Model Compression for Domain Adaptation through Causal Effect Estimation

Recent improvements in the predictive quality of NLP systems are often dependent on a substantial increase in the number of model parameters. This has led to various attempts of compressing such models, but existing methods have not considered the differences in the predictive power of various model components or in their generalizability. To understand the connection between model compression and out-of-distribution generalization, we define the task of compressing language models such that they perform best in a domain adaptation setting. We attempt to estimate the average treatment effect (ATE) of a model component on the model's predictions. Our proposed model AMoC generates many model candidates, differing by the model components that were removed. Then, we select the best candidate through a stepwise-regression model that utilizes the ATE to predict the expected performance on the target domain. AMoC outperforms strong baselines on dozens of domain pairs across three text classification tasks.