This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Dual Attention Networks for Multimodal Reasoning and Matching

Hyeonseob Nam
Search Solutions Inc.
hyeonseob.nam@navercorp.com

Jung-Woo Ha
NAVER Corp.
jungwoo.ha@navercorp.com

Jeonghee Kim
NAVER LABS Corp.
jeonghee.kim@naverlabs.com

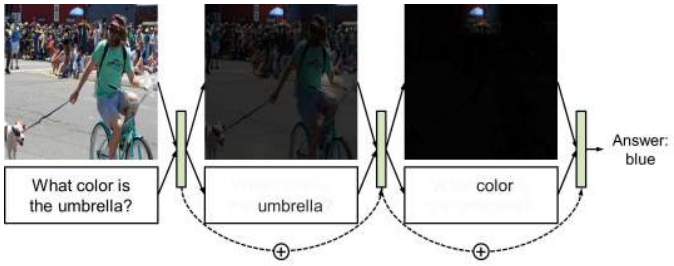
Abstract

We propose Dual Attention Networks (DANs) which jointly leverage visual and textual attention mechanisms to capture fine-grained interplay between vision and language. DANs attend to specific regions in images and words in text through multiple steps and gather essential information from both modalities. Based on this framework, we introduce two types of DANs for multimodal reasoning and matching, respectively. The reasoning model allows visual and textual attentions to steer each other during collaborative inference, which is useful for tasks such as Visual Question Answering (VQA). In addition, the matching model exploits the two attention mechanisms to estimate the similarity between images and sentences by focusing on their shared semantics. Our extensive experiments validate the effectiveness of DANs in combining vision and language, achieving the state-of-the-art performance on public benchmarks for VQA and image-text matching.

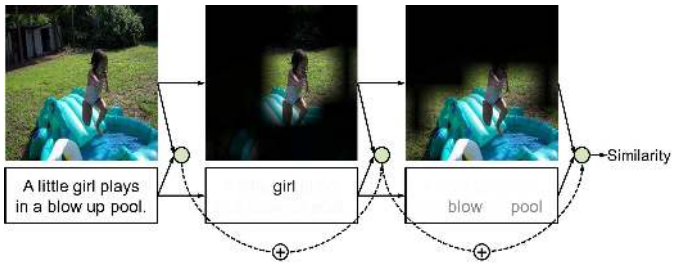
1. Introduction

Vision and language are two central parts of human intelligence to understand the real world. They are also fundamental components in achieving artificial intelligence, and a tremendous amount of research has been done for decades in each area. Recently, dramatic advances in deep learning have broken the boundaries between vision and language, drawing growing interest in their intersection, such as visual question answering (VQA) [3, 37, 23, 35], image captioning [33, 2], image-text matching [8, 11, 20, 30], visual grounding [24, 9], etc.

One of the recent advances in neural networks is the attention mechanism [21, 4, 33]. It aims to focus on certain aspects of data sequentially and aggregate essential information over time to infer the results, and has been successfully applied to both areas of vision and language. In computer vision, attention based methods adaptively select a sequence of image regions to extract necessary features [21, 6, 33]. Similarly, attention models for natural language processing highlight specific words or sentences



(a) DAN for multimodal reasoning. (r-DAN)



(b) DAN for multimodal matching. (m-DAN)

Figure 1: Overview of Dual Attention Networks (DANs) for multimodal reasoning and matching. The brightness of image regions and darkness of words indicate their attention weights predicted by DANs.

to distill information from input text [4, 25, 15]. These approaches have improved the performance of wide applications in conjunction with deep architectures including convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

Despite the effectiveness of attention in handling both visual and textual data, it has been hardly attempted to establish a connection between visual and textual attention models which can be highly beneficial in various scenarios. For example, the VQA problem in Figure 1a with the question `What color is the umbrella?` can be efficiently solved by simultaneously focusing on the region of `umbrella` and the word `color`. In the example of image-text matching in Figure 1b, the similarity between the image and sentence can be effectively measured by at-

tending to the specific regions and words sharing common semantics such as `girl` and `pool`.

In this paper, we propose *Dual Attention Networks* (DANs) which jointly learn visual and textual attention models to explore the fine-grained interaction between vision and language. We investigate two variants of DANs illustrated in Figure 1, referred to as *reasoning-DAN* (r-DAN) and *matching-DAN* (m-DAN), respectively. The r-DAN collaboratively performs visual and textual attentions using a joint memory which assembles the previous attention results and guides the next attentions. It is suited to the tasks requiring multimodal reasoning such as VQA. On the other hand, the m-DAN separates visual and textual attention models with distinct memories but jointly trains them to capture the shared semantics between images and sentences. This approach eventually finds a joint embedding space which facilitates efficient cross-modal matching and retrieval. Both proposed algorithms closely connect visual and textual attention mechanisms into a unified framework, achieving outstanding performance in VQA and image-text matching problems.

