ICLR 2021

Remembering for the Right Reasons: Explanations Reduce Catastrophic Forgetting

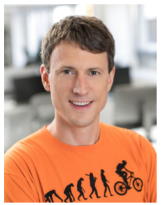

Sayna Ebrahimi
UC Berkeley


Suzanne Petryk
UC Berkeley


Akash Gokul
UC Berkeley


William Gan
UC Berkeley


Joseph Gonzalez
UC Berkeley


Marcus Rohrbach
Facebook AI Research


Trevor Darrell
UC Berkeley

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

FACEBOOK AI

Continual Learning

Definition:

learning a sequence of tasks without *catastrophic forgetting*

Continual Learning

Definition:

learning a sequence of tasks without *catastrophic forgetting*

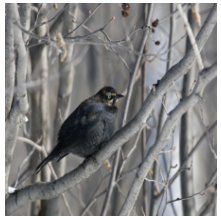
Hypothesis:

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.

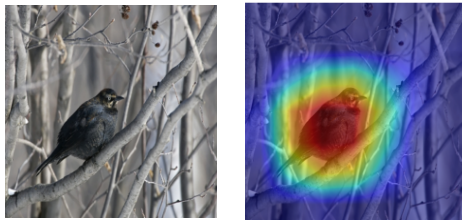


Task t

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.



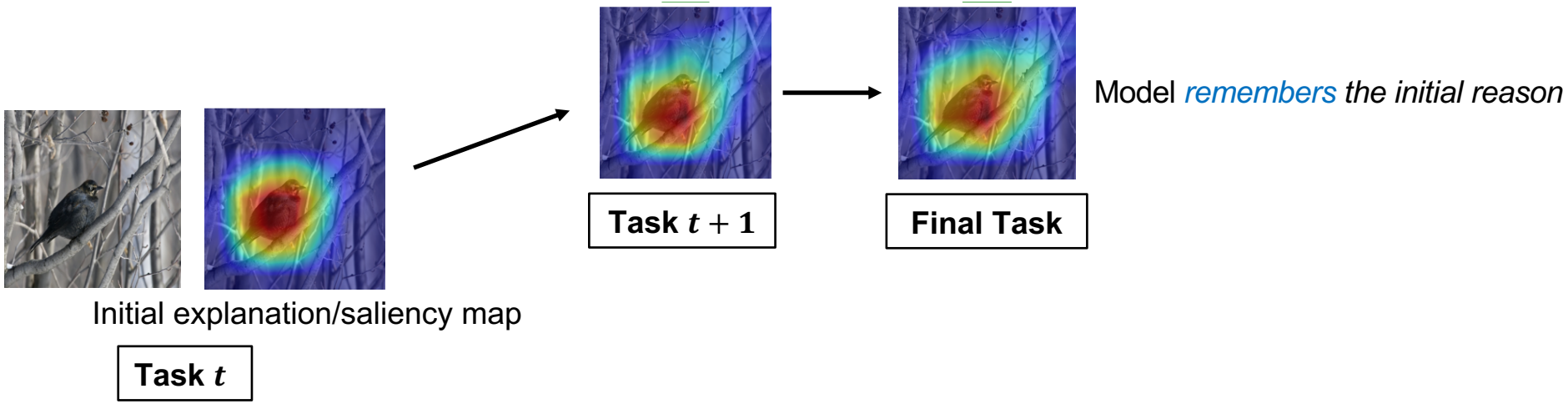
Initial explanation/saliency map

Task t

eXplainable AI for Continual Learning

Hypothesis:

