Remembering for the Right Reasons: Explanations Reduce Catastrophic Forgetting

Sayna Ebrahimi[1], Suzanne Petryk[1], Akash Gokul[1], William Gan[1], Joseph Gonzalez[1], Marcus Rohrbach[2], and Trevor Darrell[1]

[1]University of California Berkeley
[2]Facebook Inc

June 25, 2021

Abstract

The goal of continual learning (CL) is to learn a sequence of tasks without suffering from the phenomenon of catastrophic forgetting. Previous work has shown that leveraging memory in the form of a replay buffer can reduce performance degradation on prior tasks. We hypothesize that forgetting can be further reduced when the model is encouraged to remember the *evidence* for previously made decisions. As a first step towards exploring this hypothesis, we propose a simple novel training paradigm, called Remembering for the Right Reasons (RRR), that additionally stores visual model explanations for each example in the buffer and ensures the model has "the right reasons" for its predictions by encouraging its explanations to remain consistent with those used to make decisions at training time. Without this constraint, there is a drift in explanations and increase in forgetting as conventional continual learning algorithms learn new tasks. We demonstrate how RRR can be easily added to any memory or regularization-based approach and results in reduced forgetting, and more importantly, improved model explanations. We have evaluated our approach in the standard and few-shot settings and observed a consistent improvement across various CL approaches using different architectures and techniques to generate model explanations and demonstrated our approach showing a promising connection between explainability and continual learning. Our code is available at \url{https://github.com/SaynaEbrahimi/Remembering-for-the-Right-Reasons}

Hosted file

`Remembering_for_the_Right_Reasons__Explanations_Reduce_Catastrophic_Forgetting.pdf` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

Hosted file

`math_commands.tex` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

Hosted file

`wileyNJD-AMA.tex` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

Hosted file

`appendix.tex` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

Hosted file

`wileyNJD-AMA.bib` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

Hosted file

`WileyNJD-v2.cls` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

Hosted file

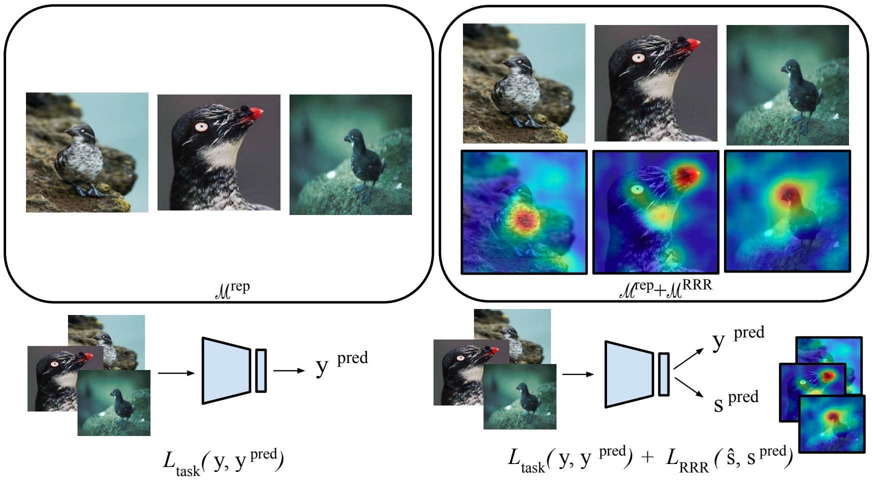
`NJDnatbib.sty` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

