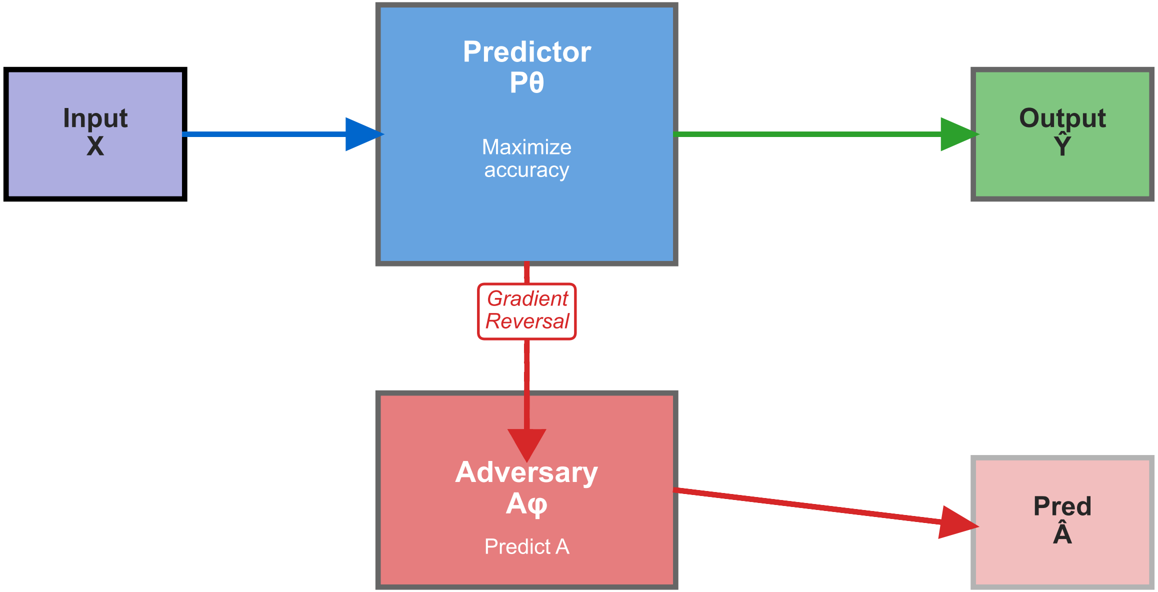
Adversarial Debiasing Architecture



Predictor Loss:

$L_P = -\mathrm{Acc}(Y, \hat{Y})$

Adversary Loss:

$L_A = -\mathrm{Acc}(A, \hat{A})$

Combined:

$\min_{\theta} \max_{\phi} L_P - \lambda L_A$