To summarize, the main contributions of our work are as follows:

- We propose an integrated framework of visual and textual attentions, where critical regions and words are jointly located through multiple steps.
- Two variants of the proposed framework are implemented for multimodal reasoning and matching, and applied to VQA and image-text matching.
- Detailed visualization of the attention results validates that our models effectively focus on vital portions of visual and textual data for the given task.
- Our framework demonstrates the state-of-the-art performance on the VQA dataset [3] and the Flickr30K image-text matching dataset [36].

2. Related Work

2.1. Attention Mechanisms

Attention mechanisms allow models to focus on necessary parts of visual or textual inputs at each step of a task. Visual attention models selectively pay attention to small regions in an image to extract core features as well as reduce the amount of information to process. A number of methods have recently adopted visual attention to benefit image classification [21, 28], image generation [6], image captioning [33], visual question answering [35, 26, 32], etc. On the other hand, textual attention mechanisms generally aim to find semantic or syntactic input-output alignments under an encoder-decoder framework, which is especially effective in handling long-term dependency. This approach has been successfully applied to various tasks including machine translation [4], text generation [16], sentence summarization [25], and question answering [15, 32].

2.2. Visual Question Answering (VQA)

VQA is a task of answering a question in natural language regarding a given image, which requires multimodal reasoning over visual and textual data. It has received a surge of interest since Antol *et al*. [3] presented a large-scale dataset with free-form and open-ended questions. A simple baseline by Zhou *et al*. [37] predicts the answer from a concatenation of CNN image features and bag-of-word question features. Several methods adaptively construct a deep architecture depending on the given question. For example, Noh *et al*. [23] impose a dynamic parameter layer on a CNN which is learned by the question, while Andreas *et al*. [1] utilize a compositional structure of the question to assemble a collection of neural modules.

One limitation of the above approaches is that they resort to a global image representation which contains noisy or unnecessary information. To address this problem, Yang *et al*. [35] propose stacked attention networks which perform multi-step visual attention, and Shih *et al*. [26] use object proposals to identify regions relevant to the given question. Recently, dynamic memory networks [32] integrate an attention mechanism with a memory module, and multimodal compact bilinear pooling [5] is exploited to expressively combine multimodal features and predict attention over the image. These methods commonly employ visual attention to find critical regions, but textual attention has been rarely incorporated into VQA. Although HieCoAtt [18] applies both visual and textual attentions, it independently performs each step of co-attention without reasoning over previous co-attention outputs. On the contrary, our method moves and refines both attentions via multiple reasoning steps based on the memory of previous attentions, which facilitates close interplay between visual and textual data.

2.3. Image-Text Matching

The core issue with image-text matching is measuring the semantic similarity between visual and textual inputs. It is commonly addressed by learning a joint space where image and sentence feature vectors are directly comparable. Hodosh *et al*. [8] apply canonical correlation analysis (CCA) to find embeddings that maximize the correlation between images and sentences, which is further improved by incorporating deep neural networks [14, 34]. A recent approach by Wang *et al*. [30] includes structure-preserving constraints within a bidirectional loss function to make the joint space more discriminative. In contrast, Ma *et al*. [19] construct a CNN to combine an image and sentence fragments into a joint representation, from which the matching

score is directly inferred. Image captioning frameworks are also exploited to estimate the similarity based on the inverse probability of sentences given a query image [20, 29].

To the best of our knowledge, no study has attempted to learn multimodal attention models for image-text matching. Even though Karpathy *et al*. [11, 10] propose to find the alignments between image regions and sentence fragments, they explicitly compute all pairwise distances between them and estimate the average or best alignment score, which leads to inefficiency. On the other hand, our method automatically attends to the shared concepts between images and sentences while embedding them into a joint space, where cross-modal similarity is directly obtained by a single inner product operation.

3. Dual Attention Networks (DANs)

We present two structures of DANs to consolidate visual and textual attention mechanisms: r-DAN for multimodal reasoning and m-DAN for multimodal matching. They share a common framework but differ in their ways of associating visual and textual attentions. We first describe the common framework including input representation (Section 3.1) and attention mechanisms (Section 3.2). Then we illustrate the details of r-DAN (Section 3.3) and m-DAN (Section 3.4) applied to VQA and image-text matching, respectively.

3.1. Input Representation

Image representation The image features are extracted from 19-layer VGGNet [27] or 152-layer ResNet [7]. We first rescale images to 448×448 and feed them into the CNNs. In order to obtain feature vectors for different regions, we take the last pooling layer of VGGNet (pool5) or the layer beneath the last pooling layer of ResNet (res5c). Finally the input image is represented by $\{\mathbf{v}_1, \cdots, \mathbf{v}_N\}$, where N is the number of image regions and $\mathbf{v}_n$ is a 512 (VGGNet) or 2048 (ResNet) dimensional feature vector corresponding to the n-th region.

