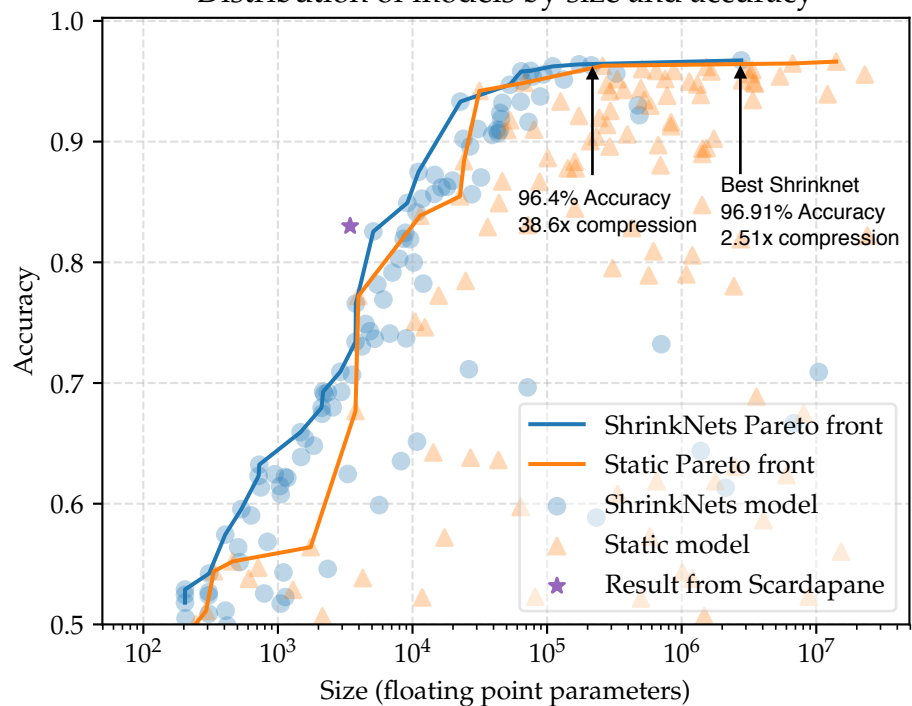
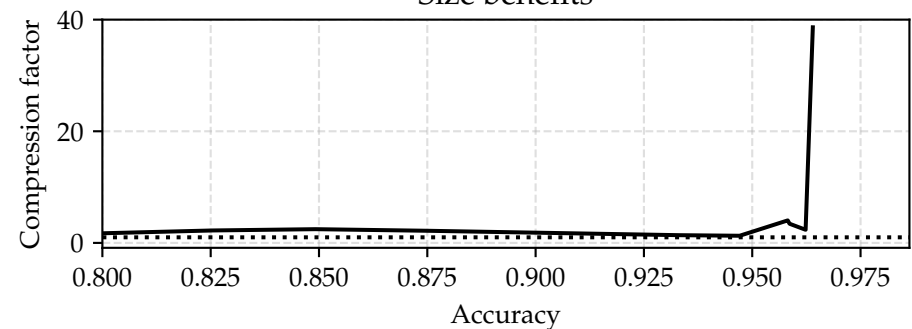


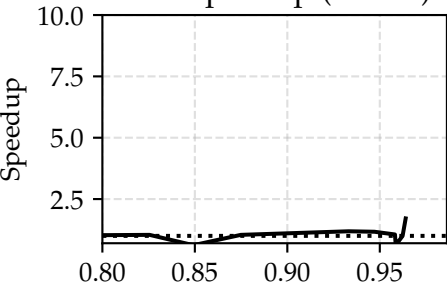
Distribution of models by size and accuracy



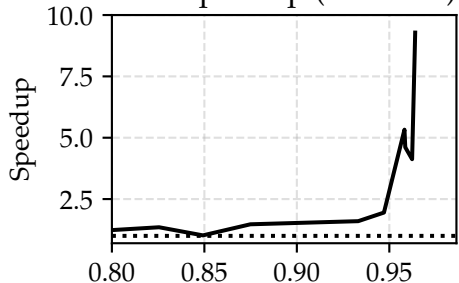
Size benefits



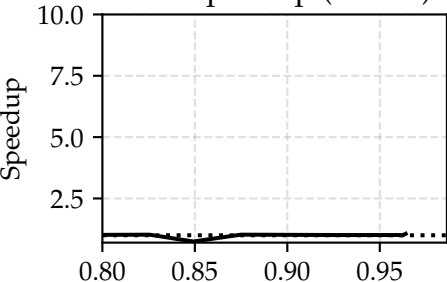
CPU speedup ($bs = 1$)



CPU speedup ($bs = 512$)



GPU speedup ($bs = 1$)



GPU speedup ($bs = 4096$)

