Contents lists available at ScienceDirect

High-Confidence Computing

journal homepage: www.sciencedirect.com/journal/high-confidence-computing

Research article

Coreference resolution helps visual dialogs to focus

Tianwei Yue[1], Wenping Wang [*,1], Chen Liang, Dachi Chen, Congrui Hetang, Xuewei Wang

Carnegie Mellon University, Pittsburgh 15213, USA

ARTICLE INFO

Article history:
Received 30 August 2023
Revised 21 September 2023
Accepted 2 October 2023
Available online 22 November 2023

Keywords:
Multi-model machine learning
Visual dialog
Co-reference resolution

ABSTRACT

Visual Dialog is a multi-modal task involving both computer vision and dialog systems. The goal is to answer multiple questions in conversation style, given an image as the context. Neural networks with attention modules are widely used for this task, because of their effectiveness in reasoning the relevance between the texts and images. In this work, we study how to further improve the quality of such reasoning, which is an open challenge. Our baseline is the Recursive Visual Attention (RVA) model, which refines the vision-text attention by iteratively visiting the dialog history. Building on top of that, we propose to improve the attention mechanism with contrastive learning. We train a Matching-Aware Attention Kernel (MAAK) by aligning the deep feature embeddings of an image and its caption, to provide better attention scores. Experiments show consistent improvements from MAAK. In addition, we study the effect of using Multimodal Compact Bilinear (MCB) pooling as a three-way feature fusion for the visual, textual and dialog history embeddings. We analyze the performance of both methods in the discussion section, and propose further ideas to resolve current limitations.

1. Introduction

Visual Dialog was originally proposed by [1]. The goal of this task is to answer a sequence of questions grounded in an image, in the style of a conversation [2]. It can be seen as an extended form of visual question answering (VQA) [3,4], requiring multiple rounds of question-answering. Visual Dialog is a multi-modal task, involving challenges in both computer vision and natural language processing, and is being actively researched since its proposal.

The application value of visual dialogue systems is immense. Solving the visual dialogue task enables people to build interactive communicative AI agents that could further solve tasks like object retrieval, transport and more [2]. They may help visually impaired people understand their surroundings, enable self-driving cars to behave more like human, make it easy to extract information from large-scale security camera data [1,5,6], to name a few.

A state-of-the-art model for Visual Dialog is the Recursive Visual Attention (RVA) model [7,8]. The RVA model is a neural network that heavily utilizes the attention mechanism. The highlighted contribution of RVA is its unique method for co-reference resolution. In short, the questions in natural conversations commonly involve pronouns like "it, this, them", which are ambiguous without the context and hurt question-answering accuracy. RVA addresses this by tracing back the dialog history until such ambiguity has been fully resolved, and iteratively refining the visual attention (the weight assignment of image features aiming to up-weight the most relevant ones) alongside the process.

In this work, we adopt the overall architecture of RVA, and make modifications to its components to study the effects. We also experiment with our novel modules and demonstrate improvements.

First, we observe that in the original RVA model, the attention module (ATT) that calculates the visual attention scores uses the vanilla softmax attention design, and has headroom in its capacity. We propose a Matching-Aware Attention Kernel (MAAK), which is a superior cross-modality similarity measurement that better evaluates the similarity of visual and textual features. Experiments demonstrate that the MAAK architecture itself already improves performance. Further, we propose a joint-trained contrastive learning objective that aligns the embeddings of images and their captions, which also improves the cross-modal alignment and further boosts overall performance.

Second, the original RVA model uses simple concatenation to fuse multi-modal features (image, question, history) before the decoder. To study the effects of different feature fusion strategies, we experimented with Multimodal Bilinear Compact Pooling (MBC) as an alternative, together with other options. Their performance are evaluated and discussed.

2. Related work

In contrast to classical vision-and-language tasks such as image captioning and VQA that only involve a single round of

[*] Corresponding author.
E-mail address: wpw436@gmail.com (W. Wang).

[1] The authors contributed equally to this work.

https://doi.org/10.1016/j.hcc.2023.100184

natural language interaction, visual dialog tasks require not only visual grounding of linguistic expressions but also capturing semantic information from several rounds of human conversation context. To solve this task, we need to tackle classical VQA challenges (Section 2.2) as well as visual co-reference resolution (Section 2.5) with the conversation history. Improving attention on images (Section 2.4) and multi-modal feature fusion (Section 2.3) are effective ways improve visual dialog performance.

2.1. Visual dialog

A branch of attention-based approaches for visual dialog were primarily proposed to address these challenges, including memory networks [1,9], a generative-discriminative model for history-conditioned image attentive encoder [10,11], sequential co-attention using adversarial learning [12], synergistic networks [13, 14], *etc.*

Visual question answering is fundamentally compositional in nature. Simple questions can be composed to form more complex ones. So another branch of research towards solving visual dialog problems develops module networks to solve a sequence of subtasks in visual dialog [15–18]. In addition, instead of supervised ways towards building the dialogue systems, it is also reasonable to utilize deep reinforcement learning method for visual dialog tasks [19,20].