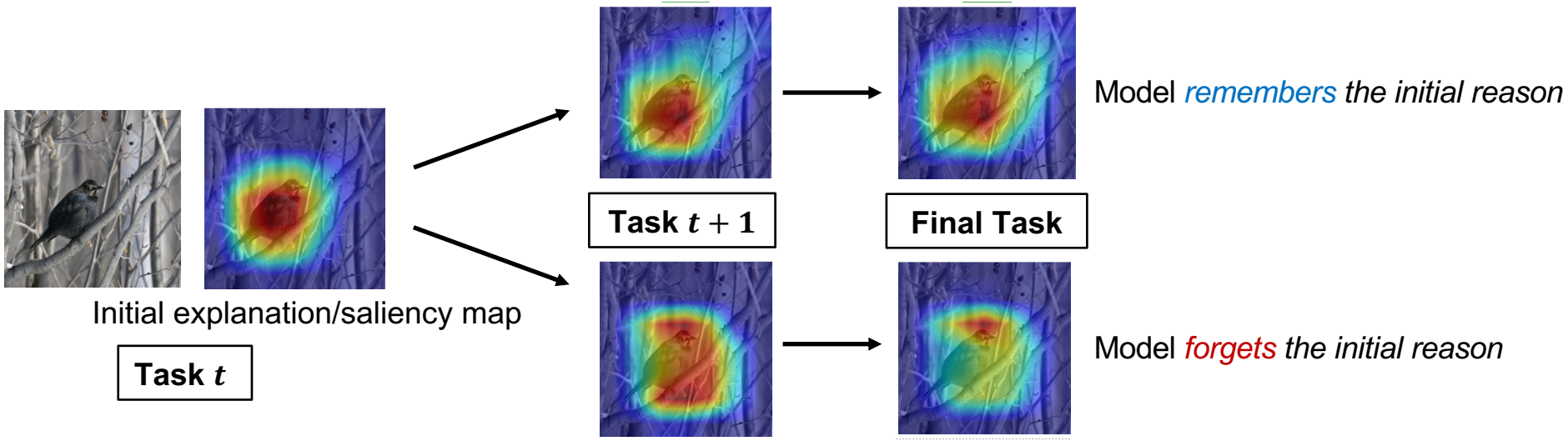
Catastrophic forgetting is due in part to not being able to rely on the **same reasoning** as was used for a previously seen observation.



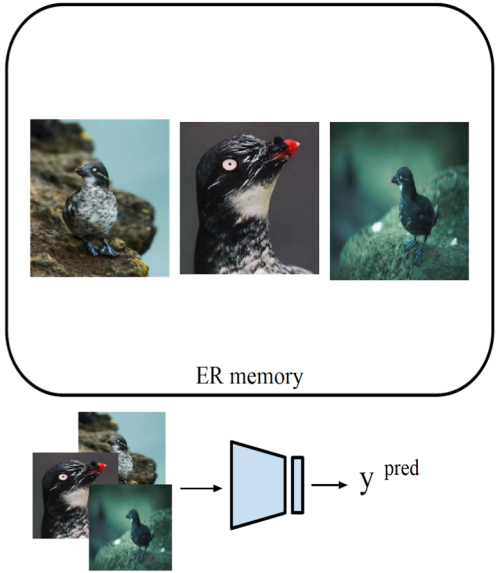
eXplainable AI for Continual Learning

Hypothesis:

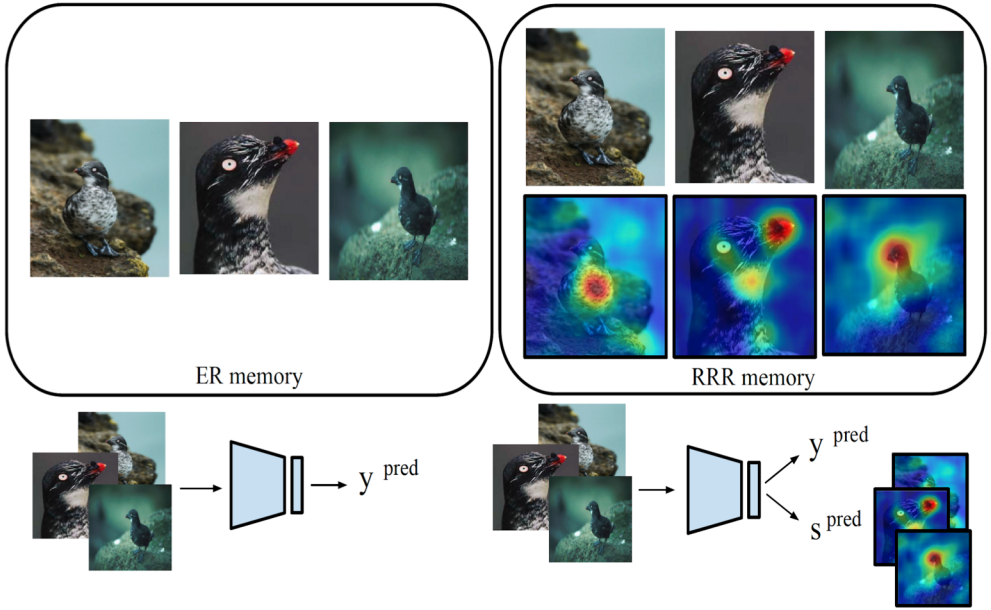
Catastrophic forgetting is due in part to not being able to rely on the **same reasoning** as was used for a previously seen observation.



