Adversarially Diversified Rehearsal Memory (ADRM): Mitigating Memory Overfitting Challenge in Continual Learning

Dr. Hikmat Khan (Presenter)
Department of Electrical and Computer Engineering,
Rowan University,
Glassboro, NJ, USA

Dr. Nidhal Carla Bouaynaya
Department of Electrical and Computer Engineering,
Rowan University,
Glassboro, NJ, USA

Dr. Ghulam Rasool
Department of Machine Learning,
Moffitt Cancer Center and Research Institute,
Tampa, FL, USA

IEEE WCCI 2024 Yokohama Japan

Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

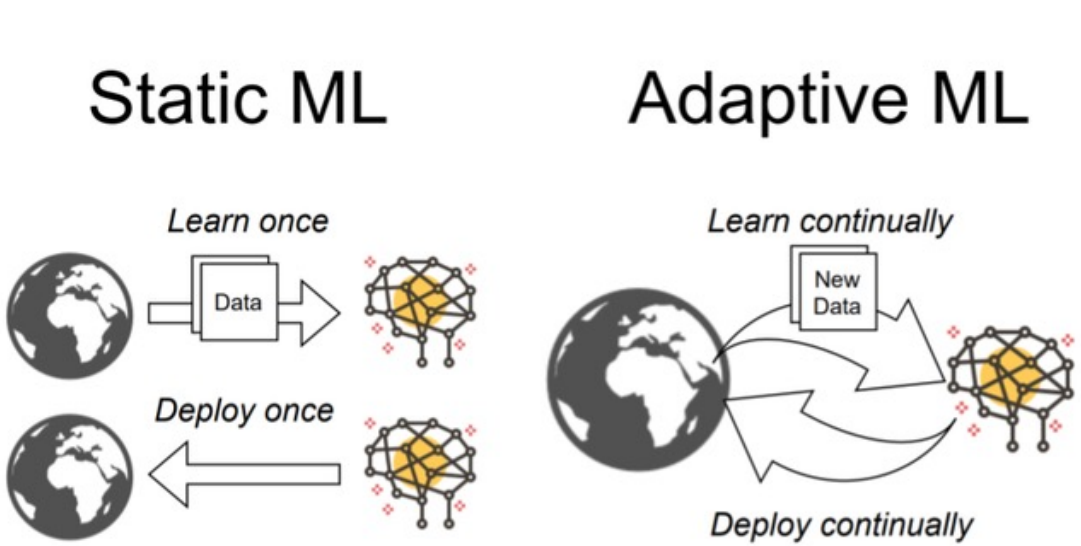
Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Deep Learning – The Promise

- Deep Learning (DL)/Artificial Intelligence (AI) models have achieved or even surpassed human-level accuracy in several areas, including computer vision and pattern recognition.

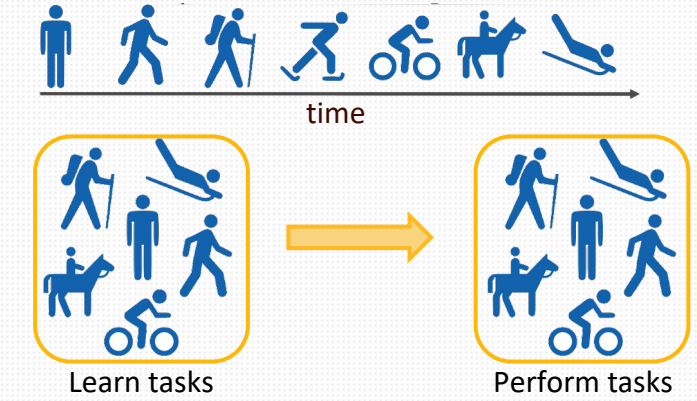
The World is Non-Stationary [5]

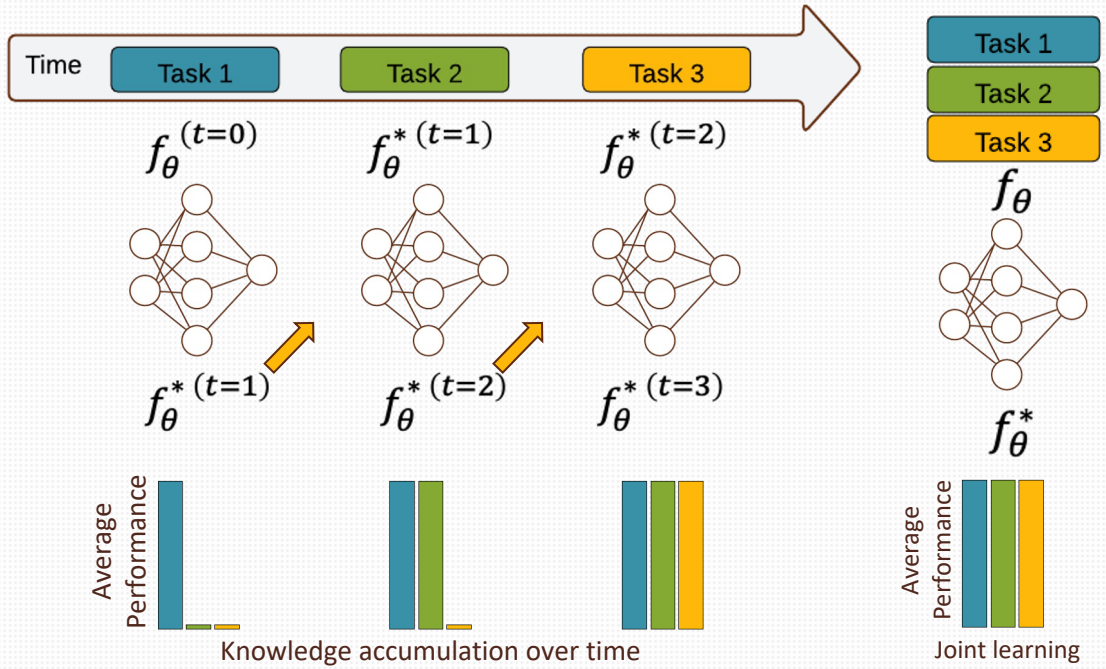


https://medium.com/gooroo-official/if-youre-not-learning-you-re-dying-b2b4d54eace4 | https://senicessm.live/product_details/47625942.html

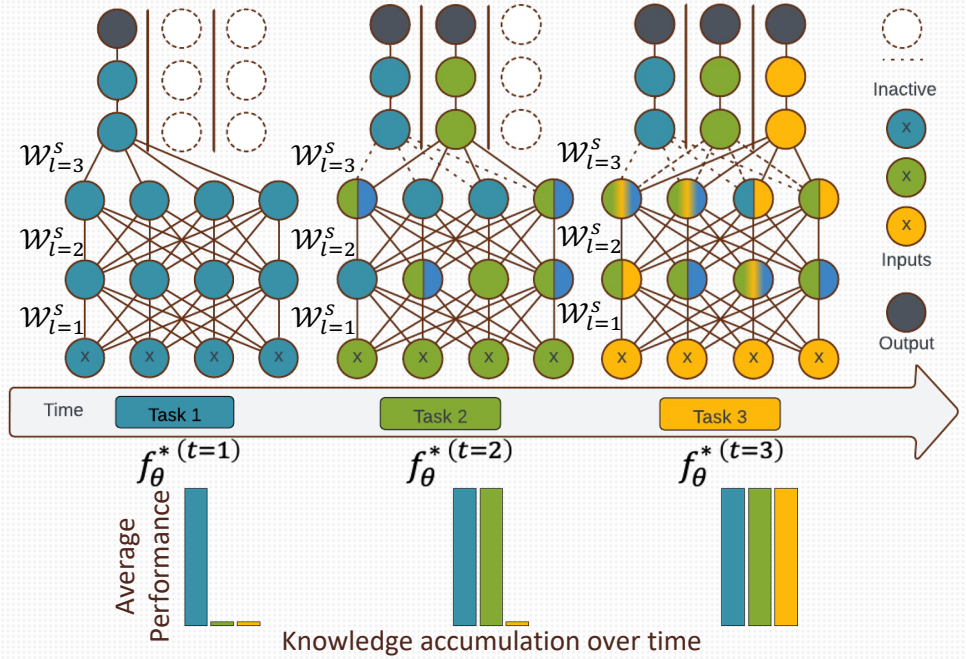
Continual Learning (CL)

- Learning tasks over time[1].
  - The tasks can be any classification or segmentation.
- Offline learning [1, 2].
  - Learn all tasks simultaneously.



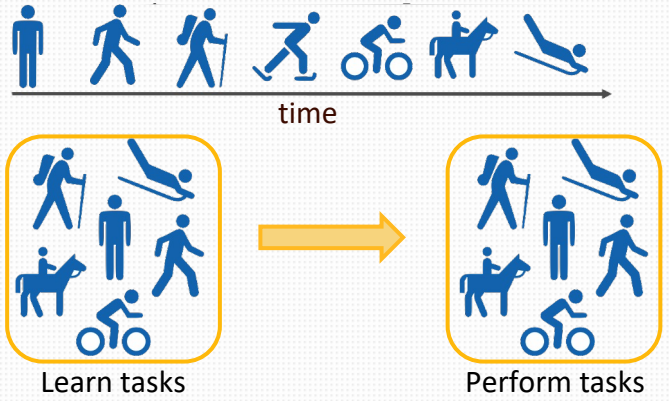


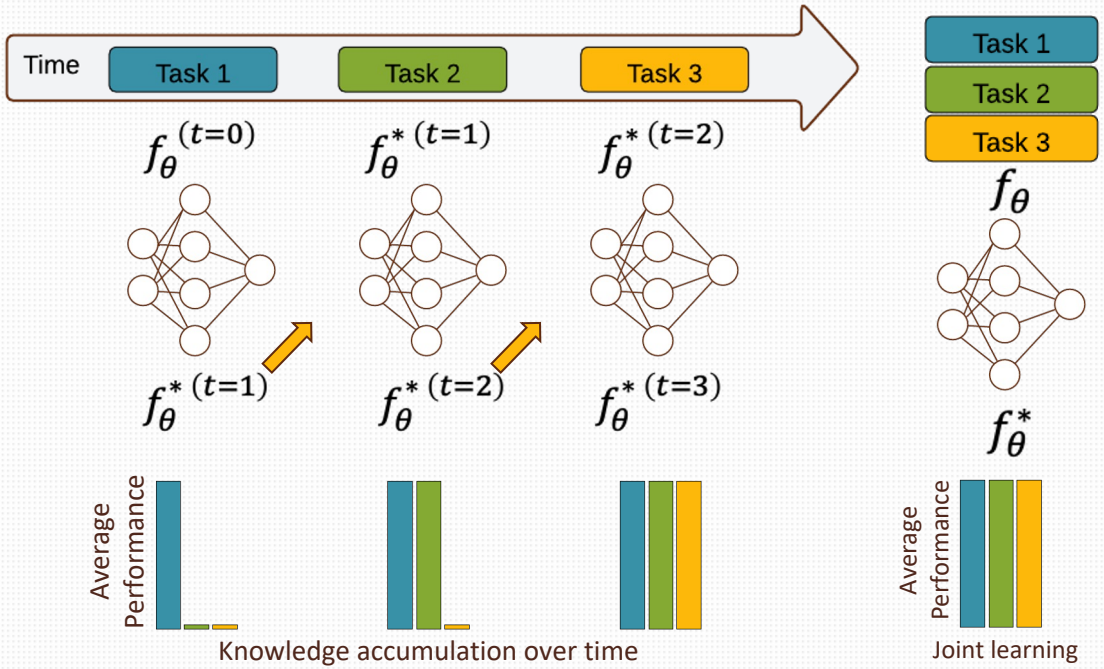
Task Incremental Learning [2]



