Densely Connected Attention Flow for Visual Question Answering

Fei Liu[1,2] , **Jing Liu**[1*] , **Zhiwei Fang**[1,2] , **Richang Hong**[3] , **Hanqing Lu**[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]School of Computer and Information, Hefei University of Technology
liufei2017@ia.ac.cn, {jliu, zhiwei.fang}@nlpr.ia.ac.cn, hongrc.hfut@gmail.com, luhq@nlpr.ia.ac.cn

Abstract

Learning effective interactions between multi-modal features is at the heart of visual question answering (VQA). A common defect of the existing VQA approaches is that they only consider a very limited amount of interactions, which may be not enough to model latent complex image-question relations that are necessary for accurately answering questions. Therefore, in this paper, we propose a novel DCAF (*Densely Connected Attention Flow*) framework for modeling dense interactions. It densely connects all pairwise layers of the network via *Attention Connectors*, capturing fine-grained interplay between image and question across all hierarchical levels. The proposed *Attention Connector* efficiently connects the multi-modal features at any two layers with symmetric co-attention, and produces interaction-aware attention features. Experimental results on three publicly available datasets show that the proposed method achieves state-of-the-art performance.

1 Introduction

Recently, the Visual Question Answering (VQA) task has gained increasing attention in both computer vision and natural language processing communities. It aims at answering a natural language question about a given image. The progress of VQA has been mainly brought about by two lines of works, the development of better attention mechanisms and the improvement in multi-modality fusion approaches.

Attention mechanisms have been widely used in VQA, and a number of methods have been developed so far. These methods are categorized into two classes. One is the class of methods that use the question as guidance to generate attention on image regions (*i.e.* question-guided attention). For instance, Yang *et al.* [Yang *et al.*, 2016] proposed stacked attention network that produces multiple attention maps on the image sequentially. Kim *et al.* [Kim *et al.*, 2016] extended this idea by introducing residual learning to produce better attention. The other is the class of methods that additionally consider image-guided attention on question words (*i.e.* co-attention) [Lu *et al.*, 2016; Nguyen and Okatani, 2018]. Co-attention mechanisms can model bidirectional inter-modality information flow, thus perform better than question-guided attention mechanisms. In particular, Nguyen *et al.* [Nguyen and Okatani, 2018] proposed a novel co-attention mechanism that can model interactions between *any* image region and *any* question word, achieving state-of-the-art performance on VQA 1.0 and VQA 2.0 datasets.

Meanwhile, multimodal fusion approaches have been explored extensively. In early studies, simple fusion methods such as the element-wise add, product, and concatenation of the visual and language features are employed. To capture high-level interactions between the two modalities, some bilinear pooling methods, including MCB [Fukui *et al.*, 2016], MLB [Kim *et al.*, 2017], MUTAN [Ben-Younes *et al.*, 2017] and MFB [Yu *et al.*, 2017], were proposed. In [Nguyen and Okatani, 2018], the authors proposed Dense Co-attention Network (DCN), which fuses multi-modal features by multiple applications of the symmetric co-attention.

We point out that the existing VQA approaches only perform a limited amount of interactions between language and vision domains. Some methods [Fukui *et al.*, 2016; Kim *et al.*, 2017; Yu *et al.*, 2017; Anderson *et al.*, 2018] perform only one interaction (*i.e.* multimodal fusion) in the latter model stage, and the fused features are then fed into the classifier to obtain the scores of candidate answers. The other methods [Yang *et al.*, 2016; Nguyen and Okatani, 2018; Gao *et al.*, 2018] perform multi-step interactions by stacking several interaction modules. However, such a design of stacking multiple modules may impair gradient flow and feature propagation, making networks harder to optimize. Consequently, at most four stacked modules are employed in these methods, which perform at most four interactions between the two modalities. We argue that this can be a significant limitation of the existing approaches. Limited interactions possibly fail to model complex image-question relations that are necessary for answering questions correctly.

