Learning effective fusion of multi-modality features is at the heart of visual question answering. We propose a novel method of dynamically fuse multi-modal fea