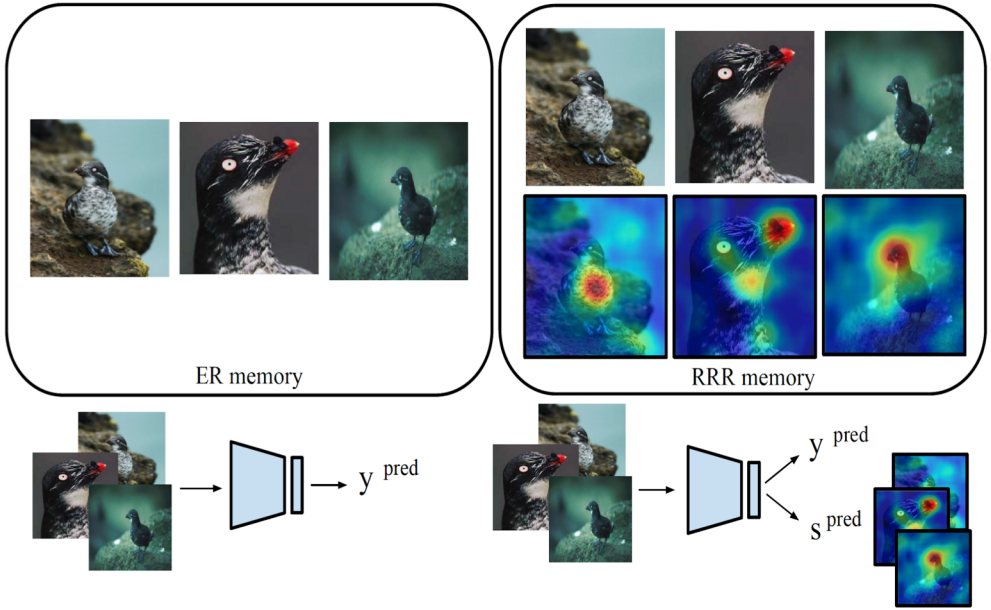
Experience Replay (ER) in CL



Remembering for the Right Reasons (RRR)