2.2. Visual question answering tasks

Going beyond the current trend of improving both visual and language components of VQA models, [21,22] took a further step to support VQA tasks with known facts. Their main idea is to construct a unified knowledge base which links visual concepts to corresponding textual concepts in the knowledge base.

In contrast, [23] proposed to augment VQA tasks with visual question generation (VQG) tasks. They introduced VQG as an auxiliary task towards improving VQA performance.

2.3. Feature fusion

Visual dialog requires utilizing both the information from visual and language features. Current approaches in VQA or visual grounding rely on vector concatenations or element-wise sum or product of vectors. In contrast, [24] proposed to rely on Multimodal Compact Bilinear pooling (MCB) to get a joint representation. The benefits of bilinear pooling is that it computes the outer product between two vectors, which allows a multiplicative interaction between all elements of both vectors. However, the outer product will result in vectors of much higher dimension compared to vector sum or product. Therefore, the author applied Count Sketch projection [25] to project the outer product to a lower-dimension space which is applicable as the input of subsequent neural networks.

2.4. Attentions on images

Answering questions in conversations requires both attention on images and attention on the dialog history. [26] proposed to focus on informative image regions for VQA. Concretely, this method first learns to embed the textual question and the set of visual image regions into a latent space where the inner product yields a relevance weighting for each region, then it obtains a weighted average of concatenated vision and language features to feed into a two-layer network that outputs confidence scores for candidate answers.

To extract salient image regions to attend to, top-down visual attention mechanisms have been widely used in image captioning and VQA systems. They train the neural network to figure out the important image features from a convolutional neural network (CNN). However, they do not address the fact that CNN extracts feature map from an image uniformly, while only some salient image regions matter in solving problems. [27] proposed a bottom-up mechanism based on faster R-CNN to produce salient image regions for attention. Faster R-CNN generates regions-of-interest (RoI) around the foreground objects, and each RoI contributes a feature embedding through RoI pooling. This ensures the image features to be informative for QA.

2.5. Visual co-reference resolution

Currently, there are generally two branches of methods for co-reference resolution. One is attention-based. [28] used attention memory to store all previous image attentions at a sentence level. Different from this soft attention mechanism over all the memorized attention maps, [7] proposed recursive prediction of discrete attention over topic-related history. [29] proposed a Dual Attention Networks (DAN) to frame a visual dialog task as an encoder–decoder architecture: the encoder, consisting of `REFER` and `FIND` modules, jointly embeds the image, questions and dialog history. Specifically, `REFER` learns the latent relationships between Q and a dialog history by multi-head attention (reference-aware representations). `FIND` takes image features and reference-aware representations as input to perform visual grounding. The decoder that converts the embedded representation into the ranked list of candidate answers. The other branch for visual co-reference resolution is based on neural module networks [15,30–32].

3. Baseline methods

In this section, we first formulate the visual dialog task in Section 3.1, then describe three multimodal baselines models, with the highlighted method, recursive visual attention, described in Section 3.5.

3.1. Problem statement

A visual dialog task consists of t rounds of question-answering, while the input of each round can be defined as a given image V, a sequence of follow-up questions as Q_t, and a dialog history $\mathcal{H}_t$ until round $t-1$,

$$\mathcal{H}_t = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1^{gt})}_{H_1}, \ldots, \underbrace{(Q_{t-1}, A_{t-1}^{gt})}_{H_{t-1}})$$

A_t^{gt} denotes the ground truth answer at round t. While $\mathcal{H}_t$ denotes dialog history up to round t, H_t denotes the single (Q_t, A_t^{gt}) pair of round t. C denotes the image caption.

In each round of dialog, given $\{V, Q_t, \mathcal{H}_t\}$, the model should return a textual answer A_t (generative decoder) or a list of ranked answer choices (discriminative decoder) $\mathcal{A} = \{A_t^1, \ldots, A_t^n\}$.

3.2. Baseline description

We consider three baseline models in our experiments: Late-Fusion Generative model (LF_GEN), Late-Fusion Discriminative model (LF_DISC), and Recursive RVA. All of these models follow an encoder–decoder structure. The encoder converts the input $\{V, Q_t, \mathcal{H}_t\}$ into a fused vector representation, and the decoder converts the fused vector representation into the final output. The decoder can be either generative or discriminative, which are described in Section 3.4 with more details. The RVA model is described in Section 3.5.