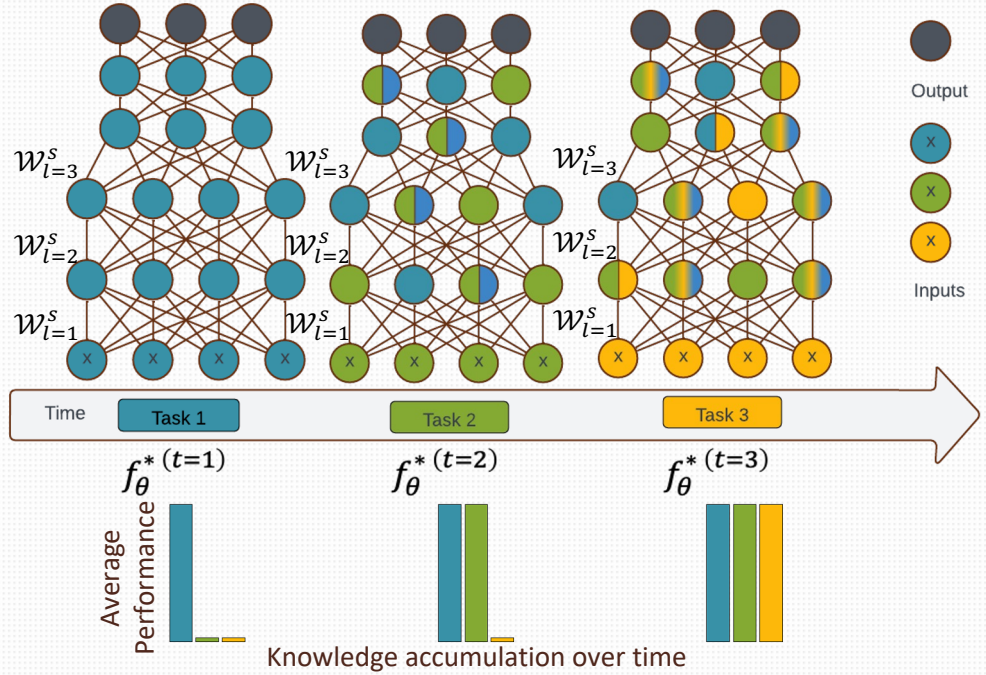
Continual Learning (CL) (Continued)

- Learning tasks over time[1].
  - The tasks can be any classification or segmentation.
- Offline learning [1, 2].
  - Learn all tasks simultaneously.



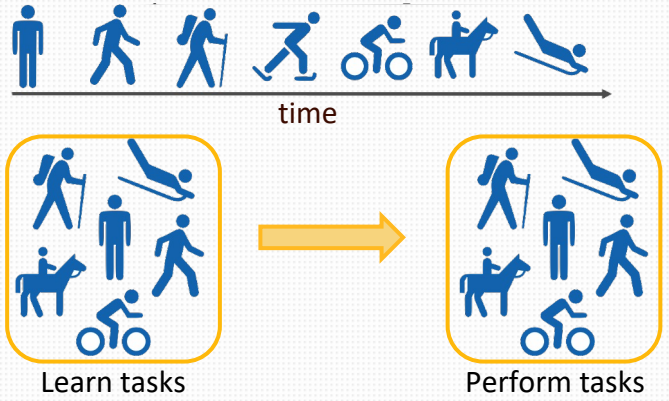


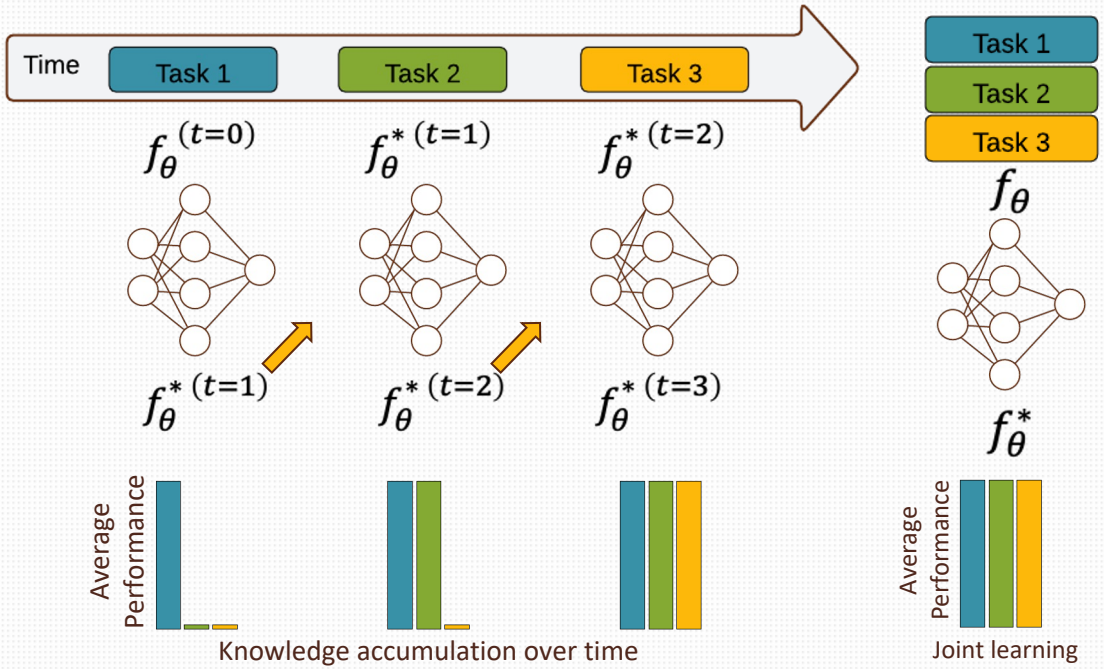
Domain Incremental Learning [2]



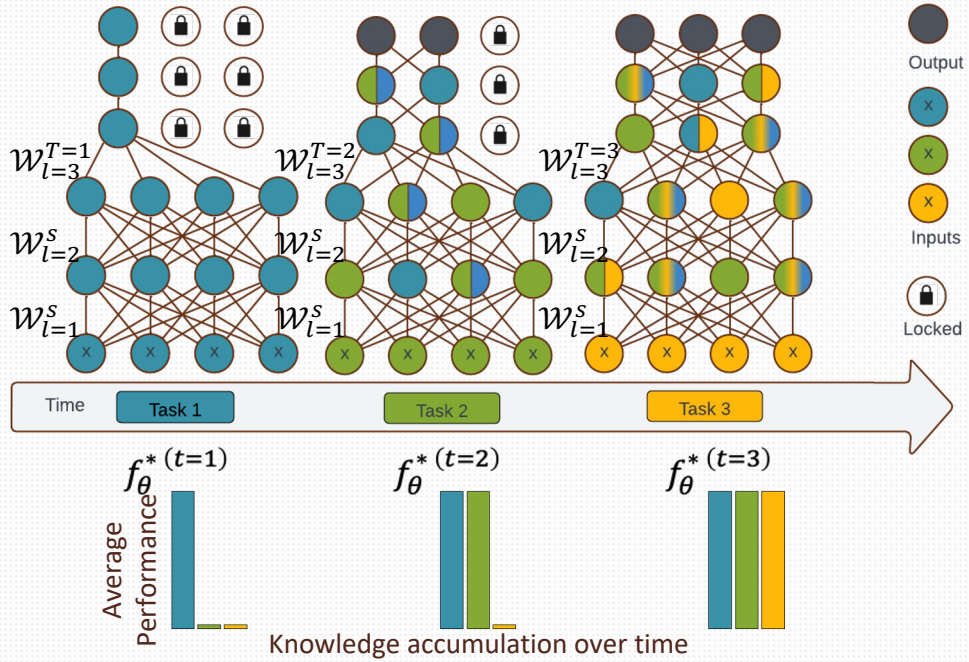
Continual Learning (CL) (Continued)

- Learning tasks over time[1].
  - The tasks can be any classification or segmentation.
- Offline learning [1, 2].
  - Learn all tasks simultaneously.