Text representation We employ bidirectional LSTMs to generate text features as depicted in Figure 2. Given one-hot encoding of T input words $\{\mathbf{w}_1, \cdots, \mathbf{w}_T\}$, we first embed the words into a vector space by $\mathbf{x}_t = \mathbf{M}\mathbf{w}_t$, where $\mathbf{M}$ is an embedding matrix. Then we feed the vectors into the bidirectional LSTMs:

$$\mathbf{h}_t^{(f)} = \mathbf{LSTM}^{(f)}(\mathbf{x}_t, \mathbf{h}_{t-1}^{(f)}), \quad (1)$$

$$\mathbf{h}_t^{(b)} = \mathbf{LSTM}^{(b)}(\mathbf{x}_t, \mathbf{h}_{t+1}^{(b)}), \quad (2)$$

where $\mathbf{h}_t^{(f)}$ and $\mathbf{h}_t^{(b)}$ represent the hidden states at time t from the forward and backward LSTMs, respectively. By adding the two hidden states at each time step, *i.e.* $\mathbf{u}_t = \mathbf{h}_t^{(f)} + \mathbf{h}_t^{(b)}$, we construct a set of feature vectors $\{\mathbf{u}_1, \cdots, \mathbf{u}_T\}$ where $\mathbf{u}_t$ encodes the semantics of the t-th word in the context of the entire sentence. Note that the models discussed here including the word embedding matrix and the LSTMs are trained end-to-end.

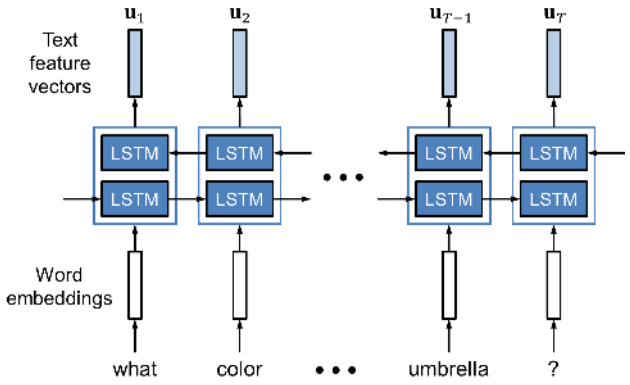


Figure 2: Bidirectional LSTMs for text encoding.

3.2. Attention Mechanisms

Our method performs visual and textual attentions simultaneously through multiple steps and gathers necessary information from both modalities. In this section, we explain the underlying attention mechanisms employed at each step, which serve as building blocks to compose the entire DANs. For simplicity, we shall omit the bias term $\mathbf{b}$ in the following equations.

Visual Attention. Visual attention aims to generate a context vector by attending to certain parts of the input image. At step k, the visual context vector $\mathbf{v}^{(k)}$ is given by

$$\mathbf{v}^{(k)} = \mathbf{V\_Att}(\{\mathbf{v}_n\}_{n=1}^N, \mathbf{m}_{\mathbf{v}}^{(k-1)}), \quad (3)$$

where $\mathbf{m}_{\mathbf{v}}^{(k-1)}$ is a memory vector encoding the information that has been attended until step $k-1$. Specifically, we employ the soft attention mechanism where the context vector is obtained from a weighted average of input feature vectors. The attention weights $\{\alpha_{\mathbf{v},n}^{(k)}\}_{n=1}^N$ are computed by a 2-layer feed-forward neural network (FNN) and the softmax function:

$$\mathbf{h}_{\mathbf{v},n}^{(k)} = \tanh\left(\mathbf{W}_{\mathbf{v}}^{(k)}\, \mathbf{v}_n\right) \odot \tanh\left(\mathbf{W}_{\mathbf{v},\mathbf{m}}^{(k)}\, \mathbf{m}_{\mathbf{v}}^{(k-1)}\right), \quad (4)$$

$$\alpha_{\mathbf{v},n}^{(k)} = \mathrm{softmax}\left(\mathbf{W}_{\mathbf{v},\mathbf{h}}^{(k)}\, \mathbf{h}_{\mathbf{v},n}^{(k)}\right), \quad (5)$$

$$\mathbf{v}^{(k)} = \tanh\left(\mathbf{P}^{(k)} \sum_{n=1}^{N} \alpha_{\mathbf{v},n}^{(k)}\, \mathbf{v}_n\right), \quad (6)$$

where $\mathbf{W}_{\mathbf{v}}^{(k)}$, $\mathbf{W}_{\mathbf{v},\mathbf{m}}^{(k)}$, and $\mathbf{W}_{\mathbf{v},\mathbf{h}}^{(k)}$ are the network parameters, $\mathbf{h}_{\mathbf{v},n}^{(k)}$ is a hidden state, and $\odot$ is element-wise multiplication. In Equation 6, we introduce an additional layer with

the weight matrix $\mathbf{P}^{(k)}$ in order to embed visual context vectors into a compatible space with textual context vectors, as we use pretrained image features $\mathbf{v}_n$.

Textual Attention. Textual attention computes a textual context vector $\mathbf{u}^{(k)}$ by focusing on specific words in the input sentence every step:

$$\mathbf{u}^{(k)} = \mathbf{T\_Att}(\{\mathbf{u}_t\}_{t=1}^{T}, \mathbf{m}_{\mathbf{u}}^{(k-1)}), \tag{7}$$

where $\mathbf{m}_{\mathbf{u}}^{(k-1)}$ is a memory vector. The textual attention mechanism is almost identical to the visual attention mechanism. In other words, the attention weights $\{\alpha_{\mathbf{u},t}^{(k)}\}_{t=1}^{T}$ are obtained from a 2-layer FNN and the context vector $\mathbf{u}^{(k)}$ is calculated by weighted averaging:

$$\mathbf{h}_{\mathbf{u},t}^{(k)} = \tanh\left(\mathbf{W}_{\mathbf{u}}^{(k)}\,\mathbf{u}_t\right) \odot \tanh\left(\mathbf{W}_{\mathbf{u},\mathbf{m}}^{(k)}\,\mathbf{m}_{\mathbf{u}}^{(k-1)}\right), \tag{8}$$

$$\alpha_{\mathbf{u},t}^{(k)} = \mathrm{softmax}\left(\mathbf{W}_{\mathbf{u},\mathbf{h}}^{(k)}\,\mathbf{h}_{\mathbf{u},t}^{(k)}\right), \tag{9}$$

$$\mathbf{u}^{(k)} = \sum_t \alpha_{\mathbf{u},t}^{(k)}\,\mathbf{u}_t. \tag{10}$$

where $\mathbf{W}_{\mathbf{u}}^{(k)}$, $\mathbf{W}_{\mathbf{u},\mathbf{m}}^{(k)}$, and $\mathbf{W}_{\mathbf{u},\mathbf{h}}^{(k)}$ are the network parameters, $\mathbf{h}_{\mathbf{u},t}^{(k)}$ is a hidden state. Unlike the visual attention, it does not need an additional layer after the last weighted averaging because the text features $\mathbf{u}_t$ are already trained end-to-end.

3.3. r-DAN for Visual Question Answering

VQA is a representative problem which requires joint reasoning over multimodal data. For this purpose, the r-DAN maintains a joint memory vector $\mathbf{m}^{(k)}$ which accumulates the visual and textual information that has been attended until step k. It is recursively updated by

$$\mathbf{m}^{(k)} = \mathbf{m}^{(k-1)} + \mathbf{v}^{(k)} \odot \mathbf{u}^{(k)}, \tag{11}$$

where $\mathbf{v}^{(k)}$ and $\mathbf{u}^{(k)}$ are the visual and textual context vectors obtained from Equation 6 and 10, respectively. This joint representation concurrently guides the visual and textual attentions, *i.e.* $\mathbf{m}^{(k)} = \mathbf{m}_{\mathbf{v}}^{(k)} = \mathbf{m}_{\mathbf{u}}^{(k)}$, which allows the two attention mechanisms to closely cooperate with each other. The initial memory vector $\mathbf{m}^{(0)}$ is defined based on global context vectors $\mathbf{v}^{(0)}$ and $\mathbf{u}^{(0)}$ as

$$\mathbf{m}^{(0)} = \mathbf{v}^{(0)} \odot \mathbf{u}^{(0)}, \tag{12}$$

$$\text{where } \mathbf{v}^{(0)} = \tanh\left(\mathbf{P}^{(0)}\frac{1}{N}\sum_n \mathbf{v}_n\right), \tag{13}$$

$$\mathbf{u}^{(0)} = \frac{1}{T}\sum_t \mathbf{u}_t. \tag{14}$$

By repeating the dual attention (Equation 3 and 7) and memory update (Equation 11) for K steps, we effectively focus on the key portions in the image and question, and gather relevant information for answering the question. Figure 3 illustrates the overall architecture of r-DAN in case of $K = 2$.

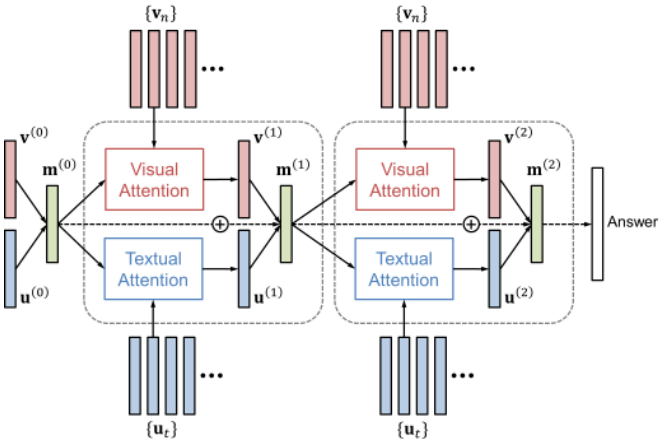


Figure 3: r-DAN in case of $K = 2$.

The final answer is predicted by multi-way classification to the top C frequent answers. We employ a single-layer softmax classifier with cross-entropy loss where the input is the final memory $\mathbf{m}^{(K)}$:

$$\mathbf{p}_{ans} = \mathrm{softmax}\left(\mathbf{W}_{ans}\,\mathbf{m}^{(K)}\right), \tag{15}$$

where $\mathbf{p}_{ans}$ represents the probability over the candidate answers.

3.4. m-DAN for Image-Text Matching

Image-text matching tasks usually involve comparison between numerous images and sentences, where effective and efficient computation of cross-modal similarities is crucial. To achieve this, we aim to learn a joint embedding space which satisfies the following properties. First, the embedding space encodes the shared concepts that frequently co-occur in image and sentence domains. Moreover, images and sentences are autonomously embedded into the joint space without being paired, so that arbitrary image and sentence vectors in the space are directly comparable.

Our m-DAN jointly learns visual and textual attention models to capture the shared concepts between the two modalities, but separates them at inference time to provide generally comparable representations in the embedding space. Contrary to the r-DAN which uses a joint memory, the m-DAN maintains separate memory vectors for visual and textual attentions as follows:

$$\mathbf{m}_{\mathbf{v}}^{(k)} = \mathbf{m}_{\mathbf{v}}^{(k-1)} + \mathbf{v}^{(k)}, \tag{16}$$

$$\mathbf{m}_{\mathbf{u}}^{(k)} = \mathbf{m}_{\mathbf{u}}^{(k-1)} + \mathbf{u}^{(k)}, \tag{17}$$

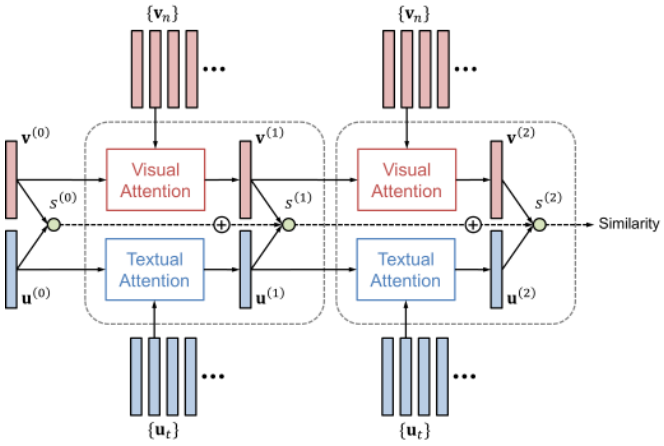


Figure 4: m-DAN in case of $K = 2$.

which are initialized to $\mathbf{v}^{(0)}$ and $\mathbf{u}^{(0)}$ defined in Equation 13 and 14, respectively. At each step, we compute the similarity $s^{(k)}$ between visual and textual context vectors by their inner product:

$$s^{(k)} = \mathbf{v}^{(k)} \cdot \mathbf{u}^{(k)}. \tag{18}$$

After performing K steps of the dual attention and memory update, the final similarity S between the given image and sentence becomes

$$S = \sum_{k=0}^{K} s^{(k)}. \tag{19}$$

The overall architecture of this model when $K = 2$ is depicted in Figure 4.

This network is trained with bidirectional max-margin ranking loss, which is widely adopted for multimodal similarity learning [11, 10, 13, 30]. For each correct pair of an image and a sentence $(\mathbf{v}, \mathbf{u})$, we additionally sample a negative image $\mathbf{v}^-$ and a negative sentence $\mathbf{u}^-$ to construct two negative pairs $(\mathbf{v}^-, \mathbf{u})$ and $(\mathbf{v}, \mathbf{u}^-)$. Then, the loss function becomes:

$$\mathcal{L} = \sum_{(\mathbf{v},\mathbf{u})} \Big\{ \max\left[0, m - S(\mathbf{v},\mathbf{u}) + S(\mathbf{v}^-,\mathbf{u})\right] + \max\left[0, m - S(\mathbf{v},\mathbf{u}) + S(\mathbf{v},\mathbf{u}^-)\right] \Big\}, \tag{20}$$

where m is a margin constraint. By minimizing this function, the network is trained to focus on the common semantics that only appears in correct image-sentence pairs through visual and textual attention mechanisms.

At inference time, an arbitrary image or sentence is embedded into the joint space by concatenating its context vectors:

$$\mathbf{z}_\mathbf{v} = [\mathbf{v}^{(0)}; \cdots; \mathbf{v}^{(K)}], \tag{21}$$

$$\mathbf{z}_\mathbf{u} = [\mathbf{u}^{(0)}; \cdots; \mathbf{u}^{(K)}], \tag{22}$$

where $\mathbf{z}_\mathbf{v}$ and $\mathbf{z}_\mathbf{u}$ are the representations for image $\mathbf{v}$ and sentence $\mathbf{u}$, respectively. Note that these vectors are obtained via separate pipelines of visual and textual attentions, *i.e.* learned shared concepts are revealed from an image or sentence itself, not from an image-sentence pair. The similarity between two vectors in the joint space is simply computed by their inner product, *e.g.* $S(\mathbf{v}, \mathbf{u}) = \mathbf{z}_\mathbf{v} \cdot \mathbf{z}_\mathbf{u}$, which is equivalent to the output of the network in Equation 19.

4. Experiments

4.1. Experimental Setup

We fix all the hyper-parameters applied to both r-DAN and m-DAN. The number of attention steps K is set to 2 which empirically shows the best performance. The dimension of every hidden layer—including word embedding, LSTMs, and attention models—is set to 512. We train our networks by stochastic gradient descent with a learning rate 0.1, momentum 0.9, weight decay 0.0005, dropout ratio 0.5, and gradient clipping at 0.1. The network is trained for 60 epochs, where the learning rate is dropped to 0.01 after 30 epochs. A minibatch for r-DAN and m-DAN consists of 128 pairs of ⟨image, question⟩ and 128 quadruplets of ⟨positive image, positive sentence, negative image, negative sentence⟩, respectively. The number of possible answers C for VQA is set to 2000, and the margin m for the loss function in Equation 20 is set to 100.

4.2. Evaluation on Visual Question Answering

4.2.1 Dataset and Evaluation Metric

We evaluate the r-DAN on the Visual Question Answering (VQA) dataset [3], which contains approximately 200K real images from MSCOCO dataset [17]. Each image is associated with three questions, and each question is labeled with ten answers by human annotators. The dataset is typically divided into four splits: *train* (80K images), *val* (40K images), *test-dev* (20K images), and *test-std* (20K images). We train our model using *train* and *val*, validate with *test-dev*, and evaluate on *test-std*. There are two forms of tasks, open-ended and multiple-choice, which require to answer each question without and with a set of candidate answers, respectively. For both tasks, we follow the evaluation metric used in [3] as

$$\mathrm{Acc}(\hat{a}) = \min\left\{ \frac{\#\text{humans that labeled } \hat{a}}{3}, 1 \right\}, \tag{23}$$

where $\hat{a}$ is a predicted answer.

4.2.2 Results and Analysis

The performance of r-DAN compared with state-of-the-art VQA systems is presented in Table 1, where our method

Table 1: Results on the VQA dataset compared with state-of-the-art methods.

	Test-dev					Test-standard				
	Open-Ended				MC	Open-Ended				MC
Method	Y/N	Num	Other	All	All	Y/N	Num	Other	All	All
iBOWIMG [37]	76.5	35.0	42.6	55.7	61.7	76.8	35.0	42.6	55.9	62.0
DPPnet [23]	80.7	37.2	41.7	57.2	62.5	80.3	36.9	42.2	57.4	62.7
VQA team [3]	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
SAN [35]	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
NMN [1]	81.2	38.0	44.0	58.6	-	-	-	-	58.7	-
ACK [31]	81.0	38.4	45.2	59.2	-	81.1	37.1	45.8	59.4	-
DMN+ [32]	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
MRN (ResNet) [12]	82.3	38.8	49.3	61.7	66.2	82.4	**38.2**	49.4	61.8	66.3
HieCoAtt (ResNet) [18]	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
RAU (ResNet) [22]	81.9	39.0	53.0	63.3	67.7	81.7	**38.2**	52.8	63.2	67.3
MCB (ResNet) [5]	82.2	37.7	**54.8**	64.2	68.6	-	-	-	-	-
DAN (VGG)	82.1	38.2	50.2	62.0	67.0	-	-	-	-	-
DAN (ResNet)	**83.0**	**39.1**	53.9	**64.3**	**69.1**	**82.8**	38.1	**54.0**	**64.2**	**69.0**