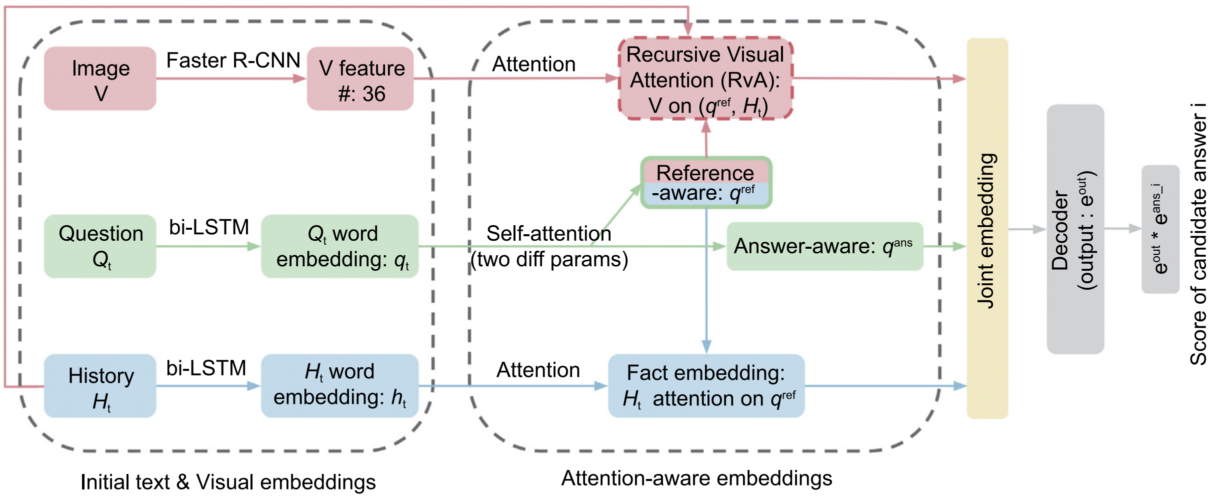


Image V
Faster R-CNN
V feature #: 36
Attention
Recursive Visual Attention (RvA): V on (q^{ref}, H_t)
Reference -aware: q^{ref}
Question Q_t
bi-LSTM
Q_t word embedding: q_t
Self-attention (two diff params)
Answer-aware: q^{ans}
History H_t
bi-LSTM
H_t word embedding: h_t
Attention
Fact embedding: H_t attention on q^{ref}
Joint embedding
Decoder (output : e^{out})
e^{out} * e^{ans_i}
Score of candidate answer i
Initial text & Visual embeddings
Attention-aware embeddings

Fig. 1. Baseline model description.

3.3. Feature processing

As a first step to solving the multi-modal visual dialog problem, inputs of different modalities need to be encoded into feature vectors. To encode the image, we use faster R-CNN [33] to extract the salient regions for attention [27]. The attention scores usually depend on their relevance to the current task. Based on the scores, we take a weighted average of the per-region features to summarize the image, a common practice as shown by [5,26,34,35]. To encode the texts, which include the history $\mathcal{H}_t$ and question Q_t, we applied two LSTMs [4] with separate parameters.

3.4. Late-fusion encoder models

The Late-Fusion encoder is a trivial baseline method for fusing the input modalities. It simply concatenates the three types of features $(V, Q, \mathcal{H}_t)$ to obtain the joint representation. Here, the embedding of each input feature before the joint embedding layer is a vector in $\mathbb{R}^{512}$.

After this simple encoder, there are two decoder options. The generative decoder use the encoded representation as the initial state of an LSTM language model. The training objective is to maximize the log-likelihood of the corresponding ground truth answer text (trained end-to-end). The discriminative decoder computes dot product similarity between the encoded representation and an LSTM encoding of each answer option, then feed that into a softmax to compute the posterior probability over options. The training objective is to maximize the log-likelihood of the correct option.

3.5. Recursive visual attention (RVA)

The RVA model refines the attention weights on visual features through a recursive attention mechanism. An illustration of RVA model is shown in Fig. 1. After obtaining the initial text and visual embeddings, the model further uses history attentions, image attentions and self-attention of the question to produce reference-aware and answer-aware question embeddings. The highlighted module is RVA for recursively calculating image attentions by backtracking the dialog history.

Specifically, given a set of the Faster R-CNN features, an ATT module will calculate the attention weights on the visual features based the current question. However, because the questions involve co-reference, the current question alone may not be enough to decide which visual feature to focus on. Therefore, given the current question, an INFER module calculates a boolean and a scalar λ. The boolean decides whether the model needs to keep browsing backwards, whereas the λ decides the importance for the attention weights calculated from the current question. If the boolean is true, the model is confident that the question can be answered unambiguously with the information gathered so far, and will return the visual attention as of now. If it is false, the visual attention of the current question will be accounted with weight λ, and the model will backtrack to one of the previous questions to get better-informed visual attention. A PAIR module decides from the dialog history which exact previous question to go to, given the current question. This recursive process stops either when the agent fully understands Q_t, or until it reaches the beginning of the dialog. The pseudocode for RVA is given in Algorithm 1.

Algorithm 1 Recursive Visual Attention (RVA)

```
function RVA($V, Q_t, H_t, t$)
    $bool$, $\lambda$ = UNDERSTAND($Q_t$,t)
    if $bool = True$ then
        return ATT($V, H_t, t$)
    else
        $t_p$ = PAIR($Q_t, H_t, t$)
        return (1-$\lambda$) $\cdot$ RVA($V, Q_t, H_t, t_p$)
                  + $\lambda \cdot$ ATT($Q_t, H_t, t$)
    end if
end function
```

4. Proposed approach

To improve upon the original RVA model, we experimented two modifications. The first is a Matching-Aware Attention Kernel (MAAK) which is jointly trained with RVA (Section 4.1). The second is a tri-modal feature fusion module using compact bilinear pooling (Section 4.2).