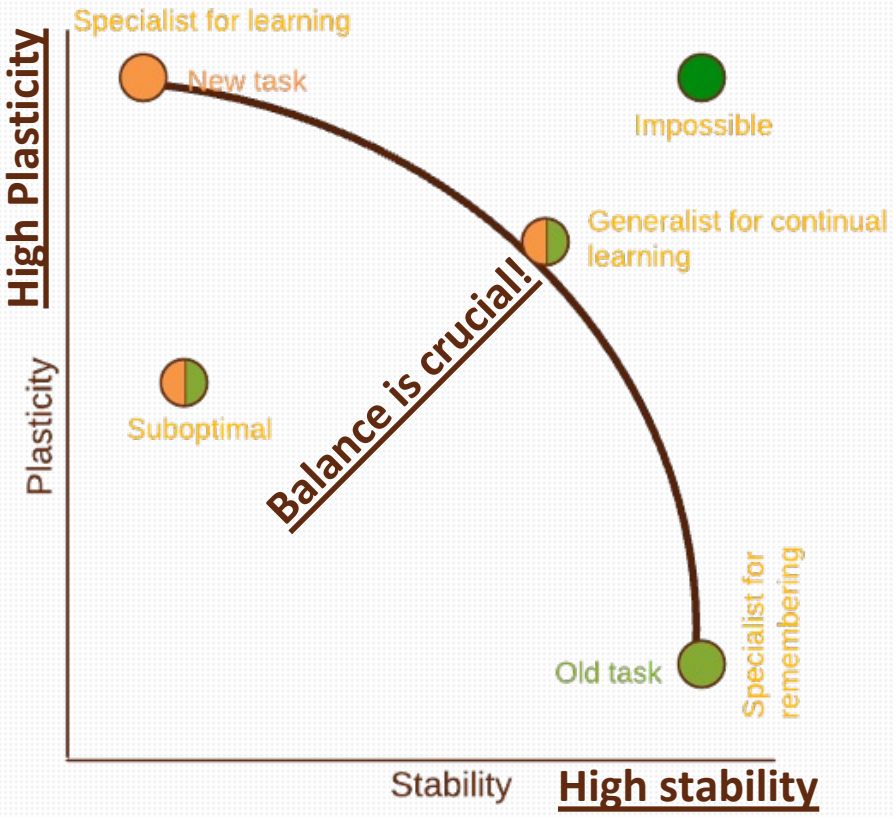


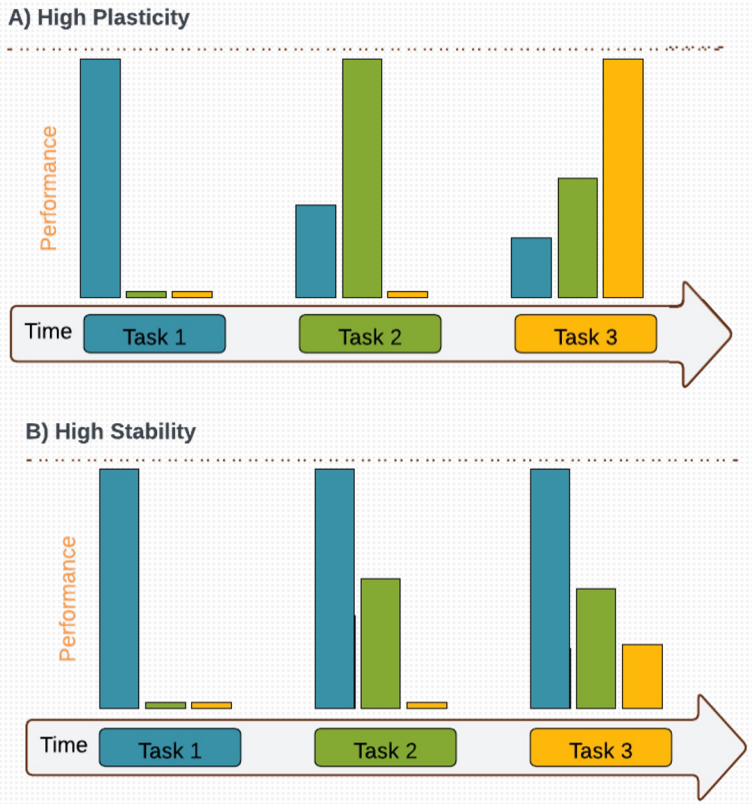
Class Incremental Learning [2]



The Challenges Stopping CL

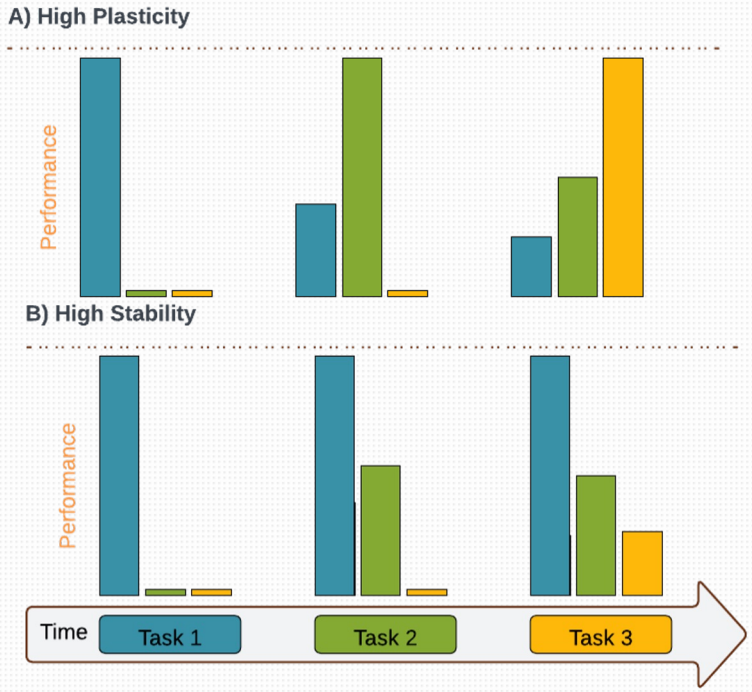
- Plasticity-stability dilemma [3].
  - The plasticity-stability dilemma is a fundamental challenge in both neuroscience and deep learning design.





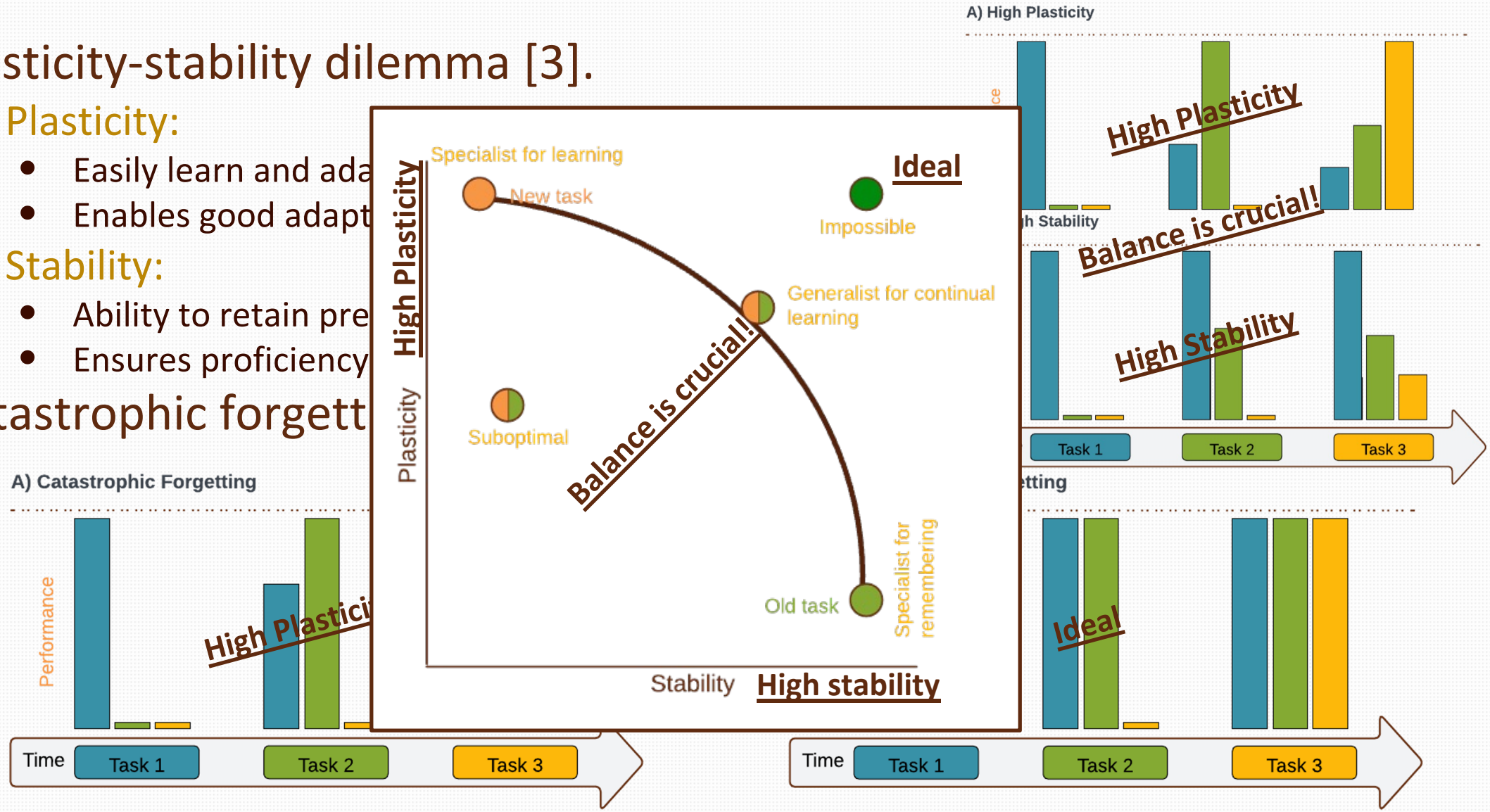
The Challenges Stopping CL (Continued)

- Plasticity-stability dilemma [3].
  - Plasticity:
    - Easily learn and adapt to new experiences.
    - Enables good adaptation.
  - Stability:
    - Ability to retain previously learned experiences.
    - Ensures proficiency in solving old tasks.
- Catastrophic forgetting [4].
  - If a system is too plastic, it risks overriding old knowledge (aka catastrophic forgetting)
  - A phenomenon where a model forgets previously learned knowledge after acquiring new knowledge.



The Challenges Stopping CL (Continued)

- Plasticity-stability dilemma [3].
  - Plasticity:
    - Easily learn and ada
    - Enables good adapt
  - Stability:
    - Ability to retain pre
    - Ensures proficiency
- Catastrophic forgett

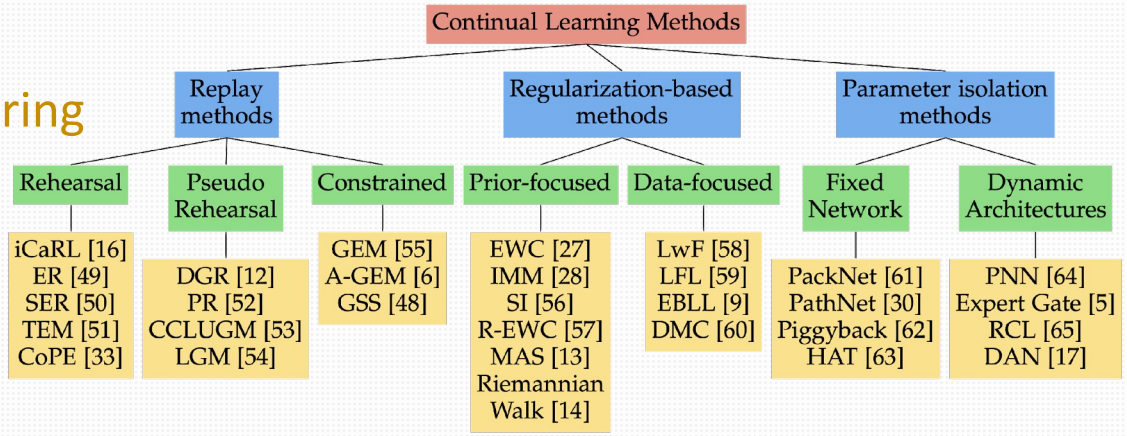


Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Related Work

1. Rehearsal-based Approaches [1, 5]
   - Store a subset of the old data and rehearse it during the learning of the subsequent tasks.
2. Regularization-based Approaches [1, 5]
   - Penalize (important) parameter updates.
3. Architectural-based Approaches [1, 5]
   - Increase the model capacity for every new task.
   - Explicitly identify important parameters for each task.
4. Knowledge Distillation-based Approaches [1, 5]
   - Take inspiration from the Knowledge Distillation.
   - Use the previous task model as a teacher.
5. Dataset Distillation-based Approaches [1, 5]
   - Create representative subsets that can be rehearsed during the learning of the subsequent tasks.
6. Brain-inspired Approaches [1, 5]
   - Incorporate inspiration from neuroscience in designing the CL
7. Hybrid Approaches [1, 5]
   - Combine multiple approaches to create more robust solutions against catastrophic forgetting.



