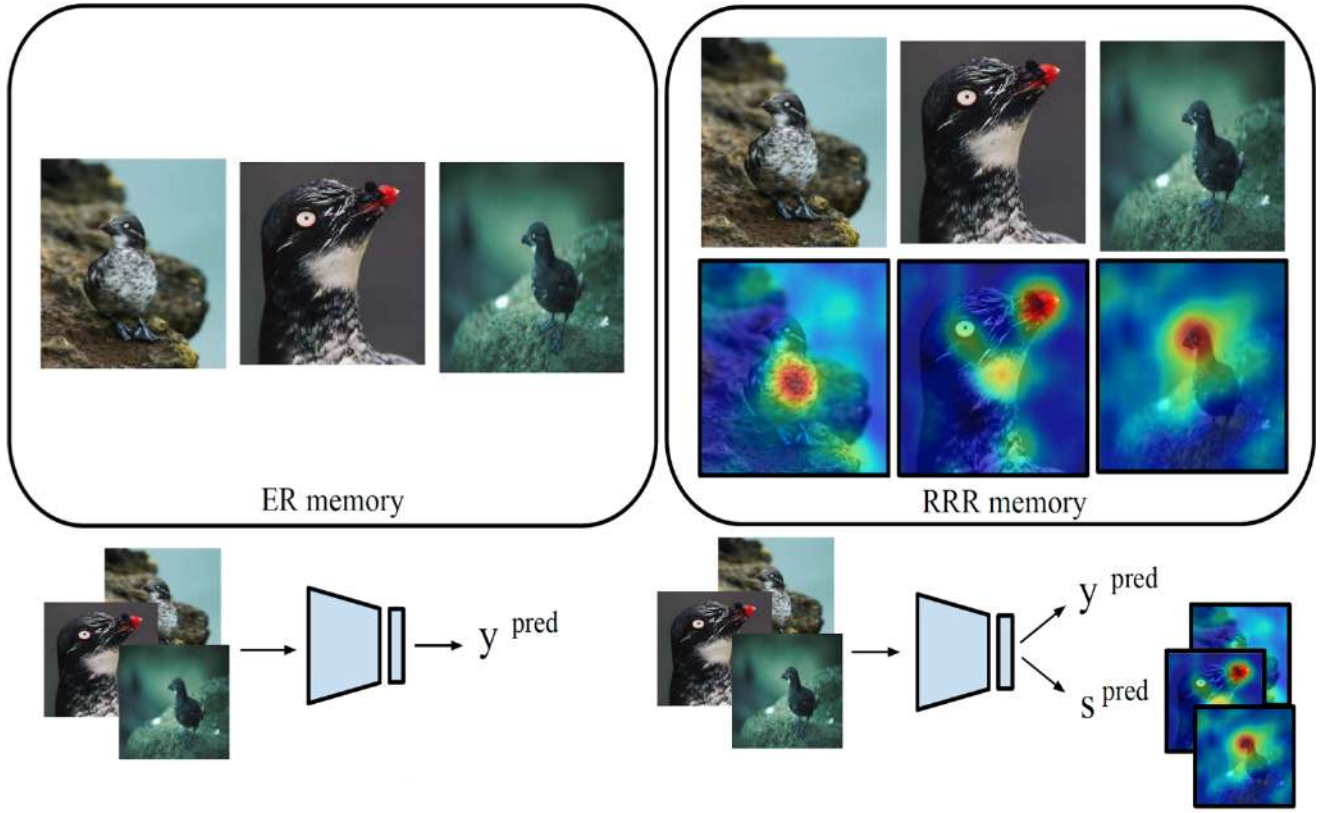
Experience Replay (ER) in CL



Remembering for the Right Reasons (RRR)

