STAR: Stability-Inducing Weight Perturbation for Continual Learning

Masih Eskandar, Tooba Imtiaz, Davin Hill, Zifeng Wang, Jennifer Dy

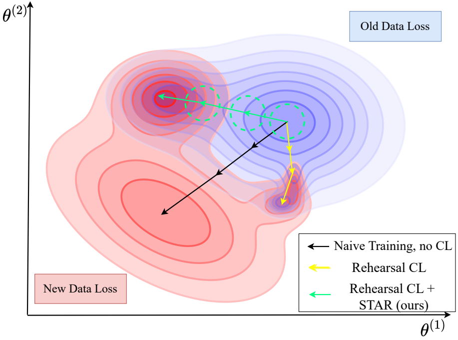
Northeastern University

Motivation

- Humans learn sequentially without forgetting.
- Neural networks suffer from **catastrophic forgetting**.
- Rehearsal-based methods help but are limited by buffer size.
- Goal: Improve stability of model predictions over time.

Key Idea: STAR

- Introduce **STAR**: Stability-Inducing Parameter-space Regularization.
- Promotes consistency of model outputs under **worst-case local parameter perturbations**.
- Plug-and-play with any rehearsal-based continual learning method.



How STAR Works

1. Identify correctly classified buffered samples.
2. Apply perturbation δ that maximizes KL divergence of outputs.
3. Use this worst-case δ to regularize training:

$$\mathcal{L}_{\text{STAR}} = \max_{\|\delta\| \leq d} \sum_{(x,y) \in M^*} \text{KL}(q_\theta(x) \parallel q_{\theta+\delta}(x))$$

4. Final objective:

$$\mathcal{L} = \mathcal{L}_{\text{CL}} + \lambda \mathcal{L}_{\text{STAR}}$$

Results Summary

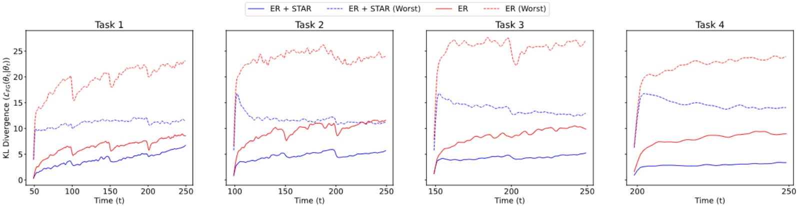
- STAR improves accuracy by up to **15%** across methods and datasets.
- Effective across buffer sizes and baselines (ER, DER++, ER-ACE, X-DER).
- Especially beneficial when buffer is small.

Table 1: Comparison of adding STAR to baseline rehearsal-based CL methods in terms of average accuracy.

Method	Split-Cifar10			Split-Cifar100			Split-miniImageNet		
Sequential	19.67			9.29			4.51		
Joint	92.38			73.29			53.55		
Buffer Size	100	200	500	200	500	2000	1000	2000	5000
ER	36.39	44.79	57.74	14.35	19.66	36.76	8.37	16.49	24.17
+ STAR (ours)	**51.5**	**59.3**	**70.70**	**19.64**	**29.64**	**44.65**	**11.83**	**16.64**	**25.83**
ER-ACE	52.95	61.25	71.16	29.22	38.01	49.95	17.95	22.60	27.92
+ STAR (ours)	**60.69**	**67.58**	**75.44**	**30.38**	**40.20**	**51.67**	**21.06**	**24.9**	**31.01**
DER++	57.65	64.88	72.70	25.11	37.13	52.08	18.02	23.44	30.43
+ STAR (ours)	**61.76**	**68.60**	**76.52**	**27.64**	**39.77**	**53.24**	**22.4**	**28.19**	**33.36**
X-DER (RPC)	52.75	58.48	64.77	37.23	**48.53**	57.00	23.19	26.38	29.91
+ STAR (ours)	**58.85**	**65.94**	**69.19**	**38.15**	47.56	**57.55**	**24.6**	**27.95**	**32.6**

Ablations and Insights

- Gradient-based perturbations outperform random ones.
- Using only buffer samples in the STAR loss is best.
- Empirically validate that STAR reduces local-worst case change in distribution



Conclusion

- STAR is a simple yet powerful regularization strategy for CL.
- Enhances rehearsal-based methods without needing task boundaries.
- Open-sourced: `github.com/Gnomy17/STAR_CL`