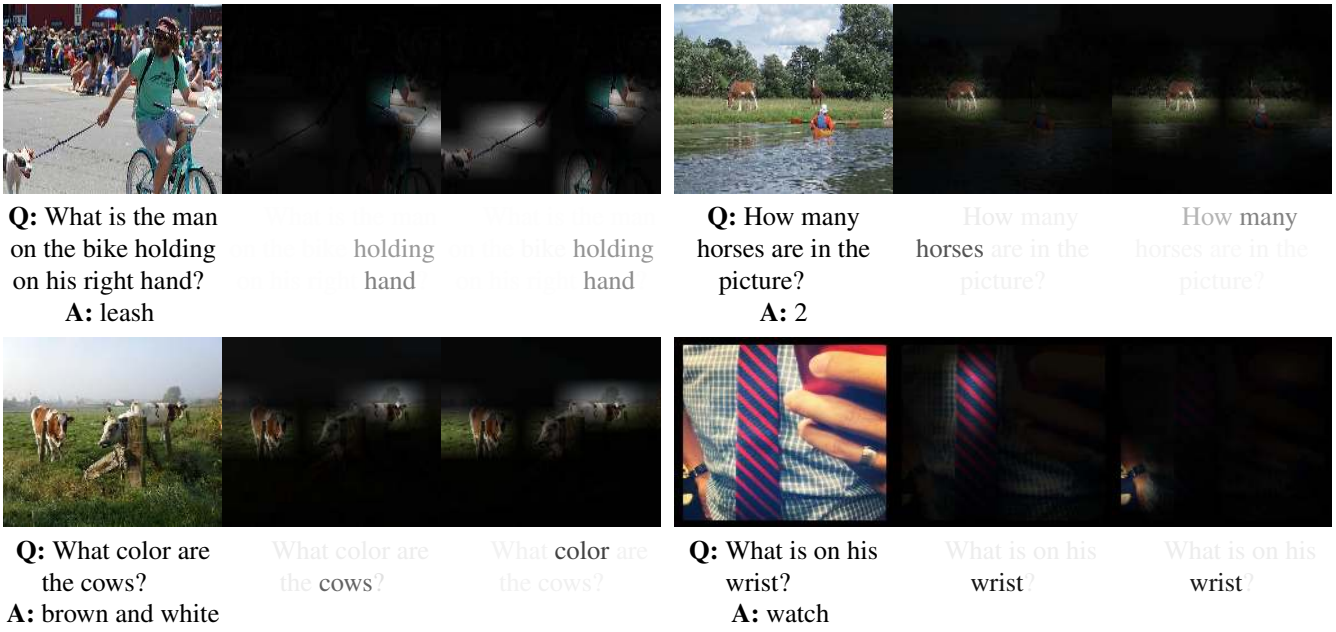


Figure 5: Qualitative results on the VQA dataset with attention visualization. For each example, the query image, question, and the answer by DAN are presented from top to bottom; the original image (question), the first and second attention maps are shown from left to right. The brightness of images and darkness of words represent their attention weights.

achieves the best performance in both open-ended and multiple-choice tasks. For fair evaluation, we compare single-model accuracies obtained without data augmentation, even though [5] reported better performance using model ensembles and additional training data. Figure 5 describes the qualitative results from our approach with visualization of the attention weights. Our method produces correct answers to challenging problems which require fine-grained reasoning, as well as successfully attends to the specific regions and words which facilitate answering the questions. Specifically, the first and fourth examples in Figure 5 illustrate that the r-DAN moves its visual attention to the proper regions indicated by the attended words, while the second and third examples show that it moves its textual attention to divide a complex task into sequential subtasks—finding target objects and extracting certain attributes.

Table 2: Bidirectional retrieval results on the Flickr30K dataset compared with state-of-the-art methods.

	Image-to-Text				Text-to-Image			
Method	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
DCCA [34]	27.9	56.9	68.2	4	26.8	52.9	66.9	4
mCNN [19]	33.6	64.1	74.9	3	26.2	56.3	69.6	4
m-RNN-VGG [20]	35.4	63.8	73.7	3	22.8	50.7	63.1	5
GMM+HGLMM FV [14]	35.0	62.0	73.8	3	25.0	52.7	66.0	5
HGLMM FV [24]	36.5	62.2	73.3	-	24.7	53.4	66.8	-
SPE [30]	40.3	68.9	79.9	-	29.7	60.1	72.1	-
DAN (VGG)	41.4	73.5	82.5	2	31.8	61.7	72.5	3
DAN (ResNet)	**55.0**	**81.8**	**89.0**	**1**	**39.4**	**69.2**	**79.1**	**2**

Figure 6: Qualitative results from image-to-text retrieval with attention visualization. For each example, the query image and the top two retrieved sentences are shown from top to bottom; the original image (sentence), the first and second attention maps are shown from left to right. (+) and (-) indicate ground-truth and non ground-truth sentences, respectively.

4.3. Evaluation on Image-Text Matching

4.3.1 Dataset and Evaluation Metric

We employ the Flickr30K dataset [36] to evaluate the m-DAN for multimodal matching. It consists of 31,783 real images with five descriptive sentences for each, and we follow the public splits by [20]: 29,783 training, 1,000 validation and 1,000 test images. We report the performance of m-DAN in bidirectional image and sentence retrieval using the same metrics as previous work [34, 19, 20, 30]. Recall@K (K=1, 5, 10) represents the percentage of the queries where at least one ground-truth is retrieved among the top K results and MR measures the median rank of the top-ranked ground-truth.