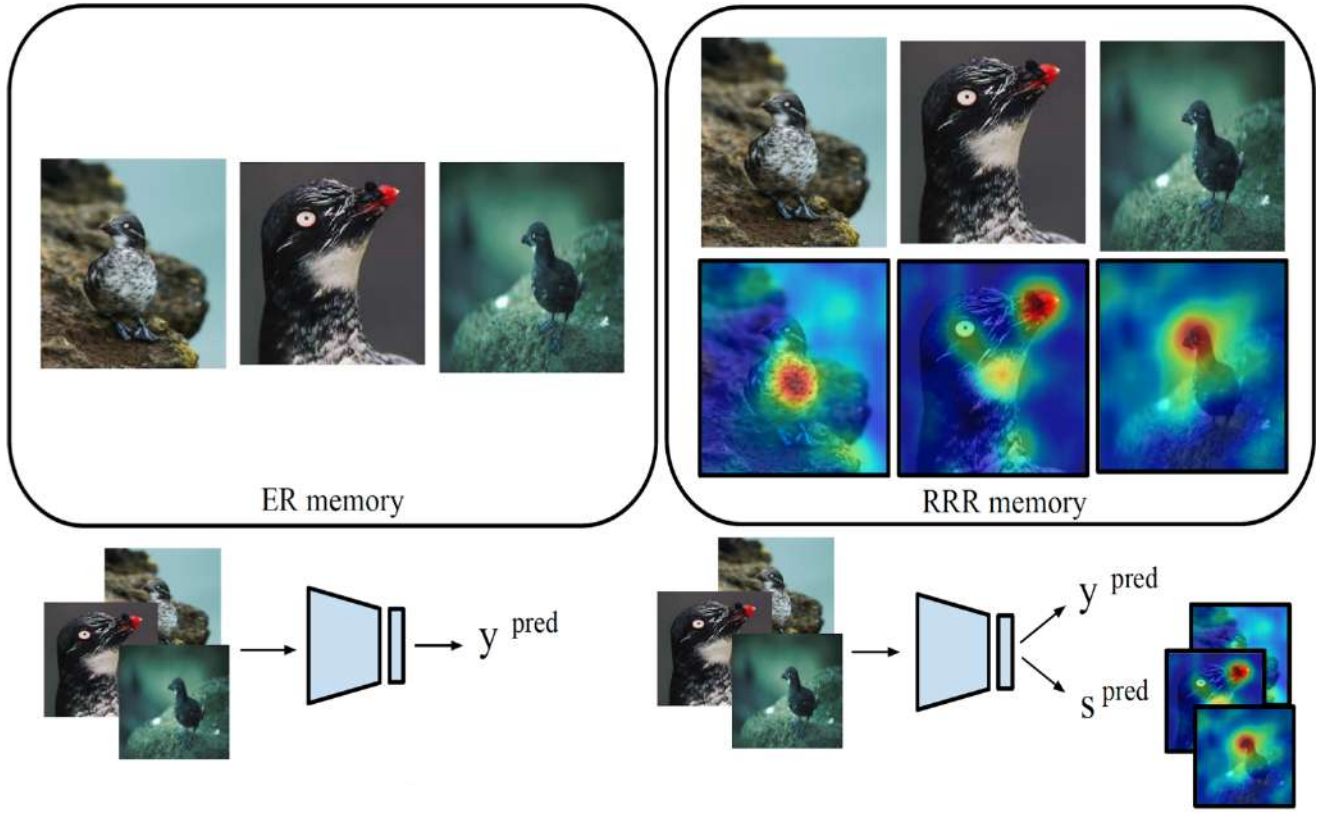


Remembering for the Right Reasons (RRR)



$$\mathcal{L}_{\text{RRR}}(f_\theta, \mathcal{M}^{\text{rep}}, \mathcal{M}^{\text{RRR}}) = \mathbb{E}_{((x,y),\hat{s})\sim(\mathcal{M}^{\text{rep}},\mathcal{M}^{\text{RRR}})} ||\mathcal{XAI}(f_\theta^k(x)) - \hat{s}||_1$$

Remembering for the Right Reasons (RRR)



