Dynamic Fusion with Intra- and Inter-modality Attention Flow for Visual Question Answering

Peng Gao[1], Zhengkai Jiang[3], Haoxuan You[4],
Pan Lu[4], Steven Hoi [2], Xiaogang Wang [1], Hongsheng Li [1]
[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Singapore Management University [3]NLPR, CASIA [4]Tsinghua University
{*1155102382@link, xgwang@ee, hsli@ee*}*.cuhk.edu.hk*

Abstract

Learning effective fusion of multi-modality features is at the heart of visual question answering. We propose a novel method of dynamically fusing multi-modal features with intra- and inter-modality information flow, which alternatively pass dynamic information between and across the visual and language modalities. It can robustly capture the high-level interactions between language and vision domains, thus significantly improves the performance of visual question answering. We also show that the proposed dynamic intra-modality attention flow conditioned on the other modality can dynamically modulate the intra-modality attention of the target modality, which is vital for multimodality feature fusion. Experimental evaluations on the VQA 2.0 dataset show that the proposed method achieves state-of-the-art VQA performance. Extensive ablation studies are carried out for the comprehensive analysis of the proposed method.

1. Introduction

Visual Question Answering [2] aims at automatically answering a natural language question related to the contents of a given image. It has extensive applications in practice, such as assisting blind people [12] and education of young children, and therefore become a hot research topic recently. The performance of Visual Question Answering (VQA) has been substantially improved in recent years thanks to three lines of works. Firstly, better visual and language feature representations are at the core for boosting VQA performance. The feature learning capability from VGG [35], ResNet [13], FishNet [36] to the recent bottom-up & top-down features [1] increases the VQA performance significantly. Secondly, different variants of attention mechanisms [40] can adaptively select important features which can help deep learning achieve better recognition accuracy. Thirdly, better multi-modality fusion approaches, such Bilinear Fusion [9], MCB [7] and MUTAN [4], have been proposed for better capturing the high-level interactions between language and visual features.

Despite being studied extensively, most existing VQA approaches focus on learning inter-modality relations between visual and language features. Bilinear feature fusion approaches [9] focus on capturing the higher order relations between language and visual modalities by feature outer product. Co-attention [39, 28, 24] or bilinear attention-based approaches [19] learn the inter-modality relations between word-region pairs to identify key pairs for question answering. On the other hand, there exist computer vision and natural language processing algorithms focusing on learning intra-modality relations. Hu *et al*. [14] proposed to explore intra-modality object-to-object relations to boost object detection accuracy. Yao *et al*. [42, 26] modeled intra-modality object-to-object relations for improving image captioning performance. In the recently proposed BERT algorithm [6] for natural language processing, intra-modality word relations are modelled by self-attention mechanism to learn state-of-the-art word embedding. However, the inter- and intra-modality relations were never jointly investigated in a unified framework for solving the VQA problem. We argue that, for the VQA problem, the intra-modality relations within each modality is complementary to the inter-modality relations, which were mostly ignored by existing VQA methods. For instance, for the image modality, each image region should obtain information not only from its associate words/phrases in the question but also from related image regions to infer the answer of the question. For the question modality, better understanding of question can be acquired by inferring other words. Such cases motivate us to propose a unified framework for modelling both inter- and intra-modality information flow.

To overcome the limitations, we propose a novel Dynamic Fusion with Intra- and Inter-modality Attention Flow

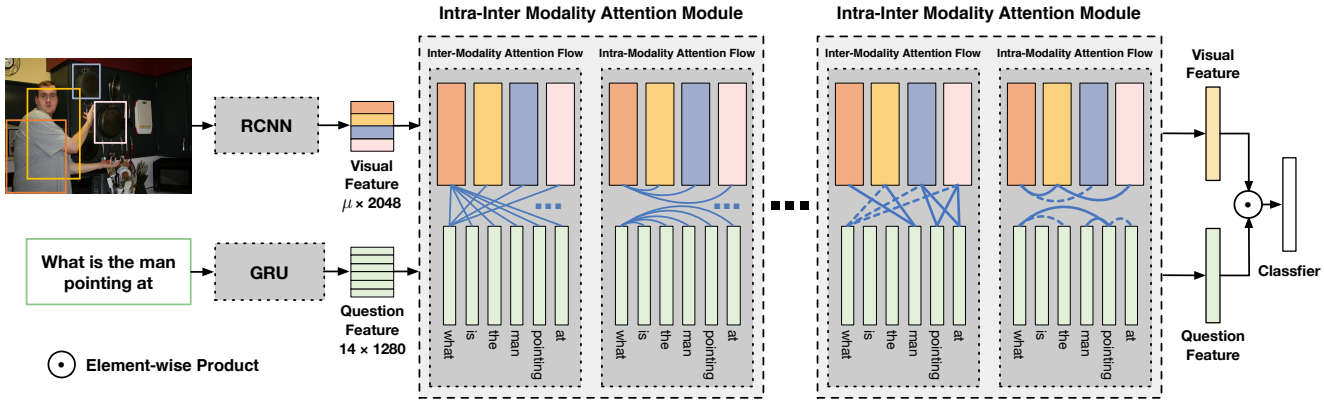

Figure 1: Illustration of the proposed Dynamic Fusion with Intra- and Inter-modality Attention Flow (DFAF) for visual question answering. Each DFAF module contains one Inter-Modality Attention Flow and one of Intra-Modality Attention Flow Module. Stacking several blocks of DFAF can help the network gradually focus on important image regions , question words and the latent alignments.