Motivated by this, we propose a novel Densely Connected Attention Flow (DCAF) framework for modeling dense interactions between the visual and language modalities (see Fig. 1). Our DCAF framework possesses several additional image and question encoding layers in the early model stage. These layers produce image and question features of different hier-

*Jing Liu is the corresponding author.

archies, and are densely connected, connecting every layer of question encoding branch with every layer of image encoding branch via Attention Connector (AC). In the AC module, the symmetric co-attention [Nguyen and Okatani, 2018] is performed to enable every interaction between *any* image region and *any* question word. The attention outputs of all AC modules are collectively propagated to the end of the encoding layers, and concatenated with the summation of all hierarchical features. Dense connections would incur a massive increase in the number of the attention outputs, thus increase in the representation size in subsequent layers (due to the *concat* operation). To this end, we compress the attention outputs so that they can be small enough to propagate. Our design of DCAF not only greatly increases the number of interaction interfaces between *question* and *image* but also facilitates information flow and gradient flow.

Our contributions are summarized as follows: **(1)** We propose a novel Densely Connected Attention Flow (DCAF) framework for visual question answering. It can perform dense multi-modal interactions, and capture fine-grained multi-modal information. The dense interactions structure is shown to be superior to other interactions structures (*e.g.* sequential interactions structure). **(2)** We propose efficient Attention Connector (AC) to connect two modalities, modeling fine-grained interplay between image and question via co-attention. AC also serves as skip-connector, which effectively connects shallow layers to deeper layers. **(3)** Extensive experiments conducted on the VQA 1.0, VQA 2.0 and TDIUC datasets show the effectiveness of the proposed DCAF. Our approach outperforms the state-of-the-art methods on the three datasets.

2 Related Work

2.1 Attention Mechanisms

Attention mechanisms allow the models to focus on the most relevant image regions and question words. Initially, [Chen *et al.*, 2015] proposed one-step attention to locate relevant image regions. Furthermore, [Yang *et al.*, 2016; Xu and Saenko, 2016] proposed multi-step attention to update relevant image regions and infer the answer progressively. Additionally, [Lu *et al.*, 2016; Nguyen and Okatani, 2018] proposed co-attention, which locates not only the relevant image regions but also question words. Recently, [Fukui *et al.*, 2016; Kim *et al.*, 2017] used bilinear fusion in attention mechanisms to generate more accurate attention weights. Different from these works, we embed the co-attention mechanism into Attention Connectors and apply them across all hierarchical levels. Besides, we also employ the co-attention mechanism similar to [Yu *et al.*, 2017] in the latter model stage.

2.2 Multimodal Fusion

In early VQA methods, simple concatenation or element-wise product between visual and language features are used for multi-modal feature fusion. Recently, bilinear pooling methods are introduced for VQA to capture high-level interactions between visual and language features. Multimodal Compact Bilinear Pooling (MCB) [Fukui *et al.*, 2016] projects the visual and language features into a higher dimensional space

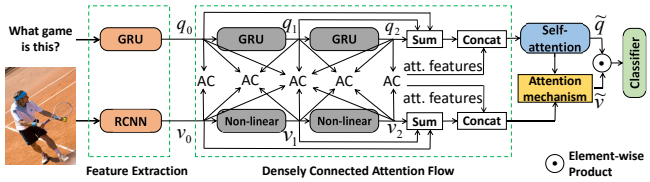

Figure 1: Overview of the proposed Densely Connected Attention Flow (DCAF) framework. AC denotes Attention Connector, which connects the multi-modality features with attention and outputs attention features. DCAF models dense interactions between vision and language by multiple AC modules. Note that the figure shows 2-layer DCAF.

and convolves them in the Fast Fourier Transform space. In Multimodal Low-rank Bilinear (MLB) [Kim *et al.*, 2017], the multi-modal features are obtained as the Hadamard product of the linear-projected visual and language features. Pointing out that MLB suffers from slow convergence rate, Yu *et al.* [Kim *et al.*, 2017] proposed Multimodal Factorized Bilinear (MFB) pooling, which computes fused features with a matrix factorization trick to reduce the number of parameters and improve convergence rate. Ben-younes *et al.* [Ben-Younes *et al.*, 2017] proposed the Multimodal Tucker Fusion (MUTAN), which unifies MCB and MLB into the same framework. The weights are decomposed according to the Tucker decomposition. MUTAN achieves better performance than MLB and MCB with fewer parameters. In [Nguyen and Okatani, 2018], the authors proposed a novel co-attention mechanism for improved fusion of visual and language features. It considers every interaction between any image region and any question word, and can be stacked to enable muti-step interactions between the image-question pair. Unlike the above methods that only perform a limited amount of interactions sequentially, our method models cross-modal interactions between any two hierarchical levels, forming dense interactions.

