Multimodal attentional networks are currently state-of-the-art models for Visual Question Answering (VQA) tasks involving real images. Although attention allows