4.1. Joint-training matching-aware attention kernel

Given a set of Faster R-CNN RoI features and a question, our model has to decide which visual feature to focus on to best

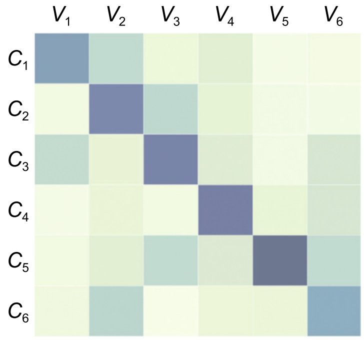


Fig. 2. Matching aware attention Kernel.

answer the question. For example, the RVA model formulates this attention calculation as following:

$$\begin{aligned} \boldsymbol{z}_{t,i}^{A} &= \mathrm{L2Norm}\left(f_q^A\left(\boldsymbol{q}_t\right) \odot f_v^A\left(\boldsymbol{v}_i\right)\right) \\ \boldsymbol{\alpha}_t^A &= \mathrm{softmax}\left(W^A Z_t^A\right) \end{aligned} \tag{1}$$

where $\boldsymbol{q}_t$ is the question embedding for the tth question, $\boldsymbol{v}_i$ is the Faster-RCNN embedding for the ith region proposal, and f_q^A and f_v^A are linear transformations to project visual and textual feature into the same embedding space. Further, in the original RVA model, given the tth question of m words $w_{t1}, w_{t2}, \ldots, w_{tm}$, the question embedding q_t is calculated as a weighted average of word embeddings w_{ti}:

$$q_t = \Sigma_i^m w_{ti} * \alpha_i \tag{2}$$

where α_i is calculated by passing e_{ti} through an LSTM, projecting the output to dimension of 1, and finally normalizing across i.

However, because visual feature space and textual feature space are significantly different, we cannot expect two linear layers to be able to find an ideal joint embedding space. Moreover, there are no explicit training signal for attention scores (signals only come from the final answer predictions). To deal with these challenges, we propose to jointly-train [36–39] a matching-aware attention kernel, which will measure the similarity of textual and visual features. In other words, we will use another joint training objective to directly supervise the training of f_q^A and f_v^A, which are the nonlinear transformations to map visual and textual features into a single joint embedding space. We propose to leverage the naturally existing image-caption pairs already included in our dataset. The overall architecture of our proposed kernel is visualized in Fig. 2.

To measure the similarity of a caption with m words $c_1, c_2, \ldots, c_m$ and an image with 36 proposal features $v_1, v_2, \ldots, v_{36}$, it is important to note that not every word in the caption will match every proposal feature. Thus we want to summarize caption embedding and image embedding to give more weights to the matching parts, so that our joint training objective will not affect the embeddings of other parts. This attention weighting is done in two steps.

First, on both the caption and image sides, we apply uni-modal pre-weighting. On the caption side, the caption embedding is summarized by attention weights calculated by an LSTM, which is used to weighted-average the word embeddings. (This part is the same as the original RVA model). On the image side, we propose to also use an LSTM to give a weight to each proposal:

$$v_i = v_i * \alpha_i \tag{3}$$

where α_i is calculated by passing v_i through an LSTM, projecting the output to dimension of 1, and finally normalizing across i. This weighted visual feature v_i is fed into the attention module instead, which is shown to improve model performance in our later sections.

Second, we apply matching-aware multi-modal weighting. At a high level, the matching-aware multi-modal weighting will give a weight β_{ci} to each word the caption w_i based on a query visual embedding v_{query} summarized from visual features alone. Then, final caption embedding is summarized by weighting with β_{ci} instead. Similarly, final image embedding is summarized by weighted averaging with β_{vi} calculated based on query caption embedding c_{query}. So in addition to the f_q^A and f_v^A, we propose another set of transformations f_q^Q and f_v^Q to obtain the query embeddings. The formula for calculating the final caption embedding e^t are as follows:

$$v_{query} = \Sigma_i^{36} f_v^Q(v_i) * \alpha_i \tag{4}$$

$$\beta_{ci} = softmax_i(f_q^A(w_i) \cdot v_{query}) \tag{5}$$

$$e^t = \Sigma_i^m f_q^A(w_i) * \beta_{ci} \tag{6}$$

The formulas for calculating the final image embedding e^t are as follows:

$$c_{query} = \Sigma_i^m f_q^Q(c_i) * \alpha_i \tag{7}$$

$$\beta_{vi} = softmax_i(f_v^A(v_i) \cdot c_{query}) \tag{8}$$

$$e^v = \Sigma_i^{36} f_v^A(v_i) * \beta_{vi} \tag{9}$$

Note that the calculation is symmetric for caption and image embeddings. Then, we calculate the matching score between $\boldsymbol{e}^v$ and $\boldsymbol{e}^t$ with dot product:

$$score = e^v \cdot e^t \tag{10}$$

The pre-training process of the proposed MAAK module is to minimize the distance between an image with its corresponding caption, while maximizing the distance from other captions. At each iteration, we sample a batch of n, then calculate the triplet loss as Fig. 3:

$$loss = \sum_{i,j} max(0, (S_{ij} + 1) - S_{ii}) +, max(0, (S_{ij} + 1) - S_{jj}) \tag{11}$$

In the above loss function, the attention matrix is denoted as S and C_i is corresponding caption of image V_i. The intuition of this loss function is that S_{ii} should be the max of the row/column.