3 Proposed Approach

Our DCAF framework for VQA is depicted in Fig. 1. Given the input question and image, the *feature extraction* module produces the initial question and image features. The DCAF module further abstracts the features through several encoding layers, and performs the interaction between question and image features at arbitrary layers. Following the DCAF module, a *self-attention* mechanism is used to learn the attention weights of every word in the question, and obtain the attended question representation (using a weighted sum of the word vectors). Then, a *visual attention* mechanism is performed to produce the attended image representation. The two representations are merged via element-wise product, then fed into the *classifier* to determine the final answer prediction.

3.1 Feature Extraction

Similar to [Anderson *et al.*, 2018], we extract the initial image features using Faster RCNN [Ren *et al.*, 2015]. We select a total of top 36 object proposals whose 2048 dimensional feature vectors are obtained from the ROI pooling layer in the Region Proposal Network. The obtained visual region features (with ℓ_2 normalization) are denoted as $v_0 \in \mathbb{R}^{2048\times 36}$.

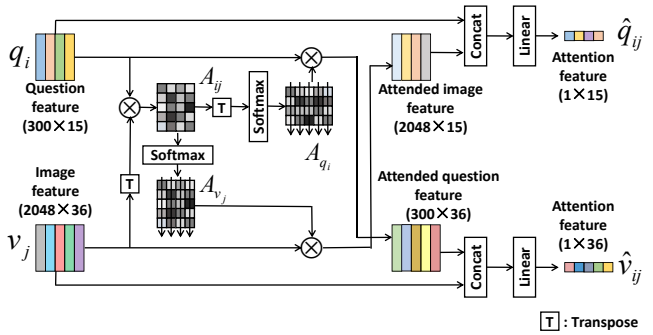

Figure 2: Illustration of the proposed Attention Connector (AC). AC takes as inputs question and image features from any two layers, then performs co-attention mechanism, and finally produces the compressed attention features that can be propagated to deeper layers of the network.

To extract the question features, we first pad or truncate all questions to the same length 15. Then, each word is embedded into a 300-dimensional GloVe vector [Pennington *et al.*, 2014]. The word embeddings of the question are inputted into a two-layer GRU with residual connections and batch normalization on the output of each layer, producing the initial question representation $q_0 \in \mathbb{R}^{300\times15}$.

3.2 Densely Connected Attention Flow

The DCAF module is the core part of our proposed framework. It has some encoding layers. We adopt a GRU with 300 hidden units as the question encoding layer and the non-linear layer used in [Anderson *et al.*, 2018] as the image encoding layer. The *i*-th non-linear layer is defined as follows:

$$v_i{}' = \tanh(W_i v_{i-1}), \quad g = \text{sigmoid}(W_i{}' v_{i-1}) \tag{1}$$

$$v_i = v_i{}' \odot g \tag{2}$$

where $W_i, W_i{}' \in \mathbb{R}^{2048\times2048}$ are learned weights, $\odot$ represents element-wise product, and $v_{i-1}, v_i \in \mathbb{R}^{2048\times36}$ are the input and output of the *i*-th non-linear layer respectively. The matrix g acts as a gate on the intermediate activation $v_i{}'$, which is inspired by similar gating operations within GRUs and LSTMs. For clearness, we do not explicitly represent the bias term in our paper.

Starting from the initial image and question features, the encoding layers in two branches generate a series of hierarchical features, $v_0, v_1, ..., v_k$ and $q_0, q_1, ..., q_k$, respectively. k denotes the number of the layers. We apply our Attention Connectors to densely connect all hierarchical features of image and question:

$$\hat{q}_{ij}, \hat{v}_{ij} = \text{AC}(q_i, v_j) \quad \forall\, i, j = 0, 1, \cdots, k \tag{3}$$

where AC represents the Attention Connector module (see Fig.2; explained later). $\hat{q}_{ij} \in \mathbb{R}^{1\times15}$ and $\hat{v}_{ij} \in \mathbb{R}^{1\times36}$ represent the generated attention features for each ij connection. In total, we obtain $(k+1)^2$ attention features for each word and each region. Intuitively, these features capture fine-grained relationships between the image and question at different stages of the network flow. For each branch, we sum up *all* hierarchical features, then concatenate them with *all* attention features as the final output. These skip connections share the similar motivation with DenseNet [Huang *et al.*, 2017]. The process is denoted as