(DFAF) framework for efficient multi-modality feature fusion to accurately answer visual questions. The overall diagram is shown in Figure 1. Our DFAF framework integrates cross-modal self-attention and cross-modal co-attention mechanisms to achieve effective information flows within and between the image and language modalities. Given visual and question features encoded by deep neural networks, the DFAF framework first generates inter-modality attention flow (InterMAF) to pass information between image and language. In the InterMAF module, visual and language features generate a joint-modality co-attention matrix. Each visual region would select question features according to the joint-modality co-attention matrix and vice versa. The InterMAF module fuses and updates each image region and each word's features according to the attention-weighted information flows from the other modality. Following the InterMAF module, DFAF calculates the dynamic intra-modality attention flow (DyIntraMAF) for passing information flows within each modality to capture the complex intra-modality relations. Visual regions and sentence words generate self-attention weights and aggregate attention-weighted information from other instances in the same modality. More importantly, although the information are only propagated within the same modalities, information of the other modality is considered and used to modulate intra-modality attention weights and flows. With such an operation, the attention flows within each modality are dynamically conditioned on the other modality and is the key difference compared with existing intra-modality message passing methods on object detection [14] and image captioning [42]. DyIntraMAF is shown to be substantially better than its variant using only internal information for intra-modality information flow and is the key to the success of the proposed framework. We alternatively use InterMA and DyIntraMA modules to create the basic blocks of the DFAF. Multiple stacks of DFAF blocks are shown to further improve the VQA performance.

Our contributions can be summarized into threefold. (1) A novel Dynamic Fusion with Intra- and Inter-modality Attention Flow (DFAF) framework is proposed for multi-modality fusion by interleaving intra- and inter-modality feature fusion. Such a framework for the first time integrates inter-modality and dynamic intra-modality information flow in a unified framework for tackling the VQA task. (2) Dynamic Intra-modality Attention Flow (DyIntraMAF) module is proposed for generating effective attention flows within each modality, which are dynamically conditioned on the information of the other modality. It is one of the core novelties of our proposed framework. (3) Extensive experiments and ablation studies are performed to examine the effectiveness of the proposed DFAF framework, in which state-of-the-art VQA performance is achieved by our proposed DFAF framework.

2. Related Work

Representation learning for VQA. The recent boost of VQA performance is due to the success of deep representation learning. In the early stage of VQA methods, the VGG [35] network was commonly used. With the introduction of ResNet [13], the VQA community shift to ResNet networks, which outperform VGG by large margins. Recently, the bottom-up and top-down network [1] derived from faster RCNN [33] are shown to be suitable for VQA and image captioning tasks. Feature learning is an essential component for the development of VQA algorithms.

Bilinear Fusion for VQA. Solving VQA requires un-

derstanding of visual and language contents as well as the relation between them. In early VQA methods, simple concatenation or element-wise multiplication [45] between visual and language are used for cross-modal feature fusion. To capture the high-level interactions between the two modalities, Bilinear Fusion [9] has been proposed to adopt bilinear pooling to fuse features from the two modalities. To overcome the limitation of high computational cost of bilinear pooling, many approximated fusion methods, including MCB [7], MLB [20] and MUTAN [4], were proposed, which have shown better performance than bilinear fusion [9] with much fewer parameters.

Self-attention-based methods. The attention mechanism in deep learning tries to mimic how human vision works. By automatically ignoring irrelevant information from the data, neural networks can selectively focus on important features. This approach has achieved great success in Natural Language Processing (NLP) [3], image captioning [40] and VQA [46]. There are many variants of the attention mechanism. Our approach are mainly motivated by self-attention and co-attention based methods. The self-attention mechanism [37] transforms features into query, key and value features. The attention matrix between different features are then calculated by the inner product of query and key features. After acquiring the attention matrix, features are aggregated as the attention-weighted summation of the original features. Motivated by the self-attention mechanism, many vision tasks' performances were improved significantly. Non-local neural network [38] proposed a non-local module for aggregating information between different frames within one video and achieved state-of-the-art performance in video classification. Relation Network learn [14] the relationship between object proposals by adopting the self-attention mechanism. The in-place module can boost Faster RCNN [33] and Non-Maximum-Suppression (NMS) performance.

