Deep neural networks consist of multiple hidden layers that contain large amounts of parameters as high-order tensors. Because of this, low-power/edge devices are unable to execute data-heavy processes due to the limited amount of memory available. To limit the computational requirement, tensor decomposition is a common technique used for compressing data in neural network layers by reshaping the tensor's shape and extracting latent information. In this paper, we study the effect a tensor's shape has within neural networks and how folding parameters can affect its overall compression and accuracy. Our aim is to develop an optimization algorithm to maximize the compression ratio for tensor decomposition within a general neural network. This research paper discusses the results found when compressing weights for deep neural networks to find the desired compression ratio, ultimately enabling our method to be utilized universally for different neural networks regardless of their architecture or application.