Taxonomy of CL approaches [1]

Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Problem Statement

- Rehearsal-based approaches are widely adopted to mitigate catastrophic forgetting in CL models.
- However, these approaches suffer from rehearsal memory overfitting, where the CL model becomes overly specialized on a limited set of memory samples, resulting in catastrophic forgetting.
- Leading to a progressive decay in the effectiveness of the rehearsal memory, ultimately causing the model to forget previously learned tasks.
- **Question!**
  - Can rehearsal memory be diversified to prevent rehearsal memory overfitting in the CL model?

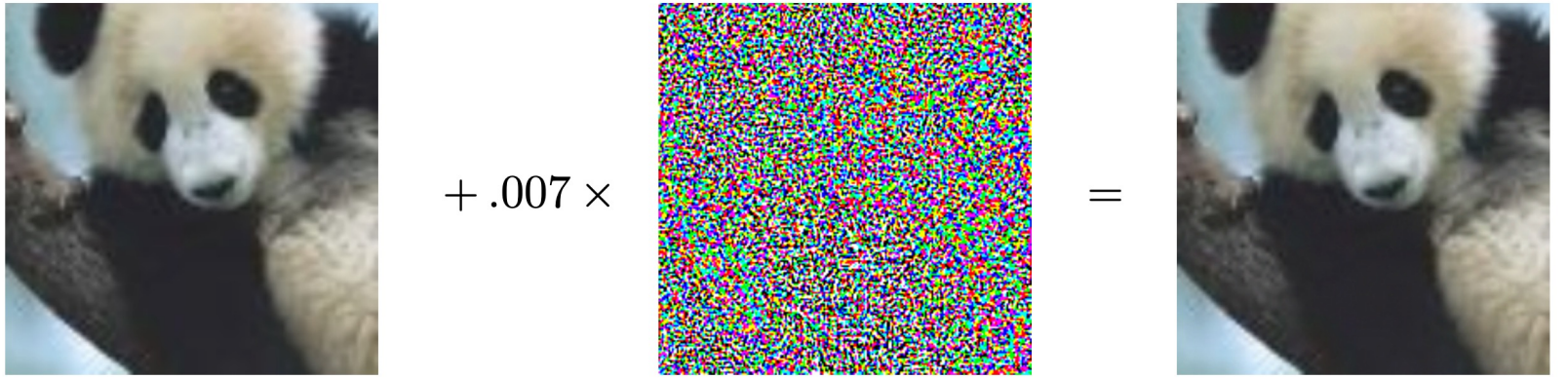
Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

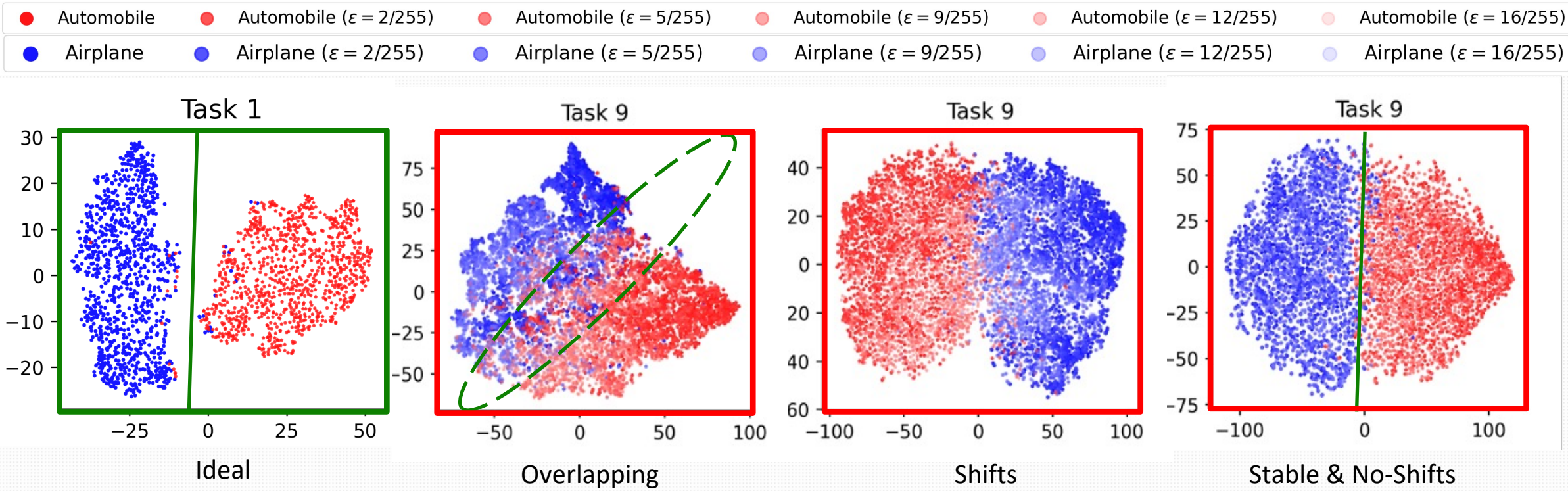
Adversarial Attacks (Quick Refresher)

- Fast Gradient Sign Method (FGSM) [10]

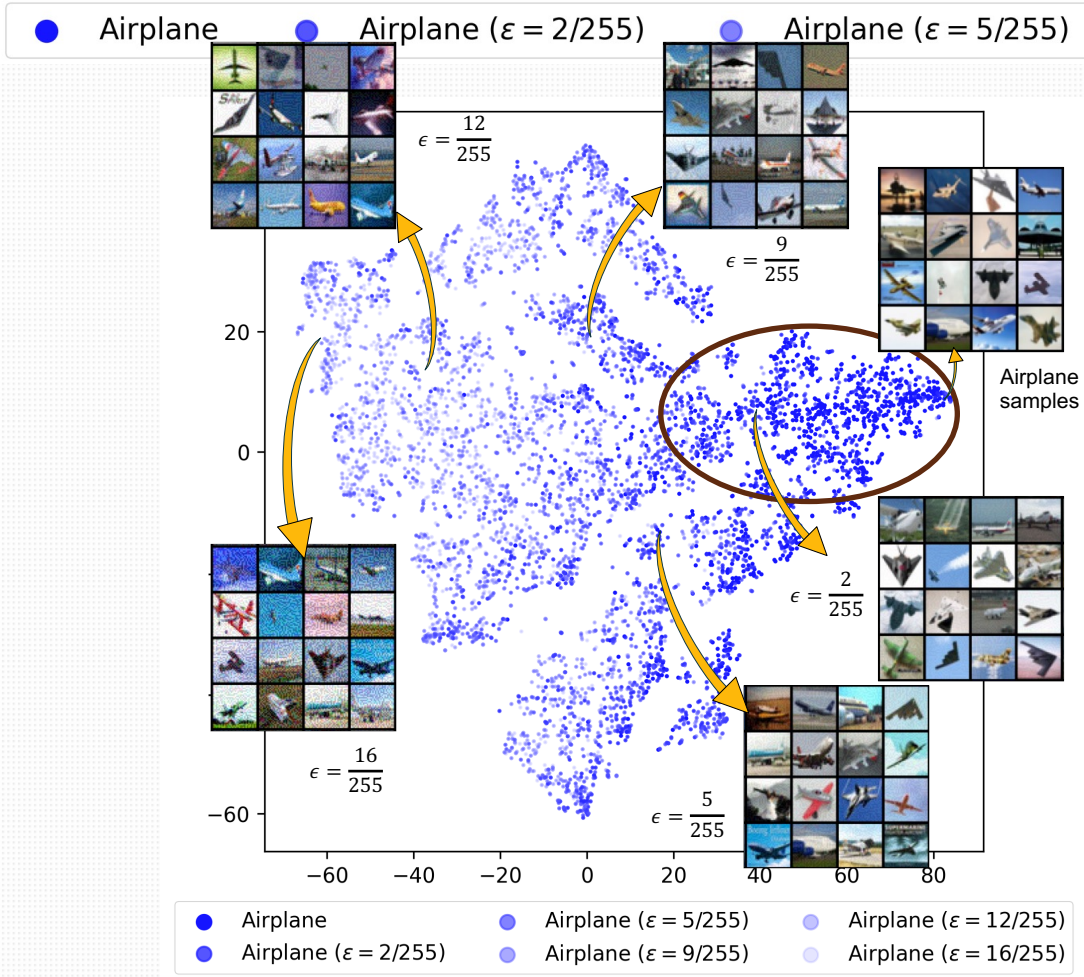
$$x_{adv} = x + \epsilon * sign(\Delta_x \, J(\theta, x, y))$$



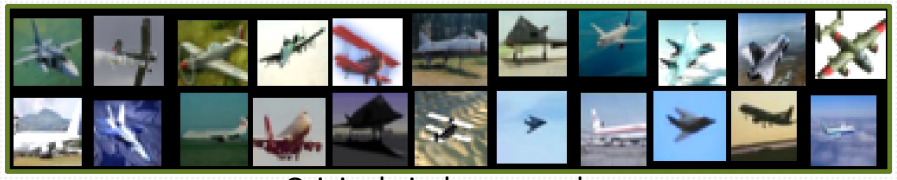
Idea Illustration: Adversarially Diversified Rehearsal Memory



Idea Illustration: Adversarially Diversified Rehearsal Memory

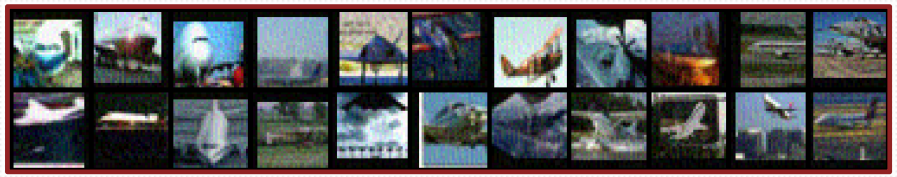


t-SNE visualization showing the dispersion of 'airplane' class features under increasing adversarial perturbations



Original airplane samples

Correctly classified adversarially diversified airplane samples



Incorrectly classified adversarially diversified airplane samples

Adversarially Diversified Rehearsal Memory (ADRM)

- Rehearsal-Based Continual Learning

$$\min_{\theta \in \Theta} \left[\mathbb{E}_{(x_t, y_t) \sim \mathcal{D}_t} \overbrace{\mathcal{L}(\theta, x_t, y_t)}^{\text{Current task}} + \mathbb{E}_{(x_m, y_m) \sim \mathcal{M}_t} \overbrace{\mathcal{L}(\theta, x_m, y_m)}^{\text{Memory}} \right]$$

- Aims to increase the complexity of memory samples using FGSM [10]

$$x_{diversified} = x_m + \epsilon \,.\, sign(\nabla_{x_m} J(\theta,\ x_m,\ y_m))$$

- Benefits of ADRM:
  - Prevents rehearsal memory overfitting.
  - Maintains the effectiveness of rehearsal memory samples.
  - Enables CL models to learn new tasks without performance degradation.