We also calculate the batch bidirectional retrieval accuracy (BRacc) of image-caption matching as

$$BRacc = (hori\_acc + vert\_acc)/2 \tag{12}$$

We expect to observe increase in BRacc compared with MAAK without pre-training.

4.2. Bilinear compact joint embeddings

The proposed model uses inputs from three sources, *i.e.* image $\mathcal{V}$, question $\boldsymbol{q}_t$ and dialog history $\mathcal{H}_t$, thus we need to fuse these multi-modal features (shown as yellow in Fig. 1). Common approaches in VQA or visual grounding rely on concatenations, element-wise sum or product of vectors. In our study, we propose to consider more fine-grained pairwise interactions. The intuition comes from Multi-modal Compact Bilinear pooling (MCB) [24]. Specifically, we perform outer product among features of three modalities, *i.e.*, $(\boldsymbol{e}^{\mathcal{V}} \otimes \boldsymbol{e}^{\boldsymbol{q}_t}) \otimes \boldsymbol{e}^{\mathcal{H}_t}$. A caveat with outer product is its very high-dimensional 3D tensor output. We address this by approximation with Count Sketches and convolution theorem [40]

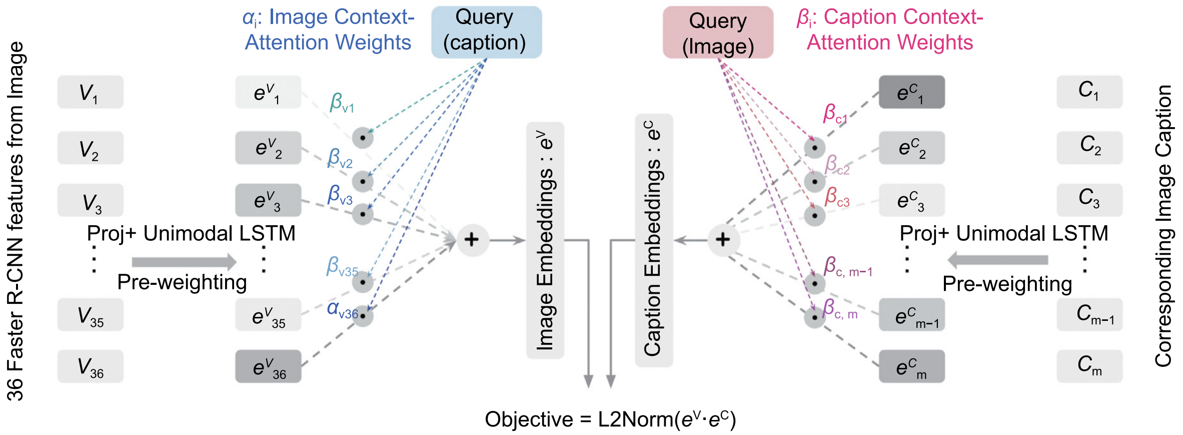


Fig. 3. Matching-aware attention kernel batch pre-training.

to make sure the fused output still fits as input for subsequent neural nets.

Given the encoded embeddings of image, question, history as $\boldsymbol{e}^{\mathcal{V}}, \boldsymbol{e}^{\boldsymbol{q}_t}, \boldsymbol{e}^{\mathcal{H}_t} \in \mathbb{R}^{512}$, the joint representation is calculated as

$$\boldsymbol{e}^{m} = \mathrm{FFT}^{-1}(\mathrm{FFT}(\Psi_d(\boldsymbol{e}^{\mathcal{V}})) \odot \mathrm{FFT}(\Phi_d(\boldsymbol{e}^{\boldsymbol{q}_t}))),$$

$$\boldsymbol{e}^{joint} = \mathrm{FFT}^{-1}(\mathrm{FFT}(\Psi_d(\boldsymbol{e}^{m})) \odot \mathrm{FFT}(\Psi_d(\boldsymbol{e}^{\mathcal{H}_t}))),$$

where FFT denotes the Fast Fourier Transform, Ψ_d denotes the Count Sketch projection [25] that projects a vector $v \in \mathbb{R}^n$ to $y \in \mathbb{R}^d$. Specifically, to calculate $\Psi_d(v)$, we first initialize y as $\mathbf{0} \in \mathbb{R}^d$ and two randomly uniformly sampled vectors $s \in \{-1, 1\}^n$ and $h \in \{1, \dots, d\}^n$, then add $v[i] \cdot s[i]$ to $y[j]$ where $j = h[i]$ is looked up using h for all dimension of v.

The direct outer product among three vectors of dimension n is of $O(n^3)$ complexity. After using the trick of Count Sketches and convolution theorem, it takes $O(n)$ for Count Sketches projection, $O(d)$ for element-wise product in the frequency domain, and $O(d \log d)$ for FFT, totaling $O(n + d \log d)$ operations.

