Enhancing Adversarial Robustness for Deep Metric Learning

Mo Zhou Vishal M. Patel
Johns Hopkins University

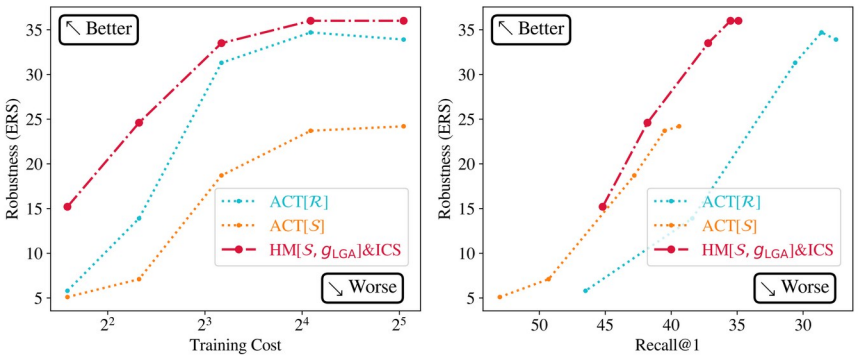
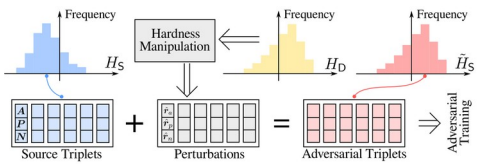
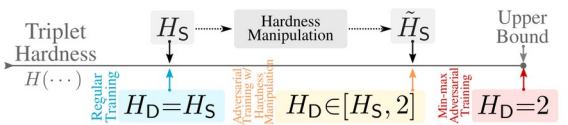


Figure 1. Comparison in robustness, training cost, and recall@1 between our method (*i.e.*, "HM[$\mathcal{S}$, g_{LGA}]&ICS") and the state-of-the-art method (*i.e.*, "ACT[$\mathcal{R}$]" and "ACT[$\mathcal{S}$]") on the CUB Dataset.

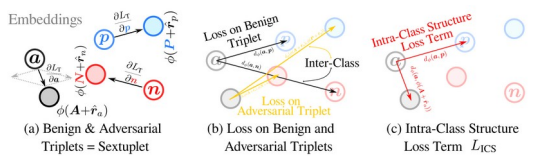
arXiv  Github 

* Deep Metric Learning is vulnerable to adversarial attacks.
* Existing defense methods learn from weak adversary in Order to avoid model collapse, which is inefficient.
* We propose Hardness Manipulation (HM), an flexible and Efficient tool for creating adversarial example triplets.
* We propose Linear Gradual Adversary (LGA) as a Pseudo-hardness function for HM to balance training Objectives during adversarial training.





* We propose Intra-Class Structure (ICS) loss term to further Improve model robustness and adversarial training efficiency.



(a) Benign & Adversarial Triplets = Sextuplet (b) Loss on Benign and Adversarial Triplets (c) Intra-Class Structure Loss Term L_{ICS}

* The proposed method overwhelmingly outperforms SoTA by a large margin.