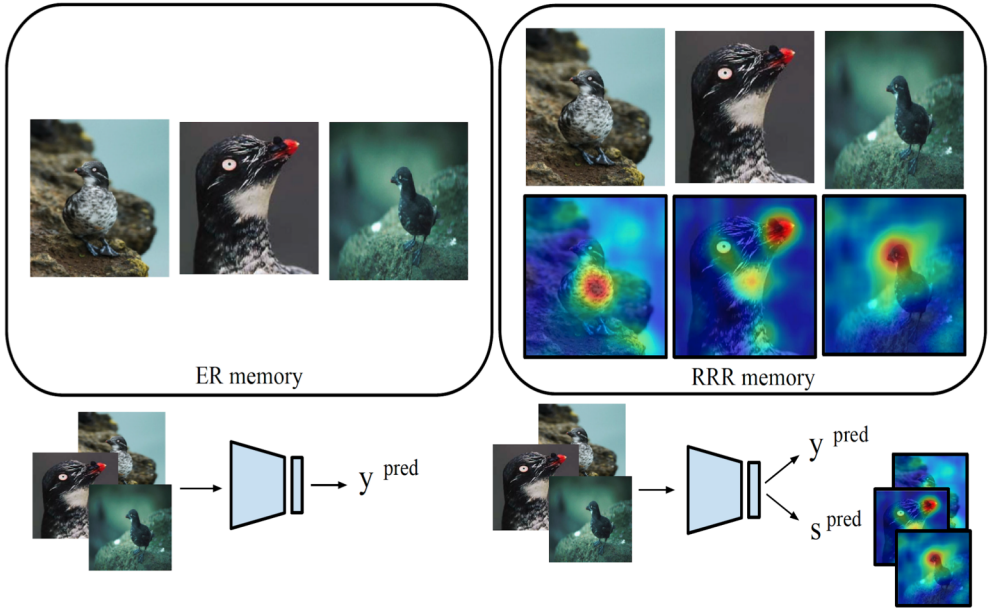


Remembering for the Right Reasons (RRR)



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

Remembering for the Right Reasons (RRR)



Saliency maps

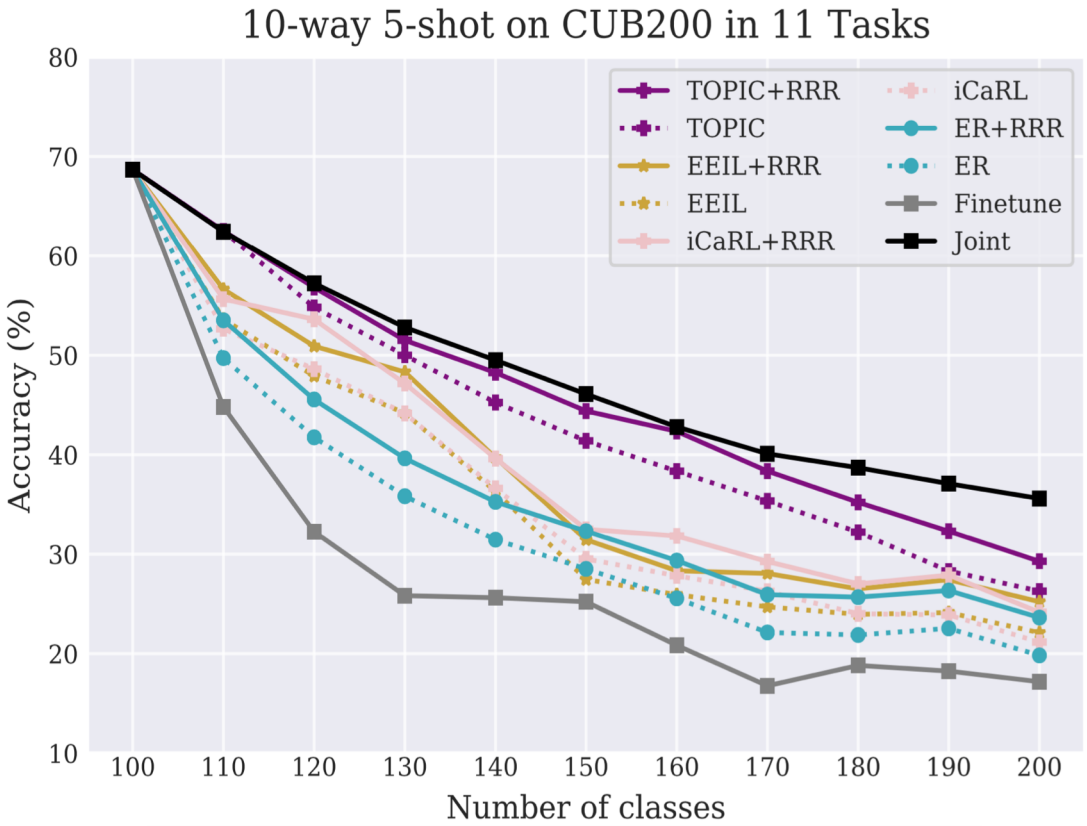
Vanilla Backprop
(Zeiler & Fergus, 2014)