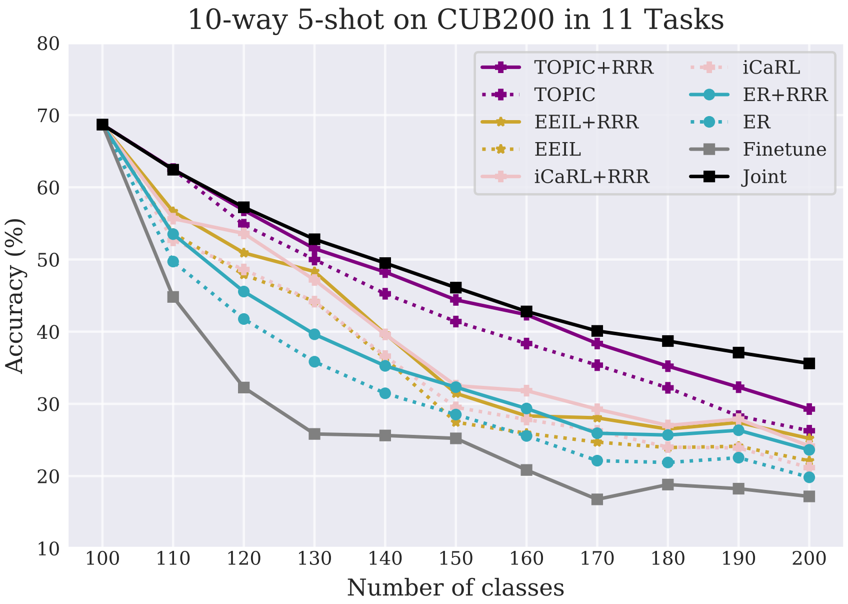
Hosted file

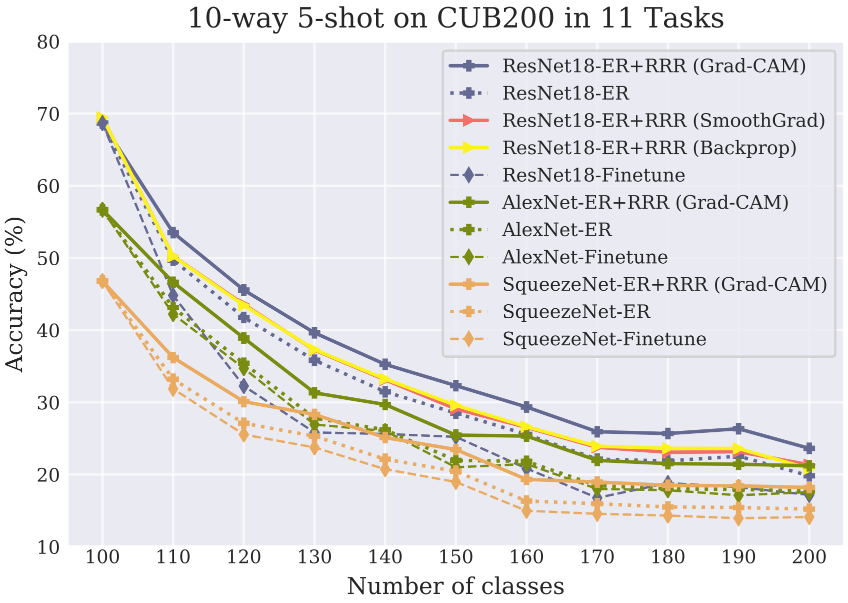
`WileyNJD-AMA.bst` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

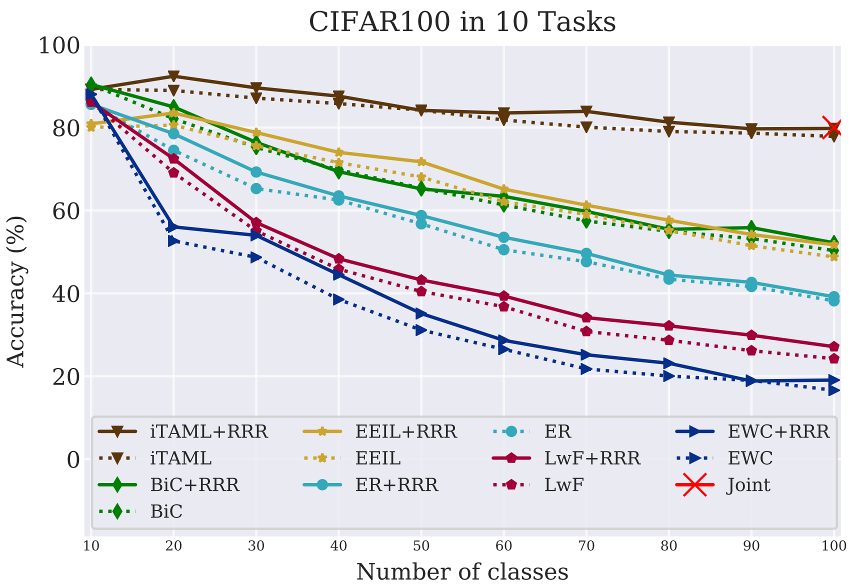
Hosted file

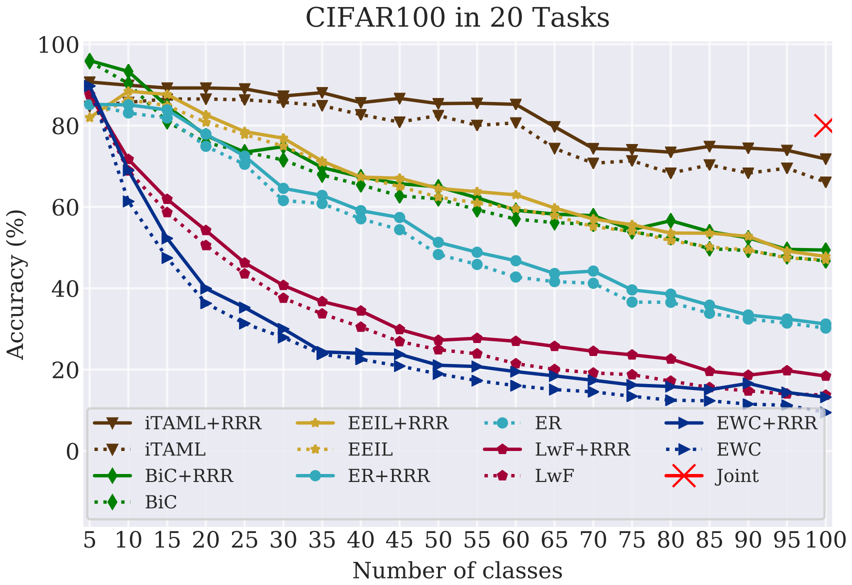
`wileyNJD-AMA.pag` available at https://authorea.com/users/422048/articles/527830-remembering-for-the-right-reasons-explanations-reduce-catastrophic-forgetting

