高效的神经网络设计，包括模型压缩和模型加速

深度学习模型压缩与加速设计 <http://www.tensorinfinity.com/paper_167.html>