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

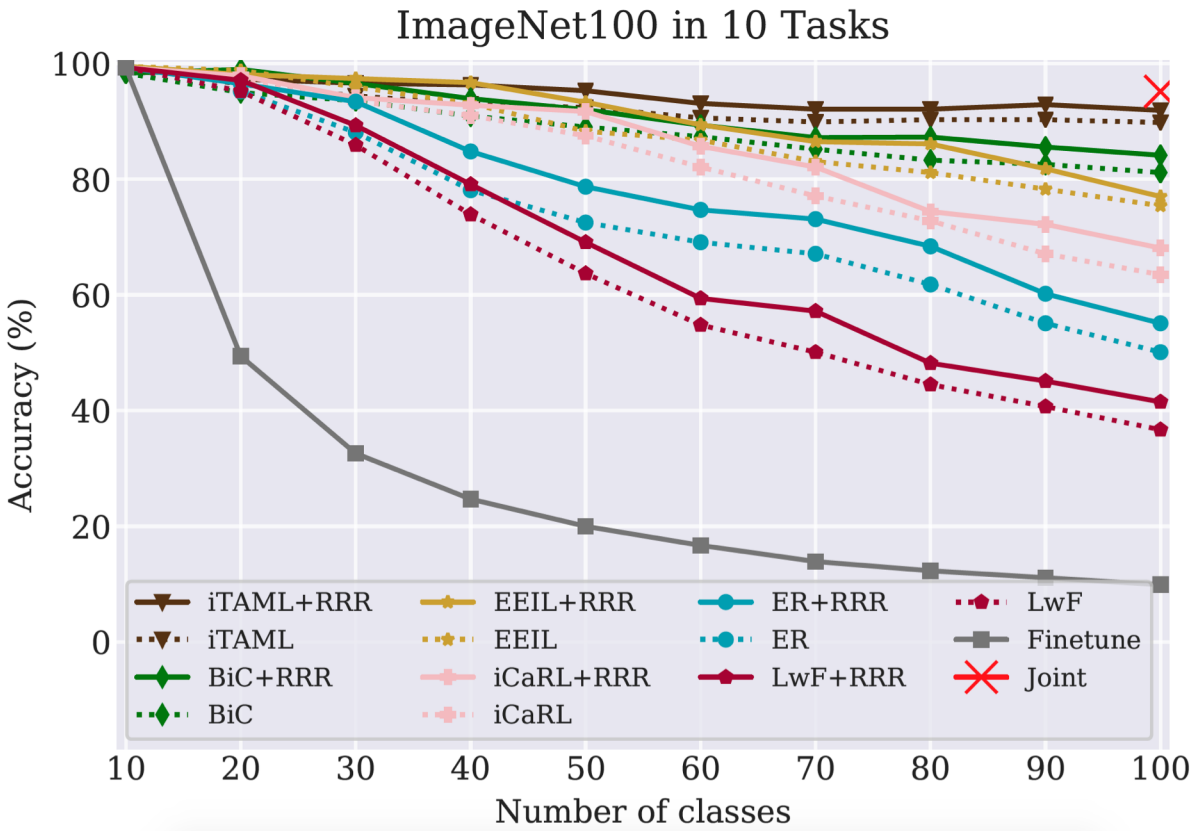
Grad-CAM
(Selvaraju et al., 2017)

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

Results: Few-Shot Class Incremental Learning



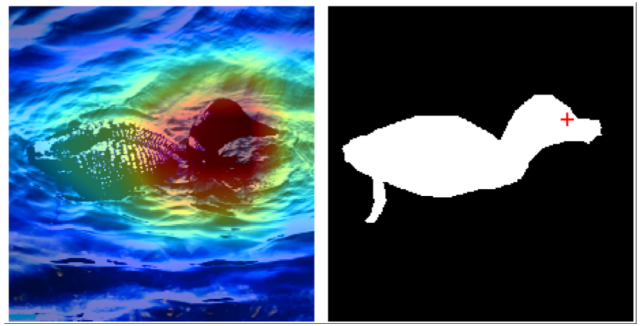
Results: Regular Class Incremental Learning



Effect on Model Explanations

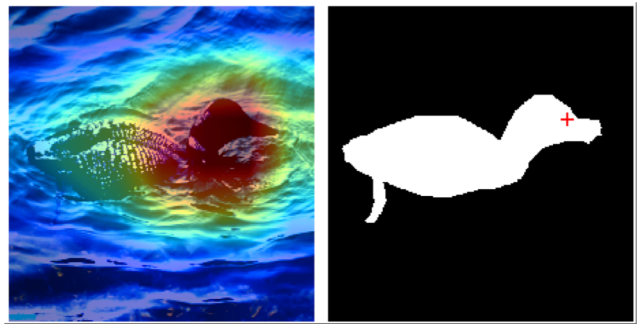
Effect on Model Explanations

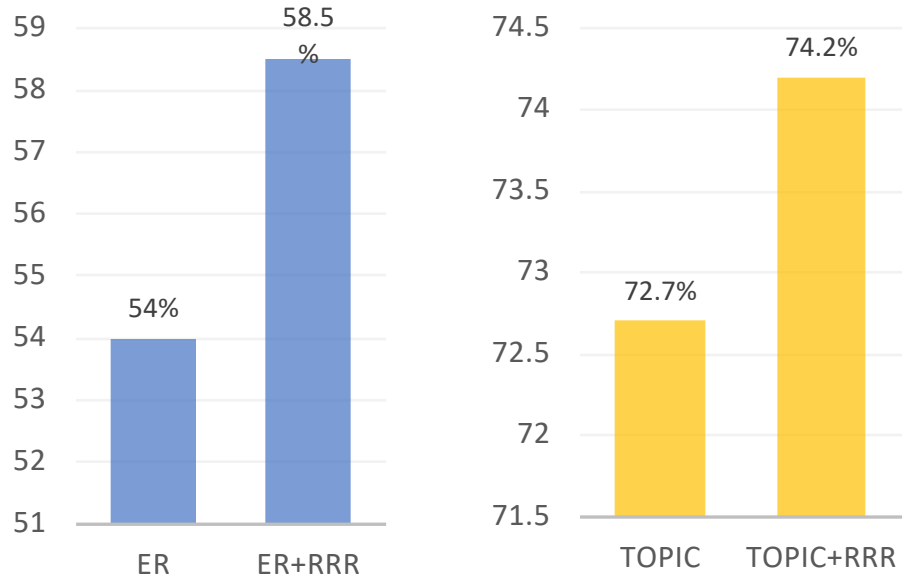
Pointing Game (PG) experiment *(Zhang et al., 2018)*



Effect on Model Explanations

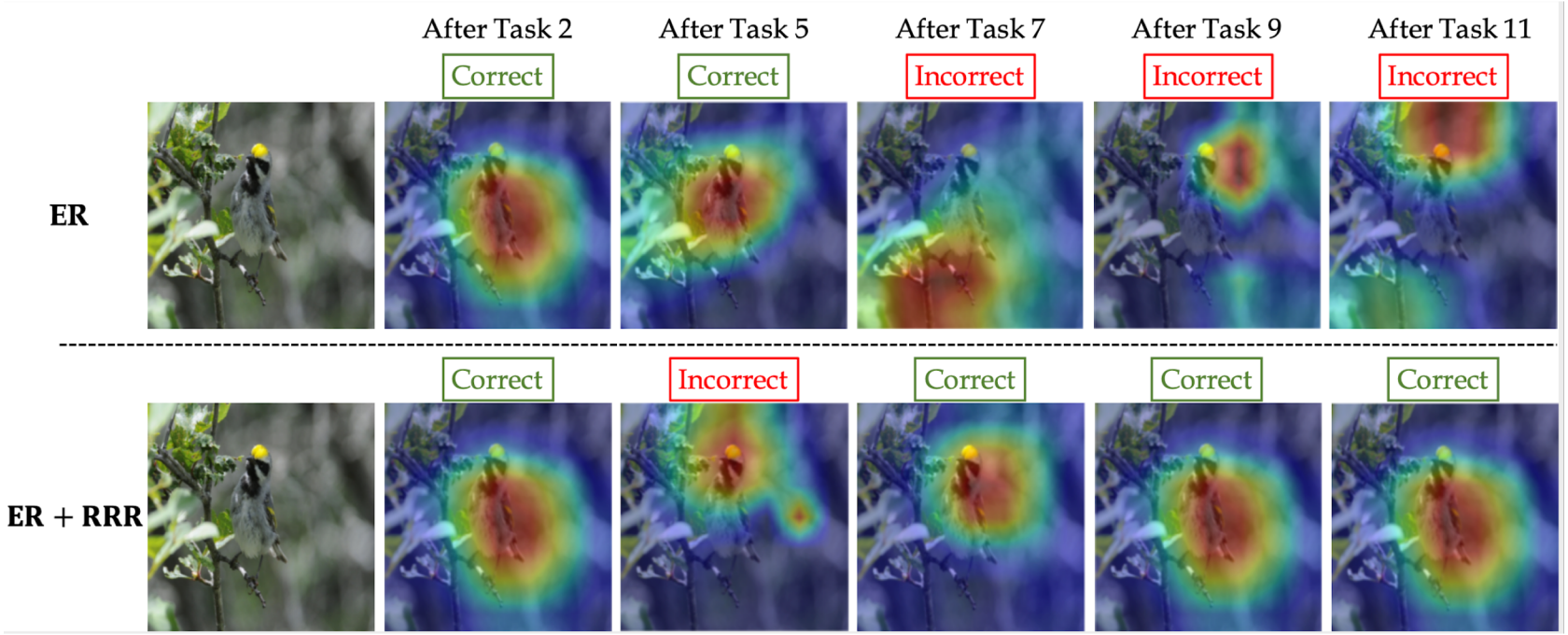
Pointing Game (PG) experiment *(Zhang et al., 2018)*