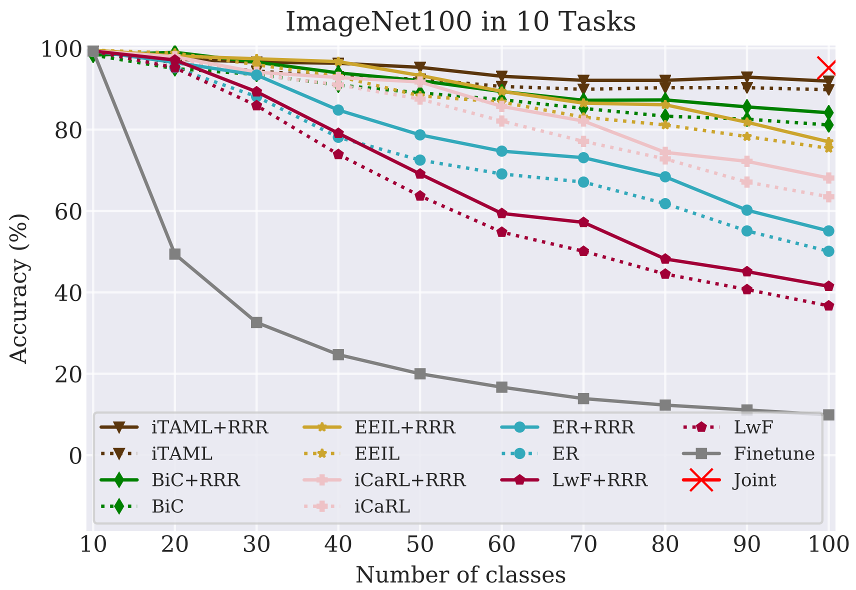

$\mathcal{M}^{\text{rep}}$
$\mathcal{M}^{\text{rep}}+\mathcal{M}^{\text{RRR}}$
y pred
$L_{\text{task}}(\text{y}, \text{y}^{\text{pred}})$
y pred
s pred
$L_{\text{task}}(\text{y}, \text{y}^{\text{pred}}) + L_{\text{RRR}}(\hat{\text{s}}, \text{s}^{\text{pred}})$

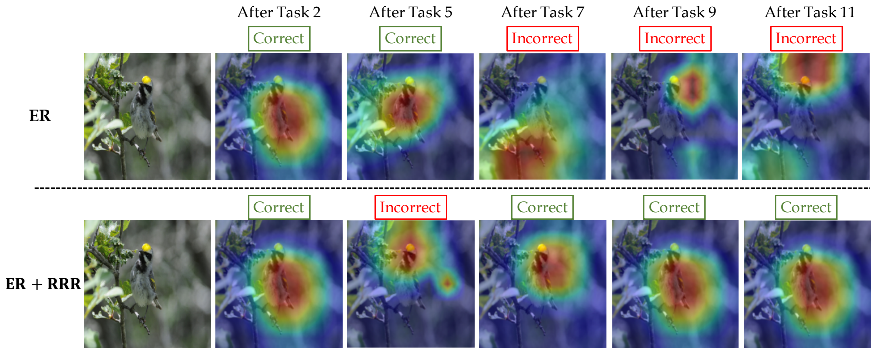

10-way 5-shot on CUB200 in 11 Tasks


10-way 5-shot on CUB200 in 11 Tasks


CIFAR100 in 10 Tasks
Accuracy (%)
100
80
60
40
20
0
10
20
30
40
50
60
70
80
90
100
Number of classes
iTAML+RRR
iTAML
BiC+RRR
BiC
EEIL+RRR
EEIL
ER+RRR
ER
LwF+RRR
LwF
EWC+RRR
EWC
Joint


CIFAR100 in 20 Tasks
Accuracy (%)
100
80
60
40
20
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Number of classes
iTAML+RRR
iTAML
BiC+RRR
BiC
EEIL+RRR
EEIL
ER+RRR
ER
LwF+RRR
LwF
EWC+RRR
EWC
Joint


ImageNet100 in 10 Tasks
Accuracy (%)
0
20
40
60
80
100
Number of classes
10
20
30
40
50
60
70
80
90
100
iTAML+RRR
iTAML
BiC+RRR
BiC
EEIL+RRR
EEIL
iCaRL+RRR
iCaRL
ER+RRR
ER
LwF+RRR
LwF
Finetune
Joint


After Task 2
After Task 5
After Task 7
After Task 9
After Task 11
Correct
Correct
Incorrect
Incorrect
Incorrect
ER
Correct
Incorrect
Correct
Correct
Correct
ER + RRR