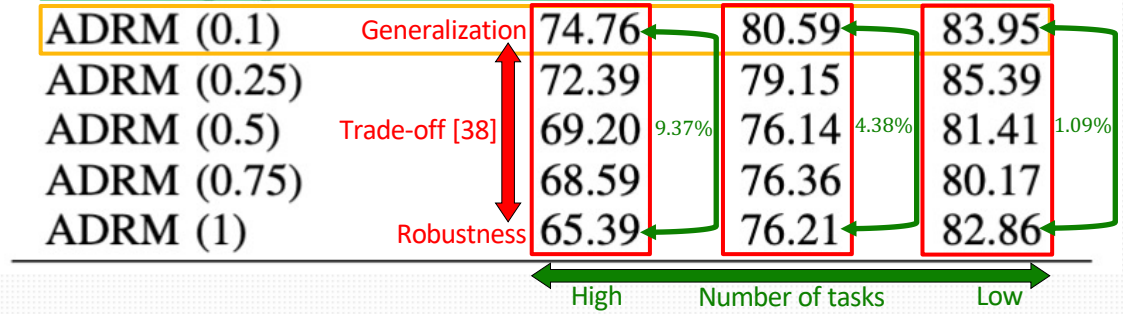
Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Results and Baselines

- The ADRM outperforms several established CL approaches and achieves comparable results to state-of-the-art CL approaches.
- ADRM, when rehearsed with 10% adversarially diversified memory, demonstrates superior performance over variants incorporating 25%, 50%, 75%, and 100% diversified memory.
- The 10% adversarially diversified memory suggests that a minimal level of adversarial diversification, can prevent the rehearsal memory overfitting in CL.

CL Methods	Split-CIFAR10		
	9 steps	5 steps	2 steps
Joint		94.8	
Fine-tune	11.11	18.54	45.31
Experience Replay [6]	67.62	75.97	80.26
iCaRL [33]	69.25	74.85	79.98
BiC [26]	53.11	71.01	86.57
PODNet [34]	72.41	76.96	87.64
WA [25]	75.56	81.67	85.34
DER [35]	74.33	78.14	82.57
SimpleCIL [27]	52.67	54.54	75.01
FOSTER [36]	74.61	80.40	83.89
FETRIL [37]	65.26	67.51	84.52
MEMO [38]	80.60	85.93	87.81
ADRM (0.1)	74.76	80.59	83.95
ADRM (0.25)	72.39	79.15	85.39
ADRM (0.5)	69.20	76.14	81.41
ADRM (0.75)	68.59	76.36	80.17
ADRM (1)	65.39	76.21	82.86



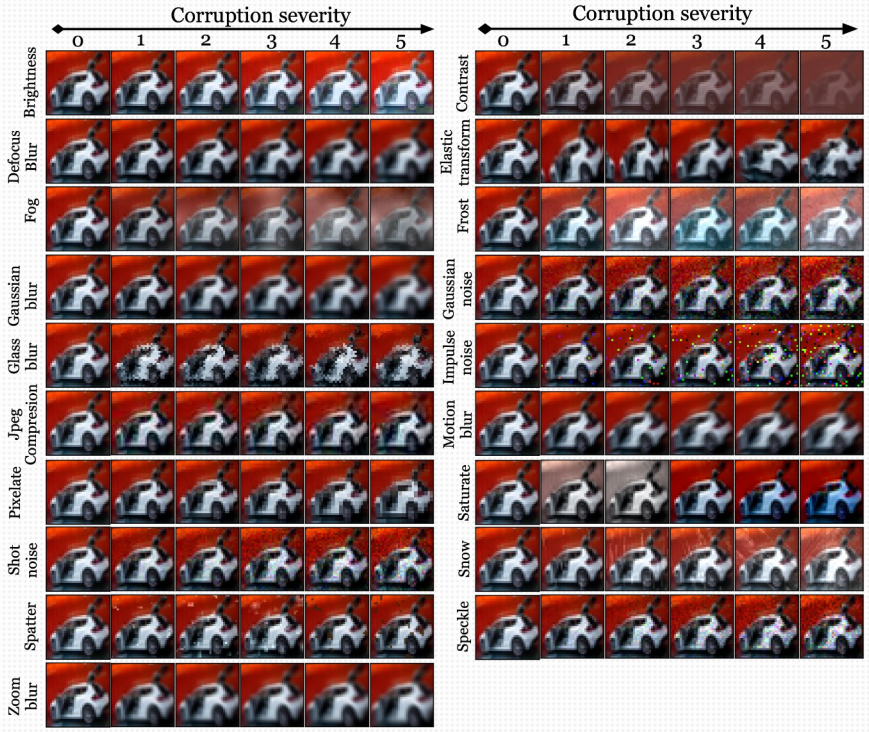
Comparative performance of CL methods on the Split-CIFAR10 dataset over 9 steps, 5 steps, and 2 steps.

Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

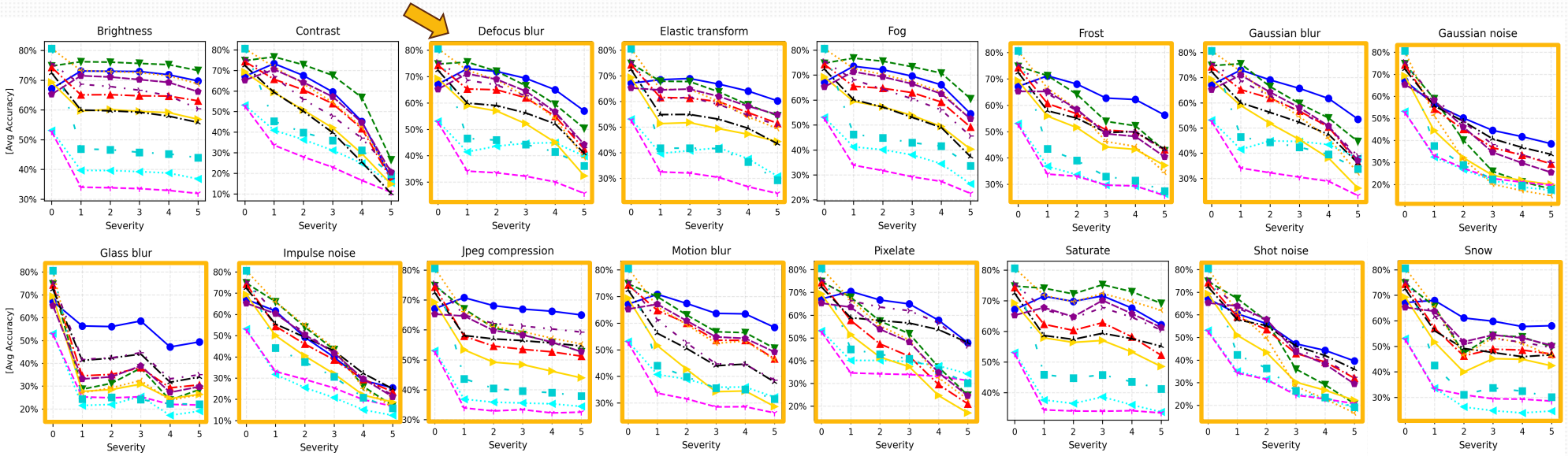
CIFAR10-C (Corrupted) Dataset

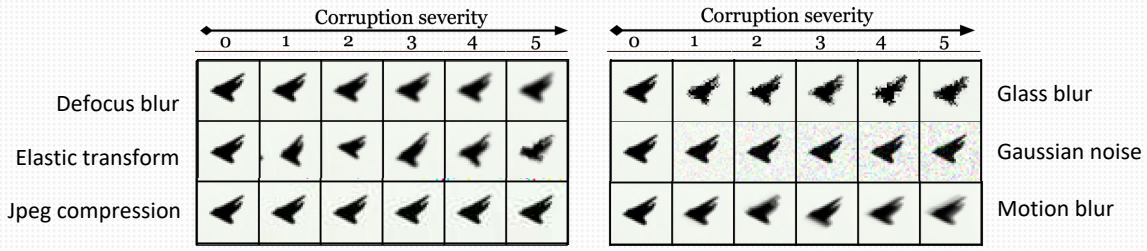
- The CIFAR10-C dataset is a variant of the original CIFAR10 dataset, characterized by the inclusion of corrupted images designed to challenge and evaluate the robustness of models [21].
  - Corrupted images are generated by applying nineteen different types of corruption, such as brightness, fog, and others.
  - We used the CIFAR-10-C dataset to evaluate the robustness and catastrophic forgetting for all benchmark CL models.



Evaluating Against Common Corruptions

- CIFAR10-C (Corrupted) is a variant of the original CIFAR10 dataset that contains corrupted images [6, 21].
- The ADRM model performed relatively best (in 15 out of 19 cases) in various noisy conditions.



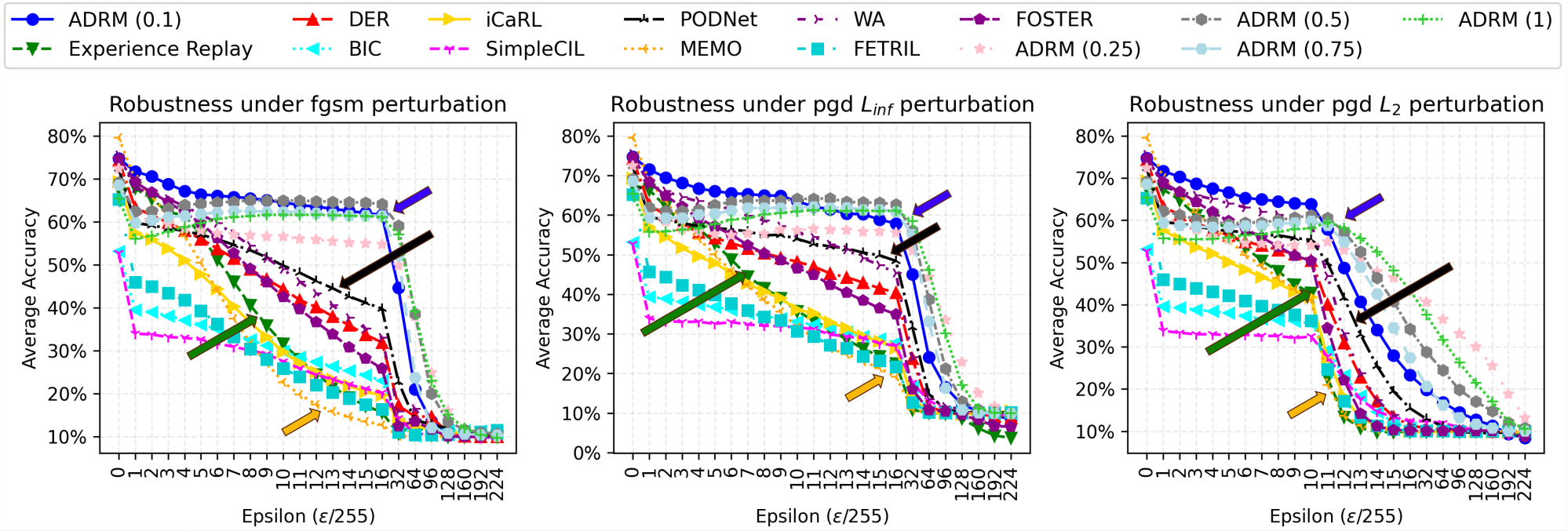


Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Evaluating Against Adversarial Attacks

- All ADRM variants experienced less catastrophic forgetting in adversarial conditions.

