Synopsis of Papers read

Edward, Sagar, Kapil

April 1, 2025

- [3] Smaller Models trained via KD rely more on spurious correlations than the teacher model. this leads to loss in worst group performance / group fairness metrics. The dedier to solve this. the claim is that this correlation are learnt in the early layers of the student network and a readouts from these help. the readouts from early layers are more confident for the problematic instances and hence weighing the kd loss uisng them would be helpful based on the confidence margin.

$$\mathcal{L}_{student} = \sum_{D_w}(1-\lambda)\cdot l_{ce} + \lambda \cdot \mathbf{wt} \cdot l_{ke}$$

  where $\mathbf{wt} = \exp^{\beta.\mathbf{cm}.\alpha}$ Evaluations done on 4 debiasing benchmarks: CelebsA: Blond/Non Blond, Waterbirds: Landbird / Waterbird, MultiNLI: Entails/neutral/contradict, CivilComments-WILDS: toxic/ or not. Baselines used- Train twice: reweighing the losses from first classifier, Group DRO, Reused teacher heads.

  Model architecture used RESNETS t-50,student-18, bert-T, distillbert student.

- [2] Talks about prediction sets in Early exit network that are consistent in some sense.

- [1] Having acces to a large model that generalises well, the genrelisation can be induced to a smaller model by distillation. Here in combination with the cross entropy loss we also minimise the kl divergence of last layer logits with the teacher model The intuition being that even the negative examples give info about dataset that is not captured by the cross entropy loss. To read: HMM, mariginalisation, Large scale distributed deep networks, mixture of experts.

References

[1] Hinton, G., Vinyals, O., and Dean, J. Distilling the Knowledge in a Neural Network, Mar. 2015. arXiv:1503.02531 [cs, stat].

[2] Jazbec, M., Forré, P., Mandt, S., Zhang, D., and Nalisnick, E. Early-Exit Neural Networks with Nested Prediction Sets, June 2024. arXiv:2311.05931 [cs, stat].

[3] Tiwari, R., Sivasubramanian, D., Mekala, A., Ramakrishnan, G., and Shenoy, P. Using Early Readouts to Mediate Featural Bias in Distillation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (Waikoloa, HI, USA, Jan. 2024), IEEE, pp. 2626–2635.