Co-attention-based methods. The co-attention based [39, 28] vision and language methods model the interactions across the two modalities. For each word, every image region features are aggregated to the word according to the co-attention weights. The co-attention mechanism has been widely used in NLP and VQA tasks. In [29], Dense Symmetric Co-attention (DCN) has been proposed. It achieved state-of-the-art performances on VQAv1 and VQAv2 datasets without using any bottom-up and top-down features. The success of DCN is due to dense concatenation [16] of symmetric co-attention.

Other works for language and vision tasks. Beyond above mentioned methods, many algorithms have also been proposed for fusion of cross-modal language and visual features. Dynamic Parameter Prediction [30] and Question-guided Hybrid Convolution [8] utilized dynamically predicted parameters for feature fusion. Adaptive attention [27] introduced a visual sensual which can skip attention during image captioning. Structured attention [21] adopted the MRF model over attention maps for better modelling better spatial attention distributions. Locally weighted deformable neighbours [18] proposed to predict offset and modulation weight.

3. Dynamic Fusion with Intra- and Inter-modality Attention Flow for VQA

3.1. Overview

The proposed approach consists of a series of DFAF modules. The whole pipeline is illustrated at figure 1 Visual and language features between the two modalities are first weighted with the co-attention mechanism and aggregated between the modalities to each image region and each word by the proposed Inter-modality Attention Flow (InterMAF) module, which learns the cross-modal interactions between the image regions and question words. Following the inter-modality module, to model the relationships within each modality, *i.e.*, word-to-word relations and region-to-region relations, the Dynamic Intra-modality Attention Flow (DyIntraMAF) module is adopted. It weights words and regions within each modality and aggregates their features to the words and regions again, which could be viewed as passing information flows within each modality. Importantly, in our proposed intra-modality module, the attention flows are dynamically conditioned on the information from the other modality, which is a key difference compared with existing self-attention based methods. Such InterMAF and DyIntraMAF modules could be stacked multiple times to pass the information flows among words and regions iteratively to model the latent alignments for visual question answering.

3.2. Base visual and language feature extraction

To obtain base visual and language features, we extract image features from bottom-up & top-down attention model [1]. The visual region features are obtained from a Faster RCNN [33] model pretrained on Visual Genome [23] dataset. For each image, we extract 100 region proposals and its associated region features. Given the input image I, the obtained region visual features are denoted as $R \in \mathbb{R}^{\mu \times 2048}$, where the i^{th} region feature is denoted as $r_i \in \mathbb{R}^{2048}$ and there are μ object regions in total. The object visual features are fixed during training.

We adopt GLoVe word embeddings [32] as the inputs of the Gated Recurrent Unit (GRU) [5] for encoding question word features. Given the question Q, we obtain word-level features $E \in \mathbb{R}^{14 \times 1280}$ from GRU, where the j^{th} word feature is denoted as $e_j \in \mathbb{R}^{1280}$ and all questions are padded and truncated to the same length 14.

The obtained visual object region features R and ques-

tion features E could be denoted as

$$R = \text{RCNN}(I; \theta_{\text{RCNN}}), \tag{1}$$

$$E = \text{GRU}(Q; \theta_{\text{GRU}}). \tag{2}$$

where visual feature parameters θ_{RCNN} are fixed while question features θ_{GRU} are learned from scratch and updated together when training our proposed framework.

3.3. Inter-modality Attention Flow

The Inter-modality Attention Flow (InterMAF) module as shown in Figure 1 first learns to capture the importance between each pair of visual region and word features. It then passes information flows between the two modalities according to the learned importance weights and aggregate features to update each word feature and image region feature. Such an information flow process is able to identify cross-modal relations between visual regions and words.

Given visual region and word features, we first calculate the association weights between every pair of visual region and word. Each visual region and word features are first transformed into query, key and value features following [34, 41], where the transformed region features are denoted as R_K, R_Q, $R_V \in \mathbb{R}^{\mu \times dim}$; Transformed word features are denoted as E_K, E_Q and $E_V \in \mathbb{R}^{14 \times dim}$,

$$R_K = \text{Linear}(R; \theta_{RK}), \quad E_K = \text{Linear}(R; \theta_{EK}), \tag{3}$$

$$R_Q = \text{Linear}(R; \theta_{RQ}), \quad E_Q = \text{Linear}(E; \theta_{EQ}), \tag{4}$$

$$R_V = \text{Linear}(R; \theta_{RV}), \quad E_V = \text{Linear}(E; \theta_{EV}). \tag{5}$$

where "Linear" denotes a fully-connected layer with parameter θ, and dim represents the common dimension of transformed features from both modalities.

By calculating the inner product $R_Q E_K^T$ between every pair of visual region feature R_Q and word key feature E_K, we obtain the raw attention weights for aggregating information from word features to each of the visual features, and vice versa. After normalizing the raw weights with the square root of the dimension number and a softmax non-linearity function, we obtain the two sets of attention weights, $\text{InterMAF}_{R \leftarrow E} \in \mathbb{R}^{\mu \times 14}$ for weighting information flow transmitted from words to image regions, and $\text{InterMAF}_{R \rightarrow E} \in \mathbb{R}^{14 \times \mu}$ for weighting information flow transmitted from image regions to sentence words,

$$\text{InterMAF}_{R \leftarrow E} = \text{softmax}(\frac{R_Q E_K^T}{\sqrt{dim}}), \tag{6}$$

$$\text{InterMAF}_{R \rightarrow E} = \text{softmax}(\frac{E_Q R_K^T}{\sqrt{dim}}). \tag{7}$$

The inner product values are proportional to the dimension of hidden feature space, thus need to be normalized by the square root of hidden dimension. The softmax non-linearity function is applied row-wisely.

The two bi-directional InterMAF matrices capture the importances between every image region and word pairs. Take the $\text{InterMAF}_{R \leftarrow E}$ for example, each row stands for the attention weights between one visual region and all word embeddings. Information from all word embeddings to this one image region feature could be aggregated as the weighted summation of the word value features E_V. We denote the information flows to update visual region features and word features by the InterMAF module as $R_{update} \in \mathbb{R}^{\mu \times dim}$ and $E_{update} \in \mathbb{R}^{14 \times dim}$, respectively,

$$R_{update} = \text{InterMAF}_{R \leftarrow E} \times E_V, \tag{8}$$

$$E_{update} = \text{InterMAF}_{R \rightarrow E} \times R_V. \tag{9}$$

where E_V and R_V are the un-weighted information flows(value features) to update visual region features and word features in Eq. (5), and the two InterMAF matrices are used to weight such information flows.

After acquiring the updated visual and word features, we concatenate them with original visual features R and word features E. A fully connected layer is utilized to transform the concatenated features into output features,

$$R = \text{Linear}([R,\ R_{update}]^T; \theta_{RT}), \tag{10}$$

$$E = \text{Linear}([E,\ E_{update}]^T; \theta_{ET}). \tag{11}$$

The output features by the InterMAF module would then be fed into the following Dynamic Intra-modality Attention Flow module for learning intra-modality information flows to further update the visual region and word features for capturing region-to-region and word-to-word relations.

3.4. Dynamic Intra-modality Attention Flow

The input visual regions and word features of DyIntraMAF have encoded cross-modal relations between visual regions and words. However, we argue that relationships within each modality are complementary to the cross-modal relations and should be taken into account for improving the VQA accuracy. For example, for the question, "who is above the skateboard ?", the intra-modality module should relate the region above the skateboard and the skateboard region to infer the final answer. Therefore, we propose the Dynamic Intra-modality Attention Flow (DyIntraMAF) module for modelling such within-modality relations with a dynamic attention mechanism. The implementation of DyIntraMAF is illustrated at Figure 2.

The naive intra-modality matrices to capture the importance between regions and between words could be defined similarly to Eq. (5) as,

$$\text{IntraMAF}_{R \leftarrow R} = \text{softmax}(\frac{R_Q R_K^T}{\sqrt{dim}}), \tag{12}$$

$$\text{IntraMAF}_{E \leftarrow E} = \text{softmax}(\frac{E_Q E_K^T}{\sqrt{dim}}). \tag{13}$$

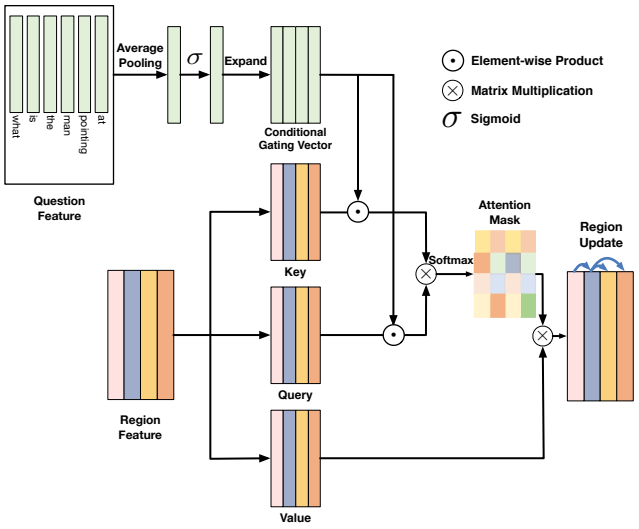

Figure 2: Illustration of the proposed Dynamic Intra-Modality Attention Flow module. Only intra-modality attention flow within the visual modality conditioned on question are shown. By average pooling over question features, the conditional gating vector can be acquired for controlling the information flows among region features. Attention will focus on question related information flows. Row-wise softmax is applied to obtain the attention weight.

