Structural pruning of neural network parameters reduces computational, energy, and memory transfer costs during inference. We propose a novel method that e