We experiment two settings for the bilinear compact joint compact. The first one concatenate history and question vector together to build a dialog vector. Then this dialog vector is fed into bilinear pooling together with the visual features. The output from the pooling is the fused feature. The input dimension of history and question vector are 1024 and 300 respectively, with dialog vector being 1324-dim. The input of visual vector is 2048, from the pre-extracted feature of Faster-rcnn. The output of the pooling is 3372-dim.

The second setting directly takes all three inputs for bilinear pooling. All history, question and visual features are hashed by count-min sketch method, and converted into frequency domain by FFT. During the tri-modal outer product step, three modalities are multiplied directly. The result of this dot product is computed with reverse FFT and becomes the final result of bilinear pooling. The input and output dimensions are the same as the first setting.

5. Experiment setup

5.1. Dataset and input modalities

We use the VisDial dataset [1] for this project. The VisDial dataset consists of 120k images extracted from Common Objects in Context (COCO) dataset. For each image, there is one dialog which consists of 10 rounds of question–answers, and each question is provided with 100 candidate answers. The input of the model is an image, a dialog history and a question. The output of the model is a ranking of the answer candidates. An example from the VisDial dataset is shown Fig. 4.

Table 1
Summary of hyper parameters.

Model	Training
encoder: RVA	batch_size:32
decoder: Discriminative	num_epochs: 15
img_feature_size: 2048	initial_lr: 0.01
word_embedding_size: 300	lr_gamma: 0.1
lstm_hidden_size: 512	lr_milestones: 5
lstm_num_layers: 2	warmup_factor: 0.2
dropout_fc: 0.3	warmup_epochs: 1
dropout: 0.5	–
relu: ReLU	–

Images for the training dataset are from COCO train2014 and val2014. It contains 123,287 images, and 10 question–answer pairs per image. The validation set contains 2,064 images × 10 QA rounds. Due to the size of the training dataset, we randomly sampled 1/10 of the training set, and evaluate on the whole validation set.

The modalities involved in this task are images and natural language. The visual modality are the input images and the textual modality includes the image caption and the dialog. The images and their captions are used to train the proposed Matching-Aware Attention Kernel. The images and the dialogues are used for training and evaluating the visual dialog task.

5.2. Experimental methodology

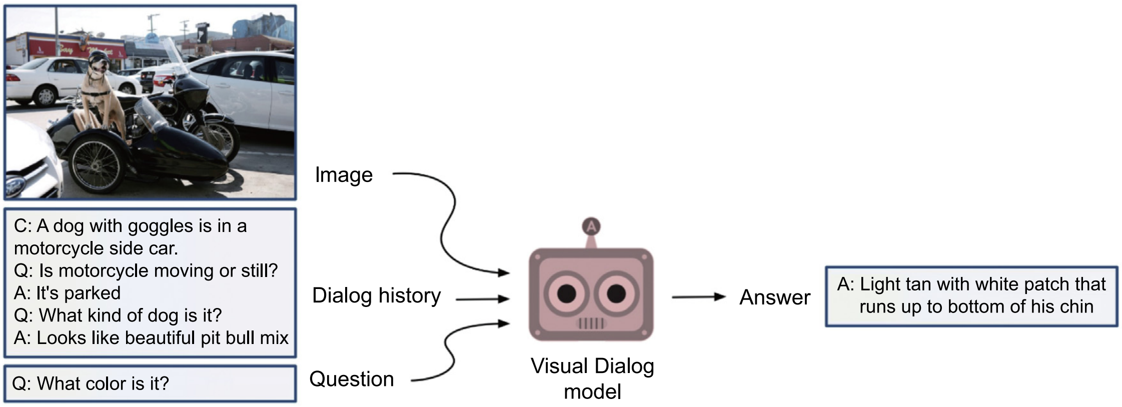
The hyper-parameters for our model are shown in Table 1:

In the above table, `lr_milestones` refers to the number of epochs when lr is updated to`lr` * `lr_gamma`. The warmup_factor refers for the first `warmup_epoch` epochs, the learning rate for training is `warmup_factor` * `initial_lr`, which is 0.002 for the first epoch in our case.

The evaluation metrics for our model include: (1) NDCG (Normalized Discounted Cumulative Gain): penalizes the lower rank of answers with high relevance; (2) MRR (Mean Reciprocal Rank); (3) Recall for the correct answer in the top 1, 5 and 10 answers; (4) Mean rank of the correct answer.

6. Results and discussion

In this section, we show the results of the baseline models and our proposed approaches. In addition, we analyzed why or why do not the proposed approaches improve the model performance.


C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix

Q: What color is it?

A: Light tan with white patch that runs up to bottom of his chin

Fig. 4. Visual dialog data example in VisDial.

Table 2
Performance of baseline models.

Models	LF_GEN	LF_DISC	**RVA**
R@1	0.2063	0.3688	**0.4164**
R@5	0.4219	0.6562	**0.7233**
R@10	0.4464	0.7685	**0.8276**
MEAN	33.26	8.0922	**6.2830**
MRR	0.3221	0.5051	**0.5583**
NDCG	0.2688	0.4582	**0.5170**

6.1. Baseline model comparison

Table 2 compares the current result of Late-Fusion encoder with generative decoder, Late-Fusion encoder with discriminative decoder and Recursive Visual Attention model. From these qualitative results, RVA model performs much better than the other two models. Therefore, we build our improvements upon the RVA model.