The dot products are utilized to estimate their within-modality importance between the same modality's query and key features. Such weight matrices could then be used to weight the information flows transmitted within each modality. Modelling within-modality relationships have been shown to be effective in object detection [14], image captioning and BERT word embedding pretraining [6].

However, the naive IntraMAF module only utilizes within-modality information for estimating the region-to-region and word-to-word importance. Some relations are important but could only be identified conditioned on information from the other modality. For instance, even for the same input image, relations between different visual region pairs should be weighted differently according to different questions. We therefore propose a Dynamic Intra-modality Attention Flow (DyIntraMAF) module to estimate within-modality relation importance conditioned on the information from the other modality.

To summarize the conditioning information from the other modality, we average pool the visual region features along the object-index dimension and the word features along the word-index dimension. The average pooled features of both modalities are then transformed to a dim-dimensional feature vector to match the dimension of the query and key features R_Q, R_K, E_Q, E_K. The dim-dimensional feature vector of each modality is then processed by a sigmoid non-linearity function $\sigma(\cdot)$ to generate channel-wise conditioning gates for the other modality,

$$G_{R\to E} = \sigma(\text{Linear}(\text{Avg\_Pool}(R); \theta_{RP})), \quad (14)$$

$$G_{R\leftarrow E} = \sigma(\text{Linear}(\text{Avg\_Pool}(E); \theta_{EP})). \quad (15)$$

The query and key features from both modalities are then modulated by the conditional gates from the other modality

$$\hat{R}_Q = (1 + G_{R\leftarrow E}) \odot R_Q, \;\; \hat{R}_K = (1 + G_{R\leftarrow E}) \odot R_K,$$
$$\hat{E}_Q = (1 + G_{R\leftarrow E}) \odot R_Q, \;\; \hat{E}_K = (1 + G_{R\leftarrow E}) \odot E_K. \quad (16)$$

where $\odot$ denotes element-wise multiplication. Channels of query and key features would be activated or deactivated by channel-wise gates conditioned on the other modality. Such a design of the two gating vectors share the similar spirit with Squeeze and Excitation Network [15] and the Gated Convolution [10]. The key difference is that the channel-wise gating vector is created based on cross-modal information.

The dynamic intra-modality attention flow matrices $\text{DyIntraMAF}_{R\leftarrow R} \in \mathbb{R}^{\mu\times\mu}$ and $\text{DyIntraMAF}_{E\leftarrow E} \in \mathbb{R}^{14\times 14}$ are then obtained by the gated query and key features to weight different within-modality relations,

$$\text{DyIntraMAF}_{R\leftarrow R} = \text{softmax}(\frac{\hat{R}_Q \hat{R}_K^T}{\sqrt{dim}}), \quad (17)$$

$$\text{DyIntraMAF}_{E\leftarrow E} = \text{softmax}(\frac{\hat{E}_Q \hat{E}_K^T}{\sqrt{dim}}). \quad (18)$$

Visual region and word features are then updated by the weighted value features R_V and E_V via residual,

$$R = \text{Linear}(R + R_{update}; \theta_{RD}), \quad (19)$$

$$E = \text{Linear}(E + E_{update}; \theta_{ED}). \quad (20)$$

where

$$R_{update} = \text{DyIntraMAF}_{R\leftarrow R} \times R_V, \quad (21)$$

$$E_{update} = \text{DyIntraMAF}_{E\leftarrow E} \times E_V. \quad (22)$$

Note that here we only make key and query features conditioned on the other modality to adaptively weight within-modality information flows. In our ablation studies, we observe that the proposed DyIntraMAF module by a large margin outperforms the naive IntraMAF module.

3.5. The Framework with Intra- and Inter-modality Attention Flow

In this section, we introduce how to integrate intra- and inter-modality attention flow modules into our proposed framework. The whole pipeline is illustrated in Figure 1.The proposed framework first extracts visual region

features and word features from the input image and question by utilizing the Faster RCNN and GRU models, respectively. Faster R-CNN model weights are fixed during training our proposed framework, while GRU weights are updated with our framework from scratch.

After visual region features and word features being transformed into vectors of the same dimension by fully connected layers. The InterMAF module passes information flows between each pair of visual region and question word and aggregates updated features to each region and each word. Such aggregated features integrate information from the other modality to update the visual and word features according to the cross-modal relations.

Given the InterMAF outputs, the DyIntraMAF module is utilized for dynamically passing information flows within each modality. The visual region and word features would be updated again with information within the same modality via residual connections.

We use one InterMAF module followed by one DyIntraMAF module to form a basic block in our proposed DFAF framework. Multiple blocks could be stacked thanks to the feature concatenation and residual connection in the feature updating procedures. Very deep intra- and inter-modality information flows can be effectively trained with stochastic gradient descent. In addition, we utilize multi-head attention in practice. The original features are split along channel dimensions into groups and different groups would generate parallel attentions to update visual and word features in different groups independently.

