

# Question-Guided Hybrid Convolution for Visual Question Answering

Peng Gao<sup>1</sup>, Hongsheng Li<sup>1</sup>, Shuang Li<sup>1</sup>, Pan Lu<sup>1</sup>, Yikang Li<sup>1</sup>, Steven C.H. Hoi<sup>2</sup>, and Xiaogang Wang<sup>1</sup>

<sup>1</sup> CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong  
{penggao, hsl i , s l i , p l u , y k l i , x g w a n g}@ee. cuhk. edu. hk

<sup>2</sup> School of Information Systems, Singapore Management Univeristy  
chhoi @smu. edu. sg

**Abstract.** In this paper, we propose a novel Question-Guided Hybrid Convolution (QGHC) network for Visual Question Answering (VQA). Most state-of-the-art VQA methods fuse the high-level textual and visual features from the neural network and abandon the visual spatial information when learning multi-modal features. To address these problems, question-guided kernels generated from the input question are designed to convolute with visual features for capturing the textual and visual relationship in the early stage. The question-guided convolution can tightly couple the textual and visual information but also introduce more parameters when learning kernels. We apply the group convolution, which consists of question-independent kernels and question-dependent kernels, to reduce the parameter size and alleviate over-fitting. The hybrid convolution can generate discriminative multi-modal features with fewer parameters. The proposed approach is also complementary to existing bilinear pooling fusion and attention based VQA methods. By integrating with them, our method could further boost the performance. Experiments on VQA datasets validate the effectiveness of QGHC.

**Keywords:** VQA · Dynamic Parameter Prediction · Group Convolution

## 1 Introduction

Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) have shown great success in vision and language tasks [12, 39, 46]. Recently, CNN and RNN are jointly trained for learning feature representations for multi-modal tasks, including image captioning [3, 4], text-to-image retrieval [5, 34], and Visual Question Answering (VQA) [6, 26, 40]. Among the vision-language tasks, VQA is one of the most challenging problems. Instead of embedding images and their textual descriptions into the same feature subspace as in the text-image matching problem [7, 8, 27], VQA requires algorithms to answer natural language questions about the visual contents. The methods are thus designed to understand both the questions and the image contents to reason the underlying truth.

---

corresponding author



