6.2. Proposed model comparison

Table 3 summarizes the comparisons of our proposed models as well as the best performing baseline model **RVA**.

6.2.1. Joint-training MAAK

From **RVA** to **RVA+MAAK**, we observe some performance improvement. Since **RVA+MAAK** if differs from **RVA** in the LSTM pre-weighting of visual features before feeding into attention module, we conclude that the uni-modal pre-weighting was helpful. From **RVA+MAAK** to **RVA+Joint_MAAK**, we see further performance improvement, which suggests that our joint learning objective was effective. In addition, we see that from **RVA+MAAK** to **RVA+Joint_MAAK**, the *batch acc* jumps from 0.2163 to 0.8683. We conclude that in the original RVA model, dot product was not very effective in capturing the matching of visual and textual features, whereas our joint training objective improves this similarity measure. We also confirmed the correlation of better BRacc with better model performance overall.

6.2.2. Bilinear compact joint embeddings

RVA+MBC is comparable with the baseline **RVA**, however, the performance of **RVA+modal_MBC** decreases slightly compared to **RVA**.

The reasons behind this could be: *i*) MBC layer is applied too late: the MBC is applied at the last layer before feeding into the decoder. At that time, each feature is already mixed up with the information/attention from other features, then the pair-wise interactions between features of different modalities would be less expressive; *ii*) too-high-dimension features: the final embedding size of image, question, history is 2048, 300, 1024, respectively. In practice, however, the scenarios when MBC help are usually when the feature size is small. Further ideas on addressing these limitations are presented in Section 7.

7. Conclusion and future directions

So far, we have tried to improve on our baseline model RVA from two aspects. The first one is using Matching-Aware Attention Kernel (Section 4.1) and a contrastive learning task to improve visual attention. The second one is using Multi-modal Compact Bilinear pooling (Section 4.2) as the multi-modal feature fusion module. According to the experiment results, Matching-Aware Attention Kernel outperforms the baseline model on all the metrics. Multi-modal Compact Bilinear pooling did not show positive results.

Given that the MAAK method showed promising results, further improvements are possible. For example, we can design

Table 3
Performance of all proposed ideas.

Ideas	RVA	RVA+MAAK	RVA+Joint_MAAK	RVA+MBC	RVA+Tri-modal_MBC
R@1	0.4164	0.4214	**0.4218**	0.4194	0.4061
R@5	0.7233	0.7284	**0.7306**	0.7166	0.7074
R@10	0.8276	0.8314	**0.8345**	0.8262	0.8123
MEAN	6.2830	6.1063	**6.0151**	6.3148	6.7182
MRR	0.5583	0.5624	**0.5632**	0.5574	0.5456
NDCG	0.5170	0.5120	**0.5189**	0.5172	0.5068
BRacc	-	0.2163	**0.8683**	-	-

RVA+MAAK is plugging our Matching-Aware Attention Kernel into RVA but without joint training objective. **RVA+Joint_MAAK** is plugging in MAAK and using the joint training objective. **RVA+MBC** is plugging in our two-dim compact bilinear pooling. **RVA+Tri-modal_MBC** is plugging in our tri-modal bilinear pooling. The BRacc measure is the batch bidirectional retrieval accuracy calculated in Eq. (12), which we only calculated for MAAK related models.

the loss function to be aware of the similarities between captions/images. Concretely, the loss function Eq. (11) can be further refined as the following,

$$sim\_aware\_loss = max\left(0, \left(S_{ij} + \sqrt{(V_i \cdot V_j)(C_i \cdot C_j)}\right) - S_{ii}\right) + max\left(0, \left(S_{ij} + \sqrt{(V_i \cdot V_j)(C_i \cdot C_j)}\right) - S_{jj}\right) \quad (13)$$

The intuition behind this similarity-aware contrastive loss for MAAK is that, though only V_i and C_i are correlated, the correlation of V_i and C_j should be slightly higher (rather than being zero) if the two images (V_i and V_j) are similar or the two captions (C_i and C_j) are similar. We expect this definition of loss to further improve the proposed MAAK method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J.M. Moura, D. Parikh, D. Batra, Visual dialog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 326–335.

[2] S. Kottur, J.M. Moura, D. Parikh, D. Batra, M. Rohrbach, CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog, 2019, arXiv preprint arXiv:1903.03166.

[3] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't just assume; look and answer: Overcoming priors for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4971–4980.

[4] Z. Zhou, T. Yue, C. Liang, X. Bai, D. Chen, C. Hetang, W. Wang, Unlocking everyday wisdom: Enhancing machine comprehension with script knowledge integration, Appl. Sci. 13 (16) (2023).

[5] G. Song, B. Leng, Y. Liu, C. Hetang, S. Cai, Region-based quality estimation network for large-scale person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018, http://dx.doi.org/10.1609/aaai.v32i1.12305, URL https://ojs.aaai.org/index.php/AAAI/article/view/12305.