3.6. Answer Prediction Layer and Loss Function

After several blocks of feature updating by InterMAF and DyIntraMAF modules, we obtain the final visual region and word features encoding inter-modality and intra-modality relations for VQA. By average pooling over region features and over word features, we could obtain discriminative representations for image and question, respectively. Such features could then be fused via either feature concatenation, or feature element-wise product, or feature addition to obtain fused features. We experiment with the three fusion approaches in which the element-wise product between visual and language representations achieves the best performance with a trivial margin.

Similar to state-of-the-art VQA approaches, we treat VQA as a classification problem. The fused multi-modal features are transformed into a probability vector by a 2-layer multi-layer perceptron with ReLU non-linearity function between the layers and a final softmax function. The ground-truth answers are extracted from annotated answers that appear for more than 5 times. Cross-entropy loss function is adopted as the objective function.

4. Experiments

4.1. Datasets

We used VQA version 2.0 [11] for our experiment. VQA dataset contains human annotated question-answer pairs for images from Microsoft COCO dataset [25]. VQA 2.0 is an updated of previous VQA 1.0 with much more annotations and less dataset bias. VQA 2.0 is split into train, validation and test-standard sets. Among test-standard test, 25% are served as test-dev set. All questions types are divided into Yes/No, Number and other categories. Train, validation and test-standard contains 82,783, 40,504 and 81,434 images, with 443,757, 214,354, 447,793 questions,respectively. Each question contains 10 answers from different annotators. Answers with the highest frequency are treated as the ground-truth. Following previous approaches, we perform ablation studies over the validation set and utilize the train and validation splits for test.

4.2. Experimental Setup

Visual features of dimension 2048 are extracted from Faster R-CNN [33] while word features are encoded into features of dimension 1280 by GRU[5]. The visual features and word features are then embedded into 512 dimensions by a fully-connected layer, respectively. Inside InterMAF, features are transformed into 8 multi-head attention with 64 dimensions for each head. For DyIntraMAF, the average pooled features from both modality are transformed into 512 dimensions by MLP followed by element-wise sigmoid activation to obtain the conditioning gating vectors. They are then multiplied with 512 dimension visual key and query features at every position of visual and word features for dynamic attention flows. Previous approaches achieve significantly better results with sentinel and relative position information. However, sentinel and relative position do not affect the performance of our method.

All fully connected layers have the same dropout rate 0.1. All gradients are clipped to 0.25. Batch size is set as 512. Adamax optimizer [22], a variant of Adam, is used. The learning rate is set as 10^{-3} for the first 2 epoch, 2×10^{-3} for the next 8 epochs and decayed by 1/4 for the rest epochs. Our method is implemented with Pytorch [31]. All initilizations are Pytorch default initilization.

All ablation studies are conducted on the validation dataset, while train, validation datasets and extra visual genome dataset are combined for testing on test-dev.

4.3. Ablation study of DFAF

We perform extensive ablation studies on the VQA 2.0 validation dataset [11]. The results are shown in Table 1. Our default setting only has 1 block of DFAF module. Region features with 2,048 dimensions are extracted from the input image by Faster RCNN [33], word features with 1,024

Component	Setting	Accuracy
Bottom-up [1]	Bottom-up	63.37
Bilinear Attention [19]	BAN-1	65.36
	BAN-4	65.81
	BAN-12	66.04
Default	DFAF-1	66.21
# of stacked blocks	DFAF-2	66.43
	DFAF-5	66.58
	DFAF-8	**66.66**
Attention type	InterMAF only	64.37
	IntraMAF only	62.34
	DyIntraMAF only	65.51
	InterMAF + DyIntraMAF	**66.21**
Attention Direction inside InterMAF	Parallel	65.99
	$R \rightarrow E$, $E \rightarrow R$	**66.21**
	$E \rightarrow R$, $R \rightarrow E$	66.19
Embedding dimension	512	**66.21**
	1024	65.89
Cross-model feature fusion	Multiplication	**66.21**
	Addition	66.11
	Concatenation	66.14
Visual Sentinel	None	**66.21**
	1	66.01
	3	66.02
Bounding Box Embedding	None	**66.21**
	Absolute Position	65.88
	Relative Position	65.23
Parallel Heads	1 head each 512	65.84
	4 heads each 128	66.17
	8 heads each 64	**66.21**

Table 1: Ablation studies of our proposed DFAF on VQA 2.0 validation dataset. R stands for region features while E stands for word embedding features

dimensions are extracted by GRU [5]. By default, all modules inside DFAF has 512 dimensions. In the final fusion layer, feature multiplication is employed, which shows a trivial improvement. Visual sentinel [27] and bounding box position embedding are also tested which give a slight drop in the final performance. 8 parallel attention heads with dimensions 64 for each head is utilized in the default setting.

