Paper: Intra-Processing Methods for Debiasing Neural Networks (link) (github link)

Summary

- Focuses on “intra-processing” methods, which are applicable to cases where you need to fine-tune parameters for an existing model (ie. our BERT model)
- Provides algorithm outlines for each method
- Trying to maximize the objective function:

$$\phi_{\mu,\rho,\epsilon}(\mathcal{D}, \hat{\mathcal{Y}}, A) = \begin{cases} \rho \text{ if } \mu < \epsilon, \\ 0 \text{ otherwise.} \end{cases}$$

- Random Perturbation
  - Gradually apply random noise to model weights and select configuration that maximizes accuracy and minimizes bias
  - Benefits: simple and computationally light
- Layerwise Optimization
  - Instead of random perturbation, we strategically optimize one layer at a time with respect to the objective function
  - 0th order optimization (doesn’t use any derivatives) to save computing power (using methods like GBRT)
  - Benefits: can specifically target layers where bias arises, computationally light
- Adversarial Debiasing
  - Already discussed - trains an ‘adversarial network’ and optimizes the primary network to produce features that are A cannot detect as biased (ie. produces outputs that are independent of biases)
  - Benefits: inherently produces an unbiased network, as opposed to simple ‘tuning’ afterwards
- Additional notes
  - It is important to choose fairness metrics matching the context of application

Application to our model

- Adversarial debiasing is an in-process method
- Random perturbation and layerwise optimization are post-process methods
- High applicability for our model, considering that we are primarily able to tune the parameters.
- All are fairly computationally reasonable methods for fine tuning (ie. layerwise optimization does one layer at a time with 0th order optimization, which is not computationally demanding, random perturbation requires little to no real computation)

Recommendations

- These are mostly integrated into the model after the base model is complete
- We need to identify which biases we are hoping to minimize
- Should complete and analyze a model before implementing debiasing methods