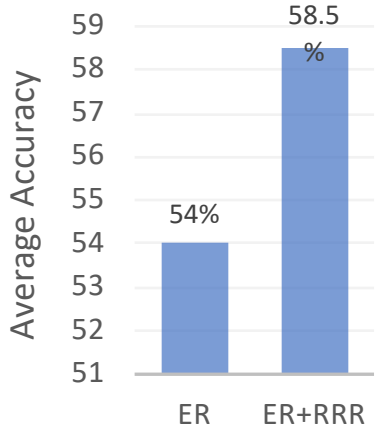
RRR Effect on Average Accuracy of Pointing Game

Effect on Model Explanations

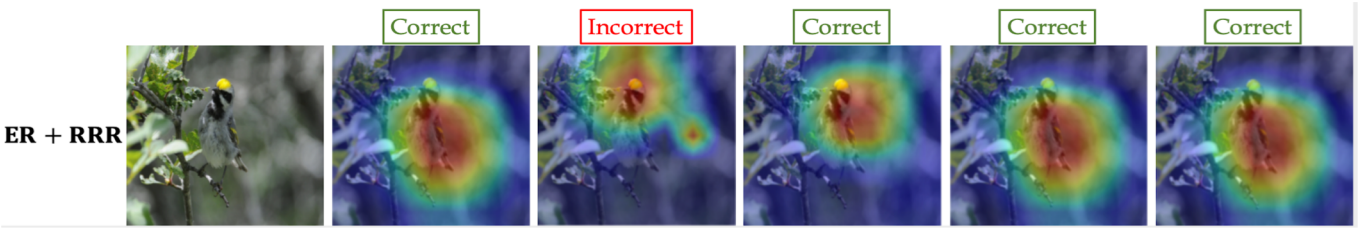


Conclusion

- Encouraging a model to remember its *evidence* will **reduce catastrophic forgetting**



- Use of eXplainable AI as part of a continual learning process makes model explanations more **accurate** and more **interpretable**



ICLR 2021

Paper: https://openreview.net/pdf?id=tHgJoMfy6nI

Code: https://github.com/SaynaEbrahimi/Remembering-for-the-Right-Reasons

Questions: sayna@berkeley.edu

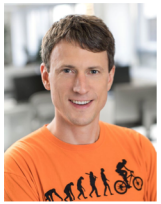
Sayna Ebrahimi
UC Berkeley

Suzanne Petryk
UC Berkeley

Akash Gokul
UC Berkeley

William Gan
UC Berkeley

Joseph Gonzalez
UC Berkeley



Marcus Rohrbach
Facebook AI Research

Trevor Darrell
UC Berkeley

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

FACEBOOK AI