We first investigate the influence of the number of stacked DFAF blocks. The default setting has one stack. As one can see from Table 1, more stacks can improve the performance thanks to the residual connection [13]. Different from ResNet, we do not employ any normalization [17] technique during residual connection. The performance of single layer DFAF is comparable with BAN-12 [19].

Then, we investigate the influence of attention type. The attention mechanism in Bottom up [1] utilizes simple attention methods. Bilinear attention network [19] proposed a bilinear attention which learns the joint attention distribution between each word and region pairs. By adding the InterMAF, performance can improve by 1% because of the modelling the inter-modality relations between image regions and question words. Integrating only the IntraMAF module would harm the performance because too many unrelated information flows hinder the learning process. By adding dynamically conditioned information flow DyIntra MAF module, we achieve a 2.15% performance improvement. By combining Intra- and Inter-modality attention flows, we significantly outperform the baseline [1] by 2.83% and previous state-of-the-art BAN-1 [19] by 0.85%.

There are several orders for passing information within the InterMAF module, namely, parallel and sequential [39, 28]. For parallel InterMAF, both region and word features are updated at the same time. For the sequential information flow, we experiment with passing attention flow from regions to words first, which updates word features, and then passing message from words to regions, which then update region features, and vice versa. We denote the first sequential order as $R \rightarrow E$, $E \rightarrow R$, and the second one as $E \rightarrow R$, $R \rightarrow E$. Sequential update outperforms parallel update way, while the specific order does not matter.

Next, we perform ablation study on the influence of embedding dimension and cross-model feature fusion. 512 dimensions result in better performance than 1024 dimensions. For the fusion method, multiplication shows a slight better performance than feature addition and concatenation.

Visual sentinel [27, 39] were utilized in many previous VQA methods, which was shown to increase the VQA accuracy. We treat sentinel as a general 512 dimension features and concatenate sentinel with all region and word features. Previous μ region features and 14 word features become into $\mu + 1$ and 15 respectively. In our experiments, adding visual sentinel do not show improvement.

The positions of bounding boxes were widely utilized as a part of image region features in previous methods. Absolute position embedding has been employed in Transformer [37], BERT [6] and Gated CNN [10] in NLP. Relative position was adopted in relation network [14] for object detection. In our experiment, adding absolute or relative positions would drop the performance.

At last, we experiment on the influence of multi-head attention [37]. We keep the overall dimensions to be 512. 1, 4 and 8 attention heads are experimented. As can be seen in Table 1, 8 attention can achieve better performance at the same number of parameters.

4.4. Visualisation of the proposed Attention Flow Weights

In Figure 3, we visualise the intra-modality attention flow weights to analyse VQA model. The attention weights modulate information flow from contextual regions(orange, blue and green) to center region(red). The left column

