Model-Agnostic and Generalizable Learning: Beyond Flatness for Domain Generalization

Jeong, Jayden (School: Paul Laurence Dunbar High School)

Post-hoc domain generalization (PDG) methods aim to generalize machine learning models to unseen target domains using only source-domain information leveraged after training. Unlike train-time DG methods, which typically require retraining to apply a new method, PDG methods operate on an existing checkpoint bank in weight space, making them 1) practical to deploy at scale and 2) model-agnostic. Further, PDG methods such as SWAD and DiWA have shown that post-hoc checkpoint averaging can be just as, and often, even more effective than train-time methods. SWAD motivates averaging through flatness and overfit-aware sampling, and DiWA motivates it through diversity-aware averaging across independently trained mergeable models. While these methods are effective, they summarize source behavior through aggregate signals and do not directly consider loss distribution across individual source domains or whether retained checkpoints complement one another, leaving useful diagnostic information unused, as recent work on QRM has shown. We propose a finer-grained deployment property called “distributional canalization”: the stability of per-domain losses under perturbations of the relative source-domain emphasis. We show that, under a local source-supported model of target shift, target risk can be bounded in terms of canalization, and that low average source loss alone does not guarantee it. Motivated by this result, we propose CDA, which implements the canalization objective via a checkpoint bank compression algorithm (Version Space Compression) and a nonuniform checkpoint weigher (CDA-BD) which work together to promote stability across source domains. On the five DomainBed benchmarks, CDA reaches 68.3 average OOD accuracy, vs. 63.3 for ERM, 66.9 for SWAD, and 68.0 for DiWA.

Awards Won: