Promoted by the evolution of artificial intelligence, more and more intelligent applications have emerged on mobile devices. As one of the most representative deep learning technologies, Convolutional Neural Networks (CNNs) have been considered as a primary tool for computer vision fields, such as image recognition and object detection. However, the heavy computation, memory, and energy demands of the CNN model restrict their deployment on resource-constrained mobile devices. This presentation will introduce my research solution that enables efficient neural network computation by compressing the CNN model. Different from previous works, I interpret the functionalities of the convolutional filters and identify the model structural redundancy as repetitive filters with similar feature preferences. I will introduce a filter functionality-oriented CNN model compression method that can achieve significant computation load reduction without losing model accuracy.