Saliency maps

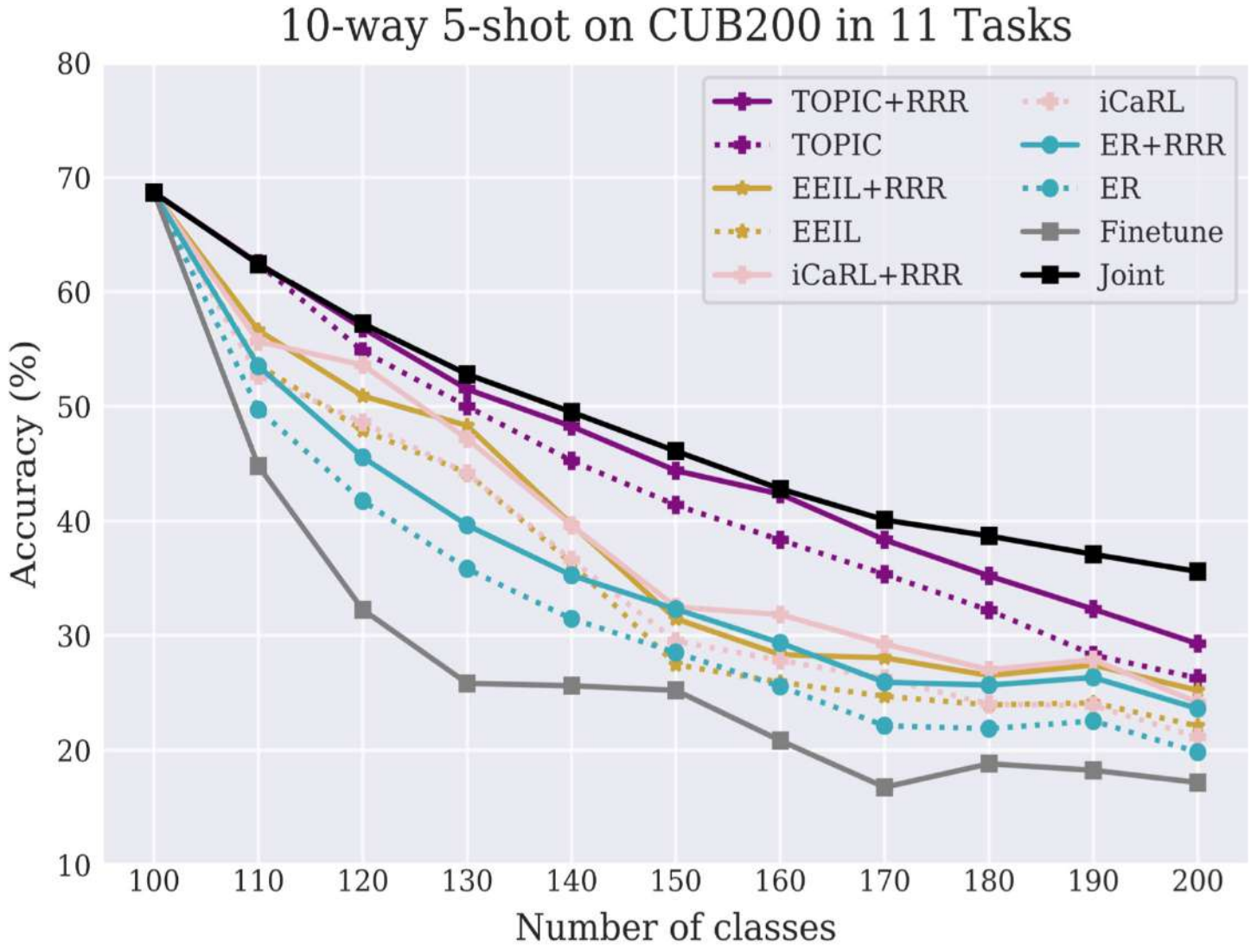
Vanilla Backprop
(Zeiler & Fergus, 2014)

Backprop w/ SmoothGrad
(Smilkov et al., 2017)

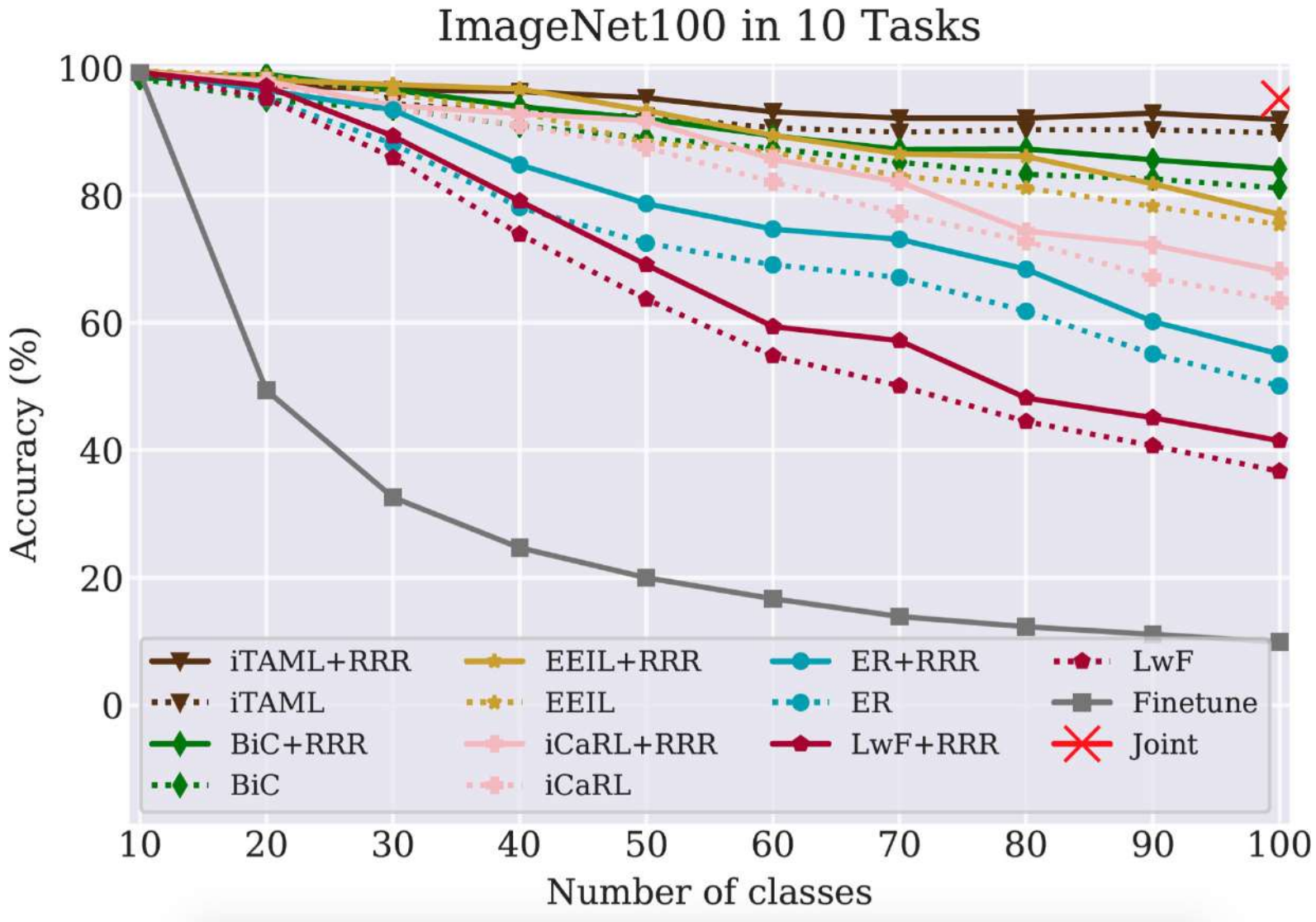
Grad-CAM
(Selvaraju et al., 2017)

$$\mathcal{L}_{\text{RRR}}(f_\theta, \mathcal{M}^{\text{rep}}, \mathcal{M}^{\text{RRR}}) = \mathbb{E}_{((x,y),\hat{s})\sim(\mathcal{M}^{\text{rep}},\mathcal{M}^{\text{RRR}})} || \mathcal{XAI}(f_\theta^k(x)) - \hat{s} ||_1$$