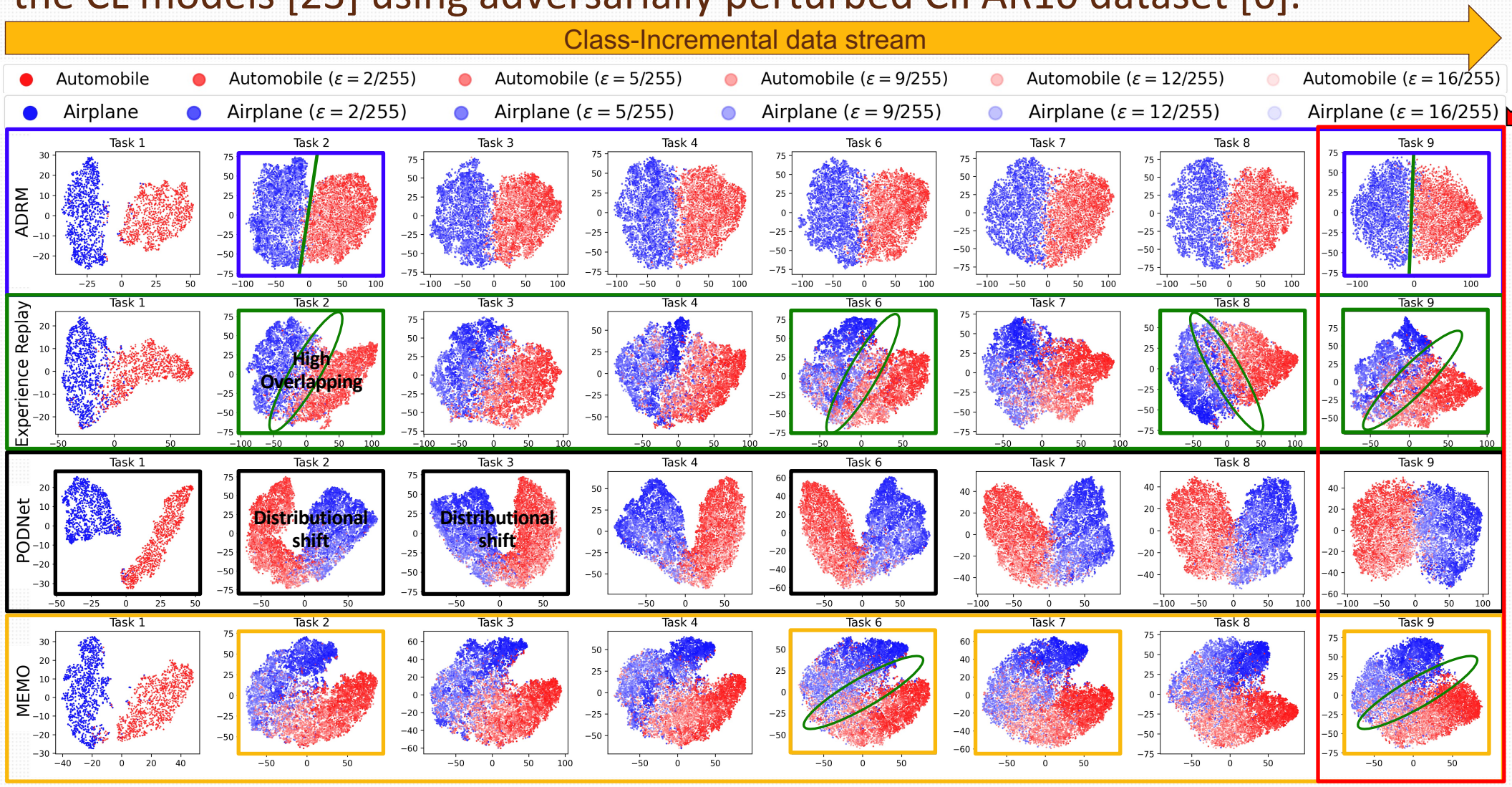


Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

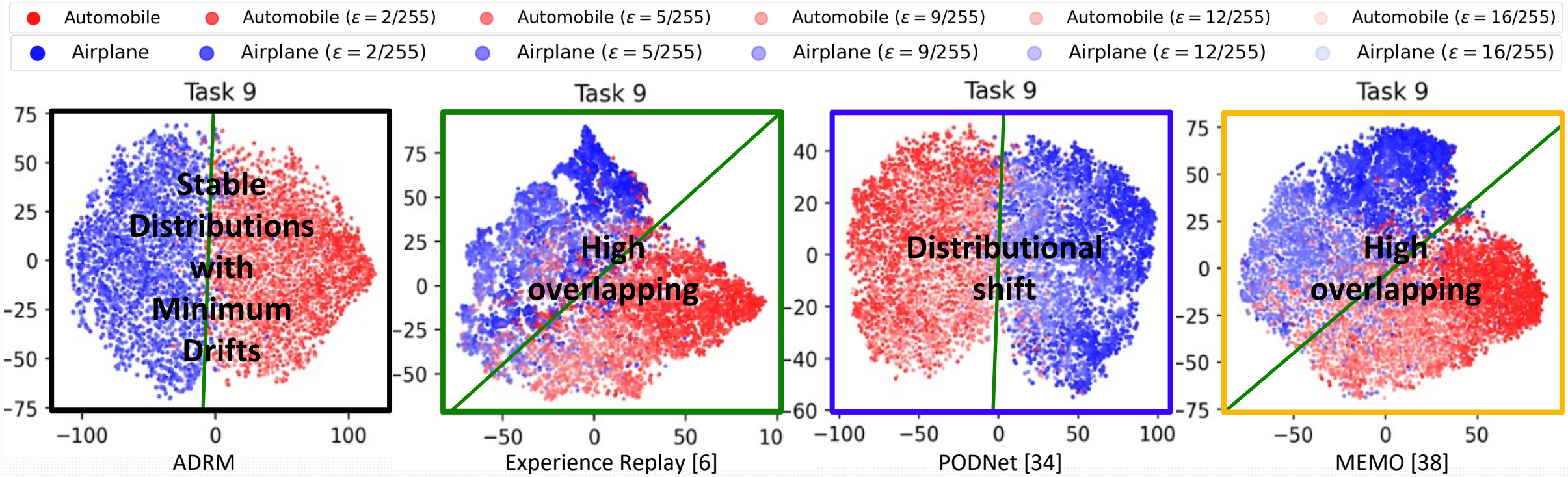
t-SNE Visualization of Latent Feature Distributions in CL

- We utilized t-SNE to visualize the distributions of latent features at the logit layer of the CL models [23] using adversarially perturbed CIFAR10 dataset [6].



t-SNE Visualization of Latent Feature Distributions in CL (Continued)

- Visualization of the learned latent feature distributions for airplanes and automobiles (from the adversarially perturbed CIFAR10 dataset [6]) at the logit layer of CL models [23].