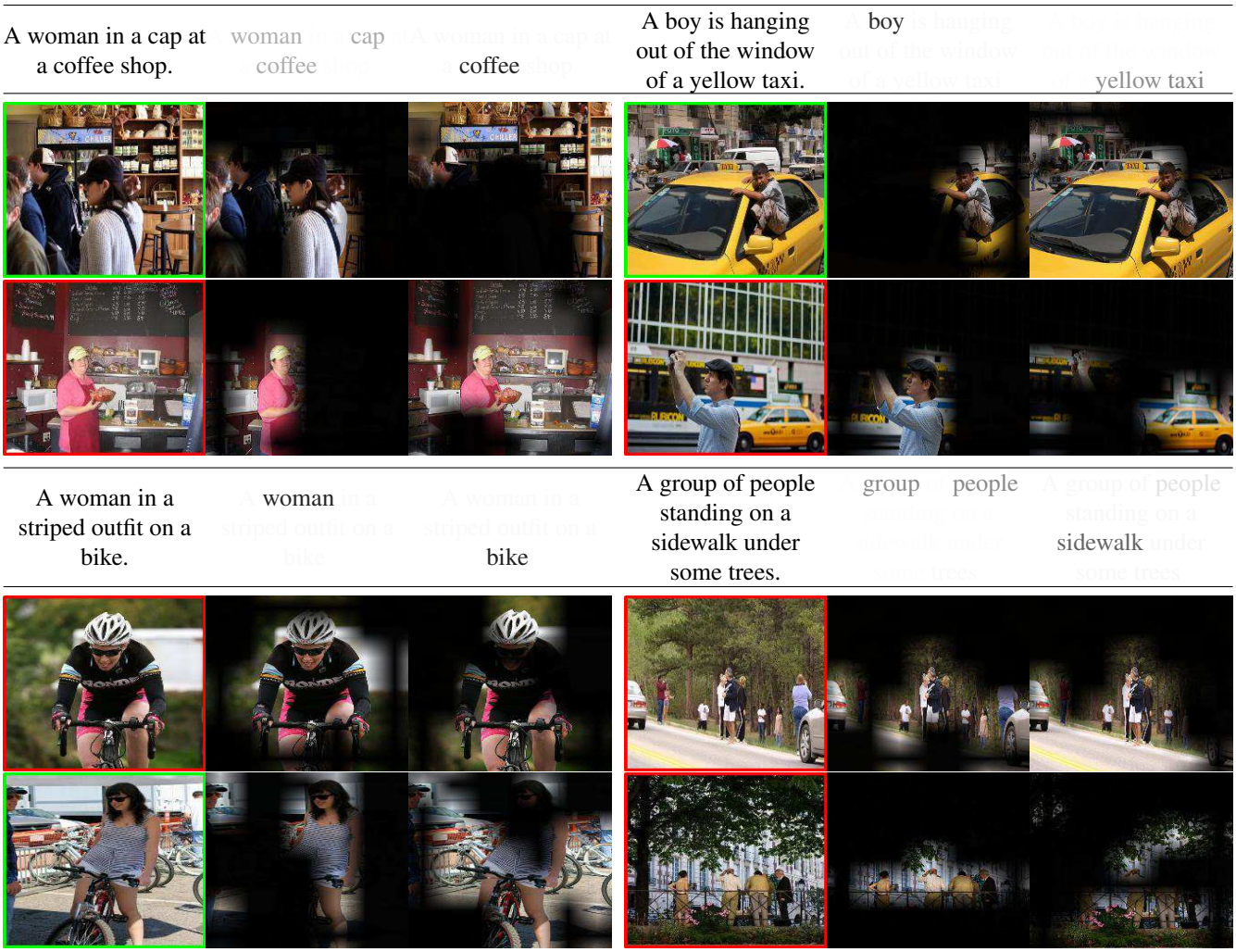


Figure 7: Qualitative results from text-to-image retrieval with attention visualization. For each example, the query sentence and the top two retrieved images are shown from top to bottom; the original sentence (image), the first and second attention maps are shown from left to right. Green and red boxes indicate ground-truth and non ground-truth images, respectively.

4.3.2 Results and Analysis

Table 2 presents the quantitative results on the Flickr30K dataset, where the proposed method outperforms other recent approaches in all measures. The qualitative results from image-to-text and text-to-image retrieval are also illustrated in Figure 6 and Figure 7, respectively, with visualization of attention outputs. At each step of attention, the m-DAN effectively discovers the essential semantics appearing in both modalities. It tends to capture the main subjects (*e.g.* `woman`, `boy`, `people`, etc.) at the first step, and figure out relevant objects, backgrounds or actions (*e.g.* `computer`, `scaffolding`, `sweeps`, etc.) at the second step. Note that this property solely comes from the training stage where visual and textual attention models are jointly learned, while images and sentences are processed independently at inference time.

5. Conclusion

We propose Dual Attention Networks (DANs) to bridge visual and textual attention mechanisms. We present two architectures of DANs for multimodal reasoning and matching. The first model infers the answers collaboratively from images and sentences, while the other one embeds them into a common space by capturing their shared semantics. These models demonstrate the state-of-the-art performance in VQA and image-text matching, showing their effectiveness in extracting essential information via the dual attention mechanism. The proposed framework can be potentially generalized to various tasks at the intersection of vision and language, such as image captioning, visual grounding, video question answering, etc.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, 2016. 2, 6

[2] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 1

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *CVPR*, 2015. 1, 2, 5, 6

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 1, 2

[5] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2, 6

[6] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *ICML*, 2015. 1, 2

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[8] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899, 2013. 1, 2

[9] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, June 2016. 1

[10] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 3, 5

[11] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 1, 3, 5

[12] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. *arXiv preprint arXiv:1606.01455*, 2016. 6

[13] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. 5

[14] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015. 2, 7

[15] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, 2016. 1, 2

[16] J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *ACL*, 2015. 2

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[18] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*, 2016. 2, 6

[19] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *CVPR*, 2015. 2, 7

[20] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015. 1, 3, 7

[21] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014. 1, 2

[22] H. Noh and B. Han. Training recurrent answering units with joint loss minimization for vqa. *arXiv preprint arXiv:1606.03647*, 2016. 6

[23] H. Noh, P. Hongsuck Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 2016. 1, 2, 6

[24] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 1, 7

[25] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, 2015. 1, 2

[26] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016. 2

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014. 3

[28] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections. In *NIPS*, 2014. 2

[29] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 3

[30] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 1, 2, 5, 7

[31] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016. 6

[32] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016. 2, 6

[33] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2

[34] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015. 2, 7

[35] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 1, 2, 6

[36] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 2, 7

[37] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 1, 2, 6