[6] Y. He, J. Qian, J. Wang, Depth-wise decomposition for accelerating separable convolutions in efficient convolutional neural networks, arXiv preprint arXiv:1910.09455 (2019).

[7] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, J.-R. Wen, Recursive visual attention in visual dialog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6679–6688.

[8] T. Chen, S. Liu, Z. Chen, W. Hu, D. Chen, Y. Wang, Q. Lyu, C.X. Le, W. Wang, Faster, Stronger, and More Interpretable: Massive Transformer Architectures for Vision-Language Tasks, 2023.

[9] W. Wang, Y. Guo, C. Shen, S. Ding, G. Liao, H. Fu, P.K. Prabhakar, Integrity and junkiness failure handling for embedding-based retrieval: a case study in social network search, arXiv preprint arXiv:2304.09287 (2023).

[10] J. Lu, A. Kannan, J. Yang, D. Parikh, D. Batra, Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model, in: Advances in Neural Information Processing Systems, 2017, pp. 314–324.

[11] K. Yu, Y. Wang, S. Zeng, C. Liang, X. Bai, D. Chen, W. Wang, InkGAN: Generative Adversarial Networks for Ink-And-Wash Style Transfer of Photographs, 2023.

[12] Q. Wu, P. Wang, C. Shen, I. Reid, A. van den Hengel, Are you talking to me? reasoned visual dialog generation through adversarial learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6106–6115.

[13] D. Guo, C. Xu, D. Tao, Image-question-answer synergistic network for visual dialog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10434–10443.

[14] C. Le, C. Hetang, A. Cao, Y. He, Euclidreamer: fast and high-quality texturing for 3d models with stable diffusion depth, arXiv preprint arXiv:2311.15573 (2023).

[15] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural module networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 39–48.

[16] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, Inferring and executing programs for visual reasoning, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2989–2998.

[17] C. Hetang, Y. Wang, Novel view synthesis from a single rgbd image for indoor scenes, arXiv preprint arXiv:2311.01065 (2023).

[18] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko, Learning to reason: End-to-end module networks for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 804–813.

[19] F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, O. Pietquin, End-to-end optimization of goal-driven and visually grounded dialogue systems, 2017, arXiv preprint arXiv:1703.05423.

[20] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, A. van den Hengel, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3674–3683.

[21] P. Wang, Q. Wu, C. Shen, A. Dick, A. van den Hengel, Fvqa: Fact-based visual question answering, IEEE Trans. Pattern Anal. Mach. Intell. 40 (10) (2018) 2413–2427.

[22] W. Wang, et al., Sentiment analysis: a systematic case study with yelp scores, Advances in Artificial Intelligence and Machine Learning 3 (3) (2023) 74.

[23] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, M. Zhou, Visual question generation as dual task of visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6116–6124.

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

[25] M. Charikar, K. Chen, M. Farach-Colton, Finding frequent items in data streams, in: International Colloquium on Automata, Languages, and Programming, Springer, 2002, pp. 693–703.

[26] K.J. Shih, S. Singh, D. Hoiem, Where to look: Focus regions for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4613–4621.

[27] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.

[28] P.H. Seo, A. Lehrmann, B. Han, L. Sigal, Visual reference resolution using attention memory for visual dialog, in: Advances in Neural Information Processing Systems, 2017, pp. 3719–3729.

[29] G.-C. Kang, J. Lim, B.-T. Zhang, Dual attention networks for visual reference resolution in visual dialog, 2019, arXiv preprint arXiv:1902.09368.

[30] S. Kottur, J.M. Moura, D. Parikh, D. Batra, M. Rohrbach, Visual coreference resolution in visual dialog using neural module networks, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 153–169.

[31] C. Hetang, Autonomous path generation with path optimization, Google Patents, 2022, US Patent App. 17/349, 450.

[32] X. Yang, et al., Linguistically-inspired neural coreference resolution, Advances in Artificial Intelligence and Machine Learning 3 (2) (2023) 66.

[33] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[34] C. Hetang, Impression network for video object detection, in: 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence, Vol. 3, ICIBA, 2023, pp. 735–743.

[35] H. Tangcongrui, Q. Hongwei, Methods and apparatuses for recognizing video and training, electronic device and medium, Google Patents, 2021, US Patent 10, 909, 380.

[36] Z. Longxiang, W. Wenping, Y. Keyi, H. Jingxian, L. Qi, X. Haoru, H. Congrui, Sliding-BERT: Striding towards conversational machine comprehension in long context, Adv. Artif. Intell. Mach. Learn. 3 (2023).

[37] R. Thibaux, D.H. Silver, C. Hetang, Stop Location Change Detection, Google Patents, 2022, US Patent App. 17/131, 232.

[38] C. Hetang, Y. Shen, Y. Zhou, J. Gao, Implementing synthetic scenes for autonomous vehicles, Google Patents, 2022, US Patent App. 17/349, 489.

[39] C. Hetang, N. Zhang, Autonomous vehicle driving path label generation for machine learning models, Google Patents, 2023, US Patent App. 17/740, 215.

[40] N. Pham, R. Pagh, Fast and scalable polynomial kernels via explicit feature maps, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 239–247.