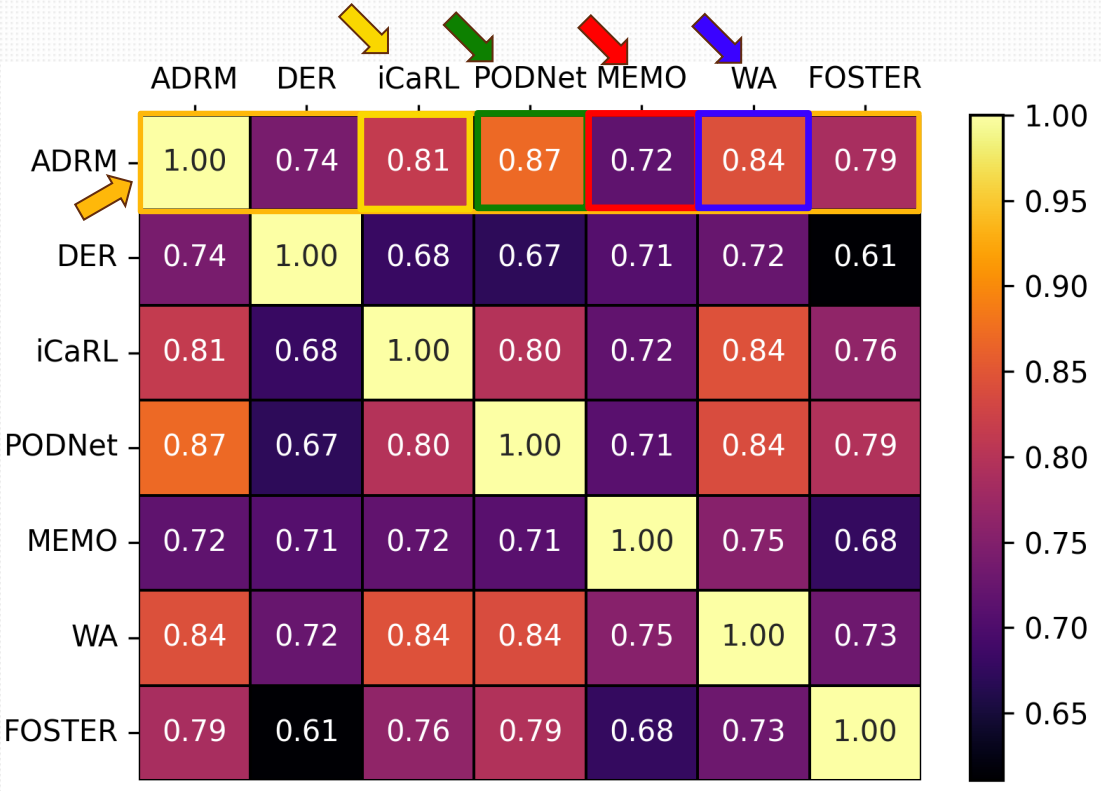


Outline

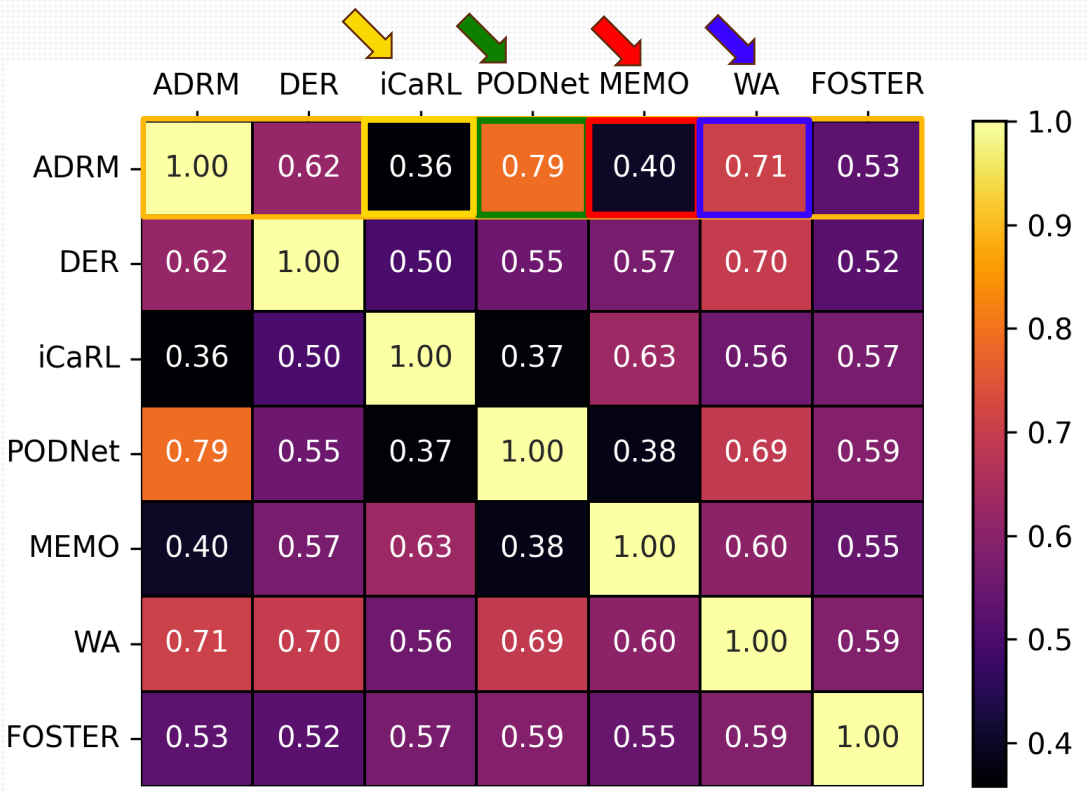
I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Features Similarity Matrices: A Central Kernel Alignment Analysis [24]

- The CL models displayed higher feature similarities (Central Kernel Alignment [24]) to ADRM on Standard CIFAR10 but differed in feature similarities on adversarially perturbed CIFAR10 [6].



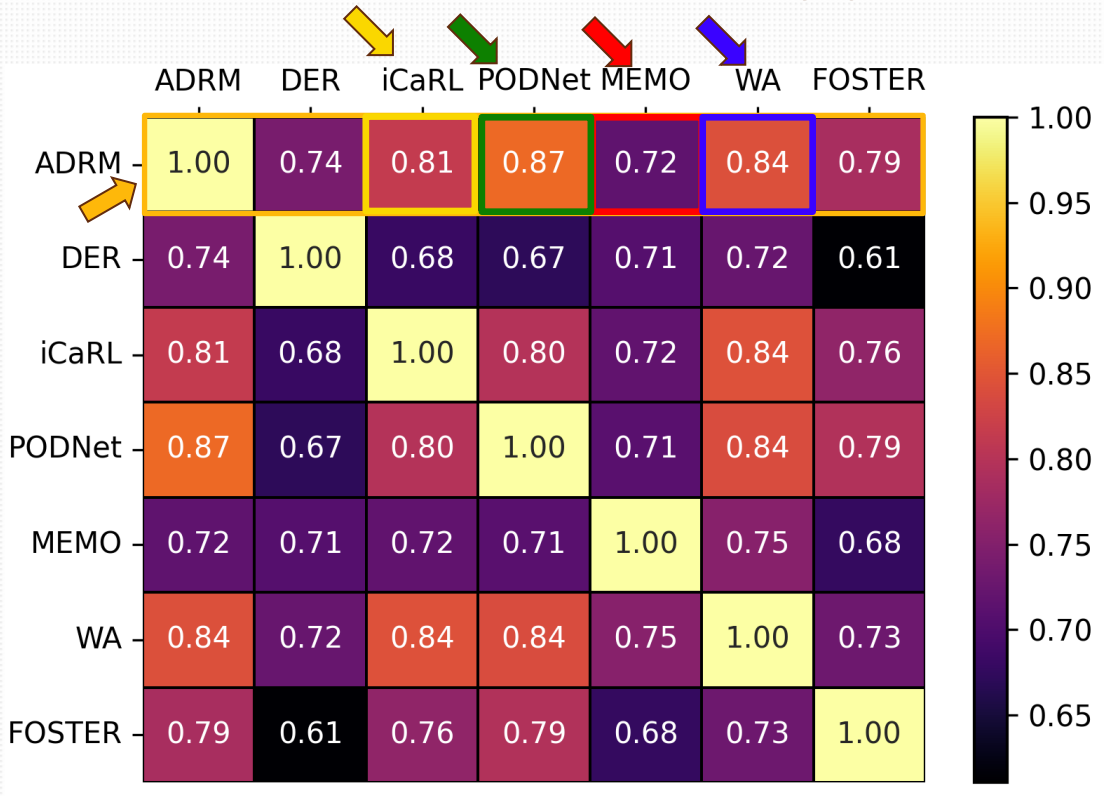
a) Feature Similarity on Standard CIFAR10



b) Feature Similarity on adversarial perturbed CIFAR10

Features Similarity Matrices: A Central Kernel Alignment Analysis [24]

- The CL models displayed higher feature similarities (Central Kernel Alignment [24]) to ADRM on Standard CIFAR10 but differed in feature similarities on adversarially perturbed CIFAR10 [6].



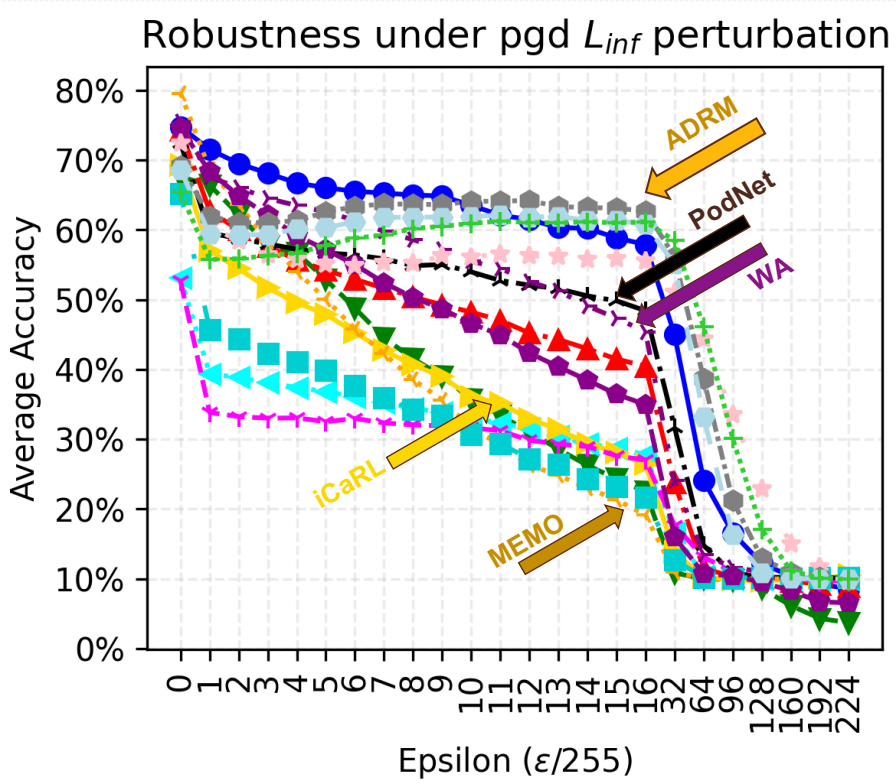
a) Feature Similarity on Standard CIFAR10

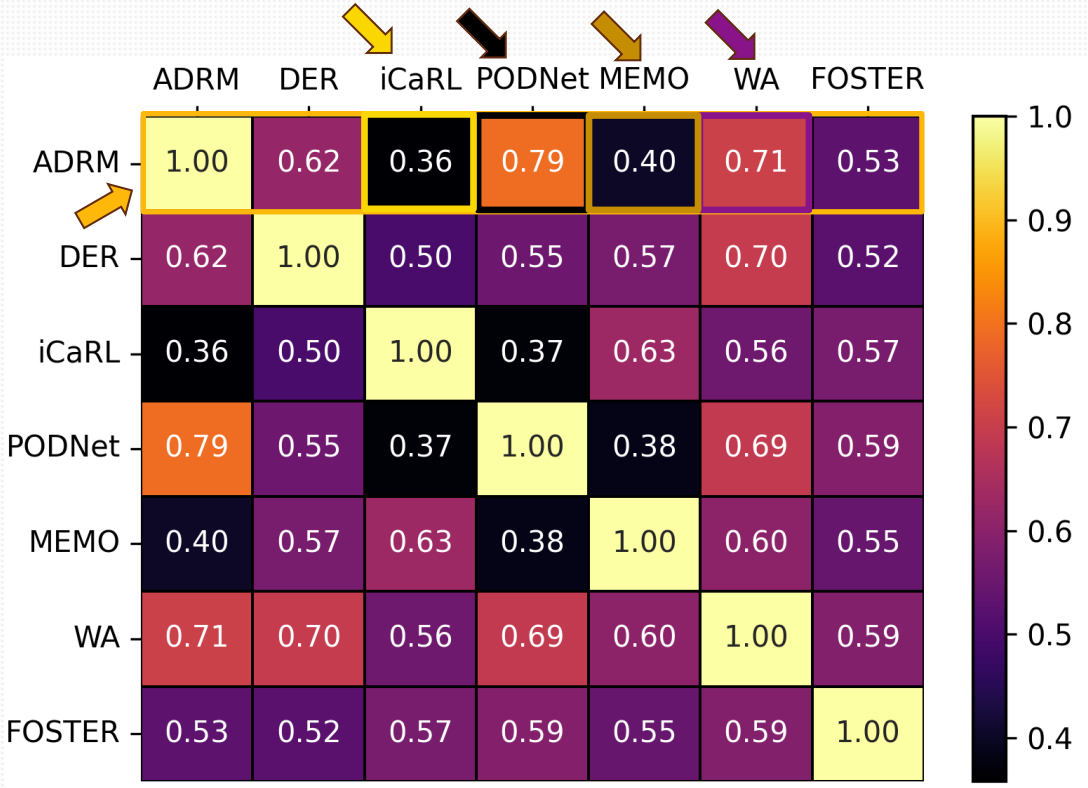
CL Methods	Split-CIFAR10		
	9 steps	5 steps	2 steps
Joint		94.8	
Fine-tune	11.11	18.54	45.31
Experience Replay [6]	67.62	75.97	80.26
iCaRL [33]	69.25	74.85	79.98
BiC [26]	53.11	71.01	86.57
PODNet [34]	72.41	76.96	87.64
WA [25]	75.56	81.67	85.34
DER [35]	74.33	78.14	82.57
SimpleCIL [27]	52.67	54.54	75.01
FOSTER [36]	74.61	80.40	83.89
FETRIL [37]	65.26	67.51	84.52
MEMO [38]	80.60	85.93	87.81
ADRM (0.1)	74.76	80.59	83.95
ADRM (0.25)	72.39	79.15	85.39
ADRM (0.5)	69.20	76.14	81.41
ADRM (0.75)	68.59	76.36	80.17
ADRM (1)	65.39	76.21	82.86

Comparative performance of CL methods on the Split-CIFAR10 dataset over 9 steps, 5 steps, and 2 steps.