$$q_{sum} = q_0 + \alpha_1 q_1 + \cdots + \alpha_k q_k \tag{4}$$

$$v_{sum} = v_0 + \beta_1 v_1 + \cdots + \beta_k v_k \tag{5}$$

$$q_{out} = [q_{sum}\,;\, \hat{q}_{00}\,;\, \hat{q}_{01}\,;\, \cdots\,;\, \hat{q}_{kk}] \tag{6}$$

$$v_{out} = [v_{sum}\,;\, \hat{v}_{00}\,;\, \hat{v}_{01}\,;\, \cdots\,;\, \hat{v}_{kk}] \tag{7}$$

where $\alpha_1, \alpha_2 \cdots \alpha_k, \beta_1, \beta_2 \cdots \beta_k$ are learned scale parameters initialized as 0, and $[\cdot\,;\cdot]$ represents the concatenation manipulation. $q_{out} \in \mathbb{R}^{[300+(k+1)^2]\times15}$ and $v_{out} \in \mathbb{R}^{[2048+(k+1)^2]\times36}$ are the outputs of the DCAF module.

Attention Connector

The proposed Attention Connector (AC) is illustrated in Fig. 2. It connects any two encoding layers of image and question in the DCAF module, and captures fine-grained interplay between image and question. The initial step in this module is a symmetric co-attention mechanism. Given q_i and v_j, an affinity matrix is first constructed via:

$$A_{ij} = {v_j}^\top W_{ij} q_i \tag{8}$$

where $W_{ij} \in \mathbb{R}^{2048\times300}$ is a learnable weight matrix. Next, we derive attention maps on question words and attention maps on image regions:

$$A_{q_i} = \text{softmax}(A_{ij}^\top), \quad A_{v_j} = \text{softmax}(A_{ij}) \tag{9}$$

Note that each column of A_{q_i} and A_{v_j} contains a single attention map. The attended representations are computed as follows:

$$\tilde{q}_{ij} = q_i \otimes A_{q_i}, \quad \tilde{v}_{ij} = v_j \otimes A_{v_j} \tag{10}$$

where $\otimes$ represents matrix multiplication. $\tilde{q}_{ij} \in \mathbb{R}^{300\times36}$ and $\tilde{v}_{ij} \in \mathbb{R}^{2048\times15}$ are the attended question feature and image feature respectively. In practice, we use multi-glimpse attention mechanism. Specifically, we construct multiple affinity matrixes, leading to multiple attended features. We *average* the multiple attended features to obtain the final attended features. In Sec. 4.2, we experiment with different number of glimpses to obtain the optimal setting.

The attended features are fused with the original features of the other modality by concatenation and then compressed to low-dimensional space by a single linear layer (FC):

$$\hat{q}_{ij} = \text{FC}([q_i\,;\,\tilde{v}_{ij}]), \quad \hat{v}_{ij} = \text{FC}([v_j\,;\,\tilde{q}_{ij}]) \tag{11}$$

where $\hat{q}_{ij}, \hat{v}_{ij}$ are the outputs of AC. We compress the features so that the representation size in subsequent layers would not increase massively, as mentioned in Sec. 1. The output dimension of each word and each region is set to 1, as it performs best in our experiments (see Table 1).

3.3 Attention Mechanisms & Answer Prediction

In this section, we introduce the attention mechanisms (including the textual self-attention and visual spatial attention) and the answer prediction.

Given the outputs q_{out} and v_{out} of the DCAF module, we first perform self-attention mechanism on q_{out} to obtain aggregated representation of the whole question:

$$s^Q = W_2^Q \text{ReLU}(W_1^Q q_{out}) \tag{12}$$

$$\alpha^Q = \text{softmax}(s^Q) \tag{13}$$

$$\widetilde{q} = \textstyle\sum_{i=1}^{15} \alpha_i^Q q_{out_i} \tag{14}$$

where $q_{out} = \{q_{out_1}, q_{out_2}, ..., q_{out_{15}}\} \in \mathbb{R}^{[300+(k+1)^2]\times 15}$, W_1^Q, W_2^Q are learnable weights, s^Q is the scores of words, α^Q is attention weights, and $\widetilde{q} \in \mathbb{R}^{300+(k+1)^2}$ is vector representation of the question.