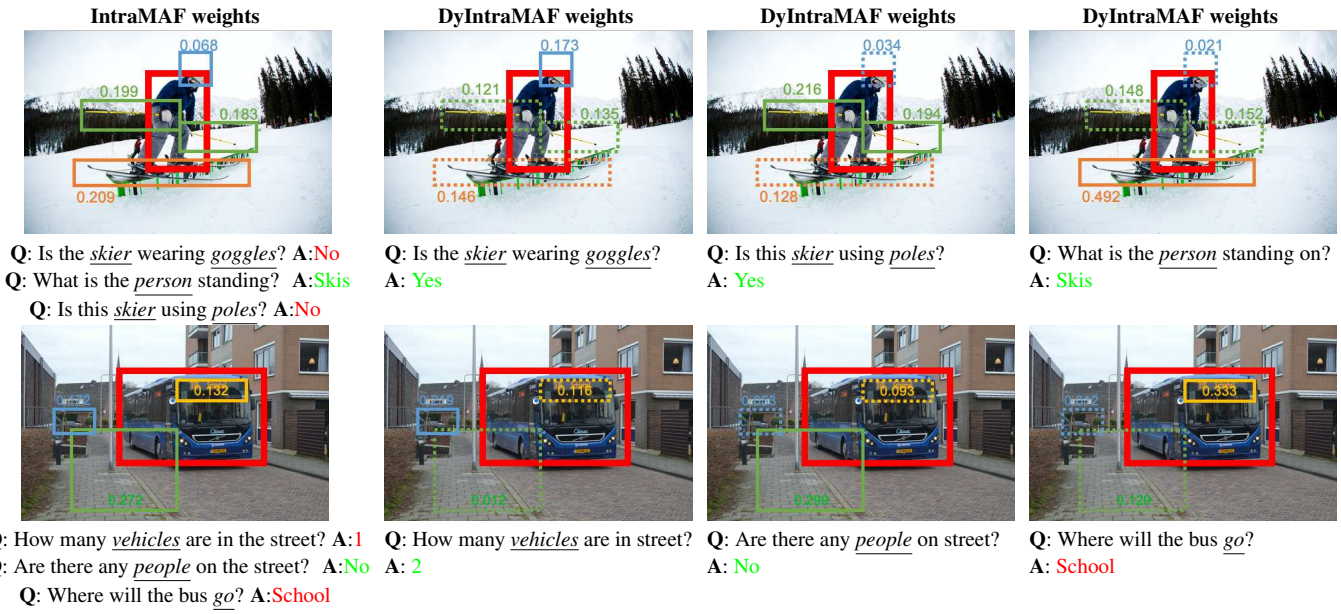

Figure 3: Visualisation of IntraMAF and DyIntraMAF attention weights between central region(red) and other related regions. (Left) The IntraMAF module treats different questions equally and generate uninformative weight for different questions. (Right) The proposed DyIntraMAF module dynamically changes attention weights according to input questions.

stands for the attention flow weights in the IntraMAF module. While the rest columns represent dynamic attention flow weights in the DyIntraMAF module. In the DyIntraMAF module, unrelated information flow are filtered out by question features and thus generate the correct answer.

4.5. Comparison with State-of-the-arts methods

Table 2 shows the performance of our proposed algorithm trained with extra visual genome dataset and state-of-the-art methods on VQA. Bottom up in Table 2 is the winner of VQA challenge 2017. This approach proposed to use features based on Faster RCNN [33] instead of ResNet [13]. Multi-modal Factorized High-order Pooling (MFH) [43] is a state-of-the-art bilinear pooling methods. Dense Co-Attention Network (DCN) [29] utilized dense stack of multiple layers of Co-attention mechanism which significantly outperform previous methods with ResNet features. Counting methods [44] are good at counting questions by utilizing the information of bounding boxes. Bilinear Attention Network (BAN) [19] is a state-of-the-art approach on VQA 2.0 which has 12 stacked blocks of BAN modules.

Model	test-dev				test-std
	Y/N	No.	Other	All	All
Bottom-up [1]	81.82	44.21	56.05	65.32	65.67
MFH [11]	n/a	n/a	n/a	66.12	n/a
DCN [29]	83.51	46.61	57.26	66.87	66.97
Counter [44]	83.14	51.62	58.97	68.09	68.41
MFH+Bottom-Up [11]	84.27	49.56	59.89	68.76	n/a
BAN+Glove [19]	85.46	50.66	**60.50**	69.66	n/a
DFAF(ours)	**86.09**	**53.32**	60.49	**70.22**	**70.34**

Table 2: Comparison with previous state-of-the-art methods on VQA 2.0 test dataset.

5. Conclusions

In this paper, we proposed a novel framework Dynamic Fusion with Intra- and Inter-modality Attention Flow (DFAF) for visual question answering. The DFAF framework alternatively passes information within and across different modalities based on an inter-modality and intra-modality attention mechanisms. The information flow inside visual features are dynamically conditioned on the question features. Stacking multiple blocks of DFAF are shown to improve the performance of VQA.

6. Acknowledgment

This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616, CUHK14208417, CUHK14239816, and in part by CUHK Direct Grant.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and

visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[4] Hedi Ben-younes, Remi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *IEEE International Conference on Computer Vision*, 2017.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016.

[8] Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven CH Hoi, and Xiaogang Wang. Question-guided hybrid convolution for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 469–485, 2018.

[9] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, 2016.

[10] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252, 2017.

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2018.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[18] Zhengkai Jiang, Peng Gao, Chaoxu Guo, Qian Zhang, Shiming Xiang, and Chunhong Pan. Video object detection with locally-weighted deformable neighbors. 2019.

[19] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018.

[20] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations*, 2017.

[21] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. In *International Conference on Learning Representations*, 2017.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[24] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1908–1917. IEEE, 2017.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[26] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *European Conference on Computer Vision*, pages 353–369. Springer, 2018.

[27] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3242–3250. IEEE, 2017.

[28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[29] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric

co-attention for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[30] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 30–38, 2016.

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *The Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[34] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *The Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 464–468, 2018.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Shuyang Sun, Jiangmiao Pang, Jianping Shi, Shuai Yi, and Wanli Ouyang. Fishnet: A versatile backbone for image, region, and pixel level prediction. In *Advances in Neural Information Processing Systems*, pages 760–770, 2018.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[39] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *International Conference on Learning Representations*, 2017.

[40] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[41] Kui Xu, Zhe Wang, Jiangping Shi, Hongsheng Li, and Qiangfeng Cliff Zhang. A2-net: Molecular structure estimation from cryo-em density volumes. *arXiv preprint arXiv:1901.00785*, 2019.

[42] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018.

[43] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, (99):1–13, 2018.

[44] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations*, 2018.

[45] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[46] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1300–1309. IEEE, 2017.