Features Similarity Matrices: A Central Kernel Alignment Analysis [24]

- The CL models displayed higher feature similarities (Central Kernel Alignment [24]) to ADRM on Standard CIFAR10 but differed in feature similarities on adversarially perturbed CIFAR10 [6].





b) Feature Similarity on adversarial perturbed CIFAR10

Outline

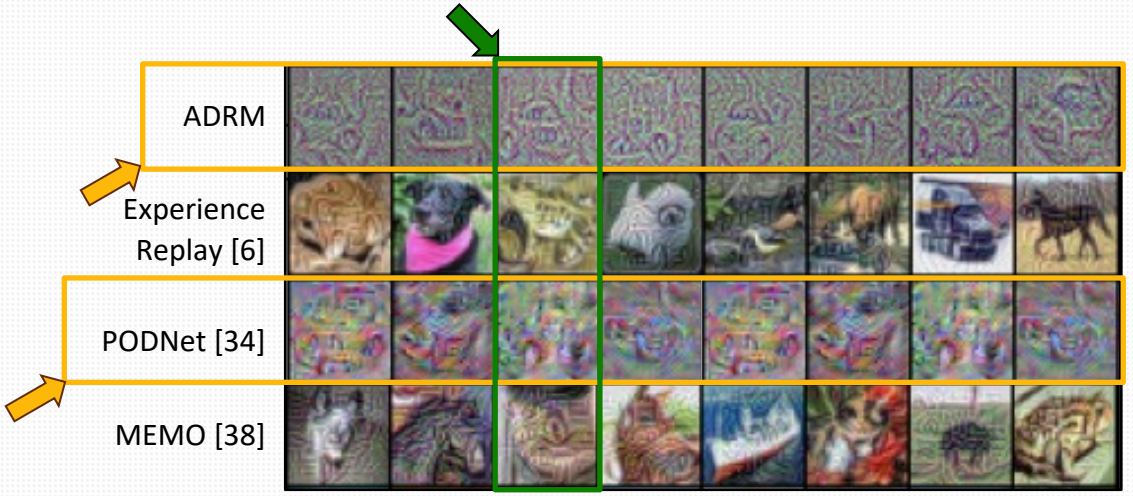
I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Visualization of CL Models' Features for Automobile (Car) Class in Input Space

- The features visualization for the CL model $f^*_{\theta_t}$ is performed by solving the below:

$$\min_{x_r} \| f^*_{\theta_t}(x_r) - f^*_{\theta_t}(x) \|_2^2, \quad x_r \sim D$$

- ADRM learned the salient features of the object as compared to the others CL models.



a) Visualization of disentangled features for car class in input space

Outline

I. Introduction
II. Related Work
III. Problem Statement
IV. Our Work: Adversarially Diversified Rehearsal Memory
V. Results and Baselines
    1. Evaluation Against Common Corruptions
    2. Evaluation Against Common Adversarial Attacks
    3. t-SNE Visualization of Latent Features Distributions in CL
    4. Features Similarity Metrics: A Central Kernel Alignment Analysis
    5. Visualization of CL Model's Features in Input Space
VI. Conclusion

Conclusions

- Rehearsal memory can be diversified to prevent rehearsal memory overfitting in the CL models.

- Rehearsal memory diversification increases the complexity of the memory samples and prevents memory overfitting,
  - Resulting in:
    - Maintaining the effectiveness of memory samples throughout learning
    - Reduced catastrophic forgetting

- Our Code is publicly available at https://github.com/hikmatkhan/ADRM

Acknowledgment

- This work was supported by National Science Foundation (NSF) Award NSF OAC 2008690.

References

[1]. De Lange, Matthias, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. "A continual learning survey: Defying forgetting in classification tasks." *IEEE transactions on pattern analysis and machine intelligence* 44, no. 7 (2021): 3366-3385.

[2]. Van de Ven, Gido M., and Andreas S. Tolias. "Three scenarios for continual learning." arXiv preprint arXiv:1904.07734 (2019).

[3]. Mermillod, Martial, Aurélia Bugaiska, and Patrick Bonin. "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects." *Frontiers in psychology* 4 (2013): 504.

[4]. French, Robert M. "Catastrophic forgetting in connectionist networks." Trends in cognitive sciences 3, no. 4 (1999): 128-135.

[5]. Wang, Liyuan, Xingxing Zhang, Hang Su, and Jun Zhu. "A comprehensive survey of continual learning: Theory, method and application." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[6]. H. Khan, N. C. Bouaynaya and G. Rasool, "Adversarially Robust Continual Learning," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9892970.
doi: 10.1109/ISCC58397.2023.10218203.

[7]. H. Khan, N. C. Bouaynaya and G. Rasool, "The Importance of Robust Features in Mitigating Catastrophic Forgetting," in 2023 IEEE Symposium on Computers and Communications (ISCC), Gammarth, Tunisia, 2023 pp. 752-757.

[8]. H. Khan, N. C. Bouaynaya and G. Rasool, "Brain-Inspired Continual Learning: Robust Feature Distillation and Re-Consolidation for Class Incremental Learning," in IEEE Access, vol. 12, pp. 34054-34073, 2024, doi: 10.1109/ACCESS.2024.33694.88.

[9]. H. Khan, N. C. Bouaynaya and G. Rasool, "Adversarially Diversified Rehearsal Memory (ADRM): Mitigating Memory Overfitting Challenge in Continual Learning," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024, pp. 1-8.

[10]. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[11]. Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial examples are not bugs, they are features." Advances in neural information processing systems 32 (2019).

[12]. Wang, Tongzhou, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. "Dataset distillation." arXiv preprint arXiv:1811.10959 (2018).

[13]. Kim, Jang-Hyun, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. "Dataset condensation via efficient synthetic-data parameterization." In International Conference on Machine Learning, pp. 11102-11118. PMLR, 2022.

[14]. Deng, Zhiwei, and Olga Russakovsky. "Remember the past: Distilling datasets into addressable memories for neural networks." Advances in Neural Information Processing Systems 35 (2022): 34391-34404.

References (Continued)

[15]. Rolnick, David, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. "Experience replay for continual learning." Advances in neural information processing systems 32 (2019).

[16]. Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan et al. "Overcoming catastrophic forgetting in neural networks." Proceedings of the national academy of sciences 114, no. 13 (2017): 3521-3526.

[17]. Yan, Shipeng, Jiangwei Xie, and Xuming He. "Der: Dynamically expandable representation for class incremental learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3014-3023. 2021.

[18]. Douillard, Arthur, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. "Podnet: Pooled outputs distillation for small-tasks incremental learning." In Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16, pp. 86-102. Springer International Publishing, 2020.

[19]. H. Khan, C. Johnson, R. University, N. Bouaynaya, G. Rasool, T. Travis, et al., "Explainable AI: Rotorcraft attitude prediction", Proc. Vertical Flight Soc. 76th Annu. Forum, Oct. 2020.

[20]. H. Khan, C. Johnson, R. University, N. Bouaynaya, G. Rasool, L. Thompson, et al., "Deep ensemble for rotorcraft attitude prediction", Proc. Vertical Flight Soc. 77th Annu. Forum, May 2021.

[21]. Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." arXiv preprint arXiv:1903.12261 (2019).

[22]. Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. 11 (2008).

[23]. Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. 11 (2008).

[24]. Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. "Similarity of neural network representations revisited." In International conference on machine learning, pp. 3519-3529. PMLR, 2019.

[25]. Zhao, Bowen, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. "Maintaining discrimination and fairness in class incremental learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13208-13217. 2020.

[26]. Wu, Yue, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. "Large scale incremental learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 374-382. 2019.

[27]. Zhou, Da-Wei, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. "Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need." arXiv preprint arXiv:2303.07338 (2023).

[28]. Wamsley, Erin J. "Dreaming and offline memory consolidation." Current neurology and neuroscience reports 14 (2014): 1-7.

[29]. Lee, Albert K., and Matthew A. Wilson. "Memory of sequential experience in the hippocampus during slow wave sleep." Neuron 36.6 (2002): 1183-1194.

References (Continued)

[30]. Wamsley, Erin J. "Dreaming and offline memory consolidation." Current neurology and neuroscience reports 14 (2014): 1-7.

[31] Wamsley, Erin J., et al. "Cognitive replay of visuomotor learning at sleep onset: temporal dynamics and relationship to task performance." Sleep 33.1 (2010): 59-68.

[33]. Rebuffi, Sylvestre-Alvise, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. "icarl: Incremental classifier and representation learning." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2001-2010. 2017.

[34]. Douillard, Arthur, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. "Podnet: Pooled outputs distillation for small-tasks incremental learning." In Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16, pp. 86-102. Springer International Publishing, 2020.

[35]. Yan, Shipeng, Jiangwei Xie, and Xuming He. "Der: Dynamically expandable representation for class incremental learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3014-3023. 2021.

[36]. Wang, Fu-Yun, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. "Foster: Feature boosting and compression for class-incremental learning." In European conference on computer vision, pp. 398-414. Cham: Springer Nature Switzerland, 2022.

[37]. Petit, Grégoire, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. "Fetril: Feature translation for exemplar-free class-incremental learning." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3911-3920. 2023.

[38]. Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness may be at odds with accuracy." arXiv preprint arXiv:1805.12152 (2018).

[39]. Zhao, Bo, Konda Reddy Mopuri, and Hakan Bilen. "Dataset condensation with gradient matching." arXiv preprint arXiv:2006.05929 (2020).

[40]. Li, Yiming, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. "Backdoor learning: A survey." IEEE Transactions on Neural Networks and Learning Systems (2022).

[41]. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[42]. Waseda, Futa, Sosuke Nishikawa, Trung-Nghia Le, Huy H. Nguyen, and Isao Echizen. "Closer look at the transferability of adversarial examples: How they fool different models differently." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1360-1368. 2023.

[43]. Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness may be at odds with accuracy." arXiv preprint arXiv:1805.12152 (2018).

[44]. Kim, Junho, Byung-Kwan Lee, and Yong Man Ro. "Distilling robust and non-robust features in adversarial examples by information bottleneck." Advances in Neural Information Processing Systems 34 (2021): 17148-17159.

[45]. Huang, Tianjin, Vlado Menkovski, Yulong Pei, and Mykola Pechenizkiy. "Bridging the performance gap between fgsm and pgd adversarial training." arXiv preprint arXiv:2011.05157 (2020).

References (Continued)

[46]. Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).

Adversarially Diversified Rehearsal Memory (ADRM): Mitigating Memory Overfitting Challenge in Continual Learning

Dr. Hikmat Khan
Department of Electrical and Computer Engineering,
Rowan University,
Glassboro, NJ, USA
Official: Khanhi83@rowan.edu
Personal: Hikmat.khan179@gmail.com