Results: Few-Shot Class Incremental Learning

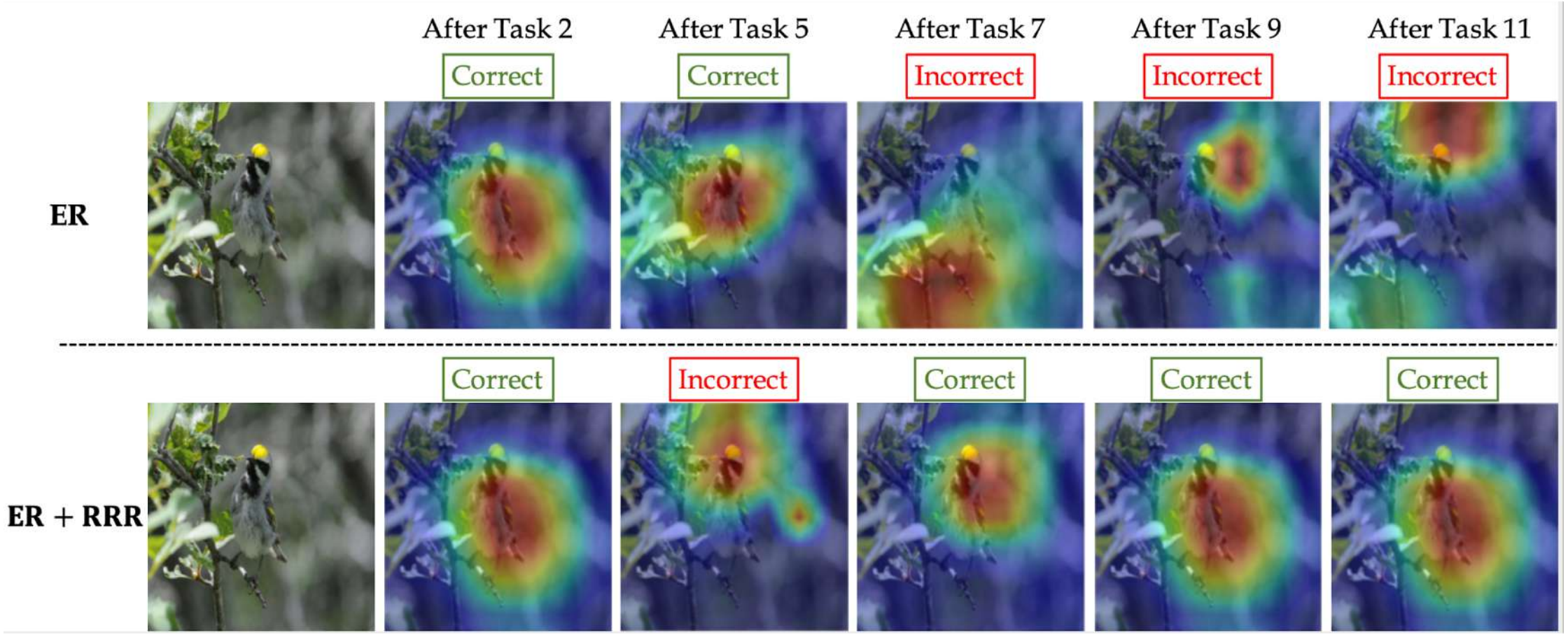


Results: Regular Class Incremental Learning



Effect on Model Explanations

Effect on Model Explanations



Explaining Model Decisions and Fixing them via Focused Feedback

http://ramprs.github.io/

rselvaraju@salesforce.com

Thank you