

Efficient neural network inference is important in a number of practical domains, such as deployment in mobile settings. An effective method for increasing infer