We then perform visual attention mechanism on v_{out} to obtain aggregated representation of the image:

$$s_i^I = W^I f_a([v_{out_i}\,;\,\widetilde{q}]) \tag{15}$$

$$\alpha^I = \mathrm{softmax}(s^I) \tag{16}$$

$$\widetilde{v} = \textstyle\sum_{i=1}^{36} \alpha_i^I v_{out_i} \tag{17}$$

where $v_{out} = \{v_{out_1}, v_{out_2}, ..., v_{out_{36}}\} \in \mathbb{R}^{[2048+(k+1)^2]\times 36}$. $f_a(\cdot)$ denotes the *gated tanh* function [Anderson *et al.*, 2018] with parameters a. It is used to project the concatenated vector to 512-dimensional space. $W^I \in \mathbb{R}^{1\times 512}$ is learnable weights, s^I is the scores of image regions, α^I is attention weights, and $\widetilde{v} \in \mathbb{R}^{2048+(k+1)^2}$ is vector representation of the image.

After obtaining $\widetilde{q}$ and $\widetilde{v}$, we project them to the same dimensional space (512 dimensions) using two *gated tanh* functions with different parameters, respectively. The features are then fused via element-wise product. Similar to [Ben-Younes *et al.*, 2017], we treat VQA as a classification problem. The fused multi-modal features are fed into the *classifier* composed of 2-layer MLP with ReLU non-linearity function between the layers and a final softmax function, outputing a class probability vector. Cross-entropy loss is adopted as the objective function for training the VQA system.

4 Experiments

4.1 Setup

Datasets. We use the VQA 1.0 [Antol *et al.*, 2015], VQA 2.0 [Goyal *et al.*, 2017] and TDIUC [Kafle and Kanan, 2017] datasets for our experiments. VQA 1.0 is built from 204,721 MSCOCO images with human annotated questions and answers. The dataset is divided into three splits: *train* (248,349 questions), *val* (121,512 questions) and *test* (244,302 questions). VQA 2.0 is an updated version of VQA 1.0. It contains more samples (443,757 *train*, 214,354 *val*, and 447,793 *test* questions) and is more balanced in term of language bias. TDIUC is a larger dataset that contains 1,654,167 samples and 12 question types. For VQA 1.0 and 2.0, we use the evaluation protocol of [Antol *et al.*, 2015] to evaluate the model. For TDIUC, we calculate the simple accuracy for each question type and Arithmetic/Harmonic mean-per-type (MPT).

Implementation details. As in [Yu *et al.*, 2017], we choose the most frequent 3,000 answers in the *train* and *val* sets to form the set of candidate answers. The model is trained using the AMSGrad [Reddi *et al.*, 2018] optimizer with an initial learning rate of 6×10^{-4}. The batch size is set to 128.

4.2 Ablation Studies

We conduct ablation studies on the VQA 2.0 to investigate factors that influence the performance of our proposed DCAF network. The models are trained on the train set and evaluated on the validation set. The results are shown in Table 1.

We first investigate the influence of k (*i.e.* the number of stacked encoding layers). As shown in the first block of the table, our models outperform the baseline by a significant margin of 1.5%∼1.9%. This demonstrates the effectiveness of the proposed DCAF. The best result is obtained when $k = 2$. There is a slight performance drop when employing 3-layer DCAF. We then investigate the influence of the number of glimpses. It can be seen from the table that 4 glimpses attains the best performance. 8 glimpses incurs a massive increase in the number of parameters (53.5M→75.6M), making the network harder to optimize. The third block of the table shows the impacts of the dimension of attention features. Increasing the dimension from 1 to 4, produces a 0.2% performance drop. When we set the dimension to be 8, 16 and 32, the performance is 65.4%∼65.7% (not reported in Table 1 due to the lack of space). Considering the efficiency, we set the dimension to be 1 in our model. The last block of the table shows the effect of dense interactions. We remove dense

Component	Setting	Accuracy
# of stacked layers	None (baseline)	63.8
	1	65.3
	2	**65.7**
	3	65.6
# of glimpses	2	65.4
	4	**65.7**
	8	65.5
Dimension	1	**65.7**
	2	65.6
	4	65.5
Dense interactions	Full model	**65.7**
	w/o dense interactions	64.5

Table 1: Ablation studies of our proposed DCAF on the VQA 2.0 *val* set.

Model	MLB(2)	MUTAN(1)	DCN(3)	DCAF(1)
#Par.	33.8M	36.2M	31.9M	32.3M
Acc.	63.3	63.6	64.9	65.3
Model	MLB(4)	MUTAN(4)	DCN(6)	DCAF(2)
#Par.	51.4M	53.8M	51.3M	53.5M
Acc.	63.3	63.7	64.7	65.7

Table 2: Comparison of different interaction structures on the VQA 2.0 *val* set. The number in brackets indicates the number of stacked layers.

	SCAF-1	SCAF-2	SCAF-3	DCAF
Sceen Recognition	94.8	93.6	93.9	**95.0**
Sport Recognition	96.6	95.5	95.7	**96.7**
Color Attributes	**75.7**	70.0	71.4	74.9
Other Attributes	**61.4**	55.3	57.5	60.4
Activity Recognition	61.9	54.0	56.6	**62.8**
Positional Reasoning	41.8	31.9	**47.4**	44.4
Object Recognition	89.4	**90.4**	88.1	90.2
Absurd	92.7	**98.1**	**98.1**	95.0
Utility&Affordances	44.1	45.5	**48.3**	**48.3**
Object Presence	96.2	**97.1**	95.5	96.7
Counting	59.3	55.5	54.9	**62.9**
Sentiment	71.8	65.5	67.3	**72.7**
Overall Accuracy	86.9	87.1	86.8	**88.0**

Table 3: Comparison of different DCAF variants under different types of questions on the TDIUC *test* set.

Model	test-dev				test-std
	Yes/No	Number	Other	Overall	Overall
Bottom-up [Anderson *et al.*, 2018]	81.8	44.2	56.1	65.3	65.7
DCN [Nguyen and Okatani, 2018]	83.5	46.6	57.3	66.9	67.0
DA-NTN [Bai *et al.*, 2018]	84.3	47.1	57.9	67.6	67.9
Counter [Zhang *et al.*, 2018]	83.1	**51.6**	59.0	68.1	68.4
CoR [Wu *et al.*, 2018]	85.2	47.9	59.2	68.6	69.1
MFH+Bottom-Up [Yu *et al.*, 2018]	84.3	49.6	59.9	68.8	-
BAN+Glove [Kim *et al.*, 2018]	85.5	50.7	60.5	69.7	-
DCAF (ours)	**86.0**	50.5	**61.1**	**70.2**	**70.4**

Table 4: Comparison with previous state-of-the-art methods on the VQA 2.0 dataset.

Model	test-dev				test-std
	Yes/No	Number	Other	Overall	Overall
QGHC [Gao *et al.*, 2018]	83.5	38.1	57.1	65.9	65.9
VKMN [Su *et al.*, 2018]	83.7	37.9	57.0	66.0	66.1
MFH [Yu *et al.*, 2018]	85.0	39.7	57.4	66.8	66.9
DCN [Nguyen and Okatani, 2018]	84.6	42.4	57.3	66.9	67.0
DA-NTN [Bai *et al.*, 2018]	85.8	41.9	58.6	67.9	68.1
CoR [Wu *et al.*, 2018]	85.7	44.1	59.1	68.4	68.5
DCAF (ours)	**86.8**	**45.5**	**61.0**	**69.9**	**70.0**

Table 5: Comparison with previous state-of-the-art methods on the VQA 1.0 dataset.

Question Type	RAU [Kafle and Kanan, 2017]	CATL-QTA^W [Shi *et al.*, 2018]	CoR [Wu *et al.*, 2018]	DCAF (ours)
Sceen Recognition	94.0	93.8	94.7	**95.0**
Sport Recognition	93.5	95.6	95.9	**96.7**
Color Attributes	66.9	60.2	74.5	**74.9**
Other Attributes	56.5	54.4	60.0	**60.4**
Activity Recognition	51.6	60.1	62.2	**62.8**
Positional Reasoning	35.3	34.7	40.9	**44.4**
Object Recognition	86.1	87.0	88.8	**90.2**
Absurd	96.1	**100.0**	94.7	95.0
Utility and Affordances	31.6	31.5	37.4	**48.3**
Object Presence	94.4	94.6	95.8	**96.7**
Counting	48.4	53.3	58.8	**62.9**
Sentiment Understanding	60.1	64.4	67.2	**72.7**
Overall (Arithmetric MPT)	67.8	69.1	72.6	**75.0**
Overall (Harmonic MPT)	59.0	60.1	65.8	**69.9**
Overall Accuracy	84.3	85.0	86.9	**88.0**

Table 6: Comparison with previous state-of-the-art methods on the TDIUC dataset.

connections while retain the encoding layers in the DCAF module. Such modifications decrease the accuracy by 1.2%. This indicates that the efficacy of DCAF is due to not only the depth of the network (*i.e.*, stacked encoding layers) but also dense interactions. Furthermore, the dense interactions contribute more to the final performance compared with the stacked encoding layers (1.2% vs. 0.7% gain).

Some methods [Yang *et al.*, 2016; Nguyen and Okatani, 2018] perform sequential interactions by stacking several interaction layers/modules. We compare this interaction structure with our dense interactions in Table 2. We implement the stack structure of MLB and MUTAN. The stack structure is proposed by SAN [Yang *et al.*, 2016], which stacks multiple attention layers. DCN [Nguyen and Okatani, 2018] is implemented by stacking multiple symmetric co-attention layers. As shown in Table 2, with a similar number of parameters, the dense interaction structure models more interactions and achieves better performance than the sequential interaction structure. Furthermore, for the sequential interaction structure, there is no or a slight performance gain when stacking more layers (*e.g.* 0%, 0.1% and -0.2% gains for MLB, MUTAN and DCN, respectively). While our DCAF(2) can achieve a 0.4% improvement compared with DCAF(1). This phenomenon is in line with our intuition. Our structure contains dense connections, which are helpful to gradient flow and information flow. This enables us to exploit more potentiality of a large network to boost the performance.

In Table 3, we compare the performance of different DCAF variants under different question types. We develop three variants, and they have different sparse connections between

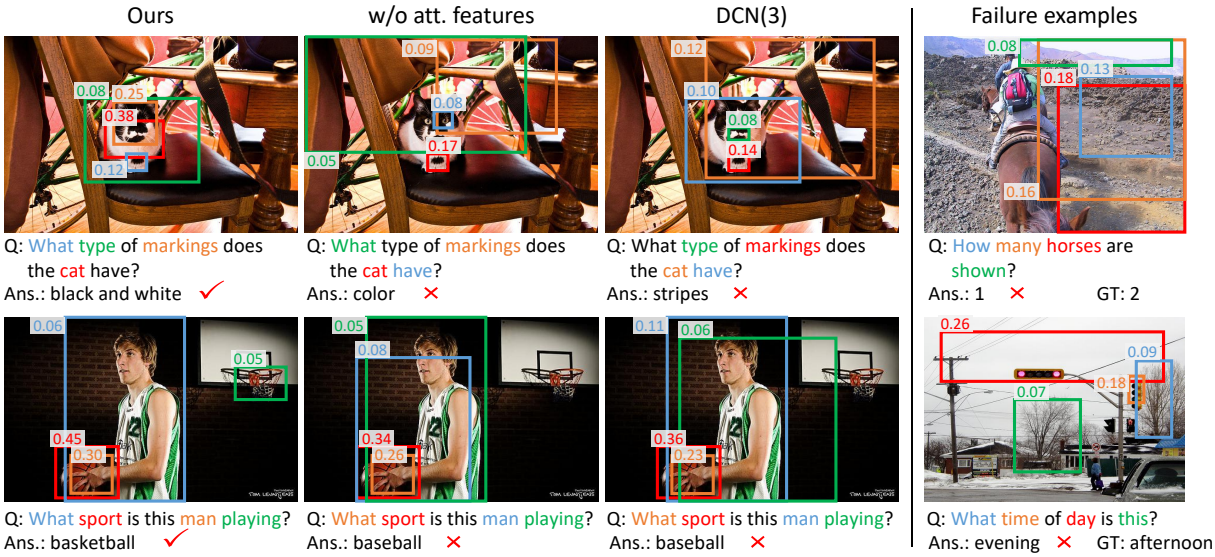

Figure 3: Visualization of attention weights. We highlight the top 4 boxes in each picture and top 4 words in each question using different colors. The box or word highlighted in red has the highest attention weight. The boxes or words highlighted in orange, blue, and green have descending attention weights. The weight value of each box is also shown in the picture. Best view in color.

the encoding layers. The first variant SCAF-1 only connects the encoding layers at the same hierarchical level (*i.e.* connecting q_0/v_0, q_1/v_1, and q_2/v_2). The second variant SCAF-2 has cross-hierarchical connections (*i.e.* connecting q_0/v_1, q_1/v_0, q_1/v_2, and q_2/v_1). The third variant SCAF-3 has different cross-hierarchical connections (*i.e.* connecting q_0/v_2, and q_2/v_0). From the table, we can find that each variant has its unique ability to answer a particular type of questions. For question type of *"Attributes"*, SCAF-1 outperforms other variants and DCAF. For question type of *"Positional Reasoning"*, SCAF-3 performs best. These variants show different strengths on different types of questions. In contrast, DCAF integrates the three types of connections above, forming dense connections. It inherits the strengths of the variants with sparse connections to a certain extent, and capture fine-grained multi-modal information, thus achieving the best performance in overall accuracy.

4.3 Comparison with State-of-the-arts

In this section, we compare our single DCAF model with the state-of-the-art models on three datasets. Firstly, Table 4 shows the results on the VQA 2.0 dataset. Remarkably, our approach outperforms other state-of-the-art methods in overall accuracy and all question types except for *"Number"*. Secondly, Table 5 shows the results on the VQA 1.0 dataset. Compared with the most recent state-of-the-art model CoR [Wu *et al.*, 2018], our single DCAF model achieves a new state-of-the-art result of 70.0% on test-std set. Thirdly, Table 6 shows the results on the TDIUC dataset. DCAF improves the overall accuracy of the state-of-the-art CoR from 86.9% to 88.0%. In particular, there is an improvement of 10.9% in *"Utility and Affordances"* and 5.5% in *"Sentiment Understanding"*. In summary, DCAF achieves consistently the best performance on all three datasets.

4.4 Qualitative Evaluation

In Figure 3, we visualize four object regions with the highest attention weights in each picture and four words with the highest attention weights in each question. The first column shows two visualization examples from our model. We can see that the model focuses on relevant object regions and words, thus generates the correct answers. To better understand the effect of attention features, we remove the attention features when visualization. As shown in the second column, the model pays attention to some background regions or neglects some relevant objects (*e.g.* basketball hoop), obtaining the wrong answers. This shows these attention features can capture some fine-grained information, and help to locate important regions accurately. In the third column, we show the visualization results from 3-layer DCN model as comparison. Some failure examples are shown in the last column. Both failure cases are due to wrong attention regions.

5 Conclusions

In this paper, we present a novel framework Densely Connected Attention Flow (DCAF) for VQA. The core of the framework is the DCAF module, which is designed to enable dense image-question interactions. The proposed Attention Connectors are used to densely connect image and question features at arbitrary layers in the DCAF module and model the interaction between image and question. The experimental results on the VQA 1.0, VQA 2.0 and TDIUC datasets confirm the effectiveness of the proposed framework, and DCAF outperforms state-of-the-art approaches on the three datasets. DCAF is applicable to other multimodal tasks.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (4192059) and National Natural Science Foundation of China (61872366 and 61472422).

References

[Anderson *et al.*, 2018] 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 *CVPR*, 2018.

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

[Bai *et al.*, 2018] Yalong Bai, Jianlong Fu, Tiejun Zhao, and Tao Mei. Deep attention neural tensor network for visual question answering. In *ECCV*, 2018.

[Ben-Younes *et al.*, 2017] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *International Conference on Computer Vision (ICCV)*, 2017.

[Chen *et al.*, 2015] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv:1511.05960*, 2015.

[Fukui *et al.*, 2016] 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 *EMNLP*, 2016.

[Gao *et al.*, 2018] 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 *European Conference on Computer Vision (ECCV)*, 2018.

[Goyal *et al.*, 2017] 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 *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Kafle and Kanan, 2017] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *International Conference on Computer Vision (ICCV)*, 2017.

[Kim *et al.*, 2016] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *NeurIPS*, 2016.

[Kim *et al.*, 2017] 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 (ICLR)*, 2017.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[Nguyen and Okatani, 2018] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*, 2018.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[Reddi *et al.*, 2018] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

[Shi *et al.*, 2018] Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. Question type guided attention in visual question answering. In *European Conference on Computer Vision (ECCV)*, 2018.

[Su *et al.*, 2018] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. Learning visual knowledge memory networks for visual question answering. In *CVPR*, 2018.

[Wu *et al.*, 2018] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. Chain of reasoning for visual question answering. In *NeurIPS*, 2018.

[Xu and Saenko, 2016] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision (ECCV)*, 2016.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[Yu *et al.*, 2017] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *International Conference on Computer Vision*, 2017.

[Yu *et al.*, 2018] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[Zhang *et al.